Gesture and Gaze: Multimodal Data in Dyadic Interactions

Bertrand Schneider[1], Marcelo Worsley[2], Roberto Martinez-Maldonado[3]

[1] Harvard University, Graduate School of Education, Cambridge, USA, bertrand_schneider@gse.harvard.edu

[2] Northwestern University, Learning Sciences and Computer Science, Evanston, USA, marcelo.worsley@northwestern.edu

[3] Monash University, Faculty of Information Technologies, Melbourne, Australia, roberto.martinezmaldonado@monash.edu

Abstract: *With the advent of new and affordable sensing technologies, CSCL researchers are able to automatically capture collaborative interactions with unprecedented levels of accuracy. This development opens new opportunities and challenges for the field. In this chapter we describe empirical studies and theoretical frameworks that leverage multimodal sensors to study dyadic interactions. More specifically, we focus on gaze and gesture sensing and how these measures can be associated with constructs such as learning, interaction and collaboration strategies in co-located settings. We briefly describe the history of the development of multimodal analytics methodologies in CSCL, the state of the art of this area of research, and how data fusion and human-centered techniques are most needed to give meaning to multimodal data when studying collaborative learning groups. We conclude by discussing the future of these developments and their implications for CSCL researchers.*

Keywords. Multimodal Sensing, Learning Analytics, eye-tracking, motion sensing, co-located collaborative learning, computational models.

Definitions & Scope

Educational researchers have argued for decades that the field needs better ways to capture process data (Werner, 1937). More recently in CSCL, Dillenbourg et al. (1996) noted that “empirical studies have started to focus less on establishing parameters for effective collaboration and more on trying to understand the role which such variables play in mediating interaction. This shift to a more process-oriented account requires new tools for analyzing and modelling interactions”. Multimodal Learning Analytics (MMLA; Blikstein & Worsley, 2016) is about creating new tools to automatically generate fine-grained process data from multimodal sensors.

More specifically, the focus of this chapter is on gesture and gaze data collected in co-located interactions. We recognize that collaboration is the result of subtle micro-behaviors, such as learners' body position, gestures, head orientation, visual attention and discourse. These actions are complex, intertwined and result in a rich choreography of behaviors that create sophisticated social interactions. Figure 1 provides a visual representation of the key constructs of this chapter:

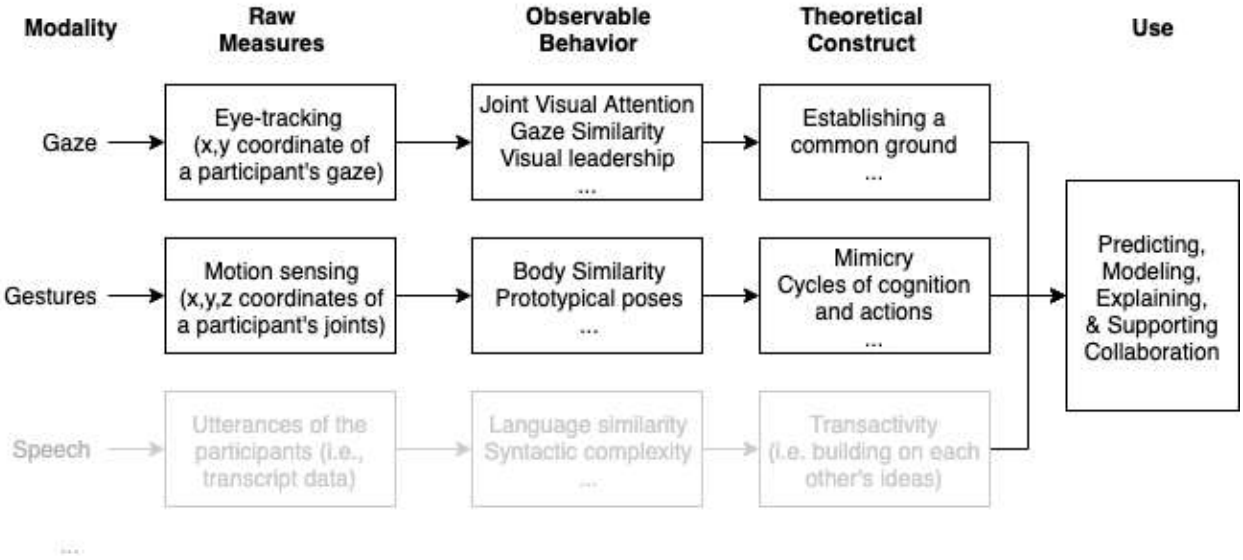

Fig. 1 How different sensor modalities can help CSCL researchers capture constructs relevant to collaborative learning, and how this can be used to predict, model, explain and support productive behaviors. In this chapter, we focus on gaze and gestures (even though other modalities - such as speech - are highly relevant in CSCL settings)

The first column shows modalities studied by CSCL researchers (e.g., gaze, gestures, speech, dialogue). These modalities provide "Raw Measures" of users' gaze or body postures. This data is then used to capture specific "Observable Behaviors", such as Joint Visual Attention (JVA) or body similarity. We can use these behaviors as proxies for "Theoretical Constructs" (Wise, Knight & Buckingham Shum, 2020), for example the quality of a group's common ground (Clark & Brennan, 1991) or the extent to which group members mimic each other (Chartrand & Bargh, 1999).

The raw measures, observables behaviors and constructs can be used to **predict** outcomes of interest (e.g., how well a group is collaborating), **model** collaborative processes (e.g., how social interactions change over time), **explain** them (e.g., contribute to theories of collaboration) or **support** collaboration (e.g., design interventions that use sensor data to support learning). In the sections below, we describe the history and development of MMLA. We then provide additional definitions for the constructs in Fig. 1 and provide concrete examples of their use.

