On the Possibility of Predicting Gaze Aversion to Improve Video-Chat Efficiency

Nigel G. Ward*, Chelsey N. Jurado, Ricardo A. Garcia, Florencia A. Ramos
University of Texas at El Paso

Abstract

A possible way to make video chat more efficient is to only send video frames that are likely to be looked at by the remote participant. Gaze in dialog is intimately tied to dialog states and behaviors, so prediction of such times should be possible. To investigate, we collected data on both participants in 6 video-chat sessions, totalling 65 minutes, and created a model to predict whether a participant will be looking at the screen 300 milliseconds in the future, based on prosodic and gaze information available at the other side. A simple predictor had a precision of 42% at the equal error rate. While this is probably not good enough to be useful, improved performance should be readily achievable.

Keywords: eye gaze, video conferencing, behavior modeling, dialog patterns, conversation

Concepts: •**Information systems** → **Web conferencing;** •**Human-centered computing** → *User models;*

1 Motivation

The principal challenge in telecommunications is maximizing the quality of experience that can be delivered for a given bandwidth. Real-time adaptation can help: channels where instantaneous quality varies over time can support a better overall user experience than constant-rate transmission. To date, adaptation is generally based on the network state. We propose instead to adapt to the state of the user, in particular to the user's attention level.

Specifically, we would like to make video chat more efficient. Video chat is popular and a significant data hog, especially for mobile users [Jana et al. 2013]. Like other forms of interactive telecommunication, it is especially resource-demanding because it is impossible to buffer significantly. Delays and packet losses are very noticeable, and to avoid these carriers generally overprovision bandwidth. This involves significant costs for carriers and, indirectly, users. These resource requirements preclude use entirely in many situations.

We would like to make video chat more efficient by transmitting less information for video frames that the remote user is unlikely to look at, exploiting the fact that people often look away (avert gaze) during dialog. A video communication channel should be able to use less bandwidth at such times without much impact on perceived quality. This paper explores the feasibility of this idea.

Section 2 overviews related research and describes two preliminary studies on gaze and attention. Section 3 describes the data collection, 4 the features and model, and 5 the results. Section 6 discusses the implications, Section 7 notes directions for future work and 8 summarizes.

*e-mail: nigelward@acm.org

ETRA '16,, March 14-17, 2016, Charleston, SC, USA
ISBN: 978-1-4503-4125-7/16/03
DOI: http://dx.doi.org/10.1145/2857491.2857497

2 Background and Preliminary Studies

It has long been known that people in dyadic conversation often gaze away from their interlocutors [Kendon 1967], with reported percentages varying from 10% to 91%, depending on the individual and other factors. In dyadic conversation, work studying the relationship between gaze, dialog activities, and prosody has shown that gaze aversion is not randomly distributed, but relates to turn-taking behavior [Cummins 2012; Kawahara et al. 2012; Jokinen et al. 2013]. For example, in one study, 73% of gaze aversions were turn-initial, with an average duration of 2.3 seconds [Andrist et al. 2013]. Turn-taking behavior in turn relates to the prosody of the utterances of the participants [Gravano and Hirschberg 2011; Raux and Eskenazi 2012], suggesting that we may be able to predict gaze from prosody.

For telecommunications applications, there have been studies of gaze in multi-party scenarios regarding the question of who a participant will be looking at [Vertegaal et al. 2001]. More detailed gaze modeling has also been done, but so far only for non-interactive video [Komogortsev 2009; Yonetani 2012; Feng et al. 2013; Kim and Kim 2013].

As a preliminary study, we examined the human ability to predict gaze. For this we made six over-the-shoulder videos of people Skyping, in a configuration with a one-way delay of about 600 milliseconds. We then had four untrained observers watch these using Elan, stopped the video at random points, and had them predict whether the remote participant would be looking at the local one 350 ms later. We thus obtained 250 informal predictions. 72% of the predictions of an upcoming gaze aversion were correct, and overall only 10% of their predictions were false positives. Thus gaze appears to be fairly predictable.

We also did a second preliminary study, to investigate the relationship between gaze aversion and the value of the displayed image to the user. In particular, since the focus of attention is not always at the gaze point, we need an estimate of how well gaze can serve as a proxy for attention. While aspects of visual acuity across the visual field have been well-studied, we are concerned specifically with viewing faces, and with sensitivity during conversation rather than in single-task experiments. We therefore investigated the relation between gaze direction and the likelihood of noticing video impairments. In this two participants communicated over Skype, and occasionally one closed his or her eyes for about a second. These stimuli were cued by a silent random-interval timer visible to that one participant. Immediately afterwards the stimulus-generating participant said "freeze" and the other froze his or her gaze and reported the location, to within 10 degrees, by naming the nearest one of a grid of labeled dots on the screen and wall. They also reported whether or not they had noticed the closing of the eyes.

