Read Between the Lines: An Annotation Tool for Multimodal Data for Learning

Daniele Di Mitri¹, Jan Schneider², Roland Klemke¹, Marcus Specht¹, and Hendrik Drachsler^{1,2}

¹ Welten Institute, Research Centre for Learning, Teaching and Technology
Open University of the Netherlands - Valkenburgerweg 177 6401 AT Heerlen, The Netherlands

² Leibniz Institute for Research and Information in Education
Rostocker Straße 6, 60323 Frankfurt am Main, Germany
{daniele.dimitri,roland.klemke,marcus.specht,hendrik.drachsler}@ou.nl,schneider.jan@dipf.de

ABSTRACT

This paper introduces the Visual Inspection Tool (VIT) which supports researchers in the annotation of multimodal data as well as the processing and exploitation for learning purposes. While most of the existing Multimodal Learning Analytics (MMLA) solutions are tailor-made for specific learning tasks and sensors, the VIT addresses the data annotation for different types of learning tasks that can be captured with a customisable set of sensors in a flexible way. The VIT supports MMLA researchers in 1) triangulating multimodal data with video recordings; 2) segmenting the multimodal data into time-intervals and adding annotations to the time-intervals; 3) downloading the annotated dataset and using it for multimodal data analysis. The VIT is a crucial component that was so far missing in the available tools for MMLA research. By filling this gap we also identified an integrated workflow that characterises current MMLA research. We call this workflow the Multimodal Learning Analytics Pipeline, a toolkit for orchestration, the use and application of various MMLA tools.

CCS CONCEPTS

• **Applied computing** → Computer-assisted instruction;

KEYWORDS

Multimodal data, Internet of Things, Learning Analytics, sensors

ACM Reference format:

Daniele Di Mitri¹, Jan Schneider², Roland Klemke¹, Marcus Specht¹, and Hendrik Drachsler^{1,2}. 2019. Read Between the Lines: An Annotation Tool for Multimodal Data for Learning. In *Proceedings of The 9th International Learning Analytics & Knowledge Conference, March, 2019, Tempe, AZ, USA, (LAK19)*, 10 pages.
<https://doi.org/10.1145/3303772.3303776>

1 INTRODUCTION

Multimodal interaction methods are becoming increasingly popular in learning science. User-computer interaction data, traditionally

derived from software logs or clickstreams are being enriched by additional data sources. These new data sources are gathered with wearable sensors, Internet of Things (IoT) ubiquitous devices or embedded systems. The driving reason for this shift is to achieve a more comprehensive evidence and analysis of learning activities taking place in the physical realm. In the last decade, research in learning science using data-driven technologies such as learning analytics, educational data mining or intelligent tutoring systems have focused primarily on online, desktop or mobile learning. In these settings, learning activities are mediated by computer devices. In contrast, settings as co-located collaborative learning, practical skills training, or dialogic classroom discussions are less addressed in data-driven learning research. Including novel data sources to investigate these learning activities happening in the physical space is in line with the overarching theme of the Learning Analytics & Knowledge conference 2019, which focuses on ‘inclusion’.

In support of these new forms of interaction, within the Learning Analytics community, a new research focus has emerged, coined as multimodal learning analytics (MMLA) [6]. The objective of MMLA is to track learning experiences by collecting data from multiple modalities and bridging complex learning behaviours with learning theories and learning strategies [38]. MMLA, however, is an emerging research field that needs more evidence: a recent survey analysed eighty-two MMLA papers, forty-six of which are empirical, while the reminder theoretical [39]. Along with more evidence, MMLA needs structured technological practice. Whilst the learning analytics community put considerable efforts into standardisation and interoperability into data collection, analysis and exchange, these efforts did not yet meet multimodal data. Researchers using multimodal interaction approaches face multiple challenges that stem from the complex nature of multimodal data [19]. For the case of MMLA, when using multimodal data for learning, the challenges have been grouped into five categories [10].

- (1) The *data collection*: the approach used for capturing, aggregating and synchronising data from multiple modalities and sensor streams;
- (2) the *data storing*: the approach used for organising multimodal data which having multiple formats and big sizes, for storing them the and logic for later retrieval;
- (3) the *data annotation*: the approach for providing meaning to portions of multimodal recordings and collecting human interpretations through expert or self-reports;

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

LAK19, , March, 2019, Tempe, AZ, USA

© 2019 Association for Computing Machinery.

ACM ISBN 978-1-4503-6256-6/19/03...\$15.00

<https://doi.org/10.1145/3303772.3303776>

- (4) the *data processing*: various steps for cleaning, aligning, integrating, extracting relevant features from the “raw” multimodal data and transforming them into a new data representation suitable for exploitation;
- (5) the *data exploitation*: the approach to ultimately support the learner during the learning process with the predictions and the insights obtained by the multimodal data.

As a recent review concluded, MMLA has yet no structured approach to handling multimodal data for learning gathered from different sources [33]. In contrast, multiple tailor-made solutions have been developed in related research adapted for specific use-cases, making the process of MMLA time consuming and expensive. This aspect constitutes, in our view, a great shortcoming for the field which hinders MMLA solutions to be adopted in the practice. Standard approaches could help researchers simplify the setup of their experiments. They would allow practitioners benefiting from more inclusive Learning Analytics that go beyond mouse and keyboard interaction, including learning activities in the physical space. To fill this gap in the current research, we propose a structured approach for MMLA. We first explain in section 2 the differences that multimodal bring about compared to traditional user-interaction data (section 2.1). We describe the challenges introduced by sensors (section 2.2) as well as the potential of the multimodal approach (section 2.3). We then review different existing multimodal tools and see how they compare them in terms of their functionalities and purposes and classifying them their features according to five MMLA challenges (section 2.4). In section 3 we introduce our prototypical solution, consisting of the design and implementation of the Visual Inspection Tool (VIT). The VIT is a generic solution for the third category of MMLA challenges, the data annotation. We list a set of functional requirements for the VIT (section 3.1) and we describe how we implement the different components of it (section 3.2). We propose a procedure for the data processing (section 3.3) and data exploitation (section 3.4) and we test the VIT in three use cases (section 3.5). In section 4, we discuss the findings of our approach positioning the VIT within an integrated workflow, the *Multimodal Learning Analytics Pipeline* which works as guideline for MMLA researchers. Finally, in section 4 we draw conclusions and point to future directions.