History & Development

While MMLA seems to be a new and exciting methodological development, there has been a long tradition of designing multimodal devices to capture human behavior. At the beginning of the 20th century, Huey (1908) designed the first eye-tracker by having participants wear contact lenses with a small opening for the pupil. Because a pointer was attached to it, Huey was able to make new discoveries on effective reading behaviors. In the 1920s, a German pedagogue, Dr. Kurt Johnen, created a device to measure expert piano players' breathing and muscular tension as a way to design better instruction for novices (Johnen, 1929). In 1977, Manfred Clynes built a device called a "sentograph" which attempted to detect emotions by extracting the length and force applied on a pressure-sensitive finger rest. There are many other examples of early "sensors" designed to capture human behaviors.

Over the last decade, however, the affordability and accessibility of multimodal sensing has opened new doors for monitoring, analyzing, visualizing and regulating a variety of learning processes. Depth cameras such as the Microsoft Kinect can collect information about a person's body joints (x, y, z coordinates), their facial expressions, and their speech 30 times per second. Researchers can obtain more than a hundred variables from this sensor, which represents +3000 data points per second for one person. This translates to roughly 10 million data points for an hour of data collection. Multiply this figure by the number of sensors (e.g., eye-trackers, galvanic skin response sensors, emotion detection tools, speech features) and number of learners to get a sense of the possibilities and challenges of combining sensor data with data mining techniques.

State of the Art

In this section, we describe the state of the art research methods for analyzing gaze and motion data from small groups in educational settings. We start with some definitions, conventions, and findings from the CSCL community and beyond. We conclude this chapter with a comparison of the state of the field for gaze and gesture sensing, comments on the future of associated methodologies and implications for CSCL researchers.

Gaze Sensing in CSCL

With sensing devices becoming more affordable, the last decades have seen an increasing number of CSCL researchers taking advantage of eye-trackers to study small collaborative groups. This line of work is grounded in the literature on joint visual attention (Tomasello, 1995). Joint attention is an important mechanism for building a common ground (i.e., "grounding", which allows group members to anticipate and prevent misunderstanding; Clark & Brennan, 1991). Educational researchers have built on this idea and extended it to learning scenarios: "*From the viewpoint of collaborative learning, misunderstanding is a learning opportunity. In order to repair misunderstandings, partners have to engage in constructive activities: they will build*

explanations, justify themselves, make explicit some knowledge which would otherwise remain tacit and therefore reflect on their own knowledge, and so forth. This extra effort for grounding, even if it slows down interaction, may lead to better understanding of the task." (Dillenbourg & Traum, 2006)

In other words, educational researchers go beyond the psycho-linguistic definition of grounding to focus on shared meaning making (Stahl, 2007). Shared 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. There is some evidence suggesting that groups with high levels of joint visual attention are more likely to iteratively sustain and refine their common understanding of a shared problem space (Barron, 2003). Because eye-trackers can provide a rigorous measure of joint visual attention, gaze sensing has become an attractive methodology for studying grounding in collaborative learning groups.

The state of the art of CSCL gaze sensing is a dual eye-tracking methodology where pairs of learners solve a problem together and learn from a shared set of resources. Early studies had two participants looking at a different computer screen equipped with an eye-tracker (Jermann et al., 2001). Participants can communicate through an audio channel and have access to the same interface. For dyadic analysis, the two eye-tracking devices need to be synchronized so that the resulting datasets can be combined to compute measures of joint visual attention (JVA).

After the data is acquired, there are established methodologies for computing JVA measures. Cross recurrence graphs (Richardson et al., 2007) are commonly used to visually inspect the joined eye-tracking datasets and identify missing data. JVA is then computed according to Richardson and Dale's findings (2005), where they found that dyad members are rarely perfectly synchronized; it takes participants +/- 2 seconds to react to an offer of joint visual attention and respond to it. Thus, for a particular gaze point to count as joint visual attention, researchers usually look at a 4 seconds time window to check whether the other participant was paying attention to the same location. This methodology provides an overall measure of attentional alignment for dyads.

One common finding is that levels of joint visual attention are positively associated with constructs that the CSCL community cares about. For example researchers have used established coding schemes to evaluate the quality of a dyad's collaboration and correlated it with measures of JVA. Meier et al. (2007) developed a coding scheme that characterizes collaboration across 9 subdimensions: sustaining mutual understanding, dialogue management, joint information processing, information pooling, reaching consensus, division, time management, technical coordination, reciprocal interaction, and individual task orientation. Among those subdimensions, JVA has been repeatedly found to be significantly associated with a group's ability to sustain mutual understanding (e.g., Schneider et al., 2015; Schneider & Pea, 2013). Some other

studies have also found positive correlations between JVA and learning gains (Schneider & Pea, 2013), which suggests that this type of collaborative process is not just beneficial to collaboration, but also to learning. This shows that, to some extent, JVA measures can be used to **predict** collaboration quality and learning.

Additional measures of JVA have been developed for specific contexts. For example, "with-me-ness" was developed to measure if students are following along with a teacher's instruction (Sharma et al., 2014). This measure is calculated by aggregating three features of gaze data: entry time, first fixation duration and the number of revisits. Entry time is the temporal lag between the time a reference pointer (gaze) appears on the screen and stops at the referred location (x, y) and the time the student first looks at the referred location (x, y). The first fixation duration is how long the student gaze stopped at the referred location for the first time and revisits are the number of times the student's gaze comes back to the referred location within four seconds.

