IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 6 February 2024, accepted 8 April 2024, date of publication 12 April 2024, date of current version 19 April 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3387932

RESEARCH ARTICLE

Long-Term Human Participation Assessment in Collaborative Learning Environments Using Dynamic Scene Analysis

WENJING SHI[1], (Member, IEEE), PHUONG TRAN[1], SYLVIA CELEDÓN-PATTICHIS[2], AND MARIOS S. PATTICHIS[1], (Senior Member, IEEE)

[1]Image and Video Processing and Communications Laboratory, Department of Electrical and Computer Engineering, The University of New Mexico, Albuquerque, NM 87131, USA

[2]Department of Curriculum and Instruction, The University of Texas at Austin, Austin, TX 78712, USA

Corresponding author: Marios S. Pattichis (pattichi@unm.edu)

This work was supported in part by the National Science Foundation under Grant 1949230, Grant 1842220, and Grant 1613637.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Office of the Institutional Review Board at the University of New Mexico (UNM OIRB) under IRB Protocol No. 2250030376R001.

ABSTRACT The paper develops datasets and methods to assess student participation in real-life collaborative learning environments. In collaborative learning environments, students are organized into small groups where they are free to interact within their group. Thus, students can move around freely causing issues with strong pose variation, move out and re-enter the camera scene, or face away from the camera. We formulate the problem of assessing student participation into two subproblems: (i) student group detection against strong background interference from other groups, and (ii) dynamic participant tracking within the group. A massive independent testing dataset of 12,518,250 student label instances, of total duration of 21 hours and 22 minutes of real-life videos, is used for evaluating the performance of our proposed method for student group detection. The proposed method of using multiple image representations is shown to perform equally or better than YOLO on all video instances. Over the entire dataset, the proposed method achieved an F1 score of 0.85 compared to 0.80 for YOLO. Following student group detection, the paper presents the development of a dynamic participant tracking system for assessing student group participation through long video sessions. The proposed dynamic participant tracking system is shown to perform exceptionally well, missing a student in just one out of 35 testing videos. In comparison, a state-of-the-art method fails to track students in 14 out of the 35 testing videos. The proposed method achieves 82.3% accuracy on an independent set of long, real-life collaborative videos.

INDEX TERMS Human participation assessment, dynamic participant tracking, occlusion detection.

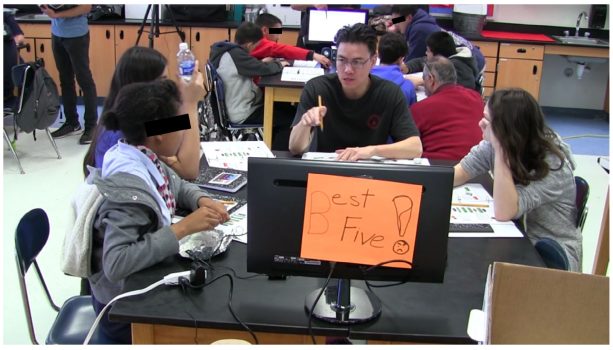
I. INTRODUCTION

Classroom video analysis requires the development of robust image processing methods that can work in very challenging environments. In this paper, we study methods for student group detection, student recognition, and assessing student participation under challenging occlusions and student movement. We demonstrate our methods on classroom videos that were collected by the Advancing Out-of-school Learning in Mathematics and Engineering (AOLME) project.

AOLME videos were recorded in actual student classrooms as demonstrated in Fig. 1. The classroom is organized into several groups of students, where multiple student groups can appear in a single video (see Fig. 1(b)). We use a single video camera for each group. Thus, our first task is to develop methods for student group detection, by detecting the students that are closest to the camera. As it is clear from Fig. 1, students need to be detected from multiple angles. Furthermore, there are significant issues with both partial and full occlusions. Students can be active participants while they remain partially or fully occluded. Thus, in order to properly assess student participation, we need to develop effective

The associate editor coordinating the review of this manuscript and approving it for publication was Ramakrishnan Srinivasan.

(a) Example with total occlusion.



(b) Example with partial occlusion, complicated background, and multiple activities.

FIGURE 1. Examples of the challenges associated with developing methods for assessing student participation based on the AOLME datasets.

methods to deal with occlusions. In addition, we also need to deal with significant student movements in and out of the frame (see Fig. 1(a)). Furthermore, AOLME video sessions are very long, ranging from 45 to 90 minutes each. As a result, we need to keep track of student participation throughout the long video sessions. We provide an extensive comparison of the AOLME video dataset against other datasets in section II.

The unique challenges associated with processing the AOLME dataset require that we consider the development of new approaches. Furthermore, due to the need to process large video datasets, we require the development of fast methods. Specifically, we need to integrate person detection, face recognition, and tracking under occlusion. These methods have to be integrated into a video analysis system that supports the long durations of the AOLME video sessions. We use the term Dynamic Participant Tracking (DPT) to refer to our approach. DPT processes the time history of group detections from individual video frames to determine a state for each student participant (e.g., occluded, inside frame, outside frame, inside and outside, and unknown). Furthermore, DPT processes a sequence of frames to determine transitions from state to state.

We provide a summary of related methods that we have adopted for our DPT. Due to its speed, we adopted the use of YOLO for person tracking. For face recognition, we adopt the use of the InsightFace system [1] that has been tested on a large number of camera-facing image datasets and a variety of loss function models. We will provide more details on related background methods in Section II.

Here, we provide a brief summary of methods that have been recently developed to track objects under occlusion. We note the use of a correlation filter in [2], a classifier approach in [3], and convolutional neural networks in [4]. More recently, a geometric approach has been developed in [5] and [6]. In [5], the authors proposed a novel algorithm that addresses occlusion by using only the location and size of detection bounding boxes. The algorithm, termed Simple Online and Real-time Tracking with Occlusion Handling (SORT_OH [5]) can predict occlusions and re-identify lost targets. This paper uses both MOT16/17 datasets for pedestrian tracking and achieved state-of-the-art results for online tracking algorithms. We will provide comparisons of our proposed approach against Simple Online and Real-time Tracking with Occlusion Handling (SORT_OH [5]) to demonstrate that we can achieve significantly better performance on the AOLME dataset.

We claim four primary contributions. First, we develop a system for student group detection using multiple image representations. As we document in our results, the use of multiple representations results in much better person detection. Second, we develop a system for video face recognition for identifying the students within the group. Our video face recognition enables face recognition from different angles. Third, we develop new methods for dynamic scene analysis system using DPT. We demonstrate that the DPT provides much better results than SORT_OH. Fourth, we introduce the use of student participation maps for visualizing the results over long video sessions.

We note that we presented preliminary results on group detection in conference publications: [7], [8], [9],, and video face recognition in [11]. While we review these earlier methods for completeness, we note that the current paper describes training and testing on the complete system over much larger datasets. Furthermore, the dynamic participant tracking methodology that is a primary focus of the current paper has never appeared in any previous publications. The paper also uses participant maps that were initially developed in [12] for tracking student activities associated with hand movements (see [13], [14]). Here, we note that the current paper does not involve any student activities that include hand movements. Overall, the complete system, including the dynamic scene analysis, has not been previously discussed in the literature.

We organize the rest of the paper into five additional sections. In Section II, we provide a detailed description of the AOLME dataset and elaborate on its challenges as we compare against other datasets. In Section III, we provide detailed background information. We then describe our proposed methods in Section IV. The results are given in Section V. We provide concluding remarks in Section VI.

TABLE 1. AOLME dataset uniqueness against common video datasets. AOLME contains real-life recordings of actual classrooms with significant challenges.

Features	MOT16/17 [15]	OTB-2015 [16]	VOT2018 [17]	LaSOT [18]	TAO [19]	AOLME
Various camera angles	✓	✓	✓	✓	✓	✓
Multiple objects and humans	✓	✓	✓	✓	✓	✓
Diverse scales of activities	✓	✓	✓	✓	✓	✓
Complicated background	✓	✓	✓	✗	✗	✓
Complete occlusion	✓	✓	✓	✗	✗	✓
Multiple activities	✓	✗	✓	✗	✗	✓
Humans are at the edge of the frame	✗	✗	✗	✓	✓	✓
Specific group detection	✗	✗	✗	✗	✗	✓
Long-term occlusion	✗	✗	✗	✗	✗	✓
Tracking specific objects throughout the video	✗	✗	✗	✗	✗	✓
Video length	$<$ 1 min 25 secs	$<$ 2 min 9 secs	$<$ 1 min 23 secs	$\approx$ 83 secs	$\approx$ 30 secs	23 min 45 secs segments

TABLE 2. AOLME student datasets.

Problem		Method	Training/Validation Dataset		Testing Dataset		Final Testing Dataset
			Dataset Source	Dataset	Dataset Source	Dataset	
Group detection	Face detection	YOLO	AOLME-G	AOLME-GY1 (2,200 images)	AOLME-G	-	AOLME-GT 13 videos 12,518,250 labels 21 hours 22 min
		Group face classifier		AOLME-GF1 112,129 images		AOLME-GF2 28,032 images	
	Back-of-the-head detection	Group back-of-the-head classifier		AOLME-GB1 45,568 images		AOLME-GB2 11,392 images	
Face recognition		Extended InsightFace	AOLME-FR	AOLME-FR1 3,968 images	-	-	AOLME-DLT 2h 21m 24s videos
Dynamic Participant Tracking		Dynamic scene analysis	-	-	AOLME-D	AOLME-DST 17m 17s videos	AOLME-DLT 2h 21m 24s videos

II. AOLME STUDENT DATASETS

We provide a comparison of the unique characteristics of the AOLME dataset as compared against related video datasets in Table 1. We begin with a summary of common datasets and then provide a summary of the characteristics that are unique to AOLME.

We begin with a summary of common datasets. Large-scale Single Object Tracking (LaSOT [18]) is used for single-object tracking with an average video length of approximately 83 seconds. The Tracking Any Object (TAO [19]) has 2907 videos and 833 classes, where each video only includes a single activity lasting around 30 seconds. Both LaSOT and TAO datasets are characterized by simple backgrounds and partial, short-term occlusions. In contrast, the AOLME video dataset is characterized by complex backgrounds with both partial and full longer-term occlusions.

The Visual Tracker Benchmark 2015 (OTB-2015 [16]) contains 100 video clips with various activities and different objects. Example objects include humans and SUVs. The entire OTB-2015 dataset contains 58,613 frames, and each video only has one type of activity. In contrast, the AOLME dataset is significantly larger with far more complex activities.

