Multimodal Visual Sensing : Automated Estimation of Engagement

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultat
der Eberhard Karls Universitat Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
M. Sc. Ömer Sümer
aus Izmir, Türkei

Tübingen
2021

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	26.02.2021
Stellvertretender Dekan:	Prof. Dr. József Fortágh
1. Berichterstatter:	Prof. Dr. Enkelejda Kasneci
2. Berichterstatter:	Prof. Dr. Andreas Schilling

To my parents...

Acknowledgements

First of all, I would like to thank my supervisors, Prof. Dr. Enkelejda Kasneci and Prof. Dr. Ulrich Trautwein, for their insightful guidance and continuous support during this dissertation. Becoming a part of the Hector Research Institute of Education Sciences and Psychology (HIB) and LEAD Graduate School & Research Network was an excellent experience, from personal and professional perspectives. I am grateful to Prof. Dr. Ulrich Trautwein for his academic guidance. Working on an interdisciplinary project is difficult, but this great environment and his continuous source of motivation facilitated and made this dissertation possible.

I would like to thank all members of my committee, Prof. Dr. Andreas Schilling, for agreeing to be on my committee and for evaluating my work.

I feel very privileged for conducting the research resulting in this dissertation within the scope of Leibniz-Institut für Wissensmedien (IWM)'s Leibniz ScienceCampus. I would like to thank Prof. Dr. Peter Gerjets for his help and critical suggestions during the project and being on my committee. I am also grateful for working with all ScienceCampus project members: Dr. Wolfgang Wagner, Dr. Richard Göllner, Prof. Dr. Kathleen Stürmer, Rainer Adolf, and my colleague Patricia Goldberg.

I would like to thank my collaborators in the different studies resulting in this dissertation, Prof. Dr. Sidney D'Mello, Prof. Dr. Olaf Kramer, Dr. Fabian Ruth, and Dr. Cigdem Beyan.

I thank my colleagues past and present in the Human-Computer Interaction (previously Perception Engineering) group. Learning is a collaborative effort, and I felt your support from the first days. I am indebted to Dr. Shahram Eivazi for his personal and professional recommendations and friendship. I especially thank my friend Efe Bozkir. It was a great chance to share the workplace and deadline stress and unquestionably discuss science together.

Finally, I want to thank my father and mother for raising me, giving me courage for new starts, believing in my decisions, and for being there for me when I needed them. I also thank my little brother and sister for their support.

Tübingen, January, 2021 Ömer Sümer

Abstract

Many modern applications of artificial intelligence involve, to some extent, an understanding of human attention, activity, intention, and competence from multimodal visual data. Nonverbal behavioral cues detected using computer vision and machine learning methods include valuable information for understanding human behaviors, including attention and engagement.

The use of such automated methods in educational settings has a tremendous potential for good. Beneficial uses include classroom analytics to measure teaching quality and the development of interventions to improve teaching based on these analytics, as well as presentation analysis to help students deliver their messages persuasively and effectively.

This dissertation presents a general framework based on multimodal visual sensing to analyze engagement and related tasks from visual modalities.

While the majority of engagement literature in affective and social computing focuses on computer-based learning and educational games, we investigate automated engagement estimation in the classroom using different nonverbal behavioral cues and developed methods to extract attentional and emotional features. Furthermore, we validate the efficiency of proposed approaches on real-world data collected from videotaped classes at university and secondary school. In addition to learning activities, we perform behavior analysis on students giving short scientific presentations using multimodal cues, including face, body, and voice features.

Besides engagement and presentation competence, we approach human behavior understanding from a broader perspective by studying the analysis of joint attention in a group of people, teachers' perception using egocentric camera view and mobile eye trackers, and automated anonymization of audio visual data in classroom studies.

Educational analytics present valuable opportunities to improve learning and teaching. The work in this dissertation suggests a computational framework for estimating student engagement and presentation competence, together with supportive computer vision problems.

Zusammenfassung

Viele moderne Anwendungen der künstlichen Intelligenz beinhalten bis zu einem gewissen Grad ein Verständnis der menschlichen Aufmerksamkeit, Aktivität, Absicht und Kompetenz aus multimodalen visuellen Daten. Nonverbale Verhaltenshinweise, die mit Hilfe von Computer Vision und Methoden des maschinellen Lernens erkannt werden, enthalten wertvolle Informationen zum Verständnis menschlicher Verhaltensweisen, einschließlich Aufmerksamkeit und Engagement.

Der Einsatz solcher automatisierten Methoden im Bildungsbereich birgt ein enormes Potenzial. Zu den nützlichen Anwendungen gehören Analysen im Klassenzimmer zur Messung der Unterrichtsqualität und die Entwicklung von Interventionen zur Verbesserung des Unterrichts auf der Grundlage dieser Analysen sowie die Analyse von Präsentationen, um Studenten zu helfen, ihre Botschaften überzeugend und effektiv zu vermitteln.

Diese Dissertation stellt ein allgemeines Framework vor, das auf multimodaler visueller Erfassung basiert, um Engagement und verwandte Aufgaben anhand visueller Modalitäten zu analysieren.

Während sich der Großteil der Engagement-Literatur im Bereich des affektiven und sozialen Computings auf computerbasiertes Lernen und auf Lernspiele konzentriert, untersuchen wir die automatisierte Engagement-Schätzung im Klassenzimmer unter Verwendung verschiedener nonverbaler Verhaltenshinweise und entwickeln Methoden zur Extraktion von Aufmerksamkeits- und emotionalen Merkmalen. Darüber hinaus validieren wir die Effizienz der vorgeschlagenen Ansätze an realen Daten, die aus videografierten Klassen an Universitäten und weiterführenden Schulen gesammelt wurden. Zusätzlich zu den Lernaktivitäten führen wir eine Verhaltensanalyse von Studenten durch, die kurze wissenschaftliche Präsentationen unter Verwendung von multimodalen Hinweisen, einschließlich Gesichts-, Körper- und Stimmmerkmalen, halten.

Neben dem Engagement und der Präsentationskompetenz nähern wir uns dem Verständnis des menschlichen Verhaltens aus einer breiteren Perspektive, indem wir die Analyse der gemeinsamen Aufmerksamkeit in einer Gruppe von Menschen, die Wahrnehmung von Lehrern mit Hilfe von egozentrischer Kameraperspektive und mobilen Eyetrackern sowie die automatisierte Anonymisierung von audiovisuellen Daten in Studien im Klassenzimmer untersuchen.

Educational Analytics bieten wertvolle Möglichkeiten zur Verbesserung von Lernen und Lehren. Die Arbeit in dieser Dissertation schlägt einen rechnerischen Rahmen zur Einschätzung des Engagements und der Präsentationskompetenz von Schülern vor, zusammen mit unterstützenden Computer-Vision-Problemen.

Contents

1 List of Publications

Accepted Articles

1. Patricia Goldberg, **Ömer Sümer**, Kathleen Stürmer, Wolfgang Wagner, Richard Göllner, Peter Gerjets, Enkelejda Kasneci, and Ulrich Trautwein. "Attentive or Not?: Toward a Machine Learning Approach to Assessing Students' Visible En- gagement in Classroom Instruction". In: *Educational Psychology Review* (2019). url: https://doi.org/10.1007/s10648-019-09514-z.
2. **Ömer Sümer**, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. "Attention Flow: End-to-End Joint Attention Estimation". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. Mar. 2020.
3. **Ömer Sümer**, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. "Automated Anonymisation of Visual and Audio Data in Classroom Studies". In: *The Workshops of the Thirty-Forth AAAI Conference on Artificial Intelligence*. Feb. 2020.
4. **Ömer Sümer**, Patricia Goldberg, Kathleen Sturmer, Tina Seidel, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. "Teachers' Perception in the Classroom". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. June 2018.

Submitted Articles

1. **Ömer Sümer**, Patricia Goldberg, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. "Multimodal Engagement Analysis from Facial Videos in the Classroom". submitted to *IEEE Transactions on Affective Computing*. 2020.
2. **Ömer Sümer**, Cigdem Beyan, Fabian Ruth, Olaf Kramer, Ulrich Trautwein, and Enkelejda Kasneci. "Estimating Presentation Competence using Multimodal Nonverbal Behavioral Cues". submitted to *ACM Transactions on Interactive Intelligent Systems*. 2021.

1.1 Scientific Contribution

This work proposes a unified framework for automated behavior understanding in the classroom. Chapter 2 summarizes the general framework of two main research problems, visual modalities required for automated analysis and learning methodology. The general framework of the dissertation, multimodal visual sensing is outlined in Chapter 3.

Six scientific publications from 2018 to 2020 approach engagement, presentation competence, and related behavior analysis problems such as end-to-end joint attention estimation, teachers' perception in the classroom, and automated anonymization of audiovisual data in classroom studies. Practical and method-based contributions to the state-of-the-art and discussion are detailed in Chapter 4.

All publications were published or are in review with renowned conferences or journals.

2 Introduction

Automated human behavior understanding is an important research problem relevant in many disciplines such as psychology, machine learning, signal processing, computer vision, and human-computer interaction. The ability to detect humans, understand actions, and perceive intentions is vital when developing intelligent systems that interact with humans in varying tasks. Computers can achieve this to some extent. Machine learning and artificial intelligence models can perceive our actions and intentions using video, audio, or physiological signals. Within the context of the environment, multimodality is essential for better sensing and understanding of human activities. This is especially important because fusing different sensor data or features yields a better representation of data and enhances the performance of machine learning tasks. Many applications of human behavior understanding directly affect outcomes in our educational, social, and professional lives. Educational sciences and psychology are particularly prominent among these applications. The use of data-driven technologies can enhance learning efficiency and result in a better educational system.

Analyzing emotions is a more developed problem in human behavior understanding. Instead, we focus on analyzing cognitive activities and learning-related behaviors, particularly student behaviors in the classroom. Student engagement is a vital prerequisite for classroom learning. When students are more attentive to the learning material or teacher, interact, and ask or reply to questions, they are more engaged. Research in educational sciences shows that teachers vary in eliciting and guiding students' attention to learning materials. Particularly novice teachers might have difficulty understanding aspects of engagement in the classroom. Thus, the use of machine learning methodologies in the classroom can benefit the estimation of student engagement in several ways. First, these methodologies can be used on recorded classroom instruction videos to analyze student behaviors as a part of teacher education. Second, they can be used in classroom management studies that aim to understand the effects of interventions (i.e., in teaching methodology) on student outcomes. The last and final objective is to validate the performance of different modalities and learning algorithms in engagement estimation within the context of an online classroom system that directly gives feedback to teachers. Student engagement estimation can be useful in both traditional

classroom situations and more dynamic group discussions.

In addition to student engagement, another important education topic is assisting students with public speaking and presentation competence. In personal, professional, and academic life, the skills and abilities necessary to competently relay a message are essential. Presentation competence can be estimated by multimodal visual sensing. Such an automated system can be beneficial for students seeking to improve public speaking and presentation competence. Our work focuses particularly on short scientific presentations by students.

In this thesis, we introduce a multimodal visual sensing framework to address both problems, student engagement and presentation competence estimation, in the learning context. In an interactive learning situation and a more active presentation setting, we show that the use of different modalities and learning algorithms can efficiently estimate behaviors in a manner effective for a human-computer interaction workflow to improve student outcomes. Furthermore, there may be students who do not consent to the monitoring of their behaviors. We present a framework to anonymize non-participants' faces and voices to make video observation studies or deployment of such systems as an interface available.

The rest of this chapter is organized as follows: We first cover the educational problems that motivate us to develop computational approaches, engagement in the learning context (2.1) and presentation competence (2.2). Subsequently, we review the computational foundations from the social and affective computing domain (2.3) and explain the nonverbal cues that we use to address both problems and the automated methods used to estimate them (2.3.2).

2.1 Engagement in the Learning Context

The etymological origin of the world, *engagement*, is French and dates back to the early 15th century. To engage meant to "to pledge" something, as security for payment. Later, the word was used in the context of agreeing formally to marriage. While for hundreds of years it was related to formal, mostly legal obligations, in time the word engagement came to mean "to attract and occupy the attention."

In the educational context, the study of student engagement started with educational psychologist Ralph Tyler who examined the relationship between how much students spent on their work and learning in the 1930s at Ohio State University and the University of Chicago. In order to express similar concepts in student learning, different words were used. For instance, C. Robert Pace's research on the quality of student efforts in the 1960s and Alexander Astin's theory of student involvement in the 1980s both employed different terms for similar behavioral traits. According to Astin, the quantity and quality of psychosocial and physical energy is decisive for the academic achievement of students [1].

Psychological and Educational Perspective. In more recent literature, school engagement is defined as a multifaceted construct. According to Fredericks et al. [2], the three dimensions of engagement are behavioral (academic and social), cognitive, and emotional engagement.

Behavioral engagement refers to students' participation in learning, including attentiveness, completing assignments in the classroom and at home, and learning through extracurricular activities. The social aspects of behavioral engagement include following written and unwritten rules of the classroom, coming to school on time, and not exhibiting antisocial behaviors. *Cognitive engagement* consists of the thoughtful activities behind comprehending concepts and ideas explained in the classroom. For instance, asking questions for clarification, reading more material than assigned, and using self-regulation strategies to guide learning can be indicators of cognitive engagement. Last, *affective engagement* is in the measure of feelings of interest, such as general satisfaction about a particular topic, teacher, or school. More detailed definitions of these dimensions and their connection to other motivational variables and contextual influences can be found in [3].

As student engagement is highly multidimensional, it incorporates a student's socio-economic background, motivational factors, and parents; however, our focus is limited to classroom instruction where audiovisual sensors have the ability to perceive behaviors. There are several motivations for student engagement analysis, and memory is one of them.

Memory can be divided into two categories: long-term and short-term (working) memory. Long-term memory incorporates a significant portion of information learned throughout one's life. Short-term memory, on the other hand, encompasses a small amount of information maintained during the execution of cognitive tasks such as classroom learning [4].

According to cognitive theories of instruction (**C**ognitive **L**oad **T**heory, CLT [5]; **C**ognitive **T**heory of **M**ultimedia **L**earning, CTML [6]), the capacity to process the data is limited. According to Mayer [6], there are auditory and visual channels for information processing; each channel has a finite capacity, and learning is an active process of filtering, organizing, and integrating new information with prior knowledge. Considering classroom instruction with visual aids, teacher, and other peers, keeping students' attention on learning-related tasks is more effective through the use of working memory.

Besides the student aspect, understanding student attention and engagement is also decisive for teaching quality [7]. In the literature, teaching quality is considered related to student engagement and learning [8, 9]. Teachers' instructional practices can be gathered in three dimensions [10]: classroom management, cognitive activation (instructional support), and constructive guidance (emotional support). Classroom management and cognitive activation are the principal mainstays of classroom learning, and analysis of students' learning behaviors (time-on-task and more) can potentially provide insight into both.

So as to avoid confusion, we should note that human behaviors are not only composed of observable components; there is a covert aspect to behaviors. On the other hand, there is

a wide range of actions that we can sense through visual activities. These constitute overt behavior. Our focus is to understand behaviors by sensing, mainly using visual modalities, with an aim to analyze overt attention and engagement.

Engagement measures. The first and most crucial part of visible engagement analysis is how we measure engagement. The most popular measures of engagement are self-reports and observational methods. Self-reports are gathered through questionnaires completed by students before or after attending a class. A teacher's rating of students can be used to validate them. In general, self-reports may vary from student to student as the internal mechanism to judge oneself and what one's understanding the questionnaire items may differ. Even though self-reports are the most widespread measure of engagement, they contain drawbacks such as social desirability bias, memory recall limitations, acquiescence bias, and halo effects [11].

Table 2.1: Learning Activities in the Classroom according to ICAP framework [12].

Situation	Mode	Examples
Listening to a lecture	**I**nteractive *(dialoguing)*	Defending and arguing a position in dyads or small groups
	Constructive *(generating)*	Reflecting out-loud; Drawing concept maps; Asking questions
	Active *(manipulating)*	Repeating or rehearsing; Copying solution steps; Taking verbatim notes
	Passive *(receiving)*	Listening without doing anything else but oriented toward instruction

Observational methods can be conducted in real time online in the classroom or performed by expert raters reviewing video recordings of classes. Some examples of observational methods (also used in our studies) are on-task off-task behavior assessment such as in the Munich Attention Inventory (MAI) [13] and a revised version for video recordings [14] and Chi & Wylie's ICAP (**I**nteractive, **C**onstructive, **A**ctive, and **P**assive) framework [12]. Table 2.1 shows the learning activities in the classroom according to ICAP.

The main drawback of observational methods is the human effort required to train observers to reliably rate student behaviors and annotation time. Human effort limits the deployment of classroom management studies that aim to understand the relationships between various teaching methods, skills, and student outcomes. The motivation for developing automated methods to estimate student engagement in this thesis arises from this point. The availability of measures that can reliably observe and analyze large numbers of students and classrooms promises to increase the scale of classroom observation studies without introducing any additional human effort.

Another measure of engagement is the Electronically Activated Recorder (EAR), a device that

samples audio clips during learning or employs physiological sensors such as electrodermal activity (EDA) recorders. However, the necessity of annotation involved introduces additional costs per student and the noisy nature of speech and human physiology make these measures difficult to scale.

As our focus is an automated, multimodal, and visual understanding of engagement, which we capture through the audio-visual recording of classroom instruction using field cameras. To show how reliable machine learning approaches are to estimate engagement, we acquire engagement measured by observational methods and report the performance of computational methods.

2.2 Presentation Competence

What makes a presentation great? How can we deliver our messages to the audience effectively? If we do not perform well when speaking in front of an audience, how can we improve our competence and learn how to communicate better in a persuasive or informative setting? These questions are highly valuable because presentation competence is positively associated with success in academic and business life while also contributing to long-term professional success. Presentation competence is also one of the core competencies in education. To some extent, the answer to all of these questions can be found in the field of rhetorical and discourse analysis. In the literature, several competence rubrics define different aspects of verbal and non-verbal speaker behaviors.

Even though observer reports based on these public speaking rubrics are a measurable and reliable way to assess the competence, reliance on human ratings has several drawbacks. Human raters have to be trained to analyze the speaker's behavior according to the assessment criteria of a competence rubric either online or by watching audio-visual recordings of a presentation. When the number of speeches increases, more human raters are required, and even the most reliable rubric can lead to variation in the perception of different evaluators. This makes rating a large number of speaking performance unfeasible. Furthermore, without a real-time assessment of presentation competence, it is impossible to utilize for self-regulation and interactive training purposes.

To tackle these problems, an automated analysis of presentation competence is a promising alternative. In the last decade, multimodal analysis of human behaviors rapidly developed by advancing deep learning and the availability of big data. Many computer vision problems such as gaze and body pose estimation can be reliably addressed even under challenging environmental conditions. Similarly, audio processing problems from speaker recognition to affective analysis of voice or speech-to-text perform well. Despite the success for low-level tasks, the literature in automated presentation competence analysis is still quite limited.

During our formal and vocational education, success in many situations relies on practical expertise in communication and presentation acquired over years of learning. Having an

Table 2.2: Comparison of Assessment Rubrics for Presentation Competence.

Assessment Rubric	**Target level**	**Item number**	**Seperate items per NFs**	**Sample** (#speech)	(**Interrater**) **Reliability**
Classroom Public Speaking Assessment Carlson et al. [15]	higher education	(Form B) 5 items/ 5-point scale	✗	2	– Cronbach coefficient: from .69 to .91
Public Speaking Competency Instrument Thomson et al. [16]	higher education	20 items/ 5-point scale	✗	1	n.a.
Competent Speaker Speech Evaluation Form Morreale et al. [17]	higher education	8 items/ 3-point scale	✗	12	– Ebel's coefficient: from .90 to .94 – Cronbach coefficient: from .76 to .84
Public Speaking Competence Rubric Schreiber et al. [18]	higher education	11 items/ 5-point scale	✗	45-50	ICC: $.54 \leq r \leq .93$
Tübingen Instrument for Presentation Competence Ruth et al. [19]	high school	22 items/ 4-point scale	✓	161 *(T1)* 94 *(T2)*	– Cronbach coefficient: from .67 to .93 – ICC > .60 for 10 out of 15 items

automated instrument to assess presentation competence can be helpful in many ways. A potential use case is educational settings. For instance, in professional teacher training and qualification, a multimodal visual sensing and analytics system built on non-verbal and verbal behavior analysis can create a useful tool to measure and improve teachers' public speaking and communication skills. We can train machines to imitate human raters' evaluations according to psychologically valid assessment instruments. Furthermore, it is possible to use the available big data and public speaking and presentation recordings on the internet to learn multimodal representations of public speaking and presentation competence.

In the literature, there are several presentation competence assessment rubrics. The most popular rubrics in educational speeches are depicted in Table 2.2. These instruments summarize the various behaviors of the presenter and most use several raters to show the validity of the proposed instruments.

The sample sizes of previous rubrics were very limited. This situation makes their validity questionable in different types of settings. With the exception of Schreiber et al. [18], they were tested on a few speeches. None have separate items for nonverbal behavioral cues such as gaze direction, facial expressions, gestures, and posture except for Tübingen Instrument for Presentation Competence (TIP).

Similar to other social and affective computing problems, there are visible and nonvisible aspects of presentation competence. As our focus is to develop an automated tool to estimate presentation competence using multimodal behavioral cues, having a reliable measure of presentation competence is essential. Even though there is a line of work in the computational

domain that aims to estimate presentation competence or give behavioral analytics to the presenter, they lack structured and psychologically valid assessment rubrics. In our studies to assess presentation competence, we used the TIP instrument, particularly the group of items representing body language and voice, which can be sensed solely using audio-visual recordings. These items measure how effective and convincing the presenter's use of body posture, gestures, eye contact, facial expressions, and voice during the presentation is.

From the human-computer interaction perspective, The ability of an automated tool to sense a presenter's behavior offers exciting potential. Such a tool can provide presenters with useful feedback and help improve presentation competence over time. There is a line of work developed either in situ interfaces to give presenters real-time feedback or in offline tools for improving communication skills. For instance, Tanveer et al. [20] proposed a Google Glass based system that provides online feedback about the volume and speed of the speaker's voice. Another example of online practices is the use of the virtual audience [21] and conversation coach [22]. There are more recent examples that combine multimodal cues and subjective ratings on an interface for collaborative training and improving communication skills [23, 24]. However, most of these semi-automated feedback systems aimed to display raw multimodal cues in an interface instead of estimating expert ratings. In other words, these tutors do not use a psychologically valid assessment tool and try to improve the users' speaking performance based on defined rhetorical principles.

2.3 Computational Perspective: Multimodal Visual Sensing

Our two main problems, estimating student engagement and presentation competence, aim to automatically sense student behaviors in mostly passive (listening to a lecture) and active (giving a short scientific presentation) settings. Despite the differences between engagement and presentation competence, a computer program that aims to automatically understand either needs to use a similar set of nonverbal behaviors. We can consider them a unified problem, human behavior understanding composed of *sensing* human actions, *perceiving* affective and cognitive states, and *interacting* and sharing the findings with users.

Automated engagement estimation approaches learning by analyzing student behaviors, whereas presentation competence indirectly addresses teaching. The integral part of learning and teaching is multimodality. Various forms of interactions happen in learning situations, and multimodal sensor data is a crucial solution to better sense learning and teaching situations. By focusing mainly on the visual modalities, we propose a computational framework, "multimodal visual sensing", to recognize human behaviors in learning-related situations.

The computational problems that are helpful when tackling both engagement and presentation competence analysis are shown in Figure 2.1. In the visual domain, object detection is the initial step of any application of multimodal visual sensing. Detection, in this context, is required to recognize human presence in the scene. Our work mainly requires *face detection* and *person detection.* Instead of still images, we usually work with temporal data. Thus associating

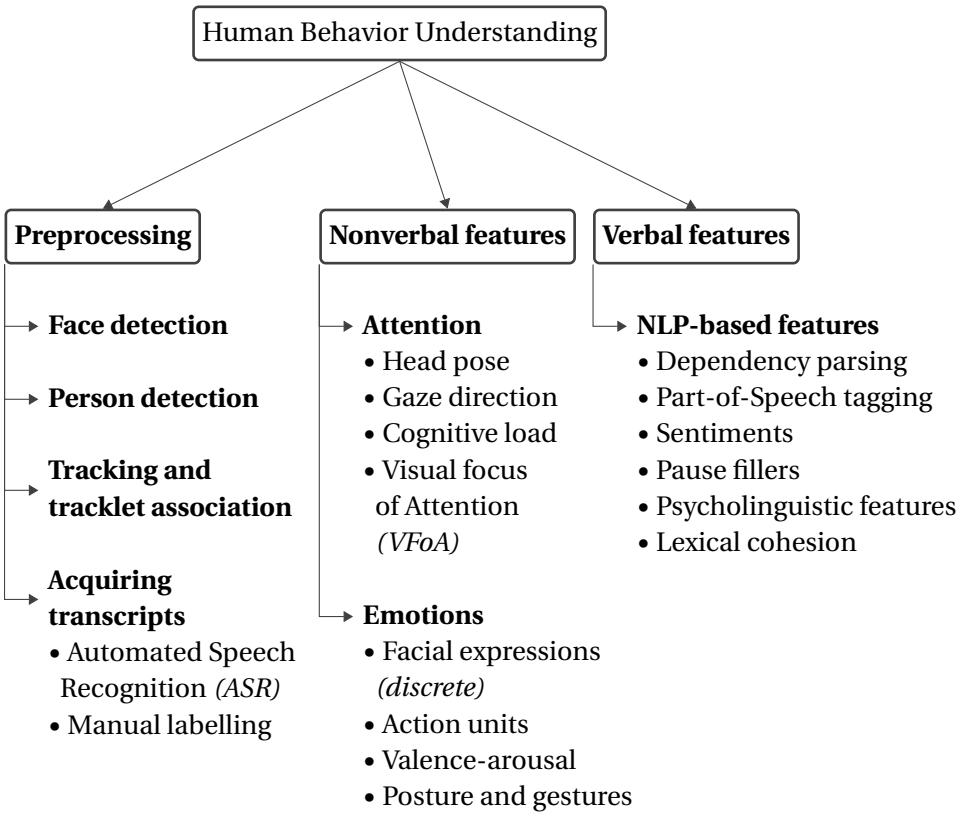


Figure 2.1: Human Behavior Understanding using nonverbal and verbal features.

detection of the same identity of two images in time is a second important step. Detections in each time point must be related to each other by *object tracking or tracklet association* methods.

After confirming face and body sequences, the next task is to identify those behaviors we aim to analyze. Even though we could perform this task by analyzing bodies, the most reliable solution is to deploy a *face verification* method.

So far, deployed detection, tracking, and identification methods can give us the behavioral data belonging to the person who we aim to study in order to better understand his or her behavior. At that point, multimodal visual features that are cues of affective and behavioral engagement can be extracted using the data acquired in previous steps. We can separate visual nonverbal features into two categories: *affective* and *attention-related*.

Both affective and attention-related features are essential to better understand the underlying behaviors. Even though there are methods to sense emotions from gait or gestures, the most reliable method is to analyze facial expressions. Facial expressions refer to either discrete categories of emotion or a more advanced measurement system based on emotion theory, for instance, the **F**acial **A**ction **C**oding **S**ystem (FACS) or valance-arousal model.

Attention-related features tell us where we attend. In both human-human or human-object interactions, attention is crucial, and it can be sensed using head pose estimation and gaze estimation methods. Attention can also be sensed from gaits. Gait sequences reveal information about our temporal attention and activities.

In the following sections, we first review the machine learning and computer vision methodologies needed for understanding engagement and presentation competence. Then, we describe the traditional and state-of-the-art approaches to estimate affective and attentional nonverbal features.

2.3.1 Foundations

In this section, we review the methodology in computer vision and the verbal and nonverbal features described in Figure 2.1.

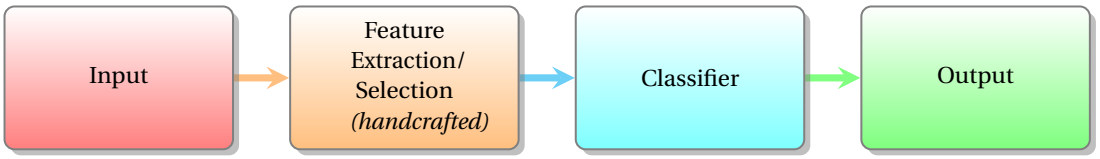
Computer Vision Methodology

The computational tasks at the heart of multimodal visual sensing are composed of computer vision tasks, mainly object detection and classification. Even though human behavior analysis tasks can exist in higher levels of abstraction than object analysis, the approach remains similar.

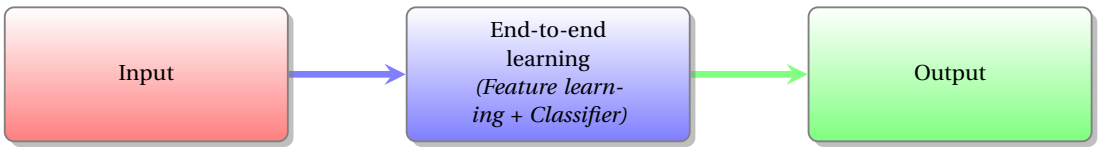
To start with an example, let us examine object detection. Since the late 1990s, the three main approaches for object detection are feature-based, template-based, and appearance-based. Feature-based methods try to find distinctive image region locations and review a small portion to the whole image in order to determine whether the geometric configuration of parts belongs to the desired object or not. Template-based approaches, such as active appearance models (AAM) [25], use templates in different scales and poses. As they are susceptible to the initialization, these approaches were used for detecting object parts, such as facial keypoint detection or human pose estimation, rather than face or pedestrian detection. Appearance-based methods scan the entire image by overlapping patches in different sizes and scales and depend on training a robust classifier to determine the difference between object or non-object patches.

One common factor in these approaches, object detection (also same in object recognition and other computer vision tasks), was approached as feature extraction and classification. As shown in Figure 2.2, the success of deep learning in object detection and recognition since 2012 [26] replaced feature extraction/selection and classifiers.

The handcrafted features extract the local or global appearance, shape, or texture statistics of an image; however, they are needed to be designed and limited in terms of representation ability. Wavelets and Gabor filters [27], Histogram of Oriented Gradients (HoG) [28] (Histogram of Oriented Optical Flows, HOOF [29] in videos), Scale-Invariant Feature Transform (SIFT)



(a) Traditional computer vision workflow.



(b) Deep learning workflow.

Figure 2.2: The traditional computer vision workflow and deep learning.

[30], Local Binary Patterns (LBP) [31] and its variants [32, 33] are examples of handcrafted features.

When handcrafted features are extracted, larger feature dimensions can also pose a problem due to the *curse of dimensionality* [34]. In practice, the variation in the target variable is predictable from a confined set of features [35, 36]. Furthermore, we assume that the real data distributions are assumed to have (locally) smoothness properties. Using manifold embedding approaches high dimensional features can be mapped into a lower-dimensional space subsequently to perform final classification.

The idea of substituting feature extraction and classification, and learning them together from the data goes back to the earliest years of pattern recognition (Rosenblatt's Perceptron algorithm, [37]).

The minimization of errors through steepest descent was first proposed by Cauchy in 1847 [38]. Training multiple layers of a neural network in the complex, nonlinear, and differential parameter space is achieved by using an objective function and updating the weights of multiple stacked layers according to the chain rule of derivatives. The use of gradient descent in neural networks was discussed in the 1960s and 70s and first used by Werbos [39]. After Rumelhart and Hinton's work [40] in 1986, it became more popular.

Another contribution that makes neural networks suitable for visual sensing is the use of convolutional filters. As image data is captured as 2-dimensional pixel intensity values in cameras (and 3-dimensional by scene reconstruction), convolutional neural networks can better model local and global relations. Fukushima and Miyake's Neocognitron [41] was the first neurophysiologically inspired CNN architecture. Later, LeCunn et al. [42] trained a Neocognitron-like architecture using backpropagation and applied it to handwritten digits classification.

In 2012, Krizhevsky, Sutskever, and Hinton's CNN (AlexNet) [26] won the ImageNet LSVRC-2010 contest to classify 1000 classes in real-world, challenging images. Beyond [42, 43], the

contributions of AlexNet were the use of rectified linear units (ReLU) as activation (instead of tanh or sigmoid), local response normalization (LRN) to normalize data and dropouts that help prevent overfitting. These novelties were adopted and are still being used in CNNs. Since 2012, neural network architectures in computer vision are in progress [44, 45, 46, 47].

Use of deep learning in affective computing. The fundamentals of understanding human behavior depend on supervised learning and require labeled data. For instance, in a more general domain, such as computer vision tasks such as object classification, face recognition, or body pose estimation, it is possible to collect large-scale datasets. However, it is not straightforward to increase the number of subjects in most affective computing problems because scraping webscale data is not a solution for problems such as engagement or presentation competence. We can compare the sample sizes in different vision tasks. For instance, there are hundreds of thousands of people in face recognition or pose estimation datasets. In contrast, previous studies in engagement analysis dealt with a limited number of subjects.

Deep learning-based methods require a vast number of training datasets and variations in the data. A CNN model trained on a limited number of subjects, even though the data scale is large, cannot generalize on different people. For this reason, a traditional computer vision workflow (in Figure2.2) is still being used even in recent studies. This situation deprives studies using limited data from the expressive power of representation learning in deep learning.

On the contrary, our approach to this dissertation's main problems is to acquire representations of relevant vision problems learned from data and transfer this to our limited data problems.

Subjective Labeling Many supervised problems in computer vision require objective labeling. The labeling of an object category, bounding box, or segmentation map does not contain any confusion or subjectivity. Alonso et al. defined three types of questions for labeling: objective, judgment, and subjective. Affective computing problems lie under judgment and subjective. There are many underlying factors contributing to the labeler's decision, such as ordering the stimuli and bias towards or against particular visual traits. Annotation of the same data by many labelers (i.e., crowdsourcing) can discard these subjective decisions and provide a probability distribution. However, this is often not a solution because of issues related to privacy and required training necessary for labelers.

The main tasks in our focus, engagement and presentation competence, can be considered judgmental or subjective; thus, we used intra-class correlation (ICC) [48] to measure interrater reliability in our studies.

2.3.2 Nonberval Behavior Analysis

Engagement or presentation competence can be considered high level of concepts. In this dissertation, we approached these problems by automatically analyzing multiple visual modalities and subsequently estimating them. Our particular focus is on nonverbal behaviors.

Head Pose Estimation

Estimating gaze is very difficult in many situations, and creating camera settings that capture eye direction reliably can be costly and affects the natural state of human behaviors. For this reason, human-computer interaction researchers have been working on estimating head pose. Stiefelhagen [49] reported that 89% of the time in meetings, head pose and gaze are focused in the same direction. This observation confirms the importance of head pose in behavior understanding. Particularly in behavior analysis of small meetings, head pose was used as the primary modality to estimate Visual Focus of Attention (VFoA).

Estimating head pose involves predicting the head's orientation in the image in terms of pitch, roll, and yaw angles in the three degrees of freedom of the head. The earliest methods used appearance template models, nonlinear regression, manifold embedding, and machine learning classification to estimate a discrete number of exemplar head poses [50]. Later, the progress in localizing facial keypoint [51] led geometric methods to become more popular [52, 53, 54]. An essential drawback of facial keypoint and perspective matching is that the mislocalization of even one keypoint can cause the failure of pose estimation.

Through the progression of deep learning, new large scale head pose datasets has become available, and regression of continuous angles using convolutional neural networks performed on par or even better than geometric methods.

Patacchiola and Cangelosi [55] trained a shallow CNN with regression loss. Ruiz et al. [56] used ResNet-50 by combining regression loss on continuous values and classification loss on discretized values. Recently, Yang et al. [57] combined multilevel feature aggregation and pooling to estimate head pose without keypoints.

Considering the educational situations such as classroom instruction recorded by field cameras, even the state-of-the-art facial keypoint estimation approach works well to align faces for recognition or expression recognition; however, this approach is far from satisfactory for reliable pose estimation. Furthermore, appearance-based CNN models can also be used as a feature representation in behavior.

Gaze Estimation and Attention Mapping

The gaze, even though it is more challenging to estimate than head pose, is one of the most prominent cues to understand the visual focus of attention during human-computer or multi-

party social interactions. Over the last two decades, mobile and remote eye trackers have been commercialized and used in many applications, including cognitive psychology, educational assessment, and market research. The recent advances in virtual reality (VR) and augmented reality (AR) technologies and growing awareness and interest in human-centered artificial intelligence, eye tracking, and gaze estimation play a significant role in understanding human behaviors.

In the literature of eye detection, tracking, and gaze estimation, the computational methods can be classified as shape-based, feature-based, appearance-based or a combination of several approaches [58]. In recent years, the collection of large datasets in real-world settings [59, 60, 61, 62, 63] and synthetic rendering created a valuable resource to train gaze estimation models that can surpass the performance of traditional methods.

In mobile eye-tracking or remote screen-based settings, pupil detection approaches such as ElSe [64] work well. Various challenges, including image quality, camera angle, uncontrolled illumination, occlusion, may prevent reliable gaze direction estimation using image processing-based methods [65]. Whereas image-processing based methods performs fast and accurate in mobile and high-speed screen-based eye tracking, more unconstraint settings require learning-based gaze estimation.

Gaze estimation is one of the most important behavioral cues in egocentric behavior analysis or screen-based interactions, HCI in driving situations [66, 67, 68], and user experience experiments. Related to gaze analysis, pupil-related measures can be used to estimate task-related workload [69, 70]. Gaze can be further analysed in combination with physiological signals [71]. However, it is more challenging to estimate gaze from distant field cameras. Besides, it is not needed in educational settings. For instance, in classroom analytics precise gaze estimation might not be crucial. Instead, understanding high-level behaviors or estimating the interacted objects or persons can be adequate.

Even though gaze is not required for interaction, it is a precursor of human-human or human-object interactions. A line of work [72, 73] aims to localize persons and their interactees in images. We need to see the eye region clearly to estimate gaze direction; however, for an observer who can see the entire scene, including persons and objects, the eye's fine-grained details are not necessary. Recasens et al. [74] created GazeFollow, the first dataset with faces and their gaze locations in the image compiled through crowdsourcing. They estimated approximate gazed locations using the entire scene, face patch, and face location grid using a multi-branch convolution network. Human-object interactions and gaze following problems aim to analyze the behavior of a single person. Beyond the single person, gaze analysis of a group, for instance, joint attention, is another research problem.

Fan et al. [75] created a joint attention dataset, VideoCoAtt, from social scenes in movies and TV series. Detecting and localizing joint attention has the potential for use in various applications. It can also be used in the analysis of learning in small groups.

The tasks of head pose estimation, gaze estimation, human-object interaction, gaze following, and joint attention analysis are all aimed toward the same purpose: to understand the attention of a person or group of people in an activity. The ideal task varies according to the use case, quality, and granularity of data, as well as several underlying factors.

Facial Expression Recognition

As described in Section 2.1, affect is one of the three dimensions of engagement. It is defined as a basic sense of feeling from unpleasant to pleasant (valence) and agitated to calm (arousal). In contrast, emotions are our perceived feelings and associated with neurophysiological changes and behaviors. Emotions are also intertwined with mood, engagement, and personality. As our focus is visual sensing, we aim to understand facial expressions using visual modalities.

Facial expressions are interesting and key to our purpose in this work because of their relationship to emotions and universality. Darwin [76] conducted the first study suggesting the universality of facial expressions. His theory stated that emotions and facial expressions were inborn and thus adaptable through evolution.

In a preliminary study, Ekman et al. [77, 78] showed a cross-cultural agreement in the perception of emotions between the faces of literate and preliterate people. Subsequently, Friesen [79] documented that people across different cultures exhibit spontaneous facial expressions in reaction to emotion-eliciting films. Since Friesen's work, many studies have replicated similar results, and there is a strong indication that universal facial expressions exist for seven emotions: anger, contempt, disgust, fear, joy, sadness, and surprise.

There is a large variance in the ways people express emotions. Despite the universality of facial expressions, they are certainly not obligatory. People can handle their emotions without showing it on their faces. In addition to faces, factors such as body postures, gestures, personality, and voice tone can be decisive when expressing emotions. Furthermore, contextual information also impacts facial expressions. A recent study [80] reviewed the literature about expressed emotions and perception in facial movements and reported a lack of reliability, specificity, and generalizability.

Considering the controversial aspect of emotions, we argue that the best way to reduce potential variations in facial expressions is to evaluate them participating in a specific task. In engagement or presentation analysis, the datasets we analyzed contain subjects from similar age groups and grades performing the same learning task.

In computer vision and affective computing, the datasets and methodology can be categorized into three groups: (i) grouping face images in six or seven discrete emotion categories (alternatively compound emotions [81]), (ii) estimating facial action units [82], and (iii) estimating valance and arousal intensities (Russell's circumplex model [83]).

Similar to other vision problems, facial expression recognition methodology began on datasets

with limited subjects, handcrafted features, and shallow classifiers and evolved into large-scale datasets scraped from the web and deep learning methodologies. In recent years, deep generative models were also used to improve facial expression recognition by generating images of faces from a low-dimensional attribute space.

3 Main Outcomes

This chapter will summarize the papers resulting from the research done during my Ph.D. studies. Under the branch of multimodal visual sensing, each paper performs another purpose. Even though my main focus was to estimate student engagement in the classroom, multimodal and visual computing constitutes a broader range of interest and is applicable to various research problems. I have described the research problems and motivations behind each, as well as a discussion of their main findings in this chapter. The full texts of all papers are included in the Appendix.

Research problems arising from the use of computational methods in learning situations can be categorized in three groups. Figure 3.1 shows the structure of these groups and papers according to their function under the framework of multimodal visual sensing.

When analyzing student behaviors, we can think of two situations where students attend a learning activity or transfer their knowledge to their peers actively. The estimation of visual engagement aims to automatically understand student behaviors in classroom instruction, whereas presentation competence analysis performs behavior analysis when students give a scientific presentation.

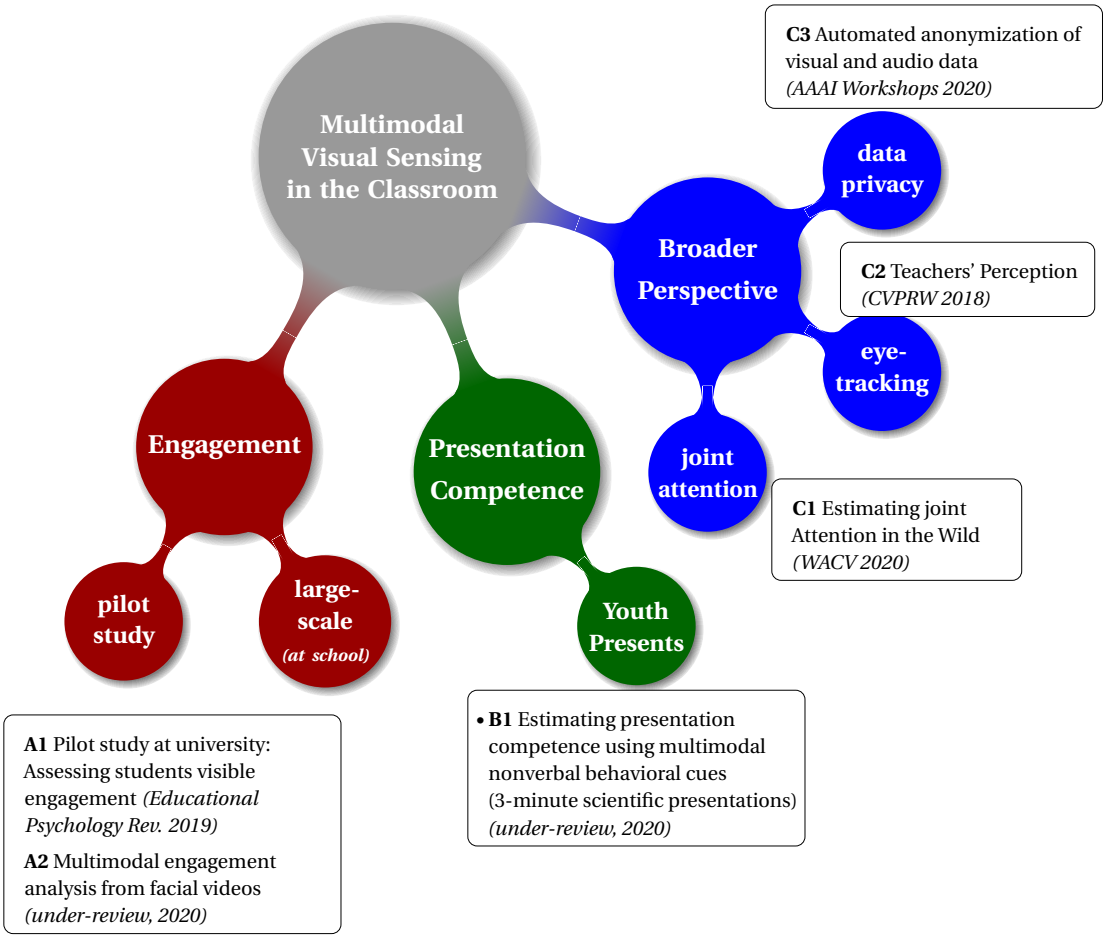


Figure 3.1: The structure of papers resulting from the research done during the dissertation.

3.1 Estimating Student Engagement

3.1.1 Attentive or Not? Toward a Machine Learning Approach to Assessing Students' Visible Engagement in Classroom Instruction

Patricia Goldberg, **Ömer Sümer**, Kathleen Stürmer, Wolfgang Wagner, Richard Göllner, Peter Gerjets, Enkelejda Kasneci, and Ulrich Trautwein. "Attentive or Not?: Toward a Machine Learning Approach to Assessing Students' Visible Engagement in Classroom Instruction". In: Educational Psychology Review (2019).

Motivation

A teacher needs to understand students' behavior properly to know their visible engagement and deploy appropriate teaching approaches accordingly. Novice teachers might not be capable of perceiving student behaviors during instruction. In order to support and train novice teachers, the audiovisual recordings of classroom instruction are being used in teacher

training. By watching and analyzing the recordings, the teachers learn to highlight verbal and nonverbal indicators of students' behaviors and develop classroom management strategies.

In educational psychology, there are manual coding procedures to assess students' behavior in video recordings. Unfortunately, manual coding necessitates expertise and takes a longer time. This drawback constitutes a critical bottleneck of video-based classroom studies. In this study, we investigated machine learning methodologies to estimate the visible engagement of students from videos. The primary motivation is to assess automated methods in engagement estimation and substitute or reduce the need for manual coding in future studies.

The manual rating system that we developed was significantly correlated with self-reported cognitive engagement, involvement, and situational interest and predicted performance on a subsequent knowledge test. Thus, we aim at predicting the manual ratings using computer vision-based nonverbal behavioral features such as head pose, gaze direction, and facial expressions. Furthermore, adding synchrony information to the engagement estimator, correlation of neighboring students' computed behavioral features with each other, improved the prediction of manual ratings, and eventually self-reported variables.

Cognitive activation, classroom management, and teacher support are regarded as fundamentals of teaching quality [84, 85]. Understanding students' (dis)engagement levels indicates much about their cognitive activities and reveals the time-on-task [86]. Teachers are usually expected to be aware of students' behaviors, take notes accordingly. They need to monitor students' attention and also keep them focused on learning tasks. Research shows, however, this may not always be the case, particularly for preservice teachers.

Monitoring students' learning related activities during instruction is difficult [87, 88, 89, 90] and teacher require support to develop the knowledge underlying these skills [87, 88, 89, 90]. The use of video recordings in a way the teachers watched and reflected their instruction showed benefits for inexperienced and experienced teachers [91, 92]. The necessity to watch hours of recordings to find the most decisive parts is demanding, and an automated tool to analyze students' engagement facilitates this process.

In the literature, there are several manual coding systems. In this study, we combined multiple indicators, more precisely, the ICAP framework [12] and the Munich Observation of Attention Inventory (MAI, [13, 14]). Two raters are first trained for our manual coding system and then continuously rated all video recordings using a joystick when watching. We regressed the continuous engagement level by using visible features acquired using computer vision.

Methods

In our study's computational phase, we are given classroom videos and continuous manual engagement labeling of each student's visible engagement during instruction. The camera views on the teacher area's left and right side that observe each student the best are dynamically picked.

We first applied face detection in videos [93], and automatically connected detections of the same students in time. Then, faces were aligned, and their representative features extracted automatically based on the OpenFace library [94]. Considering the amount of occlusion in camera views by peers, laptops, or water bottles, we used a subsample of students (N=30) in our analysis.

As the number of participants is limited, it was not possible to learn the representations for engagement from the data. Thus, we relied on two types of precomputed features: attention features that are composed of head pose and gaze direction and estimated facial action unit (FACS, [82]) intensities as affective features. The head pose features consist of way, roll, and pitch angles of heads. Gaze direction in radians in world coordinates is described by a unit vector that originated from the eyes.

