3D Tangibles Facilitate Joint Visual Attention in Dyads

Bertrand Schneider, Stanford University, schneibe@stanford.edu
Kshitij Sharma, EPFL, kshitij.sharma@epfl.ch
Sébastien Cuendet, EPFL, sebastien.cuendet@epfl.ch
Guillaume Zufferey, EPFL, guillaume.zufferey@epfl.ch
Pierre Dillenbourg, EPFL, pierre.dillenbourg@epfl.ch
Roy D. Pea, Stanford University, roypea@stanford.edu

Abstract: We report results from a dual eye-tracking study around a Tangible User Interface (TUI). Participants (N=54) worked in groups of two and solved optimization problems on the TinkerTable, a TUI designed for students in logistics. The TinkerTable features tangible shelves that students can manipulate to build and optimize the layout of a warehouse while the system provides feedback with a projector above the table. Using mobile eye-trackers, we examined participants' visual coordination when solving those problems. We describe two contributions to the CSCL community: first, we propose a methodology for synchronizing two eye-tracking goggles and computing measures of joint visual attention (JVA) in a co-located setting. Second, we report preliminary findings suggesting that participants were more likely to have moments of joint attention when looking at 3D, realistic objects compared to 2D, abstract ones. JVA was also found to be a significant predictor of students' learning gains and performance during the optimization tasks. We discuss implications of these findings for supporting interactions around a TUI.

Keywords: Tangible user interface; eye-tracking; joint attention.

Introduction

A plethora of work suggests that Tangible User Interfaces (TUIs) are beneficial for collaboration in co-located settings (reviewed in Dillenbourg & Evans, 2011). Falcao and Price (2009), for instance, identified the constructive role of interferences in tangible learning environments: they found that sharing physical objects creates opportunities for verbal negotiation and conflict resolution, and that the "present at hand" nature of the tangibles supports a balanced level of participation. Additionally, several researchers proposed frameworks for conceptualizing interactions around interactive tabletops: Hornecker and Bur (2006), for instance, proposed that TUIs are different from GUIs (Graphical User Interfaces), because they provide tactile feedback, are embedded in real space, allow for embodied interactions and combine the expressiveness of both virtual and physical material to facilitate collaboration between users. While those contributions are important stepping-stones toward understanding users' interactions around a TUI, few empirical studies describe clear quantitative differences between small groups of users. Most studies are qualitative (e.g., Falcao & Price, 2009; Hornecker & Bur, 2006) or adopt a high-level perspective by using a rating scheme for the entire collaborative episode (e.g., Schneider, Jermann, Zufferey & Dillenbourg, 2011).

The goal of this paper is to close this gap by providing highly granular data on users' visual coordination around a TUI. We describe two contributions: first, we propose a methodology for synchronizing two eye-tracking goggles and computing measures of joint visual attention in a co-located setting. Second, we report preliminary findings suggesting that participants were more likely to have moments of joint attention when looking at 3D, realistic objects compared to 2D, abstract ones. To our knowledge, this is the first study to quantitatively measure joint visual attention in a co-located setting using mobile eye-trackers.

Joint attention in collaborative settings

Joint Attention is a crucial mechanism by which people establish common ground. It is defined as "the tendency for social partners to focus on a common reference and to monitor one another's attention to an outside entity, such as an object, person, or event" (Tomasello, 1995). Joint attention has been shown to be associated with higher quality of collaboration in multiple settings, both qualitatively and quantitatively. For the focus of this literature review, however, we will limit ourselves to eye-tracking studies even though there is a rich literature on joint attention from multiple fields (developmental psychology, education, social psychology). Richardson and Dale (2005), for instance, asynchronously recorded a speaker and a listener and measured their gaze with an eye-tracker; they found that the degree of gaze recurrence between speaker-listener dyads (i.e., the proportion of times that their gazes are aligned) was correlated with the listeners' accuracy on comprehension questions when watching the video of the speaker. In a dual eye-tracking setting, Jermann, Mullins, Nuessli and Dillenbourg (2011) described how "good" programmers tend to have a higher recurrence of joint visual attention when having productive interactions compared to less proficient programmers. Schneider and Pea (2013), using two synchronized eye-trackers, showed that students who could see the gaze of their partner in real time in a remote collaboration outperformed their peers on a subsequent learning test compared to a control group; furthermore, learning gains were correlated with higher recurrence of joint visual attention. Finally, Brennan et al. (2008) studied the effect of shared gaze and speech during a spatial search task; they found that the shared gaze condition was the best of all. It was twice as fast and efficient as solitary search, and significantly faster than other collaborative conditions.

