Inferring Student Engagement in Collaborative Problem Solving from Visual Cues

Angelika Kasparova
King's College London
Department of Engineering
London, United Kingdom
angelika.kasparova@gmail.com

Oya Celiktutan
King's College London
Department of Engineering
London, United Kingdom
oya.celiktutan@kcl.ac.uk

Mutlu Cukurova
University College London
UCL Knowledge Lab
London, United Kingdom
m.cukurova@ucl.ac.uk

ABSTRACT

Automatic analysis of students' collaborative interactions in physical settings is an emerging problem with a wide range of applications in education. However, this problem has been proven to be challenging due to the complex, interdependent and dynamic nature of student interactions in real-world contexts. In this paper, we propose a novel framework for the classification of student engagement in open-ended, face-to-face collaborative problem-solving (CPS) tasks purely from video data. Our framework i) estimates body pose from the recordings of student interactions; ii) combines face recognition with a Bayesian model to identify and track students with a high accuracy; and iii) classifies student engagement leveraging a Team Long Short-Term Memory (Team LSTM) neural network model. This novel approach allows the LSTMs to capture dependencies among individual students in their collaborative interactions. Our results show that the Team LSTM significantly improves the performance as compared to the baseline method that takes individual student trajectories into account independently.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision tasks**; • **Applied computing** → **Interactive learning environments**.

KEYWORDS

Collaborative problem solving; Complex group interactions; Team Long Short-Term Memory networks

ACM Reference Format:
Angelika Kasparova, Oya Celiktutan, and Mutlu Cukurova. 2020. Inferring Student Engagement in Collaborative Problem Solving from Visual Cues. In *Companion Publication of the 2020 International Conference on Multimodal Interaction (ICMI '20 Companion), October 25–29, 2020, Virtual event, Netherlands.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3395035.3425961

1 INTRODUCTION

Today many work environments require problem solving as a team and collaboration in face-to-face contexts. For instance, the success in an operation room depends on the communication dynamics in surgical teams or engineers need to work in teams to solve many workplace problems collaboratively. Therefore, collaborative problem solving (CPS) is an essential skill for young people to acquire during their education. As a pedagogical approach, it can also reinforce knowledge, improves student attainment, and positively changes student's attitudes towards the subject studied [11]. However, positive outcomes of CPS is highly correlated with students' engagement with the CPS activities and students should be appropriately guided in their interactions to achieve the expected learning outcomes [4, 13]. Direct analysis and guidance by a teacher or an expert is desirable, yet scaling this approach to the large number of learners is often not feasible. One promising solution is to develop automatic learning analytics tools that can monitor student engagement and success in CPS activities [27]. These analytics information can be used both to adapt required support and content for students as exemplified in intelligent tutoring systems i.e [20] or to inform teachers for more appropriate and efficient pedagogical interventions [16, 25].

Nevertheless, most existing work in analysing student engagement focuses on digital learning environments and monotonous pedagogical approaches such as lecturing [24]. A few available studies have looked at student collaboration in physical spaces, yet they leverage intrusive data collection approaches such as eye tracking [19] or physiological signals [28] considering that collaboration skills in physical or blended learning settings are extremely important for the future success of young people, there is an urgent need for analysing learner behaviours in collaborative interactions leveraging nonintrusive data sources such as 2D video data only.

This paper focuses on an emerging computer vision problem, namely, the classification of student engagement in open-ended, face-to-face CPS activities using 2D video data only. We propose a novel method, called, Team LSTM model to capture dependencies of individual students' behaviours during their CPS process. Our proposed solution is inspired by the Social LSTM model [2], which was originally proposed for predicting pedestrian trajectories. However, to the best of our knowledge, this is the first implementation of a Team LSTM approach in the learning sciences domain.

2 RELATED WORK

In face-to-face, co-located collaborative learning environments, students communicate and interact with their peers via speech, facial expressions and body gestures, which can be used as indicators of their collaborative behaviours. For instance, Grover et. al [10] focused on predicting the collaboration level/strength (i.e., low, medium, high) during activities of pair programming. Similarly,

ICMI '20 Companion, October 25–29, 2020, Virtual event, Netherlands

ACM ISBN 978-1-4503-8002-7/20/10…$15.00
https://doi.org/10.1145/3395035.3425961

Worsley [29] collected gesture, speech and electro-dermal activation data from pair collaboration. The gesture data was used to learn a set of canonical clusters and the relation between students' clusters and their behaviors showed that gesture data can be used to predict students' CPS behaviours. Moreover, in some previous studies biometric data was also used to analyse students' collaborative learning. Lubold and Pon-Barry [15] conducted a study related to acoustic-prosodic features with rapport in collaborative learning and found that students' pitch may be similar when they collaborated well with each other. Similarly, Dikker et al. [8] used portable electroencephalogram (EGG) to record students' brain activities and showed that, brain-to-brain synchrony can also be a possible indicator of dynamic social interaction and effective collaboration.