2 BACKGROUND

2.1 Computer Assisted Learning without mouse and keyboard

To date, learning research communities and practitioners using data-driven approaches in learning focus on traditional and easy-to-retrieve data sources. These sources include interaction data with the learning platform, such as the Learning Management System (LMS) the Massive Online Open Course (MOOC), data crawled from social media or collected from mobile applications. The accessibility and the richness of these “traditional” data sources has in recent years motivated and inspired learning science research communities, such as the Learning Analytics and Knowledge community [15]. These traditional data sources only capture the learner’s interactions with the learning system or platform, hardly capture any other moment in which learning occurs. For instance, it was found that the usage of a mobile application used for learning was not a

reliable source for measuring the actual time the learner was spending on the learning task [35]. The ‘traditional’ learner-computer interactions are events that share similar properties. For each of those events, it is straightforward to derive *who did what*, as they are already sorted by user, action and activity. Such semantic encoding is used by interoperability standards as Experience API for specifying events such as “user likes post”, “user watches video”. With these types of events there is no need to decipher *who* did the action, *when* it was done and *what* kind of action it was. This semantic encoding is easy to be interpreted by humans and for machines. They are easily summarisable with aggregate functions and fed into visualisations or into dashboards showing the activity level and the learning effort. But what happens during the learning moments in which the interaction with a computer is not the main activity? For instance, when learning how to use a circular saw in a carpentry school, when elaborating a solution to a challenge in a co-located group discussion, or when participating in classroom discussions with teachers and peers. These are examples of learning activities frequently happening in everyday classroom or workplace settings. Psychomotor skills training, dialogic learning and co-located group interactions are dominated by physical interactions, which remain most of the time *untracked*. The physical interactions are not captured and, therefore, are not included in the datasets used for analysis. In these physical learning scenarios, mobile devices or laptops are not used as main media, but rather as side-tools, for example for taking notes or looking-up web resources. Bringing these moments into account requires a more seamless data collection that goes beyond the direct learner-computer interactions derived from clicks, keystrokes or nested software logs.

2.2 Sensors in learning

Beyond mouse and keyboard events there are multimodal sensor data, i.e. data that can be collected with sensors, computer chips, wearable trackers, microphones, cameras and other computer input devices. Consumer level sensors, wearables and IoT devices introduce new multimodal affordances that can be exploited to collect evidence of learning actions in a wide range of situations. There exist a large variety of sensor devices that can be used in learning contexts. Schneider et al. [30], have identified more than 23 different sensor devices that have been or can be used in the domain of learning. Sensors can monitor both the physical environment where the learner operates and the learner’s behaviour including 360-degrees body movements, physiological responses such as heart rate or body temperature, or interpersonal communication, student-teacher or student-peers discussions. In related literature, sensors have been used to track different modalities such as hand gestures [24, 40], gross body movements [7], eye-tracking [28, 29], facial expressions [3, 4]. Sensors were also used to monitor physiological signals such as heart rate [1, 16], galvanic skin response [14, 25], brain waves [1, 28]. The recent framework of Limbu et al. proposes how to use sensors to capture the expert performance and re-enact it through Augmented Reality using a variety of approaches, coined as Transfer Mechanisms [21]. Despite the extensive body of literature featuring sensors, still a very small percentage of studies opted for non-tailored and more generic

approaches. In addition, to the best of our knowledge, there is little use of MMLA solutions in actual educational practices. While in many sectors such as automotive, manufacturing, retail, smart living or health care, IoT and sensor-based systems are now very common, similar systems have not been developed yet for education and learning. Compared to other domains, human learning is a more complex and individual process much harder to model. The learning environment for humans is usually open-ended and highly unpredictable. There are many factors which influence learning, including the social relationships with teacher and peers, or the necessity to close the feedback loop to correct and improve learning behaviour.

2.3 Multimodal data for personalised learning

Multimodality is a theoretical assumption that can be used to provide more structure in the use of sensors for investigating learning. The idea of using multimodality in learning derives from the theory of embodied communication. According to this theory, humans use their whole body to communicate, they use a multitude of channels to exchange messages in shared contexts [36]. Similarly to human-human communication, the multimodal principle can also be employed in human-computer interaction. Sensor-based multimodal interfaces can monitor the variation of different modalities during learning activities: including dialogic learning in the classroom [12], in computer-supported collaborative learning during knowledge-sharing and group discussions [22, 27], practice-based and open-ended learning tasks [38], when understanding and executing a practical learning tasks [24] or training presentation skills [31]. The benefit of multimodal data is, first of all, enriching the learner’s digital representation as well as the ones of the learning context, environment or task. Enriched representations can shed more light on cognitive states or metacognitive factors that influence learning. Di Mitri et al. [2018a] provided a model, the *Multimodal Learning Analytics Model* that frames the different phases of MMLA and describes how MMLA outcomes can be used for personalised feedback and reflection for learners. Using the multimodal data collected during a learning task, machine learning algorithms can be trained to classify, cluster or predict learning dimensions that are “invisible” to sensors, such as learner’s emotion, learner’s cognition or outcomes. These predictions can be used to tailor different types of feedback, which can be used to guide the learner towards predefined learning goals, whether this is improving a particular task or stimulating self-reflection. Multimodal data can be used to enhance intelligent tutoring systems (ITSs) employed in physical learning activities, providing a more accurate student model representation. Figure 1 represents this aspect with a typical ITS tripartite structure 1) Knowledge domain, 2) Student model, 3) Pedagogical model [26] enhanced by multimodal data collected in the physical realm.

2.4 Tools for Multimodal Data

One relevant aspect of the MMLA research is the design of the technical infrastructure and the choice of technical tools to handle the data gathered from multiple modalities. Compared to learner-computer interaction data, sensor data pose a much bigger challenge: single behavioural particles have low semantic meaning if

Figure 1: Multimodal data can extend the digital representation of the learner in the system.

considered singularly [11]. For example, the skeleton data recorded with Microsoft Kinect can record the full body movements which translate to more than fifty sensor attributes. On the one hand, visually representing raw data values for human inspection is not informative, on the other hand, multimodal data can be processed by computer based algorithms which need more complex architectures which lack in the learning domain. There exist, however, technical tools in contiguous disciplines to MMLA as social signal processing, which can solve some specific parts of the data-informed cycle. In this section, we review seven of these tools, later we classify them into five categories previously introduced.