There are 60 sequences in the Visual Object Tracking 2018 (VOT2018 [17]) datasets at a frame rate of about 30 fps. The total duration for the dataset is only 745.2 sec. Furthermore, unlike AOLME, as for OTB-2015 and VOT2018, the dataset does not contain multiple, overlapping activities.

For human tracking, the most commonly used datasets include Multiple Object Tracking 16/17 (MOT16/17 [15]). The datasets cover short-term and full occlusions. However, unlike AOLME, each video lasts less than 90 seconds.

In summary, common datasets share videos captured from multiple video angles that can include multiple objects and humans at diverse scales. In contrast, AOLME is characterized by the need to develop methods for specific group detection, long-term occlusions, and the need to track specific objects over very long video segments. AOLME video sessions range from 45 to 90 minutes broken into shorter segments of 23 minutes and 45 seconds. Overall, the AOLME dataset contains over 950 hours of video, collected over three different cohorts, with each cohort including

TABLE 3. AOLME-DST: Short-video dataset for entire system testing of dynamic participant tracking. The test dataset includes 35 videos and 35 different students.

Group	Video	Duration	Occlusion Frames	Occlusion Time
G1 4 persons	V1	20s	434	14.5s
	V2	20s	104	3.5s
	V3	25s	188	6.3s
	V4	45s	221	7.4s
	V5	1m45s	805	26.8s
	V6	10s	23	0.7s
	V7	50s	189	6.3s
	V8	20s	5	0.2s
	V9	30s	164	5.5s
	V10	15s	222	7.4s
	V11	10s	56	1.9s
	V12	20s	339	11.3s
	V13	20s	315	10.5s
	V14	5s	29	1s
	V15	10s	28	1s
G2 6 persons	V1	10s	15	0.5s
	V2	26s	291	9.7s
	V3	32s	724	24.1s
	V4	10s	22	0.7s
	V5	20s	61	2s
	V6	15s	152	5.1s
	V7	35s	300	10s
G3 5 persons	V1	50s	90	3s
	V2	10s	119	4s
	V3	5s	36	1.2s
	V4	10s	97	3.2s
	V5	35s	219	7.3s
	V6	1m1s	975	32.5s
G4 5 persons	V1	1m15s	1605	53.5s
G5 6 persons	V1	20s	142	4.7s
	V2	15s	13	0.4s
	V3	2m26s	906	30.2s
	V4	32s	1	0.03s
	V5	10s	120	4s
	V6	15s	219	7.3s

TABLE 4. AOLME-DLT: Long-video dataset for entire system testing of dynamic participant tracking. The dataset contains videos from 22 students.

Video	No. of students in group	Duration
V1	4	23 minutes 45 seconds
V2	2	
V3	4	
V4	5	
V5	4	
V6	3	

1∼3 curriculum levels. Within each cohort, we collected 10∼12 video sessions of 10-20 students collaborating in small groups of 3 to 6 members. Thus, it is clear that we need to develop methods that detect specific groups of students and track them throughout each video.

We tackle the problem of assessing long-term student participation into three subproblems that include (i) student group detection, (ii) student face recognition within the detected group, and (iii) dynamic participant tracking. We develop separate training and testing datasets for each problem as summarized in Table 2. We use different video sessions for training and testing. At the end, we use final testing datasets for measuring the performance of the integrated system. In what follows, we provide detailed descriptions of the different datasets used to develop our system.

A. AOLME-G VIDEO DATASET FOR STUDENT GROUP DETECTION

The AOLME-G video dataset has 54 videos from 52 groups, covering two cohorts. These videos will be used for group detection. We use AOLME-G to generate separate datasets for (i) training and validation: AOLME-GY1, AOLME-GF1, AOLME-GB1, and (ii) component testing datasets: AOLME-GF2 and AOLME-GB2. We then want to test the group detection system using the massive AOLME-GT dataset. We provide separate descriptions for each dataset.

1) AOLME-GY1 FOR FACE DETECTION TRAINING AND VALIDATION

We use 1000 faces and 1200 non-face images from student groups extracted from the AOLME-G dataset to train the YOLO face detector. Among the selected face images, we use 70% of the images for training and 30% for validation. For each group, we identify the faces of each group member.

2) AOLME-GF1 FOR GROUP FACE DETECTION TRAINING AND VALIDATION

The dataset is generated from the AOLME-G videos to train the group face classifier. The augmented dataset contained 56,045 group faces and 56,084 non-group face images. We use 70% of the dataset for training and 30% for validation.

3) AOLME-GB1 FOR TRAINING FOR BACK OF THE HEAD DETECTION

For the back-of-the-head classifier, the dataset uses over 45,000 frames from AOLME-G videos. It contains 22,768 back-of-the-head images and 22,800 other images.

4) AOLME-GF2 FOR GROUP FACE DETECTION TESTING

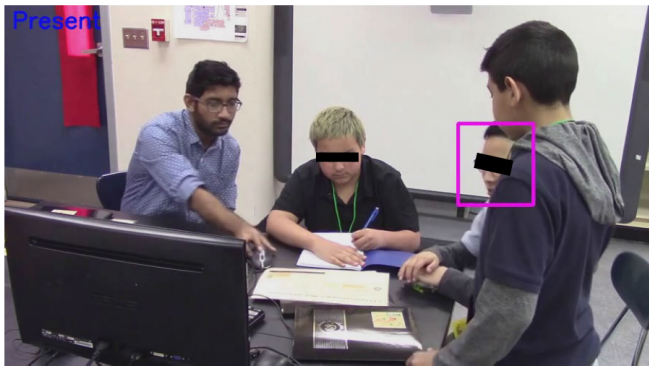
The dataset is generated from AOLME-G videos for testing the group face classifier. The dataset contains 14,011 group faces and 14,021 non-group face images. The numbers include seven-fold data augmentation performed using random rescaling, cropping, rotating, and flipping.

5) AOLME-GB2 FOR BACK OF THE HEAD TESTING

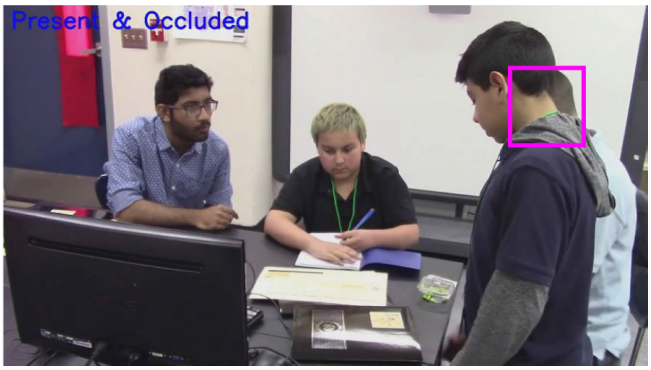
To test the back-of-the-head classifier, we used 5,710 back-of-the-heads and 5,682 others from the AOLME-G video dataset.

6) AOLME-GT: A LARGE DATASET FOR FINAL TESTING OF GROUP DETECTION

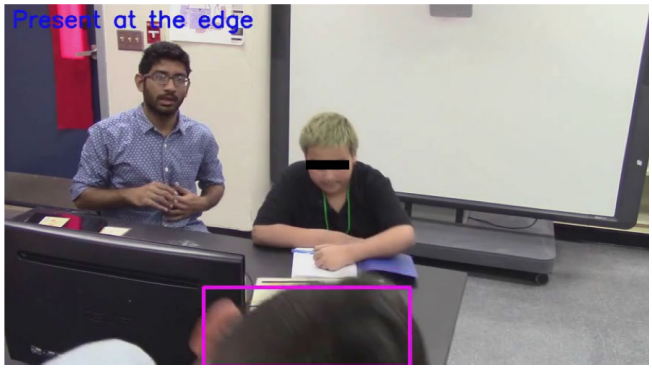
We test the group detection methodology with a set of 13 videos containing 12,518,250 student labels. The student labels identify whether a student belongs to a group or not.



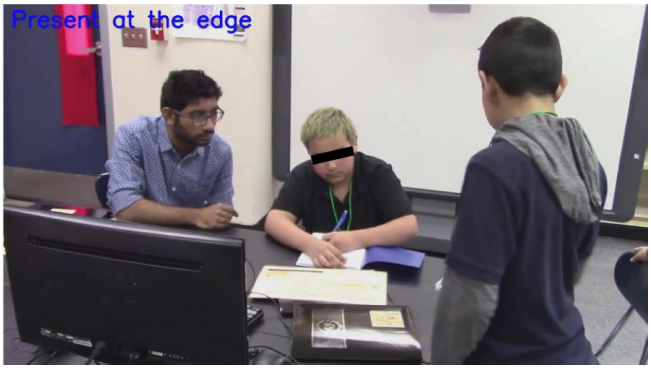
(a) The student is present appearing at an angle with partial occlusion.



(b) The student is present and fully occluded.



(c) The student is present at the edge of the frame.



(d) The student is present with his hand in the lower-right edge of the frame.

FIGURE 2. A simple example to demonstrate the issues for training and testing dynamic participant tracking. In this example, we only show annotation for a single student per image. We note that there is no bounding box for the student in (d) because he is not visible. For the training and testing datasets, in each frame, we mark all of the students for each group.

