Gesture and Gaze: Multimodal Data in Dyadic Interactions

Bertrand Schneider, Marcelo Worsley, and Roberto Martinez-Maldonado

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 colocated 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 · Colocated collaborative learning · Computational models

B. Schneider (✉)
Graduate School of Education, Harvard University, Cambridge, MA, USA
e-mail: bertrand_schneider@gse.harvard.edu

M. Worsley
Learning Sciences and Computer Science, Northwestern University, Evanston, IL, USA
e-mail: marcelo.worsley@northwestern.edu

R. Martinez-Maldonado
Faculty of Information Technologies, Monash University, Melbourne, Australia
e-mail: roberto.martinezmaldonado@monash.edu

U. Cress et al. (eds.), *International Handbook of Computer-Supported Collaborative Learning*, Computer-Supported Collaborative Learning Series 19,
https://doi.org/10.1007/978-3-030-65291-3_34

1 Definitions and 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 modeling interactions." Multimodal learning analytics (MMLA; Blikstein and 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 colocated 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:

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. These data are 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 et al. this volume), for example, the quality of a group's common ground (Clark and Brennan 1991) or the extent to which group members mimic each other (Chartrand and 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

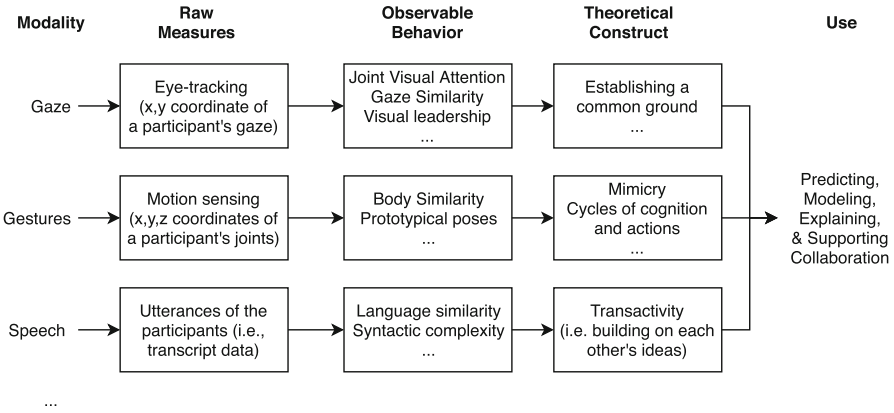

Fig. 1 How different sensor modalities can help CSCL researchers capture constructs relevant to collaborative learning, and how this can be used to predict, model, explain, and support productive behaviors. In this chapter, we focus on gaze and gestures (even though other modalities—such as speech—are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

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.

2 History and 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 twentieth 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 (Clynes 1977). 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 have 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 100 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.

3 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.

3.1 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 and 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 and Traum 2006).

In other words, educational researchers go beyond the psycholinguistic 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 are 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 (Richardson and Dale 2005), where they found that dyad members are rarely perfectly synchronized; it takes participants ± 2 s 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 s 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 nine subdimensions: sustaining mutual understanding, dialogue management, 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 and Pea 2013). Some other studies have also found positive correlations between JVA and learning gains (Schneider and 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 4 s.

In addition to these measures of JVA, CSCL researchers have also looked at the "attentional similarity" between participants (Sharma et al. 2013). For a given time window (e.g., 5 s), the proportion of time spent on different Areas of Interest (AOIs) is computed and compared across participants using a similarity metric (e.g., the cosine similarity between two vectors). Papavlasopoulou et al. (2017) found that in a pair programming task, teenagers (13- to 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 2 s) and which student responded to it (i.e., the person whose gaze was there next). 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