In addition to these measures of JVA, CSCL researchers have also looked at the "attentional similarity" between participants (Sharma, Jermann, Nüssli & Dillenbourg, 2013). For a given time window (e.g., 5 seconds), the proportion of time spent on different Areas of Interest (AOIs) is computed and compared across participants using a similarity metric (for example, the cosine similarity between two vectors). Papavlasopoulou et al. (2017) found that in a pair programming task, teenagers (13–17-year-old participants) spent more time overall working together (higher similarity gaze) than younger participants kids (8–12 year-old). While this measure is similar to others described above, it uses a less conservative operationalization of joint visual attention. These measures provide alternative ways of **modeling** joint visual attention in small groups.

It is also possible to detect asymmetrical collaboration from the eye-tracking data (Schneider et al., 2018). For each moment of joint attention, one can look at which participant initiated this episode (i.e., the person whose gaze was first present in this area during the previous two seconds) and which student responded to it (i.e., the person whose gaze was there second). The absolute value of the difference between the number of moments that each participant initiated and responded to represents 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. Schneider et al. (2018) found this measure to be negatively correlated with learning gains – meaning that groups in which one person tended to always initiate or respond to an offer of joint visual attention were less likely to achieve high learning gains. These findings can help us **explain** how specific collaborative behaviors can contribute to learning.

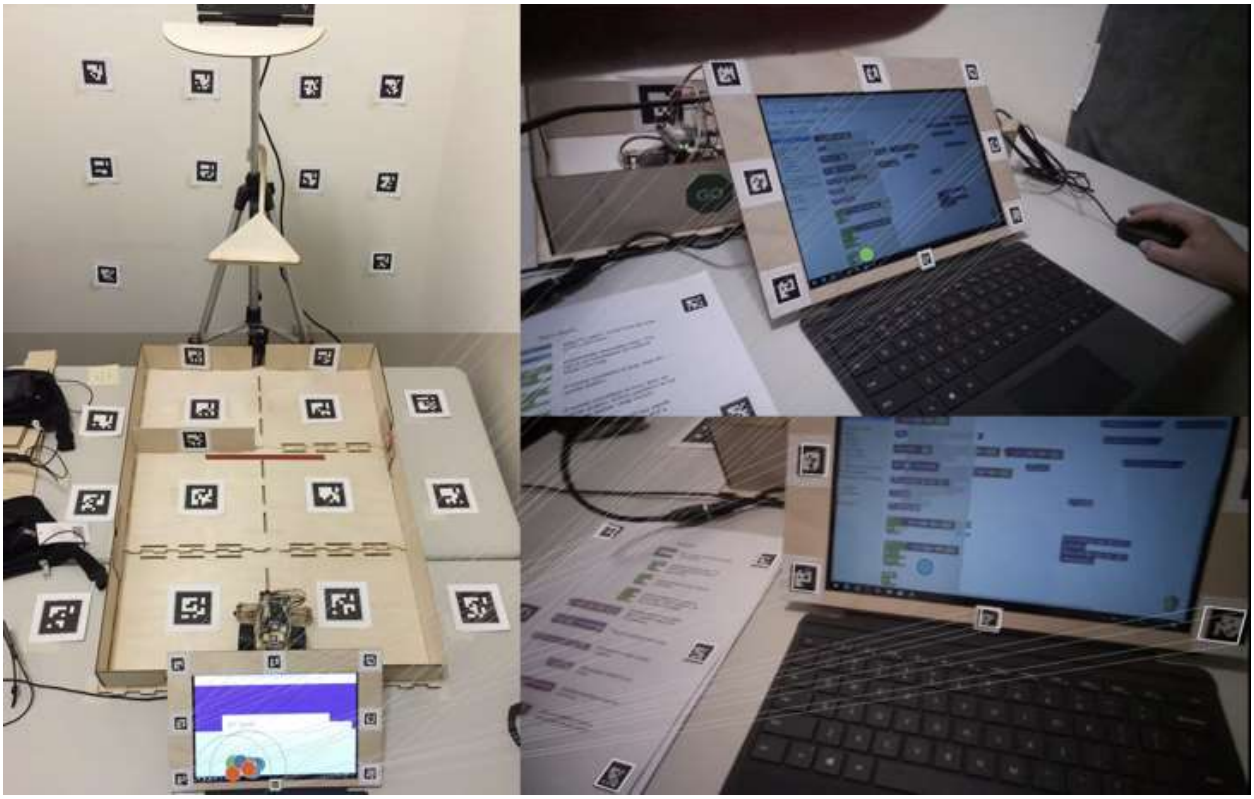

Fig. 2 (reproduced from Schneider, 2019): An example of using dual mobile eye-tracking to capture joint visual attention in a co-located setting (in this particular case, pairs of participants had to program a robot to solve a variety of mazes). The two images on the right show the perspective of the two participants; the left image shows a ground truth where gaze points are remapped using the location of the fiducial markers detected on each image (the white lines connect identical markers).

Additionally, researchers have started to go beyond remote collaboration and use dual eye-tracking in co-located settings using mobile eye-trackers (Schneider et al., 2018). In this type of setup, there is an extra step of spatially synchronizing the two eye-tracking datasets, which is usually done by remapping participants' gaze into a ground truth (i.e., a common scene that both participants look at). The remapping processes is usually accomplished by disseminating fiducial markers in the environments and using this shared set of coordinates between each participant's point of view and the ground truth (Fig. 2). When the two gaze points are remapped onto the ground truth, one can reuse the methodology described above for remote interactions and compute the same measure of joint visual attention.