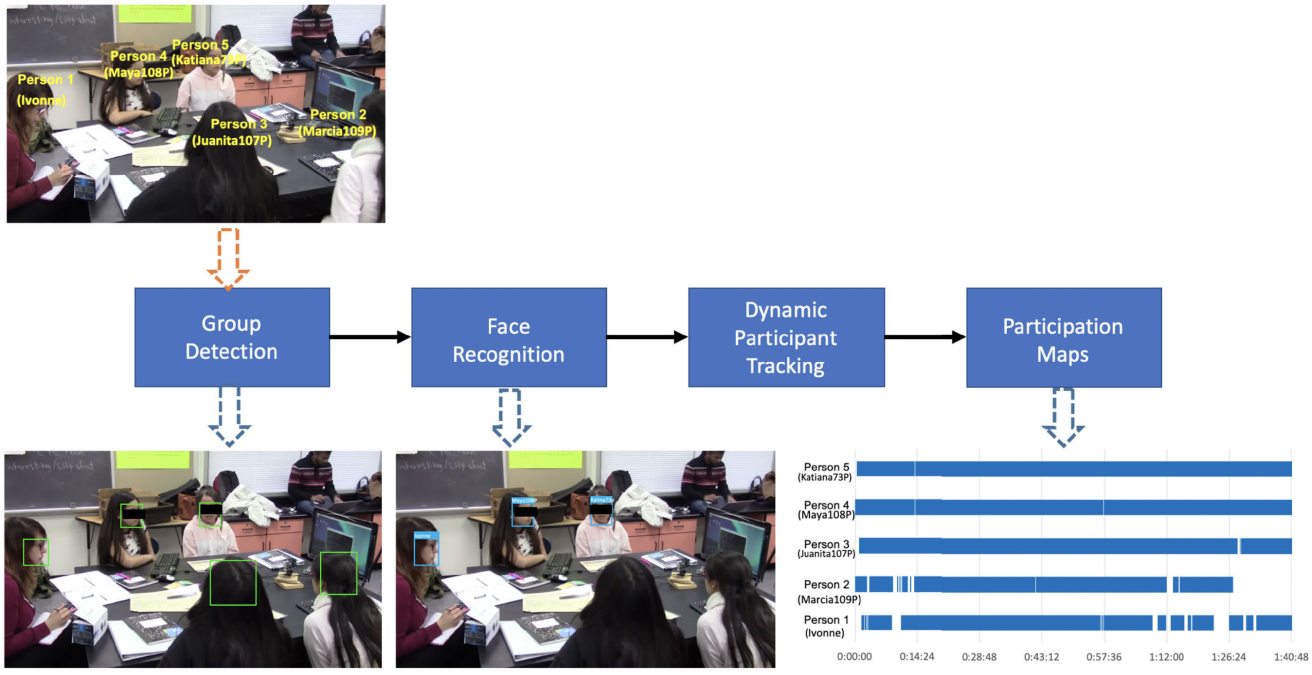


FIGURE 3. AOLME student participation analysis system. We detect groups every second. We perform face recognition and dynamic participant tracking every frame.

Overall, the combined duration of all of the AOLME-GT videos is 21 hours and 22 minutes.

B. AOLME-FR VIDEO FACE RECOGNITION TRAINING

The AOLME-FR video dataset is used for training the video face recognition algorithms. These video images were sampled from 13 sessions that cover level 1 of cohorts 2 & 3. Overall, the combinations of training and testing videos are 4 hours long.

1) AOLME-FR1 FOR TRAINING VIDEO FACE RECOGNITION

Within AOLME-FR, we separate out the AOLME-FR1 dataset that consists of 3,968 images for identifying up to 42 students and student facilitators. The dataset is used to generate face prototypes associated with each participant as described in the methodology. Each prototype is resized to 112×112 pixels.

2) AOLME-D FOR SYSTEM TESTING OF VIDEO FACE RECOGNITION FROM RAW INPUT VIDEOS

The AOLME-D video dataset has 13 different sessions of 1 to 1.5 hours each from urban and rural schools. We use AOLME-D to derive a collection of short videos (AOLME-DST) and long videos (AOLME-DLT) for final system testing.

3) AOLME-DST: SHORT VIDEOS DATASET FOR FINAL SYSTEM TESTING

The AOLME-DST dataset is summarized in Table 3. This diverse dataset contains a selection of short video samples that are used to test system performance under occlusion for 35 students from 5 groups. The AOLME-DST dataset is designed to provide exhaustive testing in many different scenarios. In the results, we will provide detailed results for each group.

4) AOLME-DLT: LONG VIDEOS DATASET FOR FINAL SYSTEM TESTING

The AOLME-DLT contains raw real-life videos as detailed in Table 4. The videos are broken into shorter videos that are 23 minutes and 45 seconds. This final dataset will be used to test all aspects of our system using different groups and a diverse set of students.

C. DATASETS FOR DYNAMIC PARTICIPANT TRACKING

The ultimate goal of dynamic participant tracking is to quantify student participation. Thus, we need to know whether a specific student is present within a group. Students are marked as present even if they do not appear in the frame due to occlusion. Thus, in order to develop ground truth for dynamic participant tracking, we review the entire video from beginning to end to eliminate false negatives due to occlusion. Furthermore, in most cases, students are partially occluded and are free to move around while remaining close to the table. In all such cases, we assume that the students are present. We only mark students as not-present if they are completely missing from several video frames over several seconds.

We present four occlusion examples in Fig. 2. In all cases, we mark the student as present. Yet, the student is partially occluded in Fig. 2(a), fully occluded in Fig. 2(b), and at the edge of the frame in Fig. 2(c). In Fig. 2(d), a small portion of his hand is visible in the lower-right edge of the video frame.

We used the Matlab video labeler to mark the presence of each student in each frame of each video. For each video frame, we carefully mark the locations of all students within each group.

1) AOLME-DST FOR SYSTEM TESTING OF DYNAMIC PARTICIPANT TRACKING

We perform both short-term and long-term testing of the ability of the system to perform dynamic participant tracking. For short-term testing, we use 35 short video segments ranging from 10 seconds to 150 seconds long at a frame rate of 30 fps. Overall, short-term testing consisted of 17 minutes and 17 seconds. The video examples include occlusion of at-least one person as detailed in Table 3.

2) AOLME-DLT FOR SYSTEM TESTING OF DYNAMIC PARTICIPANT TRACKING ON LONG-DURATION VIDEOS

We use a second dataset to test our dynamic participant tracking system over long video segments. Six long videos from different groups from the AOLME-CT video dataset are used to generate separate testing videos for AOLME-DLT. For long-term testing, each video is 23 minutes 45 seconds at a frame rate of 30 fps with 3 to 5 recognizable persons per video as described in Table 4. As for the short-term video dataset, we mark the location of every person in each video frame.

III. BACKGROUND

A. RELATED WORK

1) GROUP DETECTION

We formulate the problem of group detection as a problem of detecting the students working together and sitting at the table nearest to the camera. Beyond the classic problem of human detection, group detection requires that we detect humans at arbitrary angles, while facing the camera and also while looking away from the camera. In this subsection, we will summarize some related research done by our group, published in a conference paper, and outline the new research summarized in the current paper.

We reported on initial research of combining YOLO with AM-FM representations for group detection in a conference paper in [10]. In our basic approach, we used YOLO for face detection. YOLO generated a large number of false face detections that belonged to different student groups. To address the problem, we relied on the fact that student faces that are far away from the camera are characterized by high instantaneous-frequency components. We thus used FM feature extraction and a simple LeNet5 network to

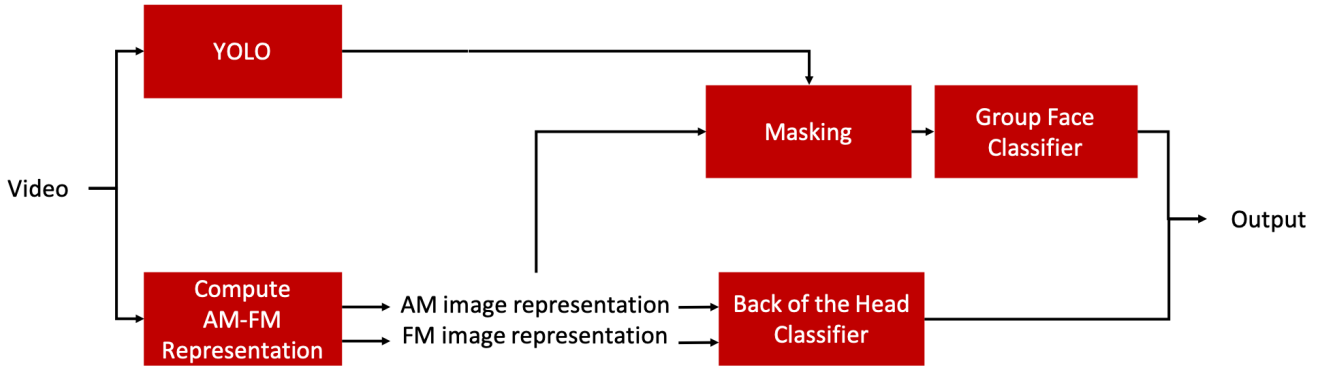
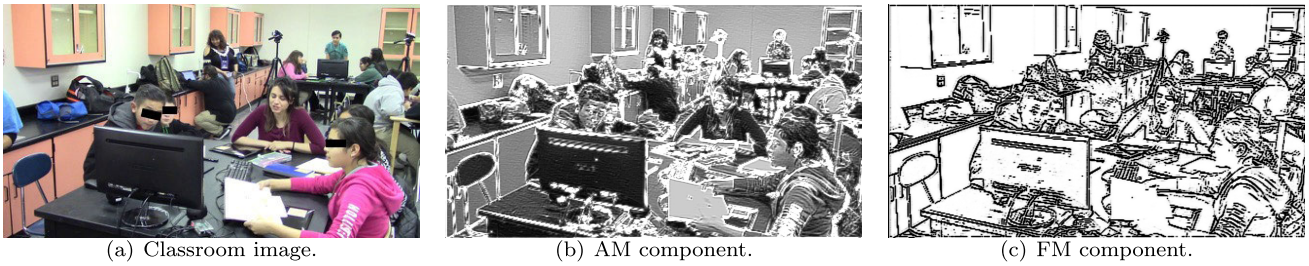


FIGURE 4. Student group detection system.



(a) Classroom image. (b) AM component. (c) FM component.

FIGURE 5. AM-FM representation of the classroom environment.

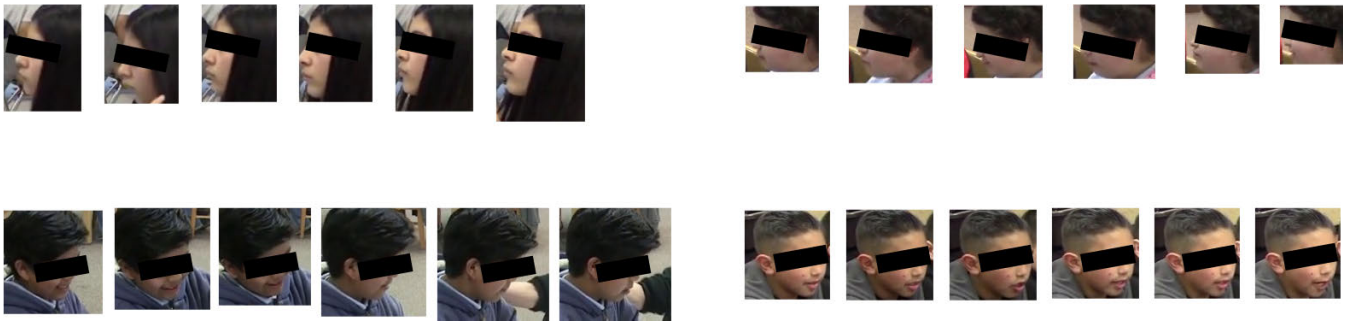


FIGURE 6. Face prototype samples of four students.

remove false face detections and also detect back of the head students facing away from the camera. Here, we note that the advantages of the FM representations come from the fact that they are explainable and provide additional image representations that go beyond the standard raw images processed by YOLO. We will employ this system for student group detection.