In FACS, facial action units are coded at five levels of intensities. We used the following 17 action units: upper face AUs are AU1 (inner brow raiser), AU2 (outer brow raiser), AU4 (brow lowerer), AU5 (upper lid raiser), AU6 (cheek raiser), and AU7 (lid tightener); the lower face AUs are AU9 (nose wrinkler), AU10 (upper lip raiser), AU12 (lip corner puller), AU14 (dimpler), AU15 (lip corner depressor), AU17 (chin raiser), AU20 (lip stretcher), AU23 (lip tightener), AU25 (lips part), AU26 (jaw drops), and AU45 (blink).

As the manual labeling is done continuously, we used the statistics of all features at 24 frames per second to predict engagement intensities. We regressed intensities using a linear Support Vector regressor. Evaluation is done in a person-independent manner by excluding the test subject and training on the remaining subjects's data. We also performed a correlation analysis of estimated mean engagement levels and self-reported variables after predicting engagement intensities per second.

Our regression analysis concluded that gaze direction or facial expression features are more representative than head pose features. Furthermore, feature-level fusion led to a better performance in terms of mean squared error and Pearson correlation. Beyond using multiple modalities, providing the behavioral alignment, cosine similarity of features between neighboring students performed the best.

Results

So far, most of the previous work in student engagement estimation approached the topic in computer-based classrooms or very limited situations. The manual coding system adopted in these studies was not standardized and used by education and psychology researchers. Our study showed that computer vision and machine learning-based approaches could successfully substitute manual labeling.

Even though regression results are good with respect to the manual coding on the test data, the relationship between estimated engagement and self-reported post-test items were best when head pose and gaze features combined with neighbor synchrony, r = .08, .43, .39, and .26 for the

knowledge test, involvement, cognitive engagement, and situational interest, respectively. The correlation of manual labels with post-test variables was .14, .64, .62, and .53 for the knowledge test, involvement, cognitive engagement, and situational interest. Thus, it is possible to argue that even the ground truth labels are far from correctly representing self-reported post-tests (for instance, the worst relation is with knowledge test). However, we should note that computer vision and machine learning methodology mimics the manual coding systems and deals with only visible cues. Furthermore, the differing assessment mechanisms in self-reports make it complicated to maintain a strong relationship.

The use of university seminars, instead of schools (i.e., secondary schools) and the automated system's failures due to low visual quality or occlusion, are the major limitations. This study proposed an objective measure to manually code student engagement and showed that it could be predicted using machine learning. Automated engagement analysis can provide simultaneous or offline feedback to teachers and makes video-based classroom studies possible with less effort.

3.1.2 Multimodal Engagement Analysis from Facial Videos in the Classroom

Ömer Sümer, Patricia Goldberg, Sidney D'Mello, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. "Multimodal Engagement Analysis from Facial Videos in the Classroom". 2021 (under review with IEEE Trans. on Affective Computing).

Motivation

During classroom instruction, automated engagement analysis can reveal many important questions about student engagement, learning outcomes, and teacher training and have been an active research topic in recent years.

Our previous study (in 3.1.1) conducted a pilot study at university-level seminars and showed that computer vision and machine learning-based methods could predict manually coded student engagement. In this study, our two main objectives are (i) to scale up our previous approach on a larger scale at schools and (ii) to improve our methodology to estimate engagement.

In addition to these two primary goals, we investigated the effect of personalization in engagement analysis. One possible application of engagement analysis can be a cognitive interface that will help the teachers perceive the classroom behaviors better. Thus, the same system needs to be deployed in a classroom many times, for instance, an entire semester. Engagement labeling is very time-consuming and requires raters to be trained. We propose the personalization of engagement models by sampling the person-specific data qualitatively, using active learning methods.

Methods

We conducted this study over the course of a month and a half at a secondary school. The ethics committee from the Faculty of Economics and Social Sciences of the University of Tübingen approved the procedure (Approval #A2.5.4-097_aa). Teachers' and parents' written consent for their kids are taken, and the students who did not want to participate in our study attended a parallel session covering the same courses.

From 47 classes from 5^{th} to 12^{th} grades, including 128 participants, we extracted the main parts of lectures where students are supposed to follow the teacher and actively participate in the instruction (excluding group works). As the manual coding of videos for all students is time-consuming, based on visibility and occurrence in different day/course recordings, we used 15 students from grade 8 (N=7) and grade 12 (N=8) in our analysis.

As a preprocessing step, we detected all faces using a single-stage face detector, RetinaFace [95], and subsequently identified them by face verification with several query images and ArcFace embedding [96]. Finally, we used 24-frame length continuous sequences.

After the preprocessing, this study's main difference from our previous study (in 3.1.1) is the feature representation of faces. OpenFace library's head pose estimation and facial action unit estimation rely on good alignment [97, 94]; however, it is most of the time, not the case in classroom videos. The difficulty of alignment (by detecting facial keypoints and Perspective-n-Point problem) causes eliminating most of the data. Even the alignment works under occlusion, and varying camera angles, facial action unit analysis might not be reliable because all action unit (AU) models were trained on nearly frontal images.

In order to tackle alignment issues and use all the data that face detection and recognition worked, we trained two separate convolutional neural networks (residual neural networks, ResNet-50 [46]): one is on head pose estimation and the other on discrete facial expression recognition problems. Head pose information tells about attention information, whereas facial expressions are affective features. Both attention and affective features are the output of the last convolutional layer of ResNet-50 trained on 300W-LP [98] and AffectNET [99] datasets, respectively.

As the number of subjects in the classroom data is limited, we first train the feature representations on open datasets. Then, these feature representations are extracted in each second of student data. Instead of low-dimensional, pre-computed features, we build our engagement classifiers on these feature embeddings.

We formulated the problem as a 3-class engagement classification: low, medium, and high engagement based on the distribution of all manually coded data. Using attention and affect features, we tested shallow classifiers (Support Vector Machines with linear and radial basis function kernels, Random Forests) and deep learning models such as Multi-Layer Perceptron Long-Short Term Memory (LSTM).

We applied an uncertainty-based sampling strategy for personalization using the Random Forest classifiers as a base classifier in addition to these person-independent models.

Results

Considering the limitation of sample size (N=15), it was not possible to train an end-to-end engagement classifier on the classroom data. We compared different classifiers based on attention and affect features trained on more diverse and unconstraint datasets. The best performing classifier is the Random Forests classifier on top of 2048-dimensional feature embeddings. SVM-rbf classifiers fall slightly behind the RF. On the other hand, using temporal models, LSTMs, improves the MLP baseline ($2-4\%$ improvement in accuracy and F1-scores); however, the base performance of MLP is low, and the use of LSTM models makes the performance only comparable to SVM and RF classifiers. A possible explanation is that our annotations were continuous. Thus, even classifiers such as RF and SVM can perform well together with majority voting, and there is not a large room for improvement for temporal models.

Using both attention and affect features, and in both grades 8 and 12, the best performing engagement classifiers achieved AUCs of .620 and .720 in Grades 8 and 12, respectively. We took the RF-based models as a baseline and tested uncertainty-based sampling for personalized engagement classifiers. We started the person-independent base classifier (RF) and in each episode picked a small batch of unlabeled data and retrained the engagement classifier. In all experiments, personal data in each experiment is restricted to 60 seconds (6 episodes of 10 seconds). Using only 60 person-specific samples yielded an average AUC improvement of .084.

We further investigated the effect of feature-level and score-level fusion. In RF classifiers, score-level fusion improved the performance of the best performing modality, Attention-Net, by a .012 of AUC in Grade 8, and is on par with the Attention-Net. On the other hand, feature-level fusion yield a comparable improvement (+.013) in Grade 8. However, it fell behind and performed .616 where the single modalities, Attention-Net and Affect-Net, performed .708 and .600, respectively.

3.2 Presentation Competence

Ömer Sümer, Cigdem Beyan, Fabian Ruth, Olaf Kramer, Ulrich Trautwein, and Enkelejda Kasneci. “Estimating Presentation Competence using Multimodal Nonverbal Behavioral Cues”. 2021 (under review with *ACM Transactions on Interactive Intelligent Systems*).

Motivation

Similar to student engagement in classroom instruction, gaining the ability to deliver good public speech or presentation is also essential for students. The studies that we focused on student engagement in the classroom dealt with behaviors, mostly in a passive manner. The main difference in presentation competence analysis is that more active nonverbal behavioral cues such as body pose or voice are needed.

In this study, we used a presentation dataset collected within the scope of a nationwide German presentation contest for secondary school students aged 12 to 20. The participating students are either giving a presentation in front of a jury on a scientific topic of their choice or prepare on a predefined topic. The approximate duration of the presentations is 3 minutes.

The main motivation for this study is as follows:

- The literature on automated methods that estimate presentation competence did not adopt a psychologically valid, objective instrument to measure competence. We aimed to investigate the performance of automated methods using the Tübingen Instrument for Presentation Competence (TIP).
- Among several nonverbal behavioral cues, which one performs better in estimating presentation competence? We examined speech, facial, and body pose features in both classification and regression settings.
- Is the rhetorical setting decisive on the performance of automated methods? As both sets of Youth Presents dataset were captured from different rhetorical settings, we investigated the effect of setting on the algorithms' performance.

Methods

Regression results might not be satisfying in terms of correlation performance, particularly in imbalanced datasets. For this reason, we approached the problem of estimating presentation competence as both classification and regression problems. The performance metrics in the regression task are mean squared error (MSE) and Pearson correlation coefficient. In classification, we reported accuracy, precision, recall, and F1-score for each modality and classifier.

In the Youth Presents dataset camera is directly looking at the speakers from a 4-5 meters distance. Similarly, a situation where presentation competence estimation will be used, such as a self-regulatory tool to develop competence for giving successful presentations, will be ideal for face and person detection. Our methodology depends on extracting three types of features:

- facial features (including head pose, gaze direction, and FACS action unit intensities),
- body pose features (two-dimensional locations of body joints),
- speech features (affective acoustic feature sets).

For all feature sets, we experimented both using global features extracted from the entire presentation (approx. 3 minutes) and local features aggregated in 16 seconds of sequences.

Tübingen Instrument for Presentation Competence (TIP) includes 22-items, and the subset of these items covering body language and voice is items 10-15. We used the average of these items as a global competence score for each presentation. The range of rated competence varies between 1 and 4. We discretized all values in the T1 and T2 sets using the median value (2.83) in the classification task to classify competence as low or high.

The feature sets stated above are low-dimensional, and the number of instances is limited in our dataset. Considering their performance in similar tasks, we used gradient boosting (GB), decision trees (DT), random forest (RF), and support vector machines (SVM) as classifiers and regressors.

As well as the contribution and comparative performance evaluation of three modalities and classification/regression methods, another fundamental question is whether different fusion strategies are effective or not. We used feature-level fusion and late-fusion using the median, product, and sum rules.

Results

The best performing feature was speech features using Geneva Minimalistic Acoustic Parameter Set (eGeMAPS). For instance, the GB classifier performed an accuracy of 65.62% and 70.66% of F1-score. Even when the features were extracted from 16-second intervals, the classification performance did not degrade too much.

Body pose features followed the speech features and outperformed facial features. RF classifier using body pose features performed an accuracy of 64.38%. Using the same classifiers, RF, facial features performed 3.76% lower than body pose features.

When global features were used, the feature-fusion of three modalities outperformed late-fusion methods (median, product, and sum rules) by a margin of 4.37% in accuracy. In local features, the performance of single modalities is on par with global features. However, the performance drop is more apparent in fusion.

The results of the regression task are consistent with the classification. The best performing modality is the speech features with Pearson correlations of .56 (SVM) and .50 (GB) in global and local features, respectively. When facial and body pose features were compared, facial features are better with global features than local features. The use of local features in body pose improved the regression performance in GB, DT, and RF regressors.

In the cross-dataset task, we trained on the T1 set and tested on the T2 set. The main difference between the datasets is that the students picked the presentation topics freely and had a longer preparation time in the T1. On the contrary, the topic and presentation material were given,

and the preparation time was limited in the T2. In both classification and regression tasks, the performance in T1→T2 was worse than in T1 results. Classification Considering the experiments in the T1 were done in a person-independent manner; on the other hand, T1→T2 is person-dependent, we would expect better results in cross-dataset.

Considering the experiments in the T1 were done in a person-independent manner; on the other hand, T1→T2 is person-dependent, we would expect better results in a cross-dataset setting. However, comparable lower results indicated that the difference in rhetorical settings made the generalization of the same classifiers or regressors more challenging.

This study showed that a psychologically valid instrument for presentation competence, TIP, can be estimated using nonverbal behavioral cues. An important limitation of our work was the sample size of the presentation dataset, in a total of 252 videos. By keeping the rhetorical setting similar, scaling the data to model all behavioral variances has the utmost priority. Besides, personalization of presentation competence models and development of recommender systems and user interfaces are also among future research topics.

3.3 Broader Perspective

The main applications of multimodal visual sensing were engagement and presentation competence analysis in the classroom. In this group of papers, we approached the problem from a broader perspective. We focused joint attention estimation, the anonymization of sensitive visual and audio data in classroom studies, and the use of egocentric computer vision and eye tracking to analyze teachers' perception.

3.3.1 Attention Flow: End-to-End Joint Attention Estimation

Ömer Sümer, Patricia Goldberg, Kathleen Stürmer, Tina Seidel, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. "Teachers' Perception in the Classroom". In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. June 2018.

Motivation

In engagement and presentation competence estimation, our analysis was based on a single person; however, a group of people's attention can be required in some situations. For instance, let us consider a situation where the joint attention of a group is needed. Video analysis can be a promising solution as a therapeutical tool for children with attention disorders. The main objective is to detect and localize joint attention.

Joint attention can be defined as social interaction when two or more people gaze at each other or interact with an object. The ability of joint attention in humans starts from the age of 3 months. An automated video observation system can help to monitor the development

of gaze following and joint attention in patients. Here, the key difference is that two or more person's attention must be analyzed.

The previous works in this problem used face detection first and then head pose or gaze estimation [74, 75, 100]. In these works, being dependent on two separate convolutional networks introduces an additional computational cost. Thus, deploying the final model in real-time becomes more difficult. Our motivation is to detect the presence of joint attention and locate the center of joint attention on the scene using a single convolutional neural network.

Understanding joint attention requires to focus on the attention of people in the scene. Another relevant problem, saliency estimation, aims to estimate where an average viewer would attend to when seeing a still image or video. The most saliency regions on the saliency map of a scene can be important for joint attention but not necessarily on the center of joint attention. In this study, we use saliency models to detect and localize joint attention better.

Methods

The difficulty of detecting joint attention is that a small difference in the gaze of persons in the scene can affect attention's presence and location. For this reason, training an end-to-end convolutional neural network to estimate joint attention does not perform well [75].

We extract saliency maps of frames in the VideoCoAtt dataset using the following saliency estimation methods: Itti and Koch [101], GBVS [102], Signature [103], and DeepGaze II [104]. The first three depend on computational attention models, whereas DeepGaze II is a data-driven approach combining different level features from a pre-trained convolutional network on image classification.

DeepGaze II's mean saliency value was 96% of the time, above the mean saliency level of the images inside the co-attention bounding boxes. It was 44%, 71%, and 77% for Itti & Koch, GVBS, and Signature. This observation proved the importance of saliency models. They can extensively reduce the search space of joint attention centers in the scene.

In this paper, the first step of our approach is saliency-based ground-truth generation. We create two-channel heatmaps (AttentionFlow), where the first channel shows the face locations in the image, and the second channel is the joint attention likelihood. By investigating these heatmaps, we can tell about the nature of social interactions in the scene. If there is no strong region above a threshold in the first channel, there are no people in the scene. In other words, the first channel acts as a face detector. The second channel will give higher values when the likelihood of two or more persons in the scene attend to the same region. When only co-attention bounding boxes are used, an end-to-end network tends to overfit. The use of saliency information to create these two-channel heatmaps regularizes and makes optimization easier.

We used a single network composed of a feature encoder and generator blocks. Besides the

basic configuration, we proposed two attention modules, channel-wise feature attention, and spatial attention.

Results

We conducted experiments on the VideoCoAtt dataset that contains videos from TV series and movies.

As an ablation study, we compared different strategies on the convolutional part (Encoder) of the AttentionFlow model. These are no learning (only depending on) on Encoder, using the same learning rate as in the Generator part, and finetuning with a slower learning rate. Finetuning performed the best in the task of joint attention localization. It shows that retaining knowledge learned in object classification was helpful for our task, too.

Looking into the effect of spatial and channel-wise attention modules, both improve the baseline without attention mechanism. When the baseline performance, mean L_2 distance between our predictions and the ground truth joint attention centers were 69.72, spatial and channel-wise attention modules perform 65.70 and 62.84, respectively. Spatial attention makes small manipulation on the output of the generator module to improve the initial estimate. On the other hand, channel-wise attention is applied to the Encoder's output that outputs high-level features and selects relevant features with joint attention task based on the given input.

In the literature, there are two works on joint attention detection and localization. The first one is Gaze Follow [74]. They estimate the location of a person gazed at in the image from the entire image, the face region of the person in the scene, and the face location grid. A simple baseline can be to apply the Gaze Follow model for all persons in the scene and accumulate them to acquire a joint attention map. This approach did not perform well (58.7% in prediction accuracy and 102 in L_2 distance). Another study [75] combined a deep network with region proposal and temporal modules and reached to 71.4% and 62 in both tasks. Our best result, even though it uses a single end-to-end network, outperformed both approaches in prediction accuracy with 78.1%. On the localization task, it is also on part with [75]'s results (62.84%).

In conclusion, we proposed two novel methods: saliency-based ground truth generation and two convolutional attention blocks for feature selection and attention map localization.

3.3.2 Teachers' Perception in the Classroom

Ömer Sümer, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. "At- tention Flow: End-to-End Joint Attention Estimation". In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Mar. 2020.

Motivation

Our previous works in engagement and presentation competence analysis were mainly on student behaviors. Also, we used field cameras located on the corner of the classroom to videotape classes or a presentation. When dealing with learning activities, the teacher's egocentric view can also be very informative to understand the teaching quality [88, 105, 106, 107].

Mobile eye trackers are commercial devices that look like standard eyeglasses. They have cameras that see egocentric views and eyes and estimate where the user gazes at the scene. Due to the egocentric movement's nature, there are various challenging situations, such as motion blur and different view angles. Eye-tracking in constraint settings on a computer (that we decide what to show the user and the exact location of the stimuli) is comparably easier to interpret. In mobile eye-tracking, the viewpoint is always changing, and the manual annotation of the region of interests is very time-consuming.

In this study, our motivation is to analyze teachers' perception during classroom instruction. More specifically, for a given egocentric video and eye-tracking data recorded by a teacher, we create spatiotemporal attention maps of teachers and find how the teacher distributed her or his attention in the class.

Methods

In this study, we used the data collected in a previous study [90]. The data was collected in a standardized teaching setting (M-Teach) with a limited amount of students; while the teacher was wearing mobile eye-tracking glasses. In total, there are seven videos in the resolution of 1280×960; each of them contains approximately 20 minutes of instruction time. The problem is to associate the eye-tracker's gaze points with students on the scene if the teacher gazed at and interact with any of the students. This task is done in the original study [90] in six months, and we know the approximate distribution of teachers' attention in the dataset.

We first applied a single shot scale-invariant face detector [93] on all egocentric sequences. Then, low-level tracklet linking is done by combining three affinities, bounding box size, location, and appearance in time. After having face sequences changing in length up to the teacher's gaze behavior, the next task is to identify students. The number of students in the classroom is known. Thus, without any manual annotation of query images, face clustering can be used to identify students.

Egocentric sequences contain varying camera angles, and alignment might not be working well in most situations. For this reason, we used a ResNet-50 representation trained in VGGFace2 that has large viewpoint and pose variations. As each input is a face tracklet, instead of a single face, and the representation is more robust to challenging cases, we used agglomerative hierarchical clustering on ResNet-50 features of sequences.

After acquiring each student's identity, we created bounding boxes around the face and body regions, and associated eye tracker's gaze points to students. The resulting output is an attention map that shows the student's name, where the teacher gazed in the class, and it can be used to understand the distribution of the teacher's attention.

Results

When we processed the entire dataset composed of seven classes, we reached the same attention distribution for all teachers reported by [90]. In this way, the use of computer vision and mobile eye-tracking together reduces months of manual labeling efforts to several hours of computation. It can increase the amount of data that can be collected in teacher training studies.

We showed that computer vision and mobile eye-tracking could also be useful as a real-time system to give teachers feedback by reflecting the summary of their interaction with each student. Then they can regulate their attention adaptively. Mobile eye-trackers are commercial devices and can be expensive to be used by many teachers in a large scale study. We argue that egocentric video capture can do a similar function in a lack of eye-trackers. When the attention maps created by the distribution of gaze points per student and the number of students whose faces are detected in the scene compared, only face processing can yield comparable distribution.

3.3.3 Automated Anonymisation of Visual and Audio Data in Classroom Studies

Ömer Sümer, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. "Automated Anonymisation of Visual and Audio Data in Classroom Studies". In: The Workshops of the Thirty-Forth AAAI Conference on Artificial Intelligence. Feb. 2020.

Motivation

In classroom studies on student or teacher behaviors that require audio or visual recording, there may be students who do not want to participate in the study. In common practice, either seating arrangement is changed, and they are taken out of video coverage or arrange a parallel session for these students. In the first option, these students still may ask questions, and their voices can be audible in the recording. On the other hand, arranging a parallel session introduces additional efforts. This study's motivation is to investigate the applicability of automated methods to anonymize sensitive audio and visual data.

Having the ability to automatically anonymize students' voices or faces who do not consent, it will be possible to keep the usual seating arrangement of those students. Either in real-time or shortly after recording the classes, the data can be anonymized.

Methods

To test automated anonymization in video and audio data, we collected classroom instruction videos from YouTube. The real instruction scenes were uploaded by school districts and recorded with a handheld camera. We scraped 18 classroom instructions over 6 hours of total duration.

The videos contain shot transitions; thus, we first detected all soft and hard transitions using TransNet [108] that uses 3D dilated convolutional networks. Then, RetinaFace [95] is applied to detect all faces in each video shot. As these sequences contain a small number of faces, we grouped faces based on bounding box intersection and minimum weight matching in bipartite graphs. Subsequently, we used an Inception ResNet model trained with triplet loss on the VGGFaceII dataset as a feature embedding. The task is face identification on the entire data. In other words, we should be able to retrieve all face bounding boxes where a student appeared from several query images.

Voice also contains biometric traits and all time intervals where the student who did not consent must be located and silenced. For this task, we used the Unbounded Interleaved-State Recurrent Neural Network (UISRNN) [109]. UIS-RNN performs speaker diarization, and similar to faces, we manually find a few second query samples and retrieve all samples belong to the same group for anonymization. In contrast to clustering-based methods (i.e., k-means or spectral clustering of d-vectors), UIS-RNN is fully supervised, uses a Bayesian non-parametric process. RNN models different speakers on time.

Results

We picked 14 students and manually annotated their faces in videos. When a single face tracklet is provided as a query, the task is to retrieve the faces belong to the same identity in evaluation. The receiver operating curve of face verification is around 95%. Despite the high performance, some faces could not be spotted in our anonymization. This can be due to the face detector's failure, mostly in motion blur situations or occlusion, or in face identification. The best option would be to use automated anonymization to speed up the procedure instead of substituting the manual task. In this way, it would be possible to verify and find rare failures in a short time.

4 Discussion

The papers within the scope of this dissertation were summarized in Chapter 3. The main contribution of this dissertation is the computational framework of multimodal visual sensing. We can present the main contributions as follows:

1. **Students (listeners).** Estimating student engagement from facial videos has a wide range of applications in educational settings. So far, classroom studies using computer vision and machine learning is limited. Most of the studies in the literature either focus on only the movements of students or aimed at estimating students' intensities of engagement but were lacking a psychologically valid measurement instrument and in limited settings. Our studies first proposed a new instrument to rate student engagement continuously (in one-second intervals), investigated the validity of engagement ratings, and showed that computer vision methodologies could predict engagement intensities from facial videos. In the second study, we conducted a large-scale classroom study at a secondary school. In this study, we further proposed a novel approach to learn Attention-Net and Affect-Net features from faces and transfer for engagement estimation. Besides, we showed that only 1-minute person-specific data could yield an average AUC improvement of .084 in engagement classifiers.
2. **Students (presenters).** In contrast to student engagement estimation during teacher-based instruction where students were mostly listeners, another essential situation is speakers' behaviors. We used the audiovisual recordings of a national contest, 3-minute scientific presentations by students, and investigated different nonverbal behavioral cues to estimate presentation competence. Previous studies that involve automated methods in presentation competence had a global competence item. Instead, we represented presentation competence using Tübingen Instrument for Presentation Competence (TIP) and by aggregating the ratings of body language and speech items separately. Speech features were the best performing. Feature-level and score-level fusion of three modalities, speech, face, and body pose, further helped in classification and regression of presentation competence. Our study is the first that investigated presentation competence on real-world data instead of acted settings. Furthermore,

we investigated the effect of different rhetorical settings and the transferability of automated methods.

3. **Teachers.** In addition to students' nonverbal behavior analysis, engagement, and presentation competence, understanding teachers' perception also tells much about classroom instruction. For teachers, the ability to distribute attention to relevant information in the complexity of classroom interaction and equally across students is very decisive for effective teaching. We developed an approach to combine face processing in an egocentric field camera view and mobile eye tracker's gaze points to create attention maps. Teachers' attention maps summarize when and whom they gazed at during the class without requiring any region of interest annotation. Our computational approach on mobile eye tracking data gave the same results that were acquired by manual area of interest labeling in months, within a few hours. This paved the way for mobile eye trackers to understand teachers' attentional processes by creating more detailed analytics without any labeling efforts.
4. **Automated anonymization.** In educational studies involving audiovisual recordings, the current data protection and privacy regulations and ethical considerations are of utmost importance. In practice, there is a demand to use automated methods to speed up the anonymization of participants appearing in video footage. We create a small-scale classroom observation dataset comprised of 6,5 hours of instruction in varying subjects and investigated both faces in videos using pre-trained face embeddings and speech segments in audio recordings using speaker diarization (the Unbounded Interleaved-State Recurrent Neural Networks, UIS-RNN). The output of automatic anonymization is provided to human observers to inspect and find any missing samples in case of failure.
5. **Joint attentional processes.** In contrast to engagement or presentation competence estimation, there are applications of multimodal visual sensing beyond single-person analysis. Joint attention refers to the situations of two or more persons looking at the same point. It can happen in dyadic (looking at each other) or triadic ways (looking at each other and an object). We proposed a novel approach to detect and localize joint attention in social scenes without any face detection or complicated processing. Our method's essential part was to use saliency information in videos to estimate saliency augmented pseudo-attention maps to estimate face and joint attention likelihood. Such an automated system that estimates joint attention contributes to the visual sensing in small meetings and therapy of children with attention disorders.

Overall, we can review this dissertation's contribution from two aspects: *practical* and *algorithmic*. The practical aspect is that we validated our methods on real-world and large-scale data, various classroom observation studies using field cameras and mobile eye trackers, and also student presentations. The algorithmic aspect can be summarized as designing deep embeddings for attention and affect features, uncertainty-based personalization strategies in engagement estimation, feature fusion in both engagement and presentation competence estimation, combining mobile eye tracking and egocentric computer vision, the use of saliency

information for joint attention.

4.1 Multimodal Visual Sensing

This dissertation proposed a framework, multimodal visual sensing, using multiple nonverbal behavioral cues to understand the attentional processes in educational settings. This section will discuss the findings of engagement and presentation competence estimation.

4.1.1 Estimating Student Engagement

Student attention is key to successful teaching and learning and also an essential dimension of engagement. There are many objects and entities around us. Given a task at hand, some parts of the world can be more relevant to the task, whereas others are irrelevant or may be disruptive. Besides, our brain is limited in resources to concurrently process the endless amount of sensory data and execute particular behaviors. Thus, there are attention mechanisms to help our brain focus a part of the sensory data on realizing the desired task.

Looking into students' behavior in the classroom, as attention is a filtering mechanism, students should be focused on learning-related tasks and spend less time on other tasks. There is an overt and covert aspect of attention; therefore, it is possible to judge the attentiveness of a student in the classroom by looking at only overt cues. Eye movements are the primary sources of information to evaluate attentiveness.

In more static situations such as computer-based activities or driving, cognitive load using eye-tracking and pupillometry is the right way to measure attention. Due to the nature of these situations, capturing eye images in good quality is possible. Pupil diameter can be used to monitor the chance in attentional processes. In classroom instruction, the visual focus of attention can move. For instance, a teacher who writes and points out something on the board can walk inside the classroom. Sometimes, gazing into a presentation material can be considered learning-related behavior, and other times, taking notes may be needed. As the focus is physically moving, the most reliable and non-invasive methods to understand attention are head pose and gaze direction.

We complemented attentional features with affective features such as facial expressions and affective acoustic audio parameters to better understand visible engagement. Our contributions to student engagement analysis are two-sided. On one side, we aimed to show the applicability of automated, computer vision, and machine learning in the classroom. Another side is the algorithmic contributions.

The majority of the computer science literature on student engagement was on computer-based learning [110, 111, 112, 113]. The underlying reasons were principally better data quality (i.e., using webcams in the approximately 1-meter distance) and the availability of other log data to measure performance or attentiveness. However, for most of our educational life, we

learn in the classroom, and the use of classroom analytics plays a vital role in making learning efficient from lower grades.

The literature in classroom analytics, for instance, mainly the works of Raca [114], were the early works that investigated the relationship between self-reported attention and some fundamental features that were extracted from visual modalities. Some examples are optical flow analysis of students' motion, the ratio of face detection (as a precursor of gaze contact to teacher area), and head pose. All these studies showed the classroom analytics as a promising research field, but their contribution stayed limited to correlational analysis. Furthermore, their analysis was based on short intervals before the students answered a questionnaire measuring attention a few times during the instruction.

Our initial research question was to what extent we could estimate students' engagement in the classroom. If we can estimate reliably, which set of visual features are required for this task? We investigated the answers of these questions in [115].

There are two major performance criteria to measure the performance of automated methods. The first is how accurate they can classify or regress manually annotated engagement labels (depending on how the problem is formulated) and the relationship between the estimated engagement intensities and self-report measures.

In this study, we used approximately 40-minute video recordings of 30 students in three sessions where the teacher covered the same topic in an undergraduate-level seminar. We formulated the problem as a regression task and compared a set of features using the same SVM-based regressor. Those modalities are head pose, gaze direction, facial expressions (action unit intensities), a feature-level fusion of head pose and gaze, and ultimately the fusion of all features. We compared both mean squared errors and correlation coefficient with ground truth labels. Our focus was to predict a single, accurate coefficient summarizing a student's engagement level during the entire instruction period in this study.

Students' distance from the camera that is located next to the whiteboard and teacher area varies from 3 to 10 meters. Thus, gaze estimation may yield very unreliable results. Despite the noisy predictions, we observed that gaze estimation outperforms only head pose features. Their Pearson correlations are .29 and .44 for head pose and gaze, respectively. Facial expressions also work on par with gaze features. Even though raters take into account the attentional cues first, it seems facial expressions of engagement and disengagement can be equally important.

When we process videos offline, or very low latency is not a priority, the fusion of several features can potentially improve engagement estimators' performance. Both the fusion of head pose and gaze and all three modalities improve single feature-based regression models. Furthermore, the correlation with the ground truth labels reached to .61. Moreover, we tested the effect of immediate neighbors' behavior on a student's engagement. It is first proposed in [116], and they performed a correlation study on the effect of synchronization. We used the

cosine similarity of attentional and emotional features of neighboring students and added this as a new set of features. In this way, the best correlation result went up to .71 ($p < 0.01$).

Looking into the second performance metric, the relationship between estimated engagement intensities and post-tests, the problem becomes more complicated because even the correlations of manual engagement ratings show .53, .62, and .63 for situational interest, cognitive engagement, and involvement. Knowledge test does not show any significant correlation with manual labels and automated methods. Our best models using a fusion of all features showed the correlations of .26, .39, and .43 with these post-test items.

From an automated human behavior understanding perspective, it is a very challenging task to estimate a person's questionnaire answers from her or his long videotaped behaviors. The performance of regression (or classification) models in different intensities are not the same, and the data is highly imbalanced. Besides, averaging a longer sequence to retrieve a global score accumulates error in time. Even though engagement raters are trained according to a special curriculum and performed higher ICC agreement on a training set before starting the actual data labeling, we cannot discard the raters' subjectivity and bias against specific visual traits. In summary, considering those points and evaluating models in shorted intervals (preferably with respect to observer rater measures) instead of global, self-reported measures will be a better way to improve engagement estimators.

Considering the potential and limitations of automated engagement analysis on our pilot study, we performed our second study on a larger-scale at schools. Instead of the same teaching curriculum and class, we conducted a longer period with different classes and grades in a secondary school. This time, we performed engagement analysis as classification in three scales: low $[-2, 0.35]$, medium $[0.35, 0.65]$, and high engagement $[0.65, 2.0]$.

In our first study [115], our approach was to extract low-dimensional features and train shallow regressors on top of them. Despite the use of data-driven methods in feature extraction, this approach was discarding the power of big data. In classroom study [117], we used deep embeddings for attention and emotion features. The main contribution of this approach is two-folded. The first is that deep embeddings require only a rough alignment and works in case of faces detected, whereas, in the previous approach, misaligned data were needed to be discarded. In this way, we can use the majority of video data in our analysis. The second is that the last layers of deep representations embed the data points in a lower embedding where small changes locally made meaningful chances in label space. This also makes predictions temporally more consistent.

We tested the feature embeddings of Attention-Net and Affect-Net branches trained to estimate head pose and classify facial expressions, respectively. The classifiers are SVM (with linear and rbf kernels), RF, MLP, LSTM. The general trend is that attention features performed $3-4\%$ better than affect features. The contribution of temporal models is limited, and we argue that this was due to having second-length manual labels. Another reason can be due to the limited sample size of the dataset prevents learning generalizable classifiers using deep learning.

Another contribution of our second study on engagement is the personalization of models. The majority of computer vision and machine learning models discard personalization. Even when it is desired, it is challenging to personalize end-to-end deep learning models. We applied an uncertainty-based batch sampling method to sample person-specific data actively. In this way, in all subjects, by requesting less than 10% of their data, we showed $10-15\%$ improvement in accuracy and F1-scores.

4.2 Presentation Competence

Presentation competence is another essential application of multimodal visual sensing. An automated system that can estimate presentation competence can be beneficial for self-reflection and competence in time.

The main difference of presentation from engagement analysis is that it contains nonverbal features changing faster than listening to a teacher that is comparatively a passive activity. Thus, even e second of presentation data is valuable, and it is more difficult to aggregate in time, for instance, an hour-length presentation. On this problem, our focus was a 3-minute scientific presentation of students acquired within the scope of a national contest named Youth Presents.

In presentation competence, there is a line of work in educational psychology and rhetorical aspects aiming to develop an objective and reliable scale to measure competence (for more details, see Section 2.2 and Table 2.2). On the other hand, studies with automated methods to estimate presentation competence used different audiovisual modalities; however, they lack a psychologically reliable measure as a ground truth.

The visual modalities that we used are facial features (head pose, gaze, and facial action unit intensities) and body features (statistical encoding of two-dimensional body joints). Additionally, we used affective audio features, the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [118], which comprises 88 features related to the audio signal. In presentation competence and similar affective computing tasks, affective audio parameters perform better than visual modalities. Thus, we took these features as a reference. The task, similar to the engagement task, is weakly supervised as it contains only video-level features.

There were three critical questions that could be answered as a result of this study. First, which modality performs better to estimate presentation competence? Even though our focus was mainly visual modalities, we observed that speech features were still beyond the face of body pose features in explaining a presentation's competence. In the same dataset results, face and body pose were comparable. However, we found that pose features were more successful in generalizing to different tasks than facial features. Particularly the use of local features in body pose (aggregated in 16 seconds). This result indicates that a more fine-grained annotation of prototypical postures might give better-performing presentation competence models in future work.

The second question was the generalization of predictive models in a person-independent setting. Our results in the T1 set reached an accuracy of 71.25% (with an F1-score of 74.54%) in classification. Similarly, the best performing feature-level fusion yielded an MSE of 0.08 and a Pearson correlation of 0.61 in the regression task. These results indicated that our approach that estimates presentation competence could work well in different subjects.

The third question was the independence from the rhetorical settings, for instance, the choice of topic (free or given) or preparation time. The T2 set is composed of a subset of students in the T1 set. The main difference is that they were assigned the presentation topic and given a limited preparation time and the same material. Our models trained on the T1 (using global features) could not generalize on the T2 set. On the other hand, the use of short 16-second sequences (local features) improved the performance of global body pose features that is slightly above the chance level up to 79.22% and 88.31% in the accuracy and F1-score, respectively. We should note that keeping the rhetorical setting as much as possible similar between training and test partitions was a vital constituent for better prediction models.

4.3 Other Related Applications of Multimodal Visual Sensing

Our contributions are not limited to engagement and presentation competence analysis. Beyond them, we investigated joint attention analysis [119], a framework to combine mobile eye-tracking and computer vision to analyze teachers' perception [120], and automated anonymization of audio and visual data in classroom studies [121].

Our approach [119] to estimate joint attention requires only RGB images and outputs face and joint attention likelihood maps. Such a system can be used as a therapeutic tool in the treatment of deficits. It can support a multimedia analysis tool to search and retrieve content that has similar social formations. Beyond these practical contributions, our study leveraged general-purpose saliency estimation models and improved their representation ability for social scenes.

Moving from student and third-person analysis, the first-person view also contains valuable information. Teachers' use of mobile eye-trackers is a pervasive approach to understanding teachers' attentional distribution in the classroom. Commercial eye-trackers do not perform computer vision and detailed attention representations. In [120], we approached this problem and combined face processing pipeline (face detection, tracking, and recognition) in the eye-tracker's field camera and gaze points. The majority of mobile eye-tracker users are non-technical persons, for instance, psychology, education, and behavior researchers, and this functionality decrease manual efforts to a great extent.

Another problem that we faced when storing and processing sensitive audio and visual data is privacy considerations. Our preliminary evaluation [121] showed that the anonymization of faces and voices could be done using computer vision and audio processing. As there is no guarantee of computer vision and machine learning systems perform without error. Thus,

instead of entirely relying on automated methods, these approaches can be used together with manual inspection and reduce human efforts.

4.4 Outlook & Future Research Directions

The works presented in this dissertation aimed to approach human behavior understanding problems under a framework of multimodal visual sensing. The main problems were exploring engagement and presentation competence, but not limited to them. We further investigated the following problems: (i) attentional processes of several persons (joint attention), (ii) a study bridging egocentric vision and mobile eye tracking, and (iii) automated anonymization of sensitive audiovisual data.

Affective computing applications, mainly using computer vision, are widespread and developed. However, automated methods to understand cognitive processes from visible nonverbal behavioral cues are quite limited. Our main contribution is to investigate the abilities and limitations of computer vision and machine learning in educational settings. In all studies within the scope of this dissertation, we processed the available data collected in previous educational studies, designed and collected our audiovisual recordings at university and secondary school, or used publicly available datasets and benchmarks.

Labeling. In engagement and presentation competence estimation, even though there are instrument metrics in education and psychology, computational studies lacked reliable measurement tools. In engagement estimation, we developed a novel measurement tool by combining the ICAP framework and on-task/off-task behavior observation systems. We validated this new tool initially at university-level seminars and subsequently in a large-scale study at a secondary school in Germany. Similarly, in the presentation competence estimation study, we adopted a recently proposed Tübingen Instrument for Presentation Competence (TIP). In both topics, annotations were done by two or more annotators with an agreement higher than .60 of intraclass correlation.

Learning with a limited number of subjects. In typical computer vision problems, i.e., image categorization, object detection, face recognition, or body pose estimation, the benchmark contains images or videos from hundreds of thousands of different resources. In contrast, the data source in affective computing and human-computer interaction is in the levels of tens. This situation is the biggest obstacle to learn end-to-end representations from the raw image data.

Considering the difficulty of learning with a limited number of subjects, we focused on extracting low-level nonverbal behavioral features. For instance, a deep learning model trained with image data belonging to 5-6 subjects tends to overfit the data. On the other hand, nonverbal features such as head pose, gaze, facial expressions, and body posture could easily generalize

to different subjects (Appendix A.1 and B). In this way, the feature extraction workflows can be improved in the future; however, our work that depends on these features will still be usable.

Both data collection and labeling in engagement were time-consuming and difficult. Furthermore, ethical considerations and data protection and privacy regulations make the design and deployment of these studies longer. We proposed novel methods, Attention-Net and Affect-Net, to learn attention and affect features from facial images (Appendix A.2). As we trained these deep embeddings on more diverse datasets, they showed outstanding performance on engagement estimation by using shallow readout classifiers.

Less supervision. In estimating joint attention, even in a more diverse, multimedia dataset (VideoCoAtt), the performance of end-to-end deep learning approaches was limited. By utilizing the saliency maps of scenes to encode face and joint attention likelihood only in the training phase, our method (Appendix C.1) outperformed previous approaches that require more supervision (i.e., face detection, LSTM/RNN models). Models estimating joint attention can be used in many educational situations, and end-to-end estimation by using a single network is an essential advantage for fast deployment.

Fusion and personalization. In both engagement and presentation competence, feature-fusion and classifier fusion strategies improved the best performing single modality performance. In the second phase of the engagement study (Appendix A.2), as we had a large amount of recording, we investigated personalized engagement classifiers. From a psychological perspective, it is known that there were variations in the visible cues of (dis)engagement. Only 60 seconds of samples selected by a margin-based uncertainty rule could show a significant improvement in classifier performance. This limited amount of person-specific data was enough to adapt base classifiers.

Mobile eye tracking. Mobile eye tracking is an innovative approach and can be used in attention and engagement studies. Considering the education settings, using mobile eye trackers in the classroom for each student would be very expensive. On the other hand, a single eye tracker worn by a teacher can capture their attentional processes. Mobile eye tracking was previously used to investigate teachers' attentional processes by manually labeling the gaze points into categories, for instance, per student. As the egocentric field of view changes in time, creating attention maps from mobile eye tracking data can be very time-consuming. Our study was the first that investigated face processing from the field camera of the eye tracker and mapped them to the students in the classroom.

Automated anonymization and privacy-aware computing. Motivating from the need to anonymize audiovisual data in classroom studies, we also investigated the feasibility of face processing and audio diarization approaches in classroom observation videos. Our approach

provided an opportunity to anonymize a selected identity in video and audio recordings.

4.4.1 Future Directions

In the general framework of multimodal visual sensing, our focus was educational applications. As the large-scale applications of visual sensing in educational analytics were limited, we examined algorithms' potential and limitations with this dissertation's scope. By following the proposed approaches, the deployment of our approach as a part of the cognitive and behavioral interface would be a future work. In engagement analysis, a user interface can summarize the engagement intensities and interactions of students. In contrast, presentation competence tools can be used for students to practice presentations themselves and learn from their mistakes and improve their competence. In both use cases, the most decisive point will be to study and validate the positive impact of these tools.

The current practices in educational analytics prevent making the datasets publicly available. As a follow-up of our automated anonymization study, the use of deep generative models (generative adversarial networks, variational autoencoders, and flow-based methods) can be used to swap faces with fictitious identities and perturb the biometric traits of speech features. This can be a possible road towards sharing video data. Alternatively, nonverbal behavioral cues are at the focus of multimodal visual sensing. Differential privacy approaches can be alternative to remove private features and store and share only non-identifying behavioral features.

A Engagement Estimation

This chapter is based on the following articles:

- P. Goldberg, Ö. Sümer, K. Stürmer, W. Wagner, R. Göllner, P. Gerjets, E. Kasneci, and U. Trautwein, "Attentive or not?: Toward a machine learning approach to assessing students' visible engagement in classroom instruction," *Educational Psychology Review*, 2019. https://doi.org/10.1007/s10648-019-09514-z
- Ö. Sümer, P. Goldberg, S. D'Mello, P. Gerjets, U. Trautwein, and E. Kasneci, "Multimodal Engagement Analysis from Facial Videos in the Classroom," Manuscript submitted for publication (*IEEE Trans. on Affective Computing*), 2020.

A.1 Attentive or not?: Toward a machine learning approach to assessing students' visible engagement in classroom instruction

Abstract

Teachers must be able to monitor students' behavior and identify valid cues in order to draw conclusions about students' actual engagement in learning activities. Teacher training can support (inexperienced) teachers in developing these skills by using videotaped teaching to highlight which indicators should be considered. However, this supposes that

- valid indicators of students' engagement in learning are known, *and*
- work with videos is designed as effectively as possible to reduce the effort involved in manual coding procedures and in examining videos.

One avenue for addressing these issues is to utilize the technological advances made in recent years in fields such as machine learning to improve the analysis of classroom videos. Assessing students' attention-related processes through visible indicators of (dis) engagement in learning might become more effective if automated analyses can be employed. Thus, in the present study, we validated a new manual rating approach and provided a proof of concept for a machine vision-based approach evaluated on pilot classroom recordings of three lessons with university students. The manual rating system was significantly correlated with self-reported cognitive engagement, involvement, and situational interest and predicted performance on a subsequent knowledge test. The machine vision-based approach, which was based on gaze, head pose, and facial expressions, provided good estimations of the manual ratings. Adding a synchrony feature to the automated analysis improved correlations with the manual ratings as well as the prediction of posttest variables. The discussion focuses on challenges and important next steps in bringing the automated analysis of engagement to the classroom.

Cognitive activation, classroom management, and teacher support are the three central tenants of teaching quality [84, 85]. The level of students' (dis)engagement in learning activities can be considered a major indicator of both cognitive activation and classroom management because it signals students' engagement in the deep processing of learning content and reveals the time on task [86] provided by the teachers for students' learning. To this end, teachers are required to take note of their students' attentional focus and make sure the students are engaging in the desired learning activities. Thus, the ability to monitor students' attention and to keep it at a high level is part of the competencies that novice teachers need to acquire. However, research has indicated that teachers might not always be aware of their students' attentional focus, and this may be particularly true for novice teachers.

In general, beginning teachers have trouble monitoring all students in the classroom evenly and noticing events that are relevant for student learning [87, 88, 89, 90]. Therefore, teacher training needs to support future teachers in developing the necessary knowledge structures that underlie these abilities (e.g., [122]). Consequently, providing an improved measurement

approach for student attention will be beneficial for research and can potentially contribute to teacher training. Research has already demonstrated that both inexperienced and experienced teachers' ability to notice relevant cues in the classroom benefits from observing and reflecting on their own videotaped teaching [91, 92]. Until now, however, instructors have typically had to watch hours of video material to select the most crucial phases of lessons. Similarly, when it comes to research on teaching effectiveness and the development of teachers' ability to notice relevant cues in classroom instruction (i.e., professional vision skills), researchers typically have to invest considerable resources, especially coding resources, to examine the association between teacher behavior and classroom processes [123]. The required effort further increases when investigating students' attention across an entire lesson and analyzing attention at the group level instead of among individuals. In this vein, attention- and engagement-related behavior during classroom instruction has rarely been studied due to the difficulty of data collection and labeling. However, learners might behave differently in naturalistic settings and show versatile behavior that cannot be found in a lab.

One potentially valuable avenue for addressing these issues is to utilize the technological advances made in recent years in fields such as computer vision and machine learning. Therefore, in an ongoing research project [124], we have been investigating whether and how the automated assessment of students' attention levels can be used as an indicator of their active engagement in learning. This automated assessment can in turn be used to report relevant cues back to the teacher, either simultaneously or by identifying and discussing the most relevant classroom situations (e.g., a situation where students' attention increases or decreases significantly) after a lesson.

In the present study, we present a proof of concept for such a machine vision-based approach by using manual ratings of visible indicators of students' (dis)engagement in learning as a basis for the automated analysis of pilot classroom recordings of three lessons with university students. More specifically, by combining multiple indicators from previous research (i.e., [12, 13, 14]), we developed a manual rating instrument to continuously measure students' observable behavior. In addition, we performed an automated analysis of the video recordings to extract features of the students' head pose, gaze direction, and facial expressions using modern computer vision techniques. Using these automatically extracted features, we aimed to estimate manually annotated attention levels for each student. Because we had continuous labeling, this could be done by training a regressor between the visible features and the manual labels. We investigated the predictive power of both the manual and automatic analyses for learning (i.e., performance on a subsequent knowledge test). To account for complexity within classrooms and enrich the automated analysis, we also considered synchronous behavior among neighboring students. In the present article, we report initial empirical evidence on the reliability and validity of our automated assessments and their association with student performance.

A.1.1 Attention in Classroom Instruction

Student attention is a key construct in research on both teaching and learning. However, definitions vary widely and are discussed from multiple perspectives. Here, we focus on describing three lines of research that inspired our research program: cognitive psychology models that describe attention as part of information processing, engagement models in which attention makes up part of a behavioral component, and teaching quality models in which student attention is a crucial factor.

