Collaboration and abstract representations: towards predictive models based on raw speech and eye-tracking data

Marc-Antoine Nüssli, Patrick Jermann, Mirweis Sangin, Pierre Dillenbourg, Ecole Polytechnique Fédérale de Lausanne, CRAFT, station 1, CH-1015 Lausanne, Switzerland
Email : marc-antoine.nuessli@epfl.ch, patrick.jermann@epfl.ch, mirweis.sangin@epfl.ch, pierre.dillenbourg@epfl.ch

Abstract: This study aims to explore the possibility of using machine learning techniques to build predictive models of performance in collaborative induction tasks. More specifically, we explored how signal-level data, like eye-gaze data and raw speech may be used to build such models. The results show that such low level features have effectively some potential to predict performance in such tasks. Implications for future applications design are shortly discussed.

Introduction

Theoretical background

We present an exploratory study about gaze patterns exhibited during collaborative interaction. We conducted an experiment to examine dyads solving induction tasks. Two tasks were chosen based upon two main criterions which were to require inductive and abstract thinking, which are known to be related to high-level cognitive processes like learning, and to be visual, in order to allow for the detection of potentially meaningful patterns in the eye-movements. Several authors (Genter 1989, Hofstadter 1995) argued that learning may proceed by analogy between multiple examples. Indeed, analogy consists of finding structural similarities between things that may appear as completely different. Thus, it corresponds to extracting abstract structural features of the concerned objects. The same sort of process may occur during conceptual knowledge learning if we consider that learners have to find similarities between examples of a particular concept to finally induce a general and abstract representation of a that concept.

Raven progressive matrices are a typical task which requires induction and the construction of abstract representations. These problems have been shown to be central to all cognitive abilities in the sense that most specific ability tests are generally well correlated with Raven matrices tests (Carpenter, Just and Shell, 1990). Carpenter and his colleagues found that gaze patterns partially reflected the solving phases of these tasks by comparing verbal reports during resolution and gaze data. They have also shown that abstraction abilities are one of the main factors which explain successful solving of the problems.

Schwartz (1995) has shown how collaborating students may outperform individuals in building abstract representations about scientific concepts. He ran two experiments in which students had to build abstract representations of a problem in order to answer a set of questions. He showed that the performance of the dyads were greater than what could be expected from a theoretical model called which he called *truth-wins* model. This model assumes that the best performance that a pair may achieve is the performance of the best of the two collaborators. This study suggests that this theoretical model may not be valid and that a dyad may be more than the sum of two individuals.

Eye-movements have been related to social interaction processes by several authors. Richardson and Dale (2005) have shown how the language and the gazes are related to each other. They have demonstrated that the coupling between two interlocutors' gazes is correlated with their level of understanding. They found a similarity in the gaze sequence of the conversants with a certain time lag. This effect is explained by the fact that speakers look at the object they are talking about before naming it and listeners do the same after hearing the word (Griffin and Bock, 2000). .

Task selection

We have chosen to explore how dyads solve two different logical games, both requiring induction and abstraction abilities. The first was also studied by Carpenter, Just and Shell (1990), namely the Raven progressive matrices (see Fig. 1 top-left). It consists in finding out the last element of a 3-by-3 matrix which exhibits certain logical patterns over its rows and columns. Performance on Raven matrices is a good predictor for performance on most specific ability tests which are generally well correlated with Raven matrices tests (Carpenter, Just and Shell, 1990).

The second kind of problem is called Bongard problems (see Fig. 1 top-right). These problems were originally designed by M. Bongard in a book entitled "Pattern Recognition" (1970). The goal was to provide examples of what pattern recognition machines should be able to solve. The goal of these problems is to find a common pattern or rule among the six images on the left (examples) and which doesn't work for the six images on the right (counter-examples). What makes these problems quite hard is that the rule may involve completely different features. It may be the relative position of the objects, their relative size, the orientation or it may also be a kind of higher level shape formed by many lower level shapes.