2) FACE RECOGNITION

In order to recognize the student participant, following face detection, we use face recognition. Here, we note that face recognition is a very mature research area for the case when the humans are facing the camera. Unfortunately, this is not the case here. We are faced with several challenges since the students are not posing for the camera. Instead, they can be at arbitrary angles. Our approach was to adopt a state of the art system face recognition system and retrain it for video face recognition for our current problem. Thus, for our baseline system, we use the InsightFace system [1] that is based on Additive Angular Margin Loss for Deep Face Recognition (ArcFace). Here, we note that ArcFace has been tested on a large number of camera-facing image datasets and a variety of loss function models. We have summarized our modified system in a conference paper in [11]. For completeness, we will provide a summary of our methodology adopted from [11] in our methods section.

3) TRACKING UNDER OCCLUSION

Following person recognition and face recognition, we are faced with the problem of tracking under occlusion. As mentioned in the introduction, previously considered methods include the use of correlation filters in [2], a classifier approach [3], convolutional neural networks in [4], and a geometric approach in [5] and [6]. As noted earlier, we will be comparing our approach to the Simple Online and Real-time Tracking with Occlusion Handling (SORT_OH [5]) which achieved state-of-the-art results on the MOT16/17 datasets for pedestrian tracking.

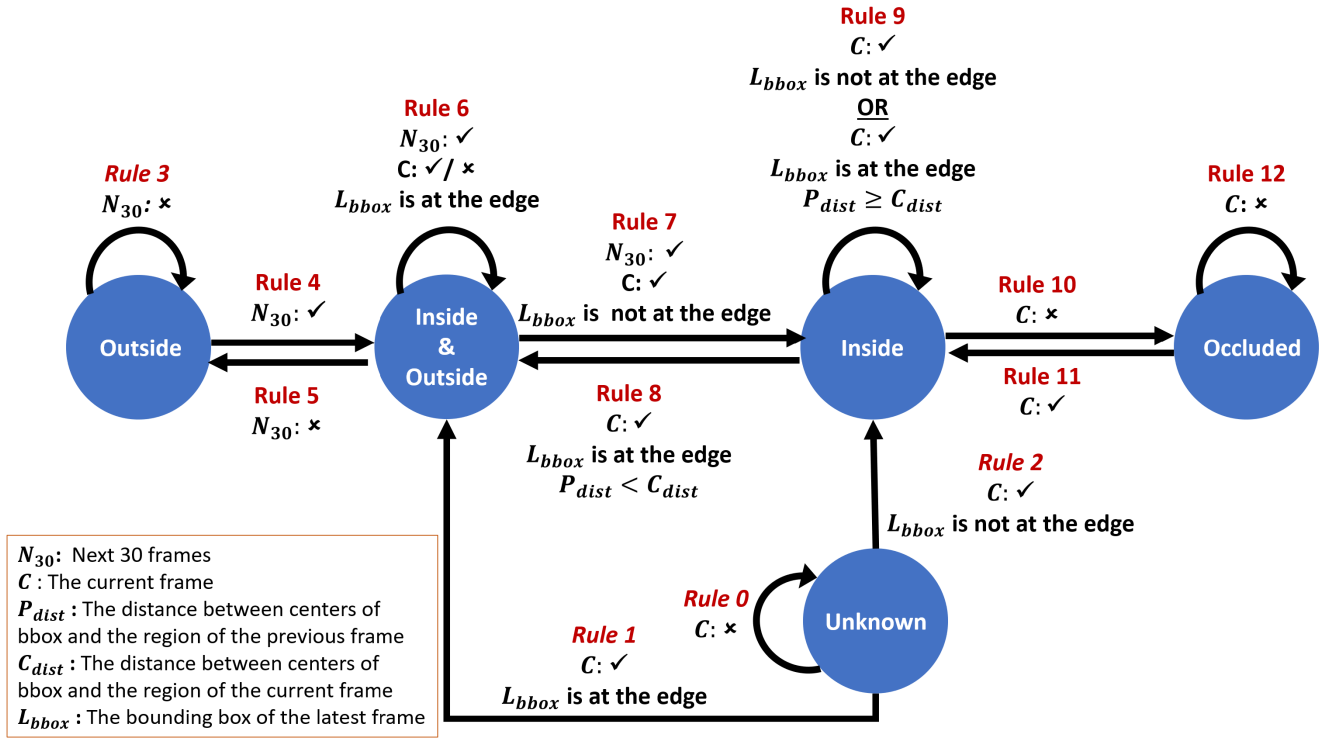


FIGURE 7. Dynamic participant tracking system. Here, bbox refers to the bounding box.

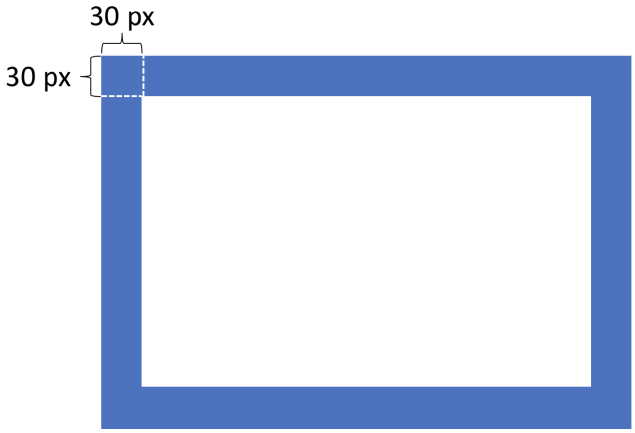


FIGURE 8. The edge of the video frame is defined as the set of pixels located within 30 pixels of the edge of the frame.

We also provide a summary of other research in this area. In [20], the authors present a novel approach for visual object tracking that discriminates occlusion from the self-deformation of the target. In [21], the authors evaluate the performance of visual object trackers in challenging occluded scenarios by creating a small dataset that includes sequences with multiple instances of hard occlusions. In [6], the authors developed a regression-based multi-pedestrian tracker that can re-track targets without an extra re-identification model. The paper reports a method for improving track management by regressing inactive tracks and also developing a method for dealing with tracks that are out of the camera's view. In [4], the authors develop an object-tracking method based on the combination of correlation filters and ResNet features. The paper describes the use of response maps by extracting features from different layers of ResNet, and then fusing response maps using the AdaBoost algorithm. In [2], the authors propose the Kernelized Correlation Filter (KCF) model to track ships in consecutive maritime images and then use the tracking to estimate ship trajectories. In [3], the authors present an integrated Circulant Structure Kernels (ICSK) tracking framework to handle occlusion by estimating target objects' translation and scale variations.

The paper describes a new method to support dynamic participant tracking that can deal with long-term occlusions and persons entering and leaving the scene. Our DPT uses a finite-state machine to track each person. Transitions between states are based on intuitive geometrical constraints. As we discuss in the results, the DPT is proven to be very effective on real-life AOLME videos.

IV. METHODOLOGY

A. OVERVIEW

We present a top-level diagram of the entire system in Fig. 3. The raw input video is first processed through group detection to identify the students with the current group while rejecting people in the background that do not

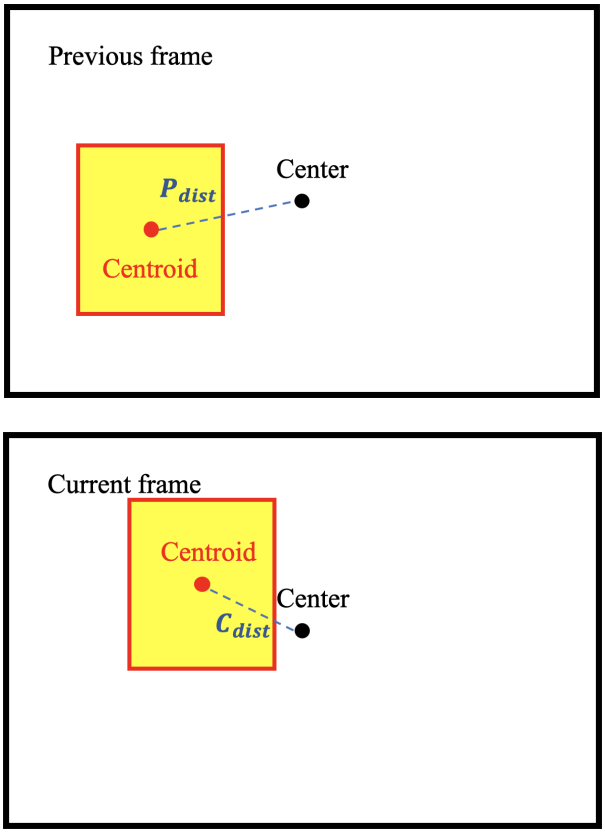


FIGURE 9. State determination is based on the distance from the center of the frame to the centroid of the object detection bounding box.

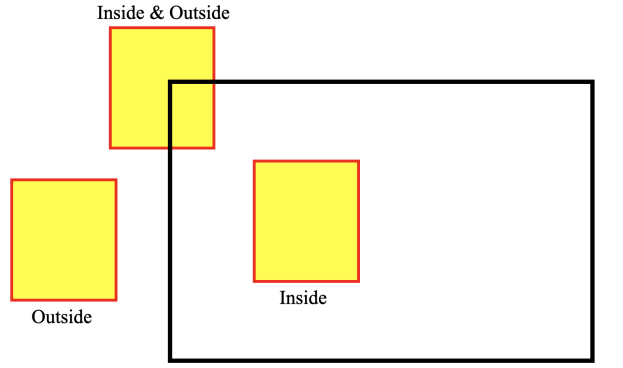


FIGURE 10. Definitions of 'Outside', 'Inside & Outside', and 'Inside' states.