2.4.1 Social Signal Interpretation & NovA. The Social Signal Interpretation (SSI) [37] is an open source framework for real-time recognition of social signals during social interactions. SSI supports synchronised data recording from a large range of sensor devices to recognise behavioural cues such as gestures, head movements, and emotional speech. SSI allows the user to collect their own training corpora and obtain personalised models. Through an XML editor, SSI allows to draft and run pipelines by connecting multiple sensors, to data processing techniques (transformations) and visualisations (consumer applications). While SSI has plugins for multiple existing sensors, it does not clarify how to connect new ones. One relevant extension, NovA (NonVerbal behaviour Analyzer) [5] allows the user to annotate behavioural cues and measure the quality of the social interactions, in terms of user’s levels of engagement and activation in the interaction. SSI is an open source software (GPL license) written in C++.

2.4.2 Lab Streaming Layer. Lab Streaming Layer (LSL) focuses on a unified collection, synchronisation and storing time series data [18]. Developed by the Swartz Center for Computational Neuroscience, LSL implements plug-ins for multiple brain-computer interface devices primarily used in neuroscience research such as electroencephalograms (EEG). LSL creates *stream outlets* where streaming data can be published in samples or chunks with regular or irregular sampling ranges. Receiver nodes can also subscribe

through *stream inlets*, topics to which computers in the same network can subscribe. LSL uses functions to discover and resolve streams of data in the network. LSL also features a built-in clock that allows assigning timestamps to the collected data samples in order of sub-millisecond accuracy using the Network Time Protocol. LSL uses a custom data format, the eXtensible Data Format (XDF). LSL is open source cross-platform software that offers interfaces in C, C++, Python, Java, C#, Matlab.

2.4.3 Data Curation Framework (DFC). Data Curation Framework (DFC) focuses on raw sensory data acquisition, curation and monitoring [2]. It implements an IoT approach for data collection in real time from multiple modalities distributed environment over a cloud-based platform. DFC implements a rule-based anomaly detection system. As the computation is performed over the cloud platform, the interesting side of DFC is the scalability. DFC introduces the concept user's lifelogs curation through continuous monitoring of the sensory data. The DFC system was created for pervasive health monitoring and, for this reason, it implements algorithms as anomaly detection based on expert hardcoded rules. DFC is written in Java and Javascript and released open source under Apache license 2.0.

2.4.4 ChronoViz. ChronoViz is a tool for the visualisation of time-based data from multiple synchronised streams of data [13]. ChronoViz allows researchers to navigate multimodal data, including video, audio, digital nodes and geographic data and add text-based annotations. It is designed to support multiple videos and support log data (e.g. from flight simulator), geographic coordinates, video and audio transcripts or notes taken on paper with a digital pen. The annotations are overlaid on top of the synchronised graphs and they can be grouped into categories of annotations that have different colours. ChronoViz runs as a local Mac OSX application, its code is open source under Apache license 2.0.

2.4.5 RepoVizz. RepoVizz is a data repository and visualisation tool for storage and user-friendly browsing of music performance multimodal recordings [23]. RepoVizz offers means for researchers to access music performance online in a shared multimodal database. RepoVizz supports several formats for audio and video. It also accepts sensor files in CSV format (e.g., motion capture or physiological signals). The data is structured into an XML (repoVizz Struct) which through metadata allows the user to organize multimodal data in a hierarchy. The XML files and all the multimodal data sources are called repoVizz *Datapack* and are uploaded into zip files. It differentiates users types into Producers and Consumers allowing different downloading, uploading and annotating rights. RepoVizz allows annotations for different sound streams to segment different notes identify different instruments playing in collaborative musical settings such as orchestras. RepoVizz is an open source software written with Javascript-HTML5 front-end, Java backend and MySQL database.

2.4.6 Generalized Intelligent Framework for Tutoring. GIFT is a framework that provides tools, methods and standards for the design and evaluation of computer-based tutoring systems [34]. GIFT is developed by the Learning in Intelligent Tutoring Environments Laboratory, part of the U.S. Army Research Laboratory. GIFT consists of a *sensor module* interfacing with commercial sensors for

processing and storing sensor data; a *domain module* providing the content to support the training and assessing the learner's performance against standards; the *pedagogical module* identifying the need for feedback. GIFT uses sensor data to identify the learner's affective, cognitive and psychomotor states. GIFT source code is available via registration.

2.4.7 Multimodal Learning Hub. The Multimodal Learning Hub (LearningHub) is a system that focuses on the data collection and data storing of multimodal learning experiences [32]. It uses the concept of *Meaningful Learning Task* (MLT) and introduces a new data format (MLT session file) for data storing and exchange. The LearningHub implements a set of specifications that shape it for certain types of learning activities. It was created to be compatible primarily with commercial devices (e.g. Microsoft Kinect, Leap Motion, Myo Armband) and other sensors with drivers running with the most common operating systems. It focuses on short and meaningful learning activities (10 minutes) and uses a distributed, client-server architecture with a master node controlling and receiving updates from multiple data-provider applications. It also handles video and audio recordings with the main purpose to support the human annotation process. The expected output of the LearningHub is one (or multiple) MLT session files including 1) *one-to-n* multimodal, time-synchronised sensor recordings; 2) a video/audio file providing evidence for retrospective annotations. The LearningHub is open source and developed in C#.

3 METHODOLOGY

Among the five categories of challenges of MMLA this paper puts its focus primarily on data annotation and consequently on data processing and exploitation to complement which, as the review of multimodal tools reveals, are the MMLA challenges which are not sufficiently addressed. To address this gap, we formulate the following research questions: RQ1) *How to design a system to annotate multimodal recordings?* RQ2) *What processing techniques can be used with annotated multimodal recordings?* RQ3) *How to exploit annotated multimodal recordings in different learning scenarios?*

3.1 Design specifications

Data annotation requires a human evaluator, typically an expert, who can be a researcher or a trainer. The user assigns meaning to portions of the multimodal sensor recording through direct or retrospective observations. This procedure is also known as *triangulation*: the user *views* the interpretable evidence (e.g. video) while *plotting* and visualising synchronised multimodal recordings; thus, the user *adds annotation* transferring meaning to multimodal recordings which are hard to interpret. The annotations should be *customised* according to the chosen annotation scheme and the output - the annotated dataset - can be *downloaded*. Most of the time, there are multiple experts who need to *shared access* to the same datasets. Depending on the richness of the sensors, processing the dataset can become computationally expensive, therefore a *scalable* and dynamic allocation of processing power is needed. This scenario can be seen in form of functional requirements:

(FR1) the user can *plot* and visualise a multimodal recording file, featuring multiple synchronised data streams;

