Bilingual Speech Recognition by Estimating Speaker Geometry from Video Data*

Luis Sanchez Tapia[1], Antonio Gomez[1], Mario Esparza[1], Venkatesh Jatla[1], Marios Pattichis[1], Sylvia Celedón-Pattichis[2], and Carlos LópezLeiva[2]

[1] Department of Electrical and Computer Engineering
The University of New Mexico, Albuquerque, NM, USA.
{luis2sancheztapia, agsuper, javesparza, venkatesh369, pattichi}@unm.edu
[2] Department of Language, Literacy, and Sociocultural Studies
The University of New Mexico, Albuquerque, NM, USA.
{sceledon, callopez}@unm.edu

Abstract. Speech recognition is very challenging in student learning environments that are characterized by significant cross-talk and background noise. To address this problem, we present a bilingual speech recognition system that uses an interactive video analysis system to estimate the 3D speaker geometry for realistic audio simulations. We demonstrate the use of our system in generating a complex audio dataset that contains significant cross-talk and background noise that approximate real-life classroom recordings. We then test our proposed system with real-life recordings.

In terms of the distance of the speakers from the microphone, our interactive video analysis system obtained a better average error rate of 10.83% compared to 33.12% for a baseline approach. Our proposed system gave an accuracy of 27.92% that is 1.5% better than Google Speech-to-text on the same dataset. In terms of 9 important keywords, our approach gave an average sensitivity of 38% compared to 24% for Google Speech-to-text, while both methods maintained high average specificity of 90% and 92%.

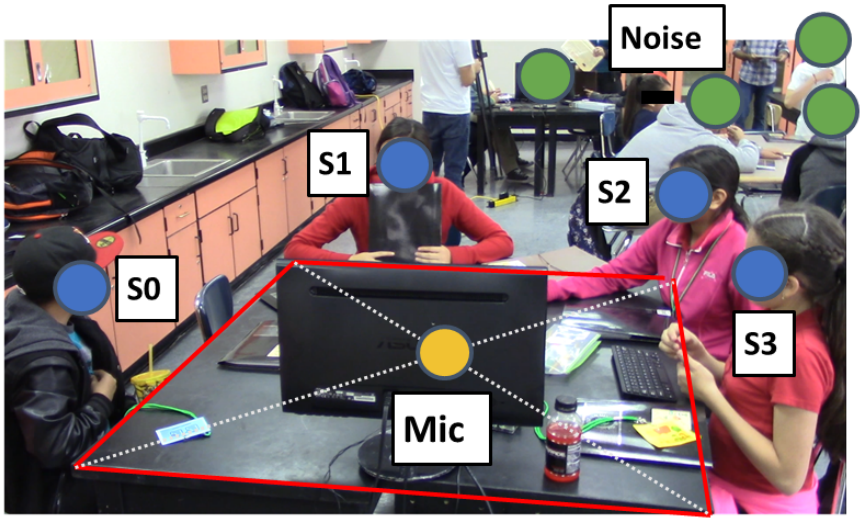
On average, sensitivity improved from 24% to 38% for our proposed approach. On the other hand, specificity remained high for both methods (90% to 92%).

Keywords: Speech Recognition · Projection Geometry · Bilingual · Video Processing.

1 Introduction

Human activity recognition can strongly benefit from the combined use of audio and video data. More recently, audio processing has been used to identify visual events [4], [9]. For our paper, we want to investigate the use of video data to

* This material is based upon work supported by the National Science Foundation under Grant No.1613637, No.1842220, and No.1949230.



S1: - "Okay, so, binary numbers only go to zero to one."

S2: - "What? Zero to one?"

S1: - "Yeah, that's all the computer knows, zero and one."

S3: - "Three... where is three? Three, where is the zero?"