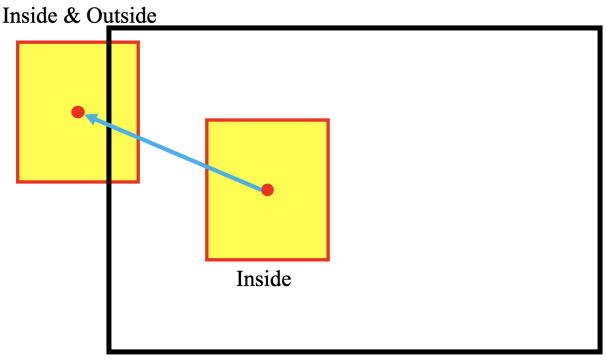


FIGURE 11. One example of DPT transition rules: rule 8.

belong to the current group. We then identify the students for whom we can detect faces based on a face recognition system. We use dynamic participant tracking for all identified students to account for cases where students may move or leave the scene. Then, we combine the information to produce participation maps documenting student participation through time. Informed consent was obtained for all study participants.

B. GROUP DETECTION

For group detection, we need to detect the students sitting at the table closest to the camera. Group detection is based on face detection for students facing the camera and back-of-the-head detection for students facing away. We present a system diagram of the group detection system in Fig 4. Due to the need for speed, we use YOLO for face detection. Back-of-the-head detection is performed based on extracted AM-FM features as described next.

AM-FM components are extracted from the grayscale (Y-component) using dominant component analysis (DCA) estimated using a 54-channel Gabor filterbank as described in [8]. Using DCA, the input image frame is approximated by: $I(x, y) \approx a(x, y) \cos \varphi(x, y)$ where $a(x, y)$ denotes the AM component and $\cos \varphi(x, y)$ denotes the FM component. Fig 5 shows an example of the extracted AM-FM components.

The FM image is masked by the results of the YOLO face detector. We apply this step to extract the FM components over students within the desired group while rejecting detections from other groups. FM components over the faces of the closest group exhibit lower frequency components than the higher frequency components associated with distant faces from other groups. To detect the group faces, we thus apply a simple, LeNet-based classifier [22] on the extracted FM components over 100×100 pixel regions.

The AM-FM components are also used to detect the hair and back-of-the-head candidate regions described in [8]. A LeNet-based classifier is used to detect the back-of-the-heads against background detections, as detailed in [8]. We detect the entire group for each video frame by concatenating the results from the face and back-of-the-head classifiers.

C. FACE RECOGNITION

We adopt the face recognition method previously described as a conference paper in [11]. We use the InsightFace [23] system to recognize faces. The face recognition system requires a set of face prototypes associated with each participant.

We combine sparse sampling and K-means clustering to compute face prototypes as given in Algorithm 1 (also see [11]).

Fig 6 displays some samples of face prototypes.

Algorithm 1 Compute Face Prototypes Using Sparse Sampling and K-Means

Input: Video clips associated with each participant
Output: facePrototypes associated with each participant
1: **for** each participant
2: **Sample** an image every 30 frames of video
3: **Apply** K-means clustering
4: **Select** cluster means
5: **Find** the nearest images from cluster centroids
6: **Align** faces to 112 × 112

1) SPARSE SAMPLING

To achieve sparsity, the algorithm extracts a single sample image per second of video with a frame rate equal to 30 fps.

2) K-MEANS CLUSTERING

We use clustering to describe different face poses. The algorithm searches for the training image that is closest to a cluster centroid to prevent the usage of centroids that might be impractical. Once the algorithm has identified a prototype image that is closest to the mean, it proceeds to align and resize every image to 112 × 112 pixels. This paper uses K-means with 64 clusters for data training.

Once the face prototypes are obtained, we use the InsightFace system to identify the students. Seven-fold data augmentation is implemented to increase the training dataset. Data augmentation is based on random rescaling, cropping, rotating, and left-to-right flipping. This process generates a total of 18,816 prototype faces for training and validation.

The algorithm for face recognition uses the MTCNN model to detect faces in the video. It then calculates the minimum distances to the face prototypes to identify each participant.

D. DYNAMIC PARTICIPANT TRACKING

We develop the Dynamic Participant Tracking (DPT) system to account for the presence of the students in relation to the camera as shown in Fig.7. More specifically, during the tracking process, a participant is classified as being 'Inside' or 'Outside' the video frame, in the process of leaving the scene ('Inside & Outside'), occluded by another object ('Occluded'), or being in an undetermined state ('Unknown'). In what follows, we begin the section by providing definitions of each state. We then describe how to determine whether a participant is in one of the states and how to transition from state to state. Here, we note that state transitions are based on the current state and the participant detection results. We also note that the DPT is applied separately for each participant.

1) EDGE

We define the edge of the video frame to be the pixels less than 30 pixels from the edge, as shown in Fig. 8.

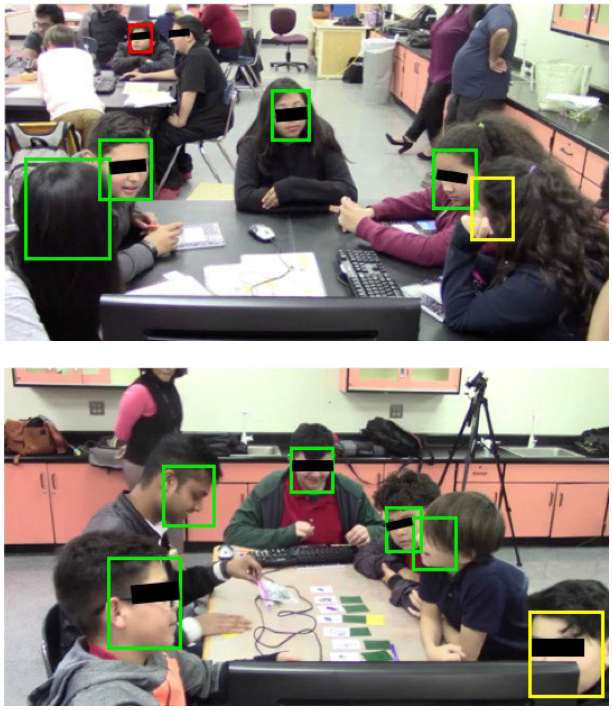


FIGURE 12. Head detection system results. True positives are bounded by green boxes. False positives are bounded by red boxes. False negatives are bounded by yellow boxes. For successful detection, we require the intersection over union (IOU) score to be at least 0.6.

2) THE DISTANCE BETWEEN CENTERS

We define the centroid distance D (Fig. 9) between the center of the frame and a bounding box that exists in the frame using:

$$D = \sqrt{|x2 - x1|^2 + |y2 - y1|^2}.$$

3) STATES

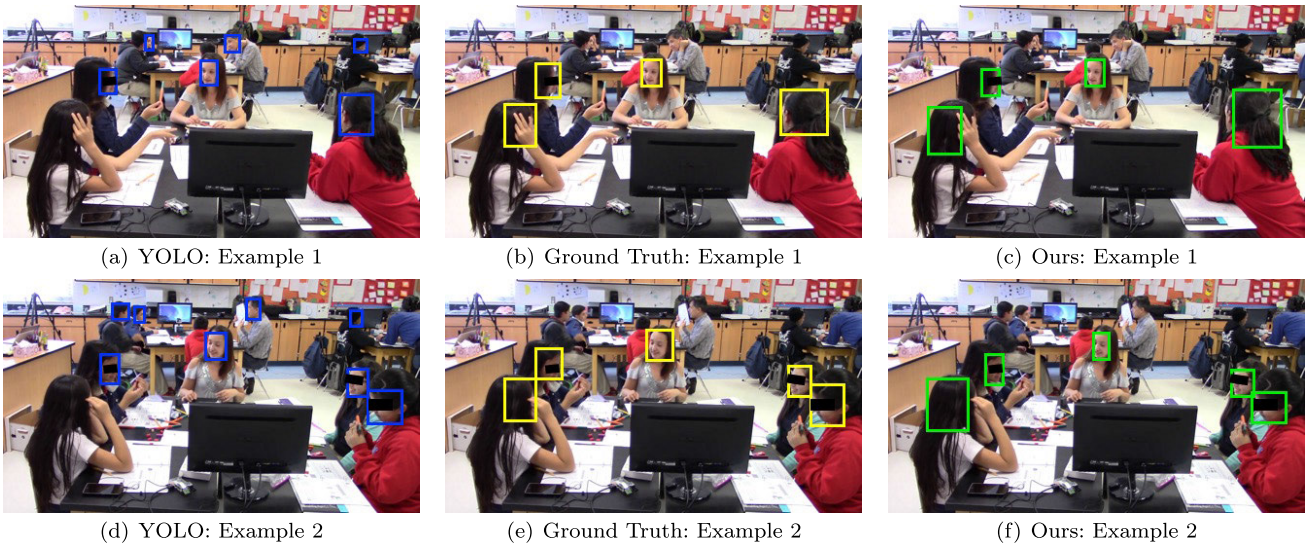
The state of each participant is based on the location of the bounding box resulting from participant detection. All participants are initially in 'Unknown' state if not detected. This state remains unknown as long as they are not detected in the current state.

If a person is detected in the initial frame, their location is used to determine their initial state. Thus, if the person is detected entirely inside the frame, their state is set as 'Inside'. If the person is detected at the edge of the frame, their initial state is set to 'Inside & Outside'.

In Fig.10, we demonstrate how the locations of the bounding boxes are used to determine the states 'Outside', 'Inside & Outside', and 'Inside'. The black rectangle represents the frame edge. The red-yellow rectangles represent the bounding boxes. This frame has no bounding box if a person is occluded because other people or objects cover them.

4) INPUTS

We define the five possible inputs for determining transitions between states as follows.



(a) YOLO: Example 1 (b) Ground Truth: Example 1 (c) Ours: Example 1

(d) YOLO: Example 2 (e) Ground Truth: Example 2 (f) Ours: Example 2

FIGURE 13. Examples from Group Detection Results. Left column ((a) and (d)) shows the results of YOLO. Middle-column ((b) and (e)) shows the ground truth. Right-column ((c) and (f)) shows the results of the proposed method.

TABLE 5. Comparative results for student group detection over 13 videos. TP, FP, and FN refer to true positives, false positives, and false negatives, respectively. F1 scores are given for each video and each method. The videos represent different student groups based on the AOLME-GT dataset. Here, recall that there is a large number of students in each image that do not belong to the current group. Hence, the number of detected persons is higher than the number of labeled students.

