Model Transfer for Event Tracking as Transcript Understanding for Videos of Small Group Interaction

Sumit Agarwal, Rosanna Vitiello, Carolyn Penstein Rose
{sumita, rvitiell, cprose}@cs.cmu.edu
Language Technologies Institute, Carnegie Mellon University

Abstract

Videos of group interactions contain a wealth of information beyond the information directly communicated in a transcript of the discussion. Tracking who has participated throughout an extended interaction and what each of their trajectories has been in relation to one another is the foundation for joint activity understanding, though it comes with some unique challenges in videos of tightly coupled group work. Motivated by insights into the properties of such scenarios, including group composition and the properties of task-oriented, goal-directed tasks, we present a successful proof-of-concept. In particular, we present a transfer experiment to a dyadic robot construction task, an ablation study, and a qualitative analysis.

1 Introduction

The broad area of transcript understanding from video encompasses more than the information communicated through discussion, especially when the video captures small group interactions. In that case, each action is meaningful in the context of a broader task. From a social perspective, actions and reactions are meaningful in relation to one another. Sequences of actions of an individual within an interaction are meaningful as an enactment of role taking within a group activity. Building on recent work in multi-object tracking, which is a paradigm of interest in the computer vision community, this paper presents a proof-of-concept for model transfer for tracking the trajectories of participants within a small group activity. In particular, we target tightly coupled group work, which is challenging due to the close proximity of participants, intermittent motion, and periodic movement in and out of view. Success tracking within such scenarios is a key enabler for joint activity understanding, which requires at the foundation tracking who has participated throughout an extended interaction and what each of their trajectories has been in relation to one another. Our results demonstrate positive impact of three different enhancements motivated by consideration of the nature of tightly coupled collaborative group activities.

In many contexts of learning and work, dyads and small groups work together to accomplish a goal. The ability to understand a video capturing this type of interaction has many real world applications. For example, video recordings of such interactions are very common forms of data for research on group learning, communication, and group work. Real time understanding of group interactions has also been used to trigger support for group behavior in order to improve outcomes. Facilitators overseeing multiple groups can use reports of this real time understanding to support decision making regarding how they divide their attention between groups.

In the remainder of the paper, we first offer a review of related work both from the computer vision community and from the multi-modal learning analytics community. Next, we present our technical approach, extending recent successes using DeepSORT for Multiple Object Tracking (MOT). We then present a successful experiment producing results demonstrating improvement over a state-of-the-art baseline, as well as an ablation study to investigate the individual effects of each enhancement and a qualitative analysis of those effects.

2 Background & Related Work

From a technical perspective, the work reported in this paper has its roots in recent directions in the Multi-Object Tracking (MOT) literature. However, as the intended application is within areas of research and practice focused on supported group work and learning, we also review work from the field of Learning Analytics.

Proceedings of the 29th International Conference on Computational Linguistic, pages 20–29
October 12–17, 2022.

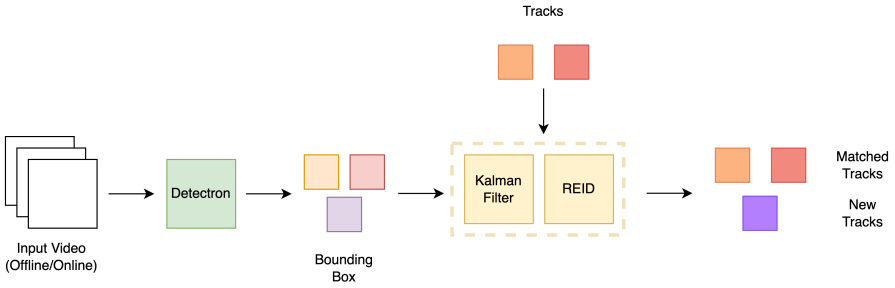


Figure 1: An overview of the DeepDSORT^{+} model architecture used in our experiments which extracts bounding boxes from frames using Detectron and then use Kalman Filter and REID based assignments to match previous tracks and create new unmatched tracks.