Fig. 1: Example setup of a typical AOLME group interaction. Blue dots mark the speaker position and the Yellow dot is assumed to be at the center of the table (marked by red). Cross-talk is expected among speakers S0 to S3, background noise is also captured by the microphone (green dots in the back). Under the picture, we depict a sample of a transcript from the current session. Keywords can be identified like "zero", "one", "computer" and "three" .

reconstruct the speaker geometry in 3D and then use this information to develop a speaker recognition system. Our approach addresses the strong need to develop a speech recognition system that can help transcribe student conversations from video recordings of collaborative learning environments.

We present an example in Figure 1. In this example, a small group of students is sitting around the table, using the keyboard to program the Raspberry Pi. The video has been recorded as part of the Advancing Out-of-School Learning in Mathematics and Engineering (AOLME) after-school program [3]. The speech recognition problem requires that we recognize what each of the students is saying as shown in the transcription of Fig. 1. More specifically, the speaker geometry requires that we identify the 3D locations of the speakers (S_0, S_1, S_2, S_3) with respect to the omnidirectional microphone placed on the center of the table.

In Fig. 1, we also see several other speakers talking in the background (refer to green dots). The students speak in both Spanish and English.

Student speech recognition in this environment is very challenging due to cross-talk, background noise, and the use of multiple languages. Current deep learning systems are hence ineffective in such environments. To address the issue, we use the estimated 3D speaker geometry and video audio transcriptions to generate a large, acoustic model based audio dataset that can be used to train a bilingual speech recognition system for this collaborative learning environment. As we demonstrate in this paper, although we train on synthetic datasets, we are still able to match and slightly exceed state-of-the-art systems.

The current paper significantly extends our previous research on analyzing such videos. More specifically, prior research has been focused on face and back of the head detection in [16], [15], [14], [13] and [12], face recognition was also targeted in [18]. Furthermore, authors in [6] provided an early approach to context-based activity detection using deep learning. The research on video activity detection was significantly extended in [7]. The object detection system developed by [17] will be the baseline system for estimating 3D speaker geometry from the AOLME videos. For completeness, we will also explain the approach in [17] in our methodology.

The paper uses video object detection and projective geometry to locate the 3D speaker geometry from still video frames. The 3D speaker geometry is input to Pyroomacoustics ([10]) to simulate how the speakers will be recorded by the omnidirectional microphone located on the center of the table. We use the audio transcriptions with the AWS text-to-speech system to generate the ground truth audio datasets for training our speech recognition system. The proposed approach obtained a 27.92% recognition rate on Spanish words that was slightly better than Google Speech-to-text [1] at 26.12%. In addition, the Bilingual Keyword Classifier obtained an average of 38% sensitivity on Spanish Keywords.

The rest of the paper is organized as follows. We define the 3D speaker geometry problem in section 2. We then describe the underlying methods in section 3. Results are given in section 4. We then provide concluding remarks in section 5.

2 3D Speaker Geometry Estimation

We use projective geometry to estimate 3D coordinates from still image frames. Our basic assumption is to use cross-ratios along the projections of 3D lines to estimate 3D distances. We begin by assuming the basic concept and showing how to apply cross-ratios to define the problem for our videos.

We illustrate the concept of cross-ratios in Fig. 2 [2]. The basic assumption is that we know the actual physical distances between three consecutive, co-linear points A, B, C. In our example, let these distances be AB and BC. Then, to estimate the distance to another point D along the same line, we use the

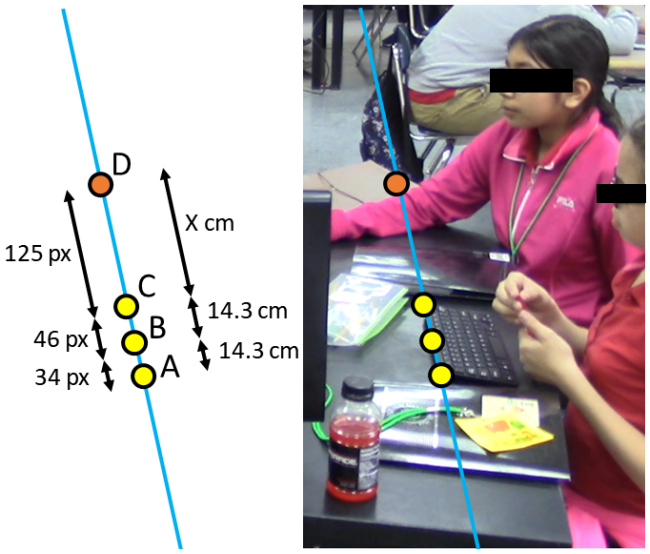


Fig. 2: Physical distance estimation using cross-ratios.