Video	Length	Labeled Persons	Method	Detected Persons	TP	FP	FN	F1
V1	96 min	1,627,320	YOLO	1,915,935	1,153,959	761,976	124,527	0.72
			Proposed Method	1,397,790	1,183,630	214,160	344,640	0.81
V2	85 min	887,700	YOLO	1,274,429	723,283	551,146	12,153	0.72
			Proposed Method	847,250	728,110	119,140	110,140	0.86
V3	117 min	1,063,300	YOLO	792,291	720,762	71,529	321,159	0.79
			Proposed Method	819,700	745,880	73,820	293,640	0.80
V4	108 min	1,139,850	YOLO	1,212,963	839,450	373,513	120,252	0.77
			Proposed Method	950,210	859,290	90,920	242,850	0.84
V5	88 min	1,233,540	YOLO	1,380,190	1,047,410	266,980	84,780	0.86
			Proposed Method	1,046,000	1,008,042	4,345	205,507	0.91
V6	103 min	1,162,740	YOLO	1,988,410	1,274,280	591,570	110,490	0.78
			Proposed Method	1,488,619	1,166,396	120,746	294,325	0.85
V7	90 min	667,500	YOLO	955,580	503,000	407,970	26,900	0.70
			Proposed Method	832,695	465,002	291,258	41,159	0.74
V8	111 min	915,370	YOLO	967,110	825,200	116,530	58,950	0.90
			Proposed Method	932,898	823,727	73,928	54,368	0.93
V9	108 min	946,710	YOLO	1,112,350	810,120	250,720	48,510	0.84
			Proposed Method	1,085,975	783,624	235,303	52,991	0.84
V10	106 min	712,050	YOLO	669,090	657,890	6,740	50,290	0.96
			Proposed Method	684,413	655,464	6,497	37,016	0.97
V11	83 min	797,580	YOLO	1,002,010	677,390	285,170	32,510	0.81
			Proposed Method	950,187	660,054	239,776	34,892	0.83
V12	106 min	829,930	YOLO	848,310	615,790	159,280	122,030	0.81
			Proposed Method	615,088	584,524	3,158	236,436	0.83
V13	81 min	534,660	YOLO	528,700	465,340	40,080	46,100	0.92
			Proposed Method	505,876	453,333	17,181	49,941	0.93
Total	21h 22m	12,518,250	YOLO	14,647,368	10,313,874	3,883,204	1,158,651	0.80
			Proposed Method	12,156,701	10,117,076	1,490,232	1,997,905	0.85

We use C to denote the detection in the current state. Thus, C : ✓ denotes successful detection. On the other hand, C : ✗ denotes failure to detect a participant in the current frame.

We use N_n to denote the detection of a participant over n frames. Thus, N_n : ✓ denotes successful detection in any of the following n frames while N_n : ✗

TABLE 6. Comparison of DPT versus SORT_OH [5] for group 1 (1 out of 5). The results are computed over the AOLME-DST dataset.

	Method	Person Label				Average
		Kenneth1P	Jesus69P	Chaitu	Javier67P	
V1	SORT_OH	27.8 %	47.6%	93.7%	99.7%	67.2%
	Ours	100%	81.7%	100%	100%	95.4%
V2	SORT_OH	82.7%	95.2%	100%	100%	94.5%
	Ours	100%	100%	100%	100%	100%
V3	SORT_OH	75%	100%	100%	100%	93.7%
	Ours	100%	100%	100%	100%	100%
V4	SORT_OH	83.6%	100%	100%	100%	95.9%
	Ours	100%	100%	100%	100%	100%
V5	SORT_OH	74.5%	100%	92.1%	100%	91.6%
	Ours	100%	100%	91.4%	100%	97.8%
V6	SORT_OH	92.4%	100%	100%	100%	98.1%
	Ours	100%	100%	100%	100%	100%
V7	SORT_OH	87.4%	100%	100%	100%	96.9%
	Ours	100%	100%	100%	100%	100%
V8	SORT_OH	99.2%	100%	100%	100%	99.8%
	Ours	100%	100%	100%	100%	100%
V9	SORT_OH	81.8%	100%	100%	100%	95.4%
	Ours	100%	100%	100%	100%	100%
V10	SORT_OH	50.8%	100%	100%	100%	87.7%
	Ours	100%	100%	100%	100%	100%
V11	SORT_OH	81.4%	100%	100%	100%	95.3%
	Ours	100%	100%	100%	100%	100%
V12	SORT_OH	43.6%	100%	100%	100%	85.9%
	Ours	100%	100%	100%	100%	100%
V13	SORT_OH	100%	100%	86%	100%	96.5%
	Ours	100%	100%	90%	100%	97.5%
V14	SORT_OH	98.7%	80.8%	100%	100%	94.9%
	Ours	100%	100%	100%	100%	100%
V15	SORT_OH	100%	90.7%	100%	100%	97.7%
	Ours	100%	100%	100%	100%	100%

TABLE 7. Comparison of DPT versus SORT_OH [5] for group 2 (2 out of 5). The results are computed over the AOLME-DST dataset.

	Method	Person Label						Average
		Kelly	Cindy14P	Carmen13P	Marina15P	Marta12P	Scott	
V1	SORT_OH	100%	100%	100%	97.5%	100%	100%	99.6%
	Ours	100%	100%	100%	100%	100%	100%	100%
V2	SORT_OH	100%	100%	100%	81.4%	100%	100%	96.9%
	Ours	100%	100%	100%	100%	100%	100%	100%
V3	SORT_OH	99.2%	100%	100%	100%	100%	100%	99.9%
	Ours	100%	100%	100%	100%	100%	100%	100%
V4	SORT_OH	100%	100%	100%	62.3%	100%	100%	93.7%
	Ours	100%	100%	100%	100%	100%	100%	100%
V5	SORT_OH	100%	100%	100%	96.3%	100%	94.8%	98.5%
	Ours	100%	93%	100%	100%	100%	99.3%	98.7%
V6	SORT_OH	100%	100%	100%	83.1%	100%	100%	97.2%
	Ours	100%	100%	100%	100%	100%	100%	100%
V7	SORT_OH	100%	100%	100%	85.7%	99.9%	100%	97.6%
	Ours	100%	100%	100%	100%	98.5%	100%	99.7%

denotes failure to detect a participant in the subsequent n frames.

We use P_{dist} to denote the distance between the centroid of the bounding box and the center of the previous frame.

We use C_{dist} to denote the distance between the bounding box's centroid and the current frame's center (See Fig.9).

We use L_{bbox} to denote the location of the latest detection with a bounding box.

TABLE 8. Comparison of DPT versus SORT_OH [5] for group 3 (3 out of 5). The results are computed over the AOLME-DST dataset.

	Method	Person Label					Average
		Shelby	Cindy14P	Cesar61P	Emily62P	Mauricio60P	
V1	SORT_OH	100%	100%	94%	100%	100%	98.8%
	Ours	100%	100%	100%	100%	100%	100%
V2	SORT_OH	100%	100%	60.5%	100%	100%	92.1%
	Ours	100%	100%	100%	100%	100%	100%
V3	SORT_OH	100%	100%	76.2%	100%	100%	95.2%
	Ours	100%	100%	100%	100%	100%	100%
V4	SORT_OH	100%	100%	67.8%	100%	100%	93.6%
	Ours	100%	100%	100%	100%	100%	100%
V5	SORT_OH	96.6%	100%	79.2%	100%	100%	95.1%
	Ours	100%	100%	100%	100%	100%	100%
V6	SORT_OH	100%	100%	45.8%	100%	100%	89.2%
	Ours	100%	100%	100%	100%	100%	100%

TABLE 9. Comparison of DPT versus SORT_OH [5] for group 4 (4 out of 5). The results are computed over the AOLME-DST dataset.

	Method	Person Label					Average
		Julia7P	Martina64P	Bernard129P	Suzie66P	Issac	
V1	SORT_OH	88%	28.7%	100%	54%	93.6%	72.9%
	Ours	86.7%	100%	100%	58.4%	97.1%	88.4%

TABLE 10. Comparison of DPT versus SORT_OH [5] for group 5 (5 out of 5). The results are computed over the AOLME-DST dataset.

	Method	Person Label						Average
		Irma	Herminio10P	Juan16P	Jorge17P	Emilio25P	Jacinto51P	
V1	SORT_OH	95.2%	100%	76.4%	100%	100%	63.6%	89.2%
	Ours	92.5%	100%	100%	100%	100%	100%	98.8%
V2	SORT_OH	100%	100%	100%	97.1%	100%	73.4%	95.1%
	Ours	100%	100%	100%	100%	100%	100%	100%
V3	SORT_OH	96.2%	100%	100%	79.3%	99.9%	99.7%	95.9%
	Ours	95.7%	100%	99.9%	100%	100%	100%	99.3%
V4	SORT_OH	93.2%	100%	85.2%	100%	100%	77.2%	92.6%
	Ours	91.1%	100%	99.9%	100%	100%	100%	98.5%
V5	SORT_OH	100%	100%	100%	100%	100%	60.1%	93.4%
	Ours	100%	100%	100%	100%	100%	100%	100%
V6	SORT_OH	97.6%	100%	68.5%	100%	100%	51.4%	86.3%
	Ours	93.6%	100%	100%	100%	100%	100%	98.9%

5) DPT TRANSITIONS

The initial states can be 'Inside', 'Unknown', or 'Inside & Outside' as described in the DPT states subsection. Here, we describe transitions among other states. We note that for each state, we consider all possible input combinations for determining how to transition to another state.

From the 'Inside' state, a participant can move to the 'Inside & Outside' state, 'Inside' state, or the 'Occluded' state as given below:

- To transition to the 'Inside & Outside' state, we detect a movement inside the frame toward outside the frame. Here, the movement is detected by requiring that 1) the person is detected in the current frame, 2) the centroid's distance of previous frame is less than the one of current frame, and 3) the person is detected at the edge of the current frame. We simplify these rules by using: C : ✓, $P_{dist} < C_{dist}$, and L_{bbox} is at the edge (rule 8) (See Fig.11).
- To remain in the 'Inside' state, we detect a movement inside the frame. Here, the movement is detected by requiring that 1) C : ✓and L_{bbox} is not at the edge, or 2) C : ✓, $P_{dist} \geq C_{dist}$, and L_{bbox} is at the edge (rule 9).
- To transition to the 'Occluded' state, we detect the movement disappears inside the frame. Here, it requires that C : ✗ (rule 10).

From the 'Outside' state, a participant can move to the 'Outside' state or the 'Inside & Outside' state as given below:

- To remain in the 'Outside' state, we detect no movement in the frame. Here, it requires that N_n : ✗ (rule 3).

TABLE 11. Final system results over the raw, real-life video dataset of AOLME-DLT. The use of DPT provided substantially better results than the frame-based results that did not use DPT. The duration of each video is 23 minutes and 45 seconds.

