Tracking Classroom Movement Patterns with Person Re-ID

Xinlu He
Worcester Polytechnic Institute
xhe4@wpi.edu

Jiani Wang
Worcester Polytechnic Institute
jwang21@wpi.edu

Viet Anh Trinh
Worcester Polytechnic Institute
vtrinh@wpi.edu

Andrew McReynolds
Worcester Polytechnic Institute
aamcreynolds@wpi.edu

Jacob Whitehill
Worcester Polytechnic Institute
jrwhitehill@wpi.edu

ABSTRACT

With the goal of supporting real-time AI-based agents to facilitate student collaboration, as well as to enable educational data-mining of group discussions, multimodal classroom analytics, and social network analysis, we investigate how to identify who-is-where-when in classroom videos. We take a person re-identification (re-ID) approach, and explore different methods of improving re-ID accuracy in the challenging environments of school classrooms. Our results on a multi-grade classroom (MGC) dataset suggest that (1) fine-tuning off-the-shelf person re-ID models (e.g AGW [26]) can deliver sizable accuracy gains (increase from 70.4% to 76.7%); (2) clustering, rather than nearest-neighbor identification, can also improve accuracy (from 76.7% to 79.4%); and (3) there is a strong benefit to re-ID accuracy in obtaining multiple enrollment images from each student.

Keywords

Group dynamics, multimodal analytics, person re-ID

1. INTRODUCTION

Research on multimodal learning analytics [15], social network analysis of classrooms [2], and classroom observation [17] relies on knowing who is where when in the classroom. This knowledge facilitates research into interactions between students and teachers. It also provides sensor input for AI-based educational agents to track contributions and collaborations. Accurate identification and location in classroom videos support research efforts [10, 1] to evaluate how a teacher distributes their attention to their students, and in real-time interactive settings it empowers AI-based educational agents to offer personalized instruction.

Classroom position tracking can be implemented either via wearable positioning sensors or with a camera and computer vision techniques. Wearable sensors offer precise (x, y) coordinate tracking but face scalability and deployability issues and incur high costs. Computer vision, on the other hand, uses cameras to capture positional data without physical tags, providing a less invasive and more scalable option. However, it faces challenges with data continuity due to occlusion and complex visual properties of classroom images.

Within the computer vision approach, there are two main techniques: tracking and person re-identification (re-ID). Tracking harnesses temporal smoothness in the position of each person over time, updating the estimated position of each person in the classroom in each frame. While intuitive, it can suffer in performance when a person suddenly "disappears" due to occlusion or leaving the field-of-view. Person re-identification (re-ID), a computer vision technique, treats each frame in the video independently; in each frame, a *query* image is matched against a *gallery* of reference images ("enrollments"). It is more robust to sudden occlusion/disappearance. Both tracking and re-ID techniques face challenges such as occlusions, lighting variations, and camera angles, and both fields have benefited from deep-learning models and feature extraction techniques to improve accuracy. Ultimately, the two approaches are likely complementary. Our paper focuses on the re-ID method.

Traditionally, re-ID has been used in public surveillance settings. Adapting person re-identification for classroom environments requires overcoming several challenges: (1) Privacy concerns limit the availability of classroom videos for public use. (2) Classrooms often have high levels of occlusion, with students blocking each other's view due to their movements and the layout. (3) Student postures vary greatly, ranging from sitting to stretching or moving around, complicating visibility of identifying features.

To address these challenges, we have developed a person re-ID system specifically with classrooms in mind. As the first step, the *enrollment* process involves capturing a short video of each individual to establish their identity label. These enrollment images constitute the *gallery*. Then, a person detector extracts person bounding boxes from images, which are processed into *embedding vectors* that encapsulate identification information – embeddings from the same person should be close together in the embedding space, and those from different people should be far apart. Finally, the embeddings from the gallery are compared with those from the queries, using identity retrieval to determine identification.

In our study, we first annotated a Multi-Grade Classroom (MGC) dataset [6] for person re-ID in classroom settings.

