Multimodal Perception and Statistical Modeling of Pedagogical Classroom Events Using a Privacy-safe Non-individual Approach

Anderson Augusma

MULTIMODAL PERCEPTION AND STATISTICAL MODELING OF PEDAGOGICAL CLASSROOM EVENTS USING A PRIVACY-SAFE NON-INDIVIDUAL APPROACH

DOCTORAL CONSORTIUM ACII2022

Anderson Augusma,
Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France

ABSTRACT

Interactions between humans are greatly impacted by their behavior. These behaviors can be characterized by signals such as smiling, speech, gaze, posture, gesture, etc. Also by the space, surroundings, time, situation, and context created for a particular activity. These signals also define emotion since they are reactions that human beings experience in response to a particular event or situation. Depending on the event or the circumstance, most of these signals can be triggered. That also happens in pedagogical activities in a classroom. Social learning is multi-modal and teaching itself is complex, these underlying cues are not entirely visible and not immediate. We are investigating Context-Aware Classroom (CAC) to provide a multi-modal perception system allowing to capture pedagogical events that occur in it, to help (young) teachers improve their teaching practices. Thanks to deep learning, which has made great progress over the past two decades, and statistical modeling, it is possible to extract and analyze the signals mentioned above to characterize these events. The main problem with this investigation is the fact that the privacy of the participants may not be preserved. From an ethical point of view, a lot of problems can be caused, i.e, privacy must be taken into account when designing artificial intelligence models. Thus, instead of monitoring individual behavior, the focus will be on global emotion, global student engagement, and the global attention level of the whole class using the signals above mentioned.

***K*eywords**: Multi-modal, Context-Aware Classroom (CAC), Deep Learning, Attention-level, Privacy-safe processing, Statistical modeling.

1 Introduction

When it comes to delivering a lesson in a classroom, several techniques can be implemented to catch the student's engagement, attention, or understanding. For example, during the lesson, teachers can use several signals which are characterized by students' behaviors such as smiles, gazes, postures, head nodding, facial expressions, etc. to know how the lesson is delivered in the classroom. In return, the students, by their side, will capture the same signals to keep the link with the teacher during the lesson. These signals also define cues that characterize the emotion of the attendees. Emotion is the main factor to express oneself in body language which is quite natural and practically in all human beings. Emotion is shared between teacher and students and between the students themselves during the lesson since it is defined by the above signals. Despite these large possibilities among interactions between teachers and students, the task of teaching is still complex. Some event cues are not fully and immediately visible to observers. Nowadays, thanks to many advances in technology, several sensors can help to capture the events that occur in a classroom. These sensors can be of various types depending on the need, such as cameras (to capture visual information like facial expression, emotion, and body posture), microphones to capture the sounds like classroom volume and speech, and eye-tracker to know who pays attention to and gaze direction, etc. Over the last decades, deep learning has made great advances allowing to capture these events, especially in the social dimension. Within a Context-Aware Classroom (CAC) equipped with several of these sensors, one can capture and analyze the events that occur during a lesson by combining deep learning and statistical modeling. From the CAC (Fig: 1) with these all sensors, two main challenges remain: The first comes from the diversity of types of recorded data and the great source of variability which can be time, space, students, teachers, class, etc. The second is that the CAC is equipped with perception systems capable of identifying the participants, capturing their faces, movements, voices, and their slightest gestures which means that the students are entirely identifiable. From privacy and ethical points of view, there are many concerns. To face the ethical issues, the proposed method will avoid monitoring individual behavior to be focused on global emotion, global engagement of students, and global attention level within the classroom.