In current models in the psychology of learning, attention denotes a filtering mechanism that determines the kind and amount of information that enters working memory [125]. This mechanism is crucial for preventing working memory overload and allows the learner to focus on the right kind of information. Only sensory information that enters working memory is encoded, organized, and linked to already existing knowledge. Thus, attention serves as a selection process for all incoming sensory information as it dictates which pieces of information will be processed further and will get the chance to be learned. Thus, attention determines the success of knowledge construction [126]. Engle [127] further proposed that executive attention, which actively maintains or suppresses current representations in working memory, is part of working memory. Certain instructional situations strongly depend on executive processes such as shifting, inhibition, or updating [128] and thus necessitate top-down attentional control. Although information processing occurs in a covert manner, some aspects of attentional processes are likely to be observed from the outside: for example, visually orienting toward a certain stimulus, which improves processing efficiency [129].

Attention is often mistaken for engagement, even though it constitutes only part of it. Engagement is defined as a multidimensional meta-construct and represents one of the key elements for learning and academic success [2]. It includes observable behaviors, internal cognitions, and emotions. Covert processes such as investment in learning, the effort expended to comprehend complex information, and information processing form part of cognitive engagement [2, 130]. Emotional engagement in the classroom includes affective reactions such as excitement, boredom, curiosity, and anger [131, 2]. Attention is considered a component of behavioral engagement alongside overt participation, positive conduct, and persistence [131, 2]. Per definition, cognitive engagement refers to internal processes, whereas only the emotional and behavioral components are manifested in visible cues. Nevertheless, all engagement elements are highly interrelated and do not occur in isolation [2]. Thus, attention plays a crucial role because it may signal certain learning-related processes that should become salient in students' behavior to some extent.

Learners' attention also plays a crucial role in research on teaching. Teachers must determine whether their students are attentive by considering visible cues, continually monitoring the course of events in order to manage the classroom successfully [105] and providing ambitious learning opportunities. A student's attention or lack thereof (e.g., when distracted or engaging in mind wandering) can signal whether she or he is on-task or off-task. This in turn can provide

hints about instructional quality and the teacher's ability to engage his or her students in the required learning activities. Thus, it is important to help teachers develop the skills needed to monitor and support student attention and engagement and adapt their teaching methods. Consequently, accounting for student attention and more broadly student engagement in teaching is considered crucial for ensuring teaching quality, including classroom management, cognitive activation, and instructional support [84, 132].

In sum, the definitions, theoretical backgrounds, and terminology used in various lines of research to describe observable aspects of students' cognitive, affective, or behavioral attention/engagement in learning are diverse, but experts agree on their importance and key role in learning. As teachers must rely on visible cues to judge their students' current attention levels [133, 134], we focused on observable aspects of attention and inferences that were based on visible indicators. In the remainder of the article, we use the term visible indicators of (dis)engagement in learning to describe these aspects. These visible indicators are highly likely to be associated with learning, but this assumption needs to be validated.

A.1.2 Previous Approaches for Measuring Visible Indicators of Engagement in Learning

The difficulty in assessing students' engagement-related processes in real-world classroom settings consists of externalizing learners' internal (covert) states through visible overt aspects to the greatest extent possible. In psychology, affective states and cognitive processes such as attentional control are usually determined from physiological signals, such as heart rate, electrodermal activity, eye tracking, or electroencephalography (e.g., [135, 136, 137, 138]). Using this kind of psychologically sound measurements makes it possible to detect covert aspects of learning-related processes; however, these measures are hardly feasible in classroom instruction, especially when teachers must be equipped with knowledge about what indicators to look for in students. Furthermore, these approaches are useful for answering very specific research questions. However, they are not sufficient for determining whether students' ongoing processes are actually the most appropriate for the situation. By contrast, overt behavior can provide visible indicators of appropriate learning-related processes in students.

Overt classroom behavior is an important determinant of academic achievement [139, 140]. Although overt behavior does not always represent a reliable indicator of covert mental processes, previous findings have demonstrated a link between cognitive activity and behavioral activity [141]. Previous studies have analyzed students' behavior and have determined its relation to achievement [13, 14, 142, 143]. Furthermore, in research on engagement, correlations between student engagement and academic achievement have been found [144]. Other studies have found opposing results (e.g., [145]); however, these studies either relied on self-reports as opposed to observer ratings or only focused on certain facets of engagement-related behavior (e.g., only active on-task behavior).

There have been various attempts to systematically assess visible indicators of engagement in classroom learning, for example, Helmke and Renkl [13] based their research on an idea by Ehrhardt et al. [146] and related observable student behavior to internal processes using time-on-task as an indicator of whether a student was paying attention to classroom-related content. Assessing observable content-related behavior is essential to this operationalization of higher order attention. Hommel [14] modified this approach and applied it to the video-based analysis of instructional situations. Rating behavior as either on- or off-task with varying subcategories demonstrated the interrelation between visual cues and achievement or reduced learning [147, 13].

However, learners can differ in their learning activities but still be engaged in a certain task. The ICAP framework proposed by Chi and Wylie [12] distinguishes between passive, active, constructive, and interactive overt behavior, which differ across various cognitive engagement activities. This framework focuses on the amount of cognitive engagement, which can be detected from the way students engage with learning materials and tasks [12]. This theoretical model provides a promising approach for further expanding the different types of on-task behavior so that variations in student behavior can be accounted for.

In sum, considering learning content has been shown to be useful; however, there is a lack of research involving the continuous analysis of attention or engagement over the course of one or more lessons. A unique feature of the present study is that we aimed to acquire a continuous assessment (i.e., a score for every student in the classroom for every second of instruction time) of students' visible indicators of (dis)engagement in learning. This temporal resolution was crucial in our approach because we aimed to provide comparable data that could be used to train a machine-learning algorithm. To reach this high level of temporal resolution, we decided to annotate learners' behavior continuously. The free software CARMA [148] enables the continuous interpersonal behavior annotation by using joysticks (see Lizdek et al. [149]). However, this new approach limited us in terms of using already existing rating instruments because existing instruments do not allow for a high enough level of temporal resolution. Furthermore, the CARMA software requires annotations on a scale rather than rating the behavior in terms of categories as already existing instruments do. When developing the new instrument, we mainly oriented on the MAI [13, 14] However, we needed to define more fine-grained indicators of student behavior to make annotations along a continuous scale possible. Therefore, we added indicators from various established instruments to extend our rating scale. We assumed that the manual observer annotations would serve only as approximations of the actual cognitive states of the students and that the averaged (i.e., intersubjective) manual annotations would reflect the "true score" of the visible indicators of (dis)engagement in learning better than a single rater could. Subsequent to the ratings, we thus calculated the mean of the raters for every second. The mean values for each second and student were used as the ground truth to train a machine-learning approach.

A.1.3 Using Machine Learning to Assess Visible Indicators of (Dis)Engagement in Learning

Machine learning and computer vision methods have made tremendous progress over the past decade and have been successfully employed in various applications. In the context of teaching, these methods might offer an efficient way to measure student engagement, thereby decreasing the need for human rating efforts. However, any machine-learning method that is aimed at estimating covert engagement-related processes in learning needs to depend on visible indicators such as head pose, gaze direction, facial action unit intensity, or body pose and gestures. State-of-the-art methodologies for the automated assessment of engagement can be divided into two categories: single-person- and classroom-based analyses.

In a single-person analysis, facial expressions can provide hints about ongoing cognitive processes and can be analyzed by considering action unit (AU) features. Related studies by Grafsgaard et al. [110] and Bosch et al. [150, 151] investigated the relations between AU features and several response items and affective states. Even though these studies found that several facial AUs were associated with engagement, they were limited to affective features and did not consider head pose or gaze direction.

In another work, Whitehill et al. [111] introduced a facial analysis approach to estimating the level of engagement on the basis of manually rated engagement levels. Although their facial analysis approach was able to predict learning just as accurately as participants' pretest scores could, the correlation between engagement and learning was moderate due to the limited amount of data and the short-term nature of the situations.

In a classroom-based analysis, the focus shifts away from single individuals onto shared features and interactions among participants. In this context, a number of notable contributions (e.g., [114, 152]) have utilized various sources of information to understand features of audience behavior, such as the amount of estimated movement and synchronized motions among neighboring students. They found that immediate neighbors had a significant influence on a student's attention, whereas students' motion was not directly connected with reported attention levels [116, 152]. Furthermore, Raca et al. [153] analyzed students' reaction time upon presentation of relevant information (sleeper's lag). In addition to estimating head pose, they considered the class period, student's row, how often faces were automatically detected (as a precursor to eye contact), head movement, and the amount of still time (i.e., 5-s periods without head movement) because these features had previously been shown to be good predictors of engagement in learning [154]. Although these results were promising, they were limited to correlational studies of reported attention levels; predictive approaches were not used due to limits in the performance of computer vision methodology.

A recent study estimated human-annotated attention levels by using 3D vision cameras to identify individuals using face and motion recognition without any physical connection to people and solely on the basis of visual features [155, 156]. Due to technological limitations associated with 3D vision cameras, the analysis was based on a single row of students rather

than the entire classroom. Fujii et al. [157] used head-up and head-down states and classroom synchronization in terms of head pose as informative tools that could provide feedback to teachers. However, they did not validate their system using educational measures (pretests, posttests, or observations) and only reported user experiences with three teachers.

In sum, few previous studies have investigated classroom-based attention and engagement beyond the single-person context due to the poor performance of computer vision approaches for face and body pose recognition in unconstrained settings (e.g., varying illumination, occlusion, motion, challenging poses, low resolution, and long distance). However, recent advances in deep learning technology have resulted in the availability of new methods for the robust extraction of such features from videos. By employing such technology in this study, we aim to bring a fine-scaled analysis of visible indicators to classroom studies and augment individual engagement analysis with another useful feature: classroom synchronization.

A.1.4 Research Questions

The present study is part of an ongoing research project in which researchers from education science, psychology, and computer science are working to create an automatic assessment of students' engagement that could one day be implemented in an interface that can be used for research as well as teacher training purposes. The present study lays the basis for achieving these goals by developing and testing an automated approach to assessing visible indicators of students' (dis)engagement in learning. Such a remote approach requires comparable data (generated by human raters) that can be used as the ground truth in order to train a classifier. However, existing instruments [13, 14] for measuring engagement-related processes in learning (a) require human observers to make a huge number of inferences and (b) require data to be collected in 30-s or 5-min intervals. This is problematic for our context because an automated analysis can only rely on visible indicators, does not consider content-specific information at all, and operates at a more fine-grained temporal resolution. Therefore, we developed a new instrument to annotate student behavior manually by applying a rating method with visible indicators over time. This manual rating served as the starting point from which to train an algorithm by applying methods from machine learning and computer vision.

The present study addressed the following research questions:

1. Is the new manual annotation of visible indicators of (dis)engagement in learning related to students' learning processes and outcomes? To validate our instrument, we examined how the manual ratings were correlated with students' self-reported cognitive engagement, involvement, and situational interest. We expected these self-reported learning activities to cover different facets of (dis)engagement in learning, and when combined, we expected them to account for cognitive parts of the construct. Furthermore, we tested whether the scores resulting from the manual annotation would predict students' performance on a knowledge test at the end of an instructional session.

2. Is it possible to adequately replicate the relation to students' learning processes and outcomes by using visible indicators of (dis)engagement in learning based on the machine-learning techniques that estimated the manual ratings? We used gaze, head posture, and facial expressions to estimate the manual ratings. To test the quality of our machine vision-based approach, we examined the associations between the scores generated from the automated approach and the manual ratings and students' self-report data regarding their learning processes, and we used the machine-learning scores to predict achievement on the knowledge test.
3. How do adding synchrony aspects of student behavior affect the automated estimations of the manual ratings? The results of previous studies have indicated that immediate neighbors have a significant influence on a student's engagement [116, 152]. As a first step toward including indicators of synchrony in our project, we added students' synchrony with the person sitting next to them as an additional variable to our prediction models, which were based on the automated assessment of student engagement.

A.1.5 Method

The ethics committee from the Leibniz-Institut für Wissensmedien in Tübingen approved our study procedures (approval #2018-017), and all participants gave written consent to be videotaped.

Sample and Procedure

We decided to conduct a study involving university students in order to validate our approach before administering it in school classrooms. A total of $N = 52$ university students (89.5% women, 8.8% men, $mean\ age = 22.33$, $SD = 3.66$) at a German university volunteered to take part in the study. The study was conducted during regular university seminar sessions on quantitative data analysis (90 min). A total of three different seminar groups were assessed. The topics of the sessions were either t tests for independent samples (sessions 1 and 2) or regressions (session 3) and ranged from 30 to 45 min. The sessions were videotaped with three cameras (one teacher camera, two cameras filming the students). If students refused to be videotaped, they were either seated outside the scope of the cameras or switched to a parallel seminar. Participants were informed in advance of the study's purpose, procedure, and ethical considerations such as data protection and anonymization. To avoid confounding effects of the teacher, the same person taught all sessions in a teacher-centered manner. Before the session started, students filled out a questionnaire on background variables (age, gender, final high school examination [Abitur] grade, school type) and individual learning prerequisites. After the session, participants completed a knowledge test on the specific topic of the session and completed another questionnaire about learning activities during the seminar.

Instruments

Individual Learning Prerequisites We used established questionnaire measures to assess three individual learning prerequisites: Dispositional interest in the session's topic was captured with four items (α = .93) adapted from Gaspard et al. (2017). Self-concept in quantitative data analysis was assessed with five items (α = .80; adapted from Marsh et al. 2006), and 13 items were used to test for self-control capacity (α = .83; Bertrams and Dickhäuser 2009). Moreover, we administered the short version of the quantitative subscale (Q3) of the cognitive abilities test (Heller and Perleth 2000). Measuring these learning prerequisites allowed us to control for potential confounding variables in the analyses.

Learning Outcomes The knowledge test consisted of 12 and 11 items that referred to participants' declarative and conceptual knowledge of the session topic, respectively. We z-standardized the knowledge test scores within each group for subsequent analysis.

Self-Reported Learning Activities After the session, we assessed students' involvement (four items, α = .61; Frank 2014), cognitive engagement (six items, α = .79; Rimm-Kaufman et al. 2015), and situational interest (six items, α = .89; Knogler et al. 2015) during the seminar session (Table A.1).

Table A.1: Item wording for learning activities

Construct	Items
Cognitive engagement	I exerted myself as much as possible during the session. I thought about different things during the session. I only paid attention when it was interesting during the session. It was important for me to really understand things during the session. I tried to learn as much as possible during the session. I pondered a lot during the session.
Involvement	During the session... ... I strongly concentrated on the situation. ... I occasionally forgot that I was taking part in a study. ... I was mentally immersed in the situation. ... I was fully engaged with the content.
Situational interest	When you think about today's session... ... the seminar session aroused your curiosity. ... the seminar session attracted your attention. ... you were completely concentrated on the seminar session. ... the seminar session was entertaining for you. ... the seminar session was fun for you. ... the seminar session was exciting for you.

A.1.6 Analysis

Continuous Manual Annotation

To develop a continuous manual annotation that included potential valid indicators of students' visible (dis)engagement in learning, we used the instruments developed by Helmke and Renkl [13] and Hommel [14] as a basis. However, these instruments label behavior in categories and thus cannot be used as a continuous scale. Therefore, we combined the idea of on-/off-task behavior and active/passive subcategories with existing scales from the engagement literature. Furthermore, we used the theoretical assumptions about students' learning processes and related activities in classrooms pointed out by the ICAP framework [12] as an inspiration to define more fine-grained differentiations within the possible behavioral spectrum. The distinction into passive, active, constructive, and interactive behavior allowed us to make subtler distinctions between the different modes of on-task behavior, and this concept could be transferred to off-task behavior (i.e., passive, active, deconstructive, and interactive) as well. By combining different approaches, we could define visible indicators of (dis)engagement in learning on a continuous scale. The resulting scale ranged from −2, indicating interruptive and disturbing off-task behavior, to +2, indicating highly engaged on-task behavior where, for example, learners ask questions and try to explain the content to fellow learners (see Figure A.1). When a person could not be seen or was not present in the classroom, the respective time points were coded as missing values in subsequent analyses.

The behavior of each observed person throughout the instructional session was coded in 1-s steps using the CARMA software (Girard 2014) and a joystick. A total of six raters annotated the videotaped seminar sessions, and each session was annotated by a total of three raters. The raters consisted of student assistants and one researcher, all of whom were trained carefully before annotating the videos. First, raters were introduced to the conceptual idea of the rating and the rating manual. They were told to concentrate on observable behavior to avoid making inferences and considering information from previous ratings. The raters focused on one student at a time in a random order. Every rater had to code one of two specific sections of the video for training, and the raters had to annotate special students who showed different types of behavior. To ensure that we could use all the video material for our analysis, raters who used video section A for training annotated video section B later and vice versa. The respective video sections used for training purposes were not included in the analysis. Only after their annotations reached an interrater reliability with an expert rating of at least $ICC(2,1) = .60$ were raters allowed to annotate the study material. We report the ICC(2,1) here as an indicator of interrater reliability because our data were coded on a metric scale level, and we had more than two raters per participant. We calculated the ICC(2,1) for every student, indicating the interrater reliability averaged across all time points, whereby values between .60 and .74 indicated good interrater reliability (Hallgren 2012); the ICC(2,1) for each student was .65 on average (absolute agreement). When the annotations between the raters deviated strongly, critical situations were discussed among the raters and recoded following consensus. The raters were not informed about the students' individual prerequisites, their learning outcomes,

or their self-reported learning activities.

Machine-Learning Approach

In addition to the manual ratings (see previous section), we employed a machine vision-based approach to estimate (dis)engagement in learning using visible indicators and analyzed the same videos with this approach. More specifically, we first detected the faces in the video (Zhang et al. 2017) and automatically connected the faces detected in the video stream to each student so that we could track their behavior. Faces were aligned, and their representative features extracted automatically based on the OpenFace library [97]. However, this procedure was not applicable to all students and all frames due to occlusions by peers, laptops, or water bottles. The subsequent analyses were therefore based on a subsample of $N = 30$ students.

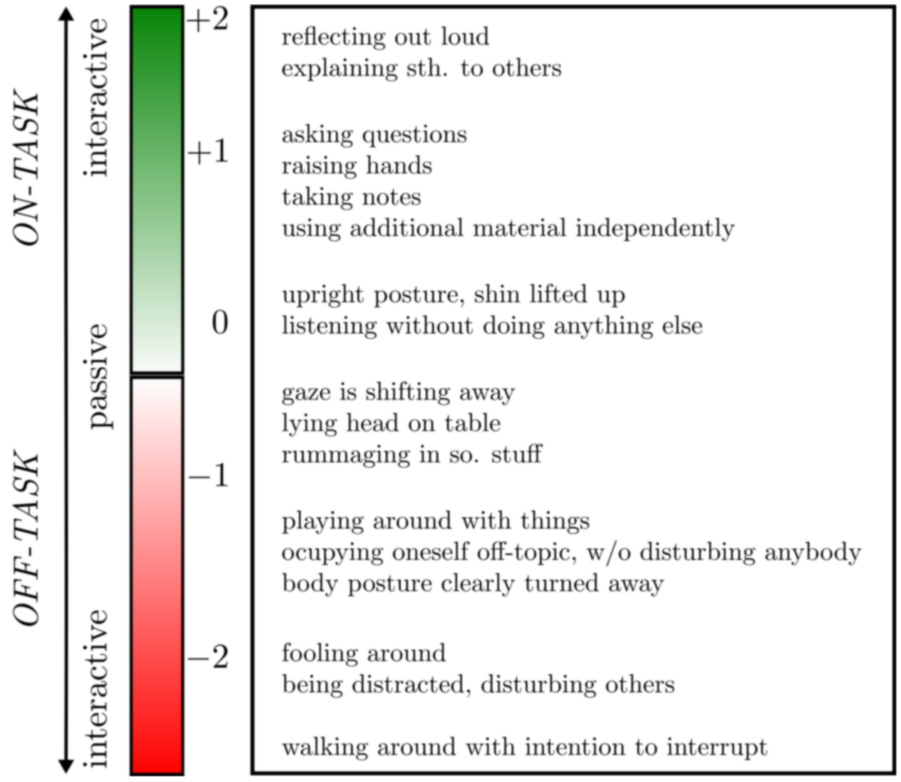


Figure A.1: Scale with exemplary behavioral indicators

In contrast to typical facial analysis tasks such as face recognition, the number of participants in classrooms is limited. We used the following three modalities as feature representations: head pose, gaze direction, and facial expressions (represented by facial action units). The head pose features consist of the head's location with respect to the camera and the rotation in radians around three axes. Gaze is represented by unit gaze vectors for both eyes and gaze direction in radians in world coordinates. Facial action units (AU) were estimated according to the Facial Action Coding System (FACS; Ekman and Friesen 1978), for which each AU can be

expressed at five intensity levels. More specifically, to estimate the occurrence and intensity of FACS AUs, we used the following 17 AUs: upper face AUs are AU1 (inner brow raiser), AU2 (outer brow raiser), AU4 (brow lowerer), AU5 (upper lid raiser), AU6 (cheek raiser), and AU7 (lid tightener); the lower face AUs are AU9 (nose wrinkler), AU10 (upper lip raiser), AU12 (lip corner puller), AU14 (dimpler), AU15 (lip corner depressor), AU17 (chin raiser), AU20 (lip stretcher), AU23 (lip tightener), AU25 (lips part), AU26 (jaw drops), and AU45 (blink). Given that our videos were recorded at 24 frames per second, and the manual annotations were conducted each second, we used the mean values of these features for time sequences of 24 frames to predict engagement intensities. More specifically, we regressed the engagement intensities using linear Support Vector Regression (Fan et al. 2008) in a subject-independent manner. Excluding the subject whose engagement intensity was to be predicted, individual regression models were trained using all other student features and labels. Subsequently, the test subject's engagement during each 1-s period was predicted. Finally, the average estimated engagement intensity during the instructional session was taken as the final descriptor for each participant.

The label space for students' manually annotated engagement was between -2 and $+2$; however, the distribution of the data was highly imbalanced. Nearly 80% of all of the annotated data ranged from 0.2 to 0.8. Therefore, we had to clip the label values to fit the range of -0.5 and 1.5 and then rescale them to 0 and 1 in our regression models.

In summary, the visible indicators we used could be differentiated into two categories: engagement-related features (i.e., head pose and gaze direction) and emotion-related features (AU intensities). In order to compare their contributions with visible indicators of (dis)engagement in learning, we used them both separately and in combination.

In order to go beyond a single-person analysis, we further integrated an indicator of synchrony. Because simultaneous (i.e., synchronous) behavior in a group of students or an entire classroom can have an impact on individual students, in this first step toward an automated approach, we considered the behavior of neighboring students sharing the same desk. First, we measured the cosine similarities between neighboring students' manual ratings ($N = 52$, 26 pairs). Second, we calculated the relation between neighbors' synchrony (cosine similarities) and their mean engagement levels during instruction. Because synchronization is a precursor to engagement, we expected the neighbors to provide valuable information for estimating (dis)engagement in learning. Therefore, in the final step of our analysis, we concatenated the feature vector of each student and his or her neighbor into a single vector and trained the same regression models as for the estimation of each individual student's engagement.

A.1.7 Results

Relation Between Continuous Manual Annotation and Student Learning

We tested the validity of our manual rating instrument in two steps. First, we investigated construct validity by correlating the manual ratings with the self-reported learning activities. The manual annotations were significantly correlated with students' self-reported cognitive engagement, situational interest, and involvement ($.49 \leq r < .62$; TableA.2).

Variable	1	2	3	4	5	6	7	8	9	10	11	12
1. Female												
2. Age	−.29* [−.52, −.01]											
3. Abitur grade	−.10 [−.36, .19]	.31* [.03, .54]										
4. School type	−.09 [−.36, .20]	−.01 [−.29, .27]	−.26 [−.51, .01]									
5. Dispositional interest	−.12 [−.39, .16]	.04 [−.24, .32]	−.07 [−.34, .21]	.01 [−.27, .29]								
6. Self-concept	.16 [−.12, .42]	−.22 [−.47, .06]	.16 [−.12, .42]	.00 [−.28, .28]	−.62** [−.77, −.41]							
7. Self-control capacity	.09 [−.20, .36]	−.10 [−.37, .18]	−.15 [−.41, .14]	−.14 [−.41, .14]	.28 [−.00, .52]	−.38** [−.60, −.12]						
8. Cognitive abilities	.04 [−.24, .31]	.07 [−.22, .34]	−.39** [−.60, −.12]	.08 [−.21, .35]	.14 [−.15, .40]	−.22 [−.47, .06]	.02 [−.30, .26]					
9. Manual rating	−.21 [−.46, .08]	.04 [−.24, .32]	.02 [−.26, .30]	−.25 [−.49, .03]	.18 [−.10, .44]	−.21 [−.46, .07]	.20 [−.08, .45]	.01 [−.27, .29]				
10. Cognitive engagement	−.14 [−.40, .14]	−.14 [−.41, .14]	.03 [−.25, .30]	−.11 [−.38, .18]	.30* [.03, .54]	−.26 [−.50, .02]	.31* [.03, .54]	−.12 [−.38, .17]	.60** [.39, .75]			
11. Situational interest	.05 [−.23, .33]	−.27 [−.51, .01]	−.06 [−.33, .22]	−.11 [−.37, .18]	.51** [.27, .69]	−.32* [−.55, −.05]	.11 [−.18, .37]	−.05 [−.32, .23]	.49** [.24, .67]	.60** [.39, .75]		
12. Involvement	−.11 [−.38, .17]	−.06 [−.34, .22]	.14 [−.15, .40]	−.23 [−.47, .06]	.30* [.02, .53]	−.28 [−.52, .00]	.35* [.07, .57]	−.17 [−.42, .12]	.62** [.42, .77]	.76** [.61, .86]	.68** [.50, .81]	
13. Knowledge test	.09 [−.20, .36]	−.07 [−.34, .21]	−.23 [−.48, .05]	−.19 [−.45, .09]	.20 [−.08, .45]	.03 [−.30, .25]	−.08 [−.35, .21]	.33* [.06, .56]	.30 [.03, .54]	.12 [−.16, .39]	.42** [.16, .62]	.21 [−.07, .46]

Table A.2: Correlations between individual characteristics, learning activities, achievement, and manual rating, with confidence intervals in brackets. To better understand the relations between individual prerequisites, learning activities, and learning outcomes, we calculated correlations across all variables. Values in square brackets indicate the 95% confidence interval for each correlation. The confidence interval is a plausible range of population correlations that could have caused the sample correlation (Cumming, 2014). Abitur grade: lower values indicate better results according to the German grading system $*p < .05$; $**p < .01$.

Table A.3: Prediction of knowledge test results (N = 52)

	Model 1			Model 2			Model 3		
	b	*SE*	*p*	*b*	*SE*	*p*	*b*	*SE*	*p*
Manual rating	1.08	0.49	.032	0.92	0.49	.067	1.00	0.48	.042
Abitur grade				-0.60	0.29	.043	-0.50	.30	.099
School type				-0.40	0.28	.159	-0.47	0.27	.087
Cognitive abilities							0.08	0.04	0.068
Dispositional interest							0.08	0.23	.066
Self-concept							0.38	0.26	.160
Self-control capacity					-0.28	0.21	.189		
R^2	.092			.184			.342		
F	4.88^{*}			$3.46^{a}st$			3.12^{**}		

Abitur grade: lower values indicate better results according to the German grading system (*p < .05; **p < .01; ***p < .001).

Additionally, we calculated a multiple linear regression with the three self-reported learning activities as regressors. Together, they explained 42.9% of the variance in the manual ratings. This corresponds to a multiple correlation of $r = .66$. Second, we examined the predictive validity of our new instrument. We inspected the intercorrelations between all variables with the knowledge test (Table A.2). The knowledge test scores (the dependent variable in this study) were significantly correlated with the manual ratings, cognitive abilities, and situational interest ($.30 \leq r < .42$). To test for effects of possible confounding variables, we calculated two additional linear regression models in which we added background variables (model 2) and learning prerequisites (model 3) into the regression and compared them with the prediction that involved only manual ratings (Table A.3). The effect of the manual ratings remained robust and still explained a significant proportion of the variance in the knowledge test results.

Reanalysis with Machine-Learning Approach

We applied our trained regression to test subjects at 1-s intervals and applied mean pooling to create a final estimation that summarized participants' engagement. Table A.4 shows the performance of different modalities for estimating (dis)engagement in learning. The performance measures were mean squared errors in the regression and the Pearson correlation coefficient between the manual annotations' mean level and our models' prediction during the instructional session.

As shown in Table A.4, the head pose modality exhibited a lower correlation with the manual ratings (r = .29) than the other features. By contrast, gaze information and facial expressions (AU intensities) were more strongly correlated with the manual annotations (r = .44). Combining head pose and gaze (r = .61) or all three modalities (r = .61) also led to substantial correlations with the manual annotations.

In addition, we tested the correlations between the posttest variables (i.e., the knowledge test and self-reported learning activities) and the different models for estimating the manual ratings (Table A.5). According to these results, regression models, which perform better with respect to MSE and lead to higher correlations with the manual ratings, seem to contain more information that is relevant for the posttest variables, particularly with respect to involvement and cognitive engagement.

Addition of Synchrony to the Machine-Learning Approach

The cosine similarities of the manual annotations between neighboring students were strongly correlated with each neighbor's mean engagement level throughout the recording ($r = .78$). More specifically, taking the synchronization into consideration improved the correlation with the manual ratings by 9%, thus showing that synchronization information is helpful for understanding (dis)engagement in learning.

Table A.4: Performance of different modalities in engagement in learning estimation depicted as mean squared error (MSE) for regression and Pearson correlations between manual ratings and our models' estimation (N = 30)

Modalities	MSE	r	p
Single students			
Head pose	0.057	.29	.126
Gaze	0.055	.44	.015
Facial expressions	0.056	.44	.014
Head pose + gaze	0.052	.61	.000
3-Combined	0.051	.61	.000
Single students + cosine similarity			
Head pose + gaze (sync)	0.029	.71	.000
3-Combined (sync)	0.050	.70	.000

The correlations between the different models for estimating the manual ratings and students' self-reported learning activities and outcomes revealed that the best models were those in which head pose and gaze features were combined with neighbor synchrony (r = .08, .43, .39, and .26 for the knowledge test, involvement, cognitive engagement, and situational interest, respectively; Table A.5). We calculated the mean correlation (based on Fisher's z-transformed correlations) of the three manual annotations (average $r = .74$) and the mean correlation of each rater and the scores from a model combining head pose, gaze features, and neighbor synchrony (average $r = .64$) for the subsample.

Because the model in which head pose and gaze were combined with neighbor's synchrony had the highest correlation with the manual rating, we calculated a linear regression to predict the posttest variables (Table 6). In order to understand the contribution of neighbor's synchrony, we trained our regression models using the same features with and without synchronization information. Adding neighbor's synchrony improved the prediction of all posttest variables

and explained at least 2% more variance. However, the manual rating remained superior.

Table A.5: Pearson correlations of different modalities in engagement in learning estimations with post-test variables (N = 30)

Modalities	Knowledge test		Involvement		Cognitive engagement		Situational interest	
	r	*p*	*r*	*p*	*r*	*p*	*r*	*p*
Single students								
Manual ratings	.14	.468	.64	.000	.62	.001	.53	.003
Head pose	-.17	.392	.05	.799	.02	.914	-.02	.913
Gaze	.11	.582	.19	.335	.16	.414	.23	.236
Facial expressions	-.09	.667	.37	.053	.23	.249	.30	.116
Head pose + gaze	-.03	.867	.41	.029	.37	.053	.21	.286
3-Combined	-.04	.827	.43	.023	.37	.055	.21	.277
Single students + similarity								
Head pose + gaze (sync)	.08	.704	.43	.023	.39	.040	.26	.175
3-Combined (sync)	- .01	.968	.45	.016	.38	.043	.26	.189

A.1.8 Discussion

The present study reported key initial results from the development of a machine vision-based approach for assessing (dis)engagement in the classroom. We were able to find empirical support for the validity of our newly developed manual rating instrument. Furthermore, the machine-learning approach proved to be effective, as shown by its correlation with the manual annotations as well as its ability to predict self-reported learning activities. Finally, as expected, including an indicator of synchrony in the automated analyses further improved its predictive power. Next, we discuss our main results in more detail before turning to the limitations of the present study and the crucial next steps.

Empirical Support for the Newly Developed Approach

The manual rating of visible indicators for (dis)engagement in learning predicted achievement on a knowledge test following a university seminar session. This prediction was robust when we controlled for individual characteristics (research question 1). In terms of validity, self-reported cognitive engagement, involvement, and situational interest were strongly correlated with the manual rating. As these self-reported learning activities reflect students' cognitive processes during the seminar session, we concluded that our manual ratings capture visible indicators that are actually related to (dis)engagement in learning. Therefore, we inferred that it is reasonable to use these manual ratings as a ground truth for our machine vision-based approach.

In the automated analyses of engagement, we used several visible features (head pose, gaze, facial expressions). More specifically, we compared their contribution with visible indicators of (dis)engagement in learning separately and in combination. Our results showed that facial expressions were more strongly correlated with the manual rating than head pose or gaze alone; however, combining the engagement-related features and combining all three visible indicators improved the correlation with the manual annotations substantially, thus emphasizing the complexity of human rating processes. However, we were not able to replicate the prediction of the knowledge test scores by considering these visible features alone (research question 2).

We expected that additional information concerning interaction with peers and similar behavioral aspects would improve the estimated model. Indeed, adding synchrony by considering the engagement patterns of students' neighbors improved the correlations with the manual rating as well as the prediction of the posttest variables (research question 3). In line with Raca et al.'s (2013) correlative results, our findings indicated that considering neighbor synchrony leads to a better understanding of engagement in predictive models. However, the manual ratings were still better at predicting the knowledge test results as well as self-reported cognitive engagement, involvement, and situational interest. Yet, the similarity between the three different manual raters ($r = .74$) differed from the similarity between the manual annotations and the machine-learning approach ($r = .64$). This difference obviously leaves some room for improvement; however, the approximation that was based on visual parameters and the synchrony with a neighbor's behavior appears to provide reliable results. This raises the question of whether human annotators should also include more than just a single person in their ratings and (unconsciously) consider additional information.

Possible Contributions of an Automated Approach for Assessing Engagement

Our machine-learning approach provides a promising starting point for reducing the effort involved in manual video inspection and annotation, which in turn would facilitate the analysis of larger numbers of individuals and longer videotaped lessons. In addition, such approaches enable the consideration of more complex information on synchronization across students in a way that goes beyond the ability of human observers. This approach is potentially fruitful for both research and practice.

Information from automated analyses of engagement can be used to provide feedback to teachers and improve their skills in monitoring and identifying relevant cues for students' attention in complex classroom interactions. When teachers can notice and identify a lack of engagement, they have the opportunity to adapt their teaching method accordingly and to encourage the students to deal with the learning content actively. Furthermore, by noticing and identifying distracting behavior, teachers get the chance to react to disruptions and ensure the effective use of instruction time. An automated analysis of videos can support novice teachers in developing professional vision skills, and it can provide feedback to teachers in general about the deep structure of their teaching. By making work with videos less effortful,

this method could allow videos to be implemented in teacher training more systematically.

Moreover, the annotation of (dis)engagement in learning over time opens up new opportunities for further investigations of classroom instruction by adding a temporal component. This method allows for the detection of crucial events that accompany similar engagement-related behavior across students and provides deeper insights into different effect mechanisms during instruction. Furthermore, this approach can be combined with additional measures. For example, tracking human raters' eye movements can provide insights into where they retrieve their information and what kinds of visible indicators they actually consider. This knowledge can further improve machine vision-based approaches by including the corresponding features. In addition, combining valid visible indicators of students' (dis)engagement in learning with eye-tracking data for the teacher, for example, makes it possible to analyze in more detail what kind of visible indicators attract novice teachers' attention (e.g., Sümer et al. 2018). This information can then be reported back in teacher training to support professional vision skills.

Challenges and Limitations

Our study has several notable limitations that need to be addressed in future research. First, face recognition was not possible for all students due to the occlusion of their faces some or most of the time. For this reason, we had to reduce the sample size for the automated analysis, which in turn reduced the statistical power. Limited data was also an issue in the study by Whitehill et al. (2014), who only found moderate correlations between engagement and learning for this reason. It can thus be assumed that increasing the number of participants recognized by face detection would further improve the linear regression models used to predict self-reported learning activities and learning outcomes. The use of mobile eye trackers for each student is an example of one solution that can provide data for individual students. However, the use of eye trackers is expensive, and when used with children who might touch the glasses too often, it deteriorates the gaze calibration and results in an erroneous analysis of attention. Besides, mobile eye trackers can affect the natural behavior of students, whereas field cameras are pervasive and do not create a significant intervention. To overcome the issue of students being occluded, different camera angles could be helpful in future studies.

Second, a challenging aspect of engagement estimation in our setting was the highly imbalanced nature of our data. Engagement levels on both outer ends of our rating scale were underrepresented. As a direct consequence of the learning setting (a teacher-focused session on statistics), few participants displayed active on-task behavior (e.g., explaining content to others); even less data were collected for visible indicators of disengagement in learning indicating active off-task behavior (e.g., walking around with the intention to interrupt). This imbalance has negative implications for the training of algorithms because greater variability in behavior typically leads to more accurate automated analyses. Whereas human raters are familiar with high levels of variance in an audience's on-task and off-task behavior and use this implicit knowledge in their annotation, the algorithms were trained using only the available data from our three sessions. However, this limitation can be overcome by recording real

classroom situations, which will be part of our future work. Although it is not possible to control the intensity of students' (dis)engagement in learning in natural classroom settings, completing more recording sessions and including more participants will eventually lead to a wider distribution of characteristics.

Third, additional research is necessary to validate our approach in schools due to the different target population. This is particularly important because high school students might exhibit a more diverse set of visible indicators of (dis)engagement in learning.

A.1.9 Conclusion

Remote approaches from the field of computer vision have the potential to support research and teacher training. For this to be achieved, valid visible indicators of students' (dis)engagement in learning are needed. The present study provides a promising contribution in this direction and offers a valid starting point for further research in this area.

Compliance with Ethical Standards

The ethics committee from the Leibniz-Institut für Wissensmedien in Tübingen approved our study procedures (approval #2018-017), and all participants gave written consent to be videotaped.

A.2 Multimodal Engagement Analysis from Facial Videos in the Classroom

Abstract

Student engagement is a key construct for learning and teaching. While most of the literature explored the student engagement analysis on computer-based settings, this paper extends that focus to classroom instruction. To best examine student visual engagement in the classroom, we conducted a study utilizing the audiovisual recordings of classes at a secondary school over one and a half month's time, acquired continuous engagement labeling per student (N=15) in repeated sessions, and explored computer vision methods to classify engagement levels from faces in the classroom. We trained deep embeddings for attentional and emotional features, training Attention-Net for head pose estimation and Affect-Net for facial expression recognition. We additionally trained different engagement classifiers, consisting of Support Vector Machines, Random Forest, Multilayer Perceptron, and Long Short-Term Memory, for both features. The best performing engagement classifiers achieved AUCs of .620 and .720 in Grades 8 and 12, respectively. We further investigated fusion strategies and found score-level fusion either improves the engagement classifiers or is on par with the best performing modality. We also investigated the effect of personalization and found that using only 60-seconds of person-specific data selected by margin uncertainty of the base classifier yielded an average AUC improvement of .084.

A.2.1 Introduction

Which students are engaged in learning during the class? What is the relationship between student engagement and the content and the quality of the learning material? And, additionally, how can we relate student engagement to learning outcomes or long-term goals? These research questions and more drew the interest of scientists from educational sciences, psychology, and similar fields to investigate student engagement.

To begin our investigation of student engagement, we must first define the term engagement and contextualize its implications in the classroom setting. Being engaged means "to involve oneself or become occupied; to participate" while engagement can be defined as "[being] actively committed". As it relates to human behavior, engagement is highly connected to commitment and involvement. In the educational context, according to Fredricks et al.'s widely accepted definition [2], engagement is a multidimensional construct that is composed of three dimensions: *behavioural*, *cognitive*, and *emotional*. Those dimensions do not reflect isolated processes, but rather dynamically interrelated factors within an individual student. Behavioral engagement focuses on the act of participation and can include behaviors such as displaying attention and concentration, or asking questions. The basis of behavioral engagement is involvement in social, academic, or extracurricular activities seen as essential for achieving positive outcomes. On the other hand, emotional engagement encompasses affective compo-

nents such as students' interest or boredom. Whereas aspects of behavioral and emotional engagement are typically observable from the outside, cognitive engagement incorporates less overt internal, cognitive processes such as psychological resource investments in learning and self-regulation [2]. It also incorporates the students' willingness to comprehend complex ideas and master skills. Importantly, previous research found positive correlations between aspects of student engagement and academic achievement, emphasizing student engagement's central role in classroom learning [144].

In the present study, we aim to evaluate student engagement based on visible indicators in learning situations and approach the analysis from an affective computing perspective. Two methods are proposed by affective computing and classroom management literature to acquire engagement levels: 1) self-reports and 2) observer ratings. Self-reports are practical, relatively cheap, and easy to administer to a large sample, making them valuable in various tasks related to engagement and beyond [158]. Despite their value, self-reports have certain drawbacks, namely a dependence on participant compliance, diligence, and a student's overall understanding of being engaged. These characteristics, however, are not always a given, especially in a high school setting.

Observer ratings are another useful assessment tool for student engagement. In general, observer ratings are systematic approaches that aim to detect and interpret certain behaviors [159]. Their deployment in large-scale studies is notably limited by the necessity of providing human raters with specialized training and the difficulty of acquiring reliable labeling. Moreover, in contrast to many other computer vision applications, crowdsourcing is not a viable option to label student engagement due to ethical considerations and the specialized training required for the raters.

Owing to the limitations of self-reports and observer ratings, automated approaches for estimating student engagement pose a challenge when increasing the sample size of classroom observation studies. A solution is to automatically estimate engagement using machine learning and computer vision. In the field of affective computing, initial studies aimed at estimating student engagement focused on computer-based learning and intelligent tutor systems (ITS) . From ITS log files such as students' reaction times, errors, and performance, preferred modalities for engagement analyses shifted to video, audio, and physiological measures (i.e., galvanic skin response, EEG, heart rate).

In computer-based learning settings, the availability of log data is an important asset. Furthermore, vision-based features can be extracted reliably using webcams. In the classroom, on the contrary, using sensors for each student can render studies expensive and intrusive and ultimately may affect student behaviors. Thus, a widely accepted practice in classrooms is to record the instruction with field cameras in the corners of the room. One drawback of this approach, however, is that the audio and visual data is noisy and may be occluded.

Contributions of the Study

In this study, we review, in detail, engagement studies in the field of affective computing. We then discuss the large-scale school study we conducted by collecting audio-visual recordings of classes during a one and half month period. Observer ratings of student engagement were acquired using an instrument previously validated in university-level seminars [115].

The current study's primary focus is to learn engagement classification from limited and unconstrained data where traditional face alignment and facial action unit estimation methods have largely failed. Following the definition by Fredricks et al. [2], behavioral and emotional aspects of student engagement can best be observed from the outside. Visual attention (subsequently referred to as attention) and affective components can thus serve as approximations of these two sub-dimensions. We propose learning attention and affect features from two convolutional neural networks trained on head pose estimation and facial expression recognition as pretask. In contrast to previous works that utilize faces based on handcrafted features in engagement analysis, the deep learning-based representations we propose work without precise facial alignment.

Our engagement classification is performed in these learned feature embeddings. We also applied feature and score level fusion on attention and affect features. Beyond the person-independent evaluation training and evaluation of engagement classifiers, we also investigated personalization because there is intrapersonal variation in students' (dis)engagement.

Although automated engagement analysis is widely studied in computer-based settings such as intelligent tutors and educational games, to our knowledge this study is one of the first to perform video-based engagement classification in the classroom on a large scale.

Table A.6: Automated Engagement Analysis in Classroom, Computer-based Learning, Human-Human/Human-Robot Intreraction (HHI/HRI) Settings

Reference	Setting	Behavioral Cues	Engagement Measurement	Predictive Models
[160]	classroom	head pose	observer reports	✓
[116, 114]	classroom	head pose, body motion	self-reports (in-class)	✗
[155, 156]	classroom	head pose, gaze, facial expressions, posture	observer reports	✓
[157]	classroom	gaze mapping (heads up/down)	–	✗
[161]	classroom	head pose, gaze, FACS action units	observer reports	✓
[162]	classroom	real-time monitoring system capable of extracting many behavioral features (i.e. smile detector, hand raising, head pose, speech analysis)	–	✗
[163]	classroom	monitoring system (head pose and gaze estimation)	–	✓
[164]	computer-based	FACS action units and ITS log features	observer ratings	✓
[165]	computer-based	FACS action units	self-reports (user engagement survey [166], NASA-TLX [167])	✓
[111]	computer-based	handcrafted features from faces	observer reports	✓
[150]	computer-based	FACS action units and appearance features	self-/observer reports (MW)	✓
[151]	computer-based	FACS action units and gross body movement	observer reports (BROMP [168])	✓
[113]	computer-based	Kinect Animation Units, facial appearance, heart rate estimated from face videos	self-reports (concurrent & retrospective)	✓
[169]	computer-based	facial appearance features	crowdsourcing	✓
[170]	computer-based	head pose and gaze direction	observer reports	✓
ELEA [171]	HHI	–	observer ratings	✗
RECOLA [172]	HHI	–	self-reports	✗
MHHRI [173]	HHI & HRI	audio, physiological, and first-person vision	self-reports	✓
[174, 175]	HRI	facial expressions, body pose, audio (in children's storrytelling and therapy with robots)	–	✓

A.2.2 Related Work

In recent years, the use of automated methods in classroom behavior analysis and engagement estimation has been on the rise. The popularity of such methods is largely due to the availability of big data and the progress of artificial intelligence. Notably, developments in deep learning have yielded significant results in social signal processing problems, including classroom and learning analytics.

We can categorize the literature of automated engagement estimation based on the following criteria:

- learning situation (computer-enabled settings, classroom: traditional formation vs. group-work, etc.)
- nonverbal features (various behavioral cues can be related to learning-related activities.)
- computational methodology (in both feature extraction and machine learning)
- final objectives (showing a statistical relation vs. fully automated predictive system, psychologically valid measurements of engagement).

In addition to these points, another consideration is the use of sensors [176]. Whereas sensor-free measurements depend on intelligent tutor systems' log files, sensor-based measurements use physical devices such as physiological sensors (i.e., EDA, EEG, heart rate sensors) and audiovisual recordings acquired from cameras and voice recorders. As our motivation is to measure engagement as seamlessly as possible without necessitating any expensive and intrusive sensors, we limit our scope to engagement analysis using only visual modalities. Table A.6 summarizes the literature of automated engagement analysis across three domains: classroom, computer-based settings (including intelligent tutors and screen-based learning games), and human-human, human-robot interactions (HHI/HRI). In the following subsections, we will review engagement studies in these three categories.

Learning Analytics in the Classroom

Despite the popularity of computer-based learning technologies, Intelligent Tutor Systems (ITS), and Massive Online Open Courses (MOOC), traditional classroom-based learning is still the dominant setting for primary through tertiary education. The popularity of classroom-based learning is primarily due to the importance of human factors and collaboration throughout the learning process. For this reason, analytics tools in the classroom that measure students' learning-related behaviors and affective and cognitive engagement play an essential role in improving the efficiency of classroom-based learning.

Learning analytics methods in the classroom may include video cameras in the corner of the room, direct recordings several students' faces and upper bodies, and external audio recorders. The quality of audio-visual feature extraction, in general, is not as fine-grained as in computer-based situations where a webcam, 1-2 meters away, captures a student's behaviors.

However, classroom analytics provide more insight into student-teacher, student-learning material, and student-student interactions.

To the best of our knowledge, Bidwell and Fuchs [160] proposed the first classroom monitoring system capable of analyzing student engagement. Although their technical report did not incorporate any quantitative results, they defined a general workflow of classroom analytics by using several color and Kinect depth-sensing cameras during a lesson in a third-grade classroom. Three observers attended the lesson and coded each students' behavior using a mobile device during 20 second intervals according to the following categories: appropriate (engaged, attentive, and transition) and inappropriate (non-productive, inappropriate, attention-seeking, resistant, and aggressive). Due to the limitations of only recording a single lesson and collecting highly imbalanced data, Bidwell and Fuchs used a Hidden Markov Model (HMM) to classify three categories (engaged, attentive, and transition) from head pose based gaze-target mappings.