X. He, J. Wang, V. A. Trinh, A. McReynolds, and J. Whitehill. Tracking classroom movement patterns with person re-id. In B. Paaßen and C. D. Epp, editors, *Proceedings of the 17th International Conference on Educational Data Mining*, pages 679–685, Atlanta, Georgia, USA, July 2024. International Educational Data Mining Society.

© 2024 Copyright is held by the author(s). This work is distributed under the Creative Commons Attribution NonCommercial NoDerivatives 4.0 International (CC BY-NC-ND 4.0) license.
https://doi.org/10.5281/zenodo.12729916

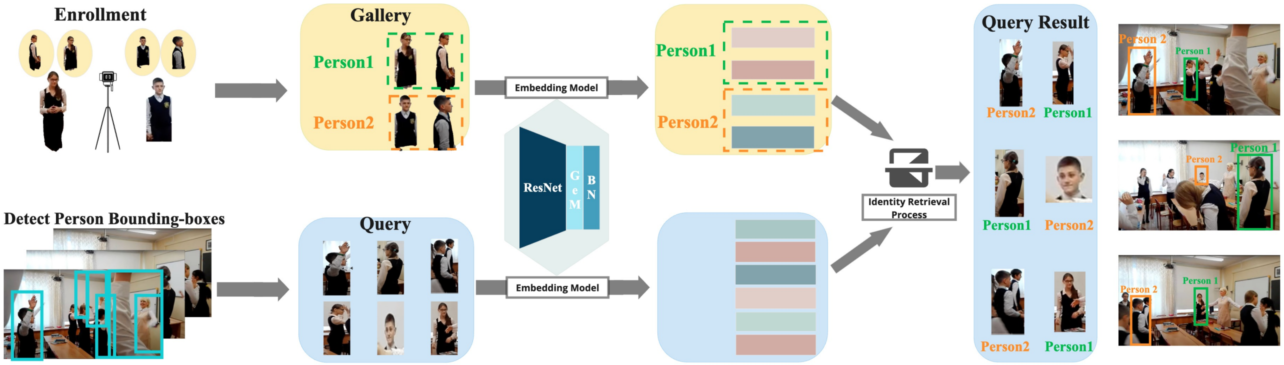

Figure 1: Person Re-Identification (ReID) applied to classroom videos involves: (a) enrolling students with identity labels to form the gallery; (b) detecting person bounding boxes in each video frame; (b) normalizing the query bounding boxes to increase similarity to the gallery; (c) extracting identity embeddings for each bounding box via the Embedding Model; and (d) matching the queries to gallery identities based on their embeddings. Classroom images are from YouTube [16].

For the embedding model, we fine-tuned the 2022 AGW model [26] and experimented with different input normalization strategies to address classroom-specific occlusions, and also explored the effect of enrolling each student multiple times on the retrieval accuracy. We also tested whether clustering methods such as k-means and spectral clustering, with different initializations and constraints, can improve the identity retrieval process.

Contributions: (1) To the best of our knowledge, we are the first to explore a re-ID workflow for classroom environments. (2) We show that fine-tuning a state-of-the-art re-ID system on classroom data can significantly improve accuracy. (3) We compared nearest neighbor retrieval to clustering methods and found that clustering, with applied with appropriate structural constraints, can yield higher accuracy.

2. RELATED WORK

In classroom person tracking, spatial analysis guides efforts, supported by tracking technology. We also review person reidentification literature to develop a ReID system for classrooms, aiming to identify individuals and analyze their positions and behaviors to improve teaching and learning.

2.1 Classroom Tracking

Spatial Analysis In the learning science community, sociospatial behavior studies have gained traction. Research [4] [14] shows that analyzing the distances and interactions between students and teachers indicates social dynamics and learning engagement. Individual work improves technical skills, while collaboration encourages sustained participation in makerspaces. Additionally, identifying various patterns of collaborative interactions enables instructors to detect struggling students early, facilitating a more tailored and effective support strategy. Lim et al. [9] introduces *Spatial Pedagogy* to understand the significance of classroom spaces and how teachers use them to adjust teaching strategies. These studies emphasize automating spatial analysis to provide teachers with accurate feedback on classroom dynamics, based on identifying *who is where when* to enhance teaching and learning.