Tool	Collection	Storing	Annotation	Processing	Exploitation	Main purpose
1. Social Signal Interpretation [37]	Multisource, Synchronised streams	No custom format	Using NovA	Custom pipelines, various ML algorithms	n.a.	Human activity recognition
2. Lab Streaming Layer [18]	Multisource, streams	Custom data format(XDF)	n.a.	n.a.	n.a.	Physiological data synch.
3. Data Curation Framework[2]	Multisource, synchronised streams	n.a.	n.a.	Anomaly detection	n.a.	Pervasive health-care monitoring
4. ChronoViz [13]	n.a.	n.a.	text-based annotations	n.a.	n.a.	Video coding human interactions
5. RepoVizz [23]	n.a.	Custom data format (repoVizz struct)	text-based annotations	n.a.	n.a.	Visual analysis of multi-user orchestration
6. GIFT [34]	Multisource, batches	Store in csv format	n.a.	Can be linked with external processing tools	Corrective and personalised feedback	Designing ITS
7. Multimodal LearningHub [32]	Multisource, synchronised batches	Custom data format (MLT)	n.a.	n.a.	Corrective feedback	Intelligent Learning Feedback

Table 1: Comparison of the existing multimodal tools. (n.a. = not available)

- (FR2) the user can *view* video of the session synchronised with the multimodal data;
- (FR3) the user can *add annotations* to single time intervals in attribute-value form;
- (FR4) the user can add *custom* annotations;
- (FR5) the user can *download* the annotations or attach them to the session file;
- (FR6) the tool should be compatible with cloud-based solutions for *scalability* and *shared access*.

For the specific challenge of data annotation, the existing tools that best fulfil these requirements are ChronoViz and repoVizz. ChronoViz fulfils FR1 and FR2, however it is not designed to handle datasets having a large number of sensor attributes. It runs locally and therefore does not fulfil FR6. It allows custom annotations but only consisting of text (not attribute-value) therefore does not fulfil FR3. Similarly, it does not allow to select time intervals but only time points. repoViz instead fulfils FR1, FR2, FR6. When it comes to annotations, however, the tool focuses more on annotating the different contributions of single modalities (e.g. musical instruments) rather than annotating a new corpus to prepare it machine learning or feedback. On this note, SSI and its extension NovA focus on creating custom corpora. One aspect implemented in DCF and repoVizz is the use of web compatible technologies which relate to FR6. repoVizz, contrarily to the other software, implements a web interface while DCF performs operations in the cloud. These decisions allow for better scalability than software running locally on a single computer.

As our solution focuses primarily on multimodal data annotation, we need to apply other solutions for data collection and storing. Among the reviewed tools, the candidates are SSI, LSL and LearningHub. The upside of LSL is the high fidelity data collection and custom format. For SSI is the wider community and support and the variety of plugins. Both do not clearly explain their extensibility with new sensors and their contribution to user feedback. The

LearningHub, despite being more recent and experimental, focuses on learning and therefore prioritises the feedback loop. We decide to adopt the LearningHub and its data format (MLT session file) providing synchronised multimodal sessions with video recordings.

3.2 Implementation

To fulfil the design requirements, we develop the Visual Inspection Tool. The VIT is developed as a server-less web application in Javascript and HTML5 running on an Internet browser. We chose this technology because it is highly compatible. The VIT allows for plotting datasets and while playing multimedia files such as videos, and high performance is not a requirement for this type of application. VIT can plot the multimodal data of the video recording of the session for triangulation. A comprehensive screenshot is represented in Figure 2. VIT contains a parser for the custom MLT data format and is currently able to handle a great number of sensors attributes. In the use cases, we have tested in section 3.5, we were able to load up to one hundred different attributes. This number can easily scale higher, since, by defaults, the attributes are not loaded into the chart unless they are clicked by the user.

3.2.1 Session file. The multimodal recordings used as input corresponds to a MLT session file of the LearningHub. The data format of these sessions is a zip compressed folder. In this version of the session folder, it can be loaded on the panel Session load (Figure 2, A, top-left) by picking a session stored on the local drive. If VIT runs on an online server, the session files could also be picked from a cloud repository. When selecting a session an asynchronous parser is activated to scan the content of each file contained in the session zip folder. Following the specs of the LearningHub, the files can be either JSON (Sensor application files) or MP4 file formats. The JSON files will be parsed as described in the next section and the video will be loaded (Figure 2, G, top-right). An example of such a file is shown in Listing 1.

Figure 2: The VIT features the following elements: (A) Session loader MLT session file custom data format; (B) Attribute list; (C) Annotation file loader; (D) Time intervals editor; (E) Annotation editor; (F) Attribute plots; (G) Video player of the recording.

```

1 {
2   "recordingID": "11H57M49S371",
3   "applicationName": "Myo",
4   "frames": [
5     {
6       "frameStamp": "00:00:00.0035019",
7       "frameAttributes": {
8         "orientation_X": "-0.6087646",
9         "accelerometer_Y": "-0.3989258",
10        "gyroscope_Z": "2.9375",
11        "EMGpod_1": "-2",
12        "EMGpod_2": "-8",
13        "EMGpod_3": "-54",
14      }
15    }, ... {

```

Listing 1: Example of one of the JSON file contained in the session file for the Myo application.

3.2.2 Sensor attributes. Sensor attributes are relevant components of multimodal data since they constitute the facets through which the sensor measures a phenomenon, for example, the electromyogram data of the learner. The sensor attributes are specified in each application file contained in the session file. When the sensor file is loaded in, the sensor attributes are hierarchically sorted in the Attribute list (Figure 2, B, bottom-left). Each sensor application (e.g. "Kinect") is one top-node on the list. The parser of the VIT can nest the sensor attributes using underscores. For example, if the parser reads an attribute named "3_Hand_Right_Tip_X", it will nest the attribute in the following hierarchy 3 > Hand > Right > Tip > X. The sensor attributes can be either numerical time-series or categorical series. Multiple sensor attributes can be selected with a check box on the Attribute list, this action will update the chart plotting all the selected attributes (Figure 2, F, centre-right).

3.2.3 Time intervals. Time intervals allows partitioning the sensor recording into shorter segments that have different purpose and meaning. The user can add a new time interval by selecting a time region under the plot F. Each time interval is represented as light yellow rectangles in the plot, which as a segment of the x-axis (time dimension). The time interval is defined by their *start-time* and *end-time*, which is the relative duration in seconds from the start of the session. All the added time intervals are listed in panel D. The user can click on one or more time intervals on this list and highlight visually the time intervals in the plot which will turn from light yellow to blue. The time intervals are added independently by the sensor attributes. The sensor attributes, however, vary within each time intervals and their value change is important for the annotations. Time intervals can also overlap.