At a relatively more mature research level, aiming to generate visualisations from automatically generated metrics of students' collaborative learning in physical spaces, Martinez-Maldonado et al. [17] used multimodal data to record students' learning activities in healthcare simulations. Similar to group's previous work by Echeverria et al. [9], they tracked and visualized how teams of students occupy the space in healthcare simulations. Although, visualisations of how students occupied space in the learning environment were created, the authors argued that teachers need the additional contextual information to have an interpretation of the students' learning process from these visualisations. On the other hand, some researchers focused on individual student data and analysis rather than the group data as presented previously. In order to identify the different performances and behaviours of individual students during collaborative learning, Oviatt et al. have conducted a series of studies [21, 22]. The group mainly explored the differences between expert and novice students, and compared their collaborative learning behaviours. The participants were asked to solve math problems in groups of three and video, audio, and written data was collected. The authors found that expert students performed more fluently in both writing and speaking during the process of collaboration. They also found that expert students had a higher ratio of using non-linguistic symbolic representations and structured diagrams to elemental marks. Similarly, Schneider et al. [26] used eye-tracking, video, and audio data to analyze individual student's learning motivation in pair collaboration. Their results show that using eye-tracking data only is not enough to fully present students' different levels of learning dynamics.

As the reviewed research above shows, multimodal data from physical learning environments can provide promising results to investigate collaborative learning in physical spaces. However, the collection and analyses of multimodal data from real-world classroom environments are challenging and it poses significant methodological, practical, and ethical challenges [5]. On the other hand, video recording is a method which is used frequently to collect data from the classrooms to study student or teacher behaviours. Due to their low financial and technical costs, video-based analytics of collaborative learning can provide valuable opportunities for immediate real-world impact. Although, there is early work investigating the potential of video data to analyze learner behaviors in collaborative learning activities through modelling learner behaviours [7], there is also a large need for developing novel computational approaches. For instance, in [7], CPS behaviours were modelled using a traditional classification approach (decision trees) and a semi-automated pipeline (active, semi-active, passive engagement values were manually coded to model CPS competence). The authors prioritised the transparency of the models over their performance and presented their results as an opportunity for humans to better interpret the models. On the other hand, here, we present a fully automated ML pipeline for the labelling of active, semi-active, and passive engagement behaviours during CPS activities.

3 DATASET

We used the dataset that was introduced by Cukurova et al. [7]. The dataset comprises video recordings of 3 students working collaboratively to produce a smart object by connecting Arduino boards and interacting with an Integrated Development Environment. All the participants were selected by their lecturers in order to obtain a balanced set of computer science abilities and alleviate the bias of existing knowledge and skill differences between students on their collaborative problem-solving performance. Therefore, the data was collected from three sequential educational interventions with 18 unique individual students at a European university (17 males and 1 female, average age 20 years). The students were divided into six groups of three students and some groups worked multiple times with the system and their interaction was recorded using a video recorder as shown in Figure 1. No time bounds were given to students performing their specific tasks. Occasionally, videos also show teachers and other people in background, just sitting or passing by. These people were excluded from the analysis.