Tracking Technology Some studies have employed wearable devices to track students' locations. Although costly, this method does not require matching identities to locations. Sensei system [18] used shoe sensors for social interactions and learning. Tatwah Mango BLE-WB200 wristband trackers [23] monitored social interactions in large spaces. Computer vision, UWB, and thermal sensing also require identity matching. UWB and thermal sensing are expensive and need multiple devices per classroom. Computer vision is increasingly popular due to its simplicity and rapid AI advances. Paul Hur et al. [10] tracked students via video postprocessing, while Chng et al. [4] used cameras and OpenPose to monitor x-y coordinates. However, these methods rely on pre-trained models without classroom-specific fine-tuning, which is our focus.

2.2 Person Re-id

Dataset Person re-identification (re-ID) in classrooms is a novel challenge, as publicly available datasets for this context are scarce. Bo Sun et al. [20] provided a dataset from classroom videos, but it focuses on student behaviors rather than identity labels and is unavailable due to privacy concerns. We are not aware of any public classroom dataset with Re-Id labels and thus collected our own (Section 3.1).

Embedding Model In person re-ID, embedding models are essential for identity representation. Global models extract features from the entire person bounding box image. Zheng et al. [29] use an ID-discriminative Embedding (IDE) CNN model, while Luo et al. [13] set a strong baseline by incorporating various training heuristics (e.g., random erasing, warmup learning). Attention mechanisms [24, 11, 3] have also improved Re-ID. Processing local features such as clothing type, color, and accessories [12, 19, 21] may improve alignment between query and gallery images, but they require accurate detection of body parts. In our work, we adopt the AGW global model [26] as our baseline.

Identity Retrieval In person re-ID, the system retrieves the item in the gallery most similar to the query to identify an individual. Zhun Zhong et al. [30] improved similarity ranking through re-ranking, and Mang Ye et al. [25] used

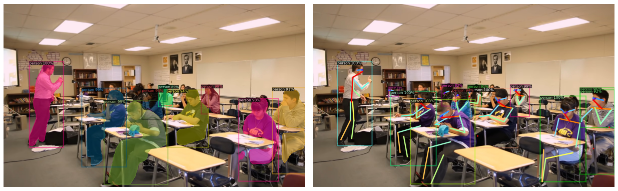

Figure 2: Detectron2 Results from Mask(left) and Keypoint(right) Modes, with image sourced from YouTube[5]

Table 1: Statistical for Bounding Boxes

	#Videos	#Annotations	#Tracks
Manual	74	56,952	915
Detectron Mask	74	75,470	1,058
Detectron Keypoint	74	84,795	1,194
Matched Data	74	39,563	895

a ranking aggregation algorithm that considers similarity and dissimilarity. In our classroom scenario, with limited labeled data, we explored clustering along with retrieval. Yunpeng Zhai et al. [27] developed AD-Cluster to boost discrimination by estimating and enhancing person clusters using unlabeled target domain samples.

3. CLASSROOM PERSON RE-ID SYSTEM

Here we describe how the person re-ID system for classroom analysis was built, including the dataset it was trained and tested on, as well as its algorithmic components.

3.1 Data and Annotation

We used the Multi-Grade Classroom dataset from [6], consisting of 74 videos (15-23 minutes each) from kindergarten to middle school classrooms in a Midwestern U.S. state. These videos vary in camera type, placement, and lighting. Teachers typically face the camera, while students face away or sideways, limiting face visibility. Each video was split into frames, and one frame per 10 sec was processed by Detectron2 [22] to detect the people in the scene. Then, annotators manually labeled the identities of persons in each frame, as well as their role (student vs. teacher). In order to investigate possible re-ID accuracy bias, they also labeled each person's skin tone (1-6) on the Fitzpatrick scale [8]. In total, 915 persons and 56,952 bounding boxes were labeled using the CVAT system (`https://cvat.ai/`) by four trained undergraduates. The sequence of bounding boxes over a video for a given person is called a *track*. The first author of this paper independently checked a sample of 50 labeled tracks for reliability, finding a 1% label error rate.