3.2.4 Annotations. The intervals define portions of the sensor recording with a specific meaning which can be assigned through custom properties. We call these properties annotations. In the case of our implementation, the annotations are provided in the time domain rather than the spatial or the frequency domain. The user provides an annotation that is valid for a time interval, i.e. between two time points of the multimodal dataset. In contrast, the user cannot provide an annotation to specific regions of the video or to certain frequencies of the sensor attributes. Each time interval can be annotated with *0-to-n* annotations. The annotations can be both loaded via the Annotation file (Figure 2, C, top-centre) or can be defined manually (Figure 2, E, centre). All the time intervals share the same set of annotations, this leads to an *Attribute-Value* representation in which the annotated dataset can be seen as a table in which each time interval is a row and each annotation constitutes

a column. The value not assigned will be treated as a null value. As time intervals can overlap also annotations can also do the same. This allows for annotations at different levels of granularity, and for example, break down activities into smaller sub-activities or atomic actions.

3.2.5 Annotation procedure. The annotation procedure consists of assigning all the time intervals to the right annotation values. To do so, the user can be helped by the video (if available) in identifying the right affiliation between time intervals and annotations. At the completion of the annotation procedure, the user can click “export” which enables three different follow-up actions for the user: download the single annotation file in JSON format. The same file format is accepted by the Load annotation panel (Figure 2, C, center-top); download the session file in zip format embedding the annotation file. The file will be renamed “*recordingID_annotated.zip*” add the annotations in the Attribute list (Figure 2, B, bottom-left) to be able to plot them along with the sensor attributes. This feature works best with numerical data and can be useful to visualise the variations of annotation values.

3.3 Data processing

3.3.1 Data transformation. As output, the VIT produces an annotated dataset “*recordingID_annotated.zip*” which can be downloaded and is ready to be transformed. We have implemented this routine in Python outside of the VIT, since, while the transformation requirements might change, the annotated dataset remains unchanged. In Figure 3, we provide a graphical representation of this transformation. The routine takes all the time intervals ($t_1 \dots t_n$) which are specified in the annotation file and make them as one row in the tabular representation. Each of this row has corresponding annotations ($y_1, \dots, y_n$). Consequently, we use the time intervals as time window on the sensor application files, each of them containing a different set of attributes. This windowed selection gives as output multiple smaller time series, one for each attribute contained in each application file. In each time series, we then apply feature selection, to extract numerical features to transform the variable length of the smaller time series. A great number of features can be extracted, spanning from the most classic time-domain features (e.g. min, max, st. deviation, variance, mean) to frequency domain features (eg. no. peaks, Fast Fourier transform). In addition, there are signal specific features, which vary depending on whether the time series describe a physical movement or a physiological response. The appropriateness of the features must be found in the literature dealing with those particular signals. At the end of this process the end result is a table with i rows (number of intervals) and $(p * w * n + m)$ columns, where p is the number of attributes for the sensor application files, w is the number of aggregating function ϕ , n is the number of applications and m the target values contained in the annotation file.

3.3.2 Model creation. The techniques for processing multimodal data highly depend on the goal of the system. Here we list a set of techniques to derive models from the transformed data.

1) Supervised machine learning models are computed *a posteriori* using the labels collected with the annotation as target attributes for prediction or classification. VIT has been designed

Figure 3: Attribute-value transformation of the annotated dataset.

particularly for using these models. The annotation allows distinguishing optimal from non-optimal learning moments, e.g. if the learning performance meets the learning goals or if the learner commits an error.

2) Unsupervised models do not use labels as the reference system for the optimal learning trajectory, but cluster similar learning moments together. This analytical approach allows discovering patterns of behaviour that are close to each other. Unsupervised techniques can be used in case no annotation scheme can be found. In that case, it could still be useful to use the VIT to segment the learning experience into meaningful parts.

3) Rule-based models, in this case the sensor thresholds for the ideal performance are presumed *a priori*. This approach is also known as constraint-based models (CBMs). CBMs use constraints in the form of tuples $\langle C_r, C_s \rangle$, where C_r is the relevant condition and C_s is the satisfaction condition. When C_r applies, then C_s is tested, if not satisfied then a feedback is triggered [17]. With this approach, VIT does not have a central role and can be bypassed.

4) Probabilistic models such as Bayesian Networks or Structural Equation Modelling. These models also require the VIT and the annotation procedure. Differently from the discriminative models such as supervised or unsupervised learning, these models learn the conditional distribution which can be used for diagnostic or for informing about the relations between different factors that influence the target attributes.

5) Expert-learner models can compare two or more performances, e.g. learner against the expert’s performance to search for the presence of erratic behaviour. This approach requires a pairwise comparison between the expert’s performance and trainee’s performance. It uses similarity algorithms, such as Dynamic Time Warping considering that sessions might have different execution times. The VIT can be used to segment the multimodal recordings in relevant activities. However, it can become complex when multimodal recordings have hundreds of attributes.

3.4 Data exploitation

The data-driven system can have multiple benefits for learning such, for instance: A) providing immediate feedback for correcting erroneous behaviour; B) tailor a more personalised learning experience; C) find

Data exploitation strategy	Data processing technique	Feedback medium	Model	Use of the VIT	Actionability	Interpretability
A) Corrective non-adaptive feedback	3) Rule-based models	ITS	Presumed	No	High	High
B) Predictive adaptive feedback	1) Supervised ML models	ITS	Computed	Intervals, annotations	High	Low
C) Patterns identification	2) Unsupervised ML models	Dashboard	Computed	Intervals	Low	Medium
D) Historical reports	Descriptive statistics	Dashboard	No model	No	Low	Low
E) Diagnostic analysis of factors	4) Probabilistic models	Dashboard	Computed	Intervals, annotations	Low	High
F) Comparison	5) Expert-learner comparison	Dashboard	Presumed	Intervals	Low	Medium

Table 2: Mapping the data processing techniques to exploitation strategies for multimodal data.

patterns in the course of action; D) provide an in-depth overview of the learning process; E) diagnose factors and triggers for learning; F) compare the learner with the expert. All these forms of feedback can be delivered differently, for example, through an ITS or via a dashboard. In Table 2, we map the strategies with the data processing techniques discussed in previous section 3.3.2. We classify these strategies based on whether the model is presumed a priori or computed a posteriori; if the output is actionable for the learner (i.e. suggest the learner what exactly to do or improve) and if the model is interpretable (i.e. explains how the model was created).

