Multimodal Pipeline: A generic approach for handling multimodal data for supporting learning.

Daniele Di Mitri[1*], **Jan Schneider**[2], **Marcus Specht**[1,3] 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]DIPF – Leibniz Institute for Research and Information in Education – Rostocker Straße 6, 60323 Frankfurt am Main, Germany
[3]Delft University of Technology – Mekelweg 5, 2628 CD Delft
daniele.dimitri@ou.nl, schneider.jan@dipf.de

Abstract

In this demo paper, we introduce the Multimodal Pipeline, a prototypical approach for the collection, storing, annotation, processing and exploitation of multimodal data for supporting learning. At the current stage of development, the Multimodal Pipeline consists of two relevant prototypes: 1) Multimodal Learning Hub for the collection and storing of sensor data from multiple applications and 2) the Visual Inspection Tool for visualisation and annotation of the recorded sessions. The Multimodal Pipeline is designed to be a flexible system useful for supporting psychomotor skills in a variety of learning scenarios such as presentation skills, medical simulation with patient manikins or calligraphy learning. The Multimodal Pipeline can be configured to serve different support strategies, including detecting mistakes and prompting live feedback in an intelligent tutoring system or stimulating self-reflection through a learning analytics dashboard.

1 Introduction

The diffusion of wearable fitness trackers, sensor-rich smartphones, mixed reality headsets, cameras and Internet of Things devices is introducing new technological affordances that can be leveraged in the field of education and learning. Educational researchers are increasingly embedding multisensor and multimodal interfaces and approaches to track learner's behaviour in authentic learning contexts. These new technologies allow moving beyond the typical human-computer interaction, where the user sits in front of a computer, and move towards more immersive and multimodal interactive experiences across spaces through the manipulation of physical and digital objects and environments. This paradigm shift allows a more careful investigation of 'psychomotor' learning activities, i.e. those practical skills that require fine coordination between body and mind. In learning science, learning analytics and human computer interaction, we are witnessing a drastic increase in the use of multisensor interfaces and multimodal data sources [Oviatt *et al.*, 2018]. Nevertheless, in these fields of research, the technological solutions chosen to gather multimodal data opted primarily for tailor-made and ad-hoc solutions. Researchers are still required to take many architectural decisions to collect their data set to reach a stage they can collect their datasets to do their investigation. To change this idea, we present our scientific contribution: the *Multimodal Pipeline*, a generic approach for systematically collect, store, annotate, process and exploit multimodal data in a learning scenario. The Multimodal Pipeline enables researchers to design their experiment and quickly obtain synchronised multimodal datasets so that they can focus on the data analysis. The Multimodal Pipeline proposes a technological solution to the different steps.

2 Technological advantages

Learning activities vary by a significant number of factors. For instance, they can take place inside or outside the classroom, they can be individualised or collaborative, more or less structured. Aiming at creating a system which can support all different combination is an ambitious task. For this reason, we restrict the number of options and better frame the contribution of the Multimodal Pipeline. We use the notion of *Meaningful Learning Task* (MLT), which is an instance of a learning activity with a clear 'start' and 'end'. In this time, we define the interval in which the sensor data have to be gathered. We focus on individual psychomotor learning activities, with a maximum of 15 minutes per recording. In the MLT session, the learning activity is recorded through one-to-n sensors having corresponding sensor applications. The learning activity needs to be structured and sequential: it should be possible in one session to identify sequences of smaller steps which can be assessed individually. The assessment or annotation scheme defines the 'goodness' of the learning performance and is highly dependent on the learning activity investigated. It is preferable that the learning task is repetitive, so that it is possible, within one session, to get multiple examples of the same action or movement (e.g. CPR procedure).

*Corresponding author

3 Current prototypes

At the current stage, Multimodal Pipeline consists of two main prototypes: 1) the Multimodal Learning Hub and 2) the Visual Inspection Tool.