With 6 participants and 150 total trials, we found, as expected, that closings were noticed less to the extent that gaze was farther from the center of the screen. In particular, when the reported gaze was greater than 15 degrees away, horizontally or vertically, from the screen center, the stimulus detection rate was 13%, far less than the

Figure 1: *Participant's view, with webcam above and gaze tracker below the screen.*

89% when gaze was closer, let alone the 93% when gaze was within 5 degrees of the center. Thus reducing video quality at gaze-away times should have only a minor impact on users.

3 Data Collection

To investigate predictability we needed data. Since gaze behavior in telecommunications differs from that in face-to-face dialog [Doherty-Sneddon and Phelps 2005] (for reasons probably including the lack of physical presence, poor image and audio quality, and degraded contingency due to delay), we created a new data collection. We set up a Skype connection between two quiet rooms in our laboratory. For each participant the Skype window filled the screen. Each sat at a comfortable distance and adjusted the gaze tracker appropriately. We asked them to not move too much because of the gaze tracker limitations. The image of the partner's face on the screen typically subtended around 7 degrees horizontally, as suggested by Figure 1.

Each pair of participants had a 4-12 minute Skype conversation, on any topics they wished to talk about. Each pair was already acquainted and many were friends. All were affiliated with the UTEP Computer Science department. There were 12 participants and 6 conversations, totaling 65 minutes.

Each participant's gaze was tracked by a Gazepoint GP3 Eye Tracker mounted below the monitor. Audio was recorded with head-mounted microphones with a Tascam DR-40. We recorded reference signals consisting of three claps in front of the gaze tracker before the conversation and again afterwards, for each side. Synchronizing with these gave a global, delay-free view of the behaviors. This was not, of course, what the participants experienced. In Skype the delay is not constant, and audio and video delays may diverge. In this setup the video delays are around 120 to 150 ms, and the audio 200 to 250 ms. Video freezes were rare.

We counted as "gaze-off" those frames where either the gaze tracker reported no gaze detected or when the gaze was away from the interlocutor's face. The former were probably mostly hard-to-predict blinks, and the latter more predicable gaze aversions. "Away from the face" was operationalized using two heuristics. First, we estimated the center of interest as the average x and y of the on-screen fixations. Second, because gaze distance from center was bimodal, with a minimum around 0.7 screen heights, and based on the informal perceptual study mentioned above, gaze aversion was inferred when the current gaze direction was more than $0.7h$ away from this center of interest, where h is the screen height. This corresponds to about 14 degrees of visual angle.

Using these definitions, the fraction of gaze-off frames averaged 29%, varying across participants from 12% to 45%.

4 Predictive Features

Given this data we investigated features that correlate with future gaze. In order to decide whether or not to send a packet at time t, we need to predict whether the remote user will be looking at the screen at time $t + d$, where d is the one-way delay. There is no need to predict the exact moments of eye movement, just the likelihood of gaze-off. Delay implies that the prediction algorithm will not have access to any recipient-side features more recent than $t - d$. Here we assume d = 300 ms, and accordingly consider only features that would be available at the local, sending, end when the prediction must be made.

We first examined prosodic features. As noted above, gaze is related to turn-taking; we have also observed that participants tend to look away while laughing, when back-channeling, when closing out a topic, when thinking, and during affect bursts. Accordingly the prosodic feature set was designed to capture information useful for turn-taking, dialog-event, and user-state prediction [Shriberg et al. 2005; Truong and van Leeuwen 2007; Ward et al. 2011; Ward and Vega 2012; Ward 2015]. Our features were intensity, pitch height, pitch range, pitch creakiness, and a speaking-rate proxy. Since we expected gaze to relate to patterns of behavior across time, we used multiple windows for each feature, together spanning the three seconds up to the prediction point. Thus we had , for example, a window for the average interlocutor volume over 1900 to 1100 ms before the prediction point t, a window for the average value over 1100 to 600 ms before, and so on. We included such features for both participants.