3.5 Technical use cases

To test the VIT in authentic learning scenarios we have used it with three different ITSs designed for three different psychomotor learning scenarios (Figure 4) 1) training how to present in public with Presentation Trainer; 2) cardiopulmonary resuscitation training with the CPR Tutor; 3) calligraphy training with the Calligraphy Tutor. For each of these three scenarios, the authors developed an ITS which uses the LearningHub in the backhand for the data collection and synchronisation. This allowed to collect session files with a structure that can be loaded into the VIT.

3.5.1 Presentation trainer. the Presentation Trainer (PT) [31] is a research prototype designed to support the development of nonverbal communication skills for public speaking (Figure 4, top). The multimodal data used are audio and skeleton via Microsoft Kinect. The feedback implemented in the PT is rule-based corrective. The model of ideal presentation has been designed a priori by interviewing presentation experts. The performance of the learner is compared against such model. PT is only able to detect gestures, not able to differentiate among types of gestures such as iconic gestures, gestures bigger than usual, deliberate gestures, unconscious gesture. The role of the VIT in the data generated by PT is to identify different types of gestures and enhance the feedback of the PT.

3.5.2 Multimodal Tutor for CPR. The Multimodal Tutor for CPR [8] was designed for training people to perform cardiopulmonary resuscitation using patient manikins (Figure 4, centre). The tutor uses a multi-sensor setup for tracking the CPR execution and generating personalised feedback. The feedback is based on the

Figure 4: Use cases for the VIT: 1) Presentation trainer, 2) CPR Tutor, 3) Calligraphy Tutor

multimodal data it predicts when the CPR correct posture for chest compression is violated. It uses supervised machine learning models to generate these predictions. The multimodal data considered are trainee's body position (with Microsoft Kinect), electromyogram (with Myo armband) and compression rates data derived from the manikin. The VIT was used in the CPR Tutor for annotating the single chest compressions, if and when errors in the from the CPR

correct posture occurred. The following binary target attributes were annotated: *compressionRate* (speed of the compression), *compressionDepth* (depth of the compression), *compressionRelease* (hands correctly released from the manikin), *armsLocked* (whether arms were kept locked), *bodyWeight* (if the whole body weight was used when compressing).

3.5.3 Calligraphy Tutor. the Calligraphy Tutor [20] application runs on Microsoft Surface tablet and is designed to record the fine motor movements of calligraphy experts (Figure 4, bottom). The multimodal data used are pen strokes with Microsoft Surface pen, Myo electromyogram data. The feedback provided guides the calligraphy trainees guides to learn calligraphy for the first time. The VIT was used for annotating letters. Letters consisting of different pen strokes were grouped into different intervals, and their execution was evaluated by a calligraphy expert.

4 DISCUSSION

We answered RQ1 by listing six functional requirements (section 3.1) and by describing how the components of the VIT fulfil these requirements (section 3.1). The testing of VIT in the three technical use cases (section 3.5) shows that VIT in combination with the LearningHub are tools that can be applied in a variety of learning tasks. There are, however, some limitations: the scenarios used as technical use cases are characterised by learning experiences conducted by one learner individually and the learner executes practical tasks which are evident for the sensors. We have not tried VIT for collaborative settings or for learning activities that require cognitive tasks which are difficult to capture with sensors, these areas require future work. To answer RQ2 we show in section 3.2 how to transform the output of the VIT in tabular representation (attribute-value) having the annotated time intervals as individual learning samples. This transformation provides the researcher a dataset ready for being further analysed with a variety of techniques, some of these summarised in section 3.3.2. Finally, we addressed RQ3 providing a set of use cases where the VIT can be applied and support specific learning tasks (section 3.5). The use cases are also enriched with list of different strategies of exploitation for multimodal data for learning. We consider that the answer to RQ3 is highly relevant for the MMLA community as it contributes to close the feedback loop that begins with multimodal data collection and ends to the learner. As we evidenced in review of existing tools in section 2.4, the data exploitation, is the part of MMLA that requires more research effort by the community. We also argue that the contribution of the VIT, together with LearningHub and other technical tools for MMLA from our review can be part of integrated workflow, which can work as a toolkit for MMLA researchers to quickly set up their experiment without having to “reinvent the wheel” and create each time new systems from scratch. We call this the *Multimodal Learning Analytics Pipeline*. The MMLA Pipeline is composed by five steps (as shown in Figure 5), corresponding to the five MMLA challenges: 1) collection, 2) storing, 3) processing, 4) annotation and 5) exploitation of multimodal data. The MMLA Pipeline is a workflow for researchers which allows multimodal tracking of learning activities using wearable sensors, IoT devices, audio and video recordings. There could be multiple routes in the MMLA Pipeline: in Figure 5, for example, we show the

Figure 5: Graphical representation of the Multimodal Learning Analytics Pipeline.

first four exploitation strategies that we discussed in section 3.4: A) Corrective feedback, B) Predictions, C) Patterns, D) Historical reports. The first two routes (A, B) can be implemented within an ITS for, respectively, instantaneous corrections and adaptation. The second two (C, D) provide the possibility to orchestrate the learning activity by detecting the frequency of patterns or to raise awareness on the historical development of the modalities. The use cases we tested in section 3.5 can be positioned into the MMLA Pipeline. The Presentation Trainer would take route A), as feedback is directly processed in the LearningHub and it requires no storing, annotation, exploitation. The Tutor for CPR will take the route B), as the VIT is used and therefore prediction model are trained. The Calligraphy Tutor will take route D or C, depending on whether the VIT used for annotating the letters.

5 CONCLUSIONS

In this paper, we introduced the Visual Inspection Tool, which addresses the data annotation challenge and facilitates the data processing and exploitation. In contrast to most of the existing MMLA architectures and tools tailored-made for specific learning tasks and sensors, the VIT allows to address data annotation in a generic way, for any type of psychomotor learning task that can be captured with a customisable set of sensors. The VIT enables the user 1) to triangulate multimodal data with video recordings; 2) to segment the multimodal data into time intervals and to add annotations to the time intervals; 3) to download the annotated dataset and use the annotations as labels for machine learning predictions. The flexibility introduced by VIT terms of dataset accepted, makes it with the LearnigHub a structured research solution solving different challenges introduced MMLA. We called this solution Multimodal Learning Analytics Pipeline. The MMLA Pipeline is a new integrated workflow that works as toolkit for supporting MMLA researchers to set up new experiments in a variety of learning scenarios. Using components from this toolkit allows to reduce developing time to set up experiments and facilitates the transfer of knowledge. In the future, we plan to continue working on the MMLA Pipeline, to further improve this structured approach for MMLA, which is able to bring the benefits of the multimodal data, machine learning and data analysis also into the learning settings beyond mouse and keyboard happening in the physical space.

