Multimodal Immersive Learning with Artificial Intelligence for Robot and Running application cases

Fernando P. Cardenas-Hernandez[1], Gianluca Romano[1] and Hendrik Drachsler[1]

[1] Information Center for Education, DIPF | Leibniz Institute for Research and Information in Education, 60323 Frankfurt am Main, Germany
cardenas@dipf.de

Abstract. In research different MMLA applications were presented that provide a solution for a particular psychomotor learning task, e.g. CPR or table tennis. A common limitation in all applications is that they are domain specific. In this sense, we present the MILKI-PSY project, whose main goal is to provide a one-for-all system across different domains. Inherently given that different psychomotor learning tasks across domains have certain aspects in common, it would make a one-for-all system possible. Additionally, we present ideas for the MMLA data collection through different sensors and its respective storage, annotation, preparation, and exploitation. The proposed ideas are with respect to two learning tasks: running in the field of sports and collaborative montage in the field of human-robot interaction. Further, we suggest that the system must give to the user the freedom to decide what sensor data to use, and which feedback to receive. Ultimately, we opt for a scalable solution that can be provided to a larger audience.

Keywords: Sensors, Multimodal Interaction, Learning.

1 Introduction

MMLA stands for Multimodal Learning Analytics which refers to the use of multiple sensory inputs at the same time to collect, analyze, and evaluate data related to the users' learning process in order to understand and enhance the learning environment in the field of educational technology. Taking this into account, we must enable the acquisition of input data of our system within an educational environment, thus, the incorporation of sensors is needed. The sensors involved in this system must be able to offer enough information to build a reliable infrastructure which provides guidance during the teaching and learning process of psychomotor skills. By using multimodal information it is expected to obtain an efficient, non-redundant, and highly significant understanding of the incoming sensory data. Normally, the use of multiple and different types of sensors lead to solving the problem of finding an efficient way to synchronize the data and making it compatible for its further analysis and representation within the system. At the end, the main objective is to allow users to be free to choose the psychomotor skill they want to learn in a self-taught way through the identification and analysis of the characteristics these psychomotor activities share in common. To carry out this objective, the MMLA Pipeline approach [1] is taken as a starting point because it offers well-structured methodology which eases the research of multimodal

experiments by designing accurate setups for the improvement of learning activities. It is also worth taking into account previous related work in this area such as the cardiopulmonary resuscitation (CPR) tutor [2], table tennis tutor [3] and the presentation trainer [4] that considered the MMLA Pipeline in their design.

2 Data Collection

The use of sensors is vital for the development of multimodal systems because they allow the data collection. Sensors used in the development of the two application cases are grouped based on their outputs. A description of their main role or application within the system is also done.

For the development of a multimodal system for the robot application case, sensors are highly flexible and efficient to react to the unpredictable human actions that occur in human-robot collaboration tasks [5]. To test the level of acceptance and comfort of the users, we recommend using both physical and contactless sensors for human-robot interactions. Sensors generate inputs to the system and they must be placed on the robot in order to equip it with the ability to collaborate and interact with and similar to humans. Given that the main goal is the learning, the robot's action sequence may be repeated or skipped considering the user's learning performance or feedback.

Vision sensors: These types of sensors include cameras whose images can be processed to extract valuable data to control the robot to avoid collision with the user or objects, and to communicate with the users by their emotion interpretation.

Audition sensors: The robot can be equipped with one or multiple microphones in charge of acquiring the verbal command dictated by the user.

Touch sensors: They allow the interaction between humans and robots by measuring the forces involved in their communication.

For the running application design, the sensors must fulfill the following requirements: they must be waterproof because sweat can damage non well isolated sensors; they must not influence significantly the normal movement of the users, otherwise, the collected learning data could not be reliable; they must have a low data delivery latency; finally, sensors worn by the learner must be easy-wearable and portable.

Vision sensors: Cameras allow to capture videos to obtain information about the learner's performance during the training and to record the environment where the learning process takes place.

Audition sensors: Microphones can register verbal commands and sounds associated with the current physical condition of the runners, for instance, coughing, gasping and exhalation. These sounds may indicate the level of fatigue and discomfort of the user.