Detectron2 has mask and keypoint modes. Mask mode creates bounding boxes and contours, while keypoint mode detects 17 keypoints (nose, eyes, shoulders, etc.) with confidence scores. Annotations, keypoint data, and mask mode results were matched (with a Intersection Over Union threshold of 0.8) so that each sample has labels, mask details, and keypoint data. See Table 2.

3.2 Embedding Model

The core of a re-ID system is its embedding model, which maps bounding boxes to an embedding space, clustering the same person's embeddings while separating different individuals. Our classroom re-ID system uses the AGW model [26]. We either used a pretrained AGW model or fine-tuned it on a subset of our dataset.

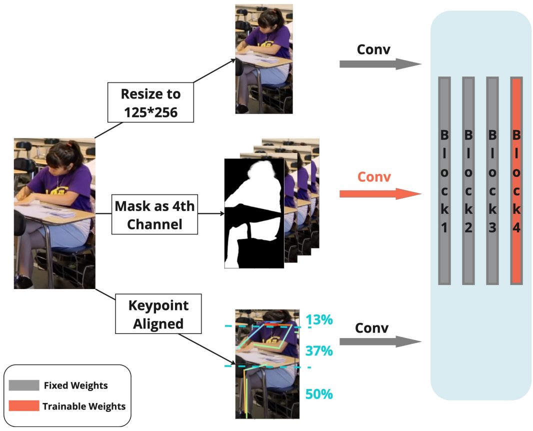

Figure 3: Different normalization methods prior to embedding: standard alignment to canonical size (top), background masking (middle), and keypoint vertical alignment (bottom).

.

We examine two important aspects of the embedding model: (a) transfer learning from the pre-trained AGW model, and (b) normalizing the input bounding box to more closely match the gallery images (see Figure 3).

3.2.1 Transfer Learning from AGW

We apply transfer learning from the pre-trained AGW model [26], which was built on a ResNet-50 backbone, and uses generalized-mean pooling, batch normalization, and a multi-class classifier for better feature extraction and classification. Training relies on three loss functions: *Cross-Entropy Loss*($\mathcal{L}_{ID}$) reduces the gap between the predicted and actual identity labels to boost classification accuracy. *Weighted Regularization Triplet Loss* ($\mathcal{L}_{\text{Triplet}}$) brings images of the same identity closer in the embedding space than those of different identities, enhancing discrimination. *Center Loss* ($\mathcal{L}_C$) minimizes the distance between class features and their centroid to increase intra-class compactness. The overall loss is formulated as: $\mathcal{L} = \mathcal{L}_{ID} + \mathcal{L}_{Triplet} + \beta\mathcal{L}_C$, β represents the balance weight for the Center Loss, set to 0.0005 in the baseline model. The embedding post-batch normalization and pre-classifier is used for identity retrieval.

3.2.2 Normalization with Background Masking

In classrooms, individuals often sit close together, leading to frequent overlapping in images. In our dataset of 41,904 annotations, over 85% (35,730) of the bounding boxes overlapped the bounding boxes of other people. Such overlaps can confuse the embedding model and lower re-ID accuracy. To tackle this problem, we incorporated mask information from Detectron into the AGW model. An additional input channel was added while maintaining the model's original structure. Fine-tuning the model follows standard procedures except for the new input channel, which is initialized with a normalized binary mask (mean=0.5, std=0.5). The first three channels use pre-trained weights, while the new fourth channel starts with random values. The AGW model is then fine-tuned with the additional channel included.

3.2.3 Normalization with Keypoint Vertical Alignment

Standard Re-ID datasets such as Market1501 [28] typically feature subjects walking or standing with a 2:1 aspect ratio. Classroom settings, however, are more diverse: students stand, sit on chairs or the floor, move frequently, and are often partially obscured, leading to aspect ratios up to 11. To address this, we used keypoint information for alignment. Detectron's 17 keypoint detections allowed images to be segmented into vertical head, body, and leg sections, resized to 128x256 pixels with a 13:37:50 distribution. This segmentation may provide better alignment despite varied classroom postures.