cross-ratio R defined by [2]:

$$R = \frac{AC}{CB} \Big/ \frac{AD}{DB} = \frac{AC \cdot BD}{BC \cdot AD} = \frac{(AB + BC) \cdot (BC + CD)}{BC \cdot (AB + BC + CD)} \tag{1}$$

where CD is the physical distance to be estimated. To estimate CD from equation (1), we first estimate the ratio R using pixel ratios of AD/DB. Then, we substitute the value for R and solve for CD.

To estimate the 3D locations of the speakers using cross-ratios, we will first need to estimate distances along 3D planes where our colinear points lie. In our example of Fig. 2, we assume that we know the physical dimensions of the keyboard (given as distance AC). Then, we estimate the midpoint B of the keyboard. We then assume that the keyboard is parallel to the sides of the table (1 to 2 or 3 to 4), and estimate the distance CD to the edge of the table using cross-ratios. Unfortunately, we cannot use the side of the keyboard to estimate the width of table that is depicted as a near-horizontal line in Fig. 1. This is because the keyboard side, compared against the table width is too small, and estimation can be very inaccurate.

We define all of the points that are needed to estimate the 3D speaker geometry in Fig. 3. Here, we estimate all physical distances along with the table defined by points $1, 2, 3, 4$ using cross-ratios. The basic idea is to define a 2D grid on the the table that is defined through the intersection of lines parallel to the keyboard (points $5, 6, 7, 8$) and the computer monitor ($9, 10, 11, 12, 13$). Here, we assume monitor points $8, 9, 10$ lie on the table to eliminate the need to map these points to the table surface. These lines are also assumed to be parallel to the corresponding sides of the table.

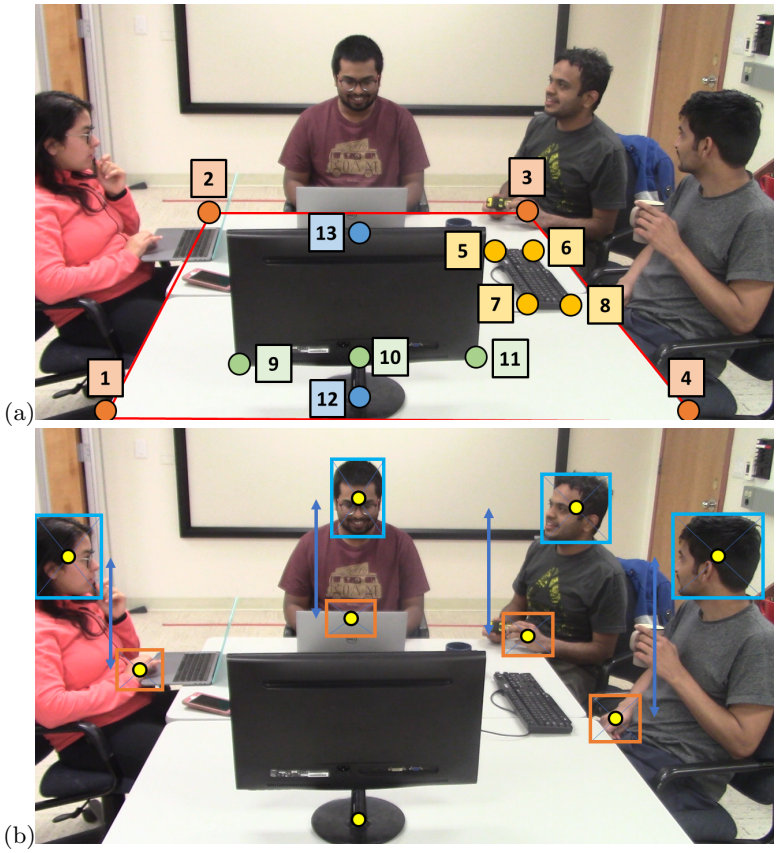


Fig. 3: Speaker geometry estimation setup.