	Label	Accuracy	
		No DPT	DPT
V1	Chaitanya	81.5%	86.8%
	Kenneth1P	47.3%	75.5%
	Jesus69P	91.3%	90.4%
	Javier67P	21.2%	53.9%
	Average	60.3%	76.7%
V2	Phuong	53.5%	64.6%
	Melly77W	97.4%	98.6%
	Average	75.5%	81.6%
V3	Bernard129P	38.0%	77.3%
	Julia7P	56.7%	81.4%
	Martina64P	31.6%	60.8%
	Suzie66P	62.3%	89.6%
	Average	47.1%	77.3%
V4	Herman78W	79.7%	87.8%
	Laura80W	40.4%	61.6%
	Lucia81W	78.8%	92.2%
	Mario130W	30.0%	84.6%
	Melly77W	93.5%	96.6%
	Average	64.5%	84.6%
V5	Herminio10P	86.9%	93.2%
	Katiana73P	86.1%	92.4%
	Guillermo72P	70.1%	88.0%
	Beto71P	38.9%	84.8%
	Average	70.5%	89.6%
V6	Ivonne	80.5%	91.3%
	Juanita107P	71.0%	90.9%
	Katiana73P	7.8%	70.2%
	Average	53.4%	84.1%

- To transition to the 'Inside & Outside' state, we detect a movement from outside the frame toward inside the frame. The movement is detected by requiring N_n : ✓ (rule 4).

From the 'Inside & Outside' state, a participant can move to the 'Outside' state, 'Inside & Outside' state, or the 'Inside' state as given below:

- To transition to the 'Outside' state, we detect a movement from the frame toward outside the frame. The movement is detected by requiring N_n : ✗ (rule 5).
- To remain in the 'Inside & Outside' state, we detect a movement around the edge of the frame. Here, the movement is detected by requiring that N_n : ✓ C : ✓/✗ and L_{bbox} is at the edge (rule 6).
- To transition to the 'Inside' state, we detect a movement from outside the frame toward inside the frame. Here, the movement is detected by requiring that N_n : ✓ C : ✓ and L_{bbox} is not at the edge (rule 7).

From the 'Occluded' state, a participant can move to the 'Inside' state or the 'Occluded' state as given below:

- To transition to the 'Inside' state, we detect a movement that appears inside the frame. The movement is detected by requiring C : ✓ (rule 11).
- To remain in the 'Occluded' state, we detect no movement inside the frame. Here, it requires that C : ✗ (rule 12).

From the 'Unknown' state, a participant can move to the 'Unknown' state, 'Inside & Outside' state, or the 'Inside' state as given below:

- To remain in the 'Unknown' state, we detect no movement inside the frame. Here, it requires that C : ✗ (rule 0).
- To transition to the 'Inside & Outside' state, we detect a movement that appears in the frame. Here, the movement is detected by requiring that C : ✓ and L_{bbox} is at the edge (rule 1).
- To transition to the 'Inside' state, we detect a movement that appears in the frame. Here, the movement is detected by requiring that C : ✓ and L_{bbox} is not at the edge (rule 2).

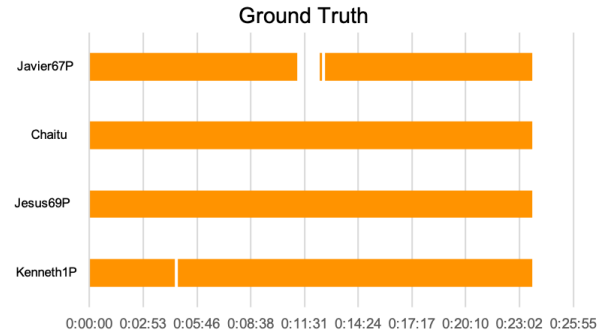
In this paper, for N_n, we set up $n = 30$ because the frame rate of video is 30 fps and students in the videos have big movement. We note that $n = 30$ represents a second. Here, it is important to note that our parameters were intuitively set for 1-second transitions.

V. RESULTS

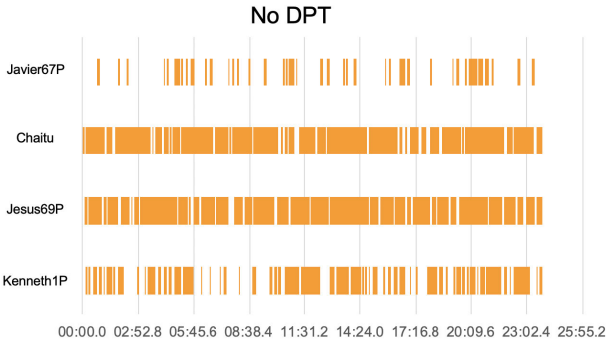
We summarize our results over the final testing datasets (see Table 2). First, we summarize our group detection results over the massive AOLME-GT dataset. For group detection, recall that we labeled 12,518,250 student instances in over 21 hours and 22 minutes of real-life videos (see section II-A6). For system testing, we will first present results over the carefully selected short videos of the AOLME-DST dataset (see section II-C1). We also summarize final system testing results over raw, real-life videos of the AOLME-DLT dataset (see section II-B4). We then present participation maps for visualizing the final results. As mentioned earlier, the testing datasets do not share any video sessions with the training and validation datasets.

A. STUDENT GROUP DETECTION TESTING USING AOLME-GT DATASET

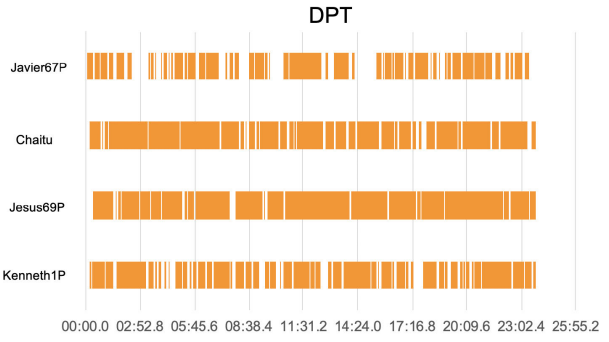
We begin with a simple example in Fig. 12. For students within the group of interest, we use green bounding boxes to indicate successful detections (true positives (TP)). We used yellow bounding boxes to denote false negatives (FN), when we fail to detect a student that belongs to the group. For students outside the group, we use red bounding boxes to indicate false positives (FP). In the top image of Fig. 12, we see that we have a false positive case for a background student facing the camera. Then, in the bottom image of Fig. 12, we have a false negative example where we could not detect a student that is partially occluded by the camera. Here, it is important to note that the false negative case can be corrected using the dynamic participant tracking algorithm. Assuming that a student was detected in an earlier frame, the DPT will



(a) Ground truth participation map.



(b) Participation map without DPT. The average accuracy is 60.3%.



(c) Participation map with DPT. The average accuracy is 76.7%.

FIGURE 14. Student participation maps on long videos of AOLME-DLT dataset (see V1 in Table 11). The results demonstrate the effectiveness of DPT (bottom) against ground truth (top). Without DPT, the lack of detection gives several false negatives (middle).

correctly classify the student as occluded and mark them as present.

We present comparative results of the proposed method and YOLO in Fig. 13. To differentiate among methods, we use blue bounding boxes for YOLO (left column), yellow bounding boxes for the ground truth (middle column), and green bounding boxes for the proposed method (right column). In this example, the proposed method successfully detected the entire group without giving any false positives. In contrast, YOLO, trained on the same dataset, gave several false positives by wrongly labeling background face detections as being part of the group. Most alarmingly, YOLO failed to detect a student that belongs to the group (see student in the lower-left part of the image in the left column). Our performance clearly benefited from the use of multiple representations and our back-of-the-head detector.

We provide comparative results over the AOLME-GT dataset in Table 5. Here, note that the numbers of detected persons are often higher than the number of labels as both methods may falsely identify background students as belonging to the current student group. Furthermore, false positives are associated with falsely labeling out of group students as being part of the group. On the other hand, false negatives are primarily due to occlusions. Our approach achieves a substantially lower number of false positives. We use the F1 score to assess overall performance (harmonic mean of precision and recall). We note that our proposed method performs better on all video examples (except for V9 where performance was the same). In many cases, the proposed method is substantially better, with over 0.07 improvement (e.g., V1 improved by 0.09, V2 improved by 0.14, V4 & V6 improved by 0.07). Overall, it yields an F1 score of 0.85 against 0.80 for YOLO.

B. DYNAMIC PARTICIPANT TRACKING AND FINAL SYSTEM TESTING RESULTS

This section provides comparative results of the DPT against SORT_OH as well as results over the raw, real-life video sessions. We also present the use of participation maps for visualizing the final results.

Following student group detection, we compare the performance of the DPT (proposed method) against SORT_OH for the AOLME-DST test dataset (see section II-C1). Our results include detailed performance analysis for each student participant in Tables 6, 7, 8, 9, and 10.

We note that our proposed method performs exceptionally on nearly every case. In the overwhelming majority of the test cases, we are able to dynamically track each participant with 100% accuracy. On the other hand, SORT_OH fails to track several students. Here, we define failure as the inability of the method to track students with more than 70% accuracy. We highlight failure examples in red in Tables 6, 7, 8, 9, and 10. In our 35 test video sequences, we can see 14 examples of failures by SORT_OH. Out of the five groups, we can see that SORT_OH has at-least one failure to track example in each student group. In comparison, our proposed method failed on just one example (see Table 9).

We believe that the efficacy of the DPT is due to its simplicity. The finite-state transitions were derived based on intuitive rules that did not require training. The only parameter used by DPT is to require the persistence of each transition rule over 30 frames (=1 second).

We report final testing results over raw, real-life video sessions of 23 minutes and 45 seconds of the AOLME-DLT dataset (see section II-C2). In Table 11, we compare DPT against not using any tracking. The results clearly illustrate the need for dynamic tracking. The overall accuracy improved from 61.9% to 82.3% when using the DPT.

We demonstrate the use of participation maps for visualizing student participation in Fig. 14. We note that the ground

truth plot of Fig. 14(a) suggests that there are long periods where the students are present. Without DPT, as shown in Fig. 14(b), tracking fails to track the top and bottom students (Javier67p + Kenneth1P). With DPT, as shown in Fig. 14(c), we see dramatic performance improvements in tracking the top and bottom students (Javier67p + Kenneth1P). In this example, the overall accuracy improved by 16.4%.