2.1 Multi-Object Tracking

In recent years, multi-object tracking has been a growing paradigm of interest in the computer vision community. The task requires the ability to detect multiple objects, mainly individual people, and consistently maintain their identities through the course of their trajectories given video input. The capability to successfully monitor trajectories grounds many high-level multimodal activities in video understanding, such as pose estimation and action recognition (Wang et al., 2013; Luo et al., 2017).

With advances in object detection and the popular MOT benchmark (Dendorfer et al., 2020), many state-of-the-art competitive tracking methods have emerged (Zhang et al., 2021; Wojke et al., 2017; Bewley et al., 2016). Offline models using batching strategies tend to perform well on the benchmark (Zhang et al., 2021). However, in domains where the goal is to achieve live video understanding, computationally efficient online tracking methods (Wojke et al., 2017; Bewley et al., 2016) that sequentially infer trajectories in real-time are preferred.

Despite advances in multi-object tracking, state-of-the-art models struggle with key issues, particularly in maintaining trajectory identities through occlusions and interactions among multiple objects. Our research in this paper shares the common goal of tackling these key issues directly, particularly in its exploration of group work in which complex interpersonal interactions are commonplace and are the necessary centerpiece to understanding dynamics of the activity.

2.2 Tracking Collaboration and Social Processes

In the field of Learning Sciences, automatic temporal analyzes of collaborative data have become essential to operationalize successful learning in student groups. Much of learning analytics has been focused around natural language data, particularly automated analysis of student discussion, which has consistently shown to be a valuable method in assessing student learning (Rosé et al., 2008; McLaren et al., 2007) and scaffolding engaging collaborative interactions (Kumar et al., 2007).

However, with the understanding that collaborative processes are innately multimodal, there is acknowledgment that traditional textual discourse may not tell the entire story. Consequently, multimodal learning analytics has become increasingly popular with the examination of visual patterns such as, in gesture, pose, and eye gaze. Recent studies have used multimodal data to detect misunderstandings during collaborative tasks (Cherubini et al., 2008), discover insights in learning processes (Spikol et al., 2018), and provide beneficial visual feedback to instructors in the classroom (Ahuja et al., 2019, 2021).

More broadly, collaborative learning analysis is one of many social processes that may benefit from precise multi-object tracking. In museums, visitor trajectories can provide curators with insights into improving interaction with content (Mezzini et al., 2020), and body tracking has been used to create immersive digital story telling exhibits (Genc and Häkkilä, 2021). Multi-object tracking is also applied in virtual reality (Uchiyama and Marchand, 2012), and provides information to create simulations for professional development, such as virtual reality for educators in the classroom (Ahuja et al., 2021). Our aim is to contribute to the abil-

ity to identify and maintain trajectories throughout videos, which provides an essential backbone and grounding for these detections in multimodal learning analytics.

3 Method

To perform tracking on videos in small group interaction, we explore the widely used online DeepSORT algorithm developed for Multiple-Object Tracking (MOT) benchmark. We extend the DeepSORT algorithm to improve transfer of the model from the task on which it was trained, namely tracking pedestrians walking on streets, to our group work setting. We begin with an explanation of the well-known DeepSORT algorithm and then discuss the extensions we have added.

3.1 DeepSORT

A tracking model must be able to detect bounding boxes, detect objects to track and continue to identify them for as long as they are in view, thus managing the lifespan of tracked objects. DeepSORT uses F-RCNN (Ren et al., 2015) or YOLO (Redmon and Farhadi, 2018), to detect bounding boxes on tracked objects. Building on SORT, DeepSORT also uses the Kalman filtering framework for track handling. Deviating from SORT, it uses CNN based appearance features for tracking as well, hence the prefix "Deep".

The algorithm considers two means for assigning tracks with bounding box detection, namely, one considering motion and the other considering appearance, captured in two different metrics, as shown in Figure 1.