Since the table is not always fully visible, we also extend the estimated depth of the table (points 1 to 2) by 5% to account for mild occlusion. Here, we note that the size of the table is needed because we assume that the microphone is located in the center of the table.

Similarly, we define 3D planes associated with each speaker (assumed to be about 4 inches away from the table edge) and assume that the mouths and hands lie on the same 3D plane that is orthogonal to the table. In terms of object recognition, we require hand detection, head detection as depicted in Figure 3(b).

We refer to [5] as a base for the assumptions at building the system of projections of parallel lines. We plan to test at the real scenario from AOLME videos (around 1000 hours).

3 Methodology

We summarize our methodology in Figure 4. Our 3D speaker geometry estimation requires detection of keyboard, hands and monitor. We based the detection on [17] with added post processing to detect necessary features to establish 3D geometry. We provide more details on our object detection methods in section 3.1.

Through the use of an interactive system, the users select specific frames, select the table corners and corners from the detected keyboard and monitor. Then, out system uses cross-ratios to reconstruct the 3D speaker geometry as summarized in section 2. As shown in the bottom branch of Figure 4(a), the AOLME transcripts are pre-processed to serve as input to the speech synthesis module. We then use the reconstructed 3D geometry and the synthesized dialogues to provide an acoustic-based generation of the audio dataset. We input the 3D speaker and microphone geometry, and synthesized speech into our acoustic simulation framework based on Pyroomacoustics [10]. The result is the acoustics-based simulated dataset for training our bilingual speech recognition system.

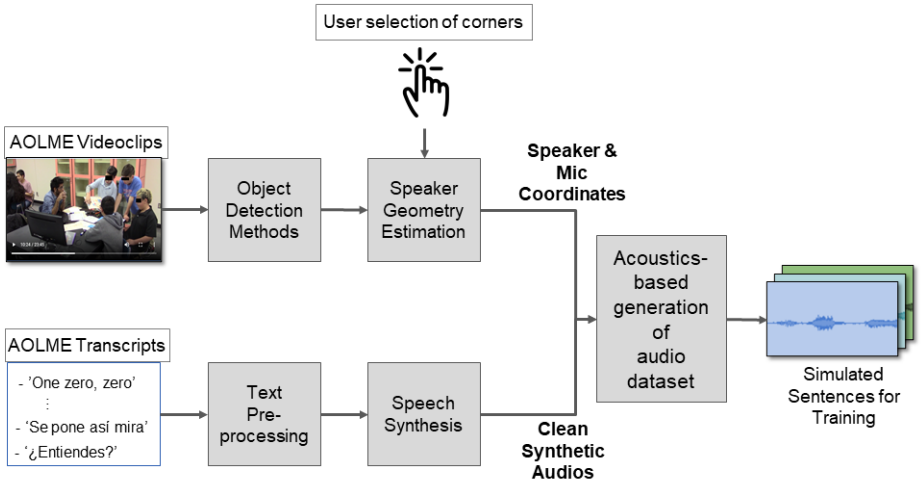
The speech recognition system is shown in Figure 4(b). The system is trained using the generated audio dataset. We provide more details of our speech recognition system in section 3.2.