Motion sensors: These sensors permit recording the persons' movements by placing them on different parts of the body, e.g., head, chest, waist, wrists, arms, thighs, ankles and toes. Accelerometers and gyroscopes are included in this group.

Physiological sensors: By measuring the body temperature, pulse rate, respiration rate and blood pressure denominated as the human body's most basic functions [6], we can gain precise information of the current physical condition of the runners. This type

of sensor is not used in the collaborative robot interaction case as this case does not involve an extenuating activity that could alter the physical condition of the people.

3 Data Storage

This step of the MMLA Pipeline corresponds to the organization of the diverse incoming multimodal data which normally require high storage capacity due to its big size. The managing and exchanging of the stored information are also parts of this step. Ideally, a data storage system must have an easy access and a minimal latency after read and write accesses, as well as, the availability to keep up with growth and the flexibility and efficiency to handle a wide range of formats coming from different sources [7].

Having these features in mind, hard drive storage and removable storage devices are discarded because they store data locally making the data sharing among users a time consuming task and their storage capacities tend not to be sufficient for systems receiving a large amount of data. Moreover, in case of a hardware malfunction the stored data is susceptible to loss. Network storage and online storage can deal with the drawbacks that local storage devices have.

Network storage: It is a network used to store data in a way by which it can be accessible to a group of devices on the network; it also manages copies of the data across the network as a backup. Normally, network storage technology is classified as *Storage Area Network* and *Network Attached Storage*.

Online storage (also called cloud storage): It permits users to delegate the storage of their data, its management, maintenance and security to online data storage services offered on the internet.

To avoid investing financial and human resources in the maintenance of the storage infrastructure using direct network storage, it is strongly recommended to use cloud storage. As there are plenty of cloud storage providers, it is very important to decide the right one(s) based on their accessibility, costs and support.

4 Data Annotation

Data has to be annotated to draw meaningful insights out of it with e.g. Machine Learning (ML) techniques. In MMLA applications different sensors are used to communicate with the learner, decoding and encoding messages between the physical and digital world [8]. Annotation can be automatically, manually or semi-automatically.

For MMLA applications the work of [9] proposes the Visual Inspection Tool (VIT) which allows data reading which is gathered from different sources, and annotating them. The VIT uses Meaningful Learning Task (MLT) [10] session files to store annotated data for different sensors.

The authors of [9] state that the VIT supports MMLA researchers in (i) triangulating multimodal data with video recordings, (ii) segmenting the multimodal data into time-intervals and adding annotations to them, and (iii) downloading the annotated dataset and using it for multimodal data analysis.

For this project we plan to use the VIT to annotate multimodal data because we expect to collect multimodal data from cameras, depth sensors and other devices. The annotated data files are directly uploaded using cloud storage. For further usage the data can be downloaded to train ML models with.

5 Data Processing

This stage deals with the extraction of the most relevant and representative features from the raw data in order to clean and reduce the amount of data that will be processed, transformed or integrated, in a way that the redundancy of information can be avoided which also leads to a reduction in the overall processing time.

Signal processing plays an important role in this step. For example, the filtering of incoming raw sensor signals can be needed to separate them from other unwanted signals coming from external sources. The filtering can be done via hardware (electrical circuits) or software (computational algorithms) methods. Data cleaning is another preparation technique involving processes like outlier removal or data normalization. Speaking of data transformation, audio signals are frequently evaluated in the frequency domain because it offers a good alternative to obtain more meaningful information. Computer vision (CV) and machine learning (ML) algorithms can solve the need to extract features to classify, merge or integrate data. For instance, principal component analysis (PCA) is an algorithm for feature extraction that reduces the high dimensional data to improve their interpretation without losing relevant information.

6 Data Exploitation

Data is meaningless without further analysis or exploitation. Thus, methods have to be applied to gain meaningful insights on the data for the learners and their experiences. In the MMLA Pipeline, [1] state three different exploitation approaches: predictions, patterns, and historical reports. It is hard to tell which prediction techniques are better than others from the start. Unique properties of the learning tasks need to be considered to see which techniques can be applied.

For example, running forward shows a periodic pattern when repeatedly moving your legs. How does velocity or acceleration affect this pattern? Consequently, sensors are needed that correctly acquire the data.

