Speaker Prediction based on Head Orientations

An Evaluation of Machine Learning and Human Performance

Rutger Rienks RIENKS@EWI.UTWENTE.NL
Ronald Poppe POPPE@EWI.UTWENTE.NL
Mannes Poel MPOEL@EWI.UTWENTE.NL
Human Media Interaction Group, Department of Electrical Engineering, Mathematics and Computer Science
University of Twente, PO Box 217, 7500 AE, Enschede, The Netherlands

Abstract

To gain insight into gaze behavior in meetings, this paper compares the results from a Naive Bayes classifier, Neural Networks and humans on speaker prediction in four-person meetings given solely the azimuth head angles. The Naive Bayes classifier scored 69.4% correctly, Neural Networks 62.3% and humans only 37.7%. None of the classifiers was able to generalize over meetings. We show that there are strong indications that human specific gaze behavior influences the fact that the models do not generalize. Additionally, we show that for all classifiers the performance of the prediction in the beginning and at the end of a speaker turn is worse than halfway through the speaker turn.

1. Introduction

Nowadays a lot of research is being done on machine processing and interpretation of signals from people or other elements present in an observed environment. Smart rooms are widely used as test environments for research in this area. Sensors in the room can capture information which can be fused, manipulated and augmented in order to achieve interpretation at desired levels (Reidsma et al., 2004).

Currently smart meeting rooms are being used for all kinds of research purposes in e.g. the AMI project, the CMU meeting Room Project and the NIST Meeting Room project. One can think of a smart meeting room with a system that creates notes by understanding speech, creates summaries (Mani & Maybury, 1999), or automatically switches microphones on and off. EasyMeeting (Chen & Perich, 2004) is an example of a system that can provide relevant services and information to meeting participants.

State of the art in computer graphics and animation of embodied agents allows us to build quite realistic 3D virtual environments in which real humans can meet virtual human-like avatars (Vilhjalmsson & Cassell, 1998; Nijholt, 2004). These virtual environments can be used as a test bed for data visualization and for studying human perception and interpretation of meeting situations.

In this paper we examine the performance of humans and machine learning techniques on the task of speaker prediction from horizontal head orientation angles (azimuth). Gaze is not solely determined by the head orientation but also by the direction of eye gaze. It has, however, been shown that there is a high correlation between gaze and head orientation (more than 85%) (Stiefelhagen, 2002).

2. Research aim

Within a meeting context, our aim is to gain insight into the nature of human gaze behavior. We can use this knowledge for both generation of gaze behavior in virtual meetings and the extraction of useful information from gaze behavior such as the speaker, the addressee and the focus of attention.

Imagine a virtual chairman who understands the meeting sufficiently to be able to structure the meeting according to an agenda by giving appropriate turns, interrupting if someone speaks for too long and keeping participants focused when they seem distracted. This might sound far-fetched, but when a machine understands where someone is looking during a meeting, it might conclude based on its models that the person is focused on the object in line with the head orientation. If a non-speaker is always looking at the ceiling instead of at the speaker, this might reveal something about his attention level or even about the personality of that person.

Autonomous agents can use the derived gaze behavior models in combination with their own beliefs, desires, intentions and emotions (Wright, 1997) to become ‘aware’ of the perceived situation and even to act according to the models. Current applications of such models lead to increased appreciation of such agents (van Es et al., 2002), since they become more lively. This directly improves remote communication (Garau et al., 2001).

2.1. Gaze behavior

In general it is believed that gaze can bear a conversational function. When someone is looked at, the person who is looking might expect a reaction from the other, either visual or vocal. According to Kendon (Kendon, 1967) gaze serves four functions: visual feedback, regulation of conversation flow, communication of emotions and relationships, and improvement of concentration by restricting visual input.

Argyle et al. (Argyle et al., 1973) define six almost similar categories: information seeking, signaling, controlling the synchronization of speech, mutual gaze and intimacy, avoiding undue intimacy, and avoiding excess input of information. We are especially interested in examining the information seeking and the conversational flow regulation. While speaking a person emits information so we expect the listeners to look at the speaker, seeking for information.

Research showed that people gaze nearly twice as much at others while listening (75%) than while speaking (41%) (Argyle & Cook, 1976). In the case of a single speaker, all listeners would be focussing more in the direction of the speaker than the speaker is focussing at all of them. In accordance with this, Vertegaal found that, in a setting with three persons, people gaze much more at the speaker (62.4%) than at others (8.5%) (Vertegaal, 1998).