3.1 Object Detection

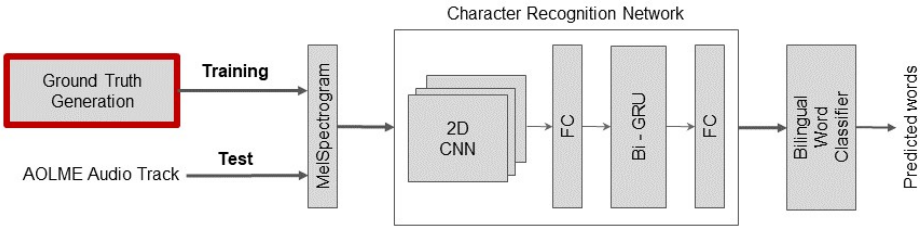
As shown in Fig. 3, we require detection of the keyboard and monitor in order to estimate the location of the speakers with respect to the table. Furthermore, to estimate the 3D locations of the speaker's mouths, we also assume that their hands and mouths are on the same 3D plane and further require hand and head detection. Here, we are only interested in hands that are located near the table as shown in Fig. 3.

We next summarize the methods that we will use to detect each object. For head detection, we use the latest version of YOLO [8] pre-trained on the crowd human data set for head detection [11]. To restrict head detection within the current student group, we use a minimum area threshold that successfully rejects smaller faces of people outside the group. For detecting hands, monitors, and keyboards, we use faster R-CNN pre-trained on the COCO dataset. The results of faster R-CNN are post-processed using clustering, time-projections (adding detections through time), and small area removal to remove distant hands (see [17] for details). Among the hand detections, we then manually select hands that lie on the table. Furthermore, we manually select the edges of the Table, the monitor, and the keyboard.

We assume that we can learn the scales, number of pixels per inch for each speaker using manual measurements during training. Later, we will look at estimating the scales for each image. Here, we note that our assumption is very restrictive. It does not account for strong scale variations when the speakers move to new positions not reflected in the training set.



(a) Acoustics-based dataset generator based on 3D speaker geometry.



(b) Speech recognition system.

Fig. 4: Bilingual speech recognition system using 3D speaker geometry estimated from the video dataset.

3.2 Speech recognition system

We summarize the speech recognition system in Fig. 4(b). The acoustic-based generated dataset is used to train a phoneme-based recognition network composed of a 2D CNN (a single layer of 8 filters of size 3×3 with stride=2) processing Mel-spectrograms, a two-layer bi-directional GRUs with 64 units per layer, and a fully connected layer with an output for each phoneme. The system generates a sequence of phonemes characters that are post-processed by a bilingual word classifier based on minimum distance.

4 Results

We first summarize results from 3D speaker geometry estimation using a baseline approach and our proposed methods. We then summarize our results for speech recognition system using the 3D speaker geometry.

Table 1: Results for 3D speaker geometry estimation. The error is given as a percentage of the distance to the microphone. All distances are given in inches.

Speaker	**Ground Truth**	**Our Method**		**Baseline**	
		Estimation	Error	Estimation	Error
S0	36.70	34.16	6.92 %	19.74	46.21 %
S1	35.59	41.27	15.96 %	24.32	31.67 %
S2	42.12	43.88	4.18 %	27.79	34.02 %
S3	34.99	29.29	16.29 %	27.79	20.58 %
Average	37.35	37.15	10.84 %	24.91	33.12 %

We define a baseline approach that does not require projective geometry or any object detection method. Assuming the keyboard and table corners are given, we assume that speakers sit around the table, equidistant from each other.

Our proposed approach performed significantly better. We present a summary of our estimates for Fig. 3 in Table 1. Our error ranges from 7% to 16%. The largest source of error comes from our estimation of the scale for each speaker (number of pixels per inch). As mentioned earlier, in future work, we will work on estimating the scale directly from each image. Overall, our interactive system gave a reduced error of 10.84% compared to 33.12% for the baseline method. In terms of the AOLME dataset, we present an example of object detection in Fig. 5. Overall we note that our proposed approach required the combination of different object detections from different video frames to establish the 3D speaker geometry.