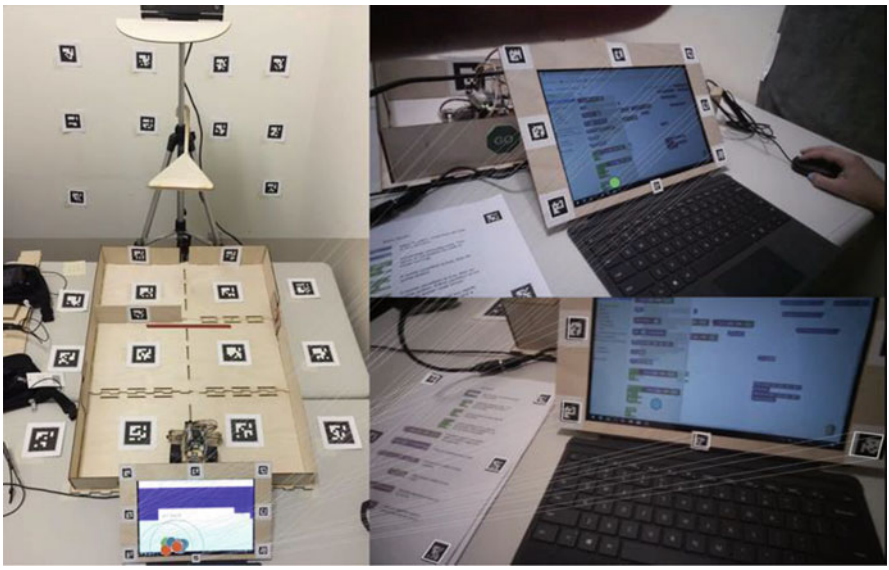

Fig. 2 (Reproduced from Schneider 2019): An example of using dual mobile eye-tracking to capture joint visual attention in a colocated setting (in this particular case, pairs of participants had to program a robot to solve a variety of mazes). The two images on the right show the perspective of the two 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)

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.

Additionally, researchers have started to go beyond remote collaboration and use dual eye-tracking in colocated 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 are 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 and 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 setting from an educational perspective. Most students still work in colocated 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.

3.2 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 et al. (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 electrodermal activation data were 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 et al. (2014b) found that body movements captured by a Kinect sensor could predict learning with 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 et al. (2014a), for example, found that nonverbal synchrony predicted creativity in 52 collaborative dyads. Models trained with synchrony scores could predict low or high scores of creativity with 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 to 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) were 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, the 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 toward 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 1, subtract 1, multiply by 10, or divide by 10. These subtasks

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.

3.3 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 and Pea 2013), attentional similarity (Sharma et al. 2013)	Body movement (Worsley and Blikstein 2013), prototypical states (Schneider and Blikstein 2015), physical synchrony (Won et al. 2014a)
Methodology	Well established; strong conventions (Richardson and 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 et al. 2014a; b); 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 and 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 et al. 2014b). 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, and 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.

3.4 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 and Pea 2015). Gestural data have 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. 2017), 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 modeling 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 modeling a meaningful

indicator of collaboration or learning. In terms of multifeature 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, multifeature 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 modeling 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. this volume).

From a teaching and learning perspective, modeling 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 et al. (2019) who proposed a modeling 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 a deeper understanding of collaborative learning, and deploying user interfaces that can provide tailored feedback on collocated activities, outweigh such challenges.

4 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 (e.g., through dashboards or awareness tools; Schneider and 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 and Schwarz (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. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 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*. Garden City: 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. doi: 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. doi: 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).

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. doi: https://doi.org/10.1145/1978942.1979230.

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

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. doi: 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. doi: 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. (2019). Unpacking collaborative learning processes during hands-on activities using mobile eye-tracking. In *The 13th International conference on computer supported collaborative learning* (Vol. 1, pp. 41–48). International Society of the Learning Sciences.

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* (Vol. 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.

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*. London: 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). Hillsdale, NJ: Lawrence Erlbaum.

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

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

Wise, A. F., Knight, S., & Buckingham Shum, S. (this volume). Collaborative learning analytics. In U. Cress, C. Rosé, A. F. Wise, & J. Oshima (Eds.), *International handbook of computer-supported collaborative learning*. Cham: 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. (2014b). 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. (2014a). 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. doi: 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. doi: https://doi.org/10.1145/2818346.2820737.

Further Readings

Abrahamson, D., Shayan, S., Bakker, A., & van der Schaaf, M. (2015). Eye-tracking Piaget: Capturing the emergence of attentional anchors in the coordination of proportional motor action. *Human Development, 58*, 218–244. https://doi.org/10.1159/000443153. This paper presents an empirical analysis that combines gaze and a gesture-based interface to surface attentional

anchors that students utilize when trying to represent fractions using their hands. The learning context is a Piagetian style interview, in which a pupil and an adult discuss fractions. The pupil is faced with a challenge, and the adult serves as a learning partner with whom the student may engage as they describe their thoughts and actions. However, the experience is also supported by a computational interface that allows the student to use gestures to control a visual display. By coupling eye tracking with the computer-supported learning interface, the authors are able to see the invisible attentional anchors that students utilize to recognize proportions and reduce the complexity of the task.