REFERENCES

- [1] Omar Alzoubi, Sidney K. D'Mello, and Rafael A. Calvo. 2012. Detecting naturalistic expressions of nonbasic affect using physiological signals. *IEEE Transactions on Affective Computing* 3, 3 (2012), 298–310. <https://doi.org/10.1109/T-AFFC.2012.4arXiv:arXiv:1011.1669v3>
- [2] Muhammad Bilal Amin, Oresti Banos, Wajahat Ali Khan, Hafiz Syed Muhammad Bilal, Jinhyuk Gong, Dinh Mao Bui, Soung Ho Cho, Shujaat Hussain, Taqdir Ali, Usman Akhtar, Tae Choong Chung, and Sungyoung Lee. 2016. On curating multimodal sensory data for health and wellness platforms. *Sensors (Switzerland)* 16, 7 (2016), 1–27. <https://doi.org/10.3390/s16070980>
- [3] Ivon Arroyo, David G. Cooper, Winslow Burleson, Beverly Park Woolf, Kasia Muldner, and Robert Christopherson. 2009. Emotion sensors go to school. *Frontiers in Artificial Intelligence and Applications* 200, 1 (2009), 17–24. <https://doi.org/10.3233/978-1-60750-028-5-17>
- [4] Kiavash Bahreini, Rob Nadolski, and Wim Westera. 2015. Improved multimodal emotion recognition for better game-based learning. *Lecture Notes in Artificial Intelligence* 9221 (2015), 107–120. https://doi.org/10.1007/978-3-319-22960-7_11
- [5] Tobias Baur, Ionut Damian, Florian Lingens, Johannes Wagner, and Elisabeth André. 2013. NovA: Automated analysis of nonverbal signals in social interactions. In *Lecture Notes in Artificial Intelligence*. Vol. 8212 LNCS. Springer, Cham, Cham, Switzerland, 160–171. https://doi.org/10.1007/978-3-319-02714-2_14
- [6] Paulo Blikstein. 2013. Multimodal learning analytics. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge - LAK '13*. ACM, New York, USA, 102–106. <https://doi.org/10.1145/2460296.2460316>
- [7] Nigel Bosch, Huili Chen, Ryan Baker, Valerie Shute, and Sidney D'Mello. 2015. Multimodal Affect Detection in the Wild. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction - ICM '15*. ACM, New York, USA, 645–649. <https://doi.org/10.1145/2818346.2823316>
- [8] Daniele Di Mitri. 2018. Multimodal tutor for CPR. In *Lecture Notes in Artificial Intelligence*, Vol. 10948 LNAI. Springer International Publishing, Cham, Switzerland, 513–516. https://doi.org/10.1007/978-3-319-93846-2_96
- [9] Daniele Di Mitri, Jan Schneider, Marcus Specht, and Hendrik Drachslers. 2018. From signals to knowledge: A conceptual model for multimodal learning analytics. *Journal of Computer Assisted Learning* 34, 4 (aug 2018), 338–349. <https://doi.org/10.1111/jcal.12288>
- [10] Daniele Di Mitri, Jan Schneider, Marcus Specht, and Hendrik Drachslers. 2018. The Big Five: Addressing Recurrent Multimodal Learning Data Challenges. In *Proceedings of the Second Multimodal Learning Analytics Across (Physical and Digital) Spaces (CrossMMLA)*, Martinez-Maldonado Roberto (Ed.). CEUR Workshop Proceedings, Aachen, 6. <http://ceur-ws.org/Vol-2163/paper6.pdf>
- [11] Pierre Dillenbourg. 2016. The Evolution of Research on Digital Education. *International Journal of Artificial Intelligence in Education* 26, 2 (2016), 544–560. <https://doi.org/10.1007/s40593-016-0106-z>
- [12] Sidney D'mello, Andrew Olney, Nathan Blanchard, Xiaoyi Sun, Brooke Ward, Borhan Samei, and Sean Kelly. 2015. Multimodal Capture of Teacher-Student Interactions for Automated Dialogic Analysis in Live Classrooms. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, New York, USA, 557–566. <https://doi.org/10.1145/2818346.2830602>
- [13] Adam Fouse, Nadir Weibel, Edwin Hutchins, and James D. Hollan. 2011. ChronoViz : A system for supporting navigation of time-coded data. In *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems - CHI EA '11*. ACM, New York, NY, USA, 299. <https://doi.org/10.1145/1979742.1979706>
- [14] Joseph F Grafsgaard. 2014. Multimodal Analysis and Modeling of Nonverbal Behaviors during Tutoring. In *International Conference on Multimodal Interaction*. ACM, New York, USA, 404–408. <https://doi.org/10.1145/2663204.2667611>
- [15] Wolfgang Greller and Hendrik Drachslers. 2012. Translating learning into numbers: A generic framework for learning analytics. *Educational Technology and Society* 15, 3 (2012), 42–57.
- [16] Md Sazzad Hussain, Hamed Monkareisi, and Rafael A. Calvo. 2012. Categorical vs. dimensional representations in multimodal affect detection during learning. *Lecture Notes in Artificial Intelligence* 7315 LNCS (2012), 78–83. https://doi.org/10.1007/978-3-642-30950-2_11
- [17] Viswanathan Kodaganallur and Rob R. Weitz. 2005. A comparison of model-tracing and constraint-based intelligent tutoring paradigms. *International Journal of Artificial Intelligence in Education* 15, 2 (2005), 117–144. <http://dl.acm.org/citation.cfm?id=1434928>
- [18] C. Kothe, M. Grivich, C. Brunner, and D. Medine. 2018. Lab Streaming Layer. (2018). <https://github.com/scen/labstreaminglayer>
- [19] D. Lahat, T. Adali, and C. Juttner. 2015. Multimodal Data Fusion: An Overview of Methods, Challenges, and Prospects. *Proc. IEEE* 103, 9 (2015), 1449–1477. <https://doi.org/10.1109/JPROC.2015.2460697>
- [20] Bibek Limbu, Jan Schneider, Roland Klemke, and Marcus Specht. 2018. Augmentation of practice with expert performance data : Presenting a calligraphy use case . In *3rd International Conference on Smart Learning Ecosystem and Regional Development - The interplay of data, technology, place and people*. Springer International Publishing, Cham, Switzerland, 1–13.