Finally, there are practical implications of using dual eye-tracking methodologies beyond quantitatively capturing collaborative processes. The last decade has seen a nascent interest for designing shared gaze visualization – i.e., displaying the gaze of one's partner on a computer screen to support joint visual attention (see review by D'Angelo & Schneider, under review). Shared gaze visualizations have been found to facilitate communication through deictic references, disambiguate vague utterances and help participants anticipate their partner's verbal contribution. This is an exciting new line of

research, because work goes beyond descriptive measure of collaboration and suggests interventions to **support** collaboration.

While the study of JVA through gaze sensing is reaching some maturity, there are obvious gaps in this area of research. Dual eye-tracking tends to be used in live remote collaboration, which is not the most ecological settings from an educational perspective. Most students still work in co-located spaces, where they work together face-to-face or side-by-side. This lack is slowly being addressed by new methodologies using mobile eye-trackers which brings more ecological validity to this field of research.

Gesture Sensing in CSCL

In contrast to eye-tracking, where researchers are looking at the x,y coordinate of a participant's gaze, gesture tracking (and more generally motion sensing in CSCL) is operationalized at varying levels of granularity. These levels of analysis range from the mere quantification of movement or the complex identification of specific gestures in dyadic interactions to localizing people in physical learning spaces. Part of this breadth in levels of analysis reflects to relative infancy of this area of study. Researchers are in the process of determining the appropriate measures and theoretical grounding for gesture sensing. In this section, we present examples along this spectrum and further note how these approaches are utilized to examine and support collaboration.

As is the case with eye tracking, the availability of low-cost gesture tracking technology has enabled researchers to develop and create interfaces that incorporate human gestures. Initially, many of these technological systems relied on an infrared camera (e.g., the Nintendo Wiimote) and an infrared source (e.g., an infrared pen, or television remote). This was, for example, used for the Mathematical Inquiry Trainer (Howison et al., 2011), a system that supports embodied learning of fractions. The next wave of gesture technology was heavily fueled by the Microsoft Kinect Sensor and supporting SDK. The Kinect Sensor V2 uses a depth camera to provide a computer vision based solution to track upper and lower body joints - as well as finger movement, head position, and even the amount of force applied to each appendage. Leong et al. (2015) provide an in-depth comparison of different depth cameras, and their capabilities. More recently, advances in computer vision have eliminated the need for specialized data capture hardware. Instead, OpenPose (Cao et al., 2017; Simon et al., 2017; Wei et al., 2016) and DensePose (Güler et al., 2018), for example, train deep neural networks for estimating human body pose, from standard web images or videos cameras. As an example, Ochoa et al. (2018) use OpenPose to provide feedback to users about their body posture during oral presentation training. The result of these technological developments is a growing opportunity to employ use gesture sensing to study collaborative learning environments, without the need for expensive, or invasive wearables.

As previously noted, research on motion sensing in CSCL operates at different levels of complexity (i.e., individual learning, small group interactions and localizing a larger number of participants in open spaces). Some studies are merely looking to quantify the

amount of movement, others examine body synchrony, while still others are concerned with recognizing specific types of gestures or body movements. The specific approaches utilized, as well as how they are operationalized are necessarily impacted by the research questions being explored.

At the individual level, several studies have looked at the potential of motion sensing for understanding learning and constructing models of the student learning experience. Schneider and Blikstein (2015), for example, tackled this question by examining prototypical body positions among pairs of learners completing an activity with a tangible user interface. The researchers categorized body postures using unsupervised machine learning algorithms and identified three prototypical states: an “active” posture (positively correlated with learning gains), a “semi-active” posture and a “passive” posture (negatively correlated with learning gains). Interestingly, the best predictor for learning was the number of times that participants transitioned between those states, suggesting a higher number of iterations between “thinking” about the problem and “acting” on it. Researchers interested in ITSs (Intelligent Tutoring Systems) have also used motion and affective sensing to predict levels of engagement, frustration and learning using supervised machine learning algorithms. Grafsgaard et al. (2014), for example, found indicators of engagement and frustration by leveraging features about face and gesture (e.g., hand-to-face gestures) and indicators of learning by using face and posture features. These two papers highlight the opportunity for motion sensing to help us better identify patterns of engagement that may be indicative of improved learning, or certain affective states. Specifically, gesture sensing can help researchers **predict** learning gains or affective states.

At the group level, the most basic uses of gesture data involve the quantification of bodily movement among pairs of students collaborating on a given task. For example, Martinez-Maldonado, Kay, Buckingham Shum & Yacef (2017) presented an application of the Kinect by locating it on top of an interactive tabletop to associate actions logged by the multitouch interface with the author of such a touch. Authors applied a sequential pattern mining algorithm on these logs to detect patterns that distinguished high from low performing small groups in a collaborative concept mapping task. Worsley and Blikstein (2013) used hand/wrist joint movement data to extract patterns of multimodal behaviors of dyads completing an engineering design activity. The gestural data, when taken in conjunction with audio and electro-dermal activation data was beneficial in codifying the types of actions students were taking at different phases of the building activity. Such information about student gestural engagement could also be used in a way that is analogous to analyses of turn-taking. Moreover, it can help answer questions about the extent of each participant’s physical contributions to a given learning activity, or, the patterns of participation that emerge between participants as they collaborate with one another. In the same vein, Won, Bailenson, and Janssen (2014) found that body movements captured by a Kinect sensor could predict learning with a 85.7% accuracy in a teacher-student dyad; the top three features were the standard deviation of the head and torso of the teacher, the skewness of students’ head and torso, and mean of teacher left arm. Other studies have looked at the relationship between body synchronization and group interaction. Won, Bailenson, Stathatos, and