Cukurova, M., Luckin, R., Millán, E., & Mavrikis, M. (2018). The NISPI framework: Analysing collaborative problem-solving from students' physical interactions. *Computers and Education, 116*, 93–109. https://doi.org/10.1016/j.compedu.2017.08.007. The Non-Verbal Index of Students' Physical Interactivity (NIPSI) framework utilizes multimodal data to model student collaborative problem-solving. The authors operationalize synchrony, individual accountability, equality, and intraindividual variability. Each construct is based on the point-wise classification of student activity as being active, semi-active, and passive. Furthermore, each construct using the point-wise student activity classifications in different ways. For example, synchrony looks at the extent to which all participants in the group are exhibiting the same level of activity, which in this paper, is based on the active state. Intraindividual accountability looks at changes in activity between adjacent data points. Accordingly, the authors find that automatic annotation of these different activity states has utility for examining a number of different constructs related to collaborative problem-solving, and these automatic codes closely align with human-generated annotations.

Martinez-Maldonado, R., Kay, J., Buckingham Shum, S., & Yacef, K. (2019). Collocated collaboration analytics: principles and dilemmas for mining multimodal interaction data. *Human-Computer Interaction, 34*(1), 1–50 In this paper, Martinez-Maldonado, Kay, Buckingham Shum, and Yacef describe six studies that showcase different ways study colocated group work considering the meaning of gestures over interactive surfaces and embodied strategies of groups of learners in the physical space. In some of these studies, they combine multiple streams of data including speech detection, activity logs, and interactions with physical objects in both experimental and authentic classroom settings. Authors recommend a series of principles that can be applied to multimodal analytics and also describe a series of dilemmas in terms of data modeling, analysis, and sensemaking.

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. In this ijCSCL paper, Schneider, Sharma, Cuendet, Zufferey, Pea, and Dillenbourg describe a methodology for studying colocated groups: mobile eye-trackers. The authors provide a comprehensive description of the data collection and analysis processes so that other researchers can take advantage of this cutting-edge technology for capturing collaborative processes. They provide empirical findings showing that imbalances in leadership behaviors (captured by eye movements) are significantly correlated with learning gains. They conclude with some implications for automatically analyzing students' interactions using dual eye-trackers.

Stahl, G. (2006). *Group cognition: Computer support for building collaborative knowledge*. Cambridge: MIT Press. To facilitate the interpretability of the results of multimodal analytics of group activity, it has been recommended to build the analysis and design on foundational CSCL work. One of such foundations that consider the evidence generated by learners during their interactions as the core unit of analysis is that of group cognition. Group cognition is focused on understanding how two or more people communicate, via spoken language and other channels of communication, and interact with artifacts within a sociocultural setting. This work can serve for underpinning the analysis and fusion of multiple sources of evidence to understand how knowledge is built as a group, and how to connect low-level data with higher level constructs in the learning sciences.

Video Data Collection and Video Analyses in CSCL Research

Carmen Zahn, Alessia Ruf, and Ricki Goldman

Abstract The purpose of this chapter is to examine significant advances in the collection and analysis of video data in computer-supported collaborative learning (CSCL) research. We demonstrate how video-based studies create robust and dynamic research processes. The chapter starts with an overview of how video analysis developed within CSCL by way of its pioneering roots. Linked throughout the chapter are the theoretical, methodological, and technological advances that keep advancing CSCL research. Specific empirical and experimental research examples will illustrate current and future advances in data collection, transformation, coding, and analysis. Research benefits and challenges that include the current state of understanding from observations of single, multiple, or 360° camera recordings will also be featured. In addition, eye-tracking and virtual reality environments for collecting and analyzing video data are discussed as they become new foci for future CSCL research.

Keywords Video data · Video analysis · Learning research · Group research · Psychological methods

1 Definitions and Scope

The particularity of rich video data compared to other data gathering methods in the learning sciences is that video data make both verbal and nonverbal social interactions in learning situations *enduringly visible and audible* to researchers. In this regard,

C. Zahn (✉) · A. Ruf
University of Applied Sciences and Arts Northwestern Switzerland, School of Applied Psychology, Olten, Switzerland
e-mail: carmen.zahn@fhnw.ch; alessia.ruf@fhnw.ch

R. Goldman
NYU Steinhardt - Educational Communication and Technology, New York, NY, USA
e-mail: ricki@nyu.edu

U. Cress et al. (eds.), *International Handbook of Computer-Supported Collaborative Learning*, Computer-Supported Collaborative Learning Series 19,
https://doi.org/10.1007/978-3-030-65291-3_35

video data differ from *outcome data* (e.g., quantitative data gathered in learning experiments systematically examining treatments and their effects), because they can open the "black box" of collaborative learning processes. The scope of this chapter is to illuminate the scholarly understanding of existing and future methods for video data collection and data analysis in CSCL research in a practical fashion. The chapter maps past, present, and future innovative advances with specific examples selected to demonstrate the methods of video data collection and data analysis that learning science and CSCL researchers in a range of fields (e.g., Zheng et al. (2014)) have been using for a better understanding of complex collaborative learning processes.

