Teachers' Perception in the Classroom

Ömer Sümer[1] Patricia Goldberg[1] Kathleen Stürmer[1]
Tina Seidel[3] Peter Gerjets[2] Ulrich Trautwein[1] Enkelejda Kasneci[1]
[1] University of Tübingen, Germany
[2] Leibniz-Institut für Wissensmedien, Germany
[3] Technical University of Munich, Germany

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.

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 crowdsourcing 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 [22], sports [1, 12], or affective computing [46]. 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 [2, 34, 48, 55]. Furthermore, the engagement of students in videos has been studied in educational settings [49, 3, 28].

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 [9, 8] in large scale data. Additionally, we propose teachers' movement and view of eye by use of flow information and number of detected faces.

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. [7]. 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. [51] 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 [50] 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 [27] 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 [42], 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) [38] 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 [45, 23, 16], the first model which measures the saliency model against human fixations in free-viewing paradigm was done by Parkhurst and Neibur [33]. 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 [30, 36], object segmentation [21], action recognition [40], zero-shot image classification [20], or image generation [37].

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 [18, 17], 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.

Fathi et al. [10] 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 [11] used a probabilistic generative model to estimate gaze points and recognize daily activities without eye tracking. Yamada et al. [54] 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. [41] proposed a framework to forecast attentional shift in wearable cameras. However, they exploited several computer vi-

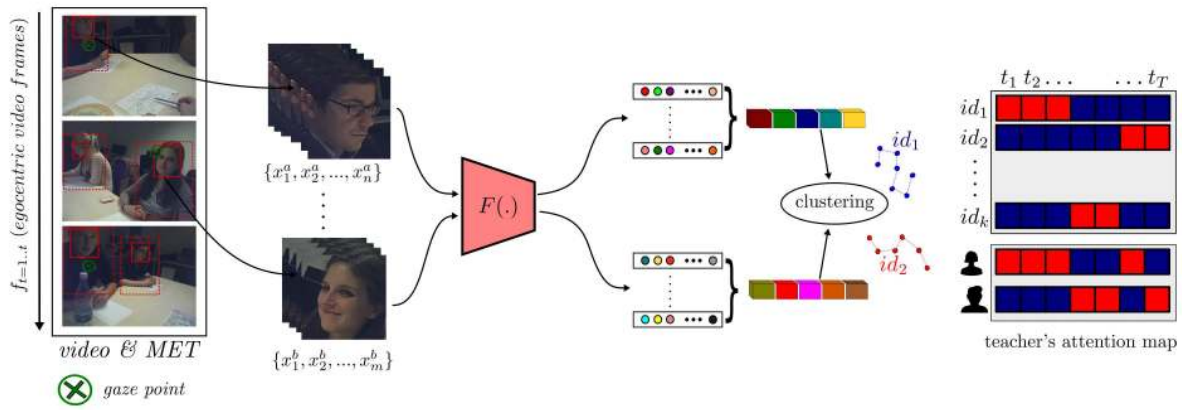

Figure 1: Teacher's attention mapping workflow. Teachers view and gaze points are recorded by a MET while instructing. In egocentric video sequences, face detection is applied, face tracklets in video are created. Then, features are extracted and aggregated by averaging along the feature dimensions. 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.

sion 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 [6, 52, 53]. 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. [19] 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 [29] 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.

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 1 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.

3.1. 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 [56] in all frames and detect faces $(x_t^i)_{t=1}^T$, where i is varying number of detected faces. Then, following [15], 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:

$$P_{(i,j)} = A_{loc.}(x_i, x_j) A_{size}(x_i, x_j) A_{app.}(x_i, x_j) \quad (1)$$

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.

3.2. Face Representation for Tracklets

Convolutional Neural Networks [24, 39, 13, 14] have become very efficient feature representation for general CV tasks and also performed well in large-scale face verification and identification tasks [44, 26]. We use and compare these methods as a face descriptor. Patricularly VGG Deep Faces [32], SphereFace [26] and VGGFace2 [4] 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 [4].

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.

3.3. 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 more discriminative. For this reason, we applied clustering on only low-level tracklets detected as described in Section 3.1.

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 [5] 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 [35] 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 2: Examples of egocentric vision in M-Teach.

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. [42] 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 [38].

The recorded videos are in the resolution of 1280×960 and they contain fast camera motion due to first person view. Figure 2 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.

4.1. 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.