From this dataset, we used 14 videos, lasting between 27 and 79 minutes with an average of 56 minutes and resulting in a total of 13 hours of recordings approximately. These videos were segmented into 30 seconds-long clips, totalling 1573 short clips, and were manually labelled by two coders using the Nonverbal Indexes of Students Physical Interactivity (NISPI) framework [6]. The coders annotated the students' behaviours with respect to three levels of engagement: (1) "active" indicates that a student is actively manipulating an object; (2) "semi-active" indicates that a student is not acting on a object but their attention is oriented towards an active peer or an object associated with the learning objective; and (3) "passive" corresponds to any other situation. To achieve reliable labels, the coders were asked to provide labelling for non-overlapping windows of 3 seconds and then a final label for each short clip was obtained by taking average over all windows. Whenever there was disagreement between the coders for a short clip, the coders were asked to revise their coding by looking at windows of 5 seconds. This process resulted in 98% agreement with the ordinal alpha value $k = 0.912$. The decision to use 30 seconds-long clips and 3 seconds-long windows for human annotation was based on the pilot work completed and previous research on interpreting collaborative problem-solving (CPS) competence from physical interactions. As shown in [6], 30 seconds-long clips annotation is granular enough to generate meaningful distinctions between low, medium, and high competence CPS groups, but also simple enough to get high interrater reliability values. 10 seconds-long, 20 seconds-long and 60 seconds-long annotations were also tested and 30 seconds-long was chosen in the final analysis.

Our manual inspection showed that the camera did not always capture the whole group and some students were largely occluded.

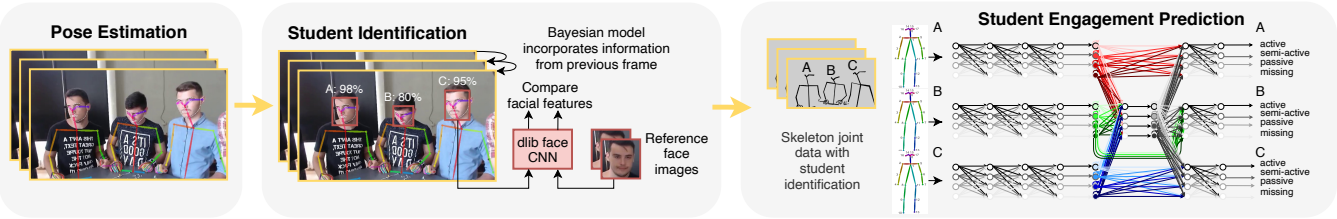

Figure 1: Our proposed framework i) detects body key points from the recordings of student interactions; ii) combines face recognition with a Bayesian model to identify and track students with a high accuracy; and iii) classifies student engagement leveraging a Team Long Short-Term Memory (Team LSTM) neural network model.

In some cases, although a student was labelled as active, they were not visible in the clip, either their hands were visible only or they were missing from the clip completely. To handle such cases, we introduced a fourth label called "missing". This resulted in the following the distribution of labels: 53% active, 24% semi-active, and 7% passive, and 16% missing.

4 PROPOSED FRAMEWORK

As shown in Figure 1, the proposed framework is composed of three main steps: (1) estimating body pose in the video and extracting their key points; (2) performing face recognition to identify and track individual students and discard other people; and (3) training a neural network model called Team Long Short-Term Memory (Team LSTM) to classify student engagement into three categories, namely, active, semi-active and passive.

4.1 Pose Estimation

Our framework builds upon OpenPose library [3] to estimate human pose from videos. OpenPose utilises a bottom up approach where individual body parts are recognised in the image and potential human poses are constructed from the recognised parts. In the clips, students are positioned in a way that objects they are working on are within the camera view most of the time. Therefore, to recognise if a student is modifying or holding an object of interest, we also estimate hand pose using OpenPose and include hand key points in the features to train the model in addition to the body key points. OpenPose output contains a numerical value representing confidence of estimation for each key point. These values are also used as an input to the network.

Taken together, each student's nonverbal cues at any time instant are represented by triplet of body and hand key point values, namely, x coordinate, y coordinate, and confidence. For each student, the network takes as input a feature vector consisting of $3 \times 25 = 75$ values for body pose and $3 \times 21 = 63$ for each hand pose, resulting in 201 features (body plus two hands) in total.

4.2 Student Identification

Our framework uses face recognition python library [1] from dlib [12] for student identification. The main improvement is that it utilises a Bayesian model to incorporate information from previous frames to improve student identification and tracking accuracy. To improve face recognition accuracy, we manually select multiple template face images (around five) for each student, comprising slightly different views. In practice, template face images can be obtained by the student standing in front of camera for few seconds.

Table 1: Student identification accuracy and statistics (%).