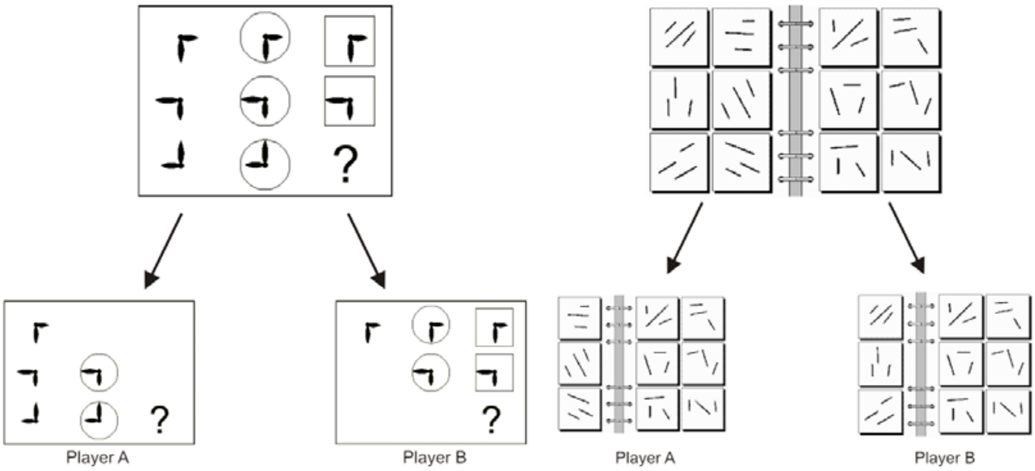
Research questions

This work explores the possibility of building predictive models in order to develop in the future gaze-sensitive groupware. Indeed, we think that eye-tracking techniques will become more and more accessible as it is possible already to build cheap eye-trackers with simple webcams. Our idea is to use real-time gaze data to support the analysis and diagnosis of collaborative learning processes. The main prerequisite to this goal is to find some gaze patterns, possibly combined with other easy-to-acquire data like raw speech, which are related to successful collaboration. Our approach to this problematic is to apply machine learning algorithms. Although these techniques generally do not yield theoretically interpretable models, they enable to build predictive models which can be very efficient and sufficiently fast to be computed in real time.

Method

Task

The two kind of problems described above have been slightly modified for the purpose of this study. In order to make them more interactive, the images have been split between the two participants. For the Raven matrices, six (out of nine) cells were shown to each participant. One saw only the upper-right part of the matrix while the other saw only the lower-left part of the matrix. Thus, they each had three personal cells which were not seen by the other participant and three shared cells (on the diagonal). One of the shared cells was the missing cell which had to be discovered by the collaborators. For the Bongard problem, the split was a bit different. Each participant could see the six counter-examples (right images) but each participant only saw three out of the six examples (left images). Thus, in both cases, three cells were not shared by the collaborators.



<u>Figure 1.</u> Examples of a Raven progressive matrix (left) and of a Bongard problem (right). The answer for the matrix would be "clock indicating nine o'clock inside a square" as there is a shape progression along the row and a clock rotation along the column. The rule of the depicted Bongard problem would be "the lines are parallels" while there is no rule for the right side. Bottom images show modifications applied to the problems to make them collaborative.

Participants

Nine dyads (ten men and eight women) were recruited among campus collaborators and students. Subjects' ages vary between 17 and 53 with a median of 27 years. None subject was aware of what a Raven matrix or a Bongard problem is before the beginning of the experiment.

Procedure

Two computers were installed in the same room separated by a shelf in order that the subjects could not see each other while still being able to speak to each other. Two eye-tracking screens (Tobii T1750) were used to record subject's eye movements. Subjects were first asked to fill in a short questionnaire about general information like age and sex and how much they know each other. The experiment was composed of 12 static images which could be passed by simply pressing the spacebar at least one time on each computer. The first and the seventh slides were instructions for the Raven matrices problems and the Bongard problems respectively. Slides 2 to 6 presented the Raven matrices and slides 8 to 12 were the Bongard problems. The problems were in order of increasing difficulty in order to allow subjects to familiarize themselves with the problems.