Dai (2014), for example, found that non-verbal synchrony predicted creativity in 52 collaborative dyads. Models trained with synchrony scores could predict low or high scores of creativity with a 86.7% accuracy. In educational contexts, Schneider and Blikstein (2015) looked for the salience of body synchronization by considering the correlation between body position similarity and learning gains. However, the results indicated no correlation between learning and body synchronization in this context. Similarly, Spikol et al. (2017) paired a number of computer vision systems to detect wrist movement and face orientation of small groups of students performing an electronic toy prototyping task in triads. Results indicated that some features, such as the distance between all learners' hands and the number of times they look at a shared screen, are promising in helping identify physical engagement, synchronicity and accountability of students' actions. Concretely, motion sensing among groups of learners can be used to **explain** success within given collaborative experience as determined through the relative participation of each individual and their level of synchrony or proximity to their peers.

Researchers are also finding ways to leverage gestural data as a means for streamlining and improving the data analysis process. In a study that involved pairs of students completing engineering design tasks, Worsley et al. (2015) was able to show that using body posture information to automatically segment data into meaningful chunks, led to analyses that provided stronger correlations with student performance and student learning. In this particular study, authors used automatically detected changes in head pose relative to learners' partners to demark the beginning of a new phase. This approach was compared to human annotation of phases, and taking a fixed window approach, with the body position based segmentation proving to be quite beneficial. Hence, the utility of gesture data does not necessarily have to be restricted to a final correlation with learning or performance. It can, instead, be used to more adequately group chunks of data into meaningful representations. In this line of work, computational methods provide ways to **model** students' behaviors.

In another emerging body of work, researchers are exploring the use of gestures, in conjunction with other modalities, to better understand embodied learning in mathematics and science. For example, Abrahamson's Mathematical Inquiry Trainers (Howison et al., 2011) and Robb Lindgren's ELASTICS (Kang et al., 2018) platforms represent computer-supported tools that help facilitate student learning with the assistance of a more knowledgeable interviewer. In both instances, the interviewer serves as a collaborator to help guide the student towards learning and articulating mathematical or scientific ideas. In the case of Abrahamson's work, students use their hands to reason about fractions, either through a touch screen interface, Nintendo Wii mote or Kinect sensor. In the case of ELASTICs, students use gestures to instantiate different mathematical operations. For example, in Kang et al. participants determine a gestural sequence that will allow them to produce a value of 431. In order to reach this value, students can complete gestures that correspond to add one, subtract one, multiply by ten or divide by ten. These sub-tasks exist within a larger task of helping students reason about exponential growth. Crucial for both Abrahamson and Lindgren's work is the opportunity to create gestural interfaces that allow for embodied

experiences, and the availability of visual representations that individuals and/or pairs can utilize to refine their thinking, and serve as a context for discussion. This kind of work exemplifies the potential of motion sensor data to **support** novel, embodied, collaborative learning.

These different examples suggest that while there are some similarities and accepted practices in how to analyze gesture data (e.g., the use of joint angles as opposed to three dimensional x, y, z data), there are still several areas where new innovations and ideas are emerging. The identification of constructs that are analogous to the joint visual attention, for example, does not yet seem to exist within the gesture space. Instead, researchers have found and explored different metrics that aim to characterize the nature of collaboration among groups or pairs of learners.

Comparison between Gaze Sensing and Gesture Sensing

In this section we compare the state of the field in gesture and gaze sensing to illustrate opportunities and challenges to studying small collaborative groups using gaze and motion sensing. Both areas of research have been evolving at different paces, and have contributed unique findings to the study of collaborative learning Table 1 summarizes the main commonalities and differences across those two methodologies:

Table 1 A comparison of the state of research using gaze and motion sensing based on the work reviewed in this chapter

	Gaze Sensing	**Motion Sensing**
Raw measures	x, y coordinates of gaze in a 2D space (e.g., remote or mobile eye-tracker)	x, y ,z coordinates of dozens of body joints in a 3D space (e.g., Kinect sensor)
Accuracy	Accurate, depending on the eye-tracker used.	More noisy and susceptible to occlusion
Constructs	Joint visual attention (Schneider & Pea, 2013), attentional similarity (Sharma et al., 2013), ...	Body movement (Worsley & Blikstein, 2013), prototypical states (Schneider & Blikstein, 2015), physical synchrony (Won, Bailenson, Stathatos & Dai, 2014), ...
Methodology	Well established; strong conventions (Richardson & Dale, 2005).	In development; currently, there are no strong conventions.
Models	Glass-box traditional statistical models (e.g., Sharma et al., 2014); higher explainability, lower predictive value.	Black-Box machine learning models (e.g., Won, Bailenson & Janssen, 2014; Won, Bailenson, Stathatos & Dai, 2014;); lower explainability, higher predictive value.

Theoretical basis	Well-documented and specific, from developmental (Tomasello, 1995) and social (Richardson et al., 2007) psychology	Emerging and less prescriptive, e.g., embodied cognition (Howison et al., 2011)

A striking difference between those two fields of research is that gaze sensing - through the study of joint visual attention - has developed well-established conventions for visualizing and capturing collaborative processes. This work leverages foundational theories in developmental psychology and has specific hypotheses about the role of visual synchronization for social interactions. Because the raw measures are simpler and the theory is more prescriptive, it has allowed researchers to use more transparent ("glass-box") statistical models (e.g., Richardson et al., 2007) and design innovative interventions to support collaborative processes - for example by building systems where participants' gaze can be displayed in real-time and shared within the group (Schneider & Pea, 2013). Motion sensing, on the other hand, offers larger and more complex datasets. Because theoretical frameworks are less specific (i.e., embodied cognition), there is a wider variety of measures and models being used, with more researchers leveraging "black box" models (i.e., supervised machine learning algorithms) to predict collaborative processes (e.g., Won, Bailenson, & Janssen, 2014). While those models are designed to provide accurate predictions, they tend to be less transparent and offer fewer opportunities for designing interventions.