CSCL video methods span the entire spectrum of the social sciences (Brauner et al. 2018), which includes qualitative research methods such as case-based fieldwork, and video ethnographic accounts, as well as quantitative methods such as experimental, and data-driven statistical research which includes learning analytics accounts. The majority of CSCL research articles in the *International Journal of Computer-Supported Collaborative Learning* (IJCSCL) as well as other related journals and volumes tend to consist of mixed methods studies, using both case-based and (quasi-)experimental research methods (e.g., Sinha et al. 2015; Zahn 2017). In this chapter, we will provide rich examples of how researchers can use video data for both deep qualitative case studies and with advanced and automated methods for complex visual analyses . Over time, we propose, they may also be used along with learning analytics.

We open the chapter with an historical overview of pioneering analog and digital video researchers in the learnings sciences and CSCL. We also delve into current research in CSCL video research to explore the benefits and challenges that exist now and will likely exist in the coming years. Questions about data collection, data transformation, data analysis, and interpretation will be followed by three examples of contemporary research studies. We will also present new approaches to using video data that allow for deeper post hoc observations of recorded learning interactions and digging into the details of knowledge co-construction and knowledge-building in and beyond CSCL research. For example, certain collaborative theoretical approaches, such as complex qualitative interaction analyses (Rack et al. 2019), focus on coordination and collaboration group processes. This interactional approach is especially enhanced by collecting and analyzing video data which can if needed, be linked to ethnographic video accounts.

The closing sections of this chapter address the current understanding of video data as observations from single or multiple cameras or 360° camera recordings. It will also look at the emergence of video data as ways of "looking through people's eyes" when eye-tracking or the use of virtual reality tools for collecting and analyzing video data are used (Greenwald et al. 2017; Sharma et al. 2017). Such tools represent promising areas for future developments. A deeper understanding of how a range of theories and collaborative methods and tools influence the research process can be found in Goldman (2007a), b, Part 1 and 4); Derry et al. (2010) as well as in Goldman et al. (2014).

2 History and Development: Pioneering Video Research

The twentieth century heralded in a range of new visual media forms such as social documentary, fictional photography, and ethnographic filmmaking. To study this topic more deeply, refer to the AMC filmsite called *The History of Film*. See: https://www.filmsite.org/pre20sintro2.html. The affordances of both photography and film were soon adopted by sociologists, anthropologists, and ethnographers around the world as tools for studying the lives of people at home, school, work, or play, in places both near and far. For example, anthropologist Margaret Mead and cybernetician Gregory Bateson used the film camera as a tool for social and cultural documentation, producing a film called *Bathing Babies in Three Cultures* in 1951 based on Mead's research comparing the bathing practices of mothers in three countries—New Guinea, Bali, and the United States. Mead, ever the futurist, imagined a time when there would be 360° cameras (Mead 1973). She thought it would take 10 years. It took 40!

2.1 Foundational Analogue and Digital Video Studies in LS and CSCL

Erickson (2011) looking back on his own early video observations of *learning processes* in groups emphasizes the central advantage which made him rely on audiovisual records to study learning in small groups: "...I could see who the speakers were addressing as they spoke—a particular individual, a subset of the group, or the whole group.... A multimodal and multiparty analysis of the locally situated ecological processes of interaction and meaning making became possible...." (p. 181). The camera he used weighed about 25 pounds and recording was done on reels that were about 16 inches in diameter.

One of the earliest breakthrough collaborative classroom studies of interpreting digital video data was conducted by Goldman-Segall (1998) at the MIT Media Lab. For over 3 years, her *digital video ethnography* at a Boston magnet school included videotaping computer activities of, and conversations with, grades 5 and 6 youth and their teachers. During the decade, Goldman-Segall developed the *Points of Viewing Theory* (1998) and the *Perspectivity Methodology* (Goldman 2007b) based on Clifford Geertz's (1973) notion of layering data to build *thick description*. Goldman along with Dong (Goldman and Dong 2007) advanced the ethnographic use of thick description to become *thick interpretations*, which were built by collaborative by layering diverse views of researchers, teachers, and students. For more than two decades, she designed digital video analysis environments with each new research study. The first environment was a simple HyperCard tool that enabled Goldman to establish categories gleaned from thematically arranged video excerpts that had been transferred onto videodiscs. By using her new tool called Learning Constellations, collaborating teachers and researchers could annotate, rate, analyze, and interpret the

video (1998). Following LC was the tool, WebConstellations in 1997, and Orion, an online digital video analysis tool for changing our perspectives as an interpretive community (Goldman 2007a). Each of these collaborative studies and the methods and tools is described in articles found in the references.