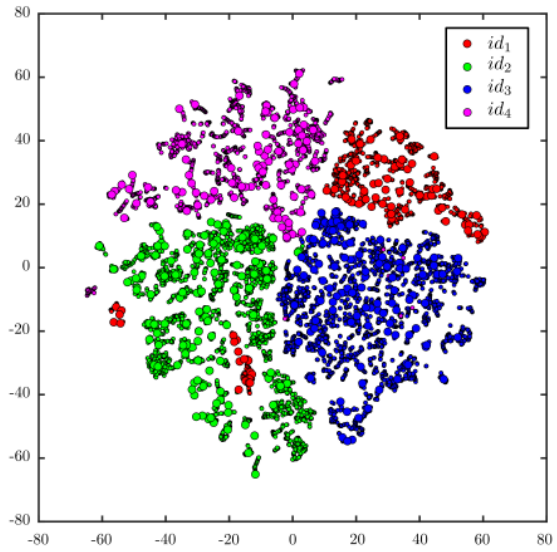

Figure 3: t-SNE distribution of face tracklets using ResNet50/VGG2 features.

A commonly used face representation, the VGG-Face [31] network is trained on VGG-Face dataset which contains 2.6 million images. He et al. [13] proposed "deep residual networks" and it performed the state-of-the-art on the ImageNet object recognition. Recently, Cao et al. [4] 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 4: Sample face tracklets which are created low-level tracklet linking.

Figure 3 shows t-SNE [47] 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.

Table 1: 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**

Figure 4 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.

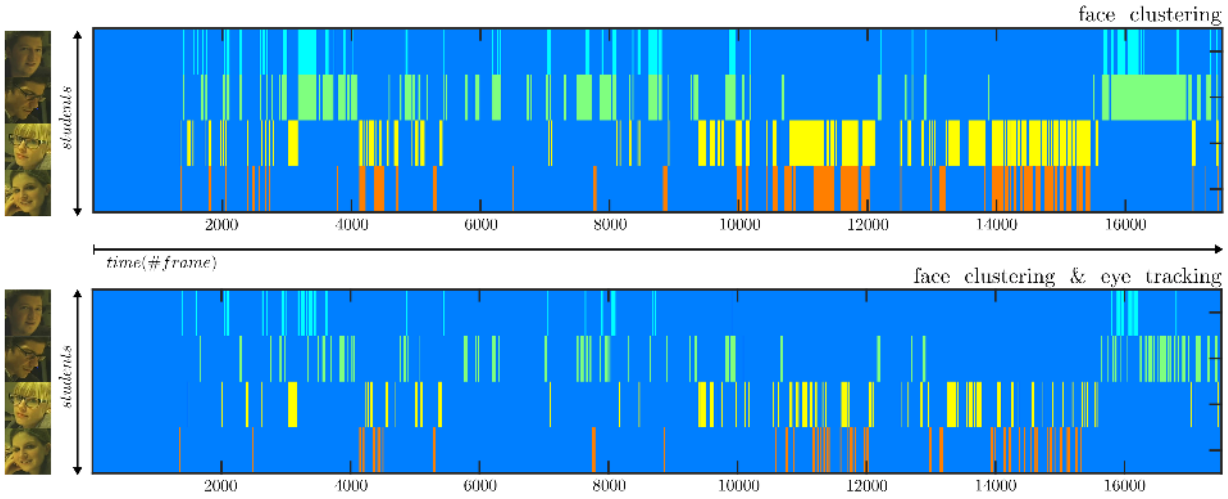

Figure 5: Attention maps. The results of face clustering during a 15-minute M-teach situation *(above)*, fixation points are assigned to the nearest identity *(below)*.

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

4.2. 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 [35]. In our analysis, only fixation events are used.

Figure 5 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 5 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 6 shows that fixation frequencies vary from 40-60% to 10%. These results are consistent with Sürmer et al.'s results [42] which were based on manually defined AOI's.

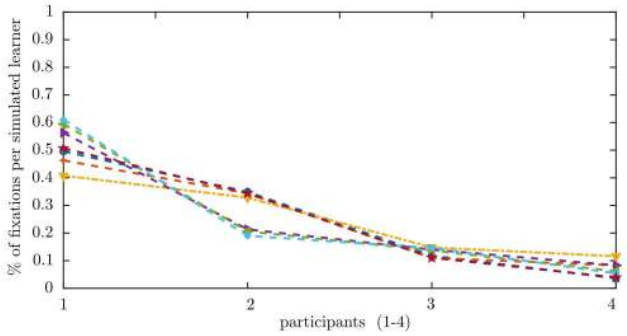

Figure 6: Ranked scores for total fixation frequencies per student in 7 M-Teach situations (in descenting order).