In summary, gaze sensing has benefited from simpler constructs, more prescriptive theoretical frameworks, accurate sensors to reach a certain level of maturity. Motion sensing, on the other hand, has an untapped potential: the technology is rapidly improving and there are new opportunities to make theoretical contributions, develop innovative measures of group interaction, and design interventions to support collaborative learning processes.

Fusion

While most of the current body of work has looked at gaze and motion sensing in isolation, there is a growing interest in combining multiple sources of data to provide a more complete depiction of complex social aspects of human activity that would be hard to model considering one modality of group interaction only. In the examples discussed above, multiple data sources have been used to model different aspects of collaborative learning. For instance, gaze sensing is commonly paired with information generated by the learning systems or with transcripts (Schneider & Pea, 2015). Gestural data has been enriched by combining them with quantitative traces of speech, such as sound level (Spikol et al., 2017) or turn-taking patterns (Martinez-Maldonado et al., 2018), to give meaning to gestures and poses. However, the process of fusing across data streams can bring a number of challenges related to low-level technical issues, such as data modelling and pattern extraction; and higher-level aspects, such as sensemaking, data interpretation and identification of implications for teaching, learning or collaboration.

Some low level challenges in fusing gaze, gesture and other sources of data are associated with deciding what features to extract from the data, and how to segment or group the multiple data streams with the purpose of jointly modelling a meaningful indicator of collaboration or learning. In terms of multi-feature extraction, researchers often overlook the opportunity to extract multiple pieces of information from a single data source. In the case of gaze data for example, multi-feature extraction includes determining fixations, saccades and pupil dilation from the single data source (i.e., the eye tracker). From skeletal tracking information one might extract pointwise velocity, angular displacement, or distance between body points. The challenge here is in giving interpretative meaning to the selected features that can be obtained from the data for particular contexts.

This challenge also applies to how the data is grouped or segmented. Summary statistics represent a simple approach for investigating multimodal data. In principle, this approach merges all of the data from a given modality into a single representation. Researchers commonly use values of mean, median, mode, range, maximum and minimum. This accomplishes fusion across time, but can grossly oversimplify the data representation. Instead researchers may wish to 'group' data into *meaningful* segments. Within this paradigm, data can be segmented into chunks that range in size from the entire dataset all the way down to individual data points. One advantage of segmentation is that it can help surface patterns and trends that are localized to particular segments. For example, Worsley and Blikstein (2017) explored the affordances of segmentation by comparing three different approaches. These authors ultimately found that having a combination of semantically meaningful segments and a large number of segments yielded the most meaningful results.

At a higher level, there are challenges in giving meaning to fused data across streams and participants. Fundamental to multimodal learning analytics is the idea that a given data stream can only be interpreted in the context of other data streams. However, a key question remains: on what basis can low-level indicators serve as proxies for higher order collaborative learning constructs? From a research perspective, this is a fundamental modelling problem that involves encoding low-level events in data representations that contain a certain amount of contextual information to facilitate higher level abstraction. This is manifested in the learning analytics and educational data mining communities in various forms such as stealth assessment (Shute and Ventura, 2013) and evidence-centered design (Mislevy et al., 2012). At the intersection between CSCL and learning analytics, this challenge has been called as mapping "from clicks to constructs" (Wise et al., 2020).

From a teaching and learning perspective, modelling group constructs from multiple data streams is a prerequisite for creating interfaces that are intelligible to teachers and learners, who commonly do not have a strong analytical background. Until now, most multimodal analytics for group activity have mainly remained the preserve of researchers (Ochoa, 2017). Imbuing traces of gaze and gesture, and other sources of data, with contextual meaning can bring teachers and students into the sensemaking and interpretation loop. One promising approach is that of Echeverria's et al. (2019)

who proposed a modelling representation to encode each modality of data into one or more of the n columns of a matrix and segments that contain instances of group behaviors into the m rows. From this representation a set of group visualizations were proposed, each presenting information related to one modality of teamwork, namely speech, arousal, positioning and logged actions.

In summary, there are numerous technical and sensemaking-related challenges related to combining multiple data sources that need to be addressed in turn. However, the potential benefits, such as the possibility of creating interpretable group models, generating deeper understanding of collaborative learning and deploying user interfaces that can provide tailored feedback on collocated activities, outweigh such challenges.

The Future

The last decade has seen an increasing number of research projects involving gaze and motion sensing. This is a positive development for the CSCL community. This methodology provides researchers with large amounts of process data and new tools to analyze them. Not only does it help automate time consuming analyses, but it also provides a new perspective to understand collaborative processes. Additionally, it provides researchers with opportunities to develop real-time interventions (for example through dashboards or awareness tools; Schneider & Pea, 2013).

These advances are not without challenges. For example, most of the work presented in this chapter is about dyads, when collaborative groups are often larger than two participants. This poses new opportunities for adapting multimodal measures of collaboration for larger groups (e.g., is JVA occurring when all the participants - or just two group members? - are jointly looking at the same place at the same time?) Researchers are slowly starting to look at larger social contexts, but this is currently an understudied area of research.