3.1 The Multimodal Learning Hub

The LearningHub [Schneider *et al.*, 2018] is a research prototype which allows controlling multiple sensor applications. The user can specify one-to-n applications running either in the local machine or in the local computer network. Hence, the user can 'start' and then 'stop' the sensor recording for all the selected applications. Each sensor application will record the data from its connected devices and, once the recording is stopped, it will return a JSON file to the LearningHub with all the sensor updates. Since the LearningHub is activating each application, it can communicate the precise timestamp to all the sensor applications, which allows obtaining the sensor data synchronised with them same clock. In addition to the JSON files, also audio and video can be recorded. All these files will be compressed into a zipped folder: the MLT session. The LearningHub is developed in C# for Windows and released under Open Source[1]. At this moment, there exist a variety of sensor applications library already connected for many existing commercial sensors (Kinect, Myo, Leap, Empatica, Android, etc.). The LearningHub is programmed that is relatively easy integrating a new sensor application.

3.2 Visual Inspection Tool

The recorded MLT sessions can be loaded into the Visual Inspection Tool (VIT) [Di Mitri *et al.*, 2019]. VIT allows the manual and semi-automatic annotation of MLT sessions enabling the researcher to 1) 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 classification or prediction. The annotations created with the VIT are saved into MLT dataformat as the other sensor files. The annotations are treated as an additional sensor application, where each frame is a time interval with relative 'startTime' and 'stopTime' instead of that a single timestamp. Using the standard MLT dataformat, the user of the VIT can both define custom annotation schemes or load existing annotation files. Also, the VIT is released with Open Source license[2].

4 Practical use cases

The Multimodal can be used for different purposes. In case of the structured task, the Multimodal Pipeline can be used in conjunction with an Intelligent Tutoring System to detect learning mistakes which can be the base for instantaneous and actionable feedback. In alternative, in the case of less-structured tasks, the data collected with the Multimodal Pipeline can also be summarised into learning analytics dashboards to stimulate reflection from the learner or the teacher. In the following sections, we report three practical use cases in which the multimodal pipeline was used in conjunction with an ITS.

[1]https://github.com/janschneiderou/LearningHub

[2]https://github.com/dimstudio/visual-inspection-tool

4.1 Cardiopulmonary Resuscitation training

We employed the Multimodal Pipeline in the pilot study for the Multimodal Tutor for CPR [Di Mitri, 2018] (figure 1). CPR is a highly standardised procedure consisting of repetitive movements. The multi-sensor setup consisted of Microsoft Kinect and Myo armband. In the pilot study, we involved 11 experts and we tracked their body position. We validated the collected data against the performance metrics derived by the ResusciAnne manikin. We also used VIT to annotate additional mistakes currently not tracked by the manikin such as correct locking of the arms and correct use of the body weight. After that, we trained multiple recurrent neural networks each of them achieving a classification accuracy of the performance indicators above 70%. With the trained models, we can implement automatic corrective feedback while the trainee is doing CPR.

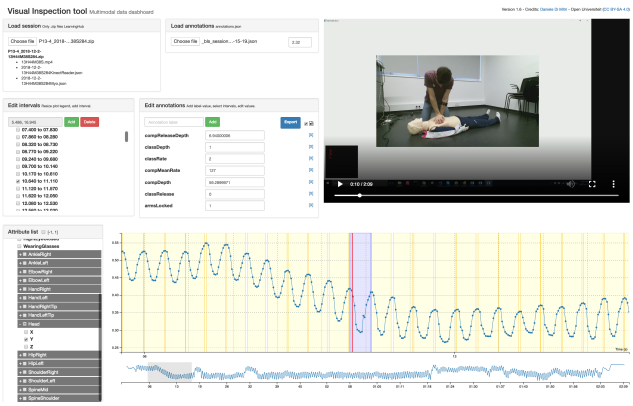

Figure 1: Screenshot of the VIT in the CPR use case

4.2 Learning a Foreign Alphabet

The second use case considered was learning how to write in a foreign alphabet using the Calligraphy Tutor [Limbu *et al.*, 2018] (figure 2). This tutor allows the expert to write a baseline sentence so that the learner can practice reproduce it with feedback by the tutor. The Calligraphy tutor uses Microsoft Surface and its capacitive pen as well as Myo. The coordinates and pressure of the pen were also combined to myogram and gaze information. The authors used these informations to study the features of optimal feedback and correlated it with the user's cognitive load.