The subjects had to solve the problems together, agree on a solution and then, say it out loud and press the spacebar to go to the next problem. The correctness of the solution was checked and recorded by the experimenter. There was a maximum time limit of 5 minutes for each problem. Speech was also recorded separately for each individual.

Data analysis

Variables

We computed several features based on the gaze data. The first feature, called number of *comparisons*, aims at detecting when subjects compare two cells. We identified every sequence of at least 3 fixations with at least one back and forth movement between one cell and another. The *comparisons* variable is the ratio of all fixations which belong to such sequences (see fig. 2, top-left). A related feature is the *comparison intensity*. For each of these comparison sequences, we computed the number of transitions between the two cells concerned and then, we averaged this number over all comparison sequences (see fig. 2, top-left).

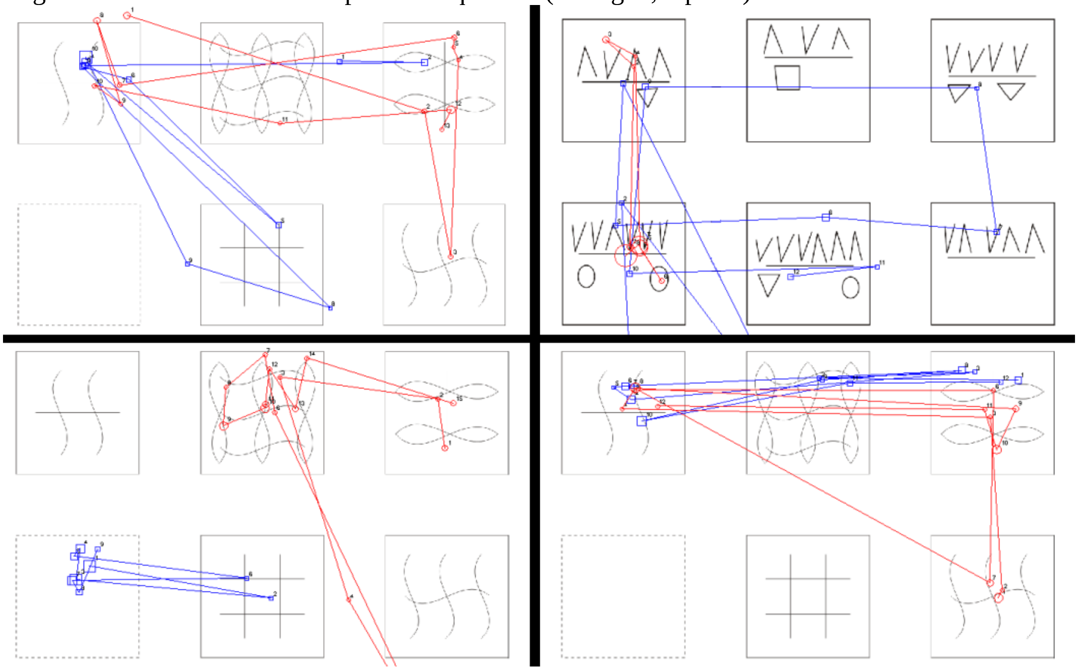


Figure 2. Illustration of gaze features. Top-left picture depicts one subject (square) doing an intense comparison between the upper-left cell and middle cell and the other subject (circle) doing a weak comparison between the upper-right cell and middle-right cell. On the top-right picture, we can see a dispersed subject (square) and one not dispersed (circle) Bottom images illustrate high gaze divergence (left) and low gaze divergence (right)

Another feature measures how much subjects look at all cells in an equivalent manner or in other words, how much their gaze is dispersed. For this, we aggregated the fixation durations in a matrix, called *cell density matrix*, representing the nine cells present on the screen and then for each cell, we took the ratio of the total aggregated durations. Finally, we computed the standard deviation of the values in this matrix as the *gaze dispersion* value (see fig. 2, bottom-right). We also used the cell density matrix to compute a dual gaze feature called *gaze divergence* (see fig. 2, bottom-left). This feature is simply the cosine between the density matrices of both subjects, which is a way to assess the similarity between two matrices.