Examining correlations, speaking rate was negatively correlated (–0.10) with upcoming gaze, perhaps because at turn-ends speakers tend both to slow down and to look at the interlocutor. Interlocutor volume and speaking rate correlated positively with upcoming gaze, perhaps because increases in rate and volume tend to signify important information and attract attention.

We next examined gaze-location features. We used similarly coarse windows and again included features for both participants. For example, it is reported that at turn end the current speaker usually looks directly at the interlocutor, and that at the subsequent turn start the new speaker looks away. Since gaze aversions in different directions may reflect different cognitive activities [Ehrlichman and Weinberger 1978], we included features for gaze-up, gaze-down, gaze-right, and gaze-left. For convenient use in our linear model, explained below, we thresholded these so that, for example, the gaze-up-distance feature value was zero when the gaze was below the center point. We also included features for gaze-distance-from-center, again thresholded, and for the fraction of the time in the window when there was gaze-on.

The mostly highly correlated feature was, unsurprisingly, the fraction of gaze-on by the speaker during the most recent available 100 ms window, namely that 700 to 600 ms before the frame to predict, with a correlation of 0.27. Gaze-away features correlated negatively, in order right $>$ down $>$ left $>$ up; perhaps because gaze aversions to the right tend to last longer. Interlocutor gaze features correlated in the same ways, although more weakly, probably reflecting the fact that gaze tends to be reciprocated. Overall, gaze features were far more informative than the prosodic features.

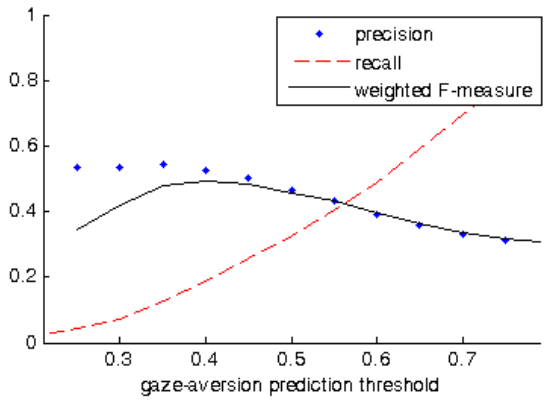

Figure 2: *Predictor Performance*

Gaze:	predicted on	predicted off	sum
actually on	67%	4%	71%
actually off	23%	6%	29%
sum	90%	10%	100%

Table 1: *Percent of Frames in Each Condition*

5 Prediction Results

As this is initial exploration, we wanted a simple model whose performance would be easy to analyze, so we chose to predict using linear regression with a threshold. Varying the threshold changes the aggressiveness of the gaze-off predictions, and thus the tradeoff between reducing transmission and preserving quality. The input was 42 gaze features and 67 prosodic features.

A gaze-off prediction is correct, of course, if it matches the actual gaze state in the data 300 ms later. We evaluated our predictor by generating predictions every 10 ms, 707830 in total. We tested, for each participant's data, the performance of a model trained on the other 11 participants. Thus each test was of the predictability of a participant unseen in the training data. The results reported are averages across all participants.

Figure 2 shows how the precision, recall, and weighted F-measure varied with the prediction threshold. The equal-error rate was 42% that is, there was an operating point at which the fraction of the gaze-off frames correctly predicted as such was 42%, and so was the fraction of the gaze-off predictions that were correct. This precision is significantly above the 29% baseline which would be obtained by random guessing. For the F-measure we weighted the precision 30 times the recall, for reasons discussed below. Varying the threshold, the best value for this measure was 49%. The best operating point varied across speakers, as did the best performance obtained, from 18% to 86%.

Examining more closely the performance, at the best threshold the average precision was 53% and the average recall 19%. The confusion matrix at this operating point is shown in Table 1.

We had hoped to obtain decent performance using prosodic features alone, as they are easy to obtain with any current hardware. However precision for feature sets without remote-speaker gaze was barely above chance.

6 Cost-Benefit Analysis

While clearly our method is able to reduce transmission rates, incorrect predictions will, from the recipient's perspective, manifest as a noticeable loss of quality. This section considers when this tradeoff is likely to be advantageous.

Unfortunately, while there are good quality-of-experience models for interactive audio and for video transmission, there are no suitable ones for video chat. User behaviors and user preferences in interactive applications such as video chat and teleconferencing clearly differ from those in passive video viewing [Whittaker and O'Conaill 1997; Bohannon et al. 2012]. Moreover, different users have different preferences, especially different cost sensitivities, and video chat can be implemented and delivered in different ways [Cermak 2009; Yu et al. 2014].