A more recent classroom monitoring system was proposed by Raca and Dillenbourg [116]. Two ideas proposed in their study were to use students' motion information during class and student orchestration between neighboring students' feature representation to estimate student attention. In [114], they handcrafted several features such as eye contact (the percentage of time where faces are detected), amount of still time (where head pose does not change significantly for a period), and head travel (normalized head pose change). As ground truth labels of attention, Raca and Dillenbourg used self-reports that students completed in approximately 10-minute intervals. These features, together with a Support Vector Machines (SVM) classifier, performed up to the accuracy of 61.86% (Cohen's $\kappa = 0.30$) to predict 3-scale attention (low, medium, and high). Their seminal work showed that student attention can be automatically measured using visible behavioral cues. However, they used considerably long intervals (10 minutes) before self-reports were obtained. Moreover, they employed only attentional features (head pose and motion), not any affective or behavioral nonverbal features.

Zalatelj and Kosir [155, 156] used a Kinect sensor and its commercial SDK to estimate body pose, facial expressions, and gaze. Subsequently, they computed behavioral cues (i.e., yawning, taking notes, etc.) from Kinect features and trained a bagged decision tree classifier to estimate observer-rated attention levels (low, medium, high). However, some nonverbal features were extracted using Kinect's commercial SDK. They also used manually-labeled behavioral features (i.e., writing, yawning, one hand's touching head). Their experimental results included only a few minutes of video recording and the number of participants in their data sets was 3 and 6 students, respectively. Additionaly, the effect range of Kinect and similar depth sensors is around 1 to 3 meters. In a typical classroom with 20-30 students, several sensors are required, potentially introducing additional cost and device synchronization issues.

Thomas and Jayagopi [161] collected video recordings of 10 students during three 12-minute intervals while they were listening to motivational video clips on YouTube. Three observers labeled the engagement of each student during 10-second intervals. They rated based on

whether a student was looking towards the screen (teacher area), talking to a neighbor, or gazing in another direction. Their approach was to use head pose, gaze direction, and facial action unit features with SVM and logistic regression. The main limitation of this study was the limited data size and the lack of engagement labeling methodology. When a student is listening to the audio, looking to a voice source should not be considered a cue of attentiveness. Students can still focus on content when looking around or taking notes.

Goldberg et al. [115] is the first study that utilizes a psychologically valid and comprehensive engagement rating system. Their continuous observer-based rating system combines Chi & Wylie's ICAP (Interactive, Constructive, Active, Passive) framework [12] and on-task/off-task behavior analysis [177]. Using attentional (head pose and gaze direction) and affective (FACS action unit intensities) sets of features, as well as SVR, they predicted continuous observer-ratings and additionally showed the correlation between estimated engagement levels and self-reports collected at the end of 40-minute teaching units (N=52). They also showed that behavioral synchrony with immediate neighbors improved the estimating of engagement.

One of the main objectives of learning analytics in the classroom is to report the estimated attention and engagement of students to teachers. For instance, Fujii et al. [157] estimated head-down (i.e., taking notes or reading learning material) and head-up (gazing at whiteboard-/teacher area) states for each student and depicted color-coded visualization to teachers with a sync rate. However, they tested the performance of the head-down/head-up detector on limited data. Also, reporting behavioral cues (looking at learning material or the teacher area) instead of engagement levels leaves teachers with minimal information.

Two recent studies [162, 163] developed smart classroom monitoring systems. Whereas Anh et al. [163] used only gaze mapping and visualized the distribution on a dashboard, Ahuja et al. [162] integrated various nonverbal features in their smart classroom, EduSense. These features included the state-of-the-art methods in face detection and alignment, body pose estimation, hand raise detection, and active speaker detection. [162] presented a technical analysis of real-time classroom monitoring systems, including the speed and latency of the system and algorithms' performance. However, they did not report on student engagement. Even though nonverbal features are essential to understand engagement, they are not easy to interpret, on their own, by a teacher.

In summary, computer vision-based classroom analytics studies are still limited. The sample sizes are small and the majority do not estimate attention or engagement levels. Besides, in studies that estimate student attention/engagement, there is no consensus regarding the assessment instrument.

Engagement Estimation in Computer-based Learning

Computer-based learning situations are more restricted than classroom situations because they only contain student-learning material interactions. These situations capture video

and audio from 1 to 2 meters away, resulting in better quality feature extraction methods. Furthermore, introducing an intervention during learning is more straightforward than in the classroom setting. For these reasons, automated engagement estimation is more prevalent in computer-based situations where the participant plays an educational game, conducts reading comprehension or writing tasks, or learns with an ITS.

The first study we reviewed that predicts the level of engagement in computer-based settings (during which the participants perform a cognitive training task) was conducted by Whitehill et al. [111]. They used appearance-based facial features (Box filters, Gabor filters, CERT FACS features) and estimated levels of engagement using several classifiers such as GentleBoost, SVM, multinomial logistic regression. They developed a manual rating system (4-scales) and annotated the video recordings at 60-sec or 10-sec intervals. The accuracy of their classifiers vary between 36-60%.

In a similar computer-based setting, a writing task, Monkarasi et al. [113] estimated engagement using Kinect face tracker ANimation Unit (ANU) features, LBP-TOP, and heart rate (estimated from videos of the face). They used concurrent self-reports (during the writing task every 2 minutes) and retrospective self-reports after the participants finished the task. Both self-reports showed high correlation ($r = 0.82$, $p \leq 0.001$). Bosch et al. [151] used estimated FACS action units as features and predicted BROMP annotations [168] using different classifiers (Bayes Net, Updateable Naive Bayes, Logistic Regression, AdaBoost, Classification via Clustering, and LogitBoost).

Another factor that plays an essential role in task engagement is mind wandering (MW). MW is defined as an attentional shift from the primary task and a subsequent decrease in task engagement [178]. For instance, in the learning context any thoughts that arise either intentionally or from boredom can be MW and linked to models of engagement.

The availability of automated methods to predict MW can reveal the covert aspect of engagement. The use of visual modalities, particularly face videos, to detect MW is preferable to eye gaze [179] and physiological signals [180] which necessitate specialized sensors or textual features [181] that may depend on speech recognition, NLP, or manual labeling and relevant to our study. Swewart et al. [182] is the first study that used the visual modality, facial action units and body motions to detect MW. They recorded facial videos while the participants watched a narrative film for 35 minutes. Each participant annotated MW by pressing a key through the screening. Facial action unit features and classifiers including logistic regression, naive Bayes, and support vector machines could spot MW in a person-independent setting with F_1 score of .390. Later, [183] showed the generalizability of MW detection when trained and tested on different tasks (reading scientific text and watching a narrative film). Bosch et al. [184] showed the applicability of MW detection in a classroom study (N=135) learning from an intelligent tutor system.

Human-Human and Human-Robot Interactions (HHI/HRI)

Another line of work is the attention analysis of human-human interactions, i.e., in group work and human-robot interactions. For example, Sanches-Cortes et al. [171] proposed an audiovisual corpus of groups with four participants during a survival task and focused on estimating group performance, apparent personality, and perceived leadership and dominance. Similarly, Rinvegal et al. [172] used a survival task during remote collaboration using audio, video, and physiological signals as well as self-reported engagement. However, although survival tasks can be useful to measure group interactions, they do not represent typical learning situations.

Looking into more recent studies, Celiktutan et al. [173] collected an audiovisual dataset during human-human and human-robot interactions using first-person cameras. They acquired self-/acquaintance-assessed personality and self-reported engagement labels. However, drawbacks that limit their setting include the scale of the dataset and interactions wherein one participant or robot asks predefined, standard questions. Another application in human-robot interactions is autism therapy for children [174, 185] and child-robot interactions (a dialogic storytelling task) [175, 186]. The distribution of engagement during children's storytelling or autism therapy is more obvious and, in these settings, it is comparably easier to differentiate between engaged and disengaged behaviors than it is in schools where most pupils learn to hide their disengagement. Despite the lack of expert-labeling criteria, these studies adopt a continuous engagement labeling approach as in our engagement annotations. Furthermore, they used deep Q learning to actively sample training data and personalize models with limited data.

To summarize, the literature in attention and engagement analysis is centered on computer-based learning settings as well as human-human and human-robot interactions. Collecting data for automated analysis in those domains is comparably more convenient than in the classroom. However, the impact of schools and classroom instruction exceeds the scope of these applications and, moreover, plays a crucial role in every student's life. Existing classroom-based studies are very limited in terms of data size. They were mostly conducted on university-level courses or on a small number of participants (mainly to test computer vision systems). While Raca and Dillenbourg [114] conducted the most comprehensive attention monitoring study in the classroom and, thusly, showed the applicability of these technologies in a school setting, their study lacked expert-labeled attention/engagement measures and predictive learning models on a larger scale.

A.2.3 Data Collection for Automated Engagement Estimation in the Classroom

The study was conducted during regular lessons at a secondary school in Germany over a one and a half month period. The ethics committee from the Faculty of Economics and Social Sciences of the University of Tübingen approved our study procedures (Approval #A2.5.4-097_aa), and all teachers and parents provided written consent for their kids to be videotaped. Students who refused to be videotaped attended a parallel session covering the same instructional

content.

Participants

We collected audio-visual recordings of 47 classes from 5^{th} to 12^{th} grades, including 128 participants overall. Each participant attended more than one class (3.84 on average). Therefore, the total number of samples across grades was over 360. The collection of labelled data for developing and benchmarking automated methods is time-consuming. Thus, we identified a sub-sample of students based on their occurrence and visibility in multiple video recordings and used a sub-sample of 15 students from grade 8 (N=7) and grade 12 (N=8) in our analysis. Each participant appears five times on average and the total number of samples in our data is 75. Classes cover a wide range of subjects, including Mathematics, Chemistry, Physics, IMP (**I**nformatics, **M**athematics, **P**hysics), History, Latin, French, German, and English.

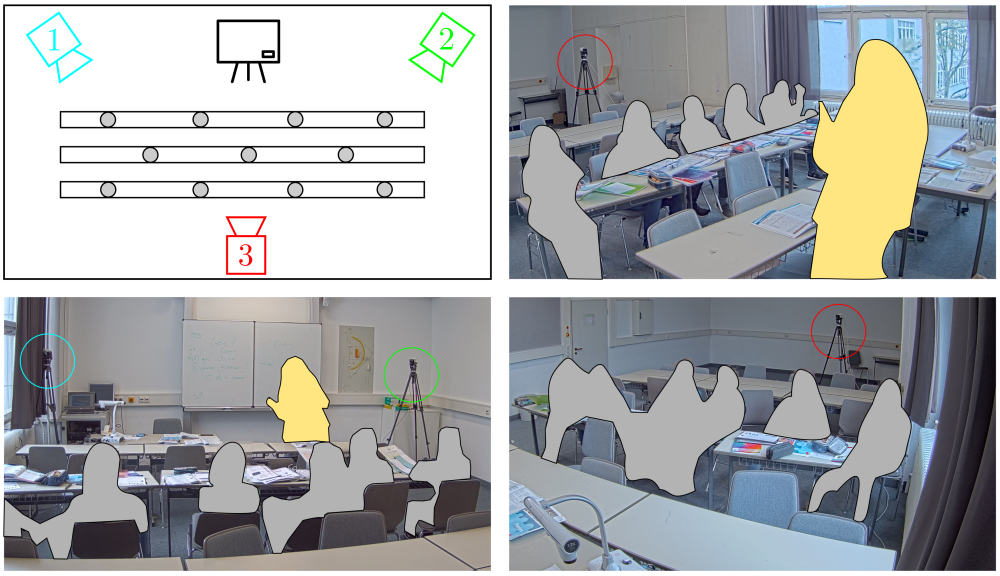


Figure A.2: Sample scene from the classroom. The synchronous cameras recorded the instruction simultaneously.

Procedure

Before classes on the first day, students filled out a questionnaire covering demographic information (age, gender) and individual prerequisites (BFI-2 XS, 15 items; [187]). After each session, students completed another questionnaire about their learning activities. Session recordings lasted between 30 and 90 minutes each. Video material during classes covered group work, individual work, and teacher-centered instruction. To best capture student attention on the instructor, we focused on teacher-centered components of the video (see Fig A.2), extracting the main part of instruction time in intervals of 15 to 20 minutes from each recording. The intervals were manually annotated by human raters.

Self-Reported Learning Activities

After each session, we assessed students' involvement (four items, $\alpha = 0.73$; [188]), cognitive engagement (six items, $\alpha = 0.78$; [189]), and situational interest (six items, $\alpha = 0.92$; [190]) during the preceding instructional period.

Continuous Manual Annotation

To manually annotate students' observable behavior, we used a one-dimensional scale in steps of seconds through the open software, CARMA [148], which enables continuous interpersonal behavior annotation via joystick [149]. We also combined the concept of on-task/off-task behavior [177, 14] with existing scales from the engagement literature. To define more fine-grained cues within the possible behavioral spectrum, **I**nteractive, **C**onstructive, **A**ctive, and **P**assive, we gained inspiration from the ICAP framework [12]. Thus, behaviors were annotated on a symmetric scale ranging from -2, indicating disturbing (i.e., interactive), off-task behavior, to +2, indicating highly engaged, interactive, on-task behavior (see Fig A.3). Values closer to 0 indicated rather unobtrusive, passive behavior. Two raters annotated the sub-set of students in all videos in random order, with inter-rater reliability ICC(2,2) for each student being 0.77 on average (absolute agreement). For subsequent analysis, the mean across the two raters is calculated for every learner in every second. For more details about the manual annotation instrument, interested readers are referred to Goldberg et al. [115].

Preprocessing

In each video recording, we had three cameras as depicted in Figure A.2. One camera was located in the rear part of the class covering the classroom and teacher and the other two cameras were placed on the left and right side of the teacher area (whiteboard) directed towards the class. We applied our computational pipeline to both the left and right camera and dynamically picked the stream where a particular student was more visible.

We used a single-stage face detector, RetinaFace [95], to detect all faces in the video streams. Subsequently, we picked several query face images that belonged to the students whose behaviors we intended to analyze. Instead of face tracking, we directly used those query images and extracted ArcFace embedding [96] for all face patches. By calculating the minimum cosine similarity between the query images and all faces, we created face tracklets for each student. Despite the challenges of occlusion and different camera angles, the face detection and recognition methods we employed could localize and recognize faces most of the time due to their training on large and unconstrained data sets. We used one-second (24 frames) continuous sequences where both face detection and recognition worked smoothly.

Table A.7 shows the number of different day recordings per student and the total length of the data where preprocessing worked. The total data length is 25,450 and 32,755 seconds in Grades 8 and 12. This amount makes over 15 hours of recording in 30 sessions. Compared to other

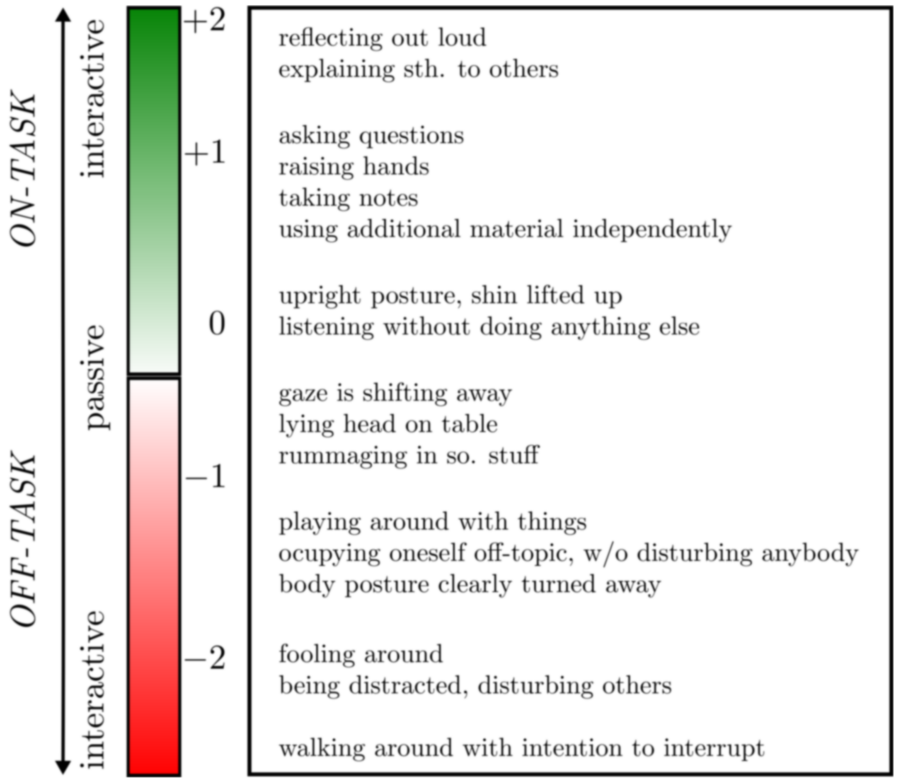


Figure A.3: Continuous scale of our manual rating instrument and visible behavioral indicators [115]

classroom-based studies, the line of work by Raca & Dillenbourg [114] used four classes in 9 sessions. Even though their study was on large-scale data, their attention analysis was based on 10-minute intervals and self-reports. When we look into the size of other engagement studies in the classroom, the results are limited: three videos of 12-minute recordings in [161], 25 minutes of video recordings in [155], 4 minutes in [156].

In the continuous labeling scale, values denoting disengagement are rarely observed and the labels are often imbalanced. Thus, we followed the previous works that discretized the continuous scale into three scales: low [-2, 0.35], medium (0.35, 0.65], and high engagement (0.65, 2.0]. Figure A.4 depicts the continuous and discrete distribution of labels in Grades 8 and 12.

A.2.4 Methodology

Problem Statement

In our setting to classify engagement level, we used video recordings of classes. Formally, we employed sequences $\mathscr{S} = \{I_1, I_2, \cdots I_n\}$ where $n = 1, \cdots, N$ denotes the time intervals of a second (24-frames). Using any of the modalities, we extracted feature vectors from each

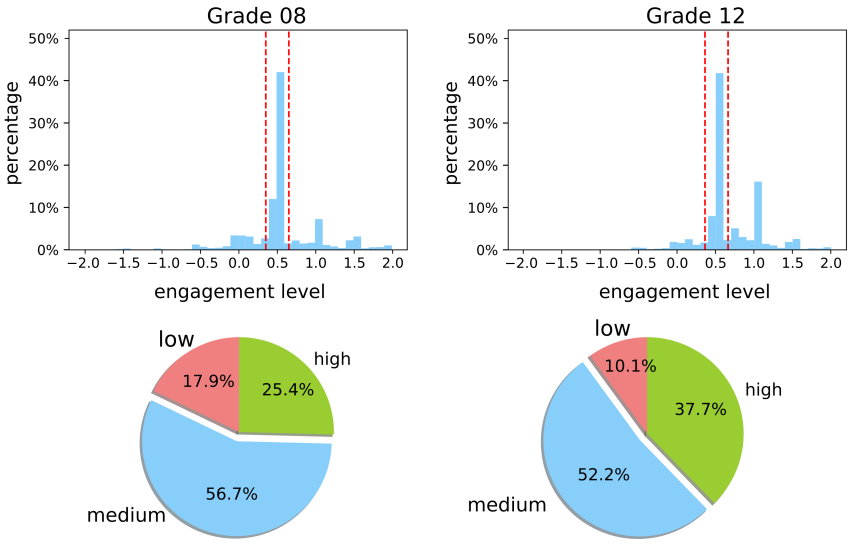


Figure A.4: The distribution of engagement labels in Grade 8 and 12. Pie charts show the percentage of quantized labels according to continuous labelling.

sequence $\mathbf{x} = \{x_1, x_2, \cdots x_n\}$ with $\mathbf{x} \in \mathscr{R}^{\mathcal{T} \times M \times D_m}$. The feature sequences are associated with engagement label $y = \{0, 1, 2\}$. To predict the engagement labels, we used either a single middle frame of a sequences or all frames in a temporal learning model.

Table A.7: The number of classes and the total duration of recording where face detection works (in seconds) per each student.

Grade 8							
student	**S4**	**S7**	**S8**	**S11**	**S13**	**S14**	**S16**
#class	2	7	7	3	6	4	6
seconds	836	5450	5309	2269	4404	2674	4508

Grade 12								
student	S1	S2	S3	S4	S5	S6	S7	S8
#class	9	8	3	3	4	3	6	4
seconds	6363	6695	2662	2708	4219	2605	3844	3659

Feature Representation

In most of the classes, students were listening to the teacher instead of speaking. Due to occlusion of the students' upper bodies in many of the recordings, nonverbal features such as speech and body pose are not always available. However, faces are usually visible and computationally faster and more reliable to detect. Consequently, our analysis depends on preprocessed faces as described in A.2.3.

Motivated by the fact that engagement is a multidimensional construct, we can extract two

different sets of information from face images: attentional and emotional features. There are several studies in the literature that used available face processing tools such as OpenFace [94] for engagement estimation [174, 170].

The main drawback of this approach, however, is that it depends on very accurate face alignment. In the classroom, camera distance from students varies between 2-10 meters, and this reduces image quality and eventually leads to poor facial keypoint localization. When we processed the classroom data using [94], it could process approximately 30-40% of a students' face in a class with high confidence. Furthermore, even though facial action unit-based approaches provide valuable information on affect, they almost always anticipate nearly frontal images. Considering these issues, we extracted affect features based on categorical facial expression recognition and attention features based on head pose estimation without depending on 68-point facial landmarks.

Figure A.5 shows the feature learning for affect and attention. In affect branch (Affect-Net), we used one of the most unconstrained and large-scale affect datasets, AffectNet [99], and trained a ResNet-50 network using softmax cross entropy loss to predict categorical models of affect (seven discrete facial expressions): neutral, happy, sad, surprise, fear, disgust, and anger. The training set of AffectNet is composed of 23,901 images, whereas the validation set has 3,500 images. We aligned all face images using five facial keypoints that were estimated by face detector [95] and aligned by similarity transform to the size of 224 × 224. The training

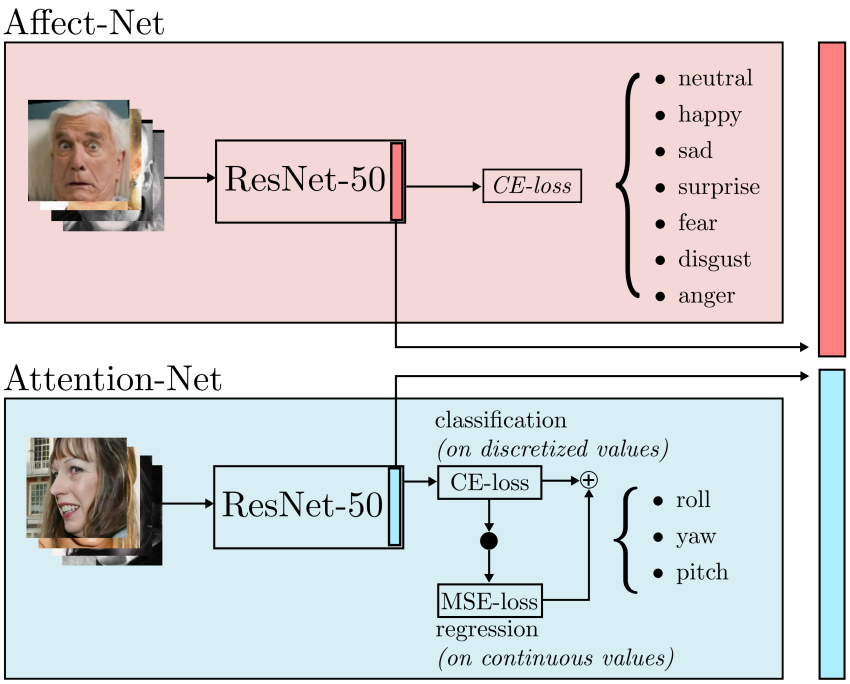


Figure A.5: Feature learning for affect and attention. Two ResNet-50 backbones are separately trained for facial expression recognition and head pose estimation. The learned features subsequently will be used for engagement estimation on classroom data.

was done using an SGD solver with an initial learning rate of 0.1 (decayed ten times in each 30 epochs) for 100 epochs. The best accuracy on the validation set reached 58%. We used the layer's feature activations before the last fully connected layer as affect embedding.

We used another ResNet-50 backbone (Attention-Net) to learn attention features. By adopting the approach in [56], we trained the network on 300W-LP [98] to estimate head pose jointly by softmax classification on discretized values and mean squared loss on continuous values. The advantage of the CNN-based approach for head pose estimation is that it is more robust than Perspective n-Point (PnP)-based methods that find correspondence between estimated facial keypoints on image and their corresponding 3D locations in an anthropological face model. In challenging cases where those methods fail, CNN-based methods can return satisfactory predictions and, more importantly, map the inputs in a continuous low-dimensional embedding according to poses.

As the training corpus is very large and contains various challenging situations in both Affect-Net and Attention-Net branches, these methods learn robust features. Compared to the handcrafted appearance features such as Local Binary Patterns or Gabor filters, deep embeddings can be extracted without precise alignment and are extendable by training with new DNN architectures on more data. We trained Attention-Net and Affect-Net representations on head pose estimation (300W-LP) and facial expression (AffectNet) datasets. To avoid overfitting due to the limited number of subjects represented in the classroom data, we did not perform any finetuning on student engagement data.

Engagement Classification

For both modalities, attention and affect, we trained several classifiers. In frame-based classifiers, we trained using only the middle frame of each 1-second sequence to avoid redundant training samples when all frames were used. We additionally reasoned that kernel-based methods take a longer time to train. In the test phase, we retrieved all 1-second (24-frame) sequences' predictions and applied majority voting.

We built our models using classifiers in two categories: shallow classifiers, and Deep Neural Networks (DNN). Shallow classifiers that we used are Support Vector Machine (SVM), and Random Forests (RF) classifiers. All model training and dimensionality reduction were conducted in a person-independent manner. Considering the behavioral differences between grades, we did all experiments separately in Grade 8 and 12.

In SVM, we tested linear SVM and also SVM with radial basis function (rbf) kernel. Training SVM-based models with a large number of instances and features (i.e., 2048-dimensional features and 20-25K training samples) increases required memory and training time. Thus, before training SVM models, we applied Principal Component Analysis (PCA) and used the principal components that explain 99% of the variance in the corresponding training set. In this way, 2048-dimensional feature embeddings were reduced to the dimension of 48. In RF,

we used feature embeddings directly without dimension reduction.

In DNN's, the first approach is to use a Multi-Layer Perceptron (MLP). Here we discuss the variation of the data to emphasize why we only trained an MLP instead of retraining the entire representation up to the first layers of ResNet-50 architecture. Even though the data subset that we acquired for manual annotation and used in our analysis is over 15 hours, we still faced a problem due to the slow nature of classroom activities and limited number of subjects. Propagating the entire network results in an easy overfitting of the data and the failure to recall previously learned features useful for understanding engagement.

We used a two-layer MLP with a hidden layer in a size of 128. Training is done in mini-batches of 256 using soft-max cross-entropy loss and SGD solver with a learning rate of 0.001. In each trial, we kept a random 10% of the training data as a validation set for early stopping. In both SVM and MLP models, we applied majority voting to acquire the prediction of 1-second sequences. In addition to those approaches, we used a recurrent neural network model, long short-term memory (LSTM) [191], to directly learn on temporal data. LSTMs showed great performance in modelling long-term dependencies in various problems such as language modelling [192], neural machine translation [193], visual recognition and video action recognition [194].

In contrast to feedforward neural networks, recurrent neural networks can learn from temporal data. In learning a problem, the memory cell of a recurrent network (here, we use LSTM cell) is defined not only by the current inputs but also by longer temporal dependencies. The key contribution of LSTMs is self-loops that produce paths through which gradients can flow, reducing the chance of an exploding or vanishing gradient. LSTMs are controlled by a hidden unit and the integration of the time scale can change.

We provided 2048-length Attention-Net or Affect-Net embeddings as input to a two-layer LSTM network with a hidden size of 128. The output of the LSTM network on the last time step is fed to a fully connected layer in size of 64, and the entire model is trained using softmax cross-entropy loss and Adam solver [195] with a learning rate of 0.001. All LSTM models are trained for 5 epochs.

Personalization of Engagement Classifiers

There are two plausible use cases for an engagement estimation system that can classify the engagement levels of all students in the classroom. Such a system could be used in classroom management studies or as part of an affective and cognitive interface to help teachers understand classroom engagement and regulate teaching styles accordingly. In order to be effective, the system needs to be used many times in the same classroom. Furthermore, engagement and disengagement during instruction can differ significantly from one student to another. Thus, engagement classifiers could benefit from personalization.

The initial step is to train the engagement classifier on person-independent training data. In

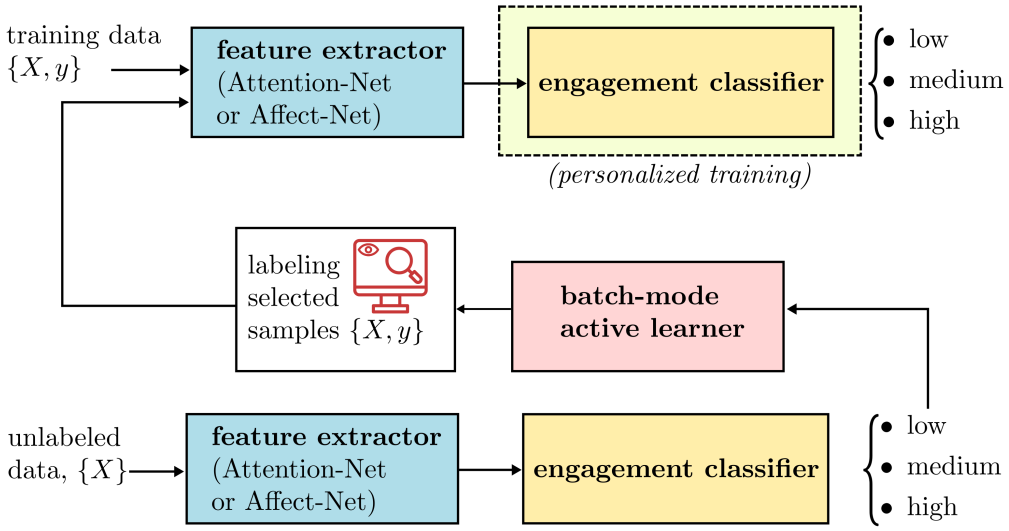


Figure A.6: Batch-mode Active Learning for Personalized Engagement Classification (The initial network is the engagement classifier trained in a person-independent manner and the weights of the feature extractors kept frozen during all experiments).

SVM-based classifiers, probability outputs can be calculated by cross-validation and Platt scaling, whereas the mean predicted class probabilities of the trees can be used in Random Forests. On the other hand, MLP and LSTM classifiers provide probability output because they were trained with softmax cross-entropy loss. These probabilities will be used to associate an uncertainty score to unlabeled instances.

Typically, traditional active learning algorithms propose a single instance to label at a time and this may result in a longer waiting time for the expert labeler during the personalization of the engagement classifier. We assume the labeler starts from an engagement classifier trained in a person-independent manner and labels a set of instances. In order to investigate the effect of

Table A.8: Performance Comparison of Engagement Classifiers on Classroom Data using Attention-Net and Affect-Net Features and Different Classifiers.

Classifier	**AUROC**			
	Grade-8		Grade-12	
	Attention-Net	*Affect-Net*	*Attention-Net*	*Affect-Net*
SVM (linear)	.560 ± .05	.570 ± .06	.656 ± .09	.563 ± .06
SVM (rbf)	.603 ± .05	.604 ± .03	.697 ± .07	.595 ± .08
RF	.620 ± .04	.608 ± .03	.708 ± .05	.600 ± .09
MLP	.615 ± .05	.597 ± .03	.701 ± .06	.622 ± .05
LSTM	.603 ± .05	.610 ± .04	.719 ± .05	.612 ± .09

personalization with a small amount of data, we utilize the margin uncertainty principle that considers the samples with the smallest margin between the first and second most likely class probabilities. These samples can be considered more difficult, and labeling them helps define a better separation among the engagement intensities. The margin uncertainty rule can be written as follows:

$$x^*_{marg} = \arg\min_x \left[P_{M_{init}}(\hat{y}_1 \mid x) - P_{M_{init}}(\hat{y}_2 \mid x) \right] \tag{A.1}$$

where $\hat{y} = P_{M_{init}}(\hat{y} \mid x)$ is the prediction with highest posterior probability, $\hat{y}_1$ and $\hat{y}_2$ are first and second most likely predictions.

Figure A.6 depicts our personalization framework using batch-mode active learning. As there is no additional training in the deep embedding part (Attention-Net and Affect-Net) on engagement at that stage, training time is not increased. Only a small batch of unlabeled data is sent to the oracle in each episode, and the classifier part is retrained. Instead of updates with a single instance, we sampled a small batch of unlabeled images to label, removed them from the pool, and retrained the initial model iteratively. In this way, each personalization step is applied on a day or a week of recording, and a batch is composed of the most qualitative samples to adapt the existing engagement classifier on a specific subject.

Results

As we report on the results of our work, it is important to note that we performed engagement classification experiments separately in grades 8 and 12 because visual engagement across grades may vary. In each grade, training and testing were conducted in a person-independent manner. With the exception of the test subject, every student in every grade was used in training and the same experiment was repeated per student, modality (affect vs. attention), and grade. Table A.8 shows the performance of various classifiers using Attention-Net and Affect-Net features. We used weighted Area Under the ROC Curve as a performance measure in the task 3 level engagement classification because it measures the performance of a classifier in different thresholds. Furthermore, it is more attune to class imbalances than metrics such as accuracy.

Engagement classification. The criteria for the manual annotation of engagement (as depicted in Figure A.3) is on a higher level and not directly related to gaze direction or facial expressions. When visual indicators were compared, Attention-Net features yielded .01 to .03 better AUC than Affect-Net in Grade 8. On the other hand, the margin between the average AUCs of Grade 12 students is more considerable; attention-net features performed .08-.11 better than Affect-Net features in Grade 12. This situation may be related to the easy distraction, movement, and increased gaze drifts characteristic of students in both grades. As a result, attention features capture engagement effective than affect features.

Another comparison is the type of classifier used to examine engagement. In the literature, shallow classifiers perform better than DNN methods in engagement and similar affective

computing problems. In our experiments, linear SVM classifiers fall behind all other classifiers (.03 to .06 in AUC). However, there is no explicit performance gain among SVM with rbf kernel, RF, and MLP classifiers across both grades and feature sets. We would expect deep learning-based methods, for instance, MLP could capture engagement better than shallow classifiers, but it is comparable to RF and SVM-rbf. This may be due to the limited sample size of the data, the multifaceted aspect of learning problems, and imbalances in feature and label distribution. As we transfer feature representations of engagement from similar tasks and large-scale corpus, better feature representations facilitate engagement classification, and the margin among the classifiers is not wide. In overall performance, the best performing engagement classifiers are the ones that depend on attention features in Grade 12 (i.e., AUC of .708 and .719 with RF and LSTM classifiers).

Looking into DNN-based classifiers, the use of temporal information during training improved the performance of MLP only in the settings of Affect-Net/Grade 8 (+.013 in AUC) and Attention-Net/Grade 12 (+.018 in AUC). The limited improvement of LSTMs can be due to the short time window (24-frame). As our continuous engagement labeling approach gathers engagement labels per second and we aim to predict engagement per second, we stuck to the same setting in all experiments and predicted 1-second intervals at a time so as not to introduce delay and produce real-time feedback for the teacher when deployed in a school setting.

Besides the average AUC performance of different feature sets and grades, there are intrapersonal variations in the performance of engagement classification. A classifier performs better in a student using affect features, whereas attention features on another student using the same classifier outperform affect features.

We tested different fusion strategies using RF engagement classifiers. Table A.9 shows the performance of feature-level and score-level fusion in Grade 8 and 12. In Grade 8, both fusion strategies yielded comparable improvement, +.012-.013 of AUC over the best performing modality (Attention-Net). On the other hand, score level fusion in Grade 12 is on par with

Table A.9: Performance Comparison of Different Fusion Strategies using Random Forest Classifiers.

Grade	**Feature Set**	Avg. AUROC
8	Attention-Net	.620
8	Affect-Net	.608
8	Feature-level Fusion	**.633**
8	Score-level Fusion	.632
12	Attention-Net	**.708**
12	Affect-Net	.600
12	Feature-level Fusion	.616
12	Score-level Fusion	.694

the performance of attention features, whereas feature-level fusion performed slightly above Affect-Net performance.

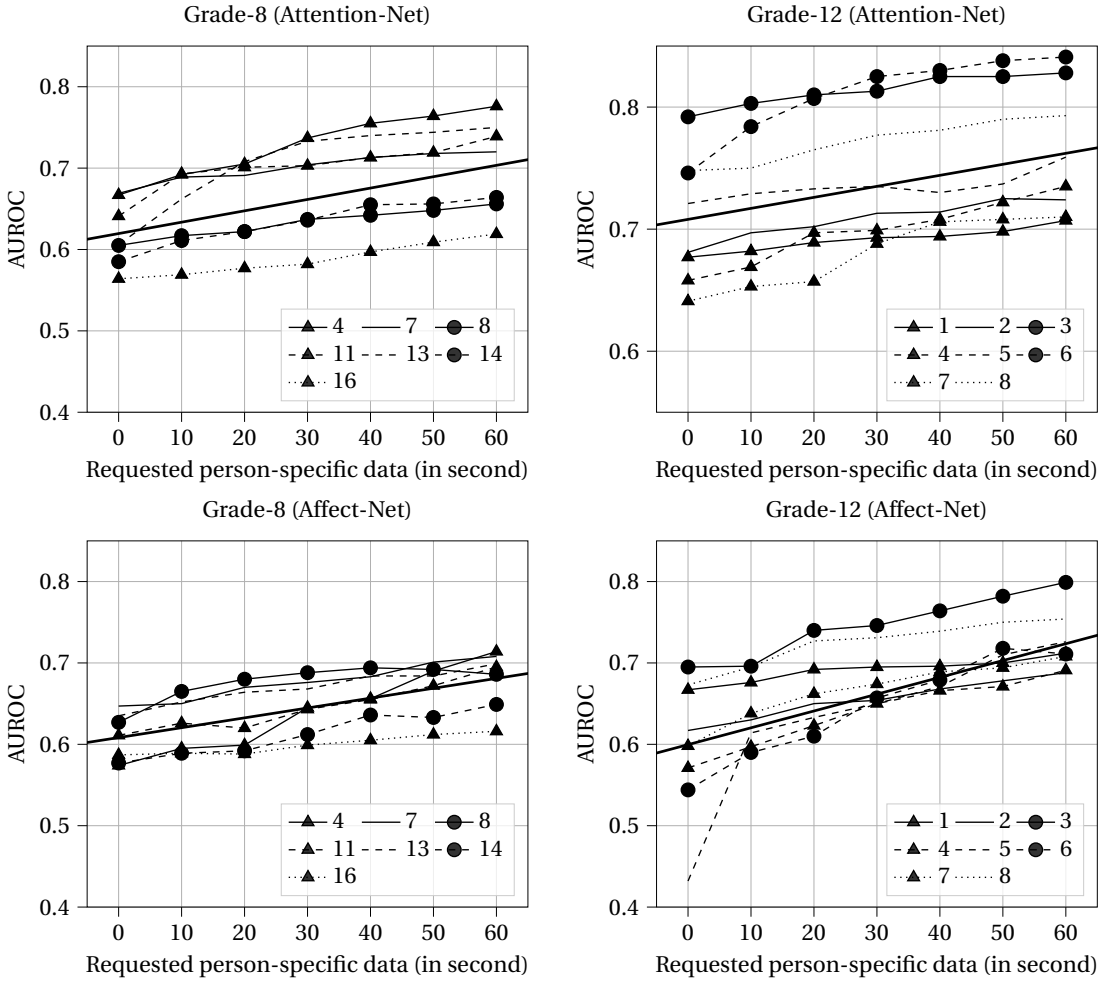


Figure A.7: The Effect of Personalization on Different Engagement Classifiers (All classifiers are based on RF. The legends show the corresponding AUC performance per student, and each thick line represents the overall trend of personalization.)

Reviewing the overall results, the average AUC ranged between .560 to .719. The performance gap between Attention-Net and Affect-Net features was rather limited (.01-0.2) in Grade 8; however, Attention-Net features outperformed Affect-Net features by a large margin, +.08-.11 of AUC. When the difficulty of interpreting a student's intensity of engagement using only facial videos is considered, these results are satisfying. Our analysis validated that student engagement could be estimated independently from the grade and course content over a long period of time.

Personalized models. In the engagement classifier's personalization, we picked RF classifiers because of their successful performance in person-independent experiments and speed in training. Both SVM classifiers with rbf kernels and DNN models take a longer time to train.

Table A.10: Confusion Matrices for the Best Person-Independent and Personalized Models.

Method	**Actual**	**Classified**			**Priors**
(Grade 12)		*low*	*medium*	*high*	
Attention-Net, RF	*low*	.099	.442	.458	.101
	medium	.053	.735	.345	.522
	high	.075	.400	.525	.377
Attention-Net, RF	*low*	.185	.387	.429	.101
(personalized)	*high*	.027	.768	.205	.522
	high	.032	.360	.608	.377

Instead of directly training and testing on person-specific data, we adapted person-independent models and checked the effect of small person-specific data in an active learning setting. The number of samples from each student varied (as depicted in Table A.7). Thus, we limited person-specific data to be requested labels by the oracle as 60 seconds for each student.

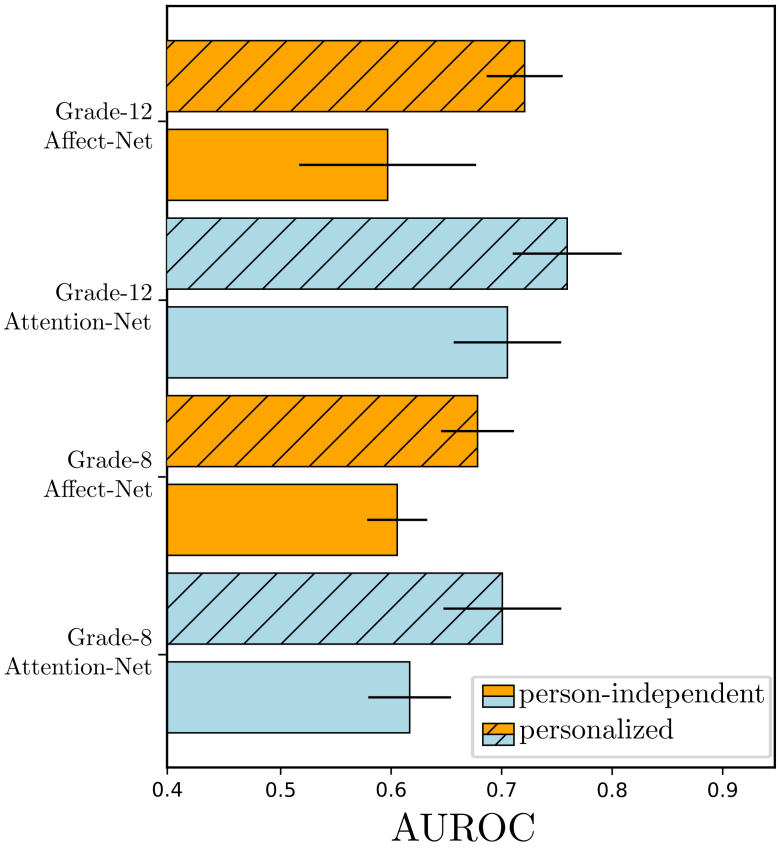


Figure A.8: The overall improvement of personalization in AUROC using Attention-Net and Affect-Net features in Grades 8 and 12.

The effect of personalization with RF engagement classifiers using Attention-Net and Affect-

Net features in Grades 8 and 12 is depicted in Figure A.7 for each student individually. Similar to previous results, we reported Area Under the ROC Curve. In each experiment, we started from the model trained in a person-independent manner, sampled 60 samples using different sampling strategies, and compared ROC performance to the initial one. The 60 samples were acquired after 6 steps by selecting only 10 samples at a time, adapting the classifier with new samples, and continuing to use the samples iteratively. As the amount of data per student changes, 60 samples correspond to different ratios of the entire person-specific data; thus, we also reported the requested (%) ratio of these samples to each student's total amount of data.

Except for one student (S4 in Grade 8), the amount of data is large enough, and the requested data (60 samples) corresponds only 2-3% of the entire data. The effect of personalization varies from .03 to .29 of AUC. Affect features performed in both grades show greater improvement after personalization. The same amount of person-specific samples causes +6.89 and +9.83 of AUROC in attention and affect features, respectively.

Even though we limited the additional data introduced during personalization to 60-seconds, personalization helped up to +.12 of AUC. Table A.10 shows the confusion matrices of RF classifier using Attention-Net features in Grade 12 before and after personalization. High engagement is misclassified mostly as medium (.400 and .360); however, low engagement's misclassification is medium and high. This can be due to the class imbalances. Of the few samples from low (10.1% in Grade 12), the worse performing class is low engagement. On the other hand, the improvement in medium and high is clearer after personalization, .735 to .768 and .525 to .608.

Powerful feature representations, Attention-Net, and Affect-Net for face images facilitated further engagement classification. Our experiments in personalization showed that performance can be improved on average +.084 by using 60 seconds of personal data. The largest improvement, as depicted in Figure A.8, is +.124 of AUC in Affect-Net features and RF classifier in Grade 12.

Labeling 60 one-second samples picked from different parts of a video is more manageable than labeling the entire recording and takes only a few minutes for an expert annotator. In return for this effort, the performance gain was substantial in both feature sets and grades. Thus, the personalization of engagement classifiers should be considered in large-scale classroom studies.

A.2.5 Discussion

Main Findings

In contrast to the previous works that used mainly handcrafted local (i.e., local binary patterns, Gabor filters) and precomputed features such as head pose or estimated facial action units, we showed that engagement as a 3-class classification problem can be predicted in the classroom.

We gathered a large-scale classroom observation dataset and collected the observer ratings of student engagement for Grades 8 and 12 (N=15). In contrast to the limited training and testing protocols in the literature, our study is the first to validate the use of automated engagement analysis in the classroom.

Our work proves that even a small amount of person-specific data could considerably enhance the performance of engagement classifiers. In comparison to the person-independent settings of many machine learning and computer vision tasks, personalization in engagement analysis significantly impacts performance. We find this to be the case because of personal differences in visible behaviors during levels of low and high engagement. Furthermore, engagement can even reveal variation in time (for instance, the indicators of engagement are not the same in different classes, i.e., math and history).

Limitations and Future Work

There are several limitations in the current paper. Permissions from educational authorities to collect audiovisual data in the classroom make the study longer, more complicated, and also limit the sample size to several classes in a school. There is no camera network available for this purpose in most school classes with the exception of some big auditoriums. Preparing the video recording setup requires approximately 20 minutes, thus limiting the number of classes that can be recorded due to tight class schedules.

The presence of cameras can put pressure on students and cause their behavior to change when they know instruction is being recorded. Collecting a significant scale of audiovisual recordings from the same classes over the course of a school year as longitudinal study could overcome these effects and allow researchers to investigate engagement in time.

Another limitation of this study is its focus on only the visible dimension of engagement. The detection of mind wandering through observation of a subject's face is a relevant emerging research topic. Combining automated methods to detect mind wandering with engagement analysis may offer a solution and yield a better understanding of students' affective and cognitive behaviors in the classroom.

Even though our study presents a step towards learning facial representations in the classroom, it was not possible to learn them on the engagement data due to the limited number of subjects. The use of self-supervision and representation learning on unlabelled classroom data may result in better representations for engagement analysis in future work. Additionally, our models failed to detect low engagement. The distribution of continuous labeling was also highly imbalanced. To solve these issues, we propose collecting more data in uncontrolled environments or, in order to obtain additional low engagement samples, employing interventions to manipulate engagement.

Applications

The reported results in this study suggest that engagement classifiers could be applied in the classroom and personalized using a small amount of data. We are hopeful that automated engagement analysis becomes a part of classroom instruction research and teacher training programs in the near future. Further improvement in the performance of engagement classification and a transition from student engagement to classroom-level analysis has the potential to make engagement analysis a more superior tool.

Data collection, storage, and privacy concerns are the most significant obstacles to large scale classroom studies. Labeling a small amount of personalized data improves the performance of engagement classifiers. Personalized models using the same feature extractor learned from the data can easily be applied to many real-time students. Instead of recording videos, such a system can record only behavioral data and substantially help increase the sample size for studies in classroom management and teacher training.

Another potential application for engagement classifiers is real-time classroom observation systems such as [162]. Affective and cognitive interfaces summarizing engagement analytics as a teaching aid can also significantly enhance teaching quality. However, ethical and societal concerns need to be carefully considered. Furthermore, advances in machine learning should address the fairness, accountability, transparency, and bias of algorithms before being deployed in such applications.