VI. CONCLUSION

The paper describes our efforts to build a system to assess student-participation in real-life collaborative learning videos. The real-life dataset presented many challenges that are not represented in standard occlusion datasets. We developed a new system to address the unique challenges. Specifically, we developed methods for student group detection using multiple representations, video face recognition, and dynamic participant tracking. We then document excellent performance by our proposed system that is significantly better than other methods. We verify our system on long videos of over 20 minutes and also provide visualization using student participation maps.

REFERENCES

[1] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4685–4694.
[2] X. Chen, X. Xu, Y. Yang, H. Wu, J. Tang, and J. Zhao, "Augmented ship tracking under occlusion conditions from maritime surveillance videos," *IEEE Access*, vol. 8, pp. 42884–42897, 2020.
[3] X. Dong, J. Shen, D. Yu, W. Wang, J. Liu, and H. Huang, "Occlusion-aware real-time object tracking," *IEEE Trans. Multimedia*, vol. 19, no. 4, pp. 763–771, Apr. 2017.
[4] Y. Yuan, J. Chu, L. Leng, J. Miao, and B.-G. Kim, "A scale-adaptive object-tracking algorithm with occlusion detection," *EURASIP J. Image Video Process.*, vol. 2020, no. 1, pp. 1–15, Dec. 2020.
[5] M. H. Nasseri, H. Moradi, R. Hosseini, and M. Babaee, "Simple online and real-time tracking with occlusion handling," 2021, *arXiv:2103.04147*.
[6] D. Stadler and J. Beyerer, "Improving multiple pedestrian tracking by track management and occlusion handling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10953–10962.
[7] W. Shi, "Human attention detection using AM-FM representations," Master's thesis, Dept. Elect. Comput. Eng., Univ. New Mex., Albuquerque, NM, USA, 2016.
[8] W. Shi, M. S. Pattichis, S. Celedón-Pattichis, and C. LópezLeiva, "Robust head detection in collaborative learning environments using AM-FM representations," in *Proc. IEEE Southwest Symp. Image Anal. Interpretation (SSIAI)*, Apr. 2018, pp. 1–4.
[9] W. Shi, M. S. Pattichis, S. Celedón-Pattichis, and C. LópezLeiva, "Dynamic group interactions in collaborative learning videos," in *Proc. 52nd Asilomar Conf. Signals, Syst., Comput.*, Oct. 2018, pp. 1528–1531.
[10] W. Shi, M. S. Pattichis, S. Celedón-Pattichis, and C. LópezLeiva, "Person detection in collaborative group learning environments using multiple representations," in *Proc. 55th Asilomar Conf. Signals, Syst., Comput.*, Oct. 2021, pp. 1109–1112.
[11] P. Tran, M. Pattichis, S. Celedón-Pattichis, and C. LópezLeiva, "Facial recognition in collaborative learning videos," in *Proc. Comput. Anal. Images Patterns, 19th Int. Conf. (CAIP)*. Berlin, Germany: Springer, Sep. 2021, pp. 252–261.
[12] V. Jatla, "Long-term human video activity quantification in collaborative learning environments," Ph.D. dissertation, Dept. Elect. Comput. Eng., Univ. New Mex., Albuquerque, NM, USA, 2023.
[13] V. Jatla, S. Teeparthi, M. S. Pattichis, S. Celedón-Pattichis, and C. LópezLeiva, "Long-term human video activity quantification of student participation," in *Proc. 55th Asilomar Conf. Signals, Syst., Comput.*, Oct. 2021, pp. 1132–1135.
[14] S. Teeparthi, V. Jatla, M. S. Pattichis, S. Celedón-Pattichis, and C. LópezLeiva, "Fast hand detection in collaborative learning environments," in *Proc. Int. Conf. Comput. Anal. Images Patterns*. Berlin, Germany: Springer, 2021, pp. 445–454.
[15] A. Milan, L. Leal-Taixe, I. Reid, S. Roth, and K. Schindler, "MOT16: A benchmark for multi-object tracking," 2016, *arXiv:1603.00831*.
[16] Y. Wu, J. Lim, and M.-H. Yang, "Online object tracking: A benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2411–2418.
[17] M. Kristan, J. Matas, A. Leonardis, T. Vojír, R. Pflugfelder, G. Fernández, G. Nebehay, F. Porikli, and L. Cehovin, "A novel performance evaluation methodology for single-target trackers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 11, pp. 2137–2155, Nov. 2016.
[18] H. Fan, L. Lin, F. Yang, P. Chu, G. Deng, S. Yu, H. Bai, Y. Xu, C. Liao, and H. Ling, "LaSOT: A high-quality benchmark for large-scale single object tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5369–5378.
[19] A. Dave, T. Khurana, P. Tokmakov, C. Schmid, and D. Ramanan, "TAO: A large-scale benchmark for tracking any object," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2020, pp. 436–454.
[20] S. Lan, J. Li, S. Sun, X. Lai, and W. Wang, "Robust visual object tracking with spatiotemporal regularisation and discriminative occlusion deformation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2021, pp. 1879–1883.
[21] T. P. Kuipers, D. Arya, and D. K. Gupta, "Hard occlusions in visual object tracking," in *Proc. Comput. Vis.–ECCV Workshops*, Glasgow, U.K. Berlin, Germany: Springer, Aug. 2020, pp. 299–314.
[22] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
[23] J. Guo, J. Deng, A. Lattas, and S. Zafeiriou, "Sample and computation redistribution for efficient face detection," 2021, *arXiv:2105.04714*.

WENJING SHI (Member, IEEE) received the M.S. degree in electrical engineering and the Ph.D. degree (Hons.) in computer engineering from The University of New Mexico, in 2016 and 2023, respectively.

Since 2017, she has been a Research Assistant with the Image and Video Processing and Communication Laboratory (ivPCL) and was a Student Facilitator and a Curriculum Developer for the Advancing Out-of-School Learning in Mathematics and Engineering (AOLME) Project. Her primary research interests include video analysis and the development of AM-FM representations and their applications in pose estimation, human detection, and activity detection. Furthermore, in 2021, she was a Research Assistant with the Mind Research Network (MRN), where her focus was on emotion detection. In Summer 2022, she pursued an opportunity as an Applied Scientist Intern with Amazon Web Services (AWS) to specialize in image semantic segmentation. She is currently with Amazon.

PHUONG TRAN received the M.S. degree in computer engineering from The University of New Mexico, Albuquerque, USA, in 2021. She is currently pursuing the Ph.D. degree in computer engineering.

She started as a Volunteer teaching middle school students in the AOLME project and later joined as a Research Assistant with the Image and Video Processing and Communication Laboatory (ivPCL). Her research interests include detecting and recognizing activities and humans for non-surveillance to assist educational researchers with analyzing students' performance and attendance. In addition to working in multiple National Science Foundations (NSF) projects, she has been an Instructor in the fundamental programming class with The University of New Mexico.

SYLVIA CELEDÓN-PATTICHIS is currently a Professor in bilingual/bicultural education with the Department of Curriculum and Instruction. She also holds the title of H.E. Heartfelder/The Southland Corporation Regents Chair of Human Resource Development (Fellow). She prepares elementary pre-service teachers in the bilingual/ESL cohort to teach mathematics and teaches graduate level courses in bilingual education. She taught mathematics with Rio Grande City High School, Rio Grande, TX, USA, for four years. Her research interests include studying linguistic and cultural influences on the teaching and learning of mathematics, particularly with bilingual students. She was a Co-Principal Investigator (PI) of the National Science Foundation (NSF)-funded Center for the Mathematics Education of Latinos/as (CEMELA). She is currently a lead-PI or a co-PI of three NSF-funded projects that broaden the participation of Latinx students in mathematics and computer programming in rural and urban contexts.

She serves as a national advisory board member for several NSF-funded projects and as an Editorial Board Member for the *Bilingual Research Journal*, *Journal of Latinos and Education*, and *Teachers College Record*. Her current work is a Special Issue on Teaching and Learning Mathematics and Computing in Multilingual Contexts through Teachers College Record. She co-edited three books published by the National Council of Teachers of Mathematics titled *Access and Equity: Promoting High Quality Mathematics in Grades PreK-2* and Grades 3-5 and *Beyond Good Teaching: Advancing Mathematics Education for ELLs*. She was a recipient of the Innovation in Research on Diversity in Teacher Education Award from American Educational Research Association and the 2011 Senior Scholar Reviewer Award from the National Association of Bilingual Education. She was also a recipient of the Regents Lectureship Award, the Faculty of Color Research Award, Chester C. Travelstead Endowed Faculty Award, and the Faculty of Color Mentoring Award to recognize her research, teaching, and service with The University of New Mexico.

MARIOS S. PATTICHIS (Senior Member, IEEE) received the B.Sc. degree (Hons.) in computer sciences, the B.A. degree (Hons.) in mathematics, the M.S. degree in electrical engineering, and the Ph.D. degree in computer engineering from The University of Texas at Austin, Austin, TX, USA, in 1991, 1991, 1993, and 1998, respectively.

He is currently a Professor and the Director of Online Programs with the Department of Electrical and Computer Engineering, The University of New Mexico. At UNM, he is also the Director of the Image and Video Processing and Communications Laboratory (ivPCL). His current research interests include digital image and video processing, video communications, dynamically reconfigurable hardware architectures, biomedical and space image-processing applications, and engineering education.

Dr. Pattichis was a fellow of the Center for Collaborative Research and Community Engagement, UNM College of Education, from 2019 to 2020. He was elected as a fellow of European Alliance of Medical and Biological Engineering and Science (EAMBES) for his contributions to biomedical image analysis. He was also elected as a Senior Member of the National Academy of Inventors. He was a recipient of the 2016 Lawton-Ellis and the 2004 Distinguished Teaching Awards from the Department of Electrical and Computer Engineering, UNM. For his development of the digital logic design laboratories with UNM, he was recognized by Xilinx Corporation, in 2003. He was also recognized with the UNM School of Engineering's Harrison Faculty Excellence Award, in 2006. He was the General Chair of the 2008 IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI) and the General Co-Chair of the SSIAI, in 2020 and 2024. He was also the General Chair of the 20th Conference on Computer Analysis of Images and Patterns, in 2023. He has served as a Senior Associate Editor for IEEE TRANSACTIONS ON IMAGE PROCESSING and IEEE SIGNAL PROCESSING LETTERS and an Associate Editor for IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, and *Pattern Recognition*. He has also served as a Guest Associate Editor for special issues published in IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE, *Teachers College Record*, IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS, and *Biomedical Signal Processing and Control*.