JrnlID 11412_ArtID 9281_Proof# 1 - 10/07/2018

Intern. J. Comput.-Support. Collab. Learn
https://doi.org/10.1007/s11412-018-9281-2

Leveraging mobile eye-trackers to capture joint visual attention in co-located collaborative learning groups

Bertrand Schneider[1] **· Kshitij Sharma**[2] **·**
Sebastien Cuendet[2] **· Guillaume Zufferey**[2] **·**
Pierre Dillenbourg[2] **· Roy Pea**[3]

Received: 25 October 2017 / Accepted: 4 July 2018

Abstract This paper describes a promising methodology for studying co-located groups: mobile eye-trackers. We provide a comprehensive description of our data collection and analysis processes so that other researchers can take advantage of this cutting-edge technology. Data were collected in a controlled experiment where 27 student dyads ($N = 54$) interacted with a Tangible User Interface. They first had to define some design principles for optimizing a warehouse layout by analyzing a set of Contrasting Cases, and build a small-scale layout based on those principles. The contributions of this paper are that: 1) we replicated prior research showing that levels of Joint Visual Attention (JVA) are correlated with collaboration quality across all groups; 2) we then qualitatively analyzed two dyads with high levels of JVA and show that it can hide a free-rider effect (Salomon and Globerson 1989); 3) in conducting this

✉ Bertrand Schneider
bertrand_schneider@gse.harvard.edu

Kshitij Sharma
kshitij.sharma@epfl.ch

Sebastien Cuendet
sebastien.cuendet@epfl.ch

Guillaume Zufferey
guillaume.zufferey@epfl.ch

Pierre Dillenbourg
pierre.dillenbourg@epfl.ch

Roy Pea
roy.pea@stanford.edu

[1] Harvard University, Cambridge, MA, USA

[2] EPFL, 1015 Lausanne, Switzerland

[3] Graduate School of Education, Stanford University, 485 Lasuen Mall, Stanford, CA, USA

analysis, we additionally developed a new visualization (augmented cross-recurrence graphs) that allows researchers to distinguish between high JVA groups that have balanced and unbalanced levels of participations; 4) finally, we generalized this effect to the entire sample and found a significant negative correlation between dyads' learning gains and unbalanced levels of participation (as computed from the eye-tracking data). We conclude by discussing implications for automatically analyzing students' interactions using dual eye-trackers.

Keywords Joint visual attention · Collaborative learning · Dual eye-tracking

Introduction

Joint visual attention is a prerequisite for virtually all social interactions across cultures and ages. This is especially true for tasks that require collaborators to build a shared problem space. This process is sometimes referred to as grounding from a psycho-linguistic perspective (Clark and Wilkes-Gibbs 1986). Building a common ground ensures that collaborators are on the same page and share a common definition of the terms used. While grounding is useful to explain short conversational events, CSCL researchers suggest going beyond this concept for educational contexts and to focus on shared *meaning making* instead (Stahl 2007). Meaning making is associated with "the increased cognitive-interactional effort involved in the transition from learning to understand each other to learning to understand the meanings of the semiotic tools that constitute the mediators of interpersonal interaction" (Baker et al. 1999, p.31). It gradually leads to the construction of new meanings and results in conceptual change. From this perspective, markers of collaborative learning are captured through iterative cycles of interactions that converge toward a shared understanding of the concepts taught. Researchers in CSCL generally agree that Joint Visual Attention (JVA) plays a central role in supporting students' conceptual convergence (Roschelle 1992). It is worth noting that joint attention is associated with many overlapping concepts in the learning sciences and CSCL - "shared cognition," "intersubjectivity," "grounding processes in conversation," "joint problem-solving," and "distributed cognition" (Barron and Roschelle 2009).

Traditionally, JVA has been studied through qualitative case studies. But we now have the means to go beyond case-by-case evidence of the importance of JVA for collaborative learning. One way of rigorously studying and quantifying this construct is to leverage new sensing technologies. Eye-trackers are becoming increasingly easier to use and more affordable and they allow researchers to collect large datasets on students' visual processes. In this paper, we show how they can be leveraged to better understand collaborative processes.

We report the following contributions to the study of JVA: first, we describe a methodology for synchronizing mobile eye-trackers to study co-located interactions. Second, we replicate previous results showing that high levels of JVA are significantly correlated with the quality of students' interactions. Third, we qualitatively analyze two dyads with high levels of JVA and high / low learning gains to understand why sometimes joint visual attention does not predict learning. Fourth, we present new ways of visualizing JVA with *augmented cross-recurrence graphs* to better understand those differences. Finally, we generalize those results to the entire sample and describe metrics that capture group members' tendency to *initiate* or *respond* to offers of JVA. We conclude by mentioning limitations of our approach and promising future work in the study of collaborative learning using mobile eye-trackers.

Literature review

Dual eye-tracking and joint visual attention in the learning sciences

Joint Visual Attention (JVA) is a central construct for any learning scientist interested in collaborative learning. It is "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, pp. 86–87). Without the ability to establish joint attention, groups are unlikely to establish a common ground, take the perspective of their peers, build on their ideas, express some empathy or solve a problem together. Autistic children, for example, are known to have difficulties in coordinating their attention with their caregivers (Mundy et al. 1990), which in turn is associated with major social disabilities.

Previous studies in CSCL (Computer-Supported Collaborative Learning) and LS (Learning Sciences) have highlighted the importance of joint attention in small groups of students. Barron (2000), for instance, carefully analyzed how groups of students who failed to achieve joint attention were more likely to ignore correct proposals and not perform as well as similar groups. More generally, there is a large body of evidence showing that JVA is a central mechanism for effective collaboration (e.g., the process of building a shared understanding has been extensively studied by psycholinguists under the name of grounding; Clark 1985). In the Learning Sciences, some have called it partner modelling (Dillenbourg et al. 2016). In education, building a shared understanding is really a mean to an end: that is, how does building a common ground lead to learning? Learning scientists suggest that it is actually the cognitive effort of building a shared understanding that produces learning, not merely the fact of sharing a common ground (Schwartz 1995). Linguists, psychologists and learning scientists all agree that Joint visual attention plays a central role in this process.

Over the last two decades, researchers have started to leverage eye-trackers to capture JVA in remote collaborations. Richardson and Dale (2005), for example, found that the degree of gaze recurrence between individual speaker–listener dyads (i.e., the proportion of times that their gazes are aligned) was correlated with the listeners' accuracy on comprehension questions. Jermann et al. (2011) conducted a similar analysis with a dual eye-tracking setup. They used cross-recurrence graphs (described below in Fig. 1) to separate "good" and "bad" collaborative learning groups. They found that productive groups exhibited a specific pattern (shown in Fig. 1's middle graphs), and that less productive groups produced a fundamentally different collaborative footprint (right graph of Fig. 1).