	Student A	Student B	Student C
Average	87	81	84
Max.	100	100	100
Min.	65	15	68
Std Dev	11	25	12

We use the output of OpenPose to locate potential face regions in a frame, and these regions are then fed into the Dlib CNN Face model to extract embeddings. These embeddings are then compared with the embeddings from template face images using Hungarian algorithm. Euclidean distance is used to calculate the distance between template embeddings and query embeddings. We assign estimated skeletons to the recognised faces based on the distance between face box and nose key point. However, this frame-by-frame basis approach yields unreliable results on some of the videos, especially when students look away from the camera or their face is blurred due to rapid movement. Therefore, following the method in [18], we build a Bayesian statistical model to improve the accuracy by incorporating information from the previous frame.

4.3 Student Engagement Prediction

Landolfi et al. [14] uses a deep neural network with recurrent LSTM layers as final stage classifier. This approach has been shown to be useful, considering temporal essence of this classification problem. Therefore, we have based our model on this design which has multiple dense layers connected to LSTM layer and a final pooling layer to obtain final classification result. However, they train a separate network for each student, and their method cannot learn interactions between students. Inspired by Social LSTM [2], we have developed a neural network model, called Team LSTM, to address this problem. More explicitly, our model includes an additional shared LSTM layer on top of individual student layers to allow the network to learn the student collaborative interactions.

Our proposed Team LSTM architecture can be summarised as follows. Features from each student are independently inputted into multiple linear neural network dense layers with rectifier activation function (ReLU). The main purpose of this initial block of layers is to learn higher level representations (such as recognising certain gesture or position) from the key points data. These layers have identical weights for each student. To prevent over-fitting, additional dropout layers are added in between the dense layers. Output of the last dense layer is used as input into LSTM layer. Again this layer have the same weights for all students, but the hidden state

and cell state are learned separately for each student. Output of the individual LSTM is used as an input to the Team LSTM layer which is shared between students. The final dense layer (pooling) combines output of the Team layer and the individual student LSTM layers and outputs 4 values - probabilities for each possible student state (i.e., active, semi-active, passive, and missing). This layer is also individual for each student, and the weights are shared.

4.3.1 Team LSTM Layer. For modelling the interactions between the students, we have taken inspiration from Social LSTM cell [2], which combines hidden states of multiple LSTM cells between multiple frames. In contrast to the problem of pedestrian trajectory tracking as presented in [2], in our case there is no need to use position based pooling as all 3 students interact with each other all the time and their locations do not change much. Initially, our model was implemented pooling all the hidden states of the student LSTM cells between frames. However, this configuration did not yield promising results in capturing the interactions in our case. Differently from [2], instead of pooling hidden states and modifying LSTM inner functioning, the outputs of all individual student LSTM layers are pooled using a dense linear layer and are inputted into the Team LSTM layer (of the same size as an individual LSTM).

5 EVALUATION

In this section, we evaluate our proposed framework on the CPS dataset introduced by [7] and compare Team LSTM model with a baseline model.

5.1 Implementation

5.1.1 Baseline Model. To evaluate the effectiveness of the Team LSTM model, we have built a model using a similar structure to the Team model but without the LSTM Team layer. This model resembles the design of Landolfi et al. [14], but it is not direct reimplementation. To reuse all the built pipelines and make the model directly comparable with the Team Model, this model takes the same form of input and produces the same output. The network is trained on the 3 students at the same time, but the networks of individual students are not connected in any way.

5.1.2 Training. The recorded videos do not have the same frame rate, thus the 30 seconds windows have variable number of frames. Therefore, for training, the frames are sampled to set the frame rate to 10 frames per second. When training the models, the 30 seconds windows (corresponding to 300 frames) are inputted into one learning step. The result of the model after processing the last frame is compared with the ground truth label. OpenPose is not able to estimate the body pose and hand pose when there are significant occlusions, or a student is missing from the frame completely. In case of missing key point the input value is set to 0.

Both models are implemented using Pytorch deep learning library [23] and our implementation is publicly available[1]. We use stochastic gradient descent with momentum for optimizing the neural network weights, where we set learning rate to 0.001, momentum to 0.9, and dropout rate to 0.2. We train models with cross entropy loss using 500 epochs. K-fold cross validation is used to measure performance of the model, with K selected as 7 leaving clips from 2 of 14 videos for testing in each fold. Since personalised models are more appropriate for applications in education, we follow a subject-independent cross-validation approach, namely, there is a chance that the same student can appear in both training set and testing set.

[1]https://github.com/angelikang/UnderstandingComplexGroupInteractionsFromVideos

Table 2: Comparison of the Team model with the baseline model in terms of accuracy (%).

