Multimodal Transcript of Face-to-Face Group-Work Activity Around Interactive Tabletops

Xavier Ochoa, Katherine Chiluiza, Roger Granda, Gabriel Falcones, James Castells and Bruno Guamán
Escuela Superior Politécnica del Litoral, ESPOL
{xavier,kchilui,roger.granda,gabriel.falcones,james.castells,bruno.guaman}@cti.espol.edu.ec

ABSTRACT: This paper describes a multimodal system around a multi-touch tabletop to collect different data sources for group-work activities. The system collects data from various cameras, microphones and the logs of the activities performed in the multi-touch tabletop. We conducted a pilot study with 27 students in an authentic classroom to explore the feasibility of capture individual and group interactions of each participant in a collaborative database design activity. From the raw data, we extracted low-level features (e.g. tabletop action, gaze interaction, verbal intervention, emotions) and generated some visualizations as annotated transcripts of what happened in a work session. We evaluated teacher's perception about how the automated multimodal transcript could potentially support the understanding of group-work activities. Results from teachers' perceptions pointed out that the multimodal transcript could become a valuable tool to understand group-work rapport and performance.

Keywords: multimodal transcripts, collaboration, group-work visualizations

1 INTRODUCTION

With the evolution of multi-user tabletop devices, the opportunities to enhance collaboration in several contexts have been extended significantly, especially in collaborative learning contexts. Several authors have studied the effect of introducing this particular technology in collaborative sessions, reporting positive results on enhancing communication skills between participants (Kharrufa, et al. 2013; Heslop, 2015). The data-capture capabilities of tabletops in learning contexts present new opportunities to better understand the collaboration and learning processes during group-work activities in the classroom. For instance, collaboration interactions gathered from a tabletop setting could help teachers by making group-work orchestration easier (Martinez-Maldonado et al., 2011) or help students to reflect about their collaboration experience. However, using only the data produced by the tabletops provide a narrow picture of those processes because students interact through a variety of modes (speech, gaze, posture, gestures, etc.) and not all of the actions are perceived or recorded by the tabletop software (Martinez-Maldonado et al. 2017). A common approach to obtain a more holistic view of the collaboration is to complement the data captured by the tabletops with the capture and analysis from other sources such as video recordings (Al-Qaraghuli, 2013). However, the visualization of these data, especially if several communication modalities want to be captured, could become cluttered and confusing for teachers who seek to provide instant feedback to students after a group-work activity.

In this sense, an automatic multimodal transcript has proven to be an efficient and comprehensive method to represent and visualize temporal information from several sources (Bezemer & Mavers,

2011). Thus, combining multimodal collaboration features into a time-based visualization could help teachers to make sense of collaboration processes through the observation of students' actions and emotions in the group-work activity (Martinez-Maldonado et al., 2011; Tang et al., 2010). Even though some studies have added new dimensions of collaboration to the data obtained from tabletop, most of the efforts to create automated transcripts from group's interactions have been focused on one or two modalities. For example, Martinez-Maldonado et al. (2013) presented an approach to identify common patterns of collaboration by mining student logs and detected speech. In another work, Adachi et al. (2015) captured and visualized gaze and talking participation of members in a co-located conversation to provide feedback that in turn would help balancing participation. These studies serve as a baseline for our analysis; however, we want to explore the potential of generating an automated transcript by combining multiple modalities to inform teachers about groups' interactions around a tabletop. Besides, in a classroom interaction research context, the focus has been almost exclusively on teacher talk or teacher–student talk and not in student-student talk in group-work e.g. student-student rapport building in group-work (Ädel, 2011). Ultimately, we want to know if it is possible for teachers to determine more evidence about collaboration, such as group rapport from the transcripts generated.