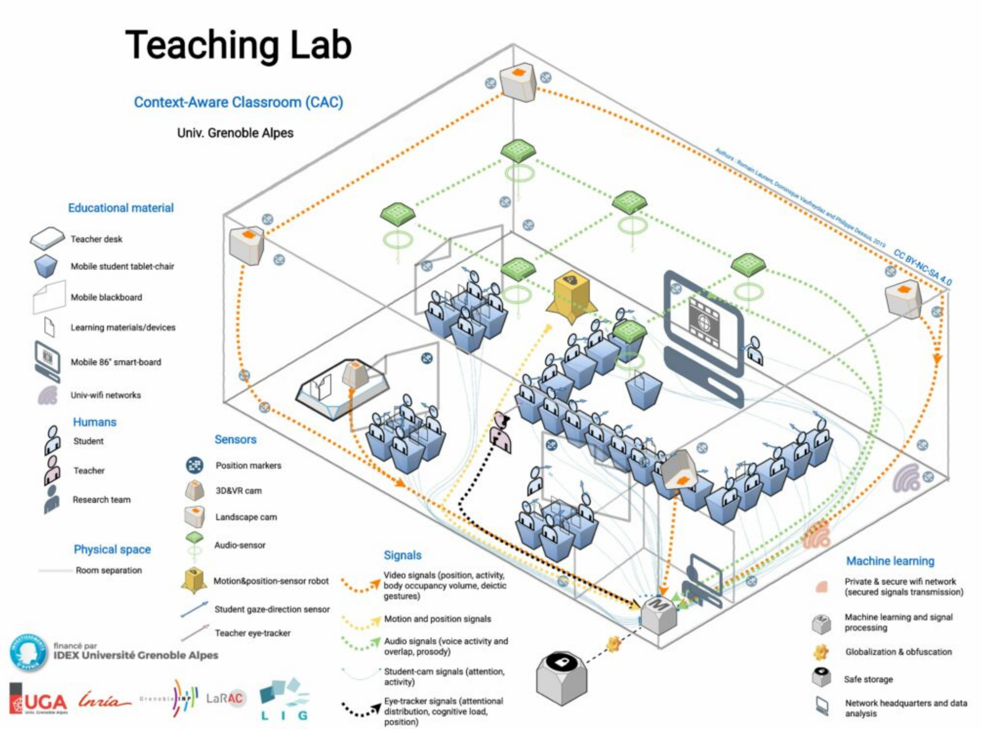
Teaching Lab
Context-Aware Classroom (CAC)
Univ. Grenoble Alpes
Educational material
Teacher desk
Mobile student tablet-chair
Mobile blackboard
Learning materials/devices
Mobile 86" smart-board
Univ-wifi networks
Humans
Student
Teacher
Research team
Physical space
Room separation
Sensors
Position markers
3D&VR cam
Landscape cam
Audio-sensor
Motion&position-sensor robot
Student gaze-direction sensor
Teacher eye-tracker
Signals
Video signals (position, activity, body occupancy volume, deictic gestures)
Motion and position signals
Audio signals (voice activity and overlap, prosody)
Student-cam signals (attention, activity)
Eye-tracker signals (attentional distribution, cognitive load, position)
Machine learning
Private & secure wifi network (secured signals transmission)
Machine learning and signal processing
Globalization & obfuscation
Safe storage
Network headquarters and data analysis
M
CC BY-NC-SA 4.0
financé par
IDEX Université Grenoble Alpes
UGA
Inria
Grenoble INP
LaRAC
L I G

Figure 1: *Teaching Lab - Context-Aware Classroom. The room is equipped with ambient cameras and microphones, a smart interactive display. Mobile chairs and tables let the teacher organize the classroom at will. (Image courtesy of R. Laurent.)*

2 Objective

This thesis aims to create new multi-modal models to capture global information via multi-view audiovisual recordings of the whole classroom without compromising the privacy of the attendees. This is a combination between machine learning and statistical modeling to describe teacher-student interactions within a CAC. The Context-Aware Classroom (CAC) contains sensors, furniture and associated perception models that can capture events that occur in it. The classroom is a genuine classroom where teachers can come and give their lessons normally. As depicted in Fig. 1, the room is equipped with ambient cameras and microphones, an intelligent interactive screen, mobile chairs and tables allow the teacher to organize the class according to his wishes. As this is a system that identifies all participants, there are privacy and ethical issues. The goal of the system is to perceive the underlying cues of teaching episodes (such as student engagement, attention level, etc.) to help teachers improve their teaching practices later, not to monitor individual behaviors. The final goal is to perceive these cues while maintaining privacy. From this, some research questions emerge:

- How to model and analyze the different cues of authentic events such as: speeches, class noises, attention distribution, interactions (with peers and surroundings), emotional traits and actions performed in the classroom without compromising the privacy of attendees?
- How do ethical concerns about personal data influence the effective analytical capabilities and performance of a CAC?