Due to space constraints, we will not exhaustively review the literature on joint attention. However, this phenomenon has been extensively studied in various fields and we can confidently say that joint attention is crucial for productive interactions between people of all ages. Additionally, the studies mentioned above suggest that eye-trackers are promising tools for understanding the factors that support a productive collaboration. In the next section, we discuss the relationship between tangible interfaces and joint attention.

Tangible interfaces and joint attention

To our knowledge there is no existing work connecting tangible interfaces and joint visual attention. Perceptual psychologists, however, have been studying the effect of physical stimuli on people's cognition for a long time. Farah, Rochlin and Klein (1994), for instance, asked subjects to recognize various 3D shapes. In one condition, they were presented with filled potato chip-like shapes. In another condition, subjects only saw their contours as wire frames. They found that participants were more accurate at remembering and recognizing shapes from the first category. This suggests that filled shapes contain perceptual information that can be used to facilitate visual recognition. In a meta-analysis, Sowell (1989) looked at 60 studies investigating the benefits of using manipulatives vs. pictorial material for teaching mathematics; she found that the long-term use of concrete material increased students' achievement and improved their attitude toward mathematics. Regarding verbal skills, Glenberg et al. (2004) asked children to read a text while manipulating toy objects (e.g., a barn, a tractor, a horse) to simulate the actions described in the text. Compared to a control group without access to those objects, this intervention resulted in markedly better (vs. rereading) comprehension of the text material. While many more studies have examined the effect of physical objects on people's cognition, it seems that in some contexts at least, tangible material has a positive effect on memory, problem-solving strategies, comprehension and learning. It is an open question whether such benefits transfer to social interactions. The goal of this paper is to explore this question in more detail using empirical measures. In the next section, we describe our experiment and dataset.

Experimental data

In this study we compared the effect of interacting with 3D or 2D objects around a tangible interface. Students interacted with an interactive simulation of a warehouse (Schneider, Jermann, Zufferey & Dillenbourg, 2011). The TinkerLamp allows users to quickly build and evaluate small-scale warehouses, thanks to a simulation projected on top of the model (Fig. 1). This learning environment has been adopted in several schools in Switzerland and is a robust pedagogical tool for teaching concepts in logistics. In this study, students' task was to analyze and optimize layouts for several warehouses while wearing eye-tracking goggles. In one condition, half of the participants used physical, 3D shelves (Fig. 1, right side); in another condition, shelves were represented by 2D paper rectangles (Fig. 1, left side). This manipulation allowed us to better control for the "representational effect" of 3D tangibles: the first group saw the warehouse as a small-scale model with realistic shelves, whereas the second group saw to a more abstract perspective with flat rectangular pieces of paper.

Methods

Subjects

Fifty-four apprentices in logistics participated in the study (28 in the "tangible" condition, mean age = 19.07, SD = 2.76; 26 in the "paper" condition, mean age = 17.96, SD = 1.56). Due to the vocational domain, few women participated (7 females; 4 in the "tangible" condition, 3 in the "paper" condition). All students who took part were following a vocational training in logistics: 16 first-year, 16 second-year, and 22 third-year (N=54).

Experimental design

We used a between-subjects experimental design with two conditions (Fig. 1): participants either worked with 2D paper shelves or 3D tangible shelves. We counter-balanced students' expertise between the two conditions: 8

first-years, 8 second-years and 12 third-years were assigned to the “tangible” condition; 8 first-years, 8 second-years and 10 third-years were assigned to the “paper” condition.

Procedure

The experiment was conducted in a private room of a Swiss professional school over a 4-day period. Participation was non-mandatory and did not count towards a grade. Upon arrival, the experimenter welcomed participants and asked them to complete a pre-test. The experimenter then told the students that their first task would be to individually memorize a warehouse layout for one minute and rebuild it from memory (Fig. 1, first row). Participants were then provided with the correct size of the warehouse—24 shelves, 2 docks—and were asked to individually recreate the layout. In the next activity, the experimenter told the students that they would have to discuss 3 layouts for 10 minutes (Fig. 1, middle row) based on three criteria (e.g., “In which warehouse would you prefer to work?”, “Which warehouse optimizes space? Why?”, and “Which warehouse minimizes the average distance from each shelf to the docks? Why?”). Finally, the experimenter provided them with two design principles about an efficient warehouse layout (in terms of optimizing storage space and of the average time to pick up an item from a shelf). For the last task, participants were instructed to use those principles to build their own warehouses and they used the interactive version of the TinkerLamp (Fig. 1, last row). They had 6 minutes to optimize storage space (e.g., insert as many shelves as possible) and 4 minutes to minimize the average distance to the docks by keeping only 9 shelves in their warehouse. Finally, participants completed a post-test (similar to the pre-test, except that the warehouse models were slightly different).