Modern technologies also allowed researchers to be more and more flexible in studying more complex learning situations comprehensively. For instance, Cobb and colleagues (e.g., Cobb and Whitenack 1996) studied children's mathematical development in long-term social contexts in a classroom study. Two cameras captured pairs of children collaborating on mathematics problem-solving over a course of 27 lessons, the authors articulate a three-stage method that begins with interpretive episode-by-episode analyses and meta-analyses resulting in integrated chronologies of children's social and mathematical developments.

Another comprehensive study was the collected videotaped records for classroom instructions from classrooms around the world called the Third International Mathematics and Science Study (TIMSS; Stigler et al. 1999). This video-based comparative study aimed at drawing comparisons between national samples. It set a standard for international sampling and video-based methods (Seidel et al. 2005): 231 eighth-grade mathematics lessons from Germany, Japan, and the United States were observed. In each classroom, one lesson was videotaped. The tapes then were encoded, transcribed, and analyzed based on a number of criteria. Analysis focused on the content and organization of the mathematics lessons and on the teaching practices that used a software especially developed for this study. According to Stigler et al., the advantages of videos compared to real-time observations make it possible for observers to work collaboratively on the video data. A further advantage described is the facilitation of communication of the research results. Similar advantages were achieved with new approaches when *digital video tools* entered the scene.

A comprehensive video workflow model for using video data in learning science research so that it can be shared was presented by Pea and Hoffert (2007). Their process model goes from the strategic planning of video research to a preproduction phase to the phases of video capturing, coding, storing, and chunking. Analysis then turns into collections of video segments, further statistical analyses, or case descriptions. The model moves from creating video as a means of observation and data collection toward decomposing video for analysis and then toward recomposing video for shared interpretation, collaboration, and discussion in a group or larger community of researchers. Pea and Hoffert thereby suggest that staying as close as possible to the video data during the research process, instead of translating results back and forth, affords an almost absolute closeness to the data during the whole process—also in sharing or presenting and discussing results. The authors introduce "WebDIVER," which is a streaming media interface for "web-based diving" into the video.

Koschmann et al. (2007) used the ethnomethodology of mini-chunks of video data to closely examine how learners form and act in collaborative communities. Their narrative methods used video to compose analytic narratives/stories from their footage.

Powell et al. (2003) developed a seven-step method through their longitudinal study of children's mathematical development within constructivist learning environments. The method starts with the researcher attentively viewing micro-video and then proceeds through stages of identifying critical events, transcribing, coding, until it ends with composing analytic narratives.

We will now discuss lessons that have been learned and how to integrate those lessons into future research practices.

3 State of the Art

Video analysis is now a common practice in learning science and CSCL research that spans across methodological approaches, be they experimental, quasi-experimental, field research, or case studies (see Derry et al. 2010). Video data are used to capture social and/or human-computer interactions, present moments of learning, and, in qualitative case studies, produce "collaborative learning accounts" (Barron 2003). In this section, we first take a generic methodological perspective that tackles the general challenges and practices of applying video analysis in the learning sciences. Then we highlight specific problematics, and solutions when video data are used with qualitative, quantitative, or mixed-methods research in CSCL settings, provide examples for CSCL video collections and analysis. Ethnographic, narrative, problem-based, and design-based methods are also included.

3.1 Benefits and Challenges of Using Video Methods in Learning Science Research

From a methodological viewpoint, researchers agree that video-based research provides highly valuable data on learning processes in collaborative settings. For instance, they provide detailed process data that can be analyzed in an event-based, but also in a time sequence-based approach (for analysis of discrete event sequences, see Chiu and Reimann this volume). At the same time, such research is highly selective and researchers' decisions determine what is being recorded and analyzed. Researchers' decisions precede the production of video data, adding their points of viewing on them at all stages of the research (Goldman-Segall 1998). On the one hand, video technologies can be beneficial in that they represent powerful ways of collecting video data with easy to use, relatively lightweight, and affordable cameras. They also constitute well-designed web-platforms for storage and for sharing video data with other researchers, and they are effective tools for deeper analysis and video editing. Despite these notable advantages, as Derry et al. (2010) specify, there are challenges posed to researchers who collect and use video records to conduct research in complex learning environments. These challenges include

developing or finding appropriate analytical frameworks and practices for given research goals; identifying available technologies and new tools for reporting, and sharing videos; and, protecting the data and rights of participants, i.e., ethics and privacy issues. Blikstad-Balas (2017) adds further key challenges: contextualization, (getting close enough to a situation to detect details, but always keeping an extra eye on the context); magnification (magnifying small details that might be irrelevant for learners in the situation, even if it may be critically important to researchers); representation (presenting data in a way that others can understand and follow scientific interpretations).