3.3 Identity Retrieval

After obtaining embeddings from the embedding model, our next step is to assign those derived from the target video to those from the gallery. We employ two alternative matching techniques: classification and clustering.

With **verification**, we independently match each query embedding to the nearest (i.e., smallest L_2 distance) gallery embedding. With **clustering**, we cluster the set of all embeddings extracted from a video all at once, followed by the Hungarian algorithm to map clusters to identities so as to maximize the number of enrollment embeddings assigned to their correct identities. The number of clusters is set to the number of individuals in the gallery, and the enrollment embeddings can optionally be used to initialize the cluster centroids. We explored k-means and spectral clustering.

Constrained Spectral Clustering: In our setting, we have the constraint that no two bounding boxes in the same video frame can represent the same identity. This constraint can be harnessed during clustering to potentially improve accuracy. While constrained k-means is known to be NP-complete [7], spectral clustering can readily incorporate constraints into its affinity matrix A, which expresses how similar two inputs x_i and x_j are to each other. In particular, we set the affinityA_{ij} between embeddings x_i and x_j to be:

$$A_{ij} = \begin{cases} \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right), & \text{if } f(i) \neq f(j), \\ \min(\alpha, \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)), & \text{otherwise.} \end{cases}$$

where $f(i)$ and $f(j)$ are the video frame indices of x_i and x_j. This ensures that the maximum affinity of two distinct embeddings from the same frame is never more than α, which is a hyperparameter.

4. EXPERIMENTS

We conducted experiments on fine-tuning the embedding model and normalizing the input bounding boxes (see Section 3.2), as well as method used for identity retrieval (Section 3.3). We also explored how the re-ID accuracy varied as a function of how much each person moved within their track (Position Variance). Finally, we assessed how accuracy may be improved by enrolling each person in the classroom multiple times (1, 3, or 5 gallery images per person). For the experiments, the dataset was divided into training, testing, and validation sets in a 6:2:2 ratio across all videos, tracks, and figures. The model was fine-tuned on training data, with optimal hyperparameters found via the validation set. Test data were then used to assess accuracy.

Table 2: Comparison of Embedding Models & Normalizations

Model	Normalization	R-1 Acc. (%)	mAP(%)
Pre-trained	Standard	70.4	78.6
Fine-tuned	Standard	**76.7**	**83.7**
Fine-tuned	Masked	72.1	80.4
Fine-tuned	Keypoint	73.7	82.0

Evaluation: We assess person re-ID using Rank-1 Accuracy and Mean Average Precision (mAP). Rank-1 Accuracy measures the probability of assigning the correct identity, checking if it's among the top gallery results. mAP calculates the model's retrieval performance by averaging the precision scores for all queries, representing the area under the precision-recall curve. This provides a comprehensive evaluation of how well the model ranks relevant samples.

4.1 Embedding Model Fine-Tuning Results

Table 2 shows normal fine-tuning achieved the best R-1 accuracy and mAP (using the verification method for identity retrieval). The fine-tuned AGW models always worked better than the pretrained model, no matter which normalization method was used. Neither the background masking nor the vertical keypoint alignment benefited over a standard warping of the bounding box to a canonical size. One possible explanation is that background image features in the classroom that are behind each person's body but consistently visible in the classroom may actually hold information that is useful for identity matching.

We also explored the correlation between *Position Variance* and Accuracy. Position Variance is defined as the average squared distance, over all frames t in which a person appears, of their location (x_t, y_t) to their mean location $(\bar{x}, \bar{y})$. Higher Position Variance means more student movement and hence less consistent background information for that person.

Figure 4 shows that R-1 accuracy drops precipitously with Positional Variance, suggesting that the background pixels of each bounding box may be driving accuracy rather than the person's appearanc per se. Also, while the standard normalization method worked best on average, the background masking improves and eventually outperforms the standard method as Positional Variance increases. Finally, the figure also shows how using multiple gallery images per person (5 instead of 1) can improve accuracy significantly.