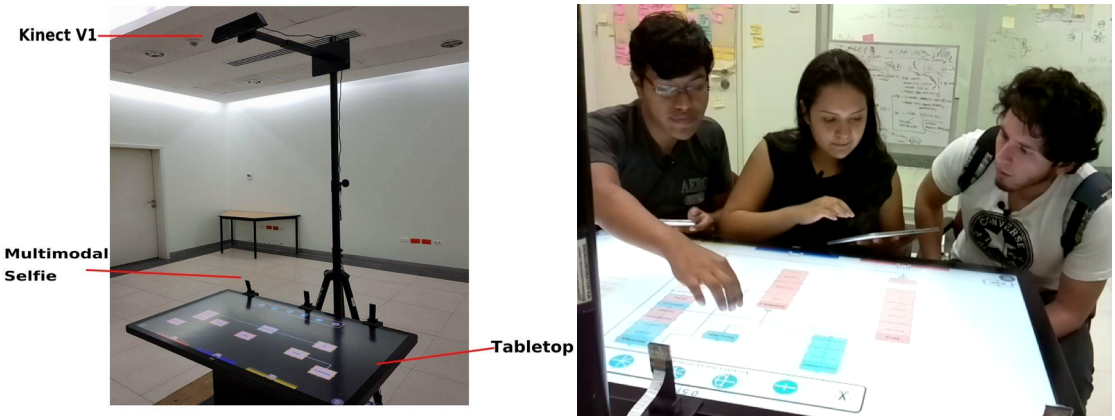

Figure 1: Components of the multimodal interactive tabletop system

2 MULTIMODAL INTERACTIVE TABLETOP SYSTEM

The system is a variation of a prototype presented by Echeverria et al. (2017), that fosters the collaborative design process. Besides the sensors used in the previous version (Kinect V1, coffee table and three tablets), this new version adds three multimodal selfies (Dominguez et al., 2015) with lapel microphones attached for capturing individual speech. The multimodal selfies are located around the tabletop to capture synchronized video and audio for each participant. Additionally, a multimodal selfie was used to capture the video of the entire group session. Figure 1 shows the components of the system and the working prototype.

The software of the system has three different applications: a tabletop application, a management web application, and a recording application. The tabletop application was designed following the design principles described in Wong-Villacres et al. (2015). It allows the participants to develop a database design through the creation and modification of several interactive objects (entities, attributes, relations). The management web application communicates with the tabletop application

and with the multimodal selfies to control the execution of the session start the recordings. It also allows the instructor to view the solution developed by the students (see Echeverria, et al. 2017 for details). The recording application was deployed in each multi-modal selfie. It controls the synchronization of the recordings by the implementation of a publish/subscribe solution using the lightweight MQTT connectivity protocol[1]. The recordings obtained are further processed to automatically tag them according to the proposed audio and video features (see section 2.1). All logs and features are stored in a relational database. In addition, raw audio and video data are saved in a NAS Server, associated with a code for identifying each student.

2.1 Multimodal Transcript

The multimodal transcript combines a set of automatic features (extracted from video, audio and tabletop action logs) into a timeline where the teacher can observe the moment each interaction took place, and how the group session developed through time. The following sections present the set of features and details of the multimodal transcript.

2.1.1 Audio Features

From the speech recorded by the system, we used a Speech-to-Text recognition software, to obtain an automated transcript of the conversation among participants in the group. Then, we extracted the speech sections from the recorded individual audio of each participant, and then converted to text using Google's Cloud Speech API[2]. In this way, we obtained the conversation between the participants along with the time each verbal interaction took place. Google's API results using Spanish language are not as accurate as results obtained using English language. In spite of that, it is still useful to retrieve sentences with words related to the design problem.

2.1.2 Video Features

Mutual gaze and smiles has been considered as non-verbal indicators of rapport in previous work, (Harrigan et al., 1985). Thus, we believe that those features could be a valuable feature to be depicted in the multimodal transcript. Key points from the face of each participant recorded by the Multimodal Selfie were extracted using the OpenPose Library (Cao et al., 2016). This library retrieves the coordinates of 20 points (e.g. eyes, nose, ears, etc.). These face key points were analyzed on every frame of the recordings, and an algorithm was developed to automatically estimate the moments when a participant is looking towards to another. This evaluation was carried out by counting false positives and false negatives from all detections made by the algorithm. To evaluate the accuracy, we considered 5 different videos of 5 minute-length from the original group sessions (see section 4 for details). According to our evaluation, this feature has an error rate of 15.1%. In addition, we extracted the emotions each participant demonstrated during the activity. Video frames of individual recordings from the Multimodal Selfie were processed using the Microsoft Emotion API[3]. Thus, for every second, one frame of the participant's face is sent to the API, which returns an array of scores determining levels of happiness, anger, disgust, among others, with values from 0 to 1. To evaluate the accuracy of the emotion recognition software, videos of three students (25 min approx.) were randomly

[1] http://mqtt.org/

[2] https://cloud.google.com/speech/

[3] https://azure.microsoft.com/en-us/services/cognitive-services/emotion/

selected from all the groups that participated in a pilot study (see section 4 for details). We selected happiness as the emotion to be evaluated because it was the most common detected emotion in the recorded sessions. A human evaluated the videos by annotating if the student was happy or not for each second. Since the API returns values between 0 and 1, we selected a threshold above 0.5 to determine if the student's emotion corresponded to happiness. Our evaluation resulted in an average error rate of 1.77%.

2.1.3 *Interactive Tabletop Logs*

All the interactions with the objects on the tabletop were recorded in a database. Each interaction is represented by the following features: type of interaction (CREATE, EDIT, DELETE), student identity, timestamp, and the type of object the student created. The solution proposed by Martínez, R. et al. (2011) was used for student differentiation while interacting with an object on the tabletop. Figure 2 shows an excerpt of a group session captured by the system. As we can see, different features of the group's interaction are represented (e.g. tabletop actions, gaze, etc.) in a vertical timeline. For instance, we can observe that in t = 1, student 1 (S1) and student 2 (S2) were looking to the right, student 3 (S3) was looking to the left, and so on.