	Fold1	Fold2	Fold3	Fold4	Fold5	Fold6	Fold7	Ave.
Baseline	17%	6%	16%	3%	15%	14%	42%	17%
Team	63%	70%	52%	23%	93%	57%	48%	**53%**

Table 3: Confusion matrix of the Team model result.

	Passive	Semi-active	Active	Missing
Passive	**8**	107	123	5
Semi-active	137	**216**	523	15
Active	372	304	**1265**	35
Missing	29	22	69	**478**

5.2 Experimental Results

5.2.1 Student Identification. In Table 1, we present our results for student identification. Proposed modifications to face recognition algorithm in dlib achieves an accuracy of 84% on average, which is reliable enough for the subsequent engagement prediction task.

5.2.2 Student Engagement Prediction. In Table 2, we compare the Team model with the baseline model in terms of accuracy (%). We present our results for the 4-class classification task (i.e., active vs. semi-active vs. passive vs. missing). Looking at the results, the Team model achieves significantly better performance compared to the baseline model on this challenging computer vision problem. The Team model is able to recognise 53% of labels correctly.

We provide the break-down of the results by folds in Table 2. The performance of individual folds differs significantly. For example, fold 5 outperforms all other folds with a performance of 93%, on the other side of the spectrum, fold 4 performs poorly with only 23% of successfully classified labels. The high variance can be explained by the challenging nature of the problem. Some clips do not show the whole group and some students are largely occluded, leading to errors in face recognition, and/or pose estimation.

We also present the confusion matrix of the Team model in Table 3. As can be seen, one of the main challenges is the unbalanced distribution of the labels, where only 7% of the labels is "passive". Consequently, our model performs better for predicting the "active" state as compared to "passive" and "semi-active" states.

5.3 Conclusion and Future Work

In this paper, we focused on an emerging computer vision problem, predicting student engagement in collaborative problem solving purely from video data. For this problem, we proposed a novel method, called Team LSTM, to capture the interactions between the students, and our results demonstrated the potential of the Team LSTM approach for automatically detecting student engagement in collaborative problem-solving from video data. Although our results show significant implications for exploring this problem further, our approach can also be improved from several aspects. Particularly, for future work, we will investigate strategies to alleviate the imbalanced data problem.

REFERENCES

[1] Ageitgey. 2020. face_recognition. https://github.com/ageitgey/face_recognition. [Online; accessed September 28, 2020].

[2] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. 2016. Social LSTM: Human Trajectory Prediction in Crowded Spaces. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016). https://doi.org/10.1109/cvpr.2016.110

[3] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. 2019. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019).

[4] Mutlu Cukurova, Judith Bennett, and Ian Abrahams. 2018. Students' knowledge acquisition and ability to apply knowledge into different science contexts in two different independent learning settings. *ReseaRch in science & Technological educaTion* 36, 1 (2018), 17–34.

[5] Mutlu Cukurova, Michail Giannakos, and Roberto Martinez-Maldonado. 2020. The promise and challenges of multimodal learning analytics. *British Journal of Educational Technology* (2020).

[6] Mutlu Cukurova, Rose Luckin, Eva Millán, and Manolis Mavrikis. 2018. The NISPI framework: Analysing collaborative problem-solving from students' physical interactions. *Computers & Education* 116 (2018), 93–109.

[7] Mutlu Cukurova, Qi Zhou, Daniel Spikol, and Lorenzo Landolfi. 2020. Modelling collaborative problem-solving competence with transparent learning analytics: is video data enough?. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*. 270–275.

[8] Suzanne Dikker, Lu Wan, Ido Davidesco, Lisa Kaggen, Matthias Oostrik, James McClintock, Jess Rowland, Georgios Michalareas, Jay J Van Bavel, Mingzhou Ding, et al. 2017. Brain-to-brain synchrony tracks real-world dynamic group interactions in the classroom. *Current Biology* 27, 9 (2017), 1375–1380.

[9] Vanessa Echeverria, Roberto Martinez-Maldonado, Tamara Power, Carolyn Hayes, and Simon Buckingham Shum. 2018. Where is the nurse? Towards automatically visualising meaningful team movement in healthcare education. In *International conference on artificial intelligence in education*. Springer, 74–78.