4.2 Identity Retrieval Results

In this experiment, we compared verification to k-means clustering (with and without gallery centroid initialization), and to spectral k-means clustering (with and without constraints). We use the fine-tuned AGW with standard normalization. Table 3 presents the results. Within each gallery size (1, 3, or 5), gallery centroid initialization outperformed random initialization, and constrained spectral clustering worked better than unconstrained. This underscores the

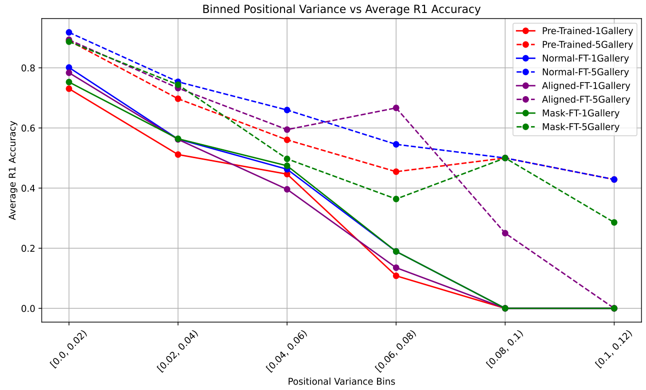

Figure 4: Average R1 accuracy across different positional variance bins for four fine-tuning strategies: Pre-trained, Normal-FT, Mask-FT, and Aligned-FT. The analysis demonstrates the impact of variance on model performance.

Table 3: Identity Retrieval Methods and Their Accuracy

Identity Retrieval Method	1-Pic	3-Pic	5-Pic
Verification	76.7	86.4	**90.2**
k-means: Random Initialization	69.2	75.3	77.0
k-means: Gallery Initialization	77.4	**86.5**	88.0
Spectral Clustering: Unconstrained	74.0	79.9	82.0
Spectral Clustering: Constrained	**79.4**	84.3	84.3

value of gallery data for clustering and ensuring identity uniqueness per frame. More gallery images increase accuracy across all methods, particularly for smaller galleries. Verification benefits more from larger galleries, while clustering methods see smaller gains.

5. SKIN TONE BIAS OF PERSON RE-ID

We evaluated how the fine-tuned AGW person-reid system might have biases in retrieval accuracy based on skin tone (1 to 6 on the Fitzpatrick scale, where 1 is lightest and 6 is darkest). Naively, one might compare R-1 accuracies across the different skin tones. However, this overlooks that some individuals might be easier to recognize exactly because of their skin tone, despite being often confused by the re-ID system with other people of a similar tone. Hence, we instead examined bias by computing distances between embeddings of persons within the same skin tone, and then calculating the probability of correctly distinguishing one person from others with the same skin tone based on these distances. For each Fitzpatrick skin tone (1 to 6), the probabilities were 0.99, 0.83, 0.95, 0.87, 0.88, and 0.99. These probabilities are not stat. sig. different ($\chi^2(5) = 5.96, p = 0.31$).

6. SOCIAL DYNAMICS VISUALIZATION

Here we illustrate how person re-ID can provide a glimpse into the movement patterns of students and teachers in a classroom. Figure 5 in appendix displays two classrooms where the AGW model was not trained or fine-tuned. Despite occasional prediction errors, the model's overall predicted paths generally align with human annotations. In the high school classroom (top half), the teacher moves around and interacts with students, while they mostly remain seated. In the elementary classroom (bottom half), students move more freely on the floor, while the teacher stays seated and reads a story. There are several obvious caveats, e.g., the camera must be stationary in order for the visualization to be meaningful, and no 2-d projection of 3-d position data can perfectly represent the classroom interactions. Nonetheless, by visualizing the trajectories of the students and teachers in the room, some information can be gleaned that may be useful feedback to teachers.