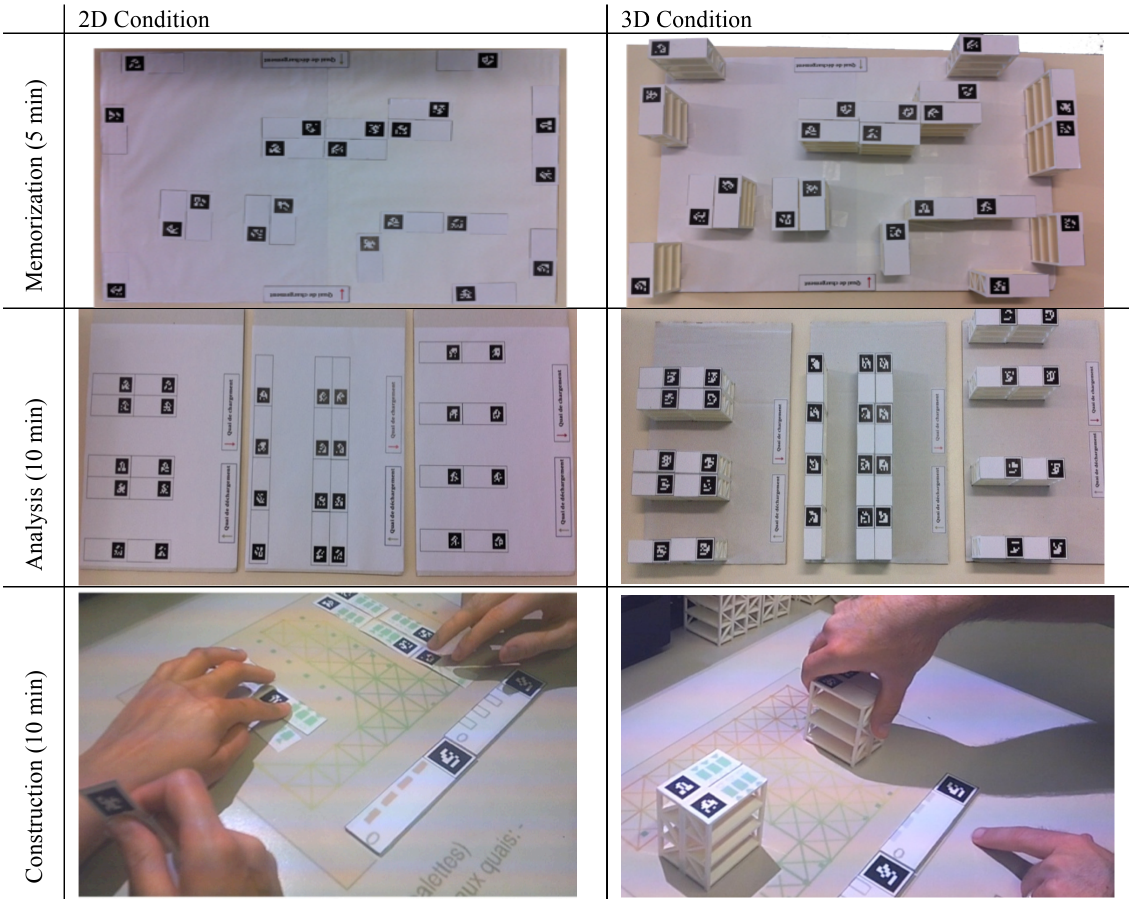


Figure 1. Our three experimental tasks: 1st row shows the layout participants had to memorize; 2nd row the contrasting cases students had to analyze; 3rd row the TUI for the construction task.

Material

The stimuli for each task are described in Figure 1. The learning tests had 4 questions: we asked students to 1) estimate the minimal distance between two shelves for a forklift to load up a pallet; 2) optimize the average distance between the docks and shelves of a given warehouse by correctly positioning two docks (arrivals and shipment); 3) optimize the average distance between the docks and shelves, except that the two docks were already positioned and participants had to move two shelves to minimize the distance to the docks; 4) select good design principles for both maximizing space and minimizing the average distance to the docks from a multiple-answer question.