Acknowledgments

Ömer Sümer is a doctoral student at the LEAD Graduate School & Research Network, which is funded by the Ministry of Science, Research and the Arts of the state of Baden-Württemberg within the framework of the sustainability funding for the projects of the Excellence Initiative II. This work is also supported by Leibniz-WissenschaftsCampus Tübingen “Cognitive Interfaces”.

B Presentation Competence Estimation

This chapter is based on the following article:

- **Ömer Sümer**, Cigdem Beyan, Fabian Ruth, Olaf Kramer, Ulrich Trautwein, and Enkelejda Kasneci. "Estimating Presentation Competence using Multimodal Nonverbal Behavioral Cues". 2021 (under review with *ACM Transactions on Interactive Intelligent Systems*).

B.1 Estimating Presentation Competence using Multimodal Nonverbal Behavioral Cues

Abstract

Public speaking and presentation competence plays an essential role in many areas of social interaction in our educational, professional, and everyday life. Since our intention during a speech can differ from what is actually understood by the audience, the ability to appropriately convey our message requires a complex set of skills. Presentation competence is cultivated in the early school years and continuously developed over time. One approach that can promote efficient development of presentation competence is the automated analysis of human behavior during a speech based on visual and audio features and machine learning. Furthermore, this analysis can be used to suggest improvements and the development of skills related to presentation competence. In this work, we investigate the contribution of different nonverbal behavioral cues, namely, facial, body pose-based, and audio-related features, to estimate presentation competence. The analyses were performed on videos of 251 students while the automated assessment is based on manual ratings according to the Tübingen Instrument for Presentation Competence (TIP). Our classification results reached the best performance with early fusion in the same dataset evaluation (accuracy of 71.25%) and late fusion of speech, face, and body pose features in the cross dataset evaluation (accuracy of 78.11%). Similarly, regression results performed the best with fusion strategies.

B.1.1 Introduction

Public speaking requires a high caliber of eloquence and persuasion in order to convey the speaker's objective while also captivating their audience. Above all, public speaking is essential to many educational and professional aspects of life, e.g., a successful thesis defense, teaching a lecture, securing a job offer, or even presenting your research at a conference. Moreover, in the context of digital transformation and with increasing online presence (e.g., online teaching courses), the demand for tutorials related to the development of presentation competence is expanding rapidly. For example, the non-profit educational organization Toastmasters International[1], which teaches public speaking through a worldwide network of clubs, currently has more than 358K members.

Besides the actual content of a speech (the verbal cues), multiple nonverbal cues, such as prosody, facial expressions, hand gestures, and eye contact, play a significant role in engaging with, convincing, and influencing the audience [196, 197]. Various public speaking performance rubrics [15, 16, 17, 18] have been used by teachers and professors to manually asses the competence of a speech. Although the rubrics above consider a speaker's nonverbal behavior, some do not differentiate between types of nonverbal behavior (acoustic or visual). For instance, Schreiber et al. [18] include nonverbal cues as a single item: "demonstrating nonverbal

[1] https://www.toastmasters.org/

behavior that reinforces the message". While it is certainly possible for a human annotator to utilize high-inference questions when rating a performance, by employing machine learning we can further investigate fine-grained nonverbal behaviors individually and provide speakers with detailed feedback to improve their presentation skills.

With this motivation in mind, our work employs a recently proposed assessment rubric, the Tübingen Instrument for Presentation competence (TIP), whose items represent nonverbal cues in detail. Having different items for behavioral cues, such as posture, gesture, facial expressions, eye contact, and audio traits, allows for a better explainability of the strengths and weaknesses of a public speech. In contrast to the sole assessment of a speech in previous works, we can, in this way, infer the underlying behavioral factors, and enable an automated assessment, which can become an asset in (self) training.

Besides their time-consuming nature, manual assessments are prone to subjectivity. Although a proper training and simultaneous rating by multiple raters might help overcome this limitation, relying on human raters limits the number of assessments that can be done at a certain time. To tackle these problems, automatic public speaking competence estimation is necessary. Some studies in the social computing domain have therefore investigated automated assessment with regard to audio-based nonverbal features (NFs) [198, 199, 200], video-based NFs [201, 202], or with a multimodal approach as in [203, 21, 204, 202, 205, 206]. Related works that performed automated public speaking competence analysis indicate that there are different types of speeches such as scientific presentations [201, 206, 207, 208], political speeches [209, 210], and video interviews [198].

In this study, we compare three major sources of nonverbal communication: *i)* speech, *ii)* face (including head pose and gaze), and *iii)* body pose, as well as the fusion of these sources, to assess public speaking competence.The experimental analyses were conducted on informational, scientific presentations performed using visual aids and in front of a two-person audience.[2]

Our main contributions are as follows:

- We conduct an in-depth analysis of nonverbal features extracted from the face, body pose, and speech for automatic presentation competency estimation in videos when features per modality are used alone or when they are fused. The features' effectiveness is examined when they are extracted from the whole video (so-called global features) and extracted from shorter video segments (so-called local features) for classification and regression tasks. These analyses are performed for a person-independent within the same dataset, and a person-specific cross-dataset setting.
- Previous studies in the computational domain used different and non-structured evaluation instruments for presentation competence. This study validates a recently proposed

[2] Different terms, such as public speaking or presentation, were used to refer a person speaking in front of a group. In our study, we prefer using presentation and presentation competence, however, to retain the original terminology used in the previous works.

evaluation metric, Tübingen Instrument for Presentation Competence (TIP). We also present Youth Presents Presentation Competence Dataset and conduct the first analysis to compare various nonverbal features and learning models in this data using TIP measures.

- 3-minute scientific presentations are emerging as an academic genre [211, 212]. Such short scientific presentations are publicly available on the internet and can also be used in combination with automated methods to estimate presentation competence. We initially validated the usability of short scientific presentations for this purpose.

The remainder of this paper is organized as follows. Section B.1.2 reviews related work on automated public speaking competence estimation and assessment rubrics. Section B.1.3 describes the data sets and presentation competence instrument used in our analysis. In Section B.1.4, we describe the proposed method in detail. Experimental analyses, the results of classification, regression and correlation analyses and cross-data experiments are provided in Section B.1.5. Lastly, we conclude the paper and discuss the limitations and future work in Section B.1.6.

B.1.2 Literature Review

Investigating the relationship between acoustic/visual nonverbal features (NFs) and public speaking performance can contribute to the development of an automated platform for speaker training and/or assessment. Below, we review social computing literature for public speaking performance analysis. There are several studies, but they are restricted to a single type of NFs, lack the adequate sample sizes, or have no differentiation in terms of speech types. Additionally, different assessment rubrics used in psychology and education domains to measure presentation quality are discussed.

Estimating Presentation Competence

Early on, Rosenberg and Hirschberg [213] found correlations between acoustic and lexical features of charismatic speech. Their defined acoustic features were the mean, standard deviation, and maximum of the fundamental frequency (f0) and speaking rate. Lexical features were defined as the number of first-person pronouns, etc. Later, Strangert and Gustafson [214] found that speakers with more dynamic f0 range were perceived more positively during political debates. Although these works [213, 214] provide preliminary research into public speaking competence, they are limited by subjective rubrics, small datasets, and few features. In addition to acoustic features (e.g., prosody and voice quality), Scherer et al. [209] examined body, head, and hand motion-based NFs to investigate their influence on the perception of political speeches. From eye-tracking data, they found that human observers mainly concentrate on speakers' faces when viewing audio-visual recordings, but concentrate on speakers' bodies and gestures when viewing visual-only recordings.

In the education domain, the *Multimodal Learning Analytics* (MLA) data corpus comprises of 40 oral presentations of students from the challenge workshop [215], including audiovisual recordings and slides. However, the manual assessment criteria/rubrics used were not published. Using this corpus, Chen et al. [202] applied a Support Vector Machine (SVM) and gradient boosting to the combination of audio intensity, pitch, the displacement of body parts detected by Kinect sensors, head pose, and slide features (e.g., the number of pictures or grammatical errors). Using the same data, Luzardo et al. [200] utilized the slide features (e.g., the number and size of text, pictures, tables) together with the audio features (e.g., pause fillers, pitch average, pitch variation) and applied an instance-based classifier. However, their approach is not suitable for public speeches without visual aids and neglects speakers' nonverbal features. Although these studies used manually extracted verbal features, they promoted efforts for semi-automatic speaking performance assessment.

Moving towards automated presentation assessment, Haider et al. [204] focused on prosodic and gestural features to categorize presentation quality as poor vs. good. In total, 6376 audio features and 42 statistical features representing hand motions were adapted for the classification of presentation classification. More important, they demonstrated that multimodal NFs perform better than using NFs of each modality alone. Specifically, it was found that presentation quality factors highly correlate with each other. In other words, it is possible to detect visual NFs with prosody features.

Continuing in the direction of multimodal features for automated assessment, Wörtwein et al. [203] developed a model to assess and improve speaker performance. Nine items measuring behavioral indicators (e.g., body pose, the flow of speech, eye contact) were defined, and audiovisual data annotated via crowd-sourcing was proposed. A relative annotation was performed by comparing two videos displayed at the same time. Correlations between extracted NFs and behavioral indicators were shown. The extracted audio-visual NFs were also used to train and make inference with ensemble classifiers. Conversely, Pfister et al. [199] claimed that highly persuasive speech requires a display of emotions consistent with verbal content. They applied affective states recognized by audio-based NFs for public speaking skill analysis and achieved 89% and 61% classification accuracy on average and within leave-one-speaker-out cross validation, respectively.

To the best of our knowledge, Chen et al. [205] and Ramanaraynanan [206] are the only studies in the literature that utilize the public speaking competence rubric (RSCP) [18], a well established assessment rubric. The public speaking performance ratings are automatically estimated using Support Vector Regression (SVR), Random Forest (RF), and generalized linear models. They use the time-aggregated statistics and histogram of co-occurrences of NFs; head pose, gaze, facial expressions, and body locations. The main drawbacks of these studies [205, 206] are the evaluation on a limited size of data and poor performance for some items in the rubric.

Table B.1: Comparison of Assessment Rubrics for Presentation Competence.

Assessment Rubric	Target level	Item number	Seperate items per NFs	Sample (#speech)	(Interrater) Reliability
Classroom Public Speaking Assessment Carlson et al. [15]	higher education	(Form B) 5 items/ 5-point scale	✗	2	– Cronbach coefficient: from .69 to .91
Public Speaking Competency Instrument Thomson et al. [16]	higher education	20 items/ 5-point scale	✗	1	n.a.
Competent Speaker Speech Evaluation Form Morreale et al. [17]	higher education	8 items/ 3-point scale	✗	12	– Ebel's coefficient: from .90 to .94 – Cronbach coefficient: from .76 to .84
Public Speaking Competence Rubric Schreiber et al. [18]	higher education	11 items/ 5-point scale	✗	45-50	ICC: $.54 \leq r \leq .93$
Tübingen Instrument for Presentation Competence Ruth et al. [19]	high school	22 items/ 4-point scale	✓	161 *(T1)* 94 *(T2)*	– Cronbach coefficient: from .67 to .93 – ICC > .60 for 10 out of 15 items

Assessment Rubrics for Presentation Competence

The ability of an automated system to decipher and report public speaking competence is incredibly valuable. One way to realize this characteristic is to use a systematic rubric that can address each possible NF as separate items. The judgments made using such a rubric can also provide better training data and can help human observers improve their confidence and rate of decision-making [16].

Carlson and Smith-Howell [15] developed three evaluation forms for informative speeches. They tested these forms on two award-winning presenters' speeches with one speech made intentionally less informative by changing the delivery and content of the speech. These speeches were evaluated by 58 individuals using the evaluation forms. Two of the three forms showed higher inter-reliability (Cronbach's α = .83 and .91). However, any of these forms include separate items representing NFs individually. Instead, visual NFs are into one item as presentation and delivery of all visual nonverbal cues.

A more recent instrument, namely, the Competent Speaker Speech Evaluation Form [17], can be used to evaluate speeches in a class environment. It can instruct students about how to prepare and present public speeches, and can generate assessment data for the accountability-related objectives of academic institutions. In this form, the acoustic NFs are defined as vocal variety in rate, pitch, and intensity, but are still represented in a single item. Visual NFs are not even defined. This kind of assessment may be suitable for classroom evaluation purposes and training automated algorithms, but it does not help to identify what is "insufficient" and can be improved in students' individual presentations.

One of the most comprehensive assessment tools for reporting indicators of objectivity, reliability, and validity is [18]. This rubric has 11-items (nine core and two optional) with a 5-point scale (*4-advanced, 3-proficient, 2-basic, 1-minimal,* and *0-deficient*). The audio-based and video-based NFs are individually considered as: "Representing how effective the speaker uses vocal expression and paralanguage[3] to engage the audience," and "demonstrating the competence of posture, gestures, facial expressions and eye contact that supports the verbal message," respectively. These items are more informative, but NFs have still not been represented individually.

Unlike the aforementioned rubrics, Thomson and Rucker [16] described individual items regarding a speaker's speech volume, gestures, and eye contact as being relaxed and comfortable as well as voice and body expressiveness. However, this rubric lacks facial expressions and posture features.

In summary, even though these rubrics provide a suitable foundation for public speaking performance assessment, there is an absence of more fine-grained items that represent various NFs separately. A more detailed comparison of the rubrics is presented in Table B.1. In the

[3] Paralanguage is the field of study that deals with the nonverbal qualities of speech (i.e. pitch, amplitude, rate, and voice quality).

current study, we use a more detailed rubric, especially for assessing NFs, which is introduced in the next section.

Table B.2: Description of Tübingen Instrument for Presentation Competence (TIP) Items.

Item	Description
	Addressing the audience
1	... addresses the audience.
2	... has a motivating introduction.
3	... takes the listeners' questions and expectations into account.
	Structure
4	... introduces the presentation convincingly.
5	... structures transitions convincingly.
6	... ends the presentation convincingly with a conclusion.
	Language use
7	... uses examples to create a tangible portrayal of the topic.
8	... uses appropriate sentence structures for oral communication.
9	... uses technical terms appropriately.
	Body language & voice
10	... has an effective posture.
11	... employs gestures convincingly.
12	... makes eye contact with the audience convincingly.
13	... uses facial expressions convincingly.
14	... uses their voice effectively (melody, tempo, volume).
15	... uses their voice convincingly (articulation, fluency, pauses).
	Visual aids
16	... uses an appropriate amount of visual information.
17	... structures visual elements appropriately.
18	... constructs an effective interplay between the speech and visual aids.
19	... creates visual aids which are visual attractive.
20	... formulated an appropriately clear scientific question.
21	... appears confident in handling information.
22	...'s reasoning is comprehensible.

B.1.3 Assessment Rubric and Data Sets

Tübingen Instrument for Presentation Competence

The items of the Tübingen Instrument for Presentation Competence (TIP) depend on rhetorical theory and cover six faces of presentation competence: *addressing the audience, structure, language use, body language & voice, visual aids*, and *content credibility*. In total there are 22 TIP items as shown in Table B.2. All items are in a 4-point Likert-type scale (1 = not true to 4 = very true).

As we aim to investigate the nonverbal behaviors for presentation competence, in the experimental analysis provided in Section B.1.4 we only used the data corresponding to items 10-15 (i.e., body language and voice). How the corresponding ratings are used for regression and classification tasks are described in Section B.1.4.

Youth Presents Presentation Competence Dataset

The Youth Presents Presentation Competence Dataset was collected during the second (T1) and third-round (T2) of the Youth Presents contest[4], a nationwide German presentation contest for secondary school students aged 12 to 20. Informed consent was obtained from all students and their parents before the study began, and the study protocol was approved by the ethics committee of the University. Students who submitted their video presentations were first pre-assessed by a jury and then selected for the second round. In this round, they were asked to give a presentation in front of a jury on a scientific topic of their choice. Their presentations were video-recorded and constituted the first set of the Youth Presents (T1). After assessing these presentations, the best performing students were invited some weeks later to the third round. The third round included an exercise presentation under standardized conditions that had no consequences for the contest. These video-recorded presentations constitute the second set of the Youth Presents (T2).

Both sets of the Youth Presents include three-minute presentations in front of a jury consisting of two people. The presenters were using analog visual aids (e.g. poster, object, experiments, notation on the blackboard). In some aspects, the presentation tasks differed between T1 and T2. Relatively speaking, students at T1 had more time to prepare: E.g., they were allowed to make analog visual aids at home and chose the scientific content of their presentation. Students at T2 were assigned the content of their presentation (microplastics in the environment) and had 40 minutes of preparation time. Additionally, they were provided a set of text materials on the topic and visualization materials (i.e., three colored pens and six white papers for a bulletin board).

Overall, 160 students delivered a presentation in the T1 condition. 91 of those presented a second time at T2. The overall number was 251 videos and the mean age of the students is

[4] https://www.jugend-praesentiert.de/ueber-jugend-praesentiert

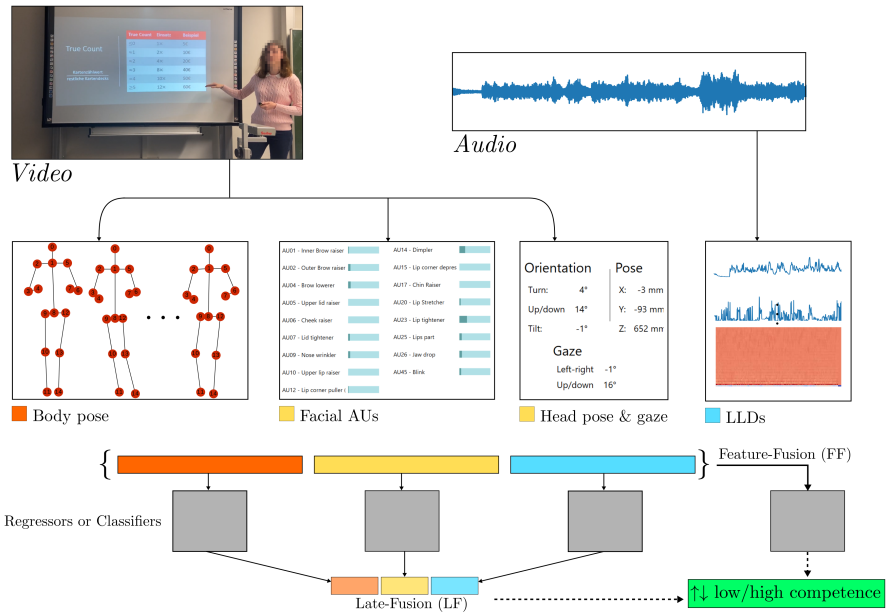


Figure B.1: Workflow of the proposed method for estimating presentation competence. Our approach uses three main modalities, body pose and facial features from the video and acoustic low-level descriptors (LLDs) from the audio. We investigate different feature fusion (FF) and late fusion (LF) strategies (The used picture is a representative of the dataset but not from the Youth Presents datasets).

15.63 years (std = 1.91). Each video was rated by four trained raters who were first introduced to the theoretical foundations of presentation competence, familiarized with the rating items, and performed exemplary ratings of video-recorded presentations that were not part of T1 and T2. During the training process, the raters discussed their ratings based on anchor examples in order to establish a common understanding of the rating items. The overall training procedure took 36 hours. After the training, each rater assessed all videos independently. The order of the videos was randomized to avoid order effects.

For each TIP item, the interrater reliability was calculated using a two-way, mixed, absolute, average-measures intraclass correlation coefficient (ICC) [216]. The results showed that among 22 items given in Table B.2, 15 items at T1 (except items 4, 8, 9, 10, 15, 17, 22) and 14 items in T2 (except 5, 7, 8, 9, 10, 17, 20, 22) exhibited ICCs above 0.60. High ICC value (> 0.60) indicates high interrater reliability and implies that the criteria rated similarly across raters.

B.1.4 Approach

This section describes our approach to estimating presentation competence from audiovisual recordings of short presentations. We formulated the problem as both classification and regression tasks. When the multimodal aspect of the problem is considered, using different

modalities is very crucial. The main features are speech features acquired from acoustic signals and facial and body pose features extracted from visual data.

Figure B.1 summarizes the main workflow of our method for estimating presentation competence. Using audio and video, we first extract nonverbal features that are relevant for the competence separately. Then, we investigate different fusion strategies, feature-level fusion (FF), and late fusion (LF) using various classifiers and regressors.

Nonverbal Feature Extraction

Speech Analysis-based NFs. Speech analysis is the most popular method to assess presentation performance [199, 202, 203, 210]. We used the state-of-the-art acoustic features extraction tool, OpenSMILE [217], to obtain the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [118], which constitutes 88 features related to the audio signal.

Facial Analysis-based NFs. Facial feature extraction consists of the following steps: face detection, facial keypoint estimation, head pose estimation, and FACS action unit occurrence and intensity estimation. We used OpenFace 2.0 [94] based on Multitask Cascaded Convolutional Networks (MTCNN) [218] for face detection, Convolutional Experts Constrained Local Model (CE-CLM) [219] for keypoint estimation and perspective n-point (PnP) matching for head pose estimation. AU analysis was performed using Histogram of Oriented Gradients (HOG) and linear kernel Support Vector Machines (SVM) on aligned face patches.

The 43 extracted facial features include the location of the head with respect to the camera in millimetres, rotation angles in radians, eye-gaze directions in radians, the estimated occurrence and intensity of the following action units: Inner brow raiser (AU1), outer brow raiser (AU2), brow lowerer (AU4), upper lid raiser (AU5), cheek raiser (AU6), lid tightener (AU7), nose wrinkler (AU9), upper lid raiser (AU10), lip corner puller (AU12), dimpler (AU14), lip corner depressor (AU15), chin raiser (AU17), lip stretcher (AU20), lip tightener (AU23), lips part (AU25), jaw drop (AU26), and blink (AU45).

Body Pose NFs. We examined the use of body pose extracted using the OpenPose algorithm [220]. OpenPose estimates the 2-dimensional locations of body joints (i.e., neck, shoulders, arms, wrists, elbows, hips) on video. Skeleton-based data is being used in various problems, for instance, video action recognition, human-computer interaction, and user interfaces, and it also helps to evaluate a presentation. Two items among the TIP labels represent body pose; these are item 10 (effective use of posture) and item 11 (employing gestures convincingly). In the context of presentation competence, using body joints instead of RGB image inputs further eliminates possible subjective bias (i.e., a presenter's visual appearance). We only used 15 joints with locations that were estimated more reliably (depicted in Figure B.1).

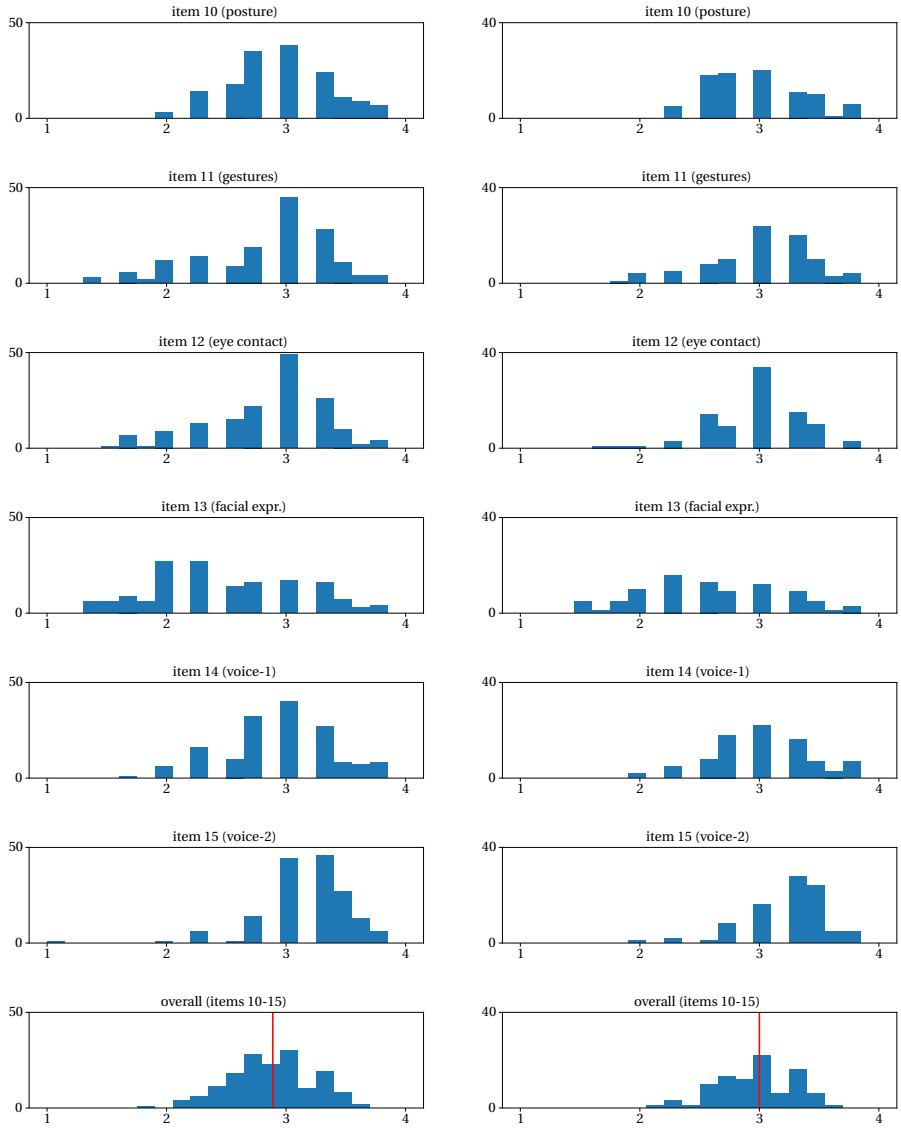


Figure B.2: Distribution of body language and voice items in T1 (on the left) and T2 (on the right) data sets. The red line on the overall plots show the median value used from discretization.

Global and Local Features

Presentation videos are rated using the TIP instrument globally per video, and the average video duration is 3 minutes. However, this duration can contain behavioral cues that contribute to improved presentation competence or vice versa. Understanding these cues in videos is extremely valuable. There are two options to achieve better understanding: use temporally global features or use temporally local features. Global features are extracted from the entire video while local features summarize behaviors during shorter intervals.

A possible use case for presentation analysis is its deployment as a recommender system in the educational domain to help students develop their presentation competence or in the field of therapy to assist people with autism spectrum disorders [22, 20, 23, 221]. In this context, localizing parts of a presentation is necessary in order to understand which parts of a presentation are effective in terms of body language and voice competence and which parts are in need of improvement. As continuous annotation of competence in videos is more time-consuming and requires raters with more advanced training, we use local features extracted from 16-second time intervals and use video-level competence items as labels.

Global features directly estimate video level competency. On the other hand, in local features we retrieve the majority vote and the median of predictions in classification and regression, respectively.

Classification & Regression

Presentation competence is a very complicated, multidimensional construct. For instance, among the TIP items shown in B.2, addressing an audience, structure, and language require some understanding of a speech's content; this is possible using natural language processing and discourse analysis. In contrast, we focus on items covering body language and voice that can be estimated through nonverbal behavior analysis.

In this study, we formulated the problem as *i)* a classification or *ii)* a regression task. While performing regression, we estimated the average of items 10-15 (i.e., the items corresponding to nonverbal communication). In classification, we discretized the ratings of the items 10-15 using the median of their distribution. In that way, we obtained two classes as high or low. In Figure B.2, the distribution of items 10-15 is given for T1 and T2 sets of Youth Presents. When T1 and T2 sets were aggregated, the median of presentation competence is 2.83; thus, we used this threshold to discretize continuous values in classification.

In total, four classifiers and regressors: Gradient Boosting (GB) [222], Decision Tree (DT) [223], Random Forest (RF) [224], and Support Vector Machines (SVM) [225, 226] were applied. These classifiers and regressors were chosen because of their use in the literature for automatic public speaking evaluation (see Section B.1.2 for more details). GB and RF were with 200 estimators. In SVM, rbf kernels and C=10 were used. In all classifiers and regressors, the data is first normalized by removing the mean and scaling to a unit variance of the training set.

Data Fusion

Estimating presentation competence necessitates understanding several modalities at the same time. Presentation competence items also cover different aspects of nonverbal behaviors. Thus, the fusion of various modalities is highly essential in the performance of presentation estimation. We compared feature-level and late fusion. Feature level fusion combines speech, face, and body pose features and trains a single classifier whereas late fusion combines decision scores of classifiers trained on different feature modalities.

We used two main fusion methods: feature fusion (FF) and late fusion (LF). In feature fusion, all input modalities are concatenated in feature-level into a single feature descriptor, and then a single classifier or regressor is trained. In late fusion, we used the median rule, product rule, and sum rule as follows:

$$\begin{aligned} P_{med}^{(i)} &= Median(P_m^i) \\ P_{prod}^{(i)} &= \prod_{m=1}^{K} P_m^i \\ P_{sum}^{(i)} &= \sum_{m=1}^{K} P_m^i \end{aligned} \tag{B.1}$$

where P is the probability retrieved from each classifier for class i. In regression tasks, we applied only median rule on the continuous predicted values from all input modalities.

B.1.5 Experimental Analysis & Results

In classification tasks, the evaluation metrics are accuracy, precision, recall, and the average F1-score. They are given as follows:

$$\begin{aligned} Accuracy &= \frac{TP+TN}{TP+TN+FP+FN} \\ Precision &= \frac{TP}{TP+FP};\ \ Recall = \frac{TP}{TP+FN} \\ F1-score &= \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \end{aligned} \tag{B.2}$$

where TP, TN, FP and FN stand for true positive, true negative, false positive and false negative, respectively. Positive class represents the high presentation competence while negative class represents the low presentation competence.

For the regression task, we used Mean Squared Error (MSE; Eq. B.3) and Pearson Correlation Coefficients ($p-values$ lower than 0.001; Eq. B.4).

$$MSE = (\frac{1}{n}) \sum_{i=1}^{n} (y_i - x_i)^2 \tag{B.3}$$

$$\rho = \frac{\text{cov}(X,Y)}{\sigma_x \sigma_y}, r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 (y_i - \overline{y})^2}} \tag{B.4}$$

where ρ; pearson coefficient value of 1 represents a perfect positive relationship, -1 a perfect negative relationship, and 0 indicates the absence of a relationship between variables x and y (i.e., distributions X and Y) while r is the Pearson Correlation estimate. In our case Y and n are the ground-truth and number of samples, respectively.

Table B.3: Estimating presentation competence using global and local features as a classification task in T1 data set (N=160). Each result is the average and the standard deviation of 10-fold cross validation. GB, DT, RF, SVM, FF, LF, S, F, BP stand for Gradient Boosting, Decision Tree, Random Forest, Support Vector Machines, feature fusion, late fusion, speech, face and body pose, respectively. The best results are emphasized in bold-face.

Classification (global features)					
Modalities	Method	Accuracy	Precision	Recall	F1-score
Speech	GB	65.62	66.23	77.36	70.66
	DT	58.13	61.75	64.49	61.61
	RF	66.25	66.98	76.49	70.40
	SVM	63.75	67.95	69.79	67.16
Face	GB	57.50	61.03	64.44	61.51
	DT	63.75	68.17	66.06	65.46
	RF	60.62	62.19	71.68	65.62
	SVM	60.00	65.42	65.07	62.48
Body Pose	GB	63.12	65.57	72.69	66.88
	DT	53.12	57.27	59.76	56.19
	RF	64.38	65.91	71.26	67.87
	SVM	59.38	60.83	68.29	63.33
Fusion, GB	FF	71.25	73.06	78.08	74.54
(**S+F+BP**)	LF (med)	66.25	66.82	79.95	71.13
	LF (prod)	66.88	67.68	79.19	71.78
	LF (sum)	66.25	66.67	79.43	71.39

Classification (local features, majority voting in video)					
Modalities	Method	Accuracy	Precision	Recall	F1-score
Speech	GB	65.62	66.35	74.46	69.49
	DT	60.00	61.91	68.24	63.69
	RF	62.50	63.85	72.28	66.84
	SVM	60.00	63.79	68.44	63.94
Face	GB	62.50	63.75	76.49	68.45
	DT	56.88	58.94	60.49	58.86
	RF	63.12	63.52	77.48	68.97
	SVM	60.00	65.25	64.81	63.00
Body Pose	GB	60.00	61.98	67.81	63.34
	DT	58.13	64.01	58.49	58.58
	RF	61.88	62.31	75.46	67.36
	SVM	65.62	67.94	71.22	67.77
Fusion, GB	FF	65.62	68.33	75.82	69.80
(**S+F+BP**)	LF (med)	66.25	65.47	84.38	72.46
	LF (prod)	66.25	66.78	83.67	72.45
	LF (sum)	65.62	65.89	83.67	72.06

The-Same-Dataset Analysis

The results reported in this section include the-same-dataset analysis such that we divided the T1 set into 10-fold so each resulting fold contains a similar number of samples belonging to high or low classes. Meanwhile, if a video (or video segment) belonging to one person exists in a training fold that person is not occurring in the corresponding test fold. Thus, the aforementioned 10-fold cross validation is person-independent.

Tables Table B.3 and Table B.4 report the classification and regression results respectively for each nonverbal feature set (speech, face, and body pose), both individually and when they are fused with feature fusion and late fusion strategies.

Presentation competence labels represent the entire video. However, the ability to estimate presentation competence in shorter time intervals is highly desirable because it can point to areas of low and high competence and would allow researchers to use the proposed methods as part of a self-regulatory tool. We chose 16-second intervals as an alternative to the global features, where all features were aggregated during the entirety of each video. Considering

that we work on 3-4 minutes presentations, using 16-second intervals is a good balance and allows having 10-15 sequences from a video on average.

The classification results in Table B.3 show that using 16-second intervals does not cause an explicit drop in classification performance. In contrast, it even further improved the accuracy and F1-scores when facial features were used. In most of the feature and classifier combinations, the best performing classifiers are GB and RF.

In feature and late fusion (Table B.3), GB classifiers are used as a reference. The performance of FF is 5.63% better in accuracy than the best performing classifier when speech features were used 65.62%. The performances of different late fusion approaches are on par. Using multimodal NFs, i.e., the fusion of all NF sets resulted in an increase in classification performance while the best results were obtained with FF.

When the effect of using global or local features is examined in terms of the best performance of each NFs group, there is no statistically significant difference. However, there is a clear performance gain when local features were used in some feature/classifier combinations, for instance, +6.24% in body pose features and SVM classifier and +5% in facial features and GB classifier.

The results of regression tasks are depicted in Table B.4. In regression, speech features are the best performing one when single modality was used. In contrast to the classification task where using local features improved the performance in some feature and classifier combinations, using local features resulted in correlation between the ground truth labels and predictions dropped significantly. In fusion, feature fusion (FF) and late fusion (LF; by using only median rule) were compared in GB regressors. FF performs better than LF (with Pearson r of 0.61 and 0.56 in both global and local features, respectively), and also beyond the best performing single modalities.

The Cross-Dataset Analysis

The cross-dataset analysis refers to using a model trained on $T1$ set to predict the $T2$ set (shown as $T1 \rightarrow T2$). The $T1 \rightarrow T2$ setting is important in order to investigate the generalizability of a model trained with the employed NFs. Additionally, we also tested the importance of rhetorical settings on the automated analysis, and, in particular, the effect of variations in presentation topics and the speakers' background as related to the presented topic. We recall here that the presentations in $T1$ set each cover different topics while $T2$ covers presentations on the same topic. In the T1 set, the speakers picked their presentation topic and had more time to prepare (implying that they might build a better background regarding the topic) while in T2 the presentation topic was assigned to the speakers with limited time to prepare.

We applied the same classifier, regressors, global, local features, FF and LF fusions for the cross-dataset experiments as in Section B.1.5. The entire T1 set was used as a training set, and the models were evaluated on 10 folds of T2 data set. Cross data set classification and

Table B.4: Estimating presentation competence using global and local features as a regression task in T1 data set (N=160). MSE is reported as the mean and the standard deviation of 10-fold cross validation. Pearson correlation coefficients are between the estimated and the ground truth values of all samples. All $p-values$ are lower than 0.001. GB, DT, RF, SVM, FF, LF, S, F, BP stand for Gradient Boosting, Decision Tree, Random Forest, Support Vector Machines, feature fusion, late fusion, speech, face and body pose, respectively.

Regression			
Modalities	Method	MSE	Pearson *r*
Speech	GB	0.09 ± 0.02	0.52
	DT	0.18 ± 0.07	0.26
	RF	0.09 ± 0.03	0.51
	SVM	0.08 ± 0.03	0.56
Face	GB	0.11 ± 0.02	0.37
	DT	0.18 ± 0.06	0.30
	RF	0.10 ± 0.02	0.44
	SVM	0.10 ± 0.03	0.46
Body Pose	GB	0.11 ± 0.04	0.37
	DT	0.20 ± 0.03	0.19
	RF	0.11 ± 0.03	0.36
	SVM	0.12 ± 0.04	0.39
Fusion, GB	FF	0.08 ± 0.02	0.61
(**S+F+BP**)	LF (med)	0.09 ± 0.03	0.51

Regression (local-features, averaged per video)			
Modalities	Method	MSE	Pearson *r*
Speech	GB	0.09 ± 0.04	0.50
	DT	0.12 ± 0.05	0.35
	RF	0.10 ± 0.04	0.43
	SVM	0.11 ± 0.04	0.38
Face	GB	0.11 ± 0.04	0.31
	DT	0.14 ± 0.07	0.19
	RF	0.10 ± 0.03	0.40
	SVM	0.11 ± 0.04	0.32
Body Pose	GB	0.10 ± 0.04	0.43
	DT	0.13 ± 0.04	0.25
	RF	0.10 ± 0.04	0.41
	SVM	0.11 ± 0.05	0.32
Fusion, GB	FF	0.08 ± 0.03	0.56
(**S+F+BP**)	LF (med)	0.09 ± 0.03	0.54

regression results are given in Table B.5 and Table B.6.

We should note that the T1 and T2 settings are different in terms of rhetorical setting; however, the T2 data set is the subset of T1 participants. Thus, our cross-dataset evaluation is not person-independent. In classification, global features' performance in all modalities is considerably lower than in the same dataset results. This is a clear sign of the effect of presentation setting on the estimation of competence.

The gap between global and local features is more visible in cross-dataset evaluation. The performance of speech and face deteriorated when local features were used. On the other hand, body pose features exhibited a 10-30% improvement in accuracy when local features composed of 16-second sequences were used. Even the weakly supervised nature of video-wise labeling is considered and the entire T1 data set is also limited in size (N=160). Using shorter trajectories further increased the size of the training set (N=1.8K) and yielded even better results than person-independent performance on the same data set, particularly in body pose features and fusion.

Looking into the cross-dataset regression results in Table B.6 using GB regressors, the use of local features negatively impacted performance (more than the performance drop from global to local features in Table B.4) with the exception of speech features which performed even better than global features. This being the case, when the problem is formulated as regression the use of local features (shorter than the length of actual labels) negatively impacts both the same-dataset and cross-dataset evaluation and should be avoided. In all regression methods,

gradient boosting regression with speech features is the best performing method that also retains a high correlation (varying from 0.50 to 0.61) with ground truth labels.

Table B.5: Classification across tasks. All models were trained on the entire T1 set and evaluated on T2 set. The average of accuracy of F1-scores in 10-folds were reported.

Modalities/Method *(global features)*	GB Accuracy / F1-score	DT Accuracy / F1-score	RF Accuracy / F1-score	SVM Accuracy / F1-score
Speech	57.89 / 56.26	56.89 / 51.99	57.00 / 48.72	66.89 / 53.83
Face	40.56 / 47.00	52.56 / 55.81	46.11 / 51.02	64.67 / 55.93
Body **P**ose	48.11 / 54.98	62.67 / 56.40	50.33 / 56.70	49.33 / 54.21
(**S+F+BP**)				
FF	48.33 / 47.55	42.67 / 42.58	49.22 / 51.81	57.00 / 51.52
LF (med)	49.22 / 57.08	62.44 / 61.21	49.22 / 53.88	63.56 / 58.91
LF (prod)	52.44 / 57.48	69.22 / 45.31	49.22 / 55.80	59.33 / 54.07
LF (sum)	50.33 / 56.75	62.44 / 61.21	49.22 / 55.80	59.33 / 54.07
Modalities/Method (local features)	GB Accuracy / F1-score	DT Accuracy / F1-score	RF Accuracy / F1-score	SVM Accuracy / F1-score
Speech	44.89 / 59.02	55.78 / 68.24	51.44 / 64.30	59.22 / 72.95
Face	60.33 / 75.09	68.00 / 79.76	68.00 / 80.67	70.22 / 82.39
Body **P**ose	79.22 / 88.31	52.78 / 66.59	79.22 / 88.31	78.11 / 87.56
(**S+F+BP**)				
FF	64.89 / 78.54	57.11 / 72.29	66.89 / 77.74	74.78 / 85.49
LF (med)	71.44 / 83.24	56.00 / 70.99	72.56 / 83.99	77.00 / 86.90
LF (prod)	68.11 / 80.75	40.44 / 50.15	70.44 / 82.41	78.11 / 87.65
LF (sum)	68.11 / 80.75	56.00 / 70.99	70.44 / 82.41	78.11 / 87.65

Table B.6: Gradient Boosting (GB) regression across task. All models were trained on the entire T1 set and evaluated on T2 set. MSE is reported as the average and standard deviation of 10-folds. Pearson correlation coefficients are between the estimated and the ground truth values of all samples in T2 data set (N=91). All $p-values$ are lower than 0.05.

Modalities	MSE	Pearson r
Global features		
Speech	0.12 ± 0.04	0.45
Face	0.14 ± 0.04	0.25
Body Pose	0.19 ± 0.04	0.21
FF	0.13 ± 0.04	0.41
LF (med)	0.13 ± 0.03	0.43
Local features		
Speech	0.12 ± 0.01	0.51
Face	0.18 ± 0.04	0.01
Body Pose	0.18 ± 0.01	0.08
FF	0.16 ± 0.01	0.25
LF (med)	0.14 ± 0.01	0.43

Which feature is better?

When all three modalities, speech, face, and body pose features, were compared, speech features outperformed face and body pose features in the same dataset evaluation. With the exception of the DT classifier or regressor, speech features consistently performed better than the other two features in both classification and regression tasks. The fact that decision trees are weaker learning models than GB, RF, and SVM is one possible explanation. Overall, speech features appear to be the most dominant nonverbal cues to estimate presentation competence.

When visual nonverbal features, face and body pose, were considered, body pose features were more efficient in most cases. The use of local features further improved the performance (for instance, GB, RF, and SVM in cross dataset classification, DT and SVM in same-dataset classification). These results indicate that finer granularity of body postures leads to a better understanding of competence. Beyond that, the labeling of prototypical body postures can further improve classification and regression performance.

B.1.6 Conclusion

This study presented an analysis of computer vision and machine learning methods to estimate presentation competence. We used audiovisual recordings of a real-world setting, the Youth Presents Presentation Competence Datasets. The dataset contained different challenges: presentation time and free selection of topics in the T1 data set and limited preparation time and predetermined topics and preparation materials in the T2 data set. We used a recently proposed instrument, Tübingen Instrument for Presentation Competence (TIP), and validated that it could be used to train automated models to estimate presentation competence.

We formulated presentation competence estimation as classification and regression tasks and conducted nonverbal analysis of presenters' behaviors. The modalities used were speech (affective acoustic parameters of voice), facial features (head pose, gaze direction, and facial action units), and body pose (the estimated locations of body joints). Classification and regression methods were gradient boosting (GB), decision trees (DT), random forests (RF), and support vector machines (SVM).

In the-same-dataset, evaluation (T1), our classification approach reached 71.25% accuracy and 74.54% F1-score when early fusion was applied. In regression, we could reach a mean squared error of 0.08 and Pearson correlation of 0.61. In both settings, the feature-level fusion strategy performed better than late fusion, combining the scores of separate models.

Training and testing in different rhetorical settings still seems difficult. Even though the T2 set contains different speeches from the same persons, having enough time to prepare and the ability to freely select a presentation topic impacts classification and regression performance.

Estimating presentation competence in a finer granularity is a key priority in the development

of recommender systems that sense the nonverbal behaviors and give feedback to the presenter. The use of shorter sequences (16-seconds) and subsequent statistics of nonverbal features aggregated in these shorter time windows does not deteriorate performance, but, rather, helps significantly in cross-dataset evaluation.

Limitation. Automated methods to estimate presentation competence can be an essential asset in education. Considering the importance of effective and successful presentation competence in academic and professional life, such systems can help students more effectively gain those competencies and provide additional support for teachers. However, the use of automated methods must comply with ethical standards and should only be deployed with the users' consent.

From the perspective of fairness, in contrast to the raw image input in many computer vision tasks, we used processed nonverbal behavioral features. For instance, the datasets and algorithms that estimate attentional features (head pose and gaze direction), emotional features (facial expressions and action units), and body pose contain various subjects representative of different demographics. Still, dataset and algorithmic fairness are highly critical issues in the current data-driven learning approaches. Beyond nonverbal feature extraction tasks, a more diverse and large-scale dataset is necessary to accurately model all behavioral differences (i.e., cultural variations) while delivering a presentation.

Future Work. In future work, we plan to increase the data scale to model all behavioral variances more accurately. The personalization of presentation competence models and development of recommender systems and user interfaces are also among future research topics.

Acknowledgements. Ömer Sümer is a doctoral student at the LEAD Graduate School & Research Network, which is funded by the Ministry of Science, Research and the Arts of the state of Baden-Württemberg within the framework of the sustainability funding for the projects of the Excellence Initiative II. This work is also supported by Leibniz-WissenschaftsCampus Tübingen "Cognitive Interfaces".

C Joint Attention, Eye-Tracking, and Data Anonymization

This chapter is based on following publications:

- Ö. Sümer, P. Gerjets, U. Trautwein and E. Kasneci, "Attention Flow: End-to-End Joint Attention Estimation," In: *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Snowmass Village, CO, USA, 2020, pp. 3316-3325, doi: 10.1109/WACV45572.2020.9093515.
- Ö. Sümer, P. Goldberg, K. Stürmer, T. Seidel, P. Gerjets, U. Trautwein, and Enkelejda Kasneci. "Teachers' Perception in the Classroom". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2018, pp. 2315-2324.
- Ö. Sümer, P. Gerjets, U. Trautwein, and E. Kasneci. "Automated Anonymisation of Visual and Audio Data in Classroom Studies". In: *AAAI Workshops*, 2020. https://arxiv.org/abs/2001.05080

C.1 Attention Flow: End-to-End Joint Attention Estimation

Abstract

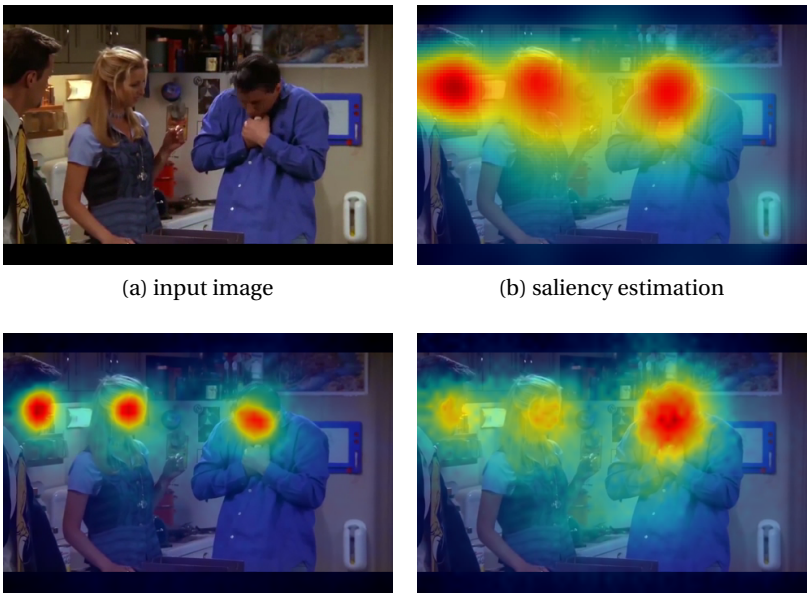
This paper addresses the problem of understanding joint attention in third-person social scene videos. Joint attention is the shared gaze behaviour of two or more individuals on an object or an area of interest and has a wide range of applications such as human-computer interaction, educational assessment, treatment of patients with attention disorders, and many more. Our method, Attention Flow, learns joint attention in an end-to-end fashion by using saliency-augmented attention maps and two novel convolutional attention mechanisms that determine to select relevant features and improve joint attention localization. We compare the effect of saliency maps and attention mechanisms and report quantitative and qualitative results on the detection and localization of joint attention in the VideoCoAtt dataset, which contains complex social scenes.

C.1.1 Introduction

Humans spend most of their lives interacting with each other. In public or private spaces such as squares, concert halls, cafes, schools, we share various aspects of everyday life with one another. Through new technologies and growing distractive effects of social media, we divide our attention and memory into separate themes and may have difficulties to focus our attention onto our primary task. In that regard, from both psychological and computer vision perspectives, understanding a person's attentional focus and particular localization of joint attention present valuable opportunities.