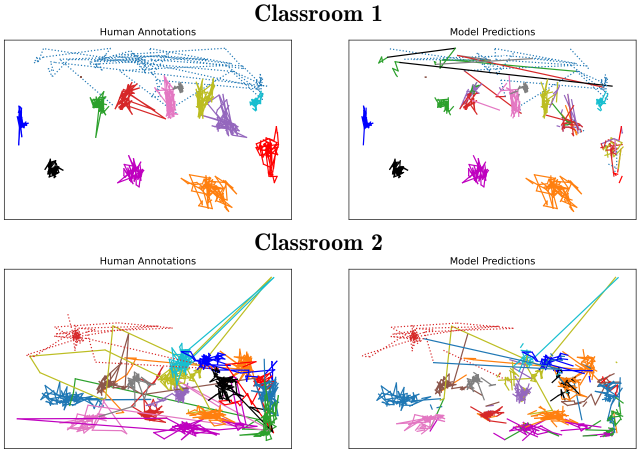

Figure 5: The movement patterns of the students and teachers in two classroom (top and bottom), for which both human annotations (left) and model predictions (right) are shown. Each colored line represents a different person, and the dotted line represents the teacher. Best viewed in color.

7. CONCLUSION

In this paper, we adapted a person re-identification (re-ID) system for the unique demands of classroom environments, and explored the accuracy of design decisions on the Multi Grade Classroom (MGC) video dataset. Our main conclusions are as follows: (1) Fine-tuning the pre-trained re-ID models was crucial, markedly boosting the system's ability to cope with classroom situations. (2) Implementing clustering techniques proved beneficial for small gallery sizes, particularly for 1 and 3-picture galleries. (3) Using more than one picture per gallery significantly improved re-ID accuracy, especially in scenarios with high positional variance among students. The long-term goal of our work is to develope-ID technology for real-world educational settings to enable educational research, and also to enable AI-based agents to facilitate student collaboration.

Acknowledgement

This research was supported by the NSF National AI Institute for Student-AI Teaming (iSAT) under grant DRL #2019805, and also from an NSF CAREER grant #2046505. The opinions expressed are those of the authors and do not represent views of the NSF.

References

[1] K. Ahuja, D. Kim, F. Xhakaj, V. Varga, A. Xie, S. Zhang, J. E. Townsend, C. Harrison, A. Ogan, and

Y. Agarwal. Edusense: Practical classroom sensing at scale. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(3):1–26, 2019.

[2] M. Bowman, S. K. Debray, and L. L. Peterson. Reasoning about naming systems. *ACM Trans. Program. Lang. Syst.*, 15(5):795–825, November 1993.

[3] G. Chen, C. Lin, L. Ren, J. Lu, and J. Zhou. Self-critical attention learning for person re-identification. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9636–9645, 2019.

[4] E. Chng, M. R. Seyam, W. Yao, and B. Schneider. Using motion sensors to understand collaborative interactions in digital fabrication labs. In I. I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, and E. Millán, editors, *Artificial Intelligence in Education*, pages 118–128, Cham, 2020. Springer International Publishing.

[5] Classroom management - Week 1, Day 1. Classroom management - Week 1, Day 1, 2021. Accessed: 2023-02-21.

[6] Z. Dai, A. McReynolds, and J. Whitehill. In search of negative moments: Multi-modal analysis of teacher negativity in classroom observation videos. *International Educational Data Mining Society*, 2023.

[7] I. Davidson and S. Ravi. Clustering with constraints: Feasibility issues and the k-means algorithm. In *Proceedings of the 2005 SIAM international conference on data mining*, pages 138–149. SIAM, 2005.

[8] T. B. Fitzpatrick. The validity and practicality of sun-reactive skin types i through vi. *Archives of dermatology*, 124(6):869–871, 1988.

[9] K. O. F.V. Lim and A. Podlasov. Spatial pedagogy: mapping meanings in the use of classroom space. *Cambridge Journal of Education*, 42(2):235–251, 2012.

[10] P. Hur and N. Bosch. Tracking individuals in classroom videos via post-processing openpose data. In *LAK22: 12th International Learning Analytics and Knowledge Conference*, LAK22, page 465–471, New York, NY, USA, 2022. Association for Computing Machinery.