Nevertheless there is information in the literature from which we can attempt to roughly characterize the tradeoff. The primary determinants of subjective video quality are packet loss and bitrate (the product of frame rate and frame quality) [Cermak 2009]. The benefit of gaze-aversion modeling is a reduction in unneeded transmission, and this can be used to increase the bitrate at times when transmission *is* needed. A 30% increase in bitrate appears to generally provide a Mean-Opinion-Score (MOS) increase of about 0.3 points, according to the graphs in [Cermak 2009]. A 1% increase in packet loss appears to generally cause a MOS decrease of about 0.3 points, according to the graphs in [Khorsandroo et al. 2012]. The ratio of these two is the weight of 30 used above.

From this we can estimate the utility of a system operating at the best threshold identified above. At this point 10% of the frames are being predicted as gaze-off, implying that, if transmission were suspended entirely at such times, the transmission reduction will be 10%. This means that the bitrate in times when transmission is needed could be increased by 11% without increasing the overall bitrate. This would indicate a MOS benefit of 0.12. At this same point, due to false negatives, 4% of the total frames are actually gaze-on frames but incorrectly not sent. The number of useful frames sent is therefore, using the noticing ratios from our preliminary study, 89% * gaze-on-frames-correctly-sent + 13% * gaze-off-frames-incorrectly-sent, or 65.5%. In comparison, the number of useful frames if all are sent is 67.1%. Thus there is a 2.4% reduction in the quantity of useful frames, equivalent from the user's perspective to a 2.4% increase in packet loss, which suggests a MOS decrement of 0.72 points, far outweighing the benefit. This estimate suggests that the performance of our current predictor is far from sufficient.

7 Future Work

Direct experimental evaluation of the potential of this technique is needed. Given the widespread acceptance of video-chat systems that periodically freeze, it is possible that the calculation above significantly underestimates the potential value of this method.

Nevertheless, it seems certain that much better gaze prediction will be needed to make gaze-adaptive transmission useful for video chat. This may be achievable. One might try better features, especially those supporting fine-grained modeling of gaze dynamics [Komogortsev et al. 2009; Kim and Kim 2013]. One might also try image features and features of the words spoken, although robustness and computational-complexity considerations may limit the utility of these in practice. Given the large individual differences, one might try speaker-specific modeling. Instead of trying to predict all gaze-off frames with one model, one might try a composite model that fuses the outputs of individual models trained to predict different gaze events, such as blinks, aversions while thinking, and

aversions while laughing. One might also try models conditioned on gender, age, relative status, dialog type, personality, dialect and language. Finally, future work should obviously use more training data and a better model.

Other types of attention prediction might also be explored. Rather than merely predicting gaze on/off, one might try to predict the actual gaze location, for example to the eyes or the mouth, so as to devote higher resolution to the predicted current region of interest. Also, while our data was collected in a distraction-free situation, people also video chat while reading email, playing games, talking to side participants, cooking, walking, and so on. Prediction of attention in non-distraction-free environments may be much easier.

8 Summary

This paper has explored the possibility of using a predictive model of video-chat participant gaze to reduce pointless transmission. It identified relevant dialog activities and states, found good predictive features, reported the first study of the predictability of gaze in dialog, and identified avenues for improvement.

Acknowledgements

We thank Victoria Bravo, Saiful Abu, David G. Novick, and Olac Fuentes. This work was supported in part by the National Science Foundation under project IIS-1241434 and by REU supplements to IIS-1914868 and IIS-1449093, and by the University of Texas System under a UT Transform award.

References

ANDRIST, S., MUTLU, B., AND GLEICHER, M. 2013. Conversational gaze aversion for virtual agents. In *Proceedings of Intelligent Virtual Agents (IVA)*.

BOHANNON, L. S., HERBERT, A. M., PELZ, J. B., AND RANTANEN, E. M. 2012. Eye contact and video-mediated communication: A review. *Displays*.

CERMAK, G. W. 2009. Subjective video quality as a function of bit rate frame rate, packet loss, and codec. In *Int'l Workshop on Quality of Multimedia Experience (QoMEx).*, IEEE, 41–46.

CUMMINS, F. 2012. Gaze and blinking in dyadic conversation: A study in coordinated behaviour among individuals. *Language and Cognitive Processes 27*, 10, 1525–1549.