Fig. 1. A cross-recurrence graph reproduced from Jermann et al. (2011). On the left, a schematic graph; in the middle picture, a cross- graph from a productive group; on the right side, from a less productive group

Q4

As a side note, it is worth mentioning that researchers have also started to think about ways to use-eye-trackers to actively support visual coordination in pairs of students. The precursor to this approach is a study by Schneider and Pea (2013), where they used two remote eye-trackers to provide students with a *gaze awareness tool*. They showed that dyads who could see the gaze of their partner in real time outperformed their peers in terms of their learning gains and quality of collaboration. This intervention was beneficial to students because they could monitor the visual activity of their partner in real time, anticipate their contribution, and more easily disambiguate vague utterances. In the same vein, Mason et al. (2015) showed that 7th graders who could see a teacher's gaze when reading an illustrated text learn significantly more than students who could not. D'Angelo and Begel (2017) designed a gaze awareness tool for programmers which shows the snippet of code their partners are looking at. They found that their intervention promoted higher levels of joint visual attention, increased the proportion of implicit to explicit references, and that participants were faster and more successful at responding to them. Finally, Schlösser et al. (2015) found that remote collaboration during a puzzle task could be improved with a gaze awareness tool compared to no gaze support. They also compared a "gaze cursor" (showing where the partner was looking) and a "gaze awareness" condition (highlighting which puzzle piece was currently looked at) and found no difference between them. While this line of work is not directly relevant for the scope of this paper, it provides examples showing that we can support collaborative learning through dual eye-tracking.

There have been very few attempts at using mobile eye-trackers in formal and informal learning environments. Gergle and Clark (2011) pioneered the use of dual eye-tracking for in-situ interactions: they used large fiducial markers in the environment to capture JVA and compared participants who were either sitting or standing. They found significant differences in terms of terms of their linguistic and visual coordination. Papavlasopoulou et al. (2017) studied "kids" (age 8–12) and "teens" (age 13–17) with mobile eye-trackers and found that kids focused more on surface features (e.g., the appearance of the characters), while teens tended to show more hypothesis-testing behavior when coding. They did not look at JVA specifically, but they measured gaze similarity – for example if pairs of participants had similar gaze behaviors (e.g., spatial dispersion). They found that teenagers exhibited higher gaze similarity than kids. Finally, Prieto et al. (2014) are working on leveraging mobile eye-tracking to capture teachers' cognitive load using a variety of measures (e.g., pupil size).

In summary, remote and mobile eye-trackers can provide us with new insights into students' learning processes, and dual eye-tracking setup can help us gain a better understanding of collaborative processes – for example by quantifying joint visual attention. Additionally, some novel work also provides evidence that eye-tracking can be used to design interventions to support communication in remote collaborations.

In the next section, we review in greater detail how researchers have been making sense of dual eye-tracking data, more specifically by generating *cross-recurrence graphs*. We describe them in detail here, because this paper builds upon (and extends) this methodology when analyzing dual eye-tracking data.

Dual eye-tracking settings and cross-recurrence graphs

Previous work done by Jermann and his colleagues (2011) used cross-recurrence graphs to make sense of dual eye-tracking datasets. A cross recurrence graph (shown on Fig. 1) is a visual representation of a dyad's visual coordination: the x-axis represents time for the first participant, the y-axis represents time for the second participant, and black pixels show moments of joint attention (for a given time slice and distance between two gazes). Thus, a black diagonal signifies that the dyad was continuously looking at the same area of interest (e.g., the middle graph in Fig. 1); black squares outside the diagonal mean that both participants looked at the same location, but not at the same time; the absence of a dark diagonal means an absence of joint visual attention (e.g., the right graph on Fig. 1). In other words, good quality interaction exhibits higher recurrence rates of JVA compared to low quality interaction. This difference can be detected visually using cross-recurrence graphs.

We use a similar methodology in this paper, with three new contributions: first, the data comes from a study that looked at co-located interactions. We captured students' gazes using two mobile eye-trackers, whereas prior work has almost exclusively looked at remote interactions (i.e., where students were looking at two different computer screens). This development is a significant improvement in terms of ecological validity, because so much of collaborative learning is among co-located individuals. Second, we augmented cross-recurrence graphs with speech data and spatial information to uncover students' visual attention when working on a problem-solving task. This provided us with compelling visualizations that guided our qualitative analysis. Third, we contrasted two groups that each exhibited high levels of joint visual attention (in contrast with comparing a productive versus a non-productive group, as Jermann and his colleague did). Our goal is to illustrate the multitude of ways that students use to successfully establish joint visual attention. We found that dual eye-tracking datasets can uncover particular types of collaboration dynamics. Finally, we extended qualitative observations to the entire sample and found that dyads' extent of participation imbalance (computed from moments of joint visual attention) was negatively correlated with learning gains.

In the section below, we describe our study and the data it generated. It should be noted that the main contribution of this paper is not the experiment itself (which was published elsewhere – see Schneider et al. 2016). The main contributions of this paper are the analyses of the eye-tracking data and the kinds of insights which were then generated. We describe the original study in more detail below.

General description of the experiment

The goal of the study was to conduct an empirical evaluation of a Tangible User Interface (TUI) designed for students following a vocational training in logistics. The system used in this study, the TinkerLamp (Fig. 4) features small-scale shelves that students can manipulate to design a warehouse layout. A camera detects their location using fiducial markers and a projector enhances the physical layout with an augmented reality layer. The original study compared the affordances of 2D, abstract-looking interfaces (Fig. 3, left side) with 3D, realistic-looking interfaces (Fig. 3, right side) and compared its effect across three levels of expertise (1st year, 2nd year and 3rd year students). See the following paper for more information (Schneider et al. 2016). In this paper, we will *not* focus on the effect of the two experimental conditions and students' expertise. Instead, we aggregate the data across all groups and search for general markers of productive interactions regardless of the interface that students interacted with (2D or 3D) or their prior knowledge.

Method

Participants

54 apprentices in logistics participated in the study (28 in the "3D" condition, mean age = 19.07, SD = 2.76; 26 in the "2D" condition, mean age = 17.96, SD = 1.56). Due to the vocational domain, few women participated (4 in the "3D" condition, 3 in the "2D" condition). All participants were following a vocational training in logistics in Switzerland: 16 first-year, 16 s-year, and 22 third-year ($N = 54$).

Procedure

The general description of the experiment is shown on Fig. 2 above. The activity lasted an hour and the goal for students was to uncover good design principles for organizing warehouses. The core of this reflection involved understanding the trade-off between the amount of merchandise that can be stored in a warehouse and how quickly one can bring an item to a loading dock (i.e., a lot of shelves make it difficult to efficiently load/unload items, while few shelves limit the size of available stock). To help students understand this relationship, we followed the Preparing for Future Learning framework (Bransford and Schwartz 1999). We designed a set of contrasting cases (Fig. 3, first row) and asked students to analyze three layouts based on the following criteria: in which warehouse they would prefer to work on a daily basis (prompt 1); which warehouse maximized space (prompt 2); and finally, which warehouse minimized the distance from each shelf to the docks (prompt 3). At the end, the experimenter revealed the numbers for those two metrics (referred to as "reveal" on Fig. 2). The contrasting cases were set up to highlight good design principles (e.g., orienting the shelves toward the docks makes them more accessible, placing them back to back so that only one side is accessible frees up space, and so on). Students were asked to use those principles in the next activity (Fig. 2, second row), where they built two warehouses: one where they had to put as many shelves as possible in a given layout, and a second one where they had to