Joint attention is very helpful in many different contexts. For example, in classroom-based learning, teachers who engage all students equally can enhance student achievement [115, 227, 228, 229]. To investigate this, educational researchers manually analyze student behaviours and especially the visual attention of students from video recordings of instructions and try to explain relationships between students' and teachers' behaviour in a very time-consuming way. Another example is in the context of attention disorders or autism research. For instance, it has been shown that joint attention and engagement, particularly in early ages, can be taught using behavioural and developmental interventions [230]. Thus, computer vision-based, automated joint attention analysis can be instrumental in behavioural psychology to develop efficient training curricula for the treatment of children with disabilities. Another useful application is in the area of human-computer interaction and especially interaction with autonomous systems. For example, robots can infer gaze direction in case of a single person or joint attention in groups and turn their heads into that direction. Such information could be further used by robots to augment their collaboration with humans [231, 232].

Although an automated analysis of joint attention might be beneficial for a variety of applications, related work in the domain of computer vision is still quite limited. Few works addressed a similar problem, namely social saliency in first and third-person view [233, 234, 235, 120].



(a) input image (b) saliency estimation

(c) face likelihood (d) co-attention likelihood

Figure C.1: Sample of a social scene in (a), and the estimated saliency map using [104] in (b). Our method, *Attention Flow* takes only the input image in (a) and estimate the face likelihood (c) and the co-attention likelihood (d).

Also, there are examples of joint attention in human-robot interaction [231, 236, 237, 238]. Whereas mapping gaze directions to a common plane [239] is a promising option in controlled settings, it does not work in more challenging multimedia data. A recent study [75] collected a large video dataset, which we use in this study, and proposed a spatiotemporal neural network to estimate shared attention. Even though we deal with the same problem, we prefer to use the term of joint attention, since shared attention, from a psychological perspective, includes further underlying cognitive processes and does not necessitate joint gaze.

In this work, we propose a new approach that relates saliency and joint attention to estimate locations of joint attention in third person images or videos. Simply explained, saliency is an estimation of fixation likelihood on an image. In fact, due to the limited capacity of our visual system, we, by the help of an attentional mechanism, focus on the most relevant parts of a scene that are more distinctive than the remaining. In essence, it is how our eye movements process a scene, by employing various eye movements (such as saccade and fixations) and visual search which is guided by various bottom-up and top-down processes. Eye tracking-based saliency information has supported many computer vision tasks such as object detection [240], zero-shot image classification [241], and image/video captioning [242, 243].

Figure C.1 shows a sample of our approach. Despite the usefulness of saliency maps, they do not necessarily represent the visual focus of people in the scene. However, during the training time, we exploit saliency maps to encode contextual information and create pseudo attention

maps by combining them with face locations and their joint attention point and learn to predict these likelihoods. Then, during the test time, we can summarise the attentional focus of people in given third-person social images or videos.

The main contributions of this paper are as follows:

1. It formulates the problem of inferring joint attention as end-to-end training. Thereby, *Attention Flow* works without additional dependencies such as face/head detection, region proposals, or saliency estimation.
2. It explicitly learns saliency and joint attention of a high-level inference task using saliency augmented pseudo attention maps and Attention Flow network with channel-wise and spatial attention mechanisms.
3. Experimental results verify the performance of our approach on large-scale social videos, namely the VideoCoAtt dataset [75]. We also present a comparative ablation analysis of saliency and attention modules.

C.1.2 Related Work

First, we review related research on gaze following and joint attention. Then, we will discuss saliency estimation and attention modules as we utilized them in our approach to infer the joint attention.

Gaze Following. Recasens *et al.* [74] proposed a neural network which predicts the locations being gazed at in a convolutional neural network using head location, an image patch from head location, and an entire image. They also created a large-scale dataset where persons' eye and gaze locations were annotated. They later extended their work to use eye locations in a video frame and to predict gazed location in future frames [244]. Gorji *et al.* [245] used a similar approach to [74]; however, they leveraged gaze information to boost saliency estimation and did not report gaze following results.

Recently, Chong *et al.* [100] proposed a method to train gaze following, head pose and gaze tasks based on a multitask learning approach by optimizing several losses on different tasks and datasets. They also included *outside* of the frame labels and predicted visual attention. Nevertheless, their approach estimates a single person's visual attention, not joint attention.

Joint vs. Shared Attention. Joint attention is a social interaction that can occur in the forms of dyadic (looking at each other) or triadic ways (looking at each other and an object). Previous research shows that infants can discriminate joint attention interactions already by the age of 3 months [246]. Joint attention is crucial for language learning and imitative learning [247, 248]. In contrast to joint attention, shared attention does not require co-attending physically or by gaze. For instance, co-attending a television broadcast when looking at another point can be an example of shared attention. An observer can understand shared attention by using

cues from the environment [249]. Shared attention is more related to the underlying cognitive processes, whereas joint attention is dyadic and triadic gaze oriented. Thus, in the following we will use the term of joint attention since computer vision relies on seen visual cues.

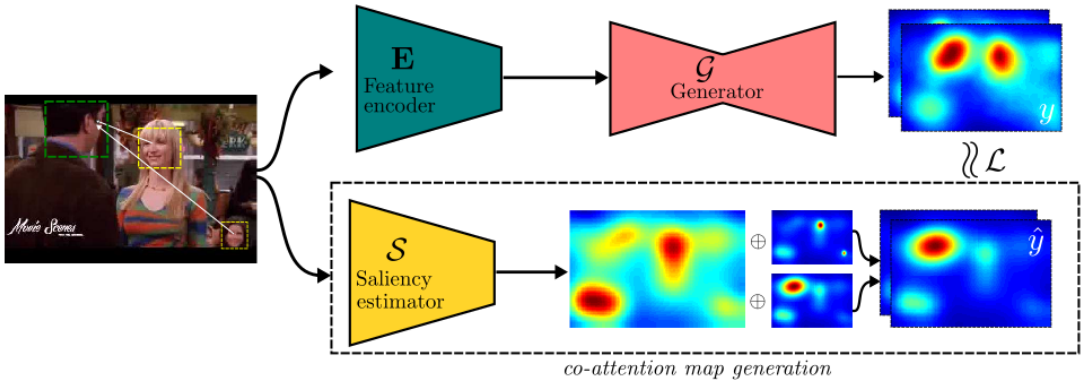


Figure C.2: *Overview of our Attention Flow* Our method is composed of three modules, (i) feature encoder, (ii) attention flow generator, and (iii) saliency-based ground truth generation. It estimates a two-channel heatmap, which encodes faces and their co-attention likelihood in the scene.

Joint Attention. Looking into studies on the analysis of attention in social interactions, [233] localized head-mounted cameras in 3D using structure from motion and triangulated joint attention. Later, they proposed a geometric model between joint attention and social formation captured from first and third person views [235]. These works are noteworthy; however, they depend on first-person views and thus cannot be applied in unconstrained third-person view images and videos. Also, they aim to predict only proximity of joint attention (social saliency) and cannot present a good understanding of joint attention.

Saliency Estimation. Saliency is a measure of spatial importance, and it characterizes the parts of the scene which stand out relative to other parts. Being salient can depend on low-level features such as luminance, color, texture, high-level features such as objectness, task-driven factor, and center bias phenomenon. In the literature of saliency estimation, two approaches exist: (a) bottom-up methods, which aim to combine relevant information without prior knowledge of the scene, and (b) top-down methods which are more goal-oriented [250]. Availability of large-scale attention datasets and deep learning approaches have surpassed all previous psychological and computational methods. Based on these recent studies, we know that humans look at humans, faces, objects, texts [251] and also emotional content [252]. The joint attention of humans in the scene is also noticeable. For this reason, we will leverage saliency information to learn joint attention.

Attention Mechanism. Computer-based estimation of attention can also be approached by means of machine-learning techniques, where models, with the help of spatial or temporal

attention mechanisms, are able to learn where, when, and what to attend. The use of First use cases are machine translation [253], image captioning [254], and action classification [255].

Looking into attention mechanisms in images, Wang *et al.* [256] incorporated attention modules into an encoder-decoder network and performed well in an image classification task. Their method learns attention jointly in 3D. Another recent work exploited inter-channel relationships. In Squeeze-and-Excitation blocks, they utilized global average-pooled features to perform a channel-wise calibration [257]. Recently, Woo et al. [258] proposed a convolutional attention module that leverages channel and spatial relations separately.

The common point of these works is that they address classification tasks by the use of spatial, temporal, or channel-wise attention. In contrast, we propose novel convolutional attention mechanisms for two purposes: the first is to learn feature selection along the channel dimension of a learned representation, and secondly, to guide a regression network to focus on more relevant areas in the spatial dimension. Instead of an architectural block in a classification task as in [257], we utilize these blocks to benefit from learned features better by applying an adaptive feature selection and apply a further refinement on top of the heatmap generation module.

C.1.3 Method

Our approach aims to infer joint attention in third person social videos, where two or more people look at another person or object. Figure C.2 shows an overview of our workflow.

For a given social image or video frame, we estimate a two-channel likelihood distribution, called *Attention Flow*. One channel represents faces in the scene, whereas the second channel is the likelihood of joint attention. In our workflow, raw images can be considered as a fusion of social presence in the scene and the center of joint attention. Figure C.2 depicts an example prediction of our approach. Our Attention Flow network takes only raw images and detects faces and their respective co-attention locations without depending on any other information. In this section, we will describe (1) the creation of pseudo-attention maps (C.1.3), which are augmented by saliency estimation; (2) learning and inference (C.1.3) by our Attention Flow network using attention mechanisms, and provide (3) implementation details (C.1.3).

Saliency Augmented Pseudo-Attention Maps

Consider persons interacting with each other in a social scene. The question we address is how to infer their visual attention focus from a third person's view? Probably the most accurate way to obtain this information would be by employing mobile eye trackers or through gaze estimation based on several high-resolution field cameras. For the majority of use cases in our daily lives, where such equipment cannot be employed, it would be very useful to be able to retrieve such information solely based on images or video material. For this reason, we first

compute pseudo-attention maps by leveraging saliency estimation.

More specifically, for an input image I, we have a number of detected head locations (x_i, y_i, w_i, h_i), where $i = 1, 2, ..., n$ and $n \geq 0$. To model social presence and the respective co-attention location we use Gaussian distributions. For a head detection or co-attention bounding box, this distribution is defined as

$$\mathscr{G}(x+\delta x, y+\delta y) = \begin{cases} \exp\{-\frac{x^2+y^2}{2\sigma^2}\} & \|\delta x\| \leq w, \|\delta y\| \leq h \\ 0, \quad otherwise \end{cases} \tag{C.1}$$

Then, we combine head locations and co-attention maps with the estimated saliency maps, which is a precursor of observer's attention. Augmented by estimated saliency maps $\mathscr{S}$, the created pseudo-attention maps can be formalized as follows:

$$\begin{aligned} \mathscr{H}_1 &= \alpha \log\Big(\sum_{i=1,..,n} \mathscr{G}_{f_i}\Big) + \beta \log(\mathscr{S}) \\ \mathscr{H}_2 &= \alpha \log(\mathscr{G}_{coatt}) + \beta \log(\mathscr{S}) \end{aligned} \tag{C.2}$$

In this way, we suppress the saliency to lower values and ensure that if there are detected faces in the scene, they and their respective co-attention point will correspond to the maximum values of pseudo-attention maps in the first and second channels.

By employing saliency estimation in our method, we leverage the information of relative importance of the regions which can be also salient for the persons in the scene. Thus, it prevents unreliable training samples, where the same object can appear as a co-attention point or zero when we use only $\mathscr{G}_f$ and $\mathscr{G}_{coatt}$.

Attention Flow Network

Our model aims to solve three problems simultaneously: (1) to locate faces in the given image, (2) to detect whether joint attention exists or not, and (3) to predict the location of joint attention.

As input we only use the raw images instead of any other computational blocks, such as face detector, object detector or proposal networks. In this way, our Attention Flow network can be used to retrieve images or videos according to their social context in an efficient and fast way. The two-channel saliency augmented pseudo-attention maps are a compressed form of these objectives and provide all necessary information. In images which do not contain faces, the first channel of the attention map will give a lower likelihood, and they can be easily omitted

from the further attention analysis.

In case of two or more persons in the scene, the first channel will represent the locations of their faces, whereas the second channel will be either estimated saliency or joint attention. Since pseudo-attention maps are a weighted summation of saliency estimation and joint attention, the typical values of maximum points are informative about the presence of joint attention. Therefore, learning pseudo-attention maps enables both detection and localization tasks simultaneously.

As it can be seen in Figure C.2, we first extract a visual representation of the scene using a pre-trained encoder network on object classification tasks. Since inferring joint attention is a complex problem even for humans, we leverage from an encoder to understand the visual focus of the persons in the image and for better generalization. The following block is a generator network, which learns attention maps from encoded representations. In order to avoid undesired outcomes of rescaling, we preserve the original aspect ratio in the input image and prefer fully convolutional architectures in both encoder and generator networks.

As a loss function, we use the Mean Squared Error (MSE) between the predicted attentions maps $\hat{\mathscr{H}}$ and ground truth pseudo saliency maps $\mathscr{H}$ (created as described in §C.1.3):

$$\mathscr{L}_{MSE} = \frac{1}{H \cdot W} \| G(E(I)), \mathscr{H} \|^2 \tag{C.3}$$

When compared to other vision tasks such as object detection, segmentation or categorization, localizing the joint attention is a very complex task because the same region, i.e., a face or an object, can be the co-attention point in a scene, but shortly for a short period of time, it might not be true. In order to deal with these situations, our Attention Flow network can be guided towards the more relevant regions. For this purpose, we propose two novel attention mechanisms, namely *channel-wise* and *spatial*, and investigate their efficiency in the localization of joint attention. Figure C.3 shows these attention mechanisms.

The encoder output is typically the output of a convolutional network which preserves spatial information in a reduced resolution and contains a higher dimension in the channel. The combination of these feature maps decides whether objects are present in the image. Using the complete encoded representation is redundant. According to the context, some channels can have more importance in the representation of the scene. Channel-wise convolutional attention performs a feature selection by weighting channels according to their contribution to the task.

On the other hand, spatial attention works as a refinement on top of the final joint attention estimations. In contrast to the spatial attention mechanisms in classification, which works as an importance map to maximize class activations, our spatial attention augments a heatmap regression task.

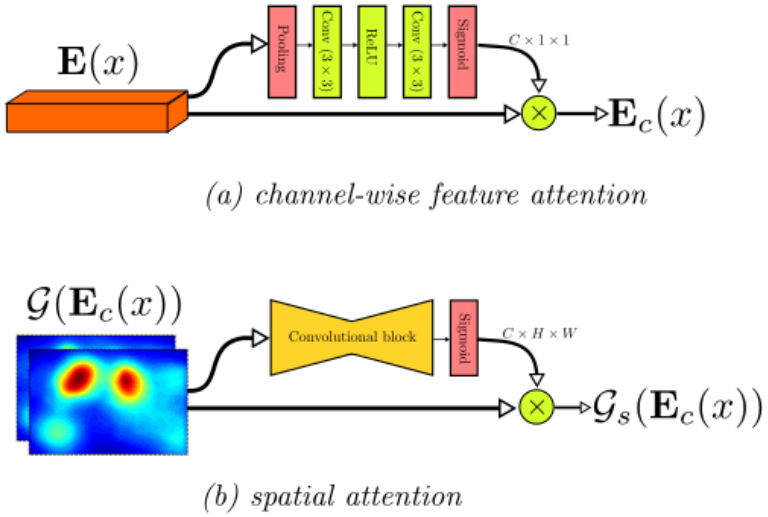


Figure C.3: Channel-wise feature attention and convolutional spatial attention blocks.

Implementation Details

The Attention Flow network is composed of three main modules: encoder, generator, and co-attention map generation blocks. In order to exploit the knowledge of large-scale object classification tasks, we use a pre-trained ResNet-50 [46] as an encoder. Our final estimation is an attention map and needs to preserve spatial relations as much as possible. Thus, we prefer dilated residual architecture, DRN-A-50 [259] trained on ImageNet and keep the resolution 1/8 at the output of the encoder.

As a generator, we used 9 residual blocks with instance normalization. It takes inputs in the number of feature channels (2048) and outputs 2-channel attention maps. Then, linear upsampling (x8) is applied.

The last block is co-attention map generation, and it is used only in training time to produce ground truth attention maps as described in §C.1.3. To estimate saliency, we used Deep Gaze II [104]. Similar to other data-driven saliency estimation methods, Deep Gaze II makes use of different level of features and has an understanding of objectness. It helps us to reduce the number of potential locations where joint attention might exist.

The layers of channel-wise feature attention are depicted in Figure C.3(a). On the other hand, in the convolutional block of spatial attention, we used a small residual network that contains 3 residual blocks. As we applied spatial attention at 1/8 resolution before upsampling, it does not introduce an extensive computational cost to the entire workflow.

At training time, we used a SGD solver with a learning rate of 0.01 in the generator block. In feature encoder, we either lock the pre-trained parameters or applied *fine-tuning* by a 10 times reduced learning rate.

C.1.4 Experiments

In this section, we first define the used dataset and performance metrics. Then, we report the ablation studies on the use of saliency estimation to create attention maps and the effect of attention mechanisms and evaluate our approach on the VideoCoAtt dataset in comparison to related approaches.

Experimental Setup

To evaluate our approach on joint attention estimation, we used the Video Co-Attention dataset [75], which is currently, to the best of our knowledge, the only available dataset on a joint attention task. The dataset contains 380 RGB video sequences from 20 different TV shows in the resolution of 320×480. There are 250,030 frames in the training set, 128,260 frames in the validation set, and 113,810 frames in the testing set. Each split comes from different TV shows, and the dataset includes varying human appearances and formation.

There are two tasks: detection and localization of joint attention. Some images might not contain human bodies or faces. In images with social content, subjects' attentional focus can be different. In the detection task, we report overall prediction accuracy in the test set of VideoCoAtt. On the other hand, localization is evaluated on the test images with joint attention locations. L_2 distance in the input resolution will be used.

By adopting the evaluation procedure from [75], we use the Structured Edge Detection Toolbox [260] to generate bounding box proposals. In the location, where our method predicts joint attention, we apply a Non-Maximum Suppression (NMS) and take the one that intersects greatest with our predicted estimation. It should be noted that our approach can locate the center of joint attention. Thus, in order to make a fair comparison with state-of-the-art methods, and we used the bounding box proposal.

Furthermore, there may be no joint attention or more than one joint attention location in an image. In order to learn the detection and localization of joint attention at the same time, we learn by all types of images without social context (body or faces), with social context but without any joint attention, and one or more joint attention. According to Eq. C.2, we limit the values of saliency to some range by natural logarithm and a scale factor. Thus, our trained network's prediction can be joint attention if and only if the predicted likelihood is greater than a threshold.

Results and Analysis

Saliency and joint attention Saliency models the attention of a third person who observes a video or image. On the other hand, joint attention analysis aims to understand from the perspectives of persons in these visual content. Due to the geometric difference between the viewpoints and human behavior in social scenes, the most salient part of images may not be

Figure C.4: Example daily life scenes from VideoCoAtt dataset [75] and their respective saliency estimations using Deep Gaze II [104]. The focus of shared visual attention does not necessarily need to be the most salient region, but contains auxiliary information to localize joint attention.

the focus of persons' attention. Thus, we investigate how the co-attention locations are salient for different saliency estimation methods.

We tested four saliency estimation methods, Itti and Koch [101], GBVS [102], Signature [103], and Deep Gaze 2 [104]. The first three were chosen as representatives of classical computational saliency methods, whereas Deep Gaze 2 is a data-driven approach that depends on pre-trained feature representations on image classification. Deep Gaze 2's mean saliency value in co-attention bounding boxes of the training images, 96% of the time, are above the mean saliency value of images, whereas it is the case in 44%, 71% and 77% for Itti & Koch, GVBS, and Signature.

In most cases, persons in the scene interact with either another person or an object. We regard that a data-driven Deep Gaze 2 can result in higher saliency in co-attention regions as it leverages a representation trained on object classification. Thus, we prefer Deep Gaze 2 when creating pseudo attention maps (§C.1.3).

Figure C.4 shows some sample images and their estimated saliency maps using Deep Gaze 2 [104], respectively. These samples show us that the most salienct regions do not necessarily

contain the possible joint attention in the social images. However, they are a precursor of observer's attention who gaze at images.

A tiny visual change in the image can cause a big change in the presence and location of joint attention. This is the main reason why we leverage the saliency information. The "raw image" results in [75] also validate our assumption. One can suppose the use of saliency as introducing noise, however, starting from the attention of observer and guiding the attention of the network towards understanding the attention of people inside the scene is a reasonable solution and also makes the problem learnable.

Use of attention mechanisms Our Attention Flow network learns joint attention by using a pre-trained representation and a generator as a regression task with mean square loss. To supplement it, we proposed two attention mechanisms. In contrast to existing attention mechanisms in the literature, such as temporal in videos or text data, or spatial in image categorization, we use two novel convolutional attention blocks for feature selection and regression tasks. We evaluate their performance on the joint attention localization task.

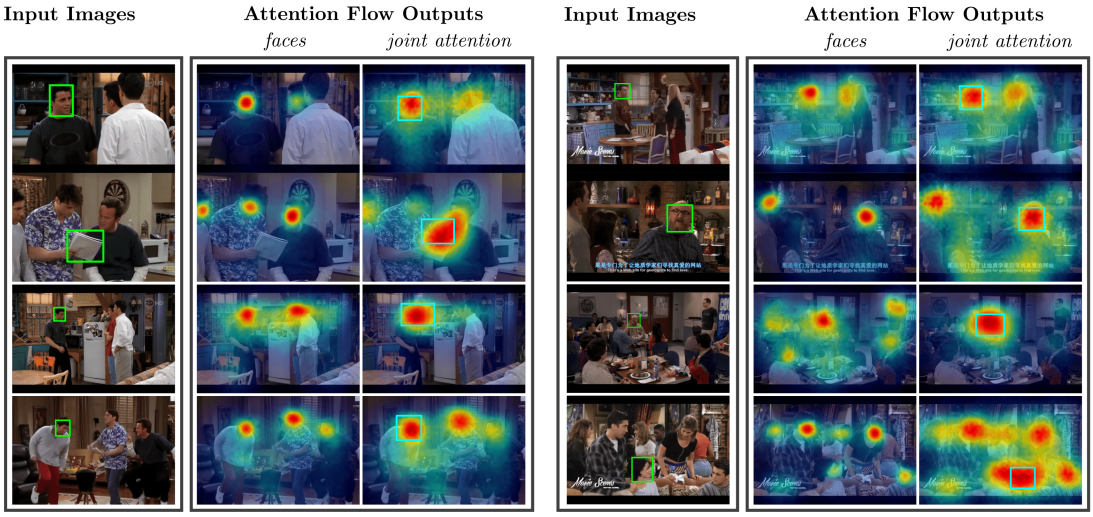


Figure C.5: **Qualitative results of Attention Flow** Bounding boxes on sample test images (green) show the ground truth attentional focus. The second and third columns are our estimated Attention Flow. In the third column, we also depicted the estimated bounding boxes (cyan). Figure best viewed in color.

The output of the dilated residual network that we used as an encoder is 1/8 resolution of the input and its channel size is 2048. The channel-wise attention module, first applies (4×6) average pooling, two convolutional layers whose kernel sizes are 3×3, 3×2 with a stride of 2 and 1, respectively. Their channel sizes are 512 and 2048, respectively. The final output is in the size of $C \times 1 \times 1$ and the original encoder output is channel-wise multiplied by these importance weights.

Table C.1 shows the results of joint attention localization over the test set of VideoCoAtt dataset.

We first tested the following options: To use the encoder as pre-trained features (no learning), to train the encoder in the same learning rate as the generator, and finetuning the encoder by a reduced (×0.1) learning rate. Channel-wise attention aims to apply feature selection in a learn representation. Thus, we freeze the encoder when training channel-wise attention and generator jointly. This approach reduces the mean L_2 distance by 10.92 and 6.88 pixels in comparison with no learning and finetuning, respectively.

Looking into our spatial attention, we applied spatial attention to the output of the generator in 1/8 resolution (40×60) before linear upsampling. Spatial attention module takes the estimation of joint attention maps ($2 \times H \times W$) and learns a spatial importance on top to better localize the co-attention point. In spatial attention, we use a 3×3 convolutional layer (64), batch normalization, a residual bottleneck module and final convolutional layer to reduce channel size back to 2. Then, a sigmoid activation is applied and the previous predictions are weighted.

Before using attention mechanisms, we show how accurate we can localize co-attention bounding boxes based on our baseline approach that is depicted in Figure C.2. After creating saliency guided pseudo-attention maps that we use as the label, our Attention Flow network has two trainable blocks: an encoder, and a generator. The encoder is initialized by ImageNet trained weights. Then, we compared the following three cases: freeze the encoder and train only generator ($E_{(lr=0)}$), train encoder and generator jointly ($E_{(base\_lr)}$), and learn encoder by transfer learning with a reduced learning rate and train generator from scratch ($E_{(finetune)}$). As it can be seen in Table C.1, transfer learning performs better than the approaches mentioned above and achieves an L_2 distance of 69.72.

Channel-wise attention, which is used between encoder and generator, can predict joint attention with a mean distance of 62.84, whereas spatial attention after the generator gives 65.70. Both attention mechanisms improve joint attention localization by 4.02 and 6.88 points with respect to our baseline network with encoder fine-tuned in Table C.1. The better performance of our channel-wise attention approach indicates that feature selection on top

Table C.1: The effect of attention mechanisms in localization of joint attention over the test set of VideoCoAtt dataset.

Method	L_2 **distance**
Attention Flow	
$E_{(lr=0)}$	73.77
$E_{(base\_lr)}$	70.47
$E_{(finetune)}$	69.72
$channel-wise\ attention$	62.84
$spatial\ attention$	65.70

of deep learning features plays an important role. Weighting features per channel improves their potential as a scene descriptor.

Table C.2 shows our results in comparison with other methods in detection and localization of joint attention. ***Random*** is acquired by drawing a Gaussian heatmap with a random mean and variance. ***Fixed Bias*** uses joint attention bias in the TV shows (averaged over the VideoCoAtt dataset) to sample predictions. An alternative to joint attention is to make prediction per person using ***Gaze Follow*** [74] and combine their attention likelihoods. Other methods are from the reference of VideoCoAtt dataset [75] and grouped into two categories: single frame and temporal models. All of these methods [75] depend on head detection bounding boxes, region proposal model or saliency estimation even in test time. In terms of used modalities, our approach is similar to their ***"Raw Image"*** approach.

Our Attention Flow network with a channel-wise attention detects joint attention with an accuracy of 78.1% over the entire test set of VideoCoAtt. Furthermore, it localizes co-attention bounding boxes with L_2 distance of 62.84. Our method performs significantly better than [75]'s single frame with region proposals and gaze estimation. Furthermore, our approach is on par with ***Gaze+RP+LSTM*** and outperforms it in terms of prediction accuracy by 6.7%.

We should note that our model makes this improvement without using any head pose/gaze estimation branch, region proposal maps, and also temporal information. [9]'s models with LSTM leverages 20-30-frame length sequences to improve and smooth prediction performance. We focused on learning an end-to-end model by using only single raw frames. Therefore, as in Table C.2, our model's performance (78.1% and 62.84) is far beyond [74] and [75]'s single frame approaches which perform at best, ***Gaze+RP***, 68.5% and 74 in joint attention detection and localization, respectively.

Table C.2: **Quantitative evaluation results** with Prediction Accuracy and *L2* Distance over the test set of VideoCoAtt dataset.

Model	**Prediction Acc.**	L_2 **distance**
Random	50.8%	286
Fixed Bias	52.4%	122
Gaze Follow [74]	58.7%	102
Raw Image [75]	52.3%	188
Only Gaze [75]	64.0%	108
Gaze+RP [75]	68.5%	74
Gaze+Saliency+LSTM [75]	66.2%	71
Gaze+RP+LSTM [75]	71.4%	**62**
Ours ***(w channel-wise att.)***	**78.1**%	62.84

Figure C.5 depicts the qualitative results of our Attention Flow network on several test images of VideoCoAtt. The ground truth co-attention locations are shown in green rectangles. Estimated face and co-attention likelihoods are overlaid on images. The first channel of our attention maps successfully locates both frontal and side faces. Looking into co-attention estimation, predictions in groups with 3-4 persons are very good. Even though their distances from ground truths are not very large, the last two examples (on the right) are relatively broad. This is due to the difficulty of scenes and a wider angle of view.

Another point that we should address is the distribution of social formations in the VideoCoAtt dataset. As the dataset is composed of acted scenes mostly from the TV shows, it does not represent the real-world formations such as in learning situations or group work. In addition, when we inspect the failures, we observed that most of them were from complicated cases where many people interest each other. Their faces were far from the camera and difficult to determine their activities (i.e., last two samples on the right side of Figure C.5). The possible direction in joint attention analysis can be to create training corpus specialized in the desired applications such as group work analysis, therapeutic situations, or children's gaze behaviors.

C.1.5 Conclusion

This study addressed a recently proposed problem, inferring joint attention in third person social videos. Our Attention Flow network infers joint attention based on only raw input images. Without using any temporal information and other dependencies such as a face detector or head pose/gaze estimator, we detect and localize joint attention better than the previous approaches. We create pseudo-attention maps by leveraging saliency information to better detect and localize joint attention. Furthermore, we propose two new convolutional attention blocks for feature selection and attention map localization. As inferring joint attention in an end-to-end fashion necessitates a high-level inference, increasing the amount of training data or the network depth will not help. We should note that these attention mechanisms, particularly channel-wise attention blocks for feature selection, are highly essential to select useful features from learned representations and improve localization performance of a heatmap regressor.

Understanding of joint attention by use of computer vision can help in a wide range of applications such as educational assessment, human-computer interactions, and therapy for attention disorders and as a future work we extend our approach to specialize in these applications and use as a tool for human behavior understanding.

Acknowledgements Ömer Sümer is a doctoral student at the LEAD Graduate School & Research Network [GSC1028], funded by the Excellence Initiative of the German federal and state governments. This work is also supported by Leibniz-WissenschaftsCampus Tübingen "Cognitive Interfaces".

C.2 Teachers' Perception in the Classroom

Abstract

The ability for a teacher to engage all students in active learning processes in classroom constitutes a crucial prerequisite for enhancing students' achievement. Teachers' attentional processes provide important insights into teachers' ability to focus their attention on relevant information in the complexity of classroom interaction and distribute their attention across students in order to recognize the relevant needs for learning. In this context, mobile eye tracking is an innovative approach within teaching effectiveness research to capture teachers' attentional processes while teaching. However, analyzing mobile eye-tracking data by hand is time consuming and still limited. In this paper, we introduce a new approach to enhance the impact of mobile eye tracking by connecting it with computer vision. In mobile eye tracking videos from an educational study using a standardized small group situation, we apply a state-of-the-art face detector, create face tracklets, and introduce a novel method to cluster faces into the number of identity. Subsequently, teachers' attentional focus is calculated per student during a teaching unit by associating eye tracking fixations and face tracklets. To the best of our knowledge, this is the first work to combine computer vision and mobile eye tracking to model teachers' attention while instructing.

C.2.1 Introduction

How do teachers manage their classroom? This question is particularly important for efficient classroom management and teacher training. To answer it, various classroom observation techniques are being deployed. Traditionally, approaches to classroom observation, such as teacher instruction and student motivation, have been from student/teacher self-reports and observer reports. However, video and audio recordings from field cameras as well as mobile eye tracking have become increasingly popular in the recent years. Manual annotation of such recorded videos and eye tracking data is very time-consuming and not scalable. In addition, it cannot be easily untangled by crowd-sourcing due to data privacy and the need of expert knowledge.

Machine learning and computer vision, with the advance of deep learning, have progressed remarkably and solved many tasks comparable with or even better than human performance. For example, literature in person detection and identification, pose estimation, classification of social interactions, and facial expressions enables us to understand fine-scale human behaviors by automatically analyzing video and audio data. Human behavior analysis has been applied to various fields, such as pedestrian analysis [261], sports [262, 263], or affective computing [264]. However, the use of automated methods in educational assessment is not so widespread.

Previous work in automated classroom behavior analysis concentrate on the activities of students using field cameras or 3D depth sensors and leveraged students' motion statistics,

head pose, or gaze [160, 154, 265, 156]. Furthermore, the engagement of students in videos has been studied in educational settings [111, 112, 113].

Students' behaviors are very important to understand the teachers' success in eliciting students' attention and keeping them engaged in learning tasks. However, the view of teachers is an underestimated perspective. How do they divide their attention among students? Do they direct the same amount of attention to all students? When a student raises her or his hands and asks a question, how do they pay attention? Such questions can be answered using mobile eye trackers and egocentric videos which are collected while instructing. Even though there are some previous studies in education sciences, they do not leverage mobile eye tracking data in depth and depend on manual inspection of recorded videos.

In this paper, we propose a framework to combine egocentric videos and gaze information provided by a mobile eye tracker to analyze the teachers' perception in the classroom. Our approach can enhance previous eye tracking-based analysis in education sciences, and also encourages future studies to work with larger sample size by providing in-depth analysis without annotation. We detect all faces in egocentric videos from teachers' eye glasses and create face tracklets from a challenging first person perspective, and eventually associate tracklets to identity. This provides us with two important information: one is whether the teacher is looking at whiteboard/teaching material or student area, and the second is which student is at the center of the teacher's attention at a specific point in time. In this way, we create the temporal statistics of a teacher's perception per student during instruction. As well as per student analysis, we integrate a gender estimation model, as an example of student characteristics, to investigate the relation between the teachers' attentional focus and students' gender [266, 267] in large scale data. Additionally, we propose teachers' movement and view of eye by use of flow information and number of detected faces.

C.2.2 Related Works

In this section we address the related works in teacher attention studies using mobile eye tracking (MET), the eye tracking in the domain of Computer Vision, attention analysis in egocentric videos, and face clustering.

Mobile eye tracking for teacher's attentional focus. The first study which links MET and high-inference assessment has been done by Cortina et al. [88]. They used fixation points and manually assigned them to a list of eight standard area of interests (e.g. black board, instructional material, student material, etc.). They investigated the variation of different skills and variables among expert and novice teachers.

Wolff et al. [105] used MET to analyze visual perception of 35 experienced secondary school teachers (experts) and 32 teachers-in-training (novices) in problematic classroom scenarios. Their work is based on Area of Interest (AOI) grid analysis, number of revisits/skips, and verbal

data (textometry). The same authors investigated in a follow-up work [106] the differences between expert and novice teacher in the interpretation of problematic classroom events by showing them short recorded videos and asking their thoughts verbally.

McIntyre and Foulsham [107] did the analysis of teachers' expertise between two cultures, in the UK and Hong Kong among 40 secondary school teachers (20 experts, 20 novices) using scanpath analysis. Scanpath is "repetitive sequence of saccades and fixations, idiosyncratic to a particular subject [person] and to a particular target pattern".

In [90], on which the paper presented here is based on their recordings, Stürmer et al. assessed the eye movements of 7 preservice teachers using fixation frequency and fixation duration in standardized instructional situations (M-Teach) [268] and real classrooms. They studied preschool teachers' focus of attention across pupils and blackboard, however their analysis also requires to predine AOI's by hand in advance.

The common point of previous studies in education sciences is that they either depend on predefined AOI's or manually annotated eye tracking output. Furthermore, none of these studies addressed the distribution of teachers' attention among students in an automated fashion. To our knowledge, none of the previous studies on teacher perception and classroom management incorporated MET and CV methodologies in order to interpret attention automatically and in a finer scale.

Eye tracking in Computer Vision. Looking into the literature, the most common use of eye tracking in CV is in the realm of saliency estimation. Saliency maps mimick our attentional focus when viewing images and are created from the fixation points of at least 20-30 observers in free-viewing or task-based/object search paradigm. Whereas initial bottom-up works in saliency estimation have used local and global image statistics go back to [269, 270, 101], the first model which measures the saliency model against human fixations in free-viewing paradigm was done by Parkhurst and Neibur [271]. The most recent state-of-the-art methods are data-driven approaches and borrow learned representations of object recognition tasks on large image datasets and adapt for saliency estimation.

Besides saliency estimation, eye tracking has been also used in order to improve the performance of various CV tasks such as object classification [240, 272], object segmentation [273], action recognition [274], zero-shot image classification [241], or image generation [275].

Attention in egocentric vision. The widespread use of mobile devices presents a valuable big data to analyze human attention during specific tasks or daily lives. Egocentric vision is an active field and there have been many works [276, 277], however there are only a few studies on gaze and attention analysis. In the realm of finescale attention analysis, particularly using eye tracking, no related work is known.

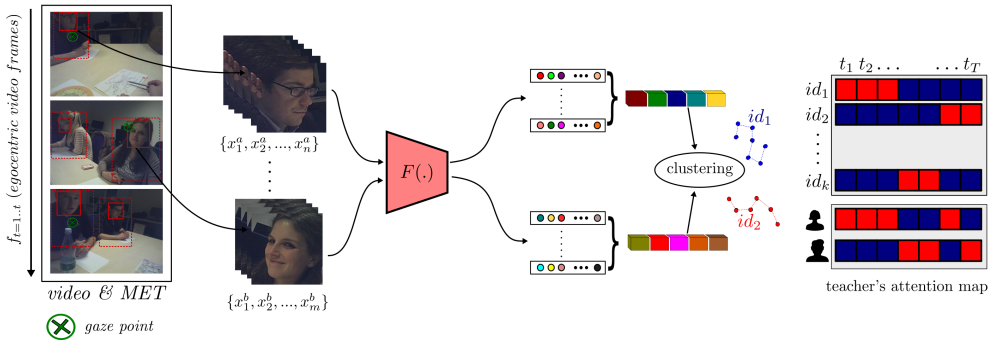


Figure C.6: Teacher's attention mapping workflow. Teachers view and gaze points are recorded by a MET while instructing. In egocentric video sequences, face detection is applied, face tracklets in video are created. Then, features are extracted and aggregated by averaging along the feature dimensions. The aggregated features are clustered. Finally, fixation points are assigned to each identity and attention maps per student identity and gender are created for whole class instruction.

Fathi et al. [278] analyzed types of social interactions (e.g. dialogue, discussion, monologue) using face detection and tracking in egocentric videos. However, their work does not include eye tracking and gaze estimation for a finescale analysis of human attention. In another work, the same authors [279] used a probabilistic generative model to estimate gaze points and recognize daily activities without eye tracking. Yamada et al. [280] leveraged bottom-up saliency and egomotion information to predict attention (saliency maps) and subsequently assesed the performance of their approach using head-mounted eye trackers. Recently, Steil et al. [281] proposed a framework to forecast attentional shift in wearable cameras. However, they exploited several computer vision algorithms as feature representation and used very specialized equipments such as stereo field cameras and head-worn IMU sensors. This make inapplicable in pervasive situations such as educational assessment.

Face clustering in videos. Face clustering is a widely studied topic and applied in still images and video tracklets, which are extracted from movies or TV series [282, 283, 284]. Many previous studies applied face detection and created low-level tracklets by merging face detections and tracking. In clustering, methods which are based on hand-crafted features exploited additional cues to create must-link and must-not-link constraints to improve representation ability of learned feature space.

The state-of-the-art deep representations are better in dealing with illumination, pose, age changes and partially occlusion and do not require external constraints. Jin et al. [285] used deep features and proposed Erdos-Renyi clustering which is based on rank-1 counts along the feature dimension of two compared images and a fixed gallery set. Recently, Nagrani and Zisserman [286] leveraged videos and voices to identify characters in TV series and movies, but they trained a classifier on cast images from IMDB or fan sites. Particularly the use of

voice, which does not happen except for question sessions and training on online cast images, make this approach unsuitable for common educational data.

Considering previous works in both fields, to the best of our knowledge this is the first work to combine mobile eye tracking and computer vision models to analyze first person social interactions for educational assessment. Furthermore, our approach presents a finescale analysis of teachers' perception in egocentric videos.

C.2.3 Method

Our goal is to detect all faces which are recorded from teacher's head mounted eye tracking glasses, create face tracklets, and cluster them by identity. Subsequently, we assign eye tracking fixations to student identities and genders when they occur in a small neighborhood of corresponding faces and body regions. Figure C.6 shows the general workflow of our proposed method. In this section, we will describe our approach to low-level tracklets linking, face representation, features aggregation, clustering, and finally, creation of teachers' attention maps while instructing.

Low-level Tracklets Linking

Students mostly sit in the same place during a classroom session, however teachers' attention is shared among whiteboard, teaching material, or a part of the student area. Furthermore, they may also walk around the classroom. Our method first start with face detection and tracklets linking.

Consider there are T video frames. We first apply Single Shot Scale-invariant Face Detector [93] in all frames and detect faces $(x_t^i)_{t=1}^T$, where i is varying number of detected faces. Then, following [287], we created face tracklets $X_K = \{x_1^{i_1}, x_2^{i_2}, ..., x_t^{i_t}\}$ are created using a two-threshold strategy. Between the detections of consecutive frames, affinities are defined as follows:

$$P_{(i,j)} = A_{loc.}(x_i, x_j) A_{size}(x_i, x_j) A_{app.}(x_i, x_j) \tag{C.4}$$

where $A(.)$ is affinities based on bounding box location, size and appearance. Detected faces between consecutive frames or shots will be associated if their affinity is above a threshold.

We adopt a low-level association, because clustering based on face tracklets instead of individial detections make subsequent face clustering more robust to outliers. Instead of a two-threshold strategy, which merges safe and reliable short tracklets, a better tracking approach can be considered. However, we observed that egocentric transition between the focuses of attention introduce motion blur and generally faces cannot be detected in succession. A significant proportion of instruction between teachers and students are in the form of dialogue or monologue. Benefiting from this situation, we can mine reliable tracklets, which contain many variations such as pose, facial expression or hand occlusion using position, size,

and appearance affinities.

Face Representation for Tracklets

Convolutional Neural Networks [26, 45, 46, 288] have become very efficient feature representation for general CV tasks and also performed well in large-scale face verification and identification tasks [289, 290]. We use and compare these methods as a face descriptor. Patricularly VGG Deep Faces [291], SphereFace [290] and VGGFace2 [292] are among the state-of-the-art methods in face recognition.

Most of these face representations require facial alignment before used in face identification. However, facial keypoint estimation is not very promising in egocentric videos. Furthermore, the image quality, even in the best scenario, is not as good as the datasets where these representations are trained. Additionally by addressing viewpoint and pose variations, we prefer ResNet-50 representation which is trained in VGGFace2 [292].

Using pre-trained networks, we extracted the feature maps of the last fully connected layers before the classifier layer. Then, feature maps are L2 normalized.

Low-level tracklets $\{X_1, ..., X_K\}$ are not of equal length. Thus, we applied element-wise mean aggregation along the feature dimension. Aggregated features are the final descriptor of tracklets and will be further used for clustering.

Face Clustering and Attention Maps

Having video face tracklets, the next step is clustering. In a general image clustering problem, number of clusters and feature representation are first needed to be decided. The number of students is given and we do not need any assumption about number of clusters (identities). When clustering, we do not leverage any must-link or must-not-link constraints, because deep feature representations are robust against various challenges such as pose, viewpoint, occlusion and illumination.

In teaching videos, we observed that the detections which cannot be associated with others in small temporal neighborhoods either belong to motion blurry frames or occluded. These samples are not representative of their identities and easily be misclassified even by human observers. On the contrary, the temporal tubes which are mined by tracklet linking have dynamics of facial features and are more discriminative. For this reason, we applied clustering on only low-level tracklets detected as described in Section C.2.3.

We used agglomerative hierarchical clustering using Ward's method. First, distance matrix between aggregated features of each tracklets $d_{ij} = d(f(X_i), f(X_j))$. Every point starts in its own cluster and greedily finds and merges closest points until there is only one cluster. Ward's linkage is based on sum-of-squares between clusters, merging cost and in each step, it keeps

the merging cost as small as possible.

We train an SVM with radial basis function [226] using aggregated tracklet features and their corresponding clustering labels. Subsequently, we predict the category of all non-tracklet detections using this model.

Having clustered tracklets and all detected faces by student identity, we can correspond teacher's focus of attention to students. MET devices deliver egocentric field video and eye tracking data. When acquiring, fixating and tracking visual stimuli, human eyes have voluntary or involuntary movements. Fixations are relatively stable moments between two saccades, fast and simultaneous movements when eye maintained gaze on a location. In attention analysis, only fixation points are used as a significant proximity of visual attention and also work load.

Eye tracking cameras are generally faster than field cameras. We use a dispersion-based fixation detection method [293] and subsequently map fixations to video frames. Then, we assign fixations to the students in case they appear in face region or body of a student. Such attention statistics enable us to better analyze and compare different teachers (i.e. expert and novice) in the same teaching situations.

Figure C.7: Examples of egocentric vision in M-Teach.

C.2.4 Experiments

To validate our approach with real teaching scenarios, we used in a first step the videos excerpts from the study of Stürmer et al. [90] in which preservice teachers' taught in standardized teaching settings (M-Teach) with a limited amount of students while wearing mobile eye tracking glasses.

7 M-Teach situations were acquired by mobile eye tracking devices (SMI - SensoMotoric

Instruments). Preservice teachers were given a topic (e.g. tactical game, transportation system) with the corresponding teaching material. Based on this material, they made preparation for instructions during 40 minutes, and then taught to a group of four students. In 20-minutes of instruction time, teachers' egocentric videos and gaze points were recorded [268].

The recorded videos are in the resolution of 1280×960 and they contain fast camera motion due to first person view. Figure C.7 depicts typical example of an egocentric sequence. In this section, our experiments will be done on this representative M-teach video about 15-minute length recorded through the eyes of a preservice teacher.

Feature Representation

Before analysis of eye tracking data, we need to identify faces of each student detected during the instruction time. To approach this, we used ResNet-50 features.

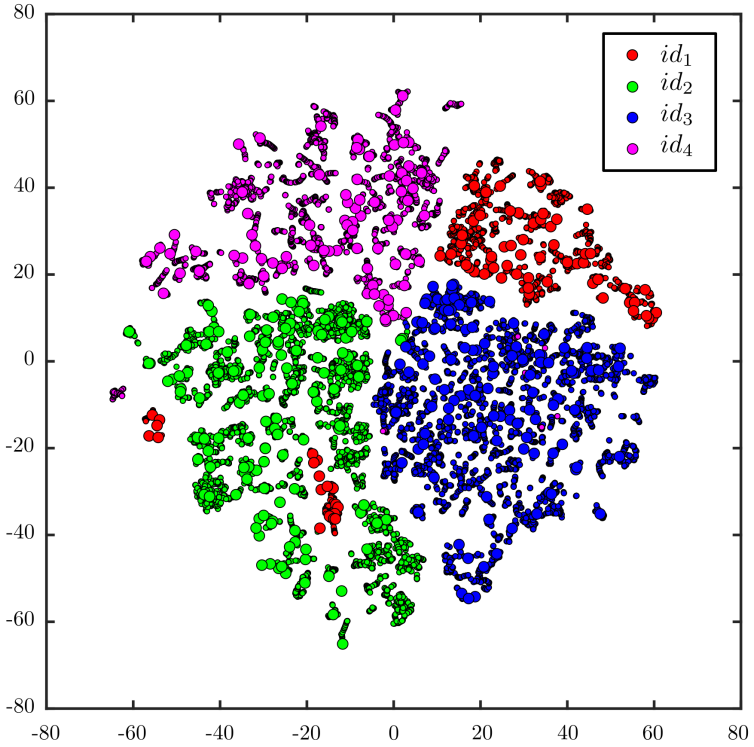


Figure C.8: t-SNE distribution of face tracklets using ResNet50/VGG2 features.

A commonly used face representation, the VGG-Face [235] network is trained on VGG-Face dataset which contains 2.6 million images. He et al. [46] proposed "deep residual networks" and it performed the state-of-the-art on the ImageNet object recognition. Recently, Cao et al. [292] collected a new face dataset, VGGFace2 whose images have large variations in pose, lightning, ethnicity, and profession. We prefered ResNet-50 network, which is pretrained on the VGGFace2 dataset. Last feature map before classification layer (2048-dimensional) is

l2-normalized and used as feature representation.

Figure C.9: Sample face tracklets which are created by low-level tracklet linking.

Figure C.8 shows t-SNE [294] distribution of faces from a M-teach instruction. Big-sized markers represent face tracklets whose deep features are aggregated by element-wise average, whereas small markers are single faces. Classroom situations are not difficult as general face recognition on unconstrained and web-gathered datasets. However, pose variation is still an issue, because the viewpoint where teachers see the students may greatly vary. Thus, we used ResNet-50 representation which is more discriminative due to the success of residual networks and also more varied training data. Feature aggregation eliminates many outliers and there are only a few misclassified tracklets in one student identity.