DOHERTY-SNEDDON, G., AND PHELPS, F. G. 2005. Gaze aversion: A response to cognitive or social difficulty? *Memory & Cognition 33*, 4, 727–733.

EHRLICHMAN, H., AND WEINBERGER, A. 1978. Lateral eye movements and hemispheric asymmetry: a critical review. *Psychological Bulletin 85*, 5, 1080.

FENG, Y., CHEUNG, G., TAN, W.-T., LE CALLET, P., AND JI, Y. 2013. Low-cost eye gaze prediction system for interactive networked video streaming. *IEEE Transactions on Multimedia*.

GRAVANO, A., AND HIRSCHBERG, J. 2011. Turn-taking cues in task-oriented dialogue. *Computer Speech and Language 25*, 601–634.

JANA, S., PANDE, A., CHAN, A., AND MOHAPATRA, P. 2013. Mobile video chat: issues and challenges. *IEEE Communications Magazine 51*, 6.

JOKINEN, K., FURUKAWA, H., NISHIDA, M., AND YAMAMOTO, S. 2013. Gaze and turn-taking behavior in casual conversational interactions. *ACM Transactions on Interactive Intelligent Systems (TiiS) 3*, 12.

KAWAHARA, T., IWATATE, T., AND TAKANASHI, K. 2012. Prediction of turn-taking by combining prosodic and eye-gaze information in poster conversations. In *Interspeech*.

KENDON, A. 1967. Some functions of gaze-direction in social interaction. *Acta Psychologica 26*, 22–63.

KHORSANDROO, S., NOOR, R. M., AND KHORSANDROO, S. 2012. A generic quantitative relationship between quality of experience and packet loss in video streaming services. In *Fourth International Conference on Ubiquitous and Future Networks (ICUFN)*, IEEE, 352–356.

KIM, J.-W., AND KIM, J.-O. 2013. Video viewer state estimation using gaze tracking and video content analysis. In *Visual Communications and Image Processing (VCIP), 2013*, IEEE, 1–6.

KOMOGORTSEV, O. V., RYU, Y. S., MARCOS, S., AND KOH, D. H. 2009. Quick models for saccade amplitude prediction. *Journal of Eye Movement Research 3*.

KOMOGORTSEV, O. V. 2009. Gaze-contingent video compression with targeted gaze containment performance. *Journal of Electronic Imaging 18*, 3, 1–10.

RAUX, A., AND ESKENAZI, M. 2012. Optimizing the turn-taking behavior of task-oriented spoken dialog systems. *ACM Transactions on Speech and Language Processing (TSLP) 9*, 1.

SHRIBERG, E., FERRER, L., KAJAREKAR, S., VENKATARAMAN, A., AND STOLCKE, A. 2005. Modeling prosodic feature sequences for speaker recognition. *Speech Communication 46*, 455–472.

TRUONG, K. P., AND VAN LEEUWEN, D. A. 2007. Automatic discrimination between laughter and speech. *Speech Communication 49*, 144–158.

VERTEGAAL, R., SLAGTER, R., VAN DER VEER, G., AND NIJHOLT, A. 2001. Eye gaze patterns in conversations: there is more to conversational agents than meets the eyes. In *SIGCHI*, ACM, 301–308.

WARD, N. G., AND VEGA, A. 2012. A bottom-up exploration of the dimensions of dialog state in spoken interaction. In *13th Annual SIGdial Meeting on Discourse and Dialogue*.

WARD, N. G., VEGA, A., AND BAUMANN, T. 2011. Prosodic and temporal features for language modeling for dialog. *Speech Communication 54*, 161–174.

WARD, N. G. 2015. Midlevel prosodic features toolkit. http://www.cs.utep.edu/nigel/midlevel/, https://github.com/nigelgward/midlevel.

WHITTAKER, S., AND O'CONAILL, B. 1997. The role of vision in face-to-face and mediated communication. In *Video-Mediated Communication*, K. E. Finn, A. J. Sellen, and S. B. Wilbur, Eds. Lawrence Erlbaum Associates.

YONETANI, R. 2012. Modeling video viewing behaviors for viewer state estimation. In *Proceedings of the 20th ACM international conference on Multimedia*, 1393–1396.

YU, C., XU, Y., LIU, B., AND LIU, Y. 2014. Can you SEE me now?, a measurement study of mobile video calls. In *Proceedings of IEEE Conference on Computer and Communications (INFOCOM)*.