4.3. 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

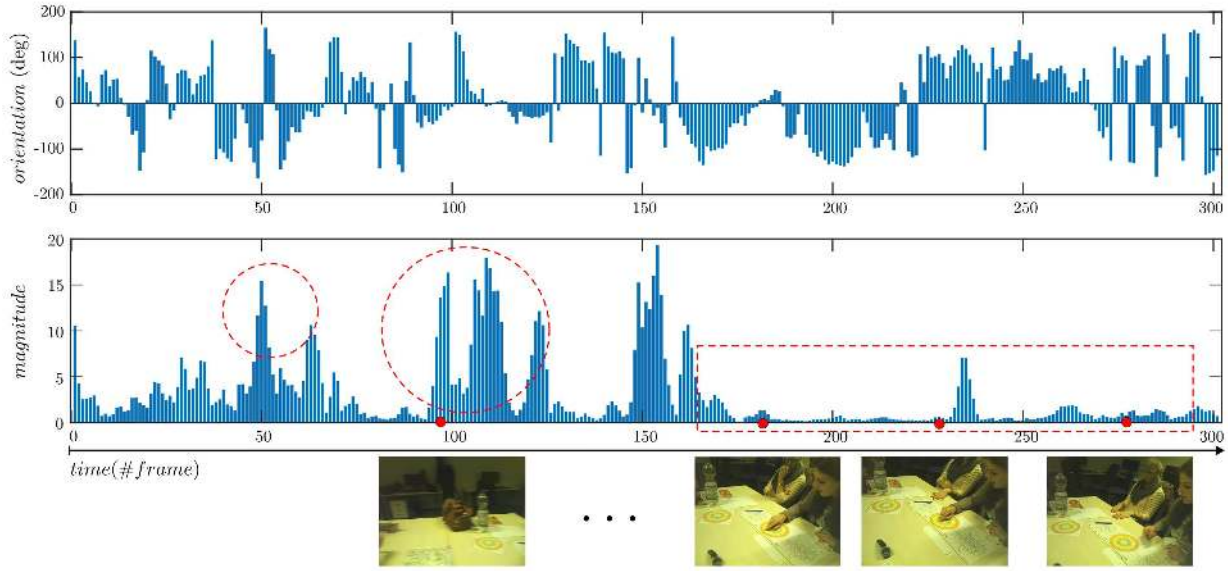

Figure 7: In a short video snippet, mean magnitude and orientation of otpical flow are shown. Large optical flow displacement indicates that teacher's attentional focus changes. In contrast, long stable areas are indicator of an interaction with a student.

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 [25] trained an AlexNet [24] on an unconstrained Adience benchmark to estimate age and gender from face images.

Using face clusters acquired as described in 4.1, we estimated gender of all face images using [25] 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 2: 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**

Table 2 provides the ground truth number of detected faces of four students, the number 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.

4.4. 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 [43]. 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 7 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 elimi-

nate falsely classified gaze points. In this way, we can concentrate long and more consistent time intervals in attention analysis.

5. 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".

References

[1] 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.

[2] J. Bidwell and F. H. Classroom analytics: measuring student engagement with automated gaze tracking. Technical report, 2011.

[3] N. Bosch. Detecting student engagement: Human versus machine. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*, UMAP '16, pages 317–320, New York, NY, USA, 2016. ACM.

[4] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. *CoRR*, abs/1710.08092, 2017.

[5] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[6] R. G. Cinbis, J. Verbeek, and C. Schmid. Unsupervised metric learning for face identification in tv video. In *2011 International Conference on Computer Vision*, pages 1559–1566, Nov 2011.

[7] 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*, 13(2):389–403, Apr 2015.

[8] T. S. Dee. A teacher like me: Does race, ethnicity, or gender matter? *The American Economic Review*, 95(2):158–165, 2005.

[9] C. Einarsson and K. Granstrm. Gender-biased interaction in the classroom: The influence of gender and age in the relationship between teacher and pupil. *Scandinavian Journal of Educational Research*, 46(2):117–127, 2002.

[10] 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*, pages 1226–1233, June 2012.