Assuming this, we expect head orientations of persons in a meeting to be good indicators for speaker identification. The remainder of this paper addresses this issue. First we describe the data collection. Then we show how well machine learning techniques are able to predict the speaker amongst four meeting participants, followed by a discussion how humans performed on this task. Finally, we compare the results for both approaches and elaborate on these results revealing more insight in gaze behavior.

3. Data collection

We used three four-person meetings with a total duration of 21 minutes that were recorded in the IDIAP smart room. Apart from the video and audio recordings, we recorded head position and orientation for all meeting participants. Flock of Bird sensors were used to accurately measure position and orientation at a rate of 50 Hz. The sensor is a small box and when mounted on top of a participant’s head is not obtrusive and does not cause any distraction, as can be seen in close-ups in Fig. 1.

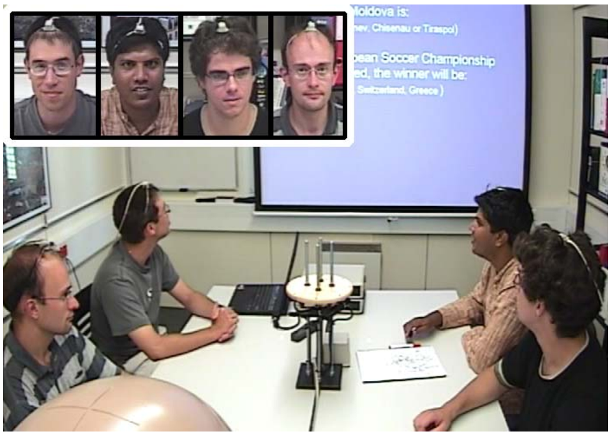

Figure 1. The setting of meeting 6, with close-ups of participants

The non-scripted meetings contain lively discussions about pre-formulated statements. To make the experiment more realistic we also incorporated a whiteboard, where statements were shown.

After recording, we analyzed both video and orientation data to discover possible biases due to incorrect mounting of the Flock sensor on the head. We corrected the orientation data for these biases. For simplicity reasons, we only used azimuth data, see Fig. 2. Since all participants reside in the same plane, parallel to the table surface, we expect that azimuth orientation contains the most relevant information.

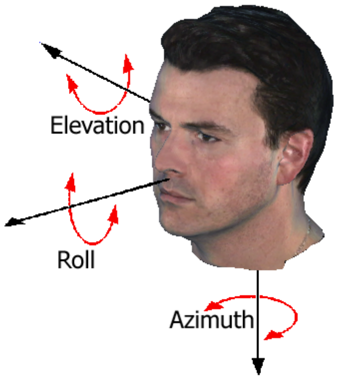

Figure 2. Azimuth, elevation and roll angles for heads

Furthermore, all occurrences with non-speech (laughter, silence, etc.) or with speech overlap were removed from the data set. The number of frames for each meeting, including *a priori* probabilities for each meeting and for all meetings, can be found in Table 1.

	M1	M2	M3	Total
Samples	11333	13078	28148	52559
A priori SP1	40.4%	26.9%	24.8%	28.7%
A priori SP2	27.3%	23.4%	9.8%	16.9%
A priori SP3	7.7%	8.6%	29.4%	19.5%
A priori SP4	24.7%	41.2%	36.0%	34.9%

Table 1. Number of samples and a priori probabilities of speakers in each meeting and all meetings. *M1* is meeting 1, *SP1* corresponds to speaker 1.

Trained	Test	ND	D
M1	M1	63.0%	82.8 %
M2	M2	51.3%	90.0 %
M3	M3	54.5%	76.6 %
M1, M2 & M3	M1, M2 & M3	50.0%	69.4 %
M1 & M2	M3	39.5%	35.4%
M1 & M3	M2	38.8%	33.5%
M2 & M3	M1	45.6%	40.4%
M1	M2 & M3	34.5%	36.5%
M2	M1 & M3	37.5%	32.7%
M3	M1 & M2	42.5%	32.3%

Table 2. Classification results for the Naive Bayes classifier without (ND) and with (D) discretization

4. Machine learning performance

In this section we compare the prediction results of Naive Bayes classifiers and Neural Networks for the task of speaker identification.

4.1. Data representation

For our machine learning experiment, we used input vectors $\mathbf{v}_t = (\alpha(1)_t, \alpha(2)_t, \alpha(3)_t, \alpha(4)_t)$, sampled at time t. Each $\alpha(i)_t$ corresponds to the azimuth angle of person i at time t. We used batch training, during which we also presented the speaker SP_t at time t, where $SP_t \in \{1, 2, 3, 4\}$.