Another major area of work is the contribution of multimodal studies to theory. Researchers are designing more sophisticated measures of visual synchronization and collaboration (e.g., leadership behaviors, with-me-ness) and turning dual eye-tracking setups into interventions to support collaborative processes. However, this kind of empirical study needs to be replicated and refined before they can be established as significant theoretical contributions to the field of CSCL. More importantly, theories of collaboration have not yet benefited from more fine-grained multimodal measures of collaborative processes.

Finally, it should be noted that most studies are unimodal or only combine two data streams together. Very few projects have attempted to combine data sources; Data fusion presents new opportunities for studying collaborative learning groups and capturing more sophisticated constructs. With these new opportunities also come increased concerns about data privacy: how should we handle questions around the collection, storage and analysis of potentially sensitive datasets? It will be important for

the CSCL researchers to carefully reflect on these concerns as they look to drive innovation and advance knowledge.

In the coming decade, we are expecting to see more affordable and accurate sensors emerge as well as easy-to-use toolkits for analyzing multimodal datasets. With an increased focus on data-driven approaches, we believe that multimodal sensing will become a common tool for educational researchers. Those new tools will provide new ways to build theories of collaboration and design interventions to support social interactions. We agree with Wise & Schwartz (2017), who argue that CSCL has to embrace those new methods if it wants to stay relevant in an increasingly data-driven world.

References

Baker, M., Hansen, T., Joiner, R., & Traum, D. (1999). The role of grounding in collaborative learning tasks. In *P. Dillenbourg (Ed.), Collaborative learning: Cognitive and computational approaches* (pp. 31-63; 223-225). Elsevier.

Barron, B. (2003). When Smart Groups Fail. *Journal of the Learning Sciences, 12*(3), 307–359.

Blikstein, P., & Worsley, M. (2016). Multimodal Learning Analytics and Education Data Mining: using computational technologies to measure complex learning tasks, *Journal of Learning Analytics, 3(2)*, 220–238.

Cao, Z., Simon, T., Wei, S.-E., & Sheikh, Y. (2017). Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 7291-7299).

Chartrand, T. L., & Bargh, J. A. (1999). The chameleon effect: the perception–behavior link and social interaction. *Journal of personality and social psychology, 76*(6), 893-910.

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. *Perspectives on socially shared cognition, 13*(1991), 127-149.

Clynes, M. (1977). *Sentics: The touch of emotions*. Anchor Press.

D'Angelo, S., & Schneider, B. (under review). *Shared Gaze Visualizations in Collaborative Work: Past, Present and Future* [Manuscript submitted for publication].

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). Emerald.

Dillenbourg, P., & Traum, D. (2006). Sharing solutions: Persistence and grounding in multimodal collaborative problem solving. *The Journal of the Learning Sciences, 15*(1), 121-151.

Echeverria, V., Martinez-Maldonado, R., & Buckingham Shum, S. (2019). Towards Collaboration Translucence: Giving Meaning to Multimodal Group Data. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Paper 39, pp. 1-16). Association for Computing Machinery. https://doi.org/10.1145/3290605.3300269

Grafsgaard, J. F., Wiggins, J. B., Vail, A. K., Boyer, K. E., Wiebe, E. N., & Lester, J. C. (2014). The Additive Value of Multimodal Features for Predicting Engagement, Frustration, and Learning During Tutoring. In *Proceedings of the Sixteenth ACM International Conference on Multimodal Interaction* (pp. 42–49). Association for Computing Machinery. https://doi.org/10.1145/2663204.2663264

Güler, R. A., Neverova, N., & Kokkinos, I. (2018). DensePose: Dense Human Pose Estimation In The Wild. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 7297-7306).

Huey, E. B. (1908). *The psychology and pedagogy of reading*. The Macmillan Company.

Howison, M., Trninic, D., Reinholz, D., & Abrahamson, D. (2011). The Mathematical Imagery Trainer: From Embodied Interaction to Conceptual Learning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing*

Systems (pp. 1989–1998). Association for Computing Machinery. https://doi.org/10.1145/1978942.1979230

Jermann, P., Mullins, D., Nuessli, M.-A., Dillenbourg, P. (2001). Collaborative Gaze Footprints: Correlates of Interaction Quality. In H. Spada, G. Stahl, N. Miyake, & N. Law (Eds.), *Connecting Computer-Supported Collaborative Learning to Policy and Practice: CSCL2011 Conference Proceedings* (Vol. 1; pp. 184-191). International Society of the Learning Sciences.

Johnen, K. (1929). Measures Energy Used In Piano. *Popular Science Monthly, 69*.

Kang, J., Lindgren, R., & Planey, J. (2018). Exploring Emergent Features of Student Interaction within an Embodied Science Learning Simulation. *Multimodal Technologies and Interaction*, *2*(3), 39.

Leong, C. W., Chen, L., Feng, G., Lee, C. M., & Mulholland, M. (2015). Utilizing depth sensors for analyzing multimodal presentations: Hardware, software and toolkits. In *ICMI 2015 - Proceedings of the 2015 ACM International Conference on Multimodal Interaction* (pp. 547–556). Association for Computing Machinery. https://doi.org/10.1145/2818346.2830605

Martinez-Maldonado, R., Kay, J., Buckingham Shum, S., & Yacef, K. (2017). Collocated Collaboration Analytics: Principles and Dilemmas for Mining Multimodal Interaction Data. *Human-Computer Interaction, 34*(1), 1-50.

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.

Mislevy, R. J., Behrens, J. T., Dicerbo, K. E., Levy, R. (2012). Design and discovery in educational assessment: Evidence-centered design, psychometrics, and educational data mining. *Journal of Educational Data Mining 4*(1), 11–48.