3 Related work

One of the ways of taking into account the distribution of attention is to analyze the distribution of the gaze of the teacher on the students in the class. It's easier to investigate since the teacher is the only one who wears the eye-tracker. As part of their research on teacher-student interaction, Dessus et al.[3] investigated the relationships between the distribution and lability of teachers' attention frequency and the general classroom climate they promote. Woolverton et al. and Kaur et al. present a survey of literature about direct student classroom behavior observation methods [9, 17]. To analyze the interrelationships between attendees of the class, a lot of layers should be taken into account. The research of Markaki et al.[13] argued about these layers such as the class climate, the physical layout of the classroom, the way the teacher interacts with students, the way instruction is delivered, and the values implicitly and explicitly demonstrated by teachers and students throughout the school day. McIntyre et al.[14] demonstrated that the attention distributions of teachers were significantly more similar to those with the same level of expertise and the same culture. Zhao et al.[20] used body posture in a real classroom video to propose a teacher-student behavioral engagement pattern (TSBEP) to synthetically measure student engagement by adding teacher behaviors. Another application of body posture in a classroom was presented in Zhang et al.[19] where they present a method to recognize student posture in a classroom with a large number of students and crowded seating. Regarding privacy-safety, Petrova et al.[15] presented a non-individual approach where they focus on the facial expression of a group of people by trying to ignore the surroundings. Zitouni et al.[21] investigated the recognition of an affective state by masking the faces of people from visual data. In this research, visual cues of body movements and background context were captured from video by masking people's faces to help preserve visual identity for privacy. A comparison of the results was made with the use of raw videos with facial expressions without individual face masking. The results show that affective state recognition can achieve comparable performance in masking and unmasking data for arousal and valence. This may be one of the possibilities that can be considered to deal with ethical and privacy issues in the classroom.

4 Methodology

The proposed method is summarized into two main parts:

1. The first one is to perceive and model real-world events in the classroom. The sensors of the CAC enable to capture features of events on some main dimensions of the instructional situation: space, social, and epistemic dimensions.

   - **Space dimension:** Space can be captured by RFID[1] sensors or video cameras.
   - **Social dimension:** Is captured by eye-trackers (who pays attention to whom), cameras and microphones (facial expressions, emotions and attention of attendees).
   - **Epistemic dimension:** The epistemic dimension is captured by eye-trackers (what knowledge content is examined), types of information recorded from microphones (who is talking about what).

   From these dimensions some inputs will need to be computed: Action/Activity, attention distribution [3], global-emotion, posture recognition features etc. It is worth noting that the emotion is one of the factors of student engagement and also direct and color our attention by selecting what attracts and holds it [1, 2].

2. The second one is to model pedagogical interactions within a CAC. Some advanced statistical modeling will be employed: multi-level models, dimensionality reduction methods with statistical methods to perform efficient feature selection and classification [11, 12].

5 First work

The first works carried out within the framework of this study, are based on the former approach proposed by Petrova et al. in their paper "*Group-level emotion recognition using a unimodal privacy-safe non-individual approach*" [15]. The approach is uni-modal as it is focused only on facial expression to capture emotion. This paper addresses an investigation of perceiving the emotion of a group of people with an ethical and privacy-safe approach, i.e. an approach excluding individual-based features. The goal of this work was to recognize the emotion of the group-level of people in the videos [16] by classifying them into 3 classes: Positive, Neutral, and Negative. This paper took the eleventh place in the EmotiW 2020 challenge [4] with an accuracy of 59.13% on the test set.

The data used comes from the VGAF database [16] for group-level emotion recognition in videos. The collected videos were downloaded from the web and have a large variation in terms of gender, ethnicity, type of social event, several people, pose, etc. The database is labeled for group-level emotion and cohesion tasks. The total number of videos for the training set is 2661 and for the validation set is 766. The duration of each video is 5 seconds. The number of frames varies across videos, the maximum one is 150 frames.

The current work began by adopting the approach proposed by Petrova [15] because it can be considered as the first step towards a more complete recognition of the atmosphere around a group of people while preserving privacy. And can be well suited in a classroom to avoid tracking a particular student but the whole class globally. And we try to get the best performance before mixing with audio information. The approach is to focus on global visual information from the whole group and not on each member of the group in particular. The method used in this approach was trained on two types of data: real data (VGAF) and synthetic data. Synthetic data (Fig: 2) consists of creating images from real faces showing the six basic emotions [6]: Angry, Happy, Sad, Disgust, Fear, Neutral and by superimposing them on an arbitrary background [18]. Sad, Fear, Angry and Disgust considered as Negative; Happy as Positive; Neutral as Neutral. It is used as a data augmentation process. The idea behind this synthetic data enhancement is to guide the neural network while training to focus on the faces in the images while trying to ignore the background and use them as data augmentation.

To implement this approach, two types of models were elaborated in parallel: a static one and a dynamic one. The static model is finetuned on the VGG-19 model by finetuning the last layer and adding new sequential linear layers. The best results for the static model are 60.97% and 60.84% of accuracy on the validation set for the choice of 1 frame and 10 frames respectively, per video. To achieve

[1] Radio-frequency identification (RFID) uses electromagnetic fields to automatically identify and track tags attached to objects or people.