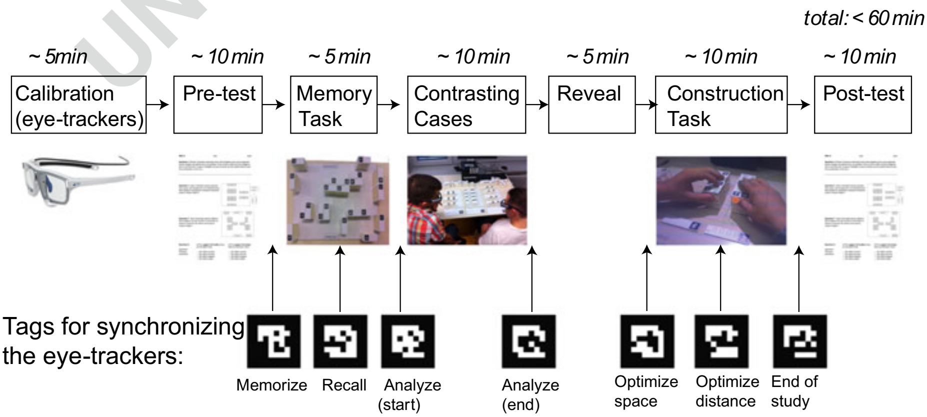

Fig. 2 The experimental design of the study. See the "procedure" section below for more details. The bottom rows show the tags presented to participants to synchronize the different data sources (i.e. the eye-trackers and the Tangible User Interface). The construction task had two subtasks: participant had to first optimize the space available in a warehouse, and then minimize the average distance between the shelves and the docks

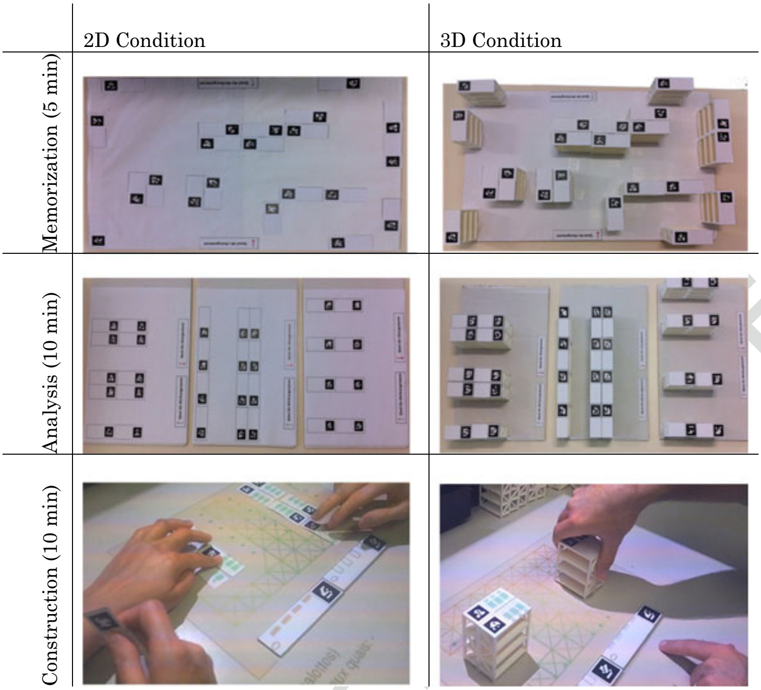

Fig. 3 The main experimental tasks: the 2nd row the contrasting cases students had to analyze; the 2rd row shows the Tangible Interface for the construction task. The "red", "green", "blue" labels are used to color-code additional analyses below (e.g., in the cross-recurrence graphs in Fig. 8)

minimize the average distance to the docks given a certain number of shelves. Before and after those two activities, we gave students a pre and post-test where we asked them to modify the layout of several warehouses to optimize them. We have two main dependent variables: how well they designed the two warehouses, and how well they answered the learning test. During those two activities, students wore two mobile eye-trackers (SMI Eye-Tracking Glasses with binocular pupil tracking at 30 Hz) and we used the scene cameras (1280 × 960 pixels) to record videos for post-processing.

Material

Students interacted with a TUI called the TinkerLamp (Fig. 4). This system combines a camera (to detect the location of the fiducial markers) and a projector (to display the augmented reality layer). The system allows students to build warehouse layouts and receive feedback in real time by combining a physical and digital representation. This system has been used in several schools in Switzerland and has been co-designed with teachers during several iterations of a Design-Based Research (DBR) program. For an exhaustive description of this TUI, please see Zufferey et al. (2009).

The stimuli for each task are described in Fig. 3. 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

JrnlID 11412_ArtID 9281_Proof# 1 - 10/07/2018

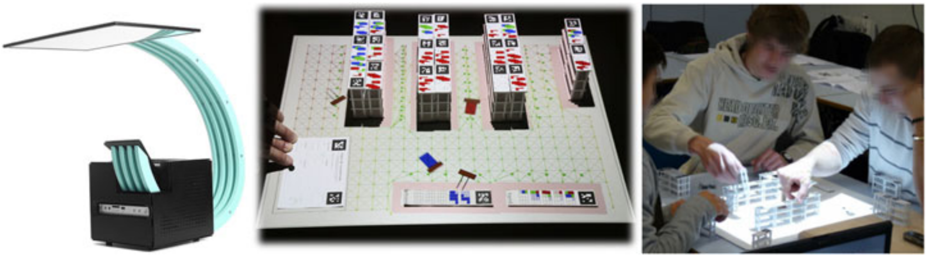

Fig. 4 The TinkerLamp. The left side shows the physical apparatus, that combines a projector and a camera. The middle image provides an example of the augmented reality layer, which shows the stocks left and the forklifts moving back and forth between the docks and the shelves. The picture on the right shows three apprentices interacting with the TinkerLamp during a classroom activity

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. The pre and post-test were similar, except that we changed the configurations of the layouts for questions 2 and 3. Question 1 and 4 were identical on the pre and post-test.

Coding

For the memory task, we counted the number of shelves and docks in the correct location as a retention score. For the construction 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.

Data pre-processing

While the previous section describes outcome measures, we describe the process data generated in this study below. Datasets came from two sources: the TUI and the mobile eye-trackers. We illustrate what the log files look like in both cases.

Data from the mobile eye-tracker

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 30 Hz. Those units are lightweight (75 g.) and can easily be used for hour-long experimental sessions. We also used the scene camera (1280 × 960 pixels) to do fiducial tracking and synchronize the two devices. The mobile eye-trackers generate log files to accompany the movies captured from the scene camera. The log files contain, for each frame of the video, the x and y location (in pixels) of the participant's gaze. Scientific packages for eye-tracking systems also provide researchers with a wealth of information about the subject's gaze: for instance, the pupil's dilation (which can be, in some circumstances, used as a proxy for cognitive load), whether the current data point is a fixation, a saccade or blink, and so on. We provide an example below (Table 1):

Where POR stands for Point of Regard, EPOS for Eye Position, GVEC for gaze vector, L for left, R for right, B for Binocular. In practice, researchers primarily use the timestamp, the x-y binocular POR, and the Event info (i.e., whether the data point is a fixation, saccade or blink). The eye position and gaze vector can be used for reconstructing the scene in three dimensions.

By processing the video from the scene camera with a fiducial tracking engine (Chilitags 2; Bonnard et al. 2016), we were able to generate the log data that described, for each frame of the video, the location of the fiducial markers on that particular frame. For example, one line of this log file would look like this:

20,881,289.29,409.847,831.83,378,852,392,812,438,814,428,852
21,881,291.501,407.195,831.671,378,848,391,815,437,815,425,850
21,915,354.014,1054.95,544.127,1027,536,1065.28,531.986,1084.18,552.626,1044,557

Those numbers correspond to the following fields: frame number, marker Id, angle, center.x, center.y, corners[0].x, corners[0].y, corners[1].x, corners[1].y, corners[2].x, corners[2].y, corners[3].x, corners[3].y. In the example above, the 20th frame of the movie file had one fiducial marker (#881), and the 21st frame had two (#881 and #915). The angle is in degrees and the location of the center and corners of the fiducial marker are in pixels.