In both experiments four series of tests were performed, each test having a different composition of training and test sets. In a first series both training and test set were obtained from a single meeting. In series two, both training and testing were performed on samples from all three meetings. In the third series, we trained on two meetings and tested on the third. Finally, we trained on a single meeting and tested on the other two. We discuss the results for the two machine learning techniques below.

4.2. Naive Bayes classifier

We trained a Naive Bayes classifier with and without supervised discretisation (Yang & Webb, 2003; Dougherty et al., 1995) using gaze vectors. We conducted the test for three different meetings using tenfold cross validation. The results for the four series are shown in Table 2. The discretized data is shown in the column *D*, the original not discretized data is shown in the column *ND*.

From the table it can be seen that within a single meeting the classifier performs quite well. However when our training and test sets are taken from different meetings the performance drops significantly. Discretization improves our results when we test on samples from the meeting on which we trained and it decreases our results when we test on samples from other meetings than those trained on. Since the discretization algorithm is supervised, the bins created by the algorithm when trained on a particular meeting do not apply for samples from a different meeting. This results in a worse instead of an increased performance.

4.3. Neural Networks

We also used Neural Networks to estimate the speaker from the azimuth angle data. The Levenberg-Marquardt algorithm (Moré, 1978) was used for training. We performed the same four series of tests used for the Naive Bayes classifier.

In each series, we experimented with different numbers of neurons in the hidden layer. In the first two series, 25 neurons were found to yield best results, in the third series we used 15 neurons and in the last series only 5 neurons were used. In the first two series, the data was divided into a training set (60%), a test set (20%) and a validation set (20%). In the last two series, the training set contained all samples from the training meeting. The validation set contained 20% of the test meeting samples. The test set contained the other 80%. In each test, 5 runs were performed and the Neural Network with the best performance on the validation set was used to obtain the test performance. In Table 3, these results are summarized.

Again, we see that when training and test sets are sampled from the same meetings, the performance is high. The results however, do not generalize over meetings.

Training	Test	Result
M1	M1	82.6%
M2	M2	81.3%
M3	M3	72.3%
M1, M2 and M3	M1, M2 and M3	62.9%
M1 and M2	M3	44.2%
M1 and M3	M2	43.7%
M2 and M3	M1	48.1%
M1	M2 and M3	38.2%
M2	M1 and M3	40.2%
M3	M1 and M2	42.4%

Table 3. Classification results for Neural Networks

5. Human performance

In this section we describe how we tested the human performance on predicting speaker turns based only on head orientations. The main problem is that interpreting the numerical $\mathbf{v}_t$ vector is hard for humans. However, presenting the video data gives more information than just $\mathbf{v}_t$, such as possible facial expressions and gestures. But there are other problems such as the fact that humans have background knowledge. This information enables them to reason about gaze behavior and use their prior knowledge about meetings.

5.1. Experiment setup

To overcome the fact that we cannot present a number of numerical vectors to humans we exploited the fact that humans have background knowledge about meetings. We created a virtual meeting room (VMR, Fig. 3), allowing precise control over the delivered stimuli. In this VMR the setting is visualized, including the locations of the participants and the white board.

One problem with this kind of controlled virtual environment is the trade-off between the ecological validity and the experimental control, resulting in sterile artificial environments (Loomis et al., 1999). However, since we are only interested in head orientations we actually want to neglect other influences. By replacing the real setting (Fig. 1) with a virtual setting we are not only able to display the necessary information but also to remove all other possibly distorting information. This makes it possible, to a minimal extent, for the humans to interpret the gaze vectors and for other possible variables to be controlled.

Participants in the experiment were shown the meeting room with the participants as well as an option panel where they were able to choose among the four speakers, being either confident or very confident. Also there was a ‘no idea’ button to prevent biased unfounded choices.