For the other learning tasks, robots do not only have to understand humans when working together but they also have to apply a sequence of actions in response. Essentially, this is communication. A learner's performance might be captured by the amount of corrective feedback of the robot over time. A robot could have the following modules: a speech-to-text/text-to-speech module to communicate, and a detection module that e.g. uses Neural Networks to detect wrong behavior. As a consequence, there are a lot of tasks the robot has to fulfill for proper communication. Thus, multiple models and sensors are required.

7 One-for-all System

Perhaps, one of the most challenging goals in our proposal for the MILKI-PSY project is the "one-for-all system across different domains" feature because every learning task has its own distinctiveness. Conversely, they may also share some common attributes with other tasks. For instance, hip rotation is involved in a lot of different psychomotor skills like dancing and martial arts. In the collaborative robot and running cases orientation of the natural head position can be used as a common aspect.

However, the way to break down the domain difference relies on finding the similarities to design or implement psychomotor learning tasks. A concise description and exemplification of the tasks, the automatization of the appropriate activities, the definition of the real-time-feedback, the classification and analysis of the tasks based on the expertise levels can help to achieve the necessary abstraction to deal with diverse domains.

The collaborative robot and running cases will function to begin extracting information to create an abstract and common framework. This framework will be gradually tested on two other similar domains (for example, running with a ball and painting) in order to obtain their common framework and to enrich the level of abstraction of the previous one. Subsequently, other two different domains will be used to extract its common framework and added to the previous one in order to update and gain more abstraction of the original framework. This iterative process is repeated until the original framework is robust enough for most of the learning tasks.

The quality of the annotations for all cases can be achieved involving two or more human annotators and computer inter-rater reliability score. Similarly the quality of the processing can be achieved by comparing the new trained model to some baselines, which is either some previously trained models or some standard existing ones.

As it is expected the system keeps growing in data volume or complexity, its scalability will require cloud computing strategies like auto-scaling.

Figure 1 shows the proposed common MMLA Pipeline steps used for the research of a single domain or psychomotor skill and figure 2 displays the suggested iterative steps to reach a common framework for different learning tasks.

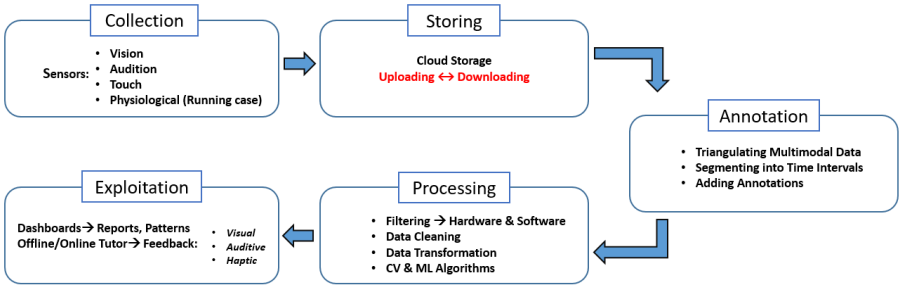

Fig. 1. MMLA Pipeline steps of a single domain.

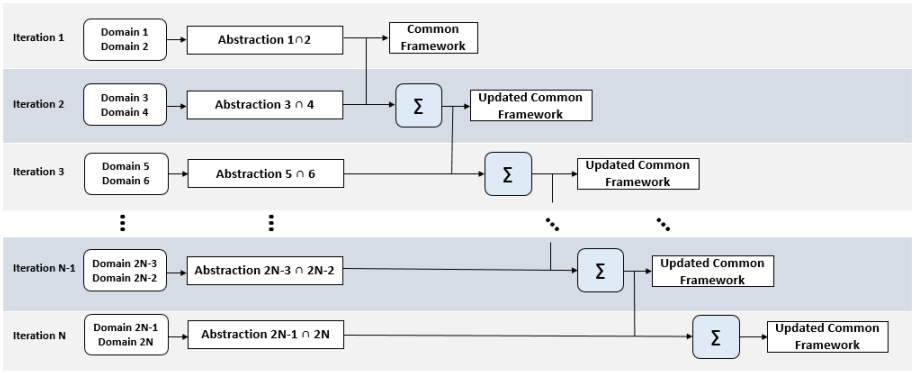

Fig. 2. Iterative steps to find a multiple domain common framework.