Data from the tangible user Interface

The TUI generated log files accompanied by screenshots captured every second from a top-down perspective. The log files contain the x and y location of the fiducial markers detected on the frames of the video (Table 2):

Challenges

The challenges of coordinating these data sources can be summarized by two main issues: aligning the logs temporally (i.e., synchronizing each dataset to make sure that events are happening at the same time) and aligning the data spatially (i.e., transposing the location of each event onto a common space, so that we can tell if they intersect or not). We describe our

Table 1 Example of a line from the eye-tracking log file

Header	Time,Type,Trial,B POR X [px],B POR Y [px],L POR X [px],L POR Y [px],R POR X [px],R POR Y [px],B Object Hit,L EPOS X,L EPOS Y,L EPOS Z,R EPOS X,R EPOS Y,R EPOS Z,L GVEC X,L GVEC Y,L GVEC Z,R GVEC X,R GVEC Y,R GVEC Z,Frame,Aux1,B Event Info
Data	20,916,024,180,SMP,1,626.41,570.39,556.52,444.18,697.15,457.43,-,30.62,0.02,-24.25,-28.19,0.09,-24.81,0.00,-0.22,0.98,0.02,-0.23,0.97,00:00:00:00,Blink

Table 2 Example of a line from the TUI log file

Header	date, timestamp, type, event, fiducial ID, top left corner, top right corner
Data	17–13-2,1,378,480,382,386,tag,added,chili,14,411.764,679.855,507.475,677.629

strategies for solving those two issues below. Figure 5 provides a visual summary of the datasets and challenges.

Temporal alignment

Making sure that the log files are temporally aligned is a challenge when dealing with multiple devices. We solved this problem by presenting fiducial markers with a specific ID at key moments (as shown on Fig. 2). We briefly presented them on the table to both participants and within the field of view of the TUI's camera. By post-processing the video of the eye-trackers' scene cameras and the snapshots of the TUI, we were able to identify *when* those markers were presented. This technique could be thought of a visual "hand clap" commonly used to synchronize movies – with the advantage that participants were not distracted or discomforted

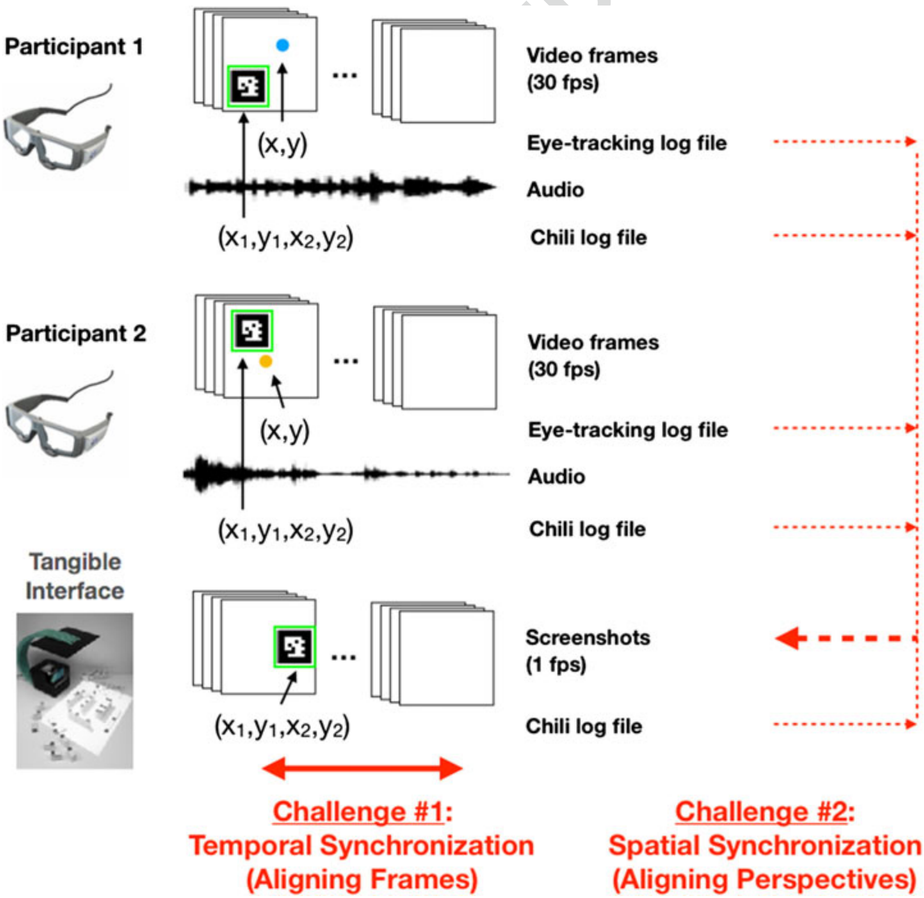

Fig. 5 Visual summary of the challenges involved with analyzing data from mobile eye-trackers

by a loud sound. All of the markers used for synchronization purposes were successfully detected by the fiducial tracking engine (Bonnard et al. 2016).

Spatial alignment

The strategy that we used to spatially align the data from each data source was to first define a *ground truth* for each task. The ground truth is a top-down representation of the scene. For the analysis task, it was straightforward to generate: we simply took a picture of the warehouses' layouts students had to analyze and used the fiducial engine to detect the location of each marker on that picture. For the construction task, we faced an additional challenge: the ground truth was changing every second. We thus needed to re-generate a new ground truth with the location of the fiducial markers every second. Figure 6 shows examples of ground truths for the analysis and construction tasks.

By using the location of the fiducial markers detected from the two mobile eye-trackers, we were able to remap participants' gaze onto the ground truths. This mathematical operation is called a *homography* and is widely used in computer vision. More concretely, having a common set of points (i.e., the location of the fiducial markers) in two different images allows the researcher to infer the location of another set of points (i.e., the gaze points) which location are only known on one perspective (i.e., the video frames of the eye-tracker's scene camera). Figure 7 summarizes those operations.