Figure C.9 are the examples of low-level tracklets. It can be seen that some tracklets are blurry, partially detected due to egocentric vision or contain difficult lightning conditions.

We applied agglomerative hierarchical clustering on 2048-dimensional ResNet-50 features. Subsequently, an SVM classifier trained on clustered data in order to assign the detections which cannot be associated with any tracklets. Table C.3 shows the performance of identification in a 15-minute length M-teach video.

As ResNet-50/VGG2 features are very discriminative even under varied pose, hierarchical clustering without leveraging any constraints performs well. Furthermore, SVM decision on

Table C.3: Confusion matrix of 4-student face clustering

	id_1	id_2	id_3	id_4
id_1	**1897**	8	13	0
id_2	9	**4428**	28	0
id_3	0	13	**4558**	5
id_4	0	0	92	**2958**

detections which could not be linked to any tracklets reduces false classified samples.

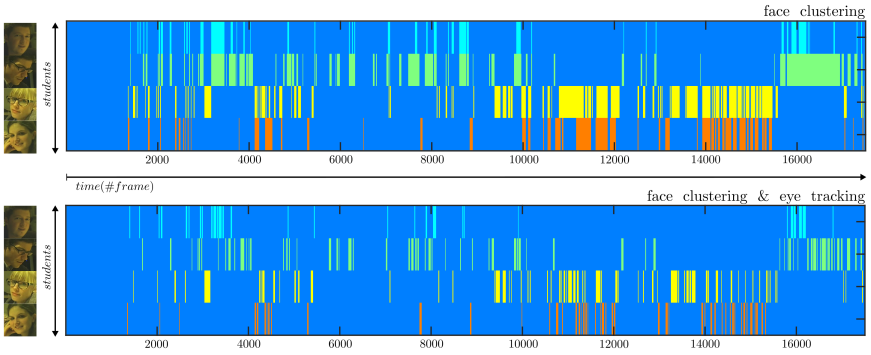


Figure C.10: Attention maps. The results of face clustering during a 15-minute M-teach situation *(above)*, fixation points are assigned to the nearest identity *(below)*.

Attention Mapping

After acquiring face tracklets, our final step is to correspond them with eye tracking data. There are four main types of eye movements: saccades, smooth pursuit movements, vergence movements, and vestibulo-ocular movements. Fixations happen between saccades and their lengths vary from 100 to 160 miliseconds. It is generally accepted that the brain processes the visual information during fixation stops. In attention analysis, therefore, mainly fixation events are used.

We extracted raw gaze points on image coordinates and calculated fixations based on a dispersion-based fixation detection algorithm [293]. In our analysis, only fixation events are used.

Figure C.10 depicts a teacher's attentional focus per student during a 15-minute M-teach instruction. First, we show the timeline of frames where each student's face is detected. In this way, we can clearly see which student(s) the teacher interacts in teaching setting. There are moments without any face detection. Teacher either looks at teaching material or explain something on the board by writing. In the second attention map of Figure C.10 represents the distribution of fixation points according to the nearest face.

After applying our workflow in 7 different M-Teach situations which were captured by different preservice teachers, we created attention maps per teacher. Then, we calculated the percentage of fixations per students from each videos separately. Figure C.10 shows that fixation frequencies vary from 40-60% to 10%. These results are consistent with Stürmer et al.'s results [90] which were based on manually defined AOI's.

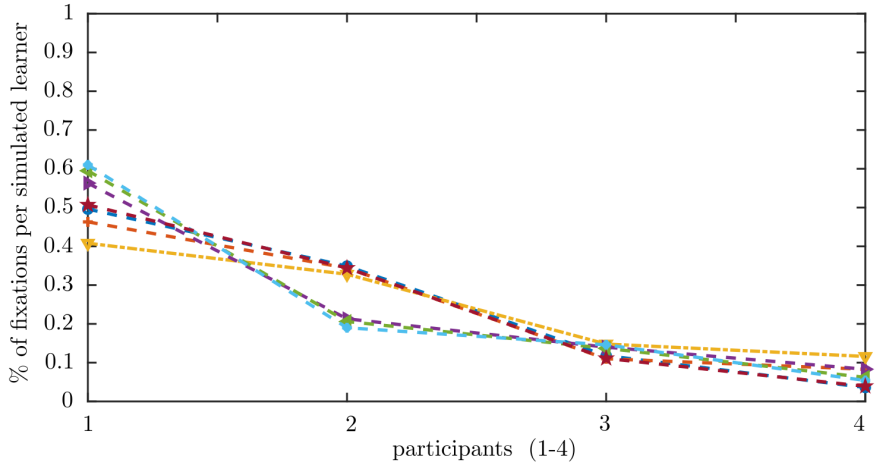


Figure C.11: Ranked scores for total fixation frequencies per student in 7 M-Teach situations (in descenting order).

C.2.5 Students' Attributes and Teacher's Attention

In automated analysis of teacher perception, another interesting question is the relation between teachers' attention and students' attributes, learning characteristics or behavior.

As an example of these attributes, we exploit gender information. Gender inequality can possibly affect the motivation and performance of students. Thus, our intuition is to extract distribution of teachers' attentional focus according to student gender as well as identity.

Having unique identity tracklets during a video recording of an instruction, one can manually label the gender of each face identity cluster. However, in large scale of data, automatic estimation of gender would be a better approach. Levi and Hassner [295] trained an AlexNet [26] on an unconstrained Adience benchmark to estimate age and gender from face images.

Using face clusters acquired as described in C.2.4, we estimated gender of all face images using [295] model. For each identity group, we consider the gender estimation of majority as our prediction and subsequently calculate the amount of teacher's eye fixations per student gender while instructing.

Table C.4 provides the ground truth number of detected faces of four students, the number

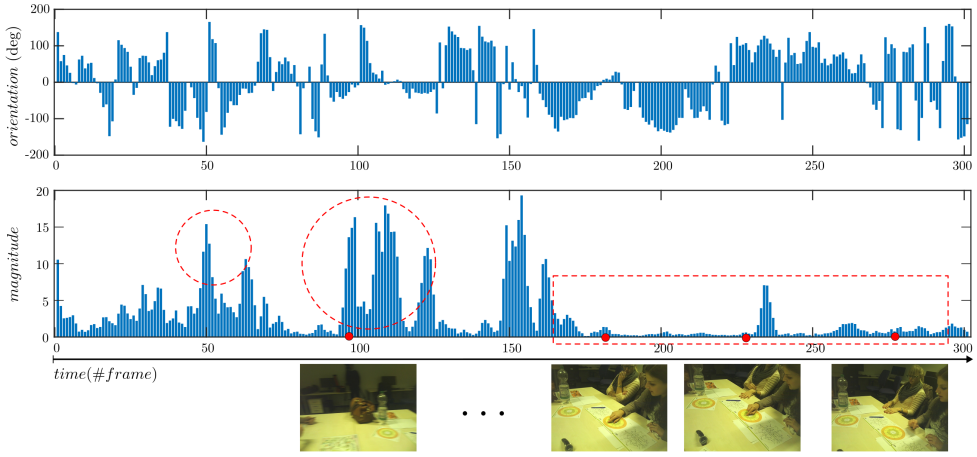


Figure C.12: In a short video snippet, mean magnitude and orientation of otpical flow are shown. Large optical flow displacement indicates that teacher's attentional focus changes. In contrast, long stable areas are indicator of an interaction with a student.

Table C.4: Gender Estimation during an M-teach video

ID/Gender	#detections (g.t.)	#predicted	gender(m/f)
ID1 (m)	1918	1906	**960**/946
ID2 (m)	4465	4449	**3321**/1128
ID3 (f)	4576	4749	879/**3870**
ID4 (f)	3050	2963	242/**2721**

of predictions from face clustering and gender estimation of all images. It can be seen that gender estimation gives accurate estimation in the majority of predicted clusters. Misclassified proportion is mainly due to blurriness of detected faces. However, we observed that gender estimation performance would be more reliable in longer sequences.

Teachers Egocentric Motion as an Additional Descriptor

As complementary to attentional focus per student identity and gender, another useful cue is teacher's egocentric motion. Some teachers may instruct without any gaze shift by looking at a constant point. Alternatively, they can move very fast among different students, teaching material and board.

Considering that M-teach situation, motion information can also give how frequent teachers' turn between left and right groups of students. For this purpose, we use mean magnitude and orientation of optical flow [296]. When using optical flow, we do not intend a high accuracy displacement between all frames of videos. Instead, we aim to spot gaze transitions between students or other source of attention. Figure C.12 shows a typical example of these cases.

Mean magnitude of optical flow becomes very large in egocentric transitions, whereas it has comparatively lower values during the dialogue with a student.

Another useful side of optical flow information is to double-check fixation points. Fixation detection methods in eye tracking can spot smooth pursuits or invalid detections as fixation. Optical flow information helps to eliminate falsely classified gaze points. In this way, we can concentrate long and more consistent time intervals in attention analysis.

C.2.6 Conclusion and Future Directions

In this study, we showed a workflow which combines face detection, tracking and clustering with eye tracking in egocentric videos during M-teach situations. In previous works in which mobile eye tracking devices were used, association of participant identities and corresponding fixations points have been done by manual processing (i.e. predefined area of interest or labeling).

We have successfully analyzed teacher's attentional focus per student while instructing. Our contribution will facilitate future works which aim at measuring teachers' attentional processes. It can also supplement previously captured mobile eye tracker recordings and provide finer scale attention maps. Furthermore, we showed that attention can be related to students' facial attributes such as gender. Our another contribution is use of flow information to discover teacher's gaze shifts and longer intervals of interaction. It particularly helps to find qualitatively important parts of long recordings.

We also aim to address following improvements on top of our proposed workflow in a future work:

1. We tested our current approach on eight 15-20 minute length M-teach videos which were recorded from the egocentric perspectives of different preservice teachers. We are planning to integrate our approach to real classroom situation which are taught by expert and novice teachers.
2. Another potential is to leverage students' levels of attention and engagement from facial images and also active speaker detection. In this manner, we can understand why teacher gazes at specific student (i.e. student asks a question or might be engaged/disengaged).
3. Fine-scale face analysis in egocentric cameras is not straightforward. In order to elude the difficulties of egocentric vision, a good solution can be to estimate viewpoint between egocentric and static field camera, and then map eye trackers gaze points into field camera. Thereby, we can exploit better quality images of stable field cameras.

Acknowledgements. Ömer Sümer and Patricia Goldberg are doctoral students at the LEAD Graduate School & Research Network [GSC1028], funded by the Excellence Initiative of the German federal and state governments. This work is also supported by Leibniz-WissenschaftsCampus

Tübingen "Cognitive Interfaces".

C.3 Automated Anonymisation of Visual and Audio Data in Classroom Studies

Abstract

Understanding students' and teachers' verbal and non-verbal behaviours during instruction may help infer valuable information regarding the quality of teaching. In education research, there have been many studies that aim to measure students' attentional focus on learning-related tasks: Based on audio-visual recordings and manual or automated ratings of behaviours of teachers and students. Student data is, however, highly sensitive. Therefore, ensuring high standards of data protection and privacy has the utmost importance in current practices. For example, in the context of teaching management studies, data collection is carried out with the consent of pupils, parents, teachers and school administrations. Nevertheless, there may often be students whose data cannot be used for research purposes. Excluding these students from the classroom is an unnatural intrusion into the organisation of the classroom. A possible solution would be to request permission to record the audio-visual recordings of all students (including those who do not voluntarily participate in the study) and to anonymise their data. Yet, the manual anonymisation of audio-visual data is very demanding. In this study, we examine the use of artificial intelligence methods to automatically anonymise the visual and audio data of a particular person.

C.3.1 Introduction

Visual and audio recording of classroom instructions has been widely used in education research for many purposes such as self-reflection, peer collaboration, teacher coaching, or classroom observation research for assessment and evaluation of teaching quality. Besides the traditional approaches in classroom observation and behaviour coding systems [13], there are recent efforts in multimodal learning analytics that aim to automate these workflows [115, 120]. The progress in the area of machine learning and artificial intelligence have additionally paved the way to conduct classroom studies in large-scale and automate these analyses. Unlike traditional types of data such as questionnaires or written reports, audio-visual recordings of a classroom cannot be easily anonymised as they lose the information which is required for further manual or automated analysis.

A typical practical challenge during such studies is that there may be students who do not want to consent in the study or their data being used by education practitioners. A common approach to overcome this limitation is to reorder the classroom by changing the seats of these persons so they can sit out of the video coverage. However, this constitutes an unnatural intervention in the usual form of classroom organisation. Furthermore, the students who do not participate in the study may raise their hands and actively participate in the ongoing instruction by speaking. Since voice also contains private information, the collection, storage, distribution, and analysis of this kind of data might cause severe violations of data protection

laws and regulations.

In fact, the current digital transformation of our society goes along with a corresponding change and regulation of data protection laws. Whereas the US data protection legislation is developing and, in comparison, Germany has the most comprehensive and the first data protection law in the world (Hessen, 1970) [297]. In Europe, the General Data Protection Regulation (GDPR) of the European Union came into effect in May 2018. According to DLA Piper[1], there are also several Asian countries which recently passed data protection laws. There is also an increasing interest in the privacy of personal data, independent of the underlying societal and individual foundations and the scope of regulations.

In August 2019, Sweden issued its first fine to a public board as they used facial recognition technology to keep track of class participation during a few weeks in a pilot study that aimed to automate the class register. Even though the data was collected and stored in local and locked computers without internet connection, according to the Swedish data protection agency, they violated the GDPR in three ways: (i) by processing personal data in a more intrusive manner than what was necessary for the purpose (monitoring of attendance), (ii) processing sensitive personal data without legal basis, and (iii) not fulfilling the requirements of data protection impact assessment and prior consultation. [298]. The Swedish example shows that even a simple use of visionary technologies may cause data breaches and that data protection in educational situations is highly critical and must be carried out after prior consultation. Contrary to the common understanding that computer vision and machine learning are used to collect and analyse private, personal data, we show that by leveraging these technologies, multimodal data can be efficiently and semi-automatically anonymised.

Let us consider an audio-visual recording of teaching situations at schools, which is typical research means in the field of classroom management since the 1990s. Under the national data protection laws and the GDPR, the current practice is to pseudo-anonymise the questionnaire data (pre- and post-tests). Pseudo-anonymisation of categorical data can be done using code words assigned to students, instead of using their names. However, audio-visual data cannot be easily anonymised because they need to either be watched by education researchers or processed by computer algorithms.

In studies that necessitates audio-visual data collection in the classroom or small group discussions, there may be subjects who do not want to participate in the research. Excluding these students may be impractical and raise other complicated issues. In contrast to longer storage and data processing, we propose to collect and store classroom data for only a short period (a few hours to max. one day). During this time, the collected audio-visual data can be automatically processed, anonymised, and validated immediately.

Our contributions. We create a small-scale classroom observation dataset that is composed of 6,5 hours of instruction in varying subjects and show the feasibility of automated anonymi-

[1] https://www.dlapiperdataprotection.com/

sation approaches on audio-visual data. Even though our focus is educational data such as classroom observation recordings our approach is applicable to any other domain of audio-visual data and makes the anonymisation much more effective.

C.3.2 Related Works

In this section, we will review the previous works that concern data anonymisation. As we mainly aim at audio-visual data anonymisation, our focus will be video and audio.

Video. Faces are perhaps the most characteristic features containing private information in video data. In the domain of computer vision, face detection, tracking, and recognition are among the widely studied problems. In the recent years, open and large-scale databases and the progress in the field of deep learning have led to fast and accurate results for face detection, tracking and recognition even under challenging conditions such as occlusion, lighting changes, eyeglasses, or make-up.

In categorical datasets, the injection of noise using differential privacy techniques is the right choice to protect private information [299]. However, these methods are not suitable for images, and in particular, for face images since the injected noise deteriorates the required face features. Even though it is far from guaranteeing privacy protection, crafting adversarial samples [300] on images can at least ensure many trained face recognisers fail on perturbed images. However, the anonymisation of faces is not only addressing privacy protection for face identification algorithms but also face recognition by humans.

In recent years, there have been many successful applications of generative models to anonymise faces, such as adversarial generative networks and variational autoencoders. For example, a face can be swapped with another person's face to hide the original identity. In [301], a face modifier network was trained together with adversarial regulariser and an action recognition network, whose task is to recognise the action even in the modified face. In this way, anonymised data can be used to train action recognition models. In contrast, [302] first filled the face that has to be anonymised with noise and applied then a conditional GAN using these patches and facial keypoints. In this manner, the authors constrained the background and head pose and guaranteed not to have any leakage of private information.

The audio-visual data is used for either automated analysis or manual inspection of behavioural data. We found two drawbacks of anonymisation based on generative models. First, they do not ensure temporally coherent generations and may induce flickering identities in some participants, which distracts the viewer who uses the material for educational purposes. Second, a face that looks like another identity does not guarantee that details in facial expressions are preserved.

In this paper, we are addressing the face in the visual domain. However, beyond faces, there are other features such as hair clothing, and body pose and they may contain private information.

In the discussion section, we present how to fulfil different levels of anonymisation.

Audio. The task of audio anonymisation aims to identify all intervals of a participant whose voice is audible. This task can be approached using speaker diarisation. In general, diarisation is composed of several sub-tasks such as speaker activity detection, speaker change detection, extraction of an embedding representation and clustering. After having speaker diarisation, we acquire clusters of speaking patterns where each of them belongs to a different participant. Then, a possible way is to ask a human annotator to locate an exemplary of the anonymised person's voice. Eventually, we can anonymise all parts assigned to the same cluster.

The focus in this field is to develop better embeddings, which in practice can be acquired from the internal representations of speaker recognition networks as in d-vectors [303] and [304]. Recently, [109] modelled speakers using multiple instances of a parameter-sharing recurrent neural network. In contrast to previous works based on clustering, they identified different speakers interleaved in time.

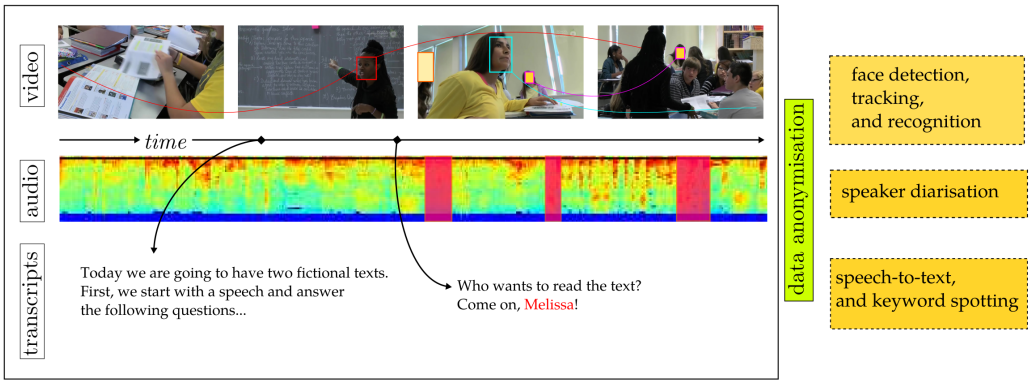


Figure C.13: Our proposed anonymisation pipeline on multimodal educational data.

Similar to changing faces with a fake identity, there exist various anonymisation approaches in the context of audio analysis. For example, [305] extracted speaker identity features from an utterance and synthesised the input by replacing the identity part with a pseudo-identity.

The most important part is to locate the speaking patterns of a person. Later, there are two possible alternatives: either to synthesise the voice by removing private acoustic parameters or directly silence these parts. We prefer the latter one because when they did not give consent to participate in the study, storage of behavioural data would not be an option even by removing identity information. Furthermore, there can be private cues in the content of their speech.

C.3.3 Approach

In practice, we have two working scenarios. (i) We are given one or several participants to anonymise their data, and (ii) we are being asked to anonymise all student data except for

the teacher's instruction (e.g. when the teacher's instruction is used for rhetorical analysis of teacher). Our audio-visual anonymisation workflow is based on multiple machine learning steps and shown in Figure C.13. The modalities that contain private information in our application are video and audio (including the spoken information during the conversations in the learning context). Depending on the level of anonymisation expected, there can be varying levels of anonymisation in visual and audio data. In this section, we describe our approach to anonymise faces in the videos and speaking patterns in the audio recordings.

Anonymising faces. Videos can be taken from handheld or static cameras. Also, they can contain footage taken from different cameras. The first step is, therefore, to locate shot boundaries and create scenes. To detect shot boundaries, including hard cuts and gradual transitions, we used the TransNet [108] approach, which is based on 3D dilated CNNs.

After detecting shot boundaries, we can define scenes, $s_i = [I^1, ..., I^{N_i}]$, where each scene s_i can be in varying length, either a few seconds or minutes. In each scene, we employ the RetinaFace face detector [95] which is trained on the WIDER face dataset [306]. On the contrary to the face detection methods in the literature that require several stages of networks, RetineFace is trained using multiple losses which combines extra-supervision and self-supervised multitask learning objectives. These additional losses include face classification, bounding box regression, extra supervision of facial landmark regression, and self-supervised mesh decoder for predicting a pixel-wise 3D shape face.

We acquire a set of faces with bounding boxes $\{[b_1^1, ..., b_{n_1}^1], ..., [b_1^t, ..., b_{n_t}^t]\}$, where t is the index of frames that belong to the scene, s_i, and n_t is a varying number of faces detected in each frame. Either in small groups or traditional classroom formation often students sit in their usual seats and do not change their seats in a scene. Furthermore, the face detector shows very decent performance even under challenging situations, such as motion blur, partial occlusion, or low image resolution. Thus, we employ a location-based tracking approach.

We first calculate the Intersection over Union (IoU) scores between the bounding boxes in the consecutive frames:

$$s(b_t, b_{t+1}) = \frac{b_t \cap b_{t+1}}{b_t \cup b_{t+1}} \tag{C.5}$$

where $b = [x, y, w, h]$'s are bounding boxes. Then, we convert the IoU scores to a cost matrix and assign the faces in the current frame to the next one using minimum weight matching in bipartite graphs that is also known as Hungarian algorithm [307]. In this way, we find correspondences that minimize the cost between the detections of consecutive frames. As we maximize the intersection between bounding boxes of the same faces, the IoU scores are negated. Even though a combination of several cost terms (i.e. additionally visual similarity between faces) can yield more reliable tracklets, in practice, we observed that the location cue is enough to create tracklets in classroom scenes.

After having face tracks, we pick only one sample of the person that we want to anonymise

Figure C.14: Snapshots from the classroom observation videos.

and use this sample as a reference for the further face verification steps. Depending on the camera angle and the subjects' movement, their head poses may show significant variation. Thus, it is crucial to pick a face track that represents these variations, for instance, the one longer and containing different head poses. As a face embedding, we use the Inception ResNet (V1) trained minimizing Triplet loss [308] on the VGGFace2 dataset [309]. Embedding vectors are L_2 normalised and their dimension is 512. To acquire the similarity of any face tracks in the video with respect to our reference, we use cosine similarities between them:

$$score = \min \frac{F(I_{(b_i,t_i)}) \cdot F(I_{(b_j,t_j)})}{||F(I_{(b_i,t_i)})||\ ||F(I_{(b_j,t_j)})||} \tag{C.6}$$

where $F(.)$ is the face embedding, $I_{b_i,t_i} \in \mathcal{T}_{(test)}$ and I_{b_j,t_j} is any face track from the same video where we want to find a specific subject whose face has to be anonymised. Eventually, we threshold the cosine similarity scores to retrieve query identity and either make these faces blurry or set them to black.

Finding speaking patterns. Similar to the face information, human voice also contains biometric traits. The main task in our context is to find the speaking patterns of a participant whose data has to be anonymised. We use the Unbounded Interleaved-State Recurrent Neural Network (UIS-RNN) [109] to diarise speaking patterns of different speakers.

An unbounded RNN network, the UIS-RNN, fed with d-vectors that is extracted from equal-length audio segments. The number of speakers is decided using a Bayesian non-parametric process, and different speakers are modelled within the states of RNN in the time domain. As the UIS-RNN approach performed better than k-means and Spectral Clustering of d-vectors on benchmark datasets, we address speaker diarisation using a pre-trained UIS-RNN model.

When we apply diarisazion to classroom data, we deploy the speaker diarisation to the spec-

trograms of the entire classroom recordings. Subsequently, we show these clustered audio parts to annotators together with the original video and ask them to pick the cluster to be anonymised.

C.3.4 Experimental Results

In order to test the feasibility of the automated anonymisation in classroom observation videos, we first created a collection of classroom observation videos from YouTube. These videos are either recorded from a static point or using handheld cameras and contain shot transitions. More specifically, the video material consists of 18 classroom instruction videos, where the taught topics include English language arts, maths, and social studies. The total duration of this instruction material is approximately 6.5 hours. All videos are available in the resolution of 1280x720px. Sample keyframes are depicted in Figure C.14.

In the face anonymisation task, we first created all face tracks. Then, on 14 participants where their faces are visible in different parts of videos, we manually labelled whether they belong to the selected test identity or not. The evaluation task is to pick a representative face track and calculate the similarity of each face track with respect to the query.

Figure C.15 shows the test results in terms of precision vs recall and receiver operating curves (ROC) per image. The Area under Curve (AUC) is 0.95, which means that our approach can recognise a given query identity with a high precision. In the performance of face identification of anonymised participant, we observed that the effect of face tracking is essential because it makes the verification more stable. Besides, using a representation that is trained on a dataset that contains large pose variations is also helpful.

With regard to the audio analysis, we segmented the audio signal of an example instruction scene from our dataset. The output of the speaker diarisation can be seen in Figure C.16. In this context, the UIS-RNN approach is quite robust to find the speaking patterns of a

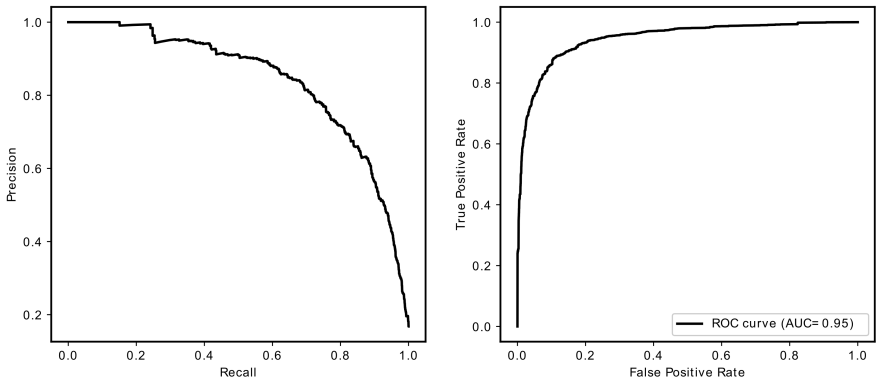


Figure C.15: Face verification results on classroom observation videos: Precision vs. Recall and ROC curves.

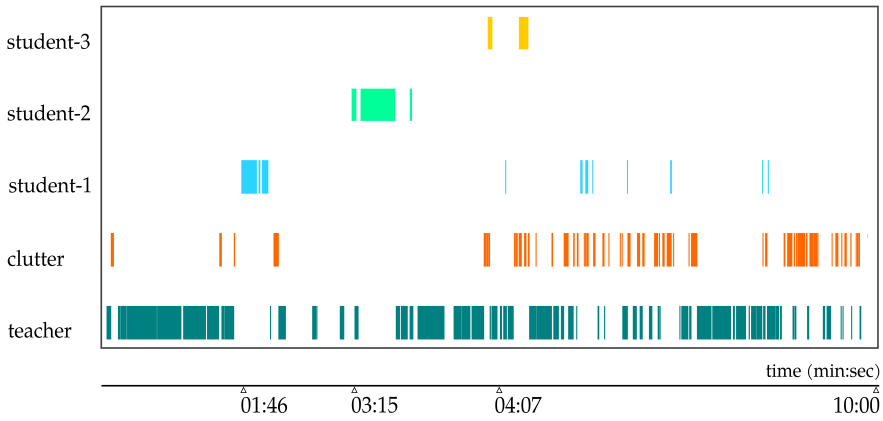


Figure C.16: An example output of speaker diarisation on a 10-minute classroom instruction from the dataset.

participant; however, in audio sequences longer than 15-20 minutes, we observed that it might over-segment some speakers, particularly the one speaking most of the time.

In order to spot the speaking pattern of a specific speaker on a long recording, such as whole day instruction with several teaching units, there are two alternative approaches. One is to use speaker diarisation and manually picking the anonymised identity in shortened clips of the audio recordings. This option requires more human effort to prove and check the quality of the diarisation. Alternatively, we may ask an annotator to pick a few samples of the anonymised person from different parts of the recording and retrieve most likely parts with respect to d-vectors. We observed that the presence of a face in the current frame is also a further precursor of the active speaker can be that person.

In addition to anonymisation, the speaking patterns of a classroom also indicate how a particular student actively participates in the learning processes during the class.

C.3.5 Discussion and Future Outlook

In this section, we first describe the levels of data anonymisation in classroom studies, summarise then our key findings and give an outlook on our future work in this context.

Different levels of data anonymisation

We focus on the automated anonymisation of video (faces) and audio data in the educational context. In practice, the level of data anonymisation and needed modalities may vary. In this part, we first address the modalities in detail and then propose different levels of anonymisation.

Questionnaires. Many educational studies, including the ones that require video recording, use questionnaires. They may contain questions regarding personal information, educational, or socio-economical background of students. Particularly personal data, such as biometric data, rate, ethnicity, and political opinion requires the highest level of protection. The most common approach to this challenge is pseudonymisation.

It should be noted, however, that the pseudonymised data may still be vulnerable to re-identification and remains therefore personal. In order to prevent re-identification, differential privacy is the current practice [310] and should be applied not only to data of specific participants who do not want to participate in the study, but to be considered for all participants.

Video. Faces in the videos are the primary modality that contains personal information. Therefore, our focus here was on finding and anonymising faces. However, there are other soft biometric traits [311] which may require anonymisation. Some examples are hair colour, body posture, and gesture patterns. When the face of an anonymised person is detected and recognised, hair can be occluded, for example by extending the face bounding box around and top of the face.

So far, we did not consider body and gestures in our analysis since this can be done by employing human pose estimation. In case of a classroom recording including approximately 20 students, deploying a multi-person pose estimation method can considerably increase the processing time. However, we will exploit this issue in our future work.

Audio. In our analysis, we tested speaker diarisation. Notably, the audio domain requires a more thorough understanding of the spoken conversation. For instance, even if we anonymised face and speaking patterns of a participant, other participants may talk about the anonymised person or content. In order to address this issue, we need a semantic understanding of the spoken text using speech-to-text approaches and efficient search techniques for the information relevant to selected keywords on the entire data.

In a manner that each level contains the anonymisation of previous modalities, we can categorise the varying levels of anonymisation for classroom studies as follows:

1. Pseudonymisation of questionnaires and personal data,
2. Anonymisation of hard biometric traits such as faces in the video and speaking time,
3. Anonymisation of soft biometric traits such as body posture, gestures, and hair appearance,
4. Anonymisation of the spoken information during the instruction unit (either by anonymised participants or other persons) such as students' and school's name or location.

The appropriate level of anonymisation can be chosen according to the nature of the data and institutional board review approval regarding ethical aspects of data collection, storage and processing.

We can summarise the key findings of our work as follows:

- We created a small classroom observation dataset that is representative of typical scenarios in educational studies.
- We investigated the feasibility of automated anonymisation on visual and audio domains. The essential modality on the visual domain is anonymisation of faces, for which we showed that it could be solved in the context of classroom observation data at high accuracy. Similarly, a speaker diarisation approach helps to find speaking clusters and those parts in the audio signal belonging to pointing to participants that have to be anonymised.

As it is the case in most vision and machine learning applications, none of the methods can guarantee completely accurate predictions. Furthermore, the owner of the data will be ethically and legally responsible for the proper and accurate anonymisation of the visual and audio modalities. Our proposed approach can be seen as an alternative to speed up anonymisation in a way that requires only quick manual inspection and correction.

For further manual inspection, we can export the output of face identification and speaker diarisation results in a compatible format with common multimedia annotation tools such as the VGG Image Annotator [312] or the ELAN software [313].

In future work, we plan to create a simple interface that is compatible with available multimedia annotation tools and can be used by educational practitioners without any technical experience and to conduct a user study on real multimodal collection and analysis process of educational data.

Acknowledgements Ömer Sümer is a doctoral student at the LEAD Graduate School & Research Network [GSC1028], funded by the Excellence Initiative of the German federal and state governments. This work is also supported by Leibniz-WissenschaftsCampus Tübingen "Cognitive Interfaces".

Bibliography

[1] R. D. Axelson and A. Flick, "Defining student engagement," *Change: The Magazine of Higher Learning*, vol. 43, no. 1, pp. 38–43, 2010. [Online]. Available: https://doi.org/10.1080/00091383.2011.533096

[2] J. A. Fredricks, P. C. Blumenfeld, and A. H. Paris, "School engagement: Potential of the concept, state of the evidence," *Review of Educational Research*, vol. 74, no. 1, pp. 59–109, 2004. [Online]. Available: https://doi.org/10.3102/00346543074001059

[3] S. L. Christenson and C. Reschly, Amy L. Wylie, *Handbook of research on student engagement.*, ser. Handbook of research on student engagement. New York, NY, US: Springer Science + Business Media, 2012. [Online]. Available: https://doi.org/10.1007/978-1-4614-2018-7

[4] N. Cowan, "Working memory underpins cognitive development, learning, and education," *Educational Psychology Review*, vol. 26, no. 2, pp. 197–223, 2014. [Online]. Available: http://www.jstor.org/stable/43549792

[5] J. Sweller, J. J. Van Merrienboer, and F. G. Paas, "Cognitive architecture and instructional design," *Educational psychology review*, vol. 10, no. 3, pp. 251–296, 1998.

[6] R. E. Mayer, *Multimedia learning, 2nd ed.*, ser. Multimedia learning, 2nd ed. New York, NY, US: Cambridge University Press, 2009. [Online]. Available: https://doi.org/10.1017/CBO9780511811678

[7] M. Kunter and U. Trautwein, "Psychologie des unterrichts (standardwissen lehramt, bd. 3895)," *Paderborn: Ferdinand Schöningh*, 2013.

[8] J. Scheerens and R. Bosker, *The foundations of educational effectiveness.* Pergamon, 1997.

[9] J. Hattie, *Visible learning: A synthesis of over 800 meta-analyses relating to achievement.* Routledge, 2009. [Online]. Available: http://books.google.co.uk/books?id=lh7SZNCabGQC

[10] B. K. Hamre and R. C. Pianta, "Classroom environments and developmental processes: Conceptualization and measurement," in *Handbook of research on schools, schooling and human development.* Routledge, 2010, pp. 43–59.

[11] S. D'Mello, E. Dieterle, and A. Duckworth, "Advanced, analytic, automated (aaa) measurement of engagement during learning," *Educational psychologist*, vol. 52, no. 2, pp. 104–123, 2017.

[12] M. T. H. Chi and R. Wylie, "The icap framework: Linking cognitive engagement to active learning outcomes," *Educational Psychologist*, vol. 49, no. 4, pp. 219–243, 2014. [Online]. Available: https://doi.org/10.1080/00461520.2014.965823

[13] A. Helmke and A. Renkl, "Das münchener aufmerksamkeitsinventar (mai): Ein instrument zur systematischen verhaltensbeobachtung der schüleraufmerksamkeit im unterricht," *Diagnostica*, vol. 38, no. 2, pp. 130–141, 1992.

[14] M. Hommel, "Aufmerksamkeitstief in reflexionsphasen–eine videoanalyse von planspielunterricht," *Wirtschaft und Erziehung*, vol. 64, no. 1, pp. 12–18, 2012.

[15] R. E. Carlson and D. Smith-Howell, "Classroom public speaking assessment: Reliability and validity of selected evaluation instruments," *Communication Education*, vol. 44, no. 2, pp. 87–97, 1995. [Online]. Available: https://doi.org/10.1080/03634529509379001

[16] S. Thomson and M. L. Rucker, "The development of a specialized public speaking competency scale: Test of reliability," *Communication Research Reports*, vol. 19, no. 1, pp. 18–28, 2002. [Online]. Available: https://doi.org/10.1080/08824090209384828

[17] S. Morreale, M. Moore, K. Taylor, D. Surges-Tatum, and L. Webster, "Competent speaker speech evaluation form," 2007.

[18] L. M. Schreiber, G. D. Paul, and L. R. Shibley, "The development and test of the public speaking competence rubric," *Communication Education*, vol. 61, no. 3, pp. 205–233, 2012. [Online]. Available: https://doi.org/10.1080/03634523.2012.670709

[19] Anonymous, "Towards a psychometrically sound assessment of students' presentation competence: The development of the tübingen instrument for presentation competence (tip)," 2020, in press.

[20] M. I. Tanveer, E. Lin, and M. E. Hoque, "Rhema: A real-time in-situ intelligent interface to help people with public speaking," in *Proceedings of the 20th International Conference on Intelligent User Interfaces*, ser. IUI '15. New York, NY, USA: ACM, 2015, pp. 286–295. [Online]. Available: http://doi.acm.org/10.1145/2678025.2701386

[21] T. Wörtwein, L. Morency, and S. Scherer, "Automatic assessment and analysis of public speaking anxiety: A virtual audience case study," in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. Xian, China: IEEE, Sep. 2015, pp. 187–193.

[22] M. E. Hoque, M. Courgeon, J.-C. Martin, B. Mutlu, and R. W. Picard, "Mach: My automated conversation coach," in *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, ser. UbiComp '13. New York, NY,

USA: ACM, 2013, pp. 697–706. [Online]. Available: http://doi.acm.org/10.1145/2493432.2493502

[23] M. I. Tanveer, R. Zhao, K. Chen, Z. Tiet, and M. E. Hoque, "Automanner: An automated interface for making public speakers aware of their mannerisms," in *Proceedings of the 21st International Conference on Intelligent User Interfaces*, ser. IUI '16. New York, NY, USA: ACM, 2016, pp. 385–396. [Online]. Available: http://doi.acm.org/10.1145/2856767.2856785

[24] R. Zhao, V. Li, H. Barbosa, G. Ghoshal, and M. E. Hoque, "Semi-automated 8 collaborative online training module for improving communication skills," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 1, no. 2, pp. 32:1–32:20, Jun. 2017. [Online]. Available: http://doi.acm.org/10.1145/3090097

[25] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Springer, 1998, pp. 484–498.

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[27] S. G. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 674–693, 1989.

[28] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, 2005, pp. 886–893 vol. 1.

[29] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal, "Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, June 2009, pp. 1932–1939.

[30] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[31] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 2037–2041, 2006.

[32] X. Tan and B. Triggs, "Enhanced local texture feature sets for face recognition under difficult lighting conditions," *IEEE Transactions on Image Processing*, vol. 19, no. 6, pp. 1635–1650, 2010.

[33] S. ul Hussain and B. Triggs, "Visual recognition using local quantized patterns," in *European conference on computer vision*. Springer, 2012, pp. 716–729.

[34] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

[35] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1157–1182, 2003.

[36] J. Tang, S. Alelyani, and H. Liu, *Feature selection for classification: A review*. CRC Press, Jan. 2014, pp. 37–64.

[37] F. Rosenblatt, *The perceptron, a perceiving and recognizing automaton Project Para*. Cornell Aeronautical Laboratory, 1957.

[38] A. Cauchy, "Méthode générale pour la résolution des systemes d'équations simultanées," *Comp. Rend. Sci. Paris*, vol. 25, no. 1847, pp. 536–538, 1847.

[39] P. Werbos, "Beyond regression:" new tools for prediction and analysis in the behavioral sciences," *Ph. D. dissertation, Harvard University*, 1974.

[40] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[41] K. Fukushima and S. Miyake, "Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition," in *Competition and cooperation in neural nets*. Springer, 1982, pp. 267–285.

[42] Y. LeCun, B. E. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. E. Hubbard, and L. D. Jackel, "Handwritten digit recognition with a back-propagation network," in *Advances in neural information processing systems*, 1990, pp. 396–404.

[43] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[44] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.

[45] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[46] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 2016, pp. 770–778. [Online]. Available: https://doi.org/10.1109/CVPR.2016.90

[47] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[48] P. E. Shrout and J. L. Fleiss, "Intraclass correlations: uses in assessing rater reliability." *Psychological bulletin*, vol. 86, no. 2, p. 420, 1979.

[49] R. Stiefelhagen, "Tracking focus of attention in meetings," in *Proceedings. Fourth IEEE International Conference on Multimodal Interfaces*, 2002, pp. 273–280.

[50] E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation in computer vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 4, p. 607–626, Apr. 2009. [Online]. Available: https://doi.org/10.1109/TPAMI.2008.106

[51] Y. Wu and Q. Ji, "Facial landmark detection: A literature survey," *Int. J. Comput. Vision*, vol. 127, no. 2, p. 115–142, Feb. 2019. [Online]. Available: https://doi.org/10.1007/s11263-018-1097-z

[52] X.-S. Gao, X.-R. Hou, J. Tang, and H.-F. Cheng, "Complete solution classification for the perspective-three-point problem," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 25, pp. 930– 943, 09 2003.

[53] V. Lepetit, F. Moreno-Noguer, and P. Fua, "Epnp: An accurate o(n) solution to the pnp problem," *Int. J. Comput. Vision*, vol. 81, no. 2, p. 155–166, Feb. 2009. [Online]. Available: https://doi.org/10.1007/s11263-008-0152-6

[54] T. Ke and S. I. Roumeliotis, "An efficient algebraic solution to the perspective-three-point problem," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4618–4626.

[55] M. Patacchiola and A. Cangelosi, "Head pose estimation in the wild using convolutional neural networks and adaptive gradient methods," *Pattern Recognition*, vol. 71, pp. 132 – 143, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320317302327

[56] N. Ruiz, E. Chong, and J. M. Rehg, "Fine-grained head pose estimation without keypoints," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018, pp. 2155–215 509.

[57] T. Yang, Y. Chen, Y. Lin, and Y. Chuang, "Fsa-net: Learning fine-grained structure aggregation for head pose estimation from a single image," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pp. 1087–1096. [Online]. Available: http://openaccess.thecvf.com/content_CVPR_2019/html/Yang_FSA-Net_Learning_Fine-Grained_Structure_Aggregation_for_Head_Pose_Estimation_From_CVPR_2019_paper.html

[58] D. W. Hansen and Q. Ji, "In the eye of the beholder: A survey of models for eyes and gaze," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 3, pp. 478–500, March 2010.

[59] K. A. Funes Mora, F. Monay, and J.-M. Odobez, "Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras," in *Proceedings of the ACM Symposium on Eye Tracking Research and Applications*. ACM, Mar. 2014.

[60] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "Appearance-based gaze estimation in the wild," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 4511–4520.

[61] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "It's written all over your face: Full-face appearance-based gaze estimation," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, July 2017, pp. 2299–2308.

[62] S. J. Garbin, Y. Shen, I. Schuetz, R. Cavin, G. Hughes, and S. S. Talathi, "Openeds: Open eye dataset," *CoRR*, vol. abs/1905.03702, 2019. [Online]. Available: http://arxiv.org/abs/1905.03702

[63] J. Kim, M. Stengel, A. Majercik, S. De Mello, D. Dunn, S. Laine, M. McGuire, and D. Luebke, "Nvgaze: An anatomically-informed dataset for low-latency, near-eye gaze estimation," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI '19. New York, NY, USA: ACM, 2019, pp. 550:1–550:12. [Online]. Available: http://doi.acm.org/10.1145/3290605.3300780

[64] W. Fuhl, T. C. Santini, T. Kübler, and E. Kasneci, "Else: Ellipse selection for robust pupil detection in real-world environments," in *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*, 2016, pp. 123–130.

[65] W. Fuhl, M. Tonsen, A. Bulling, and E. Kasneci, "Pupil detection for head-mounted eye tracking in the wild: an evaluation of the state of the art," *Machine Vision and Applications*, vol. 27, no. 8, pp. 1275–1288, 2016.

[66] C. Braunagel, W. Rosenstiel, and E. Kasneci, "Ready for take-over? a new driver assistance system for an automated classification of driver take-over readiness," *IEEE Intelligent Transportation Systems Magazine*, vol. 9, no. 4, pp. 10–22, 2017.

[67] C. Braunagel, E. Kasneci, W. Stolzmann, and W. Rosenstiel, "Driver-activity recognition in the context of conditionally autonomous driving," in *2015 IEEE 18th International Conference on Intelligent Transportation Systems*. IEEE, 2015, pp. 1652–1657.

[68] C. Braunagel, D. Geisler, W. Rosenstiel, and E. Kasneci, "Online recognition of driver-activity based on visual scanpath classification," *IEEE Intelligent Transportation Systems Magazine*, vol. 9, no. 4, pp. 23–36, 2017.

[69] T. Appel, C. Scharinger, P. Gerjets, and E. Kasneci, "Cross-subject workload classification using pupil-related measures," in *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, 2018, pp. 1–8.

[70] T. Appel, N. Sevcenko, F. Wortha, K. Tsarava, K. Moeller, M. Ninaus, E. Kasneci, and P. Gerjets, "Predicting cognitive load in an emergency simulation based on behavioral and physiological measures," in *2019 International Conference on Multimodal Interaction*, 2019, pp. 154–163.

[71] E. Kasneci, T. Kübler, K. Broelemann, and G. Kasneci, "Aggregating physiological and eye tracking signals to predict perception in the absence of ground truth," *Computers in Human Behavior*, vol. 68, pp. 450–455, 2017.

[72] C.-Y. Chen and K. Grauman, "Subjects and their objects: Localizing interactees for a person-centric view of importance," *International Journal of Computer Vision*, vol. 126, no. 2-4, pp. 292–313, 2018.

[73] G. Gkioxari, R. Girshick, P. Dollár, and K. He, "Detecting and recognizing human-object interactions," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8359–8367.

[74] A. Recasens, A. Khosla, C. Vondrick, and A. Torralba, "Where are they looking?" in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2015, pp. 199–207. [Online]. Available: http://papers.nips.cc/paper/5848-where-are-they-looking.pdf

[75] L. Fan, Y. Chen, P. Wei, W. Wang, and S.-C. Zhu, "Inferring shared attention in social scene videos," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[76] C. Darwin, *The Expression of the Emotions in Man and Animals*, 1872, the original was published 1898 by Appleton, New York. Reprinted 1965 by the University of Chicago Press, Chicago and London,.

[77] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion." *Journal of personality and social psychology*, vol. 17, no. 2, p. 124, 1971.

[78] P. Ekman, W. V. Friesen, and P. Ellsworth, *Emotion in the Human Face: Guide-lines for Research and an Integration of Findings: Guidelines for Research and an Integration of Findings*. Pergamon, 1972.

[79] W. V. Friesen, "Cultural differences in facial expressions in a social situation: An experimental test on the concept of display rules." Ph.D. dissertation, ProQuest Information & Learning, 1972.

[80] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak, "Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements," *Psychological science in the public interest*, vol. 20, no. 1, pp. 1–68, 2019.

[81] S. Du and A. M. Martinez, "Compound facial expressions of emotion: from basic research to clinical applications," *Dialogues in clinical neuroscience*, vol. 17, no. 4, pp. 443–455, Dec 2015. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/26869845

[82] P. Ekman and W. V. Friesen, *Facial action coding systems*. Consulting Psychologists Press, 1978.

[83] J. A. Russell, "A circumplex model of affect." *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980. [Online]. Available: https://doi.org/10.1037/h0077714

[84] E. Klieme, F. Lipowsky, K. Rakoczy, and N. Ratzka, *Qualitätsdimensionen und Wirksamkeit von Mathematikunterricht. Theoretische Grundlagen und ausgewählte Ergebnisse des Projekts „Pythagoras"*, 01 2006, pp. 127–146.

[85] A.-K. Praetorius, E. Klieme, B. Herbert, and P. Pinger, "Generic dimensions of teaching quality: the german framework of three basic dimensions," *ZDM*, vol. 50, no. 3, pp. 407–426, Jun 2018. [Online]. Available: https://doi.org/10.1007/s11858-018-0918-4

[86] J. B. Carroll, "A model of school learning." *Teachers college record*, 1963.

[87] D. C. Berliner, "Learning about and learning from expert teachers," *International journal of educational research*, vol. 35, no. 5, pp. 463–482, 2001.