8 Workshop Expectations

The workshop can give us the opportunity to exchange ideas with other participants in order to bring solutions to the most challenging tasks and strengthen the interdisciplinary collaboration. Besides, we expect to answer some particular questions as:

- EQ1: What sensors can be used and how many of each type?
- EQ2: How many sensors do users allow (acceptance of the user)?
- EQ3: What sensors are good to evaluate the learning process correctly?
- EQ4: How to support compatibility between different sensor manufactures and software producers on a technological level? For instance, users could just want to plug in their cameras and not care about the involved software.

For EQ1, we plan to let the workshop participants think about possible sensors for our scenario. This will help us in gathering ideas on how to use different sensors that we did not imagine before. For EQ2, we plan to attach “fake” sensors to the participants until the point they do not feel comfortable anymore and experience the attached sensor as invasive and disruptive. For EQ3 we present the audience with one of our scenarios and think together what feedback a user wishes for. Finally, for EQ4, we present an idea of integrating sensors into a system related to different feedback.

References

1. Di Mitri, D., Schneider, J., Klemke, R., Specht, M., Drachsler, H.: Multimodal Pipeline: A generic approach for handling multimodal data for supporting learning. In: First workshop on AI-based Multimodal Analytics for Understanding Human Learning in Real-world Educational Contexts, China (2019).
2. Di Mitri, D., Schneider, J., Trebing, K., Sopka, S., Specht, M., and Drachsler, H.: Real-Time Multimodal Feedback with the CPR Tutor. In: Artificial Intelligence in Education; Bittencourt, I.I., Cukurova, M., Muldner, K., Luckin, R., Millán, E. (eds.) AIED 2020. LNCS (LNAI), vol. 12163, pp. 141–152. Springer, Switzerland (2020).

3. Mat Sanusi K. A., Di Mitri, D., Limbu, B., and Klemke, R.: Table Tennis Tutor: Forehand Strokes Classification Based on Multimodal Data and Neural Networks, Sensors, 21(9): 3121, Switzerland (2021).
4. Schneider, J., Börner, D., Rosmalen, P., and Specht, M.: Presentation Trainer, your Public Speaking Multimodal Coach. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, vol. 17, pp. 539-546, USA (2015).
5. Cherubini, A., and Navarro-Alarcon, D.: Sensor-Based Control for Collaborative Robots: Fundamentals, Challenges, and Opportunities. Frontiers in Neurorobotics, vol. 14, p. 113 (2021).
6. Johns Hopkins Medicine Homepage, https://www.hopkinsmedicine.org/health/conditions-and-diseases/vital-signs-body-temperature-pulse-rate-respiration-rate-blood-pressure, last accessed 2021/07/12.
7. Strohbach, M., Daubert, J., Ravkin, H., and Lischka M.: Big data storage. In: New Horizons for a Data-Driven Economy; Cavanillas J.M., Curry E., Wahlster, W. (eds.), pp. 119–141, 1st edn., Springer, Switzerland (2016).
8. Di Mitri, D., Schneider, J., Specht, M., and Drachsler, H.: From signals to knowledge: A conceptual model for multimodal learning analytics, Journal of Computer Assisted Learning 34(4), 338–349 (2018).
9. Di Mitri, D., Schneider, J., Klemke, R., Specht, M., Drachsler, H.: Read Between the Lines: An Annotation Tool for Multimodal Data for Learning. In: Proceedings of the 9th International Conference on Learning Analytics & Knowledge - LAK19, pp. 51-60. Association for Computing Machinery, USA (2019).
10. Schneider, J., Di Mitri, D., Limbu, B., and Drachsler, H., Multimodal Learning Hub: A Tool for Capturing Customizable Multimodal Learning Experiences. In: Lifelong Technology-Enhanced Learning, Pammer-Schindler, V., Pérez-Sanagustín, M., Drachsler, H., Elferink, R., Scheffel, M., (eds.), pp. 45–58, Springer, Switzerland, (2018).