- [21] Bibek Hang Limbu, Halszka Jarodzka, Roland Klemke, and Marcus Specht. 2018. Using sensors and augmented reality to train apprentices using recorded expert performance: A systematic literature review. *Educational Research Review* 25, June 2017 (2018), 1–22. <https://doi.org/10.1016/j.edurev.2018.07.001>
- [22] Roberto Martinez, Anthony Collins, Judy Kay, and Kalina Yacef. 2011. Who did what? Who said that?. In *Proceedings of the ACM International Conference on Interactive Tabletops and Surfaces - ITS '11*. ACM Press, New York, USA, 172. <https://doi.org/10.1145/2076354.2076387>
- [23] Oscar Mayor, Quim Llimona, Marco Marchini, Panos Papiotis, and Esteban Maestre. 2013. repoVizz: A Framework for Remote Storage, Browsing, Annotation, and Exchange of Multi-modal Data. In *Proceedings of the 21st ACM international conference on Multimedia*. ACM Press, New York, USA, 415–416. <https://doi.org/10.1145/2502081.2502247>
- [24] Xavier Ochoa, Katherine Chiluiza, Gonzalo Méndez, Gonzalo Luzardo, Bruno Guaman, and James Castells. 2013. Expertise estimation based on simple multimodal features. In *Proceedings of the 15th ACM International Conference on Multimodal Interaction (ICMI '13)*. ACM Press, New York, USA, 583–590. <https://doi.org/10.1145/2522848.2533789>
- [25] H. J. Pijera-Díaz, H. Drachslers, P. A. Kirschner, and S. Järvelä. 2018. Profiling sympathetic arousal in a physics course: How active are students? *Journal of Computer Assisted Learning* 34, 4 (may 2018), 397–408. <https://doi.org/10.1111/jcal.12271>
- [26] Martha C Polson, J Jeffrey Richardson, and Elliot Soloway. 1988. *Foundations of intelligent tutoring systems*. Erlbaum Associates Inc., Hillsdale, NJ, USA, 293 pages. <http://search.ebscohost.com/login.aspx?direct=true>
- [27] Sambit Praharaj, Maren Scheffel, Hendrik Drachslers, and Marcus Specht. 2018. Multimodal Analytics for Real-time Feedback in Co-located Collaboration. *Ec-Tel* 1 (2018), 187–201. https://doi.org/10.1007/978-3-319-98572-5_15
- [28] Luis P. Prieto, Kshitij Sharma, Pierre Dillenbourg, and María Jesús. 2016. Teaching analytics. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge - LAK '16*. ACM, New York, USA, 148–157. <https://doi.org/10.1145/2883851.2883927>
- [29] Mirko Raca and Pierre Dillenbourg. 2014. Holistic Analysis of the Classroom. *Proceedings of the 2014 ACM workshop on Multimodal Learning Analytics Workshop and Grand Challenge - MLA '14* 740, Cv (2014), 13–20. <https://doi.org/10.1145/2666633.2666636> arXiv:ACM 978-1-4503-0488-7/14/11 ...\$15.00.
- [30] Jan Schneider, Dirk Börner, Peter van Rosmalen, and Marcus Specht. 2015. Augmenting the Senses: A Review on Sensor-Based Learning Support. *Sensors* 15, 2 (2015), 4097–4133. <https://doi.org/10.3390/s150204097>
- [31] Jan Schneider, Dirk Börner, Peter van Rosmalen, and Marcus Specht. 2015. Presentation Trainer, your Public Speaking Multimodal Coach. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction - ICM '15*. ACM, New York, USA, 539–546. <https://doi.org/10.1145/2818346.2830603>
- [32] Jan Schneider, Daniele Di Mitri, Bibek Limbu, and Hendrik Drachslers. 2018. Multimodal Learning Hub: A Tool for Capturing Customizable Multimodal Learning Experiences. In *Lecture Notes in Artificial Intelligence*, Vol. 11082 LNCS. Springer, Cham, Switzerland, 45–58. https://doi.org/10.1007/978-3-319-98572-5_4
- [33] Shashi Kant Shankar, Luis P. Prieto, Maria Jesus Rodriguez-Triana, and Adolfo Ruiz-Calleja. 2018. A review of multimodal learning analytics architectures. (2018), 212–214 pages. <https://doi.org/10.1109/ICALT.2018.00057>
- [34] Robert A Sottilare, Keith W Brawner, Benjamin S Goldberg, and Heather K Holden. 2012. Adaptive Tutoring & the Generalized Intelligent Framework for Tutoring. (2012). <https://doi.org/10.13140/2.1.3749.4569>
- [35] Bernardo Tabuenca, Marco Kalz, Hendrik Drachslers, and Marcus Specht. 2015. Time will tell: The role of mobile learning analytics in self-regulated learning. *Computers & Education* 89 (2015), 53–74. <https://doi.org/10.1016/j.compedu.2015.08.004>
- [36] Ipke Wachsmuth, Manuela Lenzen, and Günther Knoblich. 2012. *Embodied Communication in Humans and Machines*. Oxford University Press, Oxford, UK, 1–496 pages. <https://doi.org/10.1093/acprof:oso/9780199231751.001.0001>
- [37] Johannes Wagner, Florian Lingens, and Elisabeth André. 2011. The social signal interpretation framework (SSI) for real time signal processing and recognition. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. International Speech Communication Association, Makuhari, Chiba, JP, 3245–3248. <https://doi.org/10.1145/2502081.2502223>
- [38] Marcelo Worsley. 2014. Multimodal Learning Analytics as a Tool for Bridging Learning Theory and Complex Learning Behaviors. In *Proceedings of the 2014 ACM workshop on Multimodal Learning Analytics Workshop and Grand Challenge - MLA '14*. ACM, New York, USA, 1–4. <https://doi.org/10.1145/2666633.2666634>
- [39] Marcelo Worsley. 2018. Multimodal learning analytics' past, present, and, potential futures. In *CEUR Workshop Proceedings*, Vol. 2163. CEUR Workshop Proceedings, Aachen, Germany, 1–16. <http://crossmmla.org/wp-content/uploads/2018/02/CrossMMLA2018>
- [40] Marcelo Worsley and Paulo Blikstein. 2018. A Multimodal Analysis of Making. *International Journal of Artificial Intelligence in Education* 28, 3 (2018), 385–419. <https://doi.org/10.1007/s40593-017-0160-1>