With respect to the tension between the aforementioned benefits and challenges, Derry et al. (2010) suggest careful consideration of the different phases in the practice of using video analysis and interpretation of results. In each of these phases, researchers must be aware of the consequences of their selections, decisions, and the procedures they apply.

3.2 Specification for CSCL Research—Selected Research Examples

In CSCL, we consider specific issues related to the use of video data in computer-supported and collaborative learning. An additional challenge for CSCL settings is that researchers have to integrate or synchronize data streams on social interactions or conversations (recorded in a physical space) with further data (e.g., screen recording or logs of human–computer interactions). How can this be accomplished in practice? The following three examples illustrate possible solutions: First, a case study of a qualitative and in-depth analysis of the collaboration process based on online verbal communication, where video was used as an additive. Second, an exploratory study following $N = 5$ groups over a time span of a 6-week course where video analysis based on coding and counting was central and both conducted and reported in a very distinguished way. Third, an example from experimental research with a sample of $N = 24$ pairs of learners where video analysis was used in a complex and multileveled mixed-methods approach.

In the first example, Vogler et al. (2017) report a case study on the emerging complexity of online interactions and the way participants contribute through meaning making in a classroom discussion that took place in a CSCL environment. The research question was how meaning emerges from the collective interactions of individuals. In particular, the researchers investigated how the small groups introduced, sustained, and eventually closed a discussion topic. Therefore, computer-mediated discussions of small student groups in class were analyzed. Data were collected by means of screen recording (Camtasia software) for capturing the participants' activities on the computer (e.g., any changes that occurred on the screen display, typing, deleting, or opening of online resources). Further on, the researchers captured by means of four video cameras the activities and interactions that took

place in the physical classroom—i.e., the small groups of two to three participants were recorded (e.g., eye gaze away from the screen, body movements, and accessing offline materials). In addition, trained observers took ethnographic notes. From the collected online conversations, the researchers created transcripts, coherence maps (for an example, see Fig. 1), and then spreadsheets showing how individual comments were connected and how threads and topics evolved (Vogler et al. 2017). The authors report on microanalyses of those learners' discourses and present a detailed analysis of the life cycles of two selected discussion threads. The video recordings from the four classroom cameras were used as additional data together with the researcher's observations. The data streams were synchronized by means of a tedious process that had to be done manually prior to analysis. It would have been interesting to couple different data sources (screen recordings, written discussion threads, and video recordings of nonverbal behaviors) using complex and elaborate visual analysis methods. A point to which we will return below.

In the second example, Näykki et al. (2017) examined, in an exploratory study, the role of CSCL scripts for regulating discussions during a during a 6-weeklong environmental science course in teacher education. The scripts (i.e., prompts presented on tablet computers) aimed at supporting the planning and reflection of the collaborative process. The authors compared processes of scripted and non-scripted collaborative learning asking how socio-cognitive and socio-emotional monitoring would emerge in groups depending on the (more or less) active use of such scripts. They also investigated how monitoring activities would transfer to subsequent task work. The study took place in a classroom-like research space and video data were collected by means of a 360° recording method (for details, see: https://www.oulu.fi/leaf-eng/node/41543). The authors extracted 30 h of video data (discussions, movements, and gestures) from five student groups that were repeatedly captured five times. A multistep analysis method was applied for analysis: the video data were first segmented into 30-s events. Each 30-s segment was annotated by a researcher with a description of what had occurred within the segment resulting in a content log of each video (e.g., group finishes task; one person shows their created mind map to others, group discusses task completion, suggestions on further proceeding). The content log of each video was complemented with a comprehensive memo of the most salient observations. In a second step, each 30-s segment was observed to see if group members showed socio-cognitive and socio-emotional monitoring (i.e., the behaviors associated with the understanding and progress of the study-like task, content understanding, socio-emotional support). The subsequent development of categories and coding procedure is described thoroughly in Näykki et al. (2017). Twenty-five percent of the video data were also coded by an independent coder. Upon this data, frequency analysis was applied for further statistical hypothesis testing. Time-based video segmentation was also applied by Sinha et al. (2015) studying collaborative engagement in CSCL groups, but here the video segments were subjected to observer ratings of the quality of collaborative engagement in small groups (high, moderate, or low) and used for qualitative case studies.