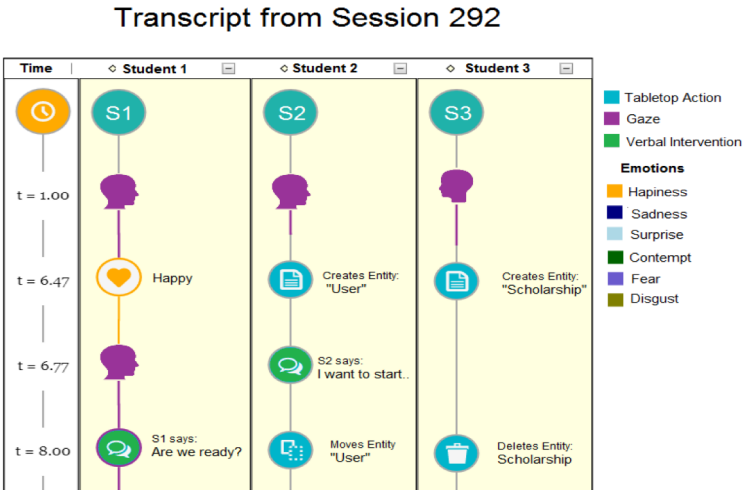

Figure 2: An excerpt of the multimodal transcript from a group

3 PILOT STUDY

The purpose of this study was to validate with teachers the results obtained from the proposed multimodal transcript gathered from groups' sessions. The study is divided in two parts. In **a first part** of the study, twenty-four undergraduate students from a Computer Science program (20 males, 4 females, average age: 24 years), enrolled in an introductory Database Systems course and were asked to participate in a collaborative session using the proposed system. Eight groups were conformed (three students each) and grouped by affinity. Each group worked approximately 30 minutes in the design session. All of the multimodal features were recorded while the students were solving the design problem. At the end of the session, the system scored the solution proposed by the group and the teacher gave feedback to students about their performance. In addition, each student reported the rapport of their group according to their *Enjoyable Interaction* and *Personal Connection* (Frisby et al., 2010).

In the **second part** of the study, four teachers (3 males, 1 female, avg. age 31) with previous experience on teaching Database Design, were invited to participate in the evaluation of the multimodal transcript. First, teachers watched a video showing how the multimodal tabletop system works. Then, they observed the multimodal transcripts from the data gathered from two groups, corresponding to the highest and lowest rapport scores from self-reported data. For purposes of simplicity, only

relevant fragments and a summary of the transcript were used in the observations. Next, the teacher answered a set of questions about the perception of *Enjoyable Interaction* and *Personal Connection* regarding each group. They assigned a score to each group for each variable using a three-point Likert Scale (Low: 1, Medium: 2. High. 3). Additionally, we included open questions about how the multimodal transcript could potentially provide support to the teacher to evaluate and recreate the group-work performance.

4 RESULTS AND DISCUSSION

As for *Enjoyable Interaction*, teachers perceived a low interaction in the group with lowest rapport, for instance they reported an average of 1.25 over 3; whereas in the group with highest rapport, they scored an average of 2.75 over 3. As for *Personal Connection*, a similar pattern was observed, the group with lowest rapport was evaluated with an average of 1.25 over 3, and the one with highest rapport scored 2.5 over 3. From these results, it seems that the multimodal transcripts were valuable for the teachers, since they mostly agreed with the rapport reported by the members of the groups. Some positive comments about the support the teachers perceived from the multimodal transcript for assessing enjoyable interaction and personal connection are presented as follows. One teacher stated*: "the combination of voice and emotions presented in the transcript give me the idea of how the students felt about the task during the session"*, another teacher said that: *"what I observed gives me evidences about the interaction among students … more specifically, the mix between actions and emotions are the evidence of such interactions"*. In addition, there were also some critical remarks. For instance, one teacher indicated that the transcript*: "did not present enough details about the emotions of the participants"*. Another teacher said that *"the emotions presented are not enough to infer the interactions that were present, I think there's the need to evidence the interrelations between the emotions of one participant with the others".*

As for the perception of teachers about how the transcript would support them to evaluate the group-work performance, all the interviewed participants answered positively to this question. Regarding the recreation of student work during the session using the multimodal transcript, three teachers answered positively and one indicated that the transcript would partially support this task. During the interviews one teacher stated that: *"I could observe whether the students were working on the task or they were debating about the task; moreover, I can observe the actions at the level of the individual. It is easy to identify who is the one who work the most or if the task was equally distributed"*. Another teacher suggested the following: *"It would be nice that the students' comment could be analyzed as well at the level of emotions"*. One teacher had a slightly reluctant reaction about the recreation of student work*: "I think it is still ambiguous what the emotions reflect in the transcript; however, the actions performed using the tabletop could help me in the recreation of the work"*.