Each second of the recorded speech data was automatically labeled as *speech* or *no-speech*. First, the audio file was split in order to have one fragment per problem and each fragment was normalized using the minimum and maximum found over the sample. Then, the root-mean square was computed for every second and if this value exceeded a threshold of 0.4, the second was considered as *speech*. The resulting feature, called *speech time*, is the ratio of seconds labeled as speech for each subject. Then, we also computed the difference of the *speech time* between the subjects of a pair in order to have an estimation of the *speech time asymmetry*. We decided to focus only on these simple raw measures of speech because it is fully automatic and thus it could be easily used in potential future application.

Finally, we analyzed two dependent variables: the success at a particular problem and the *solving time* for correctly solved problems.

Analysis methods

We tried to apply machine learning algorithms on our dataset in order to see if it is possible to predict the performance of individuals by using the gaze and speech features described above. Indeed, one of our final goals is to build gaze-sensitive applications that would detect in real-time meaningful gaze patterns, possibly combined with raw speech features, in order to give feedback to the users. Thus, machine learning techniques provide with a way to build models able to do such detection.

In this study, we compare the use of two different machine learning algorithms, one called J48, which builds binary decision trees and another called Naïve-Bayesian, which estimates probability distributions for each features.

Results

We present here results which stem from the use of two machine learning algorithms (Binary decision tree or Naïve Bayesian classifier) for each problem class separately but also for both problems classes without distinction. We also analyzed the effect of using speech only as predictors, gaze only or gaze and speech combined. Two values are always reported, the number of correctly classified cases and in parenthesis the kappa statistics, which represents how much the model is better than chance. These values have been obtained using 10-fold cross-validation procedure. Algorithms were fed with features computed on one minute duration and the minute was also used as a predictor. However, we discarded for each problem the last minute before the solution was announced in order to avoid the effect of speech which may due to the explanation of the solution. The predicted variable was the outcome of the problem: solved or unsolved. The predictors were the number of comparisons, the comparison intensities, the gaze repartition and the gaze divergence for the gaze features and speech quantity and speech asymmetry for the speech features. It is important to note that these two algorithms, like most machine learning, do not necessarily produce better results when more predictors are given.

Table 1: Results of the machine learning algorithms for both problem classes combined, kappa's are in parenthesis

	Naïve Bayesian classifier	J48 Binary decision tree
Gaze + speech	78% (50%)	79% (51%)
Speech only	77% (45 %)	86% (65%)
Gaze only	74% (35%)	68% (10%)

First, it is very interesting to note that we can obtain quite good results (50% above chance level) while we are trying to predict success in two different tasks. This is very encouraging as it suggests that there may exist some patterns in gaze and speech which are task independent. Of course, the two tasks are not completely different as they both imply some similar processes (induction and rules abstraction). We can also see that at this level, speech plays clearly a larger role than gaze. Indeed, we see that models using only gaze features are the worst for both algorithm types. However, for the Naïve Bayesian classifier, gaze seems to slightly improve the performance compared to speech only, indicating that it can still play a role.

Table 2: Results of the machine learning algorithms for Raven problems, kappa's are in parenthesis

	Naïve Bayesian classifier	J48 Binary decision tree
Gaze + speech	78% (56%)	91% (81%)
Speech only	78% (56%)	91% (81%)
Gaze only	68% (32%)	68% (34%)

The results concerning Raven problems (see table 2) only are surprisingly high, producing up to 80% above the chance level with 91% of correctly classified instances. Moreover, we can see that these results are explained only by speech features. Although it is a bit disappointing because we expected to find some patterns in gaze data, it is also very surprising to see that such raw speech features may predict so well the success on these problems. Of course, we must be very cautious in interpreting these results because like for a correlation, it does not imply that there is causality between speech quantity and asymmetry and the success.