[88] K. S. Cortina, K. F. Miller, R. McKenzie, and A. Epstein, "Where low and high inference data converge: Validation of class assessment of mathematics instruction using mobile eye tracking with expert and novice teachers," *International Journal of Science and Mathematics Education*, vol. 13, no. 2, pp. 389–403, Apr 2015. [Online]. Available: https://doi.org/10.1007/s10763-014-9610-5

[89] J. R. Star and S. K. Strickland, "Learning to observe: using video to improve preservice mathematics teachers' ability to notice," *Journal of Mathematics Teacher Education*, vol. 11, no. 2, pp. 107–125, Apr 2008. [Online]. Available: https://doi.org/10.1007/s10857-007-9063-7

[90] K. Stürmer, T. Seidel, K. Müller, J. Häusler, and K. S. Cortina, "What is in the eye of preservice teachers while instructing? an eye-tracking study about attention processes in different teaching situations," *Zeitschrift für Erziehungswissenschaft*, vol. 20, no. 1, pp. 75–92, Mar 2017. [Online]. Available: https://doi.org/10.1007/s11618-017-0731-9

[91] M. Kleinknecht and A. Gröschner, "Fostering preservice teachers' noticing with structured video feedback: Results of an online- and video-based intervention study," *Teaching and Teacher Education*, vol. 59, pp. 45 – 56, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0742051X16301007

[92] M. G. Sherin and E. A. van Es, "Effects of video club participation on teachers' professional vision," *Journal of Teacher Education*, vol. 60, no. 1, pp. 20–37, 2009. [Online]. Available: https://doi.org/10.1177/0022487108328155

[93] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li, "S3fd: Single shot scale-invariant face detector," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[94] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*. Xi'an, China: IEEE, May 2018, pp. 59–66.

[95] J. Deng, J. Guo, Y. Zhou, J. Yu, I. Kotsia, and S. Zafeiriou, "Retinaface: Single-stage dense face localisation in the wild," *CoRR*, vol. abs/1905.00641, 2019. [Online]. Available: http://arxiv.org/abs/1905.00641

[96] J. Deng, J. Guo, X. Niannan, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *CVPR*, 2019.

[97] T. Baltrušaitis, P. Robinson, and L. Morency, "Openface: An open source facial behavior analysis toolkit," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2016, pp. 1–10.

[98] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face alignment across large poses: A 3d solution," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 146–155.

[99] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2019.

[100] E. Chong, N. Ruiz, Y. Wang, Y. Zhang, A. Rozga, and J. M. Rehg, "Connecting gaze, scene, and attention: Generalized attention estimation via joint modeling of gaze and scene saliency," in *The European Conference on Computer Vision (ECCV)*, September 2018.

[101] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, Nov 1998.

[102] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Advances in Neural Information Processing Systems 19*, B. Schölkopf, J. C. Platt, and T. Hoffman, Eds. MIT Press, 2007, pp. 545–552. [Online]. Available: http://papers.nips.cc/paper/3095-graph-based-visual-saliency.pdf

[103] X. Hou, J. Harel, and C. Koch, "Image signature: Highlighting sparse salient regions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 1, pp. 194–201, Jan 2012.

[104] M. Kümmerer, T. S. A. Wallis, and M. Bethge, "Deepgaze II: reading fixations from deep features trained on object recognition," *CoRR*, vol. abs/1610.01563, 2016. [Online]. Available: http://arxiv.org/abs/1610.01563

[105] C. E. Wolff, H. Jarodzka, N. van den Bogert, and H. P. A. Boshuizen, "Teacher vision: expert and novice teachers' perception of problematic classroom management scenes," *Instructional Science*, vol. 44, no. 3, pp. 243–265, Jun 2016. [Online]. Available: https://doi.org/10.1007/s11251-016-9367-z

[106] C. E. Wolff, H. Jarodzka, and H. P. Boshuizen, "See and tell: Differences between expert and novice teachers' interpretations of problematic classroom management events," *Teaching and Teacher Education*, vol. 66, pp. 295 – 308, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0742051X17306832

[107] N. A. McIntyre and T. Foulsham, "Scanpath analysis of expertise and culture in teacher gaze in real-world classrooms," *Instructional Science*, Jan 2018. [Online]. Available: https://doi.org/10.1007/s11251-017-9445-x

[108] T. Soucek, J. Moravec, and J. Lokoc, "Transnet: A deep network for fast detection of common shot transitions," *CoRR*, vol. abs/1906.03363, 2019. [Online]. Available: http://arxiv.org/abs/1906.03363

[109] A. Zhang, Q. Wang, Z. Zhu, J. Paisley, and C. Wang, "Fully supervised speaker diarization," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Brighton, UK: IEEE, May 2019, pp. 6301–6305.

[110] J. Grafsgaard, J. B. Wiggins, K. E. Boyer, E. N. Wiebe, and J. Lester, "Automatically recognizing facial expression: Predicting engagement and frustration," in *Educational Data Mining 2013*, 2013.

[111] J. Whitehill, Z. Serpell, Y. C. Lin, A. Foster, and J. R. Movellan, "The faces of engagement: Automatic recognition of student engagementfrom facial expressions," *IEEE Transactions on Affective Computing*, vol. 5, no. 1, pp. 86–98, Jan 2014.

[112] N. Bosch, "Detecting student engagement: Human versus machine," in *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*, ser. UMAP '16. New York, NY, USA: ACM, 2016, pp. 317–320. [Online]. Available: http://doi.acm.org/10.1145/2930238.2930371

[113] H. Monkaresi, N. Bosch, R. A. Calvo, and S. K. D'Mello, "Automated detection of engagement using video-based estimation of facial expressions and heart rate," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 15–28, Jan.-March 2017. [Online]. Available: doi.ieeecomputersociety.org/10.1109/TAFFC.2016.2515084

[114] M. Raca, "Camera-based estimation of student's attention in class," Ph.D. dissertation, EPFL, Lausanne, 2015.

[115] P. Goldberg, Ö. Sümer, K. Stürmer, W. Wagner, R. Göllner, P. Gerjets, E. Kasneci, and U. Trautwein, "Attentive or not?: Toward a machine learning approach to assessing students' visible engagement in classroom instruction," *Educational Psychology Review*, 2019. [Online]. Available: https://doi.org/10.1007/s10648-019-09514-z

[116] M. Raca and P. Dillenbourg, "System for assessing classroom attention," *Proceedings of 3rd International Learning Analytics & Knowledge Conference*, 2013. [Online]. Available: http://infoscience.epfl.ch/record/185814

[117] Ö. Sümer, P. Goldberg, P. Gerjets, U. Trautwein, and E. Kasneci, "Multimodal engagement analysis from facial videos in the classroom," 2020, submitted to IEEE Transactions on Affective Computing.

[118] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[119] Ö. Sümer, P. Gerjets, U. Trautwein, and E. Kasneci, "Attention flow: End-to-end joint attention estimation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, March 2020.

[120] Ö. Sümer, P. Goldberg, K. Sturmer, T. Seidel, P. Gerjets, U. Trautwein, and E. Kasneci, "Teachers' perception in the classroom," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[121] Ö. Sümer, P. Gerjets, U. Trautwein, and E. Kasneci, "Automated anonymisation of visual and audio data in classroom studies," in *The Workshops of the Thirty-Forth AAAI Conference on Artificial Intelligence*, February 2020.

[122] A. Lachner, H. Jarodzka, and M. Nückles, "What makes an expert teacher? investigating teachers' professional vision and discourse abilities," *Instructional Science*, vol. 44, no. 3, pp. 197–203, Jun 2016. [Online]. Available: https://doi.org/10.1007/s11251-016-9376-y

[123] F. Erickson, "Video research in the learning sciences," in *Video research in the learning sciences*, R. Goldman, R. Pea, B. Barron, and S. J. Derry, Eds. Routledge, 2007, ch. Ways of seeing video: toward a phenomenology of viewing minimally edited footage, pp. 145–155.

[124] U. Trautwein, P. Gerjets, and E. Kasneci, "Cognitive interface for educational improvement: Assessing students' attentional focus in the classroom," University of Tübingen, Tech. Rep., 2017.

[125] J. Driver, "A selective review of selective attention research from the past century," *British Journal of Psychology*, vol. 92, no. 1, pp. 53–78, 2001. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1348/000712601162103

[126] R. Brünken and T. Seufert, "Video research in the learning sciences," in *Handbuch Lernstrategien*, H. Mandl and H. F. Friedrich, Eds. Hogrefe Verlag, 2007, ch. Aufmerksamkeit, Lernen, Lernstrategien, pp. 27–37.

[127] R. W. Engle, "Working memory capacity as executive attention," *Current Directions in Psychological Science*, vol. 11, no. 1, pp. 19–23, 2002. [Online]. Available: https://doi.org/10.1111/1467-8721.00160

[128] A. Miyake, N. P. Friedman, M. J. Emerson, A. H. Witzki, A. Howerter, and T. D. Wager, "The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis," *Cognitive Psychology*, vol. 41, no. 1, pp. 49 – 100, 2000. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S001002859990734X

[129] M. Posner, "Structres and functions of selective attention. master lectures in clinical neuropsychology," *Boll T, Bryant B*, 1988.

[130] P. R. Pintrich and E. V. de Groot, "Motivational and self-regulated learning components of classroom academic performance." *Journal of Educational Psychology*, vol. 82, no. 1, pp. 33–40, 1990. [Online]. Available: https://doi.org/10.1037/0022-0663.82.1.33

[131] J. P. Connell, "The self in transition: Infancy to childhood," in *The self in transition: Infancy to childhood*, D. Cicchetti and M. Beeghly, Eds. The University of Chicago Press, 1990, ch. Context, self, and action: a motivational analysis of self-system processes across the life span, pp. 61–97.

[132] R. C. Pianta and B. K. Hamre, "Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity," *Educational Researcher*, vol. 38, no. 2, pp. 109–119, 2009. [Online]. Available: https://doi.org/10.3102/0013189X09332374

[133] G. Büttner and L. Schmidt-Atzert, *Diagnostik von Konzentration und Aufmerksamkeit*. Göttingen: Hogrefe Verlag, 2004.

[134] T. Yamamoto and K. Imai-Matsumura, "Teachers' gaze and awareness of students' behavior: Using an eye tracker," *Comprehensive Psychology*, vol. 2, p. 01.IT.2.6, 2013. [Online]. Available: https://doi.org/10.2466/01.IT.2.6

[135] P. Gerjets, C. Walter, W. Rosenstiel, M. Bogdan, and T. O. Zander, "Cognitive state monitoring and the design of adaptive instruction in digital environments: lessons learned from cognitive workload assessment using a passive brain-computer interface approach," *Frontiers in Neuroscience*, vol. 8, p. 385, 2014. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnins.2014.00385

[136] T. Krumpe, C. Scharinger, W. Rosenstiel, P. Gerjets, and M. Spüler, "Unity and diversity in working memory load: Evidence for the separability of the executive functions updating and inhibition using machine learning," *Biological Psychology*, vol. 139, pp. 163 – 172, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0301051118303028

[137] M. Poh, N. C. Swenson, and R. W. Picard, "A wearable sensor for unobtrusive, long-term assessment of electrodermal activity," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 5, pp. 1243–1252, 2010.

[138] R. Yoshida, T. Ogitsu, H. Takemura, H. Mizoguchi, E. Yamaguchi, S. Inagaki, Y. Takeda, M. Namatame, M. Sugimoto, and F. Kusunoki, "Feasibility study on estimating visual attention using electrodermal activity," 09 2014.

[139] H. M. Lahaderne, "Attitudinal and intellectual correlates of attention: A study of four sixth-grade classrooms." *Journal of educational psychology*, vol. 59, no. 5, p. 320, 1968.

[140] J. D. McKinney, J. Mason, K. Perkerson, and M. Clifford, "Relationship between classroom behavior and academic achievement." *Journal of Educational Psychology*, vol. 67, no. 2, pp. 198–203, 1975. [Online]. Available: https://doi.org/10.1037/h0077012

[141] R. E. Mayer, "Should there be a three-strikes rule against pure discovery learning?" *American psychologist*, vol. 59, no. 1, p. 14, 2004.

[142] N. Karweit and R. E. Slavin, "Measurement and modeling choices in studies of time and learning," *American Educational Research Journal*, vol. 18, no. 2, pp. 157–171, 1981. [Online]. Available: https://doi.org/10.3102/00028312018002157

[143] D. Stipek, "Development of achievement motivation," A. Wigfield and J. S. Eccles, Eds. San Diego: Academic Press, 2002, ch. Good instruction is motivating, pp. 309–332.

[144] H. Lei, Y. Cui, and W. Zhou, "Relationships between student engagement and academic achievement: A meta-analysis," *Social Behavior and Personality: an international journal*, vol. 46, no. 3, pp. 517–528, 2018.

[145] C. Pauli and F. Lipowsky, "Mitmachen oder zuhören? mündliche schülerinnen-und schülerbeteiligung im mathematikunterricht," *Unterrichtswissenschaft*, vol. 35, no. 2, pp. 101–124, 2007.

[146] K. J. Ehrhardt, P. Findeisen, G. Marinello, and H. Reinartz-Wenzel, "Systematische verhaltensbeobachtung von aufmerksamkeit im unterricht: Zur prüfung von objecktivität und suverlässigkeit." *Diagnostica*, 1981.

[147] R. S. Baker, A. T. Corbett, K. R. Koedinger, and A. Z. Wagner, "Off-task behavior in the cognitive tutor classroom: When students "game the system"," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '04. New York, NY, USA: Association for Computing Machinery, 2004, p. 383–390. [Online]. Available: https://doi.org/10.1145/985692.985741

[148] J. M. Girard, "Carma: Software for continuous affect rating and media annotation," *Journal of open research software*, vol. 2, no. 1, p. e5, 2014. [Online]. Available: https://doi.org/10.5334/jors.ar

[149] I. Lizdek, P. Sadler, E. Woody, N. Ethier, and G. Malet, "Capturing the stream of behavior: A computer-joystick method for coding interpersonal behavior continuously over time," *Social Science Computer Review*, vol. 30, no. 4, pp. 513–521, 2012. [Online]. Available: https://doi.org/10.1177/0894439312436487

[150] N. Bosch, S. K. D'Mello, R. S. Baker, J. Ocumpaugh, V. Shute, M. Ventura, L. Wang, and W. Zhao, "Detecting student emotions in computer-enabled classrooms," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, ser. IJCAI'16. AAAI Press, 2016, p. 4125–4129.

[151] N. Bosch, S. K. D'mello, J. Ocumpaugh, R. S. Baker, and V. Shute, "Using video to automatically detect learner affect in computer-enabled classrooms," *ACM Trans. Interact. Intell. Syst.*, vol. 6, no. 2, Jul. 2016. [Online]. Available: https://doi.org/10.1145/2946837

[152] M. Raca, R. Tormey, and P. Dillenbourg, "Student motion and it's potential as a classroom performance metric," in *3rd International Workshop on Teaching Analytics (IWTA)*, no. POST_TALK, 2013.

[153] M. Raca, P. Dillenbourg, and R. Tormey, "Sleepers' lag - study on motion and attention," *Proceedings of the 4th Internation Conference on Learning Analytics and Knowledge*, 2014. [Online]. Available: http://infoscience.epfl.ch/record/196641

[154] M. Raca, L. Kidzinski, and P. Dillenbourg, "Translating head motion into attention - towards processing of student's body-language," *Proceedings of the 8th International Conference on Educational Data Mining*, 2015. [Online]. Available: http://infoscience.epfl.ch/record/207803

[155] J. Zaletelj and A. Kosir, "Predicting students' attention in the classroom from kinect facial and body features," *EURASIP Journal on Image and Video Processing*, vol. 2017, pp. 1–12, 2017.

[156] J. Zaletelj, "Estimation of students' attention in the classroom from kinect features," in *Proceedings of the 10th International Symposium on Image and Signal Processing and Analysis*, Sept 2017, pp. 220–224.

[157] K. Fujii, P. Marian, D. Clark, Y. Okamoto, and J. Rekimoto, "Sync class: Visualization system for in-class student synchronization," in *Proceedings of the 9th Augmented Human International Conference*, ser. AH '18. New York, NY, USA: Association for Computing Machinery, 2018. [Online]. Available: https://doi.org/10.1145/3174910.3174927

[158] S. L. Christenson, A. L. Reschly, and C. Wylie, *Handbook of research on student engagement*. Springer Science & Business Media, 2012.

[159] J. M. Girard and J. F. Cohn, "A primer on observational measurement," *Assessment*, vol. 23, no. 4, pp. 404–413, 2016, pMID: 26933139. [Online]. Available: https://doi.org/10.1177/1073191116635807

[160] J. Bidwell and F. H., "Classroom analytics: measuring stu- dent engagement with automated gaze tracking," Tech. Rep., 2011.

[161] C. Thomas and D. B. Jayagopi, "Predicting student engagement in classrooms using facial behavioral cues," in *Proceedings of the 1st ACM SIGCHI International Workshop on Multimodal Interaction for Education*, ser. MIE 2017. New York, NY, USA: Association for Computing Machinery, 2017, p. 33–40. [Online]. Available: https://doi.org/10.1145/3139513.3139514

[162] K. Ahuja, D. Kim, F. Xhakaj, V. Varga, A. Xie, S. Zhang, J. E. Townsend, C. Harrison, A. Ogan, and Y. Agarwal, "Edusense: Practical classroom sensing at scale," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 3, no. 3, pp. 71:1–71:26, Sep. 2019. [Online]. Available: http://doi.acm.org/10.1145/3351229

[163] B. Ngoc Anh, N. Tung Son, P. Truong Lam, L. Phuong Chi, N. Huu Tuan, N. Cong Dat, N. Huu Trung, M. Umar Aftab, and T. Van Dinh, "A computer-vision based application for student behavior monitoring in classroom," *Applied Sciences*, vol. 9, no. 22, 2019. [Online]. Available: https://www.mdpi.com/2076-3417/9/22/4729

[164] S. K. D'Mello, S. D. Craig, and A. C. Graesser, "Multimethod assessment of affective experience and expression during deep learning," *Int. J. Learn. Technol.*, vol. 4, no. 3/4, p. 165–187, Oct. 2009. [Online]. Available: https://doi.org/10.1504/IJLT.2009.028805

[165] J. F. Grafsgaard, R. M. Fulton, K. E. Boyer, E. N. Wiebe, and J. C. Lester, "Multimodal analysis of the implicit affective channel in computer-mediated textual communication," in *Proceedings of the 14th ACM International Conference on Multimodal Interaction*, ser. ICMI '12. New York, NY, USA: Association for Computing Machinery, 2012, p. 145–152. [Online]. Available: https://doi.org/10.1145/2388676.2388708

[166] H. L. O'Brien and E. G. Toms, "The development and evaluation of a survey to measure user engagement," *Journal of the American Society for Information Science and Technology*, vol. 61, no. 1, pp. 50–69, 2010. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.21229

[167] S. G. Hart and L. E. Staveland, "Development of nasa-tlx (task load index): Results of empirical and theoretical research," *Human mental workload*, vol. 1, no. 3, pp. 139–183, 1988.

[168] J. Ocumpaugh, R. Baker, M. A. Mercedes, and R. T., "Baker rodrigo ocumpaugh monitoring protocol (bromp) 2.0 technical and training manual," Columbia University, Tech. Rep., 2015.

[169] A. Kamath, A. Biswas, and V. Balasubramanian, "A crowdsourced approach to student engagement recognition in e-learning environments," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2016, pp. 1–9.

[170] A. Kaur, A. Mustafa, L. Mehta, and A. Dhall, "Prediction and localization of student engagement in the wild," in *2018 Digital Image Computing: Techniques and Applications (DICTA)*, Dec 2018, pp. 1–8.

[171] D. Sanchez-Cortes, O. Aran, and D. Gatica-Perez, "An audiovisual corpus for emergent leader analysis," in *Proceedings of the 10th International Symposium on Image and Signal Processing and Analysis*, 2011, p. 1–6.

[172] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, April 2013, pp. 1–8.

[173] O. Celiktutan, E. Skordos, and H. Gunes, "Multimodal human-human-robot interactions (mhhri) dataset for studying personality and engagement," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 484–497, Oct 2019.

[174] O. Rudovic, J. Lee, M. Dai, B. Schuller, and R. W. Picard, "Personalized machine learning for robot perception of affect and engagement in autism therapy," *Science Robotics*, vol. 3, no. 19, 2018. [Online]. Available: https://robotics.sciencemag.org/content/3/19/eaao6760

[175] H. W. Park, I. Grover, S. Spaulding, L. Gomez, and C. Breazeal, "A model-free affective reinforcement learning approach to personalization of an autonomous social robot companion for early literacy education," in *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*. AAAI Press, 2019, pp. 687–694. [Online]. Available: https://doi.org/10.1609/aaai.v33i01.3301687

[176] S. D'Mello, E. Dieterle, and A. Duckworth, "Advanced, analytic, automated (aaa) measurement of engagement during learning," *Educational psychologist*, vol. 52, no. 2, pp. 104–123, 2017. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/29038607

[177] A. Helmke, "Das münchner aufmerksamkeitsinventar (mai). manual für die beobachtung des aufmerksamkeitsverhaltens von grundschülern während des unterrichtes," Max-Planck-Institut für psychologische Forschung, Tech. Rep. 6, 1988.

[178] J. Smallwood and J. W. Schooler, "The restless mind." *Psychological bulletin*, vol. 132, no. 6, p. 946, 2006.

[179] D. Smilek, J. S. Carriere, and J. A. Cheyne, "Out of mind, out of sight: Eye blinking as indicator and embodiment of mind wandering," *Psychological Science*, vol. 21, no. 6, pp. 786–789, 2010. [Online]. Available: https://doi.org/10.1177/0956797610368063

[180] N. Blanchard, R. Bixler, T. Joyce, and S. D'Mello, "Automated physiological-based detection of mind wandering during learning," in *Intelligent Tutoring Systems*, S. Trausan-Matu, K. E. Boyer, M. Crosby, and K. Panourgia, Eds. Cham: Springer International Publishing, 2014, pp. 55–60.

[181] C. Mills and S. K. D'Mello, "Toward a real-time (day) dreamcatcher: Detecting mind wandering episodes during online reading," in *Proceedings of the 8th International Conference on Educational Data Mining, EDM 2015, Madrid, Spain, June 26-29, 2015*, O. C. Santos, J. Boticario, C. Romero, M. Pechenizkiy, A. Merceron, P. Mitros, J. M. Luna, M. C. Mihaescu, P. Moreno, A. Hershkovitz, S. Ventura, and M. C. Desmarais, Eds. International Educational Data Mining Society (IEDMS), 2015, pp. 69–76. [Online]. Available: http://www.educationaldatamining.org/EDM2015/proceedings/full69-76.pdf

[182] A. Stewart, N. Bosch, H. Chen, P. Donnelly, and S. D'Mello, "Face forward: Detecting mind wandering from video during narrative film comprehension," in *Artificial Intelligence in Education*, E. André, R. Baker, X. Hu, M. M. T. Rodrigo, and B. du Boulay, Eds. Cham: Springer International Publishing, 2017, pp. 359–370.

[183] A. Stewart, N. Bosch, and S. K. D'Mello, "Generalizability of face-based mind wandering detection across task contexts." *International Educational Data Mining Society*, 2017.

[184] N. Bosch and S. D'Mello, "Automatic detection of mind wandering from video in the lab and in the classroom," *IEEE Transactions on Affective Computing*, pp. 1–1, 2019.

[185] O. Rudovic, M. Zhang, B. Schuller, and R. Picard, "Multi-modal active learning from human data: A deep reinforcement learning approach," in *2019 International Conference on Multimodal Interaction*, ser. ICMI '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 6–15. [Online]. Available: https://doi.org/10.1145/3340555.3353742

[186] O. Rudovic, H. W. Park, J. Busche, B. Schuller, C. Breazeal, and R. W. Picard, "Personalized estimation of engagement from videos using active learning with deep reinforcement learning," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019, pp. 217–226.

[187] C. J. Soto and O. P. John, "Short and extra-short forms of the big five inventory–2: The bfi-2-s and bfi-2-xs," *Journal of Research in Personality*, vol. 68, pp. 69 – 81, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0092656616301325

[188] B. Frank, *Presence messen in laborbasierter Forschung mit Mikrowelten Entwicklung und erste Validierung eines Fragebogens zur Messung von Presence.* Springer-Verlag, 2015.

[189] S. E. Rimm-Kaufman, A. E. Baroody, R. A. A. Larsen, T. W. Curby, and T. Abry, "To what extent do teacher-student interaction quality and student gender contribute to fifth graders' engagement in mathematics learning?" *Journal of Educational Psychology*, vol. 107, no. 1, pp. 170–185, 2015.

[190] M. Knogler, J. M. Harackiewicz, A. Gegenfurtner, and D. Lewalter, "How situational is situational interest?: Investigating the longitudinal structure of situational interest." *Contemporary Educational Psychology*, vol. 43, pp. 39–50, 2015.

[191] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997. [Online]. Available: https://doi.org/10.1162/neco.1997.9.8.1735

[192] A. Graves, *Supervised Sequence Labelling with Recurrent Neural Networks*, ser. Studies in Computational Intelligence. Berlin: Springer, 2012. [Online]. Available: https://cds.cern.ch/record/1503877

[193] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 3104–3112. [Online]. Available: http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf

[194] J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 677–691, 2017.

[195] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1412.6980

[196] R. E. Riggio and H. S. Friedman, "Impression formation: The role of expressive behavior." *Journal of Personality and Social Psychology*, vol. 50, no. 2, pp. 421–427, 1986.

[197] M. L. Knapp, J. A. Hall, and T. G. Horgan, *Nonverbal communication in human interaction*. Wadsworth: Cengage Learning, 2013.

[198] P. Rao S. B, S. Rasipuram, R. Das, and D. B. Jayagopi, "Automatic assessment of communication skill in non-conventional interview settings: A comparative study," in *ACM ICMI*, 2017.

[199] T. Pfister and P. Robinson, "Real-time recognition of affective states from nonverbal features of speech and its application for public speaking skill analysis," *IEEE Transactions on Affective Computing*, vol. 2, no. 2, pp. 66–78, April 2011.

[200] G. Luzardo, B. Guamán, K. Chiluiza, J. Castells, and X. Ochoa, "Estimation of presentations skills based on slides and audio features," in *Proceedings of the 2014 ACM Workshop on Multimodal Learning Analytics Workshop and Grand Challenge*, ser. MLA '14. New York, NY, USA: ACM, 2014, pp. 37–44. [Online]. Available: http://doi.acm.org/10.1145/2666633.2666639

[201] R. Sharma, T. Guha, and G. Sharma, "Multichannel attention network for analyzing visual behavior in public speaking," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. Lake Tahoe, NV: IEEE, March 2018, pp. 476–484.

[202] L. Chen, C. W. Leong, G. Feng, and C. M. Lee, "Using multimodal cues to analyze mla'14 oral presentation quality corpus: Presentation delivery and slides quality," in *Proceedings of the 2014 ACM Workshop on Multimodal Learning Analytics Workshop and Grand Challenge*, ser. MLA '14. New York, NY, USA: Association for Computing Machinery, 2014, p. 45–52. [Online]. Available: https://doi.org/10.1145/2666633.2666640

[203] T. Wörtwein, M. Chollet, B. Schauerte, L.-P. Morency, R. Stiefelhagen, and S. Scherer, "Multimodal public speaking performance assessment," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ser. ICMI '15. New York, NY, USA: ACM, 2015, pp. 43–50. [Online]. Available: http://doi.acm.org/10.1145/2818346.2820762

[204] F. Haider, L. Cerrato, N. Campbell, and S. Luz, "Presentation quality assessment using acoustic information and hand movements," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016, Shanghai, China, March 20-25, 2016*. Shanghai, China: IEEE, 2016, pp. 2812–2816. [Online]. Available: https://doi.org/10.1109/ICASSP.2016.7472190

[205] L. Chen, C. W. Leong, G. Feng, C. M. Lee, and S. Somasundaran, "Utilizing multimodal cues to automatically evaluate public speaking performance," in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. Xian, China: IEEE, Sep. 2015, pp. 394–400.

[206] V. Ramanarayanan, C. W. Leong, L. Chen, G. Feng, and D. Suendermann-Oeft, "Evaluating speech, face, emotion and body movement time-series features for automated multimodal presentation scoring," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ser. ICMI '15. New York, NY, USA: ACM, 2015, pp. 23–30. [Online]. Available: http://doi.acm.org/10.1145/2818346.2820765

[207] E. Herbein, J. Golle, M. Tibus, I. Zettler, and U. Trautwein, "Putting a speech training program into practice: Its implementation and effects on elementary school children's public speaking skills and levels of speech anxiety," *Contemporary Educational Psychology*, vol. 55, pp. 176 – 188, 2018.

[208] K. Curtis, G. J. Jones, and N. Campbell, "Effects of good speaking techniques on audience engagement," in *ACM ICMI*, ser. ICMI '15, 2015, p. 35–42.

[209] S. Scherer, G. Layher, J. Kane, H. Neumann, and N. Campbell, "An audiovisual political speech analysis incorporating eye-tracking and perception data," in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA), May 2012, pp. 1114–1120. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2012/pdf/1011_Paper.pdf

[210] A. Cullen, A. Hines, and N. Harte, "Perception and prediction of speaker appeal – a single speaker study," *Computer Speech & Language*, vol. 52, pp. 23–40, 2018.

[211] G. Hu and Y. Liu, "Three minute thesis presentations as an academic genre: A cross-disciplinary study of genre moves," *Journal of English for Academic Purposes*, vol. 35, pp. 16–30, 2018.

[212] E. Rowley-Jolivet and S. Carter-Thomas, "Scholarly soundbites," *Science Communication on the Internet: Old genres meet new genres*, vol. 308, p. 81, 2019.

[213] A. Rosenberg and J. Hirschberg, "Acoustic/prosodic and lexical correlates of charismatic speech," in *INTERSPEECH 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, September 4-8, 2005*. Lisbon, Portugal: ISCA, 2005, pp. 513–516. [Online]. Available: http://www.isca-speech.org/archive/interspeech_2005/i05_0513.html

[214] E. Strangert and J. Gustafson, "What makes a good speaker? : Subject ratings, acoustic measurements and perceptual evaluations," in *Proc. Annu. Conf. Int. Speech. Commun. Assoc., INTERSPEECH :*, ser. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH. Brisbane, Australia: ISCA, 2008, pp. 1688–1691, qC 20141016.

[215] X. Ochoa, M. Worsley, K. Chiluiza, and S. Luz, "Mla'14: Third multimodal learning analytics workshop and grand challenges," in *Proceedings of the 16th International Conference on Multimodal Interaction*, ser. ICMI '14. New York, NY, USA: ACM, 2014, pp. 531–532. [Online]. Available: http://doi.acm.org/10.1145/2663204.2668318

[216] K. O. McGraw and S. P. Wong, "Forming inferences about some intraclass correlation coefficients," *Psychological Methods*, vol. 1, no. 1, pp. 30–46, 1996.

[217] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, 2010, p. 1459–1462.

[218] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multi-task cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, Oct 2016.

[219] A. Zadeh, Y. C. Lim, T. Baltrušaitis, and L. Morency, "Convolutional experts constrained local model for 3d facial landmark detection," in *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, Oct 2017, pp. 2519–2528.

[220] Z. Cao, G. Hidalgo, T. Simon, S. Wei, and Y. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," 2018. [Online]. Available: http://arxiv.org/abs/1812.08008

[221] Q. Tariq, J. Daniels, J. N. Schwartz, P. Washington, H. Kalantarian, and D. P. Wall, "Mobile detection of autism through machine learning on home video: A development and prospective validation study," *PLoS medicine*, vol. 15, no. 11, p. e1002705, 2018.

[222] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, vol. 29, pp. 1189–1232, 2000.

[223] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, "Classification and regression trees. belmont, ca: Wadsworth," *International Group*, vol. 432, pp. 151–166, 1984.

[224] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, p. 5–32, Oct. 2001. [Online]. Available: https://doi.org/10.1023/A:1010933404324

[225] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[226] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[227] A. Kent, "Synchronization as a classroom dynamic: A practitioner's perspective," *Mind, Brain, and Education*, vol. 7, no. 1, pp. 13–18, 2013. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/mbe.12002

[228] L. Prieto, K. Sharma, L. Kidzinski, M. Rodriguez-Triana, and P. Dillenbourg, "Multimodal teaching analytics: Automated extraction of orchestration graphs from wearable sensor data," *Journal of Computer Assisted Learning*, vol. 34, no. 2, pp. 193–203, 2018. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/jcal.12232

[229] K. Watanabe, "Teaching as a dynamic phenomenon with interpersonal interactions," *Mind, Brain, and Education*, vol. 7, no. 2, pp. 91–100, 2013. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/mbe.12011

[230] L. B. Olswang, P. Dowden, J. Feuerstein, K. Greenslade, G. L. Pinder, and K. Fleming, "Triadic gaze intervention for young children with physical disabilities," *Journal of Speech, Language, and Hearing Research*, vol. 57, no. 5, pp. 1740–1753, 2014. [Online]. Available: +http://dx.doi.org/10.1044/2014_JSLHR-L-13-0058

[231] S. Andrist, B. Mutlu, and A. Tapus, "Look like me: Matching robot personality via gaze to increase motivation," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, ser. CHI '15. New York, NY, USA: ACM, 2015, pp. 3603–3612. [Online]. Available: http://doi.acm.org/10.1145/2702123.2702592

[232] R. M. Aronson, T. Santini, T. C. Kübler, E. Kasneci, S. Srinivasa, and H. Admoni, "Eye-hand behavior in human-robot shared manipulation," in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 4–13. [Online]. Available: https://doi.org/10.1145/3171221.3171287

[233] H. S. Park, E. Jain, and Y. Sheikh, "3d social saliency from head-mounted cameras," in *Advances in Neural Information Processing Systems 25*, F. Pereira,

C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 422–430. [Online]. Available: http://papers.nips.cc/paper/4619-3d-social-saliency-from-head-mounted-cameras.pdf

[234] H. S. Park, E. Jain, and Y. Sheikh, "Predicting primary gaze behavior using social saliency fields," in *2013 IEEE International Conference on Computer Vision*, Dec 2013, pp. 3503–3510.

[235] H. S. Park and J. Shi, "Social saliency prediction," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 4777–4785.

[236] R. R. da Silva and R. A. F. Romero, "Modelling shared attention through relational reinforcement learning," *Journal of Intelligent & Robotic Systems*, vol. 66, no. 1, pp. 167–182, Apr 2012. [Online]. Available: https://doi.org/10.1007/s10846-011-9624-y

[237] Y. Nagai, M. Asada, and K. Hosoda, "Learning for joint attention helped by functional development," *Advanced Robotics*, vol. 20, no. 10, pp. 1165–1181, 2006.

[238] Y. Nagai, K. Hosoda, A. Morita, and M. Asada, "A constructive model for the development of joint attention," *Connection Science*, vol. 15, no. 4, pp. 211–229, 2003. [Online]. Available: https://doi.org/10.1080/09540090310001655101

[239] T. Santini, T. Kübler, L. Draghetti, P. Gerjets, W. Wagner, U. Trautwein, and E. Kasneci, "Automatic mapping of remote crowd gaze to stimuli in the classroom," in *Eye Tracking Enhanced Learning (ETEL2017)*, 09 2017.

[240] D. P. Papadopoulos, A. D. F. Clarke, F. Keller, and V. Ferrari, "Training object class detectors from eye tracking data," in *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, 2014, pp. 361–376. [Online]. Available: https://doi.org/10.1007/978-3-319-10602-1_24

[241] N. Karessli, Z. Akata, B. Schiele, and A. Bulling, "Gaze embeddings for zero-shot image classification," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 2017, pp. 6412–6421. [Online]. Available: https://doi.org/10.1109/CVPR.2017.679

[242] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "Predicting human eye fixations via an lstm-based saliency attentive model," *IEEE Trans. Image Processing*, vol. 27, no. 10, pp. 5142–5154, 2018. [Online]. Available: https://doi.org/10.1109/TIP.2018.2851672

[243] Y. Yu, J. Choi, Y. Kim, K. Yoo, S.-H. Lee, and G. Kim, "Supervising neural attention models for video captioning by human gaze data," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[244] A. Recasens, C. Vondrick, A. Khosla, and A. Torralba, "Following gaze in video," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017, pp. 1444–1452.

[245] S. Gorji and J. J. Clark, "Attentional push: A deep convolutional network for augmenting image salience with shared attention modeling in social scenes," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 3472–3481.

[246] T. Striano, V. M. Reid, and S. Hoehl, "Neural mechanisms of joint attention in infancy," *European Journal of Neuroscience*, vol. 23, no. 10, pp. 2819–2823, 2006. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1460-9568.2006.04822.x

[247] D. A. Baldwin, "Early referential understanding: Infants' ability to recognize referential acts for what they are," *Developmental Psychology*, vol. 29, pp. 832–843, 09 1993.

[248] D. S. Murray, N. A. Creaghead, P. Manning-Courtney, P. K. Shear, J. Bean, and J.-A. Prendeville, "The relationship between joint attention and language in children with autism spectrum disorders," *Focus on Autism and Other Developmental Disabilities*, vol. 23, no. 1, pp. 5–14, 2008. [Online]. Available: https://doi.org/10.1177/1088357607311443

[249] G. Shteynberg, "Shared attention," *Perspectives on Psychological Science*, vol. 10, no. 5, pp. 579–590, 2015, pMID: 26385997. [Online]. Available: https://doi.org/10.1177/1745691615589104

[250] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 185–207, Jan 2013.

[251] Z. Bylinskii, A. Recasens, A. Borji, A. Oliva, A. Torralba, and F. Durand, "Where should saliency models look next?" in *Computer Vision – ECCV 2016*. Cham: Springer International Publishing, 2016, pp. 809–824.

[252] S. Fan, Z. Shen, M. Jiang, B. L. Koenig, J. Xu, M. S. Kankanhalli, and Q. Zhao, "Emotional attention: A study of image sentiment and visual attention," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[253] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv e-prints*, vol. abs/1409.0473, Sep. 2014. [Online]. Available: https://arxiv.org/abs/1409.0473

[254] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proceedings of the 32nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, F. Bach and D. Blei, Eds., vol. 37. Lille, France: PMLR, 07–09 Jul 2015, pp. 2048–2057.

[255] S. Sharma, R. Kiros, and R. Salakhutdinov, "Action recognition using visual attention," *CoRR*, vol. abs/1511.04119, 2015. [Online]. Available: http://arxiv.org/abs/1511.04119

[256] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," *CoRR*, vol. abs/1704.06904, 2017. [Online]. Available: http://arxiv.org/abs/1704.06904

[257] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[258] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "Cbam: Convolutional block attention module," in *The European Conference on Computer Vision (ECCV)*, September 2018.

[259] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," in *Computer Vision and Pattern Recognition (CVPR)*, 2017.

[260] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 391–405.

[261] C. G. Keller and D. M. Gavrila, "Will the pedestrian cross? a study on pedestrian path prediction," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 2, pp. 494–506, April 2014.

[262] G. Bertasius, H. S. Park, S. X. Yu, and J. Shi, "Am i a baller? basketball performance assessment from first-person videos," in *The IEEE International Conference on Computer Vision (ICCV)*, October 2017.

[263] P. Felsen, P. Agrawal, and J. Malik, "What will happen next? forecasting player moves in sports videos," in *2017 IEEE International Conference on Computer Vision (ICCV)*, vol. 00, Oct. 2018, pp. 3362–3371. [Online]. Available: doi.ieeecomputersociety.org/10.1109/ICCV.2017.362

[264] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. T. Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "AVEC 2016: Depression, Mood, and Emotion Recognition Workshop and Challenge," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. Amsterdam, The Netherlands: ACM Press, Oct. 2016, pp. 3–10. [Online]. Available: http://dl.acm.org/citation.cfm?id=2988258

[265] J. Ventura, S. Cruz, and T. E. Boult, "Improving teaching and learning through video summaries of student engagement," in *Workshop on Computational Models for Learning Systems and Educational Assessment (CMLA 2016)*, IEEE. Las Vegas, NV: IEEE, 06/2016 2016.

[266] C. Einarsson and K. Granström, "Gender-biased interaction in the classroom: The influence of gender and age in the relationship between teacher and pupil," *Scandinavian Journal of Educational Research*, vol. 46, no. 2, pp. 117–127, 2002. [Online]. Available: https://doi.org/10.1080/00313830220142155

[267] T. S. Dee, "A teacher like me: Does race, ethnicity, or gender matter?" *The American Economic Review*, vol. 95, no. 2, pp. 158–165, 2005. [Online]. Available: http://www.jstor.org/stable/4132809

[268] T. Seidel, K. "Stürmer, S. Schäfer, and G. Jahn, "How preservice teachers perform in teaching events regarding generic teaching and learning components," *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, vol. 47, no. 2, pp. 84–96, 2015. [Online]. Available: https://doi.org/10.1026/0049-8637/a000125

[269] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognit Psychol*, vol. 12, no. 1, pp. 97–136, Jan. 1980. [Online]. Available: http://www.ncbi.nlm.nih.gov/sites/entrez?cmd=Retrieve&db=PubMed&list_uids=7351125&dopt=Citation

[270] C. Koch and S. Ullman, *Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry*. Dordrecht: Springer Netherlands, 1987, pp. 115–141. [Online]. Available: https://doi.org/10.1007/978-94-009-3833-5_5

[271] D. Parkhurst, K. Law, and E. Niebur, "Modeling the role of salience in the allocation of overt visual attention," *Vision Research*, vol. 42, no. 1, pp. 107–123, 2002.

[272] H. Sattar, A. Bulling, and M. Fritz, "Predicting the category and attributes of visual search targets using deep gaze pooling," in *Proc. of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2017, pp. 2740–2748. [Online]. Available: https://perceptual.mpi-inf.mpg.de/files/2017/08/sattar17_iccvw.pdf

[273] S. Karthikeyan, T. Ngo, M. P. Eckstein, and B. S. Manjunath, "Eye tracking assisted extraction of attentionally important objects from videos," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, 2015, pp. 3241–3250. [Online]. Available: https://doi.org/10.1109/CVPR.2015.7298944

[274] C. S. Stefan Mathe, "Actions in the eye: Dynamic gaze datasets and learnt saliency models for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, 2015.

[275] H. Sattar, M. Fritz, and A. Bulling, "Visual decoding of targets during visual search from human eye fixations," arXiv:1706.05993, 2017. [Online]. Available: https://arxiv.org/abs/1706.05993https://perceptual.mpi-inf.mpg.de/files/2017/12/sattar17_arxiv.pdf

[276] D. Jayaraman and K. Grauman, "Learning to Look Around: Intelligently Exploring Unseen Environments for Unknown Tasks," *ArXiv e-prints*, Sep. 2017.

[277] D. Jayaraman and K. Grauman, "Learning image representations tied to egomotion from unlabeled video," *Int. J. Comput. Vision*, vol. 125, no. 1-3, pp. 136–161, Dec. 2017. [Online]. Available: https://doi.org/10.1007/s11263-017-1001-2

[278] A. Fathi, J. K. Hodgins, and J. M. Rehg, "Social interactions: A first-person perspective," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, June 2012, pp. 1226–1233.

[279] A. Fathi, Y. Li, and J. M. Rehg, "Learning to recognize daily actions using gaze," in *Computer Vision – ECCV 2012*, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 314–327.

[280] K. Yamada, Y. Sugano, T. Okabe, Y. Sato, A. Sugimoto, and K. Hiraki, "Attention prediction in egocentric video using motion and visual saliency," in *Advances in Image and Video Technology*, Y.-S. Ho, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 277–288.

[281] J. Steil, P. Müller, Y. Sugano, and A. Bulling, "Forecasting user attention during everyday mobile interactions using device-integrated and wearable sensors," Tech. Rep., 2018. [Online]. Available: https://arxiv.org/abs/1801.06011https://perceptual.mpi-inf.mpg.de/files/2018/01/steil2018_arxiv2.pdf

[282] R. G. Cinbis, J. Verbeek, and C. Schmid, "Unsupervised metric learning for face identification in tv video," in *2011 International Conference on Computer Vision*, Nov 2011, pp. 1559–1566.

[283] B. Wu, Y. Zhang, B. G. Hu, and Q. Ji, "Constrained clustering and its application to face clustering in videos," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, June 2013, pp. 3507–3514.

[284] S. Xiao, M. Tan, and D. Xu, "Weighted block-sparse low rank representation for face clustering in videos," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 123–138.

[285] S. Jin, H. Su, C. Stauffer, and E. Learned-Miller, "End-to-end face detection and cast grouping in movies using erdos-renyi clustering," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[286] A. Nagrani and A. Zisserman, "From benedict cumberbatch to sherlock holmes: Character identification in TV series without a script," *CoRR*, vol. abs/1801.10442, 2018.

[287] C. Huang, B. Wu, and R. Nevatia, "Robust object tracking by hierarchical association of detection responses," in *Computer Vision – ECCV 2008*, D. Forsyth, P. Torr, and A. Zisserman, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 788–801.

[288] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," *CoRR*, vol. abs/1709.01507, 2017.

[289] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[290] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 6738–6746.

[291] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2015.

[292] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," *CoRR*, vol. abs/1710.08092, 2017.

[293] A. Santella and D. DeCarlo, "Robust clustering of eye movement recordings for quantification of visual interest," in *Proceedings of the 2004 Symposium on Eye Tracking Research & Applications*, ser. ETRA '04. New York, NY, USA: ACM, 2004, pp. 27–34. [Online]. Available: http://doi.acm.org/10.1145/968363.968368

[294] L. van der Maaten and G. E. Hinton, "Visualizing high-dimensional data using t-sne," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[295] G. Levi and T. Hassner, "Age and gender classification using convolutional neural networks," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) workshops*, June 2015. [Online]. Available: https://www.openu.ac.il/home/hassner/projects/cnn_agegender

[296] D. Sun, S. Roth, and M. J. Black, "Secrets of optical flow estimation and their principles," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun. 2010, pp. 2432–2439.

[297] T. Hoel and W. Chen, "Privacy and data protection in learning analytics should be motivated by an educational maxim—towards a proposal," *Research and Practice in Technology Enhanced Learning*, vol. 13, no. 1, p. 20, Dec 2018. [Online]. Available: https://doi.org/10.1186/s41039-018-0086-8

[298] "Supervision pursuant to the general data protection regulation (eu) 2016/679 – facial recognition used to monitor the attendance of students," https://www.datainspektionen.se/globalassets/dokument/beslut/facial-recognition-used-to-monitor-the-attendance-of-students.pdf, Swedish Data Protection Authority, 08 2019.

[299] Z. Ji, Z. C. Lipton, and C. Elkan, "Differential privacy and machine learning: a survey and review," *CoRR*, vol. abs/1412.7584, 2014. [Online]. Available: http://arxiv.org/abs/1412.7584

[300] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations*, 2015. [Online]. Available: http://arxiv.org/abs/1412.6572

[301] Z. Ren, Y. Jae Lee, and M. S. Ryoo, "Learning to anonymize faces for privacy preserving action detection," in *The European Conference on Computer Vision (ECCV)*, September 2018.

[302] H. Hukkelås, R. Mester, and F. Lindseth, "Deepprivacy: A generative adversarial network for face anonymization," in *Advances in Visual Computing*, G. Bebis, R. Boyle, B. Parvin, D. Koracin, D. Ushizima, S. Chai, S. Sueda, X. Lin, A. Lu, D. Thalmann, C. Wang, and P. Xu, Eds. Cham: Springer International Publishing, 2019, pp. 565–578.

[303] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *Proc. ICASSP*, 2014.

[304] D. Garcia-Romero, D. Snyder, G. Sell, D. Povey, and A. McCree, "Speaker diarization using deep neural network embeddings," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 4930–4934.

[305] F. Fang, X. Wang, J. Yamagishi, I. Echizen, M. Todisco, N. Evans, and J.-F. Bonastre, "Speaker anonymization using X-vector and neural waveform models," in *SSW 2019, 10th ISCA Speech Synthesis Workshop, 20-22 September 2019, Vienna, Austria*, Vienna, AUSTRIA, 09 2019. [Online]. Available: http://www.eurecom.fr/publication/5909

[306] S. Yang, P. Luo, C. C. Loy, and X. Tang, "Wider face: A face detection benchmark," in *CVPR*, 2016.

[307] H. W. Kuhn, "The Hungarian Method for the Assignment Problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1–2, pp. 83–97, March 1955.

[308] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, jun 2015, pp. 815–823. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR.2015.7298682

[309] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *International Conference on Automatic Face and Gesture Recognition*, 2018.

[310] M. R. Francis, Oct 2019. [Online]. Available: https://sinews.siam.org/About-the-Author/matthew-r-francis

[311] A. K. Jain, S. C. Dass, and K. Nandakumar, "Soft biometric traits for personal recognition systems," in *Biometric Authentication*, D. Zhang and A. K. Jain, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 731–738.

[312] A. Dutta and A. Zisserman, "The VIA annotation software for images, audio and video," in *Proceedings of the 27th ACM International Conference on Multimedia*, ser. MM '19. New York, NY, USA: ACM, 2019. [Online]. Available: https://doi.org/10.1145/3343031.3350535

[313] H. Sloetjes and P. Wittenburg, "Annotation by category: ELAN and ISO DCR," in *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco: European Language Resources Association (ELRA), May 2008. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2008/pdf/208_paper.pdf