The validation stage of this work points out to a promising research path. Teachers were mostly positive about the potential of the multimodal transcript to support group-work evaluation, beyond actions and scores. Teachers valued the fact that emotions were present in the transcript. They thought that the mix of this feature with voice and task would support the inference of interactions between the members of the groups. Nevertheless, this work is an on-going project that needs to

further explore how to expand the meaning of emotions in the task, as well as, the reactions between the members of the groups after some enacted emotions.

REFERENCES

Ädel, A. (2011). Rapport building in student group work. Journal Of Pragmatics, 43(12), 2932-2947. http://dx.doi.org/10.1016/j.pragma.2011.05.007.

Adachi, H., Haruna, A., Myojin, S., & Shimada, N. (2015). ScoringTalk and watchingmeter: Utterance and gaze visualization for co-located collaboration. SIGGRAPH Asia 2015

Al-Qaraghuli, A., Zaman, H. B., Ahmad, A., & Raoof, J. (2013, November). Interaction patterns for assessment of learners in tabletop based collaborative learning environment. In *Proceedings of the 25th Australian Computer-Human Interaction Conference.* (pp. 447-450). ACM.

Bezemer, J., & Mavers, D. (2011). Multimodal transcription as academic practice: a social semiotic perspective. *International Journal of Social Research Methodology*, *14*(3), 191-206.

Cao, Z., Simon, T., Wei, S. E., & Sheikh, Y. (2016). *Realtime multi-person 2d pose estimation using part affinity fields.* arXiv preprint arXiv:1611.08050.

Domínguez, F., Chiluiza, K., Echeverria, V., & Ochoa, X. (2015). Multimodal selfies: Designing a multimodal recording device for students in traditional classrooms. In *Proceedings of the 2015 ACM on Intl. Conf. on Multimodal Interaction* (pp. 567-574). ACM.

Echeverria, V., Falcones, G., Castells, J., Martinez-Maldonado, R., & Chiluiza, K. (2017). *Exploring on-time automated assessment in a co-located collaborative system.* Paper presented at the 4th Intl. Conf. on eDemocracy and eGovernment*, ICEDEG 2017,* 273-276.

Frisby, B. N., & Martin, M. M. (2010). Instructor–student and student–student rapport in the classroom. *Communication Education*, *59*(2), 146-164.

Harrigan, J. A., Oxman, T. E., & Rosenthal, R. (1985). Rapport expressed through nonverbal behavior. Journal of nonverbal behavior, 9(2), 95-110.

Heslop, Philip & Preston, Anne & Kharrufa, Ahmed & Balaam, Madeline & Leat, David & Olivier, Patrick. (2015). Evaluating Digital Tabletop Collaborative Writing in the Classroom.

Kharrufa, A., Balaam, M., Heslop, P., Leat, D., Dolan, P., & Olivier, P. (2013). *Tables in the wild: Lessons learned from a large-scale multi-tabletop deployment.* Conf. on Human Factors in Computing Systems, 1021-1030.

Martinez-Maldonado, R., Collins, A., Kay, J., and Yacef, K. (2011) Who did what? who said that? Collaid: an environment for capturing traces of collaborative learning at the tabletop. ACM *Intl. Conf. on Interactive Tabletops and Surfaces, ITS 2011*, pages 172-181.

Martinez-Maldonado, R., Kay, J., & Yacef, K. (2013). An automatic approach for mining patterns of collaboration around an interactive tabletop10.1007/978-3-642-39112-5-11

Martinez-Maldonado, R., Kay, J., Buckingham Shum, S. J., & Yacef, K. (2017). Collocated collaboration analytics: Principles and dilemmas for mining multimodal interaction data. *Human-Computer Interaction.*.

Schneider, B., Sharma, K., Cuendet, S., Zufferey, G., Dillenbourg, P., & Pea, R. (2016). Detecting collaborative dynamics using mobile eye-trackers. Proceedings of International Conference of the Learning Sciences, ICLS, 1, 522-529.

Tang, A., Pahud, M., Carpendale, S., & Buxton, B. (2010). VisTACO: visualizing tabletop collaboration. In *ACM International Conference on Interactive Tabletops and Surfaces* (pp. 29-38). ACM.

Wong-Villacres, M., Ortiz, M., Echeverría, V., & Chiluiza, K. (2015). A tabletop system to promote argumentation in computer science students. In *Proceedings of the 2015 Intl. Conf. on Interactive Tabletops & Surfaces* (pp. 325-330). ACM.