Kalman Filter - Tracking is based on an 8-dimensional state space (u, v, r, h, $\hat{u}$, $\hat{v}$, $\hat{r}$, $\hat{h}$) that includes the center of the bounding box (u, v), the aspect ratio r and height h and their respective velocities in the image coordinates. A standard Kalman filter with constant velocity motion and linear observation models is used, where the bounding box (u, v, r, h) is considered a direct observation of the object state. It uses squared Mahalanobis distances between the predicted Kalman states and the newly arrived measurements.

$$d_m(i,j) = (d_j - y_i)^T S_i^{-1} (d_j - y_i)$$

where the i-th track distribution is projected into the measurement space as (y_i, S_i) and d_j, which is the j-th bounding box detection for the current frame. We use a high threshold of 0.95 for this distance in order to filter out unassociated detections.

REID - When the motion uncertainty across frames is high, the Mahalanobis distance is not a suitable metric. Also, during occlusions it is very difficult to apply Kalman filter based approaches for continuous frame tracking. Hence, appearance based features using person REID (reidentification) models becomes essential in those scenarios. For this, we compute appearance feature for each of the bounding boxes detected using a CNN-based REID and extract an appearance feature X^i for each track i, which is a function of the current appearance feature of the track x_i and previous X_i. We have used a simple CNN-based REID model to study the effectiveness of the algorithm in zero-shot transfer in our proof-of-concept experiment.

$$x_i = \text{REID}(\text{bounding\_box}(i))$$
$$X^i = f(x_i, X^i)$$

Next, the smallest cosine distance is applied between the previously computed F^i for the i-th track and the j-th detection feature r_j for the frame in consideration, in appearance space, with an admissible threshold, which we keep as 0.2.

$$d_a(i,j) = \min(1 - X^i r_j)$$

For the initial few frames, we use Kalman filter based assignment to confirm the initial set of tracks, and then after that we try matching with the appearance based features, because they are usually consistent across frames. For later frames, only when the appearance based features aren't able to match confirmed tracks with the bounding boxes or there are bounding boxes that are left undetected, we use Kalman Filter based assignment for matching. The metrics are complementary to each other, where the Kalman metric is usually used to recover from short-term motion-based assignments that are missed by appearance metrics, whereas the appearance metric helps to recover detection of objects having been lost from view from long-term occlusion.

3.2 Additions to DeepSORT

To the original DeepSORT algorithm, we introduce a set of enhancements to improve tracking in our target group work settings. We call the model with these changes DeepSORT$^+$. We explore a modification to DeepSORT to replace YOLO with

Detectron. We refer to the revised DeepSORT with Detectron as DeepDSORT, and the version with our enhancements DeepDSORT^{+}. Our proposed algorithm extensions are motivated from insights into tightly coupled group work, in particular, that the extended interaction involves a persistent set of participants who may move in and out of view but otherwise remain stable. Motion within view is related to the group work and thus purposeful. As such, it can be expected that changes in position across frames will be consistent over stretches of time. In summary, our enhancements include no longer deleting tracks with a maximum age, putting a limit on the number of tracks to be created, and introducing a smoothed version of the appearance feature. Our model with enhancements is shown in Figure 1.

3.2.1 No Max Age

DeepSORT uses a max age to maintain the life span of a track. It deletes tracks that have not been detected for a certain number of frames. Since DeepSORT was used for the MOT benchmark, which was used to track pedestrians from surveillance camera videos, it proved to be effective in that context where the total number of objects to track is unbounded, but if a track is not viewed for an extended time, they are unlikely to return. In our setting, the number of participants who are important to track is only the direct participants in the group work, and thus bounded. However, unlike pedestrians moving through an area, they may leave for an extended time, but will nevertheless likely return to the work. In this case, allowing for an unbounded number of tracks is superfluous, and as participants move in and out of view, their movement creates opportunities for false positive detection of new tracks. However, solving the problem by imposing a max age is counter-productive since the likelihood is high that tracks will return even if they have left for some time. Thus, we remove the max age constraint.

3.2.2 Number of tracks

Complementary to removing the max age constraint, we also take advantage of the bounded number of participants in the group work. There may be other people in view, in the background, moving through the space. Bounding the number of tracks reduces the propensity to lose track of a main participant and instead begin tracking someone in the background.

3.2.3 Smoothing appearance feature

In DeepSORT, current frame detections are compared with