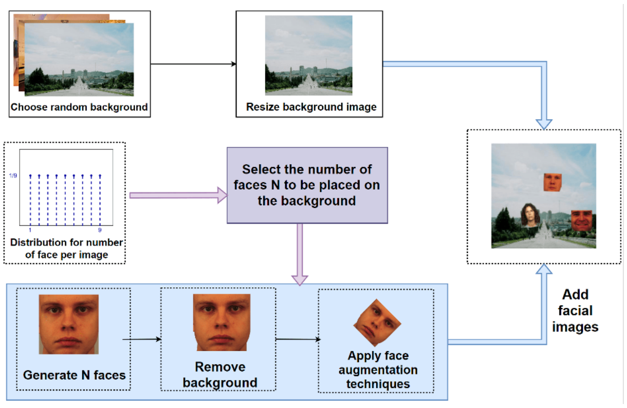
Choose random background
Resize background image
Distribution for number of face per image
Select the number of faces N to be placed on the background
Add facial images
Generate N faces
Remove background
Apply face augmentation techniques

Figure 2: Synthetic data generation processing. (Source [15])

these results, several tests and experiments have been done especially on the way the frames are chosen from videos for the training. When we chose the all frames of each video, the best accuracy is 55.2% on the validation set which is lower than results with 1 and 10 frames. This difference shows that the frames at the beginning of the video do not always carry the same information from the beginning to the end. That is normal since the facial expression of people can change throughout the video sequence. To go further in the analysis of the results on the classification of the videos, we used the best model to perform the prediction on the frames of the videos of the validation set. The overall accuracy is 49.95% for all frames and the distribution among the videos is:

- 24.8% of videos have an accuracy of 100%.
- 24.15% of videos have an accuracy greater than 50%.
- 26.63% of videos have an accuracy less than 50%.
- 24.41% of videos have an accuracy of 0.0%.

Then, with a majority vote on all the frames (for all videos), we have 75% of accuracy. That means the model can up to 75% of performance if the best representative frame is chosen for each video.

To make sure the good frames are chosen every time in the training step, we perform a similarity test approach on the validation set to see the results. As the previous analysis has shown there are variabilities between frames of a video, we decided to group the most similar frames. The main idea is that the frames that belong to the same group (or are most similar) bring the same information, and we choose the frames from the biggest ones in terms of proportion. We used the network of the model to extract features of all frames of video. After extracting these features we perform two types of classification: the first one with cosine-similarity distance which achieves 62.66% of accuracy and the second one with a clustering method achieves 63.69% of accuracy.The improvement remains minor but it shows that it is possible to increase the accuracy if we analyze this approach in depth in the training data.

The dynamic model is a CNN-BI-LSTM model by using the CNN features extraction of VGG-19. Before mixing real and synthetic data, the model is trained with only real data. Several tests were done with different timesteps. For the timestep equal to 2 and 10, the results are respectively 61.72% and 61.46% of accuracy. This result shows that the CNN-BI-LSTM model performs better than the previous one since it achieves almost the same accuracy without mixing with synthetic data.

6 Perspective

The very next work will be focused on the improvement of theses models. For the static model, the improvement perspective is summarized:

- Focus on the frame's choice similarity: choose among those which are similar or among most similar sequences. Another advantage of similarity along a sequence is representations for human action recognition by using a temporal-self similarity [5, 8]. That can be useful to recognize which actions most characterize the video or a sequence of the video.
- Force the model to learn the same weights for all frames of a video so that the frames are considered the same everywhere. Triplet-Loss [7] and Contrastive Learning [10] will be investigated to achieve this perspective.

For the dynamic model, the improvement perspective will focus on training by increasing the timestep gradually and then adding the synthetic data. Afterward, We will start mixing audio and images and more variables to go toward a multi-modal approach. In parallel, recordings in our CAC will allow us to address the statistical modeling part and thus apply the models already available to these new data.

References

[1] Manisah Mohd Ali and Noorfaziha Hassan. Defining concepts of student engagement and factors contributing to their engagement in schools. *Creative Education*, 09:2161–2170, 2018. ISSN 2151-4755. doi: 10.4236/ce.2018.914157.

[2] Aaron Ben-Ze'. The affective realm, 1997.

[3] Philippe Dessus, Olivier Cosnefroy, Vanda Luengo, and Vanda " Luengo. "keep your eyes on 'em all!": A mobile eye-tracking analysis of teachers' sensitivity to students. pages 72–84, 2016. doi: 10.1007/978-3-319-45153-4_6. URL `https://hal.archives-ouvertes.fr/hal-01362185`.