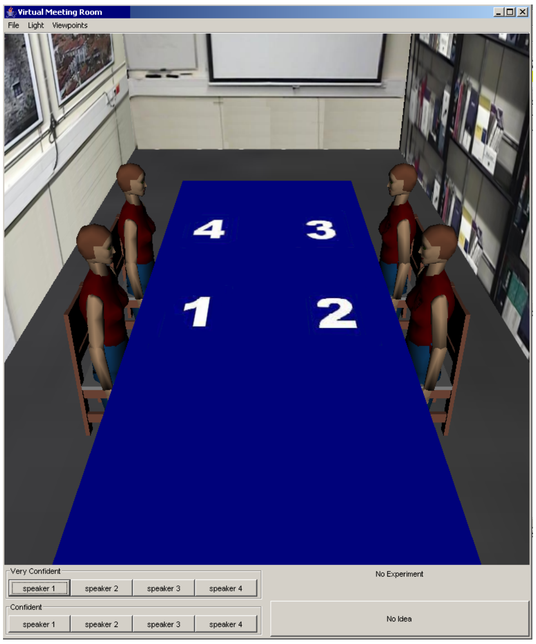

Figure 3. The virtual meeting room setting

Each experiment consisted of a session with four parts, each containing 20 samples. There were two types of sessions. Type 1 contained feedback only on the first part whereas in type 2 the feedback was omitted completely. For the first two parts of both session types two times 20 samples were randomly chosen from meeting 3, the third part contained 20 randomly chosen samples from meeting 2 and the last part contained 20 randomly chosen samples from meeting 1.

The idea behind this was twofold. In the first place it enabled us to see if the feedback was helpful to the participants. Secondly, we were able to see whether feedback on samples from one meeting influenced the results on samples from different meetings. The feedback was given by showing a red arrow above the head of the correct speaker directly after the participants had judged the sample. We asked students and employees of our department to do the test.

5.2. Results

Both session types were completed 20 times, resulting in a total of 3200 answered samples. The results are shown in Table 4

The table shows that the human performance is approximately 38%, which is lower than we expected. An interesting thing to note here is that there are sig-

Type	Part 1	Part 2	Part 3	Part 4	Total
1	47.8%	49.3%	29.8%	24.8%	37.9%
2	39.3%	42.0%	33.3%	35.3%	37.4%

Table 4. Classification results for humans per session type

nificant ($p < 0.05$ using a paired T-test) differences between the two session types. In the first place the results on the first two session parts are better with feedback than without feedback and for the last two session parts we see a significantly worse performance. Furthermore it appeared that when no feedback was given the performance remained much more stable over the different session parts.

We expect that if feedback is given humans create rules or models that work better for the meeting on which they received feedback. These could be *a priori* models that are applied when there is doubt about a possible outcome. When these models were tested on samples from different meetings they did not generalize. This seems to be in accordance with our machine learning findings.

6. Evaluation

From the results of both the machine learning techniques and the experiment with humans, it appears that the models do not generalize over all meetings. Apparently in the meetings different gaze behavior is displayed. This could be caused by different meeting topics, more or less frequent use of the whiteboard and differences in individual gaze behavior.

To gain more insight into differences between and within meetings, we examine two factors more closely in this section. First we investigate if there is a relation between the predicted speaker and the actual speaker in terms of location in the meeting. The second topic we investigate is the performance of the speaker prediction on different moments in a speaker turn.

6.1. Location effect on performance

We expect that different persons display different gaze behavior. Because we do not have sufficient meeting data in which the same persons participate, we try to find indications that there are differences between participants. We examine this by looking at differences in prediction performance for all speakers given the confusion matrices for all classifiers. Then we look at the person specific performance for neural networks on all meetings. Finally we examine whether the position with respect to other participants is of any influence.

In Table 5, 6, 7 and 8 the confusion matrices for the prediction results are shown for all classifiers. For the machine learning algorithms, training and testing is performed on samples from all three meetings (series 2 of Table 2 and 3) to obtain the most reliable model.

Actual	Estimated speaker			
speaker	SP1	SP2	SP3	SP4
SP1	26.2%	12.9%	9.9%	51.1%
SP2	9.6%	60.1%	6.0%	24.3%
SP3	8.7%	12.4%	42.5%	36.4%
SP4	11.8%	11.6%	7.6%	69.0%

Table 5. Confusion matrix for actual speakers (row) and predicted speakers (column) for Bayes classifier without discretization. Performance is 50.0%

Actual	Estimated speaker			
speaker	SP1	SP2	SP3	SP4
SP1	65.5%	8.0%	7.7%	18.9%
SP2	9.5%	71.6%	6.2%	12.7%
SP3	11.5%	7.0%	67.0%	14.5%
SP4	15.9%	4.7%	6.3%	73.2%

Table 6. Confusion matrix for actual speakers (row) and predicted speakers (column) for Bayes classifier with discretization. Performance is 69.4%

Actual	Estimated speaker			
speaker	SP1	SP2	SP3	SP4
SP1	50.6%	9.7%	11.3%	28.4%
SP2	10.1%	61.4%	8.0%	20.5%
SP3	9.8%	8.3%	60.1%	21.2%
SP4	12.3%	6.1%	6.6%	75.0%