[11] W. Li, X. Zhu, and S. Gong. Harmonious attention network for person re-identification. *CoRR*, abs/1802.08122, 2018. Accessed: 2024-05-29.

[12] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, and Y. Yang. Improving person re-identification by attribute and identity learning. *CoRR*, abs/1703.07220, 2017.

[13] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang. Bag of tricks and A strong baseline for deep person re-identification. *CoRR*, abs/1903.07071, 2019. Accessed: 2024-05-29.

[14] R. Martinez-Maldonado, V. Echeverria, J. Schulte, A. Shibani, K. Mangaroska, and S. Buckingham Shum. Moodoo: Indoor positioning analytics for characterising classroom teaching. In I. I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, and E. Millán, editors, *Artificial Intelligence in Education*, pages 360–373, Cham, 2020. Springer International Publishing.

[15] H. Ouhaichi, D. Spikol, and B. Vogel. Research trends in multimodal learning analytics: A systematic mapping study. *Computers and Education: Artificial Intelligence*, 4:100136, 2023.

[16] Pass the clap around the classroom. Pass the clap around the classroom: vocabulary game. `https://www.youtube.com/watch?v=RYTKEwrrTB4`, 2019. [Online; accessed 20-February-2023].

[17] R. C. Pianta, K. M. La Paro, and B. K. Hamre. *Classroom Assessment Scoring System™: Manual K-3.* Paul H Brookes Publishing, 2008.

[18] N. Saquib, A. Bose, D. George, and S. Kamvar. Sensei: Sensing educational interaction. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 1(4), Jan 2018.

[19] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian. Deep attributes driven multi-camera person re-identification. *CoRR*, abs/1605.03259, 2016. Accessed: 2024-05-29.

[20] B. Sun, Y. Wu, K. Zhao, J. He, L. Yu, H. Yan, and A. Luo. Student class behavior dataset: A video dataset for recognizing, detecting, and captioning students' behaviors in classroom scenes. *Neural Comput. Appl.*, 33(14):8335–8354, jul 2021.

[21] C.-P. Tay, S. Roy, and K.-H. Yap. Aanet: Attribute attention network for person re-identifications. *CoRR*, abs/1912.09021, 2019. Accessed: 2024-05-29.

[22] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019.

[23] L. Yan, R. Martinez-Maldonado, B. G. Cordoba, J. Deppeler, D. Corrigan, G. F. Nieto, and D. Gasevic. Footprints at school: Modelling in-class social dynamics from students' physical positioning traces. In *LAK21: 11th International Learning Analytics and Knowledge Conference*, LAK21, page 43–54, New York, NY, USA, 2021. Association for Computing Machinery.

[24] F. Yang, K. Yan, S. Lu, H. Jia, X. Xie, and W. Gao. Attention driven person re-identification. *CoRR*, abs/1810.05866, 2018. Accessed: 2024-05-29.

[25] M. Ye, C. Liang, Y. Yu, Z. Wang, Q. Leng, C. Xiao, J. Chen, and R. Hu. Person reidentification via ranking aggregation of similarity pulling and dissimilarity pushing. *IEEE Transactions on Multimedia*, 18(12):2553–2566, 2016.

[26] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):2872–2893, 2022.

[27] Y. Zhai, S. Lu, Q. Ye, X. Shan, J. Chen, R. Ji, and Y. Tian. Ad-cluster: Augmented discriminative clustering for domain adaptive person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9021–9030, June 2020.

[28] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1116–1124, 2015.

[29] L. Zheng, H. Zhang, S. Sun, M. Chandraker, and Q. Tian. Person re-identification in the wild. *CoRR*, abs/1604.02531, 2016. Accessed: 2024-05-29.

[30] Z. Zhong, L. Zheng, D. Cao, and S. Li. Re-ranking person re-identification with k-reciprocal encoding. *CoRR*, abs/1701.08398, 2017. Accessed: 2024-05-29.