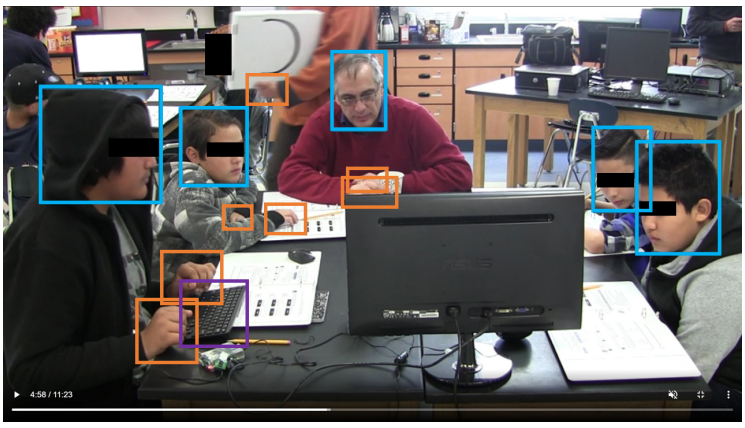


Fig. 5: Object detection for 3D speaker geometry estimation. We use blue bounding boxes for head detection, orange bounding boxes for hand detection, and purple bounding boxes for keyboard detection.

Table 2: Keyword recognition results. Here, we note that our system does not recognize accents.

	Our system		**Google Speech-to-Text**	
Keywords	**Sensitivity**	**Specificity**	**Sensitivity**	**Specificity**
uno	0.50	0.95	0.13	1.00
dos	0.24	0.91	0.06	1.00
tres	0.63	0.92	0.00	1.00
cuatro	0.30	0.99	0.00	1.00
cinco	0.25	0.99	0.23	1.00
cero	0.36	0.93	0.00	1.00
computadora	0.25	0.99	0.25	1.00
numero	0.27	0.97	0.45	1.00
Others	0.65	0.67	1.00	0.13
Average	0.38	0.92	0.24	0.90

The output of 3D speaker geometry system is the complex simulated audio dataset, used to train the speech recognition system. The training dataset was generated using audio transcriptions of 720 minutes extracted from 54 video sessions, and a typical AOLME 3D speaker geometry. For testing, we selected 517 sentences from unseen video sessions. We then assessed the character error rate for recognizing the 517 sentences. For this test, our proposed approach gave an accuracy of 27.92 % compared to 26.12% by Google speech-to-text.

We also present comparative results for the recognition of 9 Spanish keywords that were used in the number representations lessons. We summarize our results in terms of sensitivity and specificity as given in Table 2. From the results, it is clear that Google Speech-to-text fails to detect any instances of tres, cuatro, and cero. Overall, Google Speech-to-text is insensitive to the target keywords, it is prone to discard noisy samples as 'Others'. By comparison, our proposed method is much better at detecting our targeted keywords because it will try to classify even the noisy samples. On average, sensitivity improved from 24% to 38% for our proposed approach. On the other hand, specificity remained high for both methods (90% to 92%).

Our proposed approach produces more false positives and fewer false negatives than Google Speech-to-text. Hence, in terms of using our method, we note that the users would have to reject our false positive detections. On the other hand, Google Speech-to-text requires noise-free examples and fails to detect important AOLME type keywords (e.g., tres, cuatro, and cero).

5 Conclusions and Future Work

We presented an interactive system for estimating 3D speaker geometries from a single-camera video recording. We then used a typical 3D speaker geometry based on AOLME videos to generate a complex, acoustics-based, simulated

dataset based on 11.66 hours of audio dataset. Then, when tested on actual audio datasets, the proposed system slightly outperformed Google Speech-to-text. Ultimately, the detection of meaningful keywords can be used by educational researchers to identify moments of interest for further analysis.

For future work, we are currently developing multi-objective optimization methods for improving our sensitivity while maintaining high specificity.