Coding

For the memory task, we counted the number of shelves and docks in the correct location as a retention score. For the optimization task, the TinkerLamp provided us with two measures of performance: the number of accessible pallets and the average distance to the docks. Additionally, students' answers to the learning tests were evaluated as follows: for the 1st question (estimating the minimal distance between two shelves to load up a pallet), participants received 1, 2, 3 or 4 points depending on the accuracy of their measurement. Answers below the minimal distance earned 0 points. For the 2nd and 3rd question, an optimal arrangement was worth 4 points. Points were deducted based on their (dis)similarity with the best answer. Question four and five (multiple-answer question) were evaluated as right or wrong. Perfectly answering the test was worth 20 points. Learning gains were computed by subtracting the score of the pre-test from the score of the post-test.

Experimental results

Using a multivariate analysis of variance (MANOVA), we found that participants in the "tangible" condition outperformed the participants in the "paper" condition for the memory task: F(1,52) = 4.48, p = 0.039 ("paper" condition: mean = 14.08, SD = 4.23; "tangible" condition: mean = 16.21, SD = 3.14). Students also built more efficient warehouses with the tangible version of the TinkerLamp, both for the space optimization task: F(1,25) = 16.79, p < 0.001 ("paper" condition: M = 11.69, SD = 2.25; "tangible condition: M = 15.29, SD = 2.30) and for the distance optimization task: F(1,25) = 12.01, p = 0.002 ("paper" condition: M = 8.73, SD = 2.12; "tangible" condition: M = 6.61, SD = 0.84). Finally, participants in the "tangible" condition had higher learning gains compared to the participants in the "paper" condition: F(1,52), 5.21, p = 0.027 ("paper" condition: M = 0.12, SD = 3.6; "tangible" condition: M = 2.5, SD = 4.04). Since those results are not the main focus of this paper, please refer to (Schneider, 2014) for a discussion of those findings.

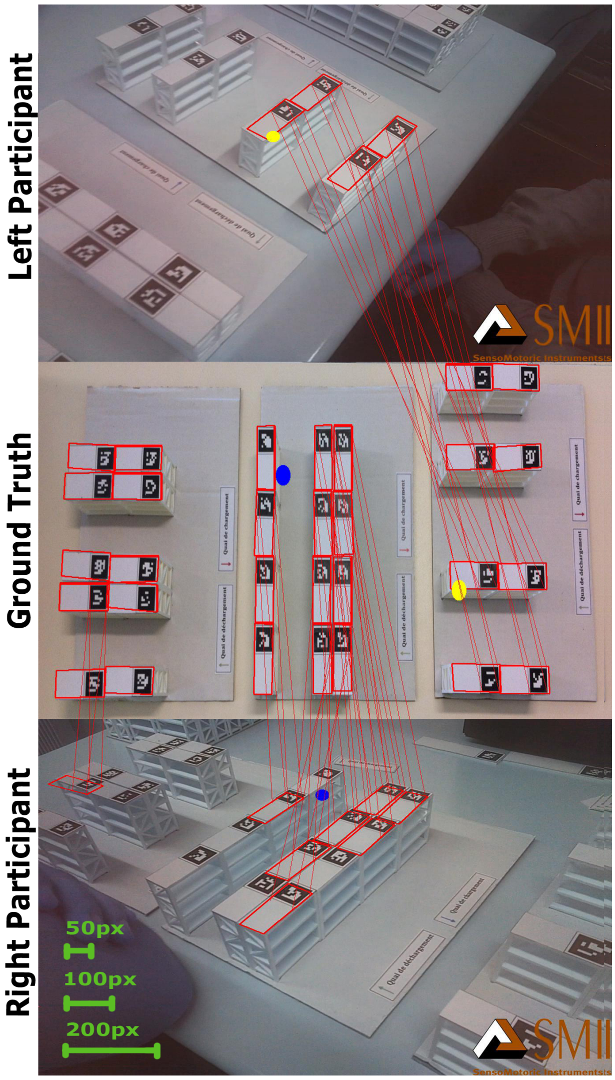
Eye-tracking analysis

Participants wore non-invasive mobile eye-trackers during the three experimental tasks. The models used were 2 SMI Eye-Tracking Glasses (ETG) with binocular pupil tracking at 30Hz. Those units are lightweight (75gr.) and can easily be used for hour-long experimental sessions. We also used the scene camera (1280x960 pixels) to do fiducial tracking and synchronize the two devices. For this paper, we limit ourselves to the eye-tracking data generated during the second task. Students did not manipulate the shelves (but merely had to analyze the layouts and identify good design principles), and there was no virtual information overlaid on the tangibles. This step was 10 minutes long, meaning that our dataset contains 30 gaze points * 60 sec. * 10 min. (=18'000 data points) for each student, and almost 1 million data points for this activity across all participants.