[11] A. Fathi, Y. Li, and J. M. Rehg. Learning to recognize daily actions using gaze. In A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, editors, *Computer Vision – ECCV 2012*, pages 314–327, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[12] 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)*, volume 00, pages 3362–3371, Oct. 2018.

[13] 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)*, pages 770–778, June 2016.

[14] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. *CoRR*, abs/1709.01507, 2017.

[15] C. Huang, B. Wu, and R. Nevatia. Robust object tracking by hierarchical association of detection responses. In D. Forsyth, P. Torr, and A. Zisserman, editors, *Computer Vision – ECCV 2008*, pages 788–801, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg.

[16] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(11):1254–1259, Nov. 1998.

[17] D. Jayaraman and K. Grauman. Learning image representations tied to egomotion from unlabeled video. *Int. J. Comput. Vision*, 125(1-3):136–161, Dec. 2017.

[18] D. Jayaraman and K. Grauman. Learning to Look Around: Intelligently Exploring Unseen Environments for Unknown Tasks. *ArXiv e-prints*, Sept. 2017.

[19] 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.

[20] 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*, pages 6412–6421, 2017.

[21] 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*, pages 3241–3250, 2015.

[22] C. G. Keller and D. M. Gavrila. Will the pedestrian cross? a study on pedestrian path prediction. *IEEE Transactions on Intelligent Transportation Systems*, 15(2):494–506, April 2014.

[23] C. Koch and S. Ullman. *Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry*, pages 115–141. Springer Netherlands, Dordrecht, 1987.

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[25] 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.

[26] 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)*, pages 6738–6746, July 2017.

[27] N. A. McIntyre and T. Foulsham. Scanpath analysis of expertise and culture in teacher gaze in real-world classrooms. *Instructional Science*, Jan 2018.

[28] 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*, 8(1):15–28, Jan.-March 2017.

[29] A. Nagrani and A. Zisserman. From benedict cumberbatch to sherlock holmes: Character identification in TV series without a script. *CoRR*, abs/1801.10442, 2018.

[30] 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*, pages 361–376, 2014.

[31] H. S. Park and J. Shi. Social saliency prediction. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4777–4785, June 2015.

[32] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2015.

[33] D. Parkhurst, K. Law, and E. Niebur. Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, 42(1):107–123, 2002.

[34] M. Raca. Camera-based estimation of student's attention in class. page 180, 2015.

[35] 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*, ETRA '04, pages 27–34, New York, NY, USA, 2004. ACM.

[36] 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)*, pages 2740–2748, 2017.

[37] H. Sattar, M. Fritz, and A. Bulling. Visual decoding of targets during visual search from human eye fixations. arxiv:1706.05993, 2017.

[38] T. Seidel, K. "Stürmer, S. Schäfer, and G. Jahn. How preservice teachers perform in teaching events regarding generic teaching and learning components. *Zeitschrift fr Entwicklungspsychologie und Pdagogische Psychologie*, 47(2):84–96, 2015.

[39] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[40] 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*, 37, 2015.

[41] J. Steil, P. Mller, Y. Sugano, and A. Bulling. Forecasting user attention during everyday mobile interactions using device-integrated and wearable sensors. Technical report, 2018.

[42] 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*, 20(1):75–92, Mar 2017.

[43] 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)*, pages 2432–2439. IEEE, June 2010.

[44] 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.

[45] A. M. Treisman and G. Gelade. A feature-integration theory of attention. *Cognit Psychol*, 12(1):97–136, Jan. 1980.

[46] 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*, pages 3–10, Amsterdam, The Netherlands, Oct. 2016. ACM Press.

[47] L. van der Maaten and G. E. Hinton. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[48] 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)*, Las Vegas, NV, 06/2016 2016. IEEE, IEEE.

[49] 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*, 5(1):86–98, Jan 2014.

[50] 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*, 66:295 – 308, 2017.

[51] 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*, 44(3):243–265, Jun 2016.

[52] 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*, pages 3507–3514, June 2013.

[53] S. Xiao, M. Tan, and D. Xu. Weighted block-sparse low rank representation for face clustering in videos. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 123–138, Cham, 2014. Springer International Publishing.

[54] K. Yamada, Y. Sugano, T. Okabe, Y. Sato, A. Sugimoto, and K. Hiraki. Attention prediction in egocentric video using motion and visual saliency. In Y.-S. Ho, editor, *Advances in Image and Video Technology*, pages 277–288, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[55] 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*, pages 220–224, Sept 2017.

[56] 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.