References

1. Google cloud speech-to-text api, https://cloud.google.com/speech-to-text
2. Brannan, D.A., Esplen, M.F., Gray, J.J.: Geometry. Cambridge University Press, 2 edn. (2011). https://doi.org/10.1017/CBO9781139003001
3. Celedón-Pattichis, S., LópezLeiva, C.A., Pattichis, M.S., Llamocca, D.: An interdisciplinary collaboration between computer engineering and mathematics/bilingual education to develop a curriculum for underrepresented middle school students. Cultural Studies of Science Education **8**(4), 873–887 (Dec 2013), https://doi.org/10.1007/s11422-013-9516-5"
4. Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K., Hassidim, A., Freeman, W.T., Rubinstein, M.: Looking to listen at the cocktail party. ACM Transactions on Graphics (2018)
5. Hartley, R.I., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge University Press, ISBN: 0521540518, second edn. (2004)
6. Jacoby, A.R., Pattichis, M.S., Celedón-Pattichis, S., LópezLeiva, C.: Context-sensitive human activity classification in collaborative learning environments. In: 2018 IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI). pp. 1–4 (April 2018). https://doi.org/10.1109/SSIAI.2018.8470331
7. Jatla, V., T.S.P.M.C.P.S., LópezLeiva, C.: Long-term human video activity quantification of student participation. Asilomar Conference on Signals, Systems, and Computers, invited. (2021)
8. Jocher, G., Stoken, A., Borovec, J., NanoCode012, Chaurasia, A., TaoXie, Changyu, L., V, A., Laughing, tkianai, yxNONG, Hogan, A., lorenzomammana, AlexWang1900, Hajek, J., Diaconu, L., Marc, Kwon, Y., oleg, wanghaoyang0106, Defretin, Y., Lohia, A., ml5ah, Milanko, B., Fineran, B., Khromov, D., Yiwei, D., Doug, Durgesh, Ingham, F.: ultralytics/yolov5: v5.0 - YOLOv5-P6 1280 models, AWS, Supervise.ly and YouTube integrations (Apr 2021). https://doi.org/10.5281/zenodo.4679653, https://doi.org/10.5281/zenodo.4679653
9. Owens, A., Efros, A.A.: Audio-visual scene analysis with self-supervised multisensory features. CoRR (2018)
10. Scheibler, R., Bezzam, E., Dokmanic, I.: Pyroomacoustics: A python package for audio room simulations and array processing algorithms. CoRR **abs/1710.04196** (2017), http://arxiv.org/abs/1710.04196
11. Shao, S., Zhao, Z., Li, B., Xiao, T., Yu, G., Zhang, X., Sun, J.: Crowdhuman: A benchmark for detecting human in a crowd. arXiv preprint arXiv:1805.00123 (2018)
12. Shi, W., P.M.C.P.S., LópezLeiva, C.: Person detection in collaborative group learning environments using multiple representations. Asilomar Conference on Signals, Systems, and Computers, accepted. (2021)

13. Shi, W., P.M.C.P.S., LópezLeiva, C.: Talking detection in collaborative learning environments. The 19th International Conference on Computer Analysis of Images and Patterns (CAIP), accepted. (2021)
14. Shi, W., Pattichis, M.S., Celedón-Pattichis, S., LópezLeiva, C.: Dynamic group interactions in collaborative learning videos. In: 2018 52nd Asilomar Conference on Signals, Systems, and Computers. pp. 1528–1531 (Oct 2018)
15. Shi, W., Pattichis, M.S., Celedón-Pattichis, S., LópezLeiva, C.: Robust head detection in collaborative learning environments using am-fm representations. In: 2018 IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI). pp. 1–4 (April 2018). https://doi.org/10.1109/SSIAI.2018.8470355
16. SHI, W.: Human Attention Detection Using AM-FM Representations. Master's thesis, University of New Mexico (2016)
17. Teeparthi, S., J.V.P.M.C.P.S., LópezLeiva, C.: Fast hand detection in collaborative learning environments. The 19th International Conference on Computer Analysis of Images and Patterns (CAIP), accepted. (2021)
18. Tran, P., P.M.C.P.S., LópezLeiva, C.: Facial recognition in collaborative learning videos. The 19th International Conference on Computer Analysis of Images and Patterns (CAIP), accepted. (2021)