[4] Abhinav Dhall, Garima Sharma, Roland Goecke, and Tom Gedeon. Emotiw 2020: driver gaze, group

emotion, student engagement and physiological signal based challenges. In Nadia Berthouze, Mohamed Chetouani, and Mikio Nakano, editors, *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 784–789, United States of America, 2020. Association for Computing Machinery (ACM). doi: 10.1145/3382507.3417973. URL `https://icmi.acm.org/2020/index.php?id=cfp,https://dl.acm.org/doi/proceedings/10.1145/3382507`. International Conference on Multimodal Interaction 2020, ICMI 2020 ; Conference date: 25-10-2020 Through 29-10-2020.

[5] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Counting out time: Class agnostic video repetition counting in the wild. 6 2020. URL `http://arxiv.org/abs/2006.15418`.

[6] Natalie C. Ebner, Michaela Riediger, and Ulman Lindenberger. Faces-a database of facial expressions in young, middle-aged, and older women and men: Development and validation. *Behavior Research Methods*, 42:351–362, 2 2010. ISSN 1554351X. doi: 10.3758/BRM.42.1.351.

[7] Weifeng Ge. Deep metric learning with hierarchical triplet loss. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[8] Imran Junejo, Emilie Dexter, Ivan Laptev, Patrick Pérez, and Imran N Junejo. View-independent action recognition from temporal self-similarities. URL `https://hal.inria.fr/hal-01064695`.

[9] Avneet Kaur, Munish Bhatia, and Giovanni Stea. A survey of smart classroom literature, 2 2022. ISSN 22277102.

[10] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper/2020/file/d89a66c7c80a29b1bdbab0f2a1a94af8-Paper.pdf`.

[11] Frédérique Letué, Marie-José Martinez, and Nathalie Henrich Bernardoni. Modelling repeated paired phonetic measures using linear mixed models with correlated errors. URL `http://www.csbigs.fr`.

[12] Frédérique Letué, Marie José Martinez, Adeline Samson, Anne Vilain, and Coriandre Vilain. Statistical methodology for the analysis of repeated duration data in behavioral studies, 3 2018. ISSN 10924388.

[13] Vassiliki Markaki and Laurent Filliettaz. Shaping participation in vocational training interactions: The case of schisming, 4 2017.

[14] Nora A. McIntyre and Tom Foulsham. Scanpath analysis of expertise and culture in teacher gaze in real-world classrooms. *Instructional Science*, 46: 435–455, 6 2018. ISSN 15731952. doi: 10.1007/s11251-017-9445-x.

[15] Anastasia Petrova, Dominique Vaufreydaz, and Philippe Dessus. Group-level emotion recognition using a unimodal privacy-safe non-individual approach. URL `https://hal.inria.fr/hal-02937871`.

[16] Garima Sharma, Shreya Ghosh, and Abhinav Dhall. Automatic group level affect and cohesion prediction in videos. pages 161–167, 2019. doi: 10.1109/ACIIW.2019.8925231. URL `http://acii-conf.org/2019/workshop-information/`. International Conference on Affective Computing and Intelligent Interaction Workshops and Demos 2019, ACIIW 2019 ; Conference date: 03-09-2019 Through 06-09-2019.

[17] Genevieve Alice Woolverton and Alisha R. Pollastri. An exploration and critical examination of how "intelligent classroom technologies" can improve specific uses of direct student behavior observation methods. *Educational Measurement: Issues and Practice*, 40:7–17, 9 2021. ISSN 17453992. doi: 10.1111/emip.12421.

[18] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. 6 2015. URL `http://arxiv.org/abs/1506.03365`.

[19] Yiwen Zhang, Tao Zhu, Huansheng Ning, and Zhenyu Liu. Classroom student posture recognition based on an improved high-resolution network. *EURASIP Journal on Wireless Communications and Networking*, 2021, 12 2021. doi: 10.1186/s13638-021-02015-0.

[20] Jian Zhao, Jiaming Li, and Jian Jia. A study on posture-based teacher-student behavioral engagement pattern. *Sustainable Cities and Society*, 67, 4 2021. ISSN 22106707. doi: 10.1016/j.scs.2021.102749.

[21] M. Sami Zitouni, Peter Lee, Uichin Lee, Leontios Hadjileontiadis, and Ahsan Khandoker. Privacy aware affective state recognition from visual data. *IEEE Access*, pages 1–1, 2022. ISSN 2169-3536. doi: 10.1109/ACCESS.2022.3165622. URL `https://ieeexplore.ieee.org/document/9751038/`.