[10] Shuchi Grover, Marie Bienkowski, Amir Tamrakar, Behjat Siddiquie, David Salter, and Ajay Divakaran. 2016. Multimodal analytics to study collaborative problem solving in pair programming. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*. 516–517.

[11] David W Johnson, Roger T Johnson, and Mary Beth Stanne. 2000. Cooperative learning methods: A meta-analysis.

[12] Davis King. 2017. High Quality Face Recognition with Deep Metric Learning. http://blog.dlib.net/2017/02/high-quality-face-recognition-with-deep.html. [Online; accessed December, 2019].

[13] Paul A Kirschner, John Sweller, and Richard E Clark. 2006. Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational psychologist* 41, 2 (2006), 75–86.

[14] L. Landolfi, E. Ruffaldi, M. cukurova, and D. Spikol. 2019. Collaboration analysis of students physical interaction based on neural networks and body pose. *Technical report* (2019).

[15] Nichola Lubold and Heather Pon-Barry. 2014. Acoustic-prosodic entrainment and rapport in collaborative learning dialogues. In *Proceedings of the 2014 ACM workshop on Multimodal Learning Analytics Workshop and Grand Challenge*. 5–12.

[16] Roberto Martinez Maldonado, Judy Kay, Kalina Yacef, and Beat Schwendimann. 2012. An interactive teacher's dashboard for monitoring groups in a multi-tabletop learning environment. In *International Conference on Intelligent Tutoring Systems*. Springer, 482–492.

[17] Roberto Martinez-Maldonado, Vanessa Echeverria, Olga C Santos, Augusto Dias Pereira Dos Santos, and Kalina Yacef. 2018. Physical learning analytics: A multimodal perspective. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*. 375–379.

[18] K.Stürmar T. Seidel P. Gerjets U. Trautwein E. Kasneci. Ö. Sümer, P.Goldberg. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2018, pp. 2315-2324. Teachers' Perception in Classroom. *ICMI 19: 2019 International Conference on Multimodal Interaction* (The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2018, pp. 2315-2324).

[19] J Olsen, K Sharma, N Rummel, and V Aleven. 2020. Using multimodal data to temporally analyze collaborative learning outcomes: Benefits and challenges. *British Journal of Educational Technology* 51, 5 (2020).

[20] Jennifer K Olsen, Daniel M Belenky, Vincent Aleven, and Nikol Rummel. 2014. Using an intelligent tutoring system to support collaborative as well as individual learning. In *International Conference on Intelligent Tutoring Systems*. Springer, 134–143.

[21] Sharon Oviatt and Adrienne Cohen. 2013. Written and multimodal representations as predictors of expertise and problem-solving success in mathematics. In *Proceedings of the 15th ACM on International conference on multimodal interaction*. 599–606.

[22] Sharon Oviatt, Kevin Hang, Jianlong Zhou, and Fang Chen. 2015. Spoken interruptions signal productive problem solving and domain expertise in mathematics. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. 311–318.

[23] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[24] Mirko Raca, Lukasz Kidzinski, and Pierre Dillenbourg. 2015. Translating head motion into attention-towards processing of student's body-language. In *Proceedings of the 8th international conference on educational data mining*.

[25] Bertrand Schneider, Patrick Jermann, Guillaume Zufferey, and Pierre Dillenbourg. 2010. Benefits of a tangible interface for collaborative learning and interaction. *IEEE Transactions on Learning Technologies* 4, 3 (2010), 222–232.

[26] Bertrand Schneider, Kshitij Sharma, Sebastien Cuendet, Guillaume Zufferey, Pierre Dillenbourg, and Roy Pea. 2016. Detecting collaborative dynamics using mobile eye-trackers. Singapore: International Society of the Learning Sciences.

[27] Daniel Spikol, Emanuele Ruffaldi, Giacomo Dabisias, and Mutlu Cukurova. 2018. Supervised machine learning in multimodal learning analytics for estimating success in project-based learning. *Journal of Computer Assisted Learning* 34, 4 (2018), 366–377.

[28] M Vujovic, D Hernàndez-Leo, S Tassani, and D Spikol. 2020. Studying collaborative learning and space design with multimodal learning analytics. *British Journal of Educational Technology* 51, 5 (2020).

[29] Marcelo Worsley. 2018. (Dis) engagement matters: Identifying efficacious learning practices with multimodal learning analytics. In *Proceedings of the 8th international conference on learning analytics and knowledge*. 365–369.