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

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.

I. MOTIVATION

The principal challenge in telecommunications is maximizing the quality of experience that can be delivered for a given bandwidth. 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 [1]. We would like to make it 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. This paper explores the feasibility of this idea.

II. BACKGROUND AND PRELIMINARY STUDIES

It has long been known that people in dyadic conversation often gaze away from their interlocutors [2], 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 [3], [4], [5] has shown that gaze aversion is not randomly distributed, but relates to turn-taking behavior. For example, in one study, 73% of gaze aversions were turn-initial, with an average duration of 2.3 seconds [6]. Turn-taking behavior in turn relates to the prosody of the utterances of the participants [7], [8], suggesting that we may be able to predict gaze from prosody.

For telecommunications applications, there have been studies of gaze in multi-party scenarios and the question of who a participant will be looking at. More detailed gaze modeling has also been done, but so far only for non-interactive video [9], [10], [11], [12].

We did a preliminary study of the human ability to predict gaze. 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, stopping the video at random points, and asking them to 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 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 normal conversation, rather than in single-task experiments. We therefore investigated the relation between gaze direction and the likelihood of noticing video impairments, again informally. 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.
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 predictable gaze aversions. “Away from the face” was determined using two heuristics. First, since the participants were free to move and because we did not capture images, we estimated the center of interest as the average x and y of the on-screen fixations. Second, 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%.

IV. 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. Delay implies, further, 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 and dialog-event prediction [14], [15], [16]. Our features were volume, pitch height, pitch range, pitch creakiness, and a speaking-rate proxy. Since we thought gaze might relate to patterns of behavior across time, we used feature averages over adjacent windows, for example, the average interlocutor volume over 1900 to 1100 ms before the prediction point, 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 of turn-end behavior, where speakers known to both tend to slow down and to tend 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, as there is clearly some interplay. 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. For the remote participant, whose gaze we were predicting, we used no features more recent than 300 ms in the past. Since gaze aversions in different directions may reflect different cognitive activities [17], we
included features to encode this information. 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. Our features were gaze-up, gaze-down, gaze-right, gaze-left, gaze-distance-from-center, and gaze-on fraction.

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, perhaps reflecting the fact that gaze tends to be reciprocated. Overall gaze features were far more predictive than the prosodic features.

V. PREDICTION RESULTS

As this is initial exploration, we chose to predict using a very simple model, linear regression with a threshold, as we wanted something easy to analyze. The input was 42 gaze features and 67 prosodic features. Varying the threshold makes it predict likely gaze-off moments more or less aggressively, thereby varying the trade-off between reducing transmission and preserving quality.

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. This we did by testing, 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 30% baseline which would be obtained by random guessing. For the F-measure we weighted the precision 30 times the recall, for reasons discussed below. The best-point value for this measure was 49%. Best-point performance varied greatly across speakers, from 18% to 86%.

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

<table>
<thead>
<tr>
<th>Gaze:</th>
<th>actually on</th>
<th>predicted on</th>
<th>actually off</th>
<th>predicted off</th>
<th>sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>predicted on</td>
<td>67%</td>
<td>4%</td>
<td>71%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>predicted off</td>
<td>23%</td>
<td>6%</td>
<td>29%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sum</td>
<td>90%</td>
<td>10%</td>
<td>100%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

TABLE I
PERCENT OF FRAMES IN EACH CONDITION

It would be convenient if decent performance could be obtained with prosodic features alone, as they are easy to compute with any current hardware, or without using remote-speaker features. However precision for feature sets without remote-speaker gaze was barely above chance.

VI. 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 video transmission and for interactive audio, there are no suitable ones for video chat. Most quality modeling has been done for passive video viewing, but user behaviors and user preferences clearly differ for interactive applications such as video chat and teleconferencing [18], [19]. Moreover, different users have different preferences, especially different cost sensitivities, and video chat can be implemented and delivered in different ways [20], [21].

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) [20]. 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 [20]. A 1% increase in packet loss appears to generally cause a MOS decrease of about 0.3 points, according to the graphs in [22]. 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 point 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 can 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 received 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.

VII. 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 estimation based on other quality models, as done 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. Better features, especially those supporting fine-grained modeling of gaze dynamics [23], [12], should be tried. In addition image features and the words spoken may be helpful, although robustness and computational-complexity considerations may limit their utility in practice. Given the large individual differences, speaker-specific modeling is also likely to greatly improve results. Instead of trying to predict all gaze-off frames with one model, one might try composite models that fuse the outputs of individual models trained to predict different gaze events, such as blinks, aversion while thinking, and aversion while laughing. Models could also be conditioned on gender, age, relative status, dialog type, personality, dialect and language. Finally, future work should obviously use more training data and better machine-learning models.

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 image 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.

VIII. 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 states and predictive features, reported the first study of the predictability of gaze in dialog, and identified avenues for improvement.

ACKNOWLEDGMENT

[omitted for anonymous submission]