Methodology for capturing joint attention in co-located settings

Working with mobile eye-trackers is challenging. Compared to standard eye-tracking setups, there is no predefined area of interest (AOI): participants are free to look 360 degrees around; they are not limited to the AOI defined by a computer screen. Thus, one requirement of our study was to have a "ground truth" to remap participants' gaze. Another requirement was to have reference points from both perspectives to compute a homography between what students saw and our ground truth (a homography is a mathematical operation that projects a point from a first plane to a second plane, using a minimum of 4 points known in both planes). A last requirement was to be able to synchronize the two eye-trackers by analyzing the videos generated by the scene camera. For the analysis task considered in this paper, the ground truth was easily defined: we simply took a picture of the three warehouses in each condition (Fig. 1, middle image). For the reference points, we used the fiducial tracker engine from the TinkerTable. We analyzed the frames of the videos generated from the eye-tracking goggles and generated log files with the location of each shelf detected in every single frame. We then used this information to remap the gaze of each subject on the ground truth using a homography. Lastly, we synchronized the two eye-tracking datasets by showing a special fiducial at the beginning and end of each activity: this provided us with a visual "hand clap" to synchronize both eye-trackers.

We spent a significant amount of time sanity checking our analyses by producing videos reconstructing the scene with both the students' perspectives and their gazes remapped onto the ground truth. Figure 2 (left side) shows one frame from such a video: the top and bottom images show the perspectives from the students, with their gaze in yellow and blue. Shelves captured by the TinkerLamp fiducial engine are shown in red. Lines between the student's perspective and the ground truth represent the points used for the homography. Each shelf provides four points (i.e., the corners of the fiducial marker). It should be noted that the fiducial engine does not perfectly capture all the tags. To make sure that we had enough data for data analysis after the homography, we show the number of data points left for analysis at the end of the process (Fig. 2, right side). In general, ~10% of the data was lost after the homography, either because students were looking away from the table or because no fiducial markers were detected. For instance in Fig. 2, if no shelf had been detected from the perspective of the left participant, we would lose this data point because no homography could be computed – we would not have any points in common with the ground truth, and thus wouldn't be able to remap the participant's gaze onto it.



	Group	Data Points Captured	After the Homography
Number of Fixations	Tangible	6295.5 (1383.92)	4817.21 (1169.32)
	Paper	6722.12 (1472.68)	5953.96 (1369.55)
Number of Saccades	Tangible	2360.07 (701.94)	1794.86 (768.01)
	Paper	2414.92 (775.68)	2149.48 (869.47)
Number of Blinks	Tangible	5439.46 (1455.97)	4250.57 (1188.4)
	Paper	5274.0 (1043.05)	4800.56 (1012.79)

Figure 2. Left side: one frame of our video analysis, used as a sanity check. The top and bottom images show the perspective from the students; the middle image shows the ground truth used to remap both gazes. Right side: Average number of data points for each student and type of gaze event (Fixation, Saccade, Blink). Standard deviations are shown in parentheses. Data retained after the homography is shown in the last column.

Computing a metric for joint visual attention

There are two main parameters to consider when measuring joint visual attention from eye-tracking data. First, participants' visual attention is rarely perfectly synchronized. In a foundational study, Richardson and Dale (2005) looked at the coupling between a speaker's and a listener's eye movements and found that a listener's eye movement most closely matched a speaker's gaze with a delay of 2 seconds. A second parameter to consider is the threshold for the distance between two gazes to qualify as 'joint attention'. Jerman, Mullins, Nuessli and Dillenbourg (2011) used a radius of 70 pixels with participants looking at a computer screen, but the size of the ellipse depends on the distance of the participants' eyes to the plane they are looking at. We build on those results to compute our own metric of joint attention: first, we looked at each gaze point from the first participant and tried to find a corresponding point from the second participant using different time windows (+/- [0,5] sec.; see Fig. 3, right side). Second, we tried radiuses between 10 and 190 pixels (Fig. 3, right side), where a threshold of 50 pixels corresponds to the shelf width. To facilitate interpretation, we show those values on the ground truth (in green on the bottom left side of Fig. 2): a distance of 50 pixels corresponds to the width of a shelf and 100 pixels to its length. We show below (Fig. 3, left side) the percentage of joint attention for each experimental group (y-axis) with different thresholds (x-axis). Finally, we also computed similar measures of JVA for the memory task (Fig. 3, left side), which provides us with a baseline showing the level of joint attention when two students looked at the same plane without collaborating. The rationale for this comparison is that there may be a bias for students to look at 3D shelves more often, which would artificially increase the likelihood of achieving joint attention (regardless of the collaboration between students).

Since the number of data points varied widely between participants (see the standard deviation reported in table 1), we divided our measure for joint attention by the total number of gaze points of a participant to obtain a *percentage* of joint attention over the entire activity. This prevented us from inflating the joint attention score of a student that had more data points captured. Finally, we discarded blinks and saccades and only focused on fixations (i.e., the pause of the eye movement on a specific area of the visual field). The eye-tracking software (SMI BeGaze) automatically detected these three events (i.e., fixations, saccades, blinks).

Eye-tracking results

Using different thresholds (i.e., distance between two gazes) for measuring joint attention

As a first pass, we used a threshold of 50 pixels for the distance between two gazes to qualify as joint attention, and a time window of +/- 2 seconds as advised by Richardson and Dale (2005). Students in the 3D condition had significantly more joint attention compared to the students in the 2D condition: $F(1,25) = 4.98$, $p = 0.04$, Cohen's $d = 0.91$ (for the 2D condition, mean=0.12, SD=0.07; for the 3D condition, mean=0.19, SD=0.07). In other words, students in the 2D condition gazed at the same location 12% of the time while students in the 3D condition jointly looked at the same area 19% of the time. Results were comparable when considering larger thresholds (see the non-overlapping error bars in Fig. 3, left side). For instance, we found similar results when using 100 pixels as a threshold: $F(1,25) = 5.64$, $p = 0.03$, Cohen's $d = 0.97$ (for the 2D condition, mean=0.25, SD=0.12; for the 3D condition, mean=0.36, SD=0.10). For the memory task, students in the 2D condition were more likely to achieve JVA, though this difference is not significant at any threshold ($F < 1$).

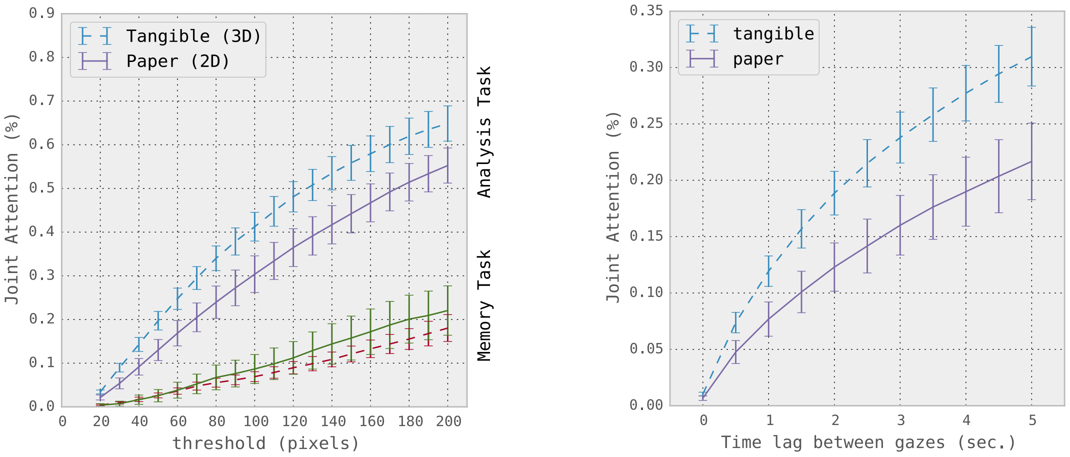


Figure 3. Difference of JVA between each experimental group when considering various distance thresholds (left side) and time lags (right side) for two gazes to qualify as joint attention.

Using different time windows (i.e., the lag between two gazes) for measuring joint attention

In this section, we explore different time lags between two gazes to qualify as joint attention. Using a delay of +/- 2 sec., we found that students in the 3D condition had significantly more joint attention compared to the students in the 2D condition ($p < 0.05$, as reported above). Results were comparable using different time windows (see the non-overlapping error bars in Fig. 3, right side).

The predictive value of joint visual attention

The JVA metric was predictive of students' performance on the subsequent optimization tasks: $r(24) = 0.431$, $p = 0.028$, which suggests that joint attention was associated with a better understanding of design principles for optimizing warehouse layouts. When computing correlations for 1st, 2nd and 3rd year students, we found different effects. Regarding 1st and 2nd year students (grouped together), joint attention was correlated with students' performance during the first optimization task (optimizing space): $r(13) = 0.59$, $p = 0.021$ and learning gains $r(16) = 0.423$, $p = 0.051$. For 3rd year students, joint attention was correlated with students' performance at the second optimization task (optimizing distance from the shelves to the docks): $r(9) = 0.618$, $p = 0.043$ and not with learning gains $r(9) = 0.101$, $p = 0.768$. Those results show different dynamics of collaboration as students become experts, and suggest differentiated effects of a 3D TUI on different samples of students.

Finally, we report an anecdotal result that needs to be confirmed by future analyses. We rated students' quality of collaboration using Meier, Spada and Rummel (2007) rating scheme (note that this measure has not been double-coded by a second judge, which is why we mention this finding as anecdotal). A researcher coded the entire collaborative episode using a five point scale on the following eight dimensions: sustaining mutual understanding, dialogue management, information pooling, reaching consensus, task division, task management, technical coordination, reciprocal interaction, individual task orientation. Additionally, an overall score was computed by averaging ratings across all these dimensions. We found the percentage of joint attention to be significantly correlated with students' overall quality of collaboration: $r(24) = 0.432$, $p = 0.027$. More specifically, joint attention was significantly correlated with students' tendency to properly manage the group dialogue $r(24) = 0.427$, $p = 0.03$, reach a consensus $r(24) = 0.517$, $p = 0.007$ and equally divide work between members of the group $r(24) = 0.424$, $p = 0.031$. For comparison, Schneider and Pea (2013) performed the same analyses in a remote collaboration and found a significant correlation between joint attention and collaboration (more specifically, at sustaining mutual understanding, reaching a consensus, managing time and pooling information). This finding replicates previous results showing that joint attention can act as a proxy for students' quality of interaction, and seems to reflect their ability to reach a consensus across different settings.

Discussion

This paper reports two contributions to the CSCL community. First, we describe a methodology for quantitatively capturing joint attention in a co-located setting: we suggest using visual "hand-claps" for synchronizing mobile eye-trackers, and advocate the use of multiple fiducial markers for remapping gazes onto a ground truth. Secondly, we present some early results on the effect of 3D tangibles on students' visual coordination: we found that 3D tangibles seemed to facilitate visual coordination between users, which in turn was associated with higher performances on an optimization task. We have two tentative explanations for this effect: first, 3D objects may have special qualities that 2D objects do not—like shadows, height, and a high level of realism; it is possible that the perceptual richness of the 3D tangible shelves act as "hot spots" that users can exploit to ground their communications. Secondly, it is also possible that 3D objects are simply bigger than 2D objects (in terms of the number of "pixels" they occupy in a visual field); this would make 3D objects more likely to attract attention when one's partner is moving a shelf. Those two hypotheses (quality versus quantity) are non-competing in the sense that both may be true. Future work should try to disentangle those two explanations by looking more closely at the eye-tracking data and the videos recorded during the study.

Additionally, we tried to control for the following confounding variable: it is possible that there is more gaze on 3D objects than on flat objects, and subsequently more chances for the two gazes to be on the same object at the same time in the 3D condition (regardless of students' interactions). Two observations seem to disconfirm this hypothesis: first we did not observe this bias when computing JVA during the memory task, when no collaboration was allowed. In fact, we saw the opposite pattern: students in the 2D condition exhibited slightly more gaze recurrence than students in the 3D condition (non significant difference). Secondly, we found a significant correlation between levels of JVA and our dependent variables (performance task, learning gains, quality of collaboration): this suggests that JVA is indeed an indicator of more productive collaboration, and not just the result of chance or a perceptual bias.

Certain limitations need to be mentioned. First, we did not study a tangible interface *per se*, but a feature common to most TUIs (3D versus 2D representations). It is unclear if our results will generalize to other settings where users can interact with physical objects. Second, detecting joint attention by looking at the distance between two gazes is arbitrary; in future analyses, we would like to detect joint attention on more specific AOIs (i.e., shelves or corridors). Finally, we only presented quantitative results; in the future, we plan to complement those analyses with qualitative observations. As a final note, we acknowledge that the one-on-one correspondence of joint visual attention with quality of collaboration is a bit simplistic. Joint attention can take many forms (e.g., a silent moment of JVA is fundamentally different from a moment of JVA accompanied by a discussion), which should be taken into account for future work. Kaplan and Hafner (2006), for instance, describe a taxonomy that distinguishes between JVA triggered by a salient event, coincidental simultaneous looking, gaze following and coordinated gaze on an object of interest. We plan to use this taxonomy in future work to get a more nuanced understanding of the relationship between joint attention and students' interactions.

Conclusions and implications

This work opens new doors for studying collaboration in co-located settings. Mounted eye-tracking devices are becoming cheaper and more accurate, and we are starting to see attempts to build low-cost mobile eye-trackers. Those technological developments suggest that eye tracking could become ubiquitous in the near future, which motivates the need for developing theoretical and practical frameworks for capturing students' gazes in co-located settings. By nature, those settings are much interesting than a remote collaboration, because most of the interactions between students currently happen in a physical classroom. Furthermore, it has long been unclear whether results from remote social eye-tracking studies generalize to co-located settings. Our findings suggest that, indeed, a higher recurrence of joint attention is correlated with students' task performance, learning gains and quality of collaboration across settings. This further motivates the need to closely pay attention to this construct when considering interactions in small groups of students.

All in all, we believe that the methodology and results described above are promising building blocks for studying visual coordination (and more generally collaborative learning) in small groups of students. We are seeing a lot of potential in this sort of approach, and believe that interesting future work can build on this first attempt at quantifying joint visual attention in a co-located setting.

References

Brennan, S. E., Chen, X., Dickinson, C. A., Neider, M. B., & Zelinsky, G. J. (2008). Coordinating cognition: The costs and benefits of shared gaze during collaborative search. *Cognition*, *106*(3), 1465–1477.

Dillenbourg, P., & Evans, M. (2011). Interactive tabletops in education. *International Journal of Computer-Supported Collaborative Learning*, *6*(4), 491–514.

Falcão, T. P., & Price, S. (2009). What Have You Done! The Role of "Interference" in Tangible Environments for Supporting Collaborative Learning. In *Proc. of the 9th International Conference on Computer Supported Collaborative Learning - Volume 1* (pp. 325–334). Rhodes, Greece.

Farah, M. J., Rochlin, R., & Klein, K. L. (1994). Orientation invariance and geometric primitives in shape recognition. *Cognitive Science*, *18*(2), 325–344.

Hornecker, E., & Buur, J. (2006). Getting a Grip on Tangible Interaction: A Framework on Physical Space and Social Interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 437–446). New York, NY, USA: ACM.

Jermann, P., Mullins, D., Nüssli, M.-A., & Dillenbourg, P. (2011). Collaborative Gaze Footprints: Correlates of Interaction Quality. *CSCL2011 Conference Proceedings.*, *Volume I - Long Papers*, 184–191.

Kaplan, F., & Hafner, V. V. (2006). The challenges of joint attention. *Interaction Studies*, *7*(2), 135–169.

Meier, A., Spada, H., & Rummel, N. (2007). A rating scheme for assessing the quality of computer-supported collaboration processes. *Int. Journal of Computer-Supported Collaborative Learning*, *2*(1), 63–86.

Richardson, D. C., & Dale, R. (2005). Looking To Understand: The Coupling Between Speakers' and Listeners' Eye Movements and Its Relationship to Discourse Comprehension. *Cog. Science*, *29*(6), 1045–1060.

Schneider, B. (2014). The Perceptual Benefits of a Tangible Interface Decrease with Users' Expertise. In *CHI '14 Ext. Abstracts on Human Factors in Computing Systems* (963–968). New York, NY, USA: ACM.

Schneider, B., Jermann, P., Zufferey, G., & Dillenbourg, P. (2011). Benefits of a Tangible Interface for Collaborative Learning and Interaction. *IEEE Transactions on Learning Technologies*, *4*(3), 222–232.

Schneider, B., & Pea, R. (2013). Real-time mutual gaze perception enhances collaborative learning and collaboration quality. *Int. Journal of Computer-Supported Collaborative Learning*, *8*(4), 375–397.

Sowell, E. J. (1989). Effects of manipulative materials in mathematics instruction. *Journal for Research in Mathematics Education*, *20*(5), 498–505.

Tomasello, M. (1995). Joint attention as social cognition. In C. Moore & P. J. Dunham (Eds.), *Joint attention: Its origins and role in development* (pp. 103–130). Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.

Acknowledgments

We gratefully acknowledge support from the National Science Foundation for this work from the LIFE Center (NSF #0835854) as well as the Leading House Technologies for Vocation Education, funded by the Swiss State Secretariat for Education, Research and Innovation. Finally, we would like to thank SMI (SensoMotoric Instruments) for their eye-tracking technology and Jacques Kurzo for his deep involvement in this project.