Table 7. Confusion matrix for actual speakers (row) and predicted speakers (column) for the experiment with Neural Networks. Performance is 62.9%

Actual	Estimated speaker			
speaker	SP1	SP2	SP3	SP4
SP1	31.8%	15.4%	24.0%	28.8%
SP2	15.3%	42.8%	24.0%	17.9%
SP3	14.3%	14.7%	51.1%	20.0%
SP4	21.7%	12.7%	22.0%	43.7%

Table 8. Confusion matrix for actual speakers (row) and predicted speakers (column) for the experiment with humans. Performance is 42.3%

It appears that there are differences in prediction results for a certain location. For example, for the Naive Bayes classifier without discretization (Table 5) correct performance for speaker 1 is 26.2% whereas the performance for speaker 4 is 69.0%. The results from the above tables are summarized over the three meetings. To find out whether different persons display different gaze behavior we examine the differences between meetings. In Table 9, the Neural Network prediction results for each meeting are summarized.

Location	M1	M2	M3	Total
1	51.4%	47.8%	51.6%	59.0%
2	67.8%	57.4%	58.7%	59.0%
3	88.7%	45.1%	59.8%	62.2%
4	73.8%	70.2%	77.8%	73.0%

Table 9. Correct Neural Network predictions for each location per meeting

Given the results of the Neural Network it appears that the differences in speaker prediction performance are also present between meetings. This shows that there are differences in performance for each person in each meeting, which is a strong indication that there are differences in gaze behavior for different persons. This might be an explanation for the fact that our models do not generalize.

If we look at the origin of the prediction errors, we can tell what mistakes are made with respect to the relative position. In Fig. 3 we see that participants 1 and 4 are sitting next to each other, whereas participants 1 and 2 are sitting opposite to each other. Finally, participants 1 and 3 are sitting diagonally to each other. Table 10 summarizes the errors in these directions.

	Next to	Opposite	Diagonal
Naive Bayes classifier (ND)	36.0%	32.4%	31.7%
Naive Bayes classifier (D)	39.2%	31.1%	29.8%
Neural Network	37.1%	31.6%	31.4%
Humans	38.2%	32.3%	29.5%

Table 10. Estimation errors in different directions for machine learning techniques and humans for all meetings

We see that there is more confusion between two persons who are sitting next to each other than between two people who are in opposite corners of the table. The results are similar for all four classifiers. We expect a relation between physical participant distance and the prediction error. The distance between participants at one side of the table is smaller than the distance between two participants on opposite sides of the table. Changing the meeting setting from a square table with participants sitting opposite to each other to a round table as is used in (Stiefelhagen, 2002) could eliminate those biases but cannot explain the differences from Table 9.

6.2. Performance within a speaker turn

We can take a closer look at the data and determine how our classifiers perform during a speaker turn. Can the speaker be determined better in the beginning, in the middle or at the end of a speaker turn? If we look at the speaker prediction scores within speaker turns longer than 1 second (92.4% of all samples), we obtain the results from Fig. 4. We ignored speaker turns shorter than 1 second, containing short utterances. Ten equally sized bins were used to assure that for each interval sufficient samples (over 300 per bins) remained.

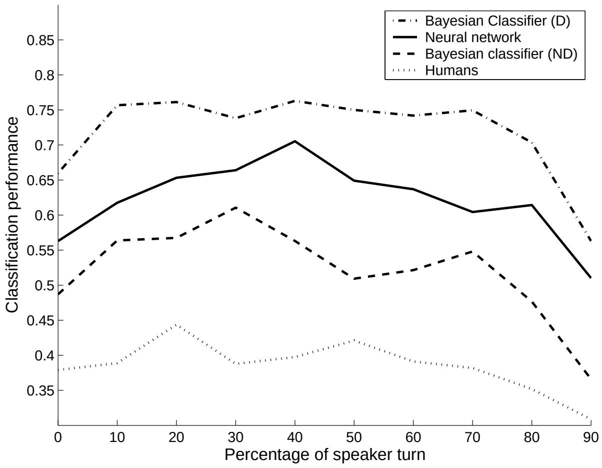

Figure 4. Speaker prediction performance during a speaker turn using 10 bins

We see a trend in the graph for all four classifiers. In the beginning and at the end of a speaker turn, the speaker is harder to determine than halfway the turn. This could be explained by assuming that participants switch from speaking to listening, from listening to speaking or start gazing at the new speaker. A similar explanation could be found for the lower identification performance at the end of a speaker turn. Participants might start gazing at the person who they expect will reply to the current speaker.