For Bongard problems (see table 3), the situation is the opposite than for Raven problems, although the results are much lower than for Raven and even lower than for both classes combined. This suggests that the good result for all problems taken together is mainly explained by the Raven problems. However, there are still some results for Bongard problems and interestingly, we can see that these performances are explained mainly by gaze features, as the best models are achieved by taken only gaze features without speech features.

Table 3: Results of the machine learning algorithms for Bongard problems kappa's are in parenthesis

	Naïve Bayesian classifier	J48 Binary decision tree
Gaze + speech	76% (34%)	75%(25%)
Speech only	76% (29%)	75% (21%)
Gaze only	77% (37%)	77% (37%)

We also tried to apply these algorithms with slightly different data to have situations closer to a real-time situation. When using only the first two minutes of solving, the results are either similar to those presented or a little bit (3 or 4%) lower. These results are maybe even more interesting because they suggest that it could be possible to detect after one or two minute if the pair will succeed or fail. Moreover, we also tried to predict the success in the next minute. Here, the results are clearly lower than the previous ones but they are still sufficiently high to be considered. We obtain kappa-scores of 40% (instead of 50%) for both problems combined. Again, such results are still more interesting for a potential future gaze-sensitive application because we could be able to predict the moment at which a pair will succeed. Also, it suggests that there exist some phases in the solving processes which are distinguishable by using the gaze patterns and this is consistent with the results found using usual statistical methods.

Discussion

It is very encouraging to see how well machine learning algorithms performed on these data. As we have seen, we can predict up to a certain point problem solving outcomes by using only raw measure of speech and gaze features. Moreover, we see that we may be able to predict the moment of resolution one minute before it happens. These results have great implications as they tend to prove that it is possible to build gaze-sensitive applications, possibly combined with simple automatic speech analysis, in order to provide meaningful feedback to users. Obviously, predicting only the solving moment or the solving outcome is not sufficient for such application but it shows that patterns may exist in gaze and raw speech and thus, we can imagine that similar patterns could be also present in other situation that may be of interest for feedback. However, one must note that gaze plays a significant role only in the Bongard case, while for Raven matrices, only speech was useful for predictions.

Of course, all these results must be taken with care. Indeed, the number of subject is very low and so, it is difficult to generalize. At this point, we cannot be sure that these models built by machine learning algorithms are really universal or if they are specific to this set of subjects.

Conclusion

We have shown that it may possible to design fully automated systems able to predict some outcomes of interaction by using signal-level features like raw speech and gaze data. This is a step towards building applications which may enhance the collaboration processes by providing real-time meaningful feedbacks. This is especially interesting because these problems require high-level thinking and thus, it suggests that similar results may be found in other high-level tasks, like collaborative learning. Of course, these results are only preliminary and we need further before being able to draw strong conclusions.

Bibliography

Bongard, M. M. (1970). Pattern Recognition. *Rochelle Park, N.J.: Hayden Book Co., Spartan Books.*

Carpenter, P. A., Just, M. A. & Shell, P. (1990) What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review 97*, 404–31.

Gentner, D. (1989). The mechanisms of analogical learning. In *Similarity and Analogical Reasoning*, S. Vosniadou and A. Ortony, Eds. Cambridge University Press, New York, NY, 197-241.

Griffin, Z. M., and Bock, K. (2000). What the eyes say about speaking. *Psychological Sciences 11*, 274--279.

Hofstadter, D. R. (1995). Fluid Concepts and Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought. *New York: Basic Books.*

Richardson, D. C., & Dale, R. (2005). Looking to understand: The coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. *Cognitive Science, 29*, 1045–1060.

Schwartz, D.L. (1995). The emergence of abstract dyad representations in dyad problem solving.*The Journal of the Learning Sciences, 4 (3)*, pp. 321-354.

Acknowledgments

This research was funded by the Swiss National Science Foundation (grant #K-12K1-117909).