Ochoa, X. (2017). Multimodal Learning Analytics. In C. Lang, G. Siemens, A. F. Wise, & D. Gaševic (Eds.), *The Handbook of Learning Analytics* (pp. 129-141). SOLAR.

Ochoa, X., Dominguez, F., Guamán, B., Maya, R., Falcones, G., & Castells, J. (2018). The RAP System: Automatic Feedback of Oral Presentation Skills Using Multimodal Analysis and Low-cost Sensors. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge* (pp. 360–364). ACM. https://doi.org/10.1145/3170358.3170406

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.

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.

Richardson, D. C., Dale, R., & Kirkham, N. Z. (2007). The Art of Conversation Is Coordination Common Ground and the Coupling of Eye Movements During Dialogue. *Psychological Science, 18*(5), 407–413.

Schneider, B., & Blikstein, P. (2015). Unraveling Students' Interaction Around a Tangible Interface using Multimodal Learning Analytics. *Journal of Educational Data Mining, 7*(3), 89-116.

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., & Pea, R. (2015). Does Seeing One Another's Gaze Affect Group Dialogue? A Computational Approach. *Journal of Learning Analytics*, *2*(2), 107–133. https://doi.org/10.18608/jla.2015.22.9

Schneider, B., Sharma, K., Cuendet, S., Zufferey, G., Dillenbourg, P., & Pea, R. (2015). 3D Tangibles Facilitate Joint Visual Attention in Dyads. In *Proceedings of the 11th International Conference on Computer Supported Collaborative Learning – Volume 1* (pp. 158–165). International Society of the Learning Sciences.

Schneider, B., Sharma, K., Cuendet, S., Zufferey, G., Dillenbourg, P., & Pea, R. (2018). Leveraging Mobile Eye-Trackers to Capture Joint Visual Attention in Co-Located Collaborative Learning Groups. *International Journal of Computer-Supported Collaborative Learning, 13*(3), 241-261.

Schneider, B. (2019). Unpacking Collaborative Learning Processes during Hands-on Activities using Mobile Eye-Tracking. In *the 13th International Conference on Computer Supported Collaborative Learning – Volume 1* (pp. 41–48). International Society of the Learning Sciences.

Sharma, K., Jermann, P., & Dillenbourg, P. (2014). "With-me-ness": A gaze-measure for students' attention in MOOCs. In *Proceedings of the 11th International Conference of the Learning Sciences* (pp. 1017-1022). ISLS.

Sharma, K., Jermann, P., Nüssli, M. A., & Dillenbourg, P. (2013). Understanding collaborative program comprehension: Interlacing gaze and dialogues. In N. Rummel, M. Kapur, M. Nathan, & S. Puntambekar (Eds.), *To See the World and a Grain of Sand: Learning across Levels of Space, Time, and Scale: CSCL 2013 Conference Proceedings Volume 1 — Full Papers & Symposia* (pp. 430-437). International Society of the Learning Sciences.

Shute, V. J., & Ventura, M. (2013). *Stealth assessment: Measuring and supporting learning in video games.* MIT Press.

Simon, T., Joo, H., Matthews, I., & Sheikh, Y. (2017). Hand Keypoint Detection in Single Images using Multiview Bootstrapping. In *Proceedings of the 2017 IEEE conference on Computer Vision and Pattern Recognition* (pp. 1145-1153). IEEE.

Spikol, D., Ruffaldi, E., & Cukurova, M. (2017). *Using multimodal learning analytics to identify aspects of collaboration in project-based learning*. International Society of the Learning Sciences.

Stahl, G. (2007). Meaning making in CSCL: Conditions and preconditions for cognitive processes by groups. In *Proceedings of the 8th international conference on Computer Supported Collaborative Learning* (pp. 652–661). ACM.

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). Lawrence Erlbaum.

Werner, H. (1937). Process and achievement—a basic problem of education and developmental psychology. *Harvard Educational Review, 7,* 353–368.

Wei, S.-E., Ramakrishna, V., Kanade, T., & Sheikh, Y. (2016). Convolutional pose machines. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (pp. 4724-4732)..

Wise, A. F., Knight, S., & Buckingham Shum, S. (2020). Collaborative Learning Analytics. In U. Cress, C. P. Rosé, A. Wise, & J. Oshima (Eds.), *International Handbook of Computer-Supported Collaborative Learning*. Springer.

Wise, A. F., & Schwarz, B. B. (2017). Visions of CSCL: eight provocations for the future of the field. *International Journal of Computer-Supported Collaborative Learning, 12(4),* 423-467.

Won, A. S., Bailenson, J. N., & Janssen, J. H. (2014). Automatic detection of nonverbal behavior predicts learning in dyadic interactions. *IEEE Transactions on Affective Computing*, *5*(2), 112–125.

Won, A. S., Bailenson, J. N., Stathatos, S. C., & Dai, W. (2014). Automatically detected nonverbal behavior predicts creativity in collaborating dyads. *Journal of Nonverbal Behavior*, *38*(3), 389–408.

Worsley, M., & Blikstein, P. (2013). Towards the development of multimodal action based assessment. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge (LAK '13)* (pp. 94–101). ACM. https://doi.org/10.1145/2460296.2460315

Worsley, M., & Blikstein, P. (2017). A Multimodal Analysis of Making. *International Journal of Artificial Intelligence in Education, 28*(3), 385-419.

Worsley, M., Scherer, S., Morency, L.-P., & Blikstein, P. (2015). Exploring behavior representation for learning analytics. In *ICMI 2015 - Proceedings of the 2015 ACM International Conference on Multimodal Interaction* (pp. 251-258). ACM. https://doi.org/10.1145/2818346.2820737