4.3 Training public skills

The Multimodal Pipeline was also used with the Presentation Trainer [Schneider *et al.*, 2015] (figure 3) a Kinect-based system which gives real-time feedback on different features of the presentation including posture, pauses, volume and hands position. The learners using the presentation trainer were enthusiastic using it as it allowed to receive feedback from practice their presentations. In the Presentation Trainer, the use of multimodal data is two-fold: it is used both for instantaneous

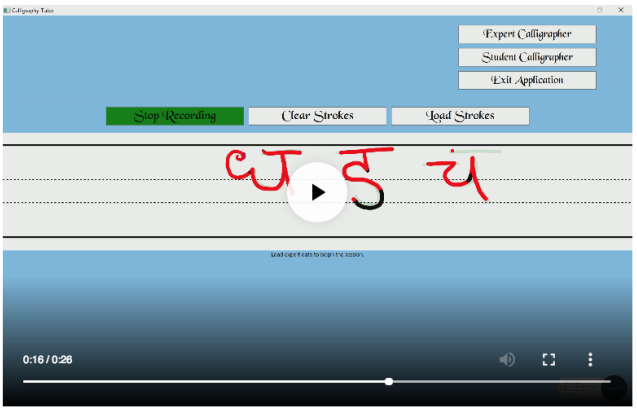

Figure 2: Screenshot of the Calligraphy Tutor

and corrective feedback and, at the end of the session, it is presented in form of visual summary for self-reflecting about the performance.

Figure 3: Screenshot of the Presentation Trainer

5 Future research directions

We plan to progressively improve the Multimodal Pipeline by refining its current components and adding additional ones. We are planning, for instance, to release a data processing called *DataFlow*, which allows to process and run machine learning on the annotated MLT sessions. We are also evaluating the possibility to have a machine learning script for each modality to stack-up modality-dependent classifiers (e.g. one for movements, one for heart rate etc). We are currently working on a runtime feedback engine, the *Multimodal Runtime Framework*, which can channel feedback across sensor applications. In this way there is a centralised interface where the researcher can set-up feedback rules depending on the task and learning design. Necessary also to point out, the Multimodal Pipeline is a research prototype which has been almost exclusively tested in laboratory settings. We do not exclude in the near future to roll it out in authentic class-room or learning environments.

In addition, with the support of the scientific community in Multimodal Learning Analytics, we are planning to develop additional use cases for the Multimodal Pipeline. For instance, one idea is to collect EEG and electrodermal activity to study visual attention in computer games. Another idea is to develop a smartphone application which can be used by students during classrooms and collects kinematic and interaction data. Moreover, we are optimising the Multimodal Pipeline to also work in collaborative learning situations, using for example multiple microphones. This requires a layer of user-identification. With this setup in mind, we are investigating how to extract features from audio and video signals. At this stage, the analysed data are restricted only the sensor data while the videos are used only by the researchers for annotation.

References

[Di Mitri *et al.*, 2019] Daniele Di Mitri, Jan Schneider, Roland Klemke, Marcus Specht, and Hendrik Drachsler. Read Between the Lines: An Annotation Tool for Multimodal Data for Learning. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge - LAK19*, pages 51–60, New York, NY, USA, 2019. ACM.

[Di Mitri, 2018] Daniele Di Mitri. Multimodal tutor for CPR. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 10948 LNAI, pages 513–516, Cham, Switzerland, 2018. Springer International Publishing.

[Limbu *et al.*, 2018] Bibeg Limbu, Jan Schneider, Roland Klemke, and Marcus Specht. 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*, pages 1–13, 2018.

[Oviatt *et al.*, 2018] Sharon. Oviatt, Björn Schuller, Philip R. Cohen, Daniel Sonntag, Gerasimos Potamianos, and Antonio Krüger. *The Handbook of Multimodal-Multisensor Interfaces: Foundations, User Modeling, and Common Modality Combinations - Volume 2*. [s.n.], apr 2018.

[Schneider *et al.*, 2015] Jan Schneider, Dirk Börner, Peter van Rosmalen, and Marcus Specht. Presentation Trainer, your Public Speaking Multimodal Coach. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction - ICMI '15*, pages 539–546, New York, USA, 2015. ACM.

[Schneider *et al.*, 2018] Jan Schneider, Daniele Di Mitri, Bibeg Limbu, and Hendrik Drachsler. Multimodal Learning Hub: A Tool for Capturing Customizable Multimodal Learning Experiences. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11082 LNCS, pages 45–58, Cham, Switzerland, 2018. Springer.