Computing a metric for joint visual attention

After solving the two challenges above, we needed to consider two additional parameters when computing joint visual attention from eye-tracking data. First, participants' visual

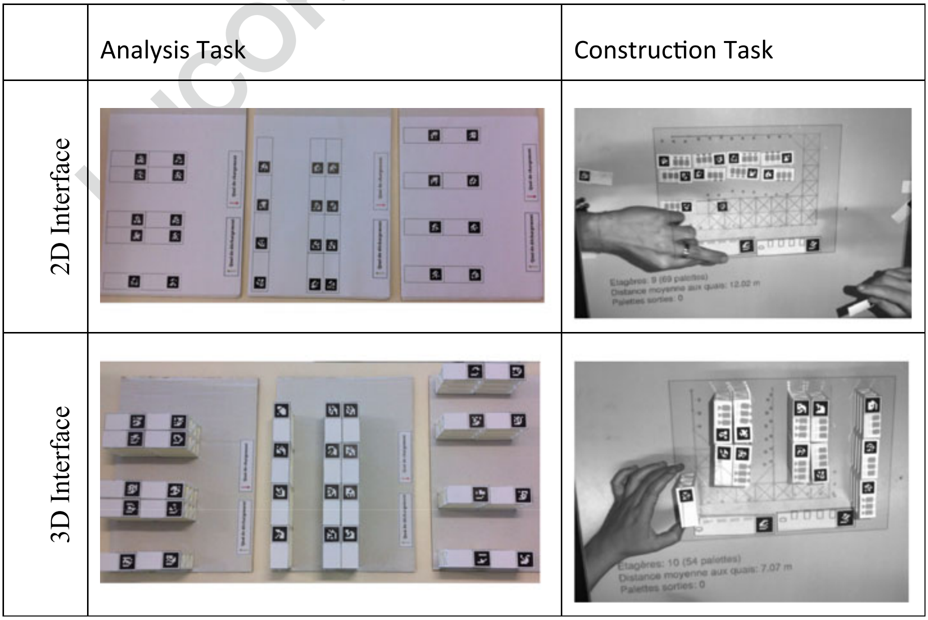

Fig. 6 Ground truths used for the analysis and the construction tasks

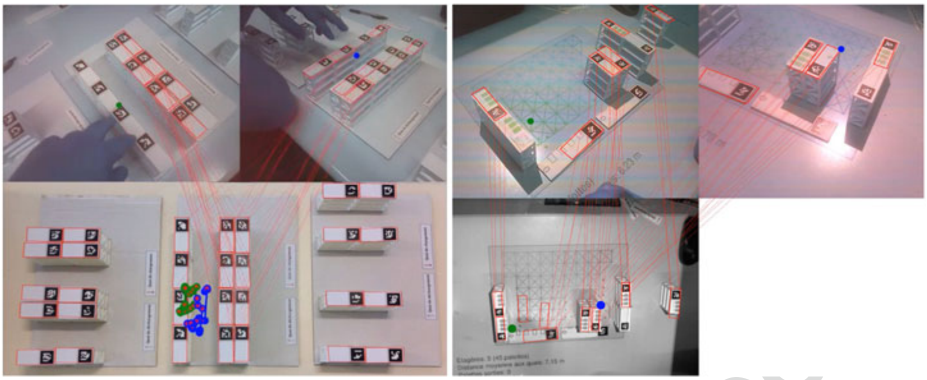

Fig. 7 Examples of a homography for each task (the analysis task is shown on the left; the construction task is shown on the right). The red lines show the common set of points across perspectives (i.e., the shelves) used for the homography and the blue / green dots show the inferred location of the gaze points onto the ground truth (bottom row)

attention is rarely perfectly synchronized. In an early 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 s. For this reason, we consider a lag of ± 2 s when computing our JVA metric. A second parameter for consideration is the threshold for the distance between two gazes. Jermann et al. (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: we looked at each gaze point from the first participant and tried to find a corresponding point from the second participant using a time window of ± 2 s in a radius of 50 pixels (which roughly corresponds to the width of a shelf). In a different publication, we experimented with different time windows and radiuses and found consistent results with different values for those parameters (Schneider et al. 2016).

Finally, since the number of data points varied widely between participants, we divided our measure for joint attention by the total number of gaze points of each participant to obtain a percentage of joint attention over the entire activity. This prevented us from inflating the joint attention score of participants who 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).

Results

Capturing JVA with mobile eye-trackers

We summarize here the main results pertaining to our metric of JVA, and how it relates to outcome measures (task performance, learning gains, quality of collaboration). Those results are more precisely described in a different publication (Schneider et al. 2016).

The JVA metric was predictive of students' performance on the construction 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 construction 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 construction 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 rated students' quality of collaboration using Meier et al. (2007) rating scheme. 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. A second judge double-coded ~20% of the videos (6 groups). Inter-reliability index using Krippendorff's alpha was 0.83 (an alpha higher than 0.8 is considered as a reliable agreement between judges; Hayes and Krippendorff 2007). We found the percentage of joint attention was 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 manage 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 consensus.

Comparing two dyads with high levels of JVA

In an earlier publication, Jermann et al. (2011) contrasted two pairs who participated in a dual eye-tracking study. Their goal was to show how good groups differed from less productive groups in terms of their visual coordination. In this section, we decided to focus on two pairs that achieved high levels of JVA in our study. We want to better understand the multitude of ways that students can synchronize their attention. We also chose two groups that performed above average on the problem-solving tasks—but *not* on the learning test. Table 3 summarizes key information about the two dyads:

After running the study, we found that the post-test was harder than the pre-test. This explains why some learning gains are negative (e.g., group 20 in Table 3).

Table 3 Data of group 13 and 20

	Task 1	Task 2	Learning gains	JVA	Speech
Group 13	13 shelves	7.47 m	7.5 points	24%	409 s
Group 20	18 shelves	7.39 m	−0.5 points	29%	540 s
All groups	M = 13.5 SD = 2.9	M = 7.6 SD = 1.9	M = 1.3 SD = 3.2	M = 15.7 SD = 8.0	M = 405.7 SD = 245.7

Augmenting cross-recurrence graphs

We used the information from the eye-tracking logs and the data from Table 3 to build a new visualization: an augmented cross-recurrence graph (Fig. 8). As a reminder, a cross-recurrence graph displays moments of joint visual attention over the entire activity. The x-axis represents time for the first participant and the y-axis represent time for the second participant. In our case, it should be noted that the origin (0,0) is on the top left corner – as opposed to the bottom left corner used in Fig. 1. Thus, a dark diagonal indicates *synchronized* moments of JVA. Dark dots outside the diagonal show when participants revisited areas of interest at a different point in time (e.g., in Fig. 8, a black dot on the top right corner means that the first participants