In the third example, $N = 24$ pairs of students were investigated when learning with advanced digital tools in history lessons (Zahn et al. 2010). Two conditions

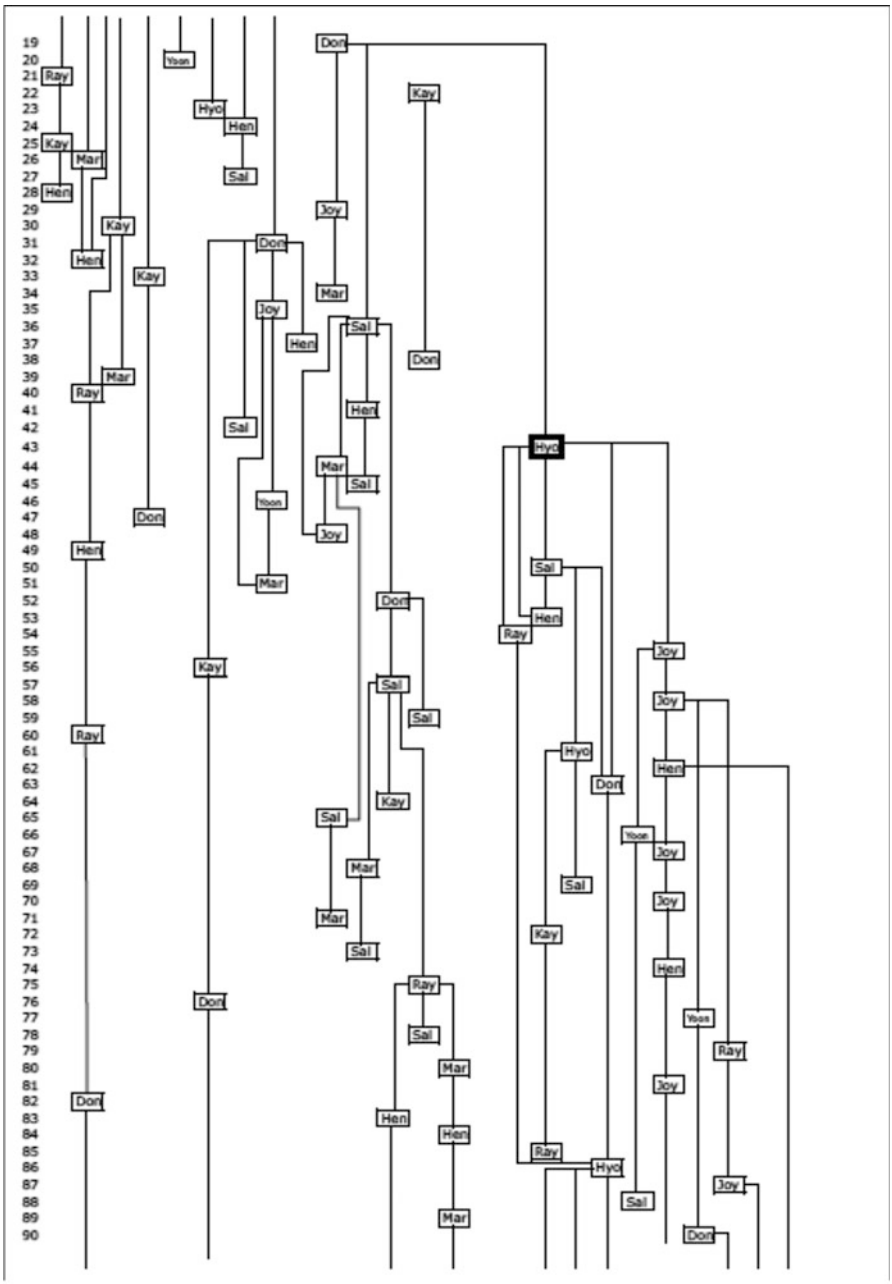

Fig. 1 Example of a coherence graph kindly with friendly permission by Jane Vogler

supporting collaborative learning were compared: one where students used an advanced web-based video tool (WebDIVER, see Pea and Hoffert 2007) and one where students used a simple video player and text tool (controls). The advanced tool allowed cutting out of details from video sequences and extracting those "pieces of video" in order to comment on the details. Students' interactions with technology were captured by means of screen recording (Camtasia Studio by TechSmith) and dyadic social interactions were recorded by means of a webcam. In order to analyze these data, a mixed-methods strategy was applied in combining both types of data in a two-step coding procedure (for subsequent quantitative analyses) and integrated activity transcripts (for subsequent qualitative case studies). Trained observers first watched the video recordings of social interaction to identify emergent behavior categories and then applied a process of coding and counting. Eight categories of verbal interactions were found in this process (e.g., content-related talk, video-related talk, technical issues talk, help seeking, etc.). The relative amounts of time spent for talking in the categories, related to total talking time, were then calculated and compared between conditions.