7. Conclusions

To gain insight into gaze behavior in meetings, a comparison of classification results for a Naive Bayes classifier, Neural Networks and humans was made on the task of speaker prediction from azimuth head angles. In four-person meetings, a Naive Bayes classifier was able to predict 69.4% correctly, Neural Networks scored 63.2%, and humans only 37.7%. The machine learning classification results do not generalize over meetings. In the experiment with humans we see similar results. The model that was learned using the feedback increased the outcome for the meeting where the feedback was given, but decreased the result for the other meetings. We showed that there are strong indications that human specific gaze behavior influences the fact that the models do not generalize. Additionally, we showed that for all classifiers the performance in the beginning and at the end of a speaker turn is worse than halfway through the speaker turn.

8. Future work

To improve insight into gaze behavior we plan to investigate whether adding more information, such as body orientation will increase the classification performance. Experiments have started where we analyze head orientations of complete speaker turns. Also, we plan to verify simple protocols possibly applied by humans when predicting the speaker.

With respect to the fact that our models do not generalize over meetings, we intend to do more research on person specific gaze behavior. Information such as the typical duration of personal speaker turns, the average head movement during a turn might reveal more cues along our path to addressee detection and focus of attention estimation. Also, more research needs to be done on the effect of meeting topics and their context on the prediction of gaze behavior.

9. Acknowledgements

This work was partly supported by the European Union 6th FWP IST Integrated Project AMI (Augmented Multi-party Interaction, FP6-506811, publication AMI-33).

References

Argyle, M., & Cook, M. (1976). *Gaze and mutual gaze*. Cambridge University Press.

Argyle, M., Ingham, R., Alkema, F., & McCallin, M. (1973). The different functions of gaze. *Semiotica*, *7*, 19–32.

Chen, H., & Perich, F. e. a. (2004). Intelligent agents meet semantic web in a smart meeting room. *Proceedings of the Third International Joint Conference on Autonomous Agents & Multi Agent Systems (AAMAS 2004).*

Dougherty, J., Kohavi, R., & Sahami, M. (1995). Supervised and unsupervised discretization of continuous features. *International Conference on Machine Learning* (pp. 192–202).

Garau, M., Slater, M., Bee, S., & Sasse., M. (2001). The impact of eye gaze on communication using humanoid avatars. *Proceedings of the SIG-CHI conference on Human factors in computing systems* (pp. 309–316). Seattle, WA USA.

Kendon, A. (1967). Some functions of gaze direction in social interaction. *Acta Psychologica*, *32*, 1–25.

Loomis, J. M., Blascovich, J. J., & Beall, A. C. (1999). Immersive virtual environment technology as a basic research tool in psyychology. *Behavior Research Methods, Instruments and Computers*, *31(4)*, 557–564.

Mani, I., & Maybury, M. T. (1999). *Advances in automatic text summarization.* MIT Press.

Moré, J. (1978). The Levenberg-Marquardt algorithm: Implementation and theory. *Lecture Notes in Mathematics 630* (pp. 104–116).

Nijholt, A. (2004). Where computers disappear, virtual humans appear. *Computers and Graphics*, *28.* to appear.

Reidsma, D., Rienks, R., & Jovanovich, N. (2004). Meeting modeling in the context of multimodal communication. *Proceedings of the MLMI'04, Martigny.*

Stiefelhagen, R. (2002). Tracking focus of attention in meetings. *Proc. of the ICMI2002.* Pittsburgh.

van Es, I., Heylen, D., van Dijk, E., & Nijholt, A. (2002). Gaze behavior of talking faces makes a difference. *Proceedings of the ACM-CHI 2002* (pp. 734–735). Minneapolis, USA.

Vertegaal, R. (1998). *Who is looking at whom.* Doctoral dissertation, University of Twente.

Vilhjalmsson, H., & Cassell, J. (1998). Bodychat: Autonomous communicative behaviors in avatars. *Proc. of the 2nd Annual ACM Int. Conf. on Autonomous Agents.* Minneapolis, USA.

Wright, I. (1997). *Emotional agents.* Doctoral dissertation, University of Birmingham.

Yang, Y., & Webb, G. I. (2003). On why discretization works for naive-bayes classifiers. *Proceedings of the 16th Australian Joint Conference on Artificial Intelligence (AI).*