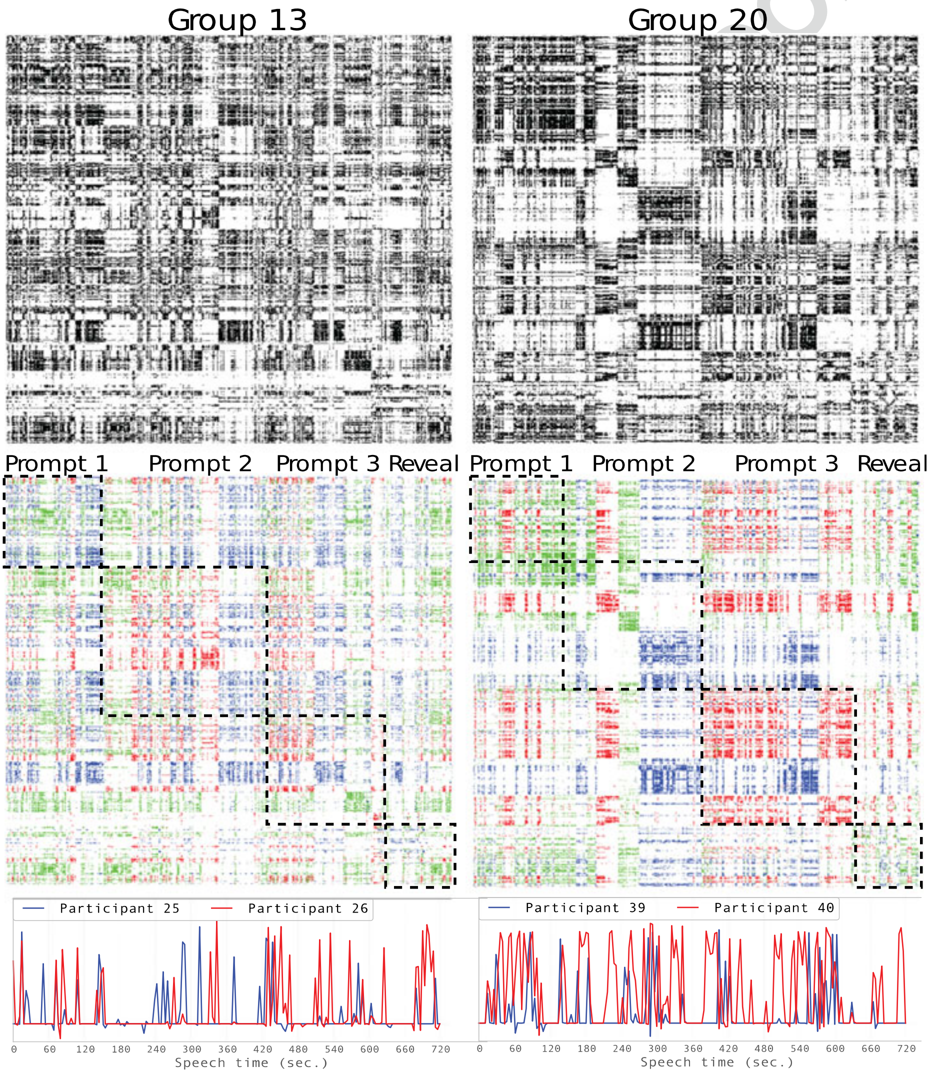

Fig. 8 Cross recurrence graphs of two high-performing groups (13 and 20). The top figures show standard cross-recurrence graphs. The middle figures show colored graphs (red pixel = JVA on the 1st warehouse; green = JVA on the 2nd one; blue = JVA on the 3rd one). The bottom figures show speech data for each dyad

looked at a particular location at the end of activity and the second participant looked at the same area at the beginning of the activity).

We extracted speech information from the audio data to collect the number of seconds that each student spoke. JVA and speech data are displayed in Fig. 8 for groups 13 and 20: on the top, we first show the traditional black and white cross-recurrence graphs used in prior work (e.g., Jermann et al. 2011); on the bottom, we show our cross-recurrence graphs augmented with spatial information (i.e., which warehouse the two students were jointly looking at) and speech data (who spoke when during the analysis task).

The graphs of Fig. 8 provide us with several interesting bits of information about groups 13 and 20. First, we can notice that the traditional cross-recurrence graphs are ideal for identifying moments of joint visual attention (black squares along the diagonal). Based on those graphs, one would predict that group 20 has a better visual coordination than group 13: the diagonal is darker, with more well-defined squares, meaning that this group jointly looked at the same area on the table more often and had longer moments of joint attention. Prior work suggests that an unfilled cross recurrence graph is likely to indicate a poor collaboration between group members. However, we would like to suggest that a "good" cross-recurrence graph (i.e., with a dark diagonal) is not necessarily indicative of a good learning group. We will illustrate this difference by exposing some of the diversity that exists between groups of highly visually-coordinated students.

Our augmented cross-recurrence graphs (middle section of Fig. 8) color-codes each pixel to show which warehouse groups 13 and 20 were looking at (red pixels represent JVA on the 1st warehouse, green on the second one, and blue on the last one). We also added dotted squares to show when the experimenter gave various prompts to the groups (i.e., prompt 1 = "in which warehouse would you prefer to work", prompt 2 = "which warehouse maximizes space, and why", prompt 3 = "which warehouse minimizes the average distance from each shelf to the docks and why", and reveal = "now I will show you numbers that answer those two questions to verify if your intuition was correct"). Two very different strategies appear: group 20 analyzed the three warehouses in a serial manner. For instance, after receiving prompt 2, they jointly looked at the first warehouse, then the second one, and finally at the last layout. Group 13, in comparison, continually compared the three warehouses. We do not have empirical evidence that one strategy is superior to the other, but common sense and the PFL framework would suggest that multiple comparisons between contrasting cases would increase students' chances of noticing important features distinguishing those layouts. One indication that this strategy might be more beneficial to learning is that group13 achieved higher learning gains on the test; indeed, they achieved the highest learning gains in the entire sample. We plan to quantify this behavior in future work to see whether it positively correlates with learning across all groups who participated in this study.

Coming back to groups 13 and 20, there is one last piece of information worth mentioning. We know from Table 3 that group 20 talked slightly more than group 13 (540 vs 409 s.). On the bottom of Fig. 8, we show participants' speech data. One can observe that group 20 had more imbalanced participation (participant 39 talked for 117 s. vs 423 s. for participant 40) compared to group 13 (in which participant 25 talked for 170 s. vs 239 s. for participant 26); thus, one group 20 student talked 78% of the time whereas one group 13 student talked 58% of the time. Might these differences be symptomatic of collaborative issues?

To illuminate those results comparing groups 13 and 20, we qualitatively examined the videos of the experiment. It should be noted that those differences (strategy used, learning gains and speech distribution) are already a striking contrast to the supposedly superior (black and white) cross-recurrence graph of group 20.

Qualitative analysis

We further compared our two groups by analyzing videos of their interactions. We generated videos by combining the scene cameras of the students (top left section of Fig. 9) and remapping their gazes onto a ground truth during task 2 (bottom left section of Fig. 9). We also added an animation of the cross-recurrence graph on the right side of Fig. 9, showing the graph up to that particular video frame. This video enabled a multi-modal analysis of the students' interactions, in particular by highlighting how they combined gestures and speech to achieve and sustain joint attention. Groups #13 and #20 were extremely similar in that regard, constantly using pointing gestures to ground their verbal contributions. It is reasonable to believe that those deictic gestures played a large role in increasing their levels of joint visual attention.

In the table below, we focus on the major differences that we observed for the two groups. Specifically, we focused on one behavior found to be an important predictor of a group's success: how peers react to one another's proposals (Barron 2000). A proposal can be accepted, rejected, challenged, or ignored. One key difference between groups 13 and 20 was that one dyad was more likely to uncritically accept proposals, while the other was much more likely to challenge them (keywords highlighting this difference are marked in bold in Table 4):

We found that this difference was a recurring pattern in the videos and transcripts of groups 13 and 20. For group 20, participant 40 was extremely verbose and dominated the interaction by using many pointing gestures to illustrate his thought process. This was crucial for the group, because it allowed his peer to maintain joint attention on the warehouse layouts. Participant 39, on other hand, was very passive and very rarely contradicted his partner. There are several quite different ways to characterize Participant 39's dyadic pattern with Participant 40. His behavior could be described by the "free rider effect" in teamwork identified by Salomon and Globerson (1989). In a disparate characterization, one could describe the subdominant dyad member as acquiescing to the control of the dominant dyad member, whose proposals reigned. Either way, the ideas generated by group 20 did not significantly change over the course of this activity because Participant 40's proposals were rarely challenged or contradicted. These two different characterizations of Participant 39 vis-à-