Transcripts of learning episodes were produced for deep analyses of selected cases and specific categories (e.g., content-related talk) from the different conditions. The transcripts synchronized the students' conversations and interactions with digital tools (e.g., typing, submitting comments, playing video, watching, stopping video, rewinding, making marks with an advanced video function, etc.). The transcripts were analyzed according to Barron (2003) as "localized accounts" of "successful learning."

Based on this qualitative approach, it would be interesting in further research to return to a quantitative strategy by counting collaboration patterns in dyads from both conditions and compare their prevalence statistically thereby testing for significance. Yet, limited resources often force research to disclaim such mixed-method approaches. Future perspectives, however, include automated analyses that could render this option feasible.

In sum, from these examples, it can be noted how a number of decisions were made in the phases of video data collection and analyses, starting from the number and types of cameras used as well as their placement in the investigated scene; to the number of groups and group sizes under scrutiny; the duration and frequency of video data collection; the decision of using transcripts for qualitative in-depth analysis versus developing categories to be coded and counted or both; the using of extra-visualizations or verbal comments for data exploration; and, to the selecting of results to be presented in a scholarly publication.

4 The Future

Video analysis has evolved rapidly alongside recent technological progress (e.g., mobile eye-tracking, social computing, virtual reality). In this section, we will look ahead and include developments such as tracking and automatic data analysis methods from social computing technologies.

4.1 Eye-Tracking in CSCL Research

Eye-tracking as a method to investigate learning behaviors has been widely used in individual learning settings in the last few years (for an overview, see Alemdag and Cagiltay 2018; Lai et al. 2013). Mobile eye-tracking, for example, was applied to research on informal learning in museums (Mayr et al. 2009; Wessel et al. 2007) where researchers could reflect on eye-tracking videos afterward together with visitors in order to gain insights into motivational factors and possible effects of exhibition design on learning during a museum visit (vom Lehn and Heath 2007).

Although using eye-tracking in CSCL is not unknown (e.g., Stahl et al. 2013), it still seems rather uncommon. Since 2013, only few studies were published that used eye-tracking as a method in CSCL research. Among these are the studies by Schneider et al. (2016), Schneider and Pea (2013, 2014), Sharma et al. (2017), and Stahl et al. (2013) that emphasize the advantages and possibilities of eye-tracking as a method to support and research collaboration. Schneider and Pea (2013) investigated collaborative problem-solving situations, where dyads saw the eye gazes of their learning partner on a screen. The authors found that this mediated joint visual attention helped dyads achieve a higher quality of collaboration and increased learning gains. These results indicate that joint visual attention in collaborations is of great importance in problem-solving settings, as it fosters an equal understanding of the problem (Stahl et al. 2013; Zemel and Koschmann 2013). In a follow-up study, Schneider and Pea (2014) examined collaborative learning processes in dyads working remotely in different rooms. Similar to their previous study (Schneider and Pea 2013), participants were able to see the gaze of their learning partner on the screen. Using eye-tracking data, Schneider and Pea (2014) could roughly predict collaboration quality with an accuracy between 85% and 100%. Hence, joint attention (which involves gaze) is an important nonverbal predictor and indicator for successful collaboration. In addition, Schneider et al. (2016) investigated the way users memorize, analyze, collaborate, and learn new concepts on a tangible user interface (TUI) in a 2D versus 3D interactive stimulation of a warehouse. Eye-tracking goggles were used as a method to further investigate collaboration processes in colocated settings. Results suggested that 3D interfaces fostered joint visual attention which significantly predicted task performance and learning gains. The little existing research about using gaze in CSCL has demonstrated that eye-tracking data contributes highly relevant and important insights into collaborative processes (see also Schneider et al. this volume). Sharma et al. (2017) elaborated that "eye-tracking provides an automatic way of analyzing and assessing collaboration, which could gain deeper and richer understandings of collaborative cognition. With the increasing number of eye-tracking studies, in collaborative settings, there is a need to create a shared body of knowledge about relations found between gaze-based variables and cognitive constructs" (p. 727).

With eye-tracking devices, especially mobile eye-trackers, becoming cheaper and widely available, we expect increases in eye-tracking studies in future CSCL research. For this reason, theoretical frameworks for eye-tracking research in