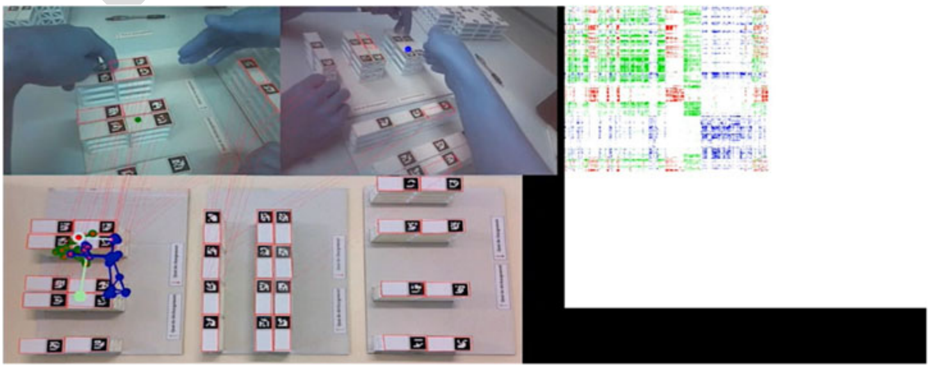

Fig. 9 A video frame generated for the qualitative analysis. On the top left, we show the perspectives of the students with their gazes in blue and green. On the bottom left, we remapped their gazes onto a ground truth (red dots show moments of JVA). On the right side, we show the cross-recurrence graph up to that frame

t4.1 **Table 4** Excerpts of group 13 and 20's dialogues. (e.g., E = experimenter, 25 = participant 25)

	Group 13	Group 20
t4.2	Group 13	Group 20
t4.3	08:24 -- > 09:38 E: The first question is, in which warehouse would you rather work, and why? 25: This one! 26: Yeah this one! 25: if you look at this one, you have less palettes than that one 26 Because of the width here... 25: Not really! 26: yes, it's wider 25: but wait, here that's 4 shelves 26: Here you can't get it out 25: but you can't get from behind 26: yes, that's what I'm saying […] 25: 1,2,3,4 you're crazy. You have more space here.	00:08 -- > 01:30 E: The first question is, in which warehouse would you prefer to work, and why? 40: I think this one is good (2nd warehouse), because you can use half of the warehouse for loading and the other half for unloading 39: yes but you can go faster in this one (1st warehouse) 40: yes because in this one you only have space for two palettes in front of the docks. It's a little tight. I think that I would prefer to work in this one (2nd warehouse). Maybe I would like this one too 39: there's enough space between the shelves 40: yes, and in this one [pointing at the 1st warehouse] the shelves are too tight
t4.4	12:40 -- > 13:52 E: if you could change something in each warehouse, how would you improve them? 26: Hmmm so... 25: I would add two shelves, like that, there. Otherwise... 26: What if we put those like that... to add more shelves 25: No; you can add two here, that's 18 additional palettes 26: Otherwise... one here 25: No, because then if you take a palette from here, you have to back off like that, even if someone's coming from that direction	08:12 -- > 09:11 E: if you could modify those warehouses to minimize the average distance to each shelf, what would you do? 40: What would I do? For instance, in this one (3rd warehouse), I would move those shelves to the corners 39: yes right here 40: This one and that one, I put them here, and those two (in the middle), I would put them there 39: yeah 40: No no, this doesn't make sense. It doesn't change anything. I was thinking, those two you put them there 39: yes 40: but then you're too far away from this shelf 39: well, it's not too bad.

vis Participant 40 have quite different connotations. The 'free-rider' label has been interpreted broadly as 'loafing behavior' (Salomon and Globerson 1989), a negative attribution, whereas an acquiescence interpretation is not a critique of the relatively passive dyad member, but a highlighting of the dominance relationship in their pattern of interaction and reasonably interpretable as conflict avoidance. Different strategies of intervention (or re-mediation) would be warranted for these two distinctive interpretations, 'free-rider' effect or 'partner dominance' effect. A 'free rider' interpretation could simply prompt the passive participant about the need for balanced participation, whereas an acquiescence or 'partner dominance' interpretation would recommend taking effort to ensure that your viewpoint is reflected in your dyad's joint work.

Group 13 provides a strong contrast to Group 20's asymmetrical dynamic. In this group, we found that participant 26 tended to act like participant 40 (i.e., by driving the interaction and assuming a leadership role). This was mostly manifested by the amount of speech shown on the bottom of Fig. 8. But participant 25 was not the kind of relatively passive participant that participant 39 was; whenever he did not agree with a proposal, he challenged it until the group reached a consensus. Those challenges were often initiated by using observations made on other layouts, which explains why the group was more likely to have some back and forth on different warehouse layouts (as shown on the colored cross-recurrence graph in Fig. 8). Thus, a continuous refinement of their proposals seemed to be beneficial to their conversation, while group 20’s reflection stayed on a more superficial level.

Detecting imbalances of participation in the eye-tracking data

The next question is whether this asymmetrical collaboration dynamic behavior pattern can be detected from the eye-tracking data. We propose one approach here; other measures might provide the same result. We started by identifying, for each moment of joint attention (i.e., each red dot on Fig. 9), which student initiated that episode. In our case, we define the initiator as the person whose gaze was most present in this area during the previous second. While we recognize that this definition is arbitrary, it is arguably a reasonable first step in capturing leadership behaviors. We then computed the proportion of the JVA moments that each student initiated, and then applied this proportion to the percentage of JVA moments achieved by the group. We took the absolute value of the difference between the score of each participant in a group to represent the (im)balance of a group’s “visual leadership”. As an illustration, a group may achieve joint attention during 25% of their time collaborating together; let us say that one student initiated 5% of those moments of JVA, while the other student initiated 20% of those moments. Following the computation above, this group would receive a score of abs(0.05–0.20) = 0.15. This measure is shown on the x-axis of Fig. 10.

Points on the right side of the graphs (with higher values) represent groups where one person was more likely to initiate more moments of JVA; points on the left side of the graph (with values closer to zero) represent groups where both students were equally likely to either respond or initiate a moment of JVA. The y-axis shows learning gains computed at the group level. We found a negative correlation between dyads’ learning gains and the absolute difference of students’ visual leadership during the analysis task: $r(23) = -0.33$, $p = 0.10$ and the construction task: $r(23) = -0.47$, $p = 0.02$. This negative correlation suggests that groups who did not equally share the responsibilities of initiating and responding to offers of JVA were less likely to learn. *This result shows that we could potentially recognize a form of the “free-rider”or “partner dominance” effect by looking at the eye-tracking data in a collaborative setting.*

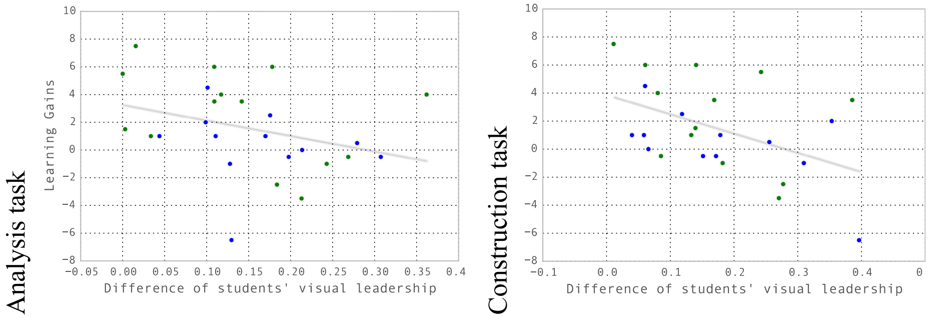

Fig. 10 Negative correlation between students’ learning gains and their visual leadership (i.e., the difference between the percentage of moments of JVA initiated by each participant). Left side shows the scatter plot for the analysis task ($r = -0.33$, $p = 0.10$) right side shows the scatter plot for the construction task ($r = -0.47$, $p = 0.02$). Green dots are dyads in the 3D condition, blue dots dyads in the 2D condition

Discussion

This paper makes four contributions to the study of collaborative learning groups. First, it presents a methodology to process the data coming from two mobile eye-trackers and capture levels of Joint Visual Attention in co-located settings. Previous work (e.g., Jermann et al. 2011; Richardson and Dale 2005) studied remote collaborations where participants were in different physical spaces and collaborated remotely. This makes our contribution more ecological and opens new doors for analyzing face-to-face and side-by-side interactions. Second, we qualitatively analyzed the interactions of two dyads with high levels of JVA and described a case where it produced low learning gains: we found that the second student was less likely to challenge his peer's proposals, which prevented the dyads from refining their understanding of warehouse management. Third, we showed that augmenting cross-recurrence graphs with spatial and verbal information provides researchers with new insights regarding students' strategies and interactions: color-coding them suggests whether a dyad is working in a serial or parallel manner when they are analyzing contrasting cases. We hypothesize that the latter strategy is more beneficial to learning, as measured by our pre and post-tests. Furthermore, adding speech data to the graphs was crucial in our analyses. It allowed us to visually detect imbalances in the group's interactions and dig deeper into their discussion. This observation spawned the paper's fourth contribution: we found that highly coordinated dyads (as measured by dual eye-trackers) were not necessarily the best learning groups. Augmented cross-recurrence graphs revealed imbalances in students' verbal contributions, which can be characteristic of the free-rider effect where one student does most the work while his/her partner stays passive. Finally, we extended those results to the entire sample and found a negative relationship between students' learning gains and their tendency to share the responsibility of initiating and responding to offers of joint visual attention. This shows that learning is promoted by productive interactions between participants; and, more specifically, that balanced levels of "visual leadership", a form of dyadic mutuality, seems to indicate exchanges where both group members are significantly contributing to the discussion. We thus build on collaborative learning work by Barron (2000) who defines a 'mutuality' form of coordination in groups as 'reciprocity with potential for all members to meaningfully contribute' (p. 429).

Conclusions and implications

The implications of this work are that cross-recurrence graphs are highly valuable for distinguishing between productive and unproductive groups. But they should ideally be complemented with spatial and verbal information to provide a more refined multi- representation of the multiple modalities of a group's interactions. Past a certain threshold, high levels of joint visual attention make higher learning gains possible, but they do not guarantee them. There are multiple ways in which student dyads can establish and maintain joint visual attention, which is a necessary but not sufficient condition for productive interactions (for example socio-cognitive conflicts). This paper makes a first step in detecting the absence of those conflicts in visually coordinated students—by identifying *imbalances* in students' tendency to initiate or respond to offers of joint visual attention. This metric could potentially be used to detect groups who would benefit from formative feedback or interventions to regulate group behavior.

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.

References

Barron, B. (2000). Achieving coordination in collaborative problem-solving groups. *The Journal of the Learning Sciences, 9*(4), 403–436.

Barron, B., & Roschelle, J. (2009). Shared cognition. In E. Anderman (Ed.), *Psychology of classroom learning: An encyclopedia* (pp. 819–823). Detroit: Macmillan Reference USA.

Bonnard, Q., Lemaignan, S., Zufferey, G., Mazzei, A., Cuendet, S., Li, N., & Dillenbourg, P. (2016). Chilitags 2: Robust fiducial markers for augmented reality and robotics.

Bransford, J. D., & Schwartz, D. L. (1999). Rethinking transfer: A simple proposal with multiple implications. *Review of Research in Education, 24*, 61–100.

Clark, H. H. (1985). Language use and language users. In G. Lindzey & E. Aronson (Eds.), *Handbook of social psychology* (3rd ed., pp. 179–231). New York: Harper and Row.

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L. B. Resnick, J. Levine, & S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 127–149). Washington, DC: American Psychological Association.

Dillenbourg, P., Baker, M., Blaye, A., & O'Malley, C. (1996). The evolution of research on collaborative learning. In P. Reimann & H. Spada (Eds.), *Learning in humans and machine: Towards an interdisciplinary learning science* (pp. 189–211). Oxford: Elsevier.

Dillenbourg, P., Lemaignan, S., Sangin, M., Nova, N., & Molinari, G. (2016). The symmetry of partner modelling. *International Journal of Computer-Supported Collaborative Learning, 11*(2), 227–253.

Gergle, D., & Clark, A. T. (2011). See what i'm saying?: Using dyadic mobile eye tracking to study collaborative reference. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work* (pp. 435–444). ACM.

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication methods and measures, 1*(1), 77–89.

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.

Mason, L., Pluchino, P., & Tornatora, M. C. (2015). Eye-movement modeling of text and picture integration during reading: Effects on processing and learning. *Contemporary Educational Psychology, 14*, 172–187.

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

Mundy, P., Sigman, M., & Kasari, C. (1990). A longitudinal study of joint attention and language development in autistic children. *Journal of Autism and Developmental Disorders, 20*(1), 115–128.

Papavlasopoulou, S., Sharma, K., Giannakos, M., & Jaccheri, L. (2017). Using eye-tracking to unveil differences between kids and teens in coding activities. In *Proceedings of the 2017 Conference on Interaction Design and Children* (pp. 171–181). ACM.

Prieto, L. P., Wen, Y., Caballero, D., Sharma, K., & Dillenbourg, P. (2014). Studying teacher cognitive load in multi-tabletop classrooms using mobile eye-tracking. In *Proceedings of the Ninth ACM International Conference on Interactive Tabletops and Surfaces* (pp. 339–344). ACM.

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*(6), 1045–1060.

Roschelle, J. (1992). Learning by collaborating: Convergent conceptual change. *The Journal of the Learning Sciences, 2*(3), 235–276.

Salomon, G., & Globerson, T. (1989). When teams do not function the way they ought to. *International Journal of Educational Research, 13*(1), 89–99.

Schlösser, C., Schlieker-Steens, P., Kienle, A., & Harrer, A. (2015). *Using real-time gaze based awareness methods to enhance collaboration. In CYTED-RITOS International Workshop on Groupware* (pp. 19–27). Cham: Springer.

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

Schneider, B., Sharma, K., Cuendet, S., Zufferey, G., Dillenbourg, P., & Pea, R. (2015). 3D tangibles facilitate joint visual attention in dyads. *CSCL2015 Conference Proceedings., I*, 158–165.

Schneider, B., Sharma, K., Cuendet, S., Zufferey, G., Dillenbourg, P., & Pea, R. (2016). Unpacking the perceptual benefits of a tangible Interface. *ACM Transactions on Computer-Human Interactions (TOCHI), 23*(6), 39.

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

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: Lawrence Erlbaum.

Zufferey, G., Jermann, P., Do-Lenh, S., & Dillenbourg, P. (2009, September). Using augmentations as bridges from concrete to abstract representations. In Proceedings of the 23rd *British HCI Group Annual Conference on People and Computers: Celebrating People and Technology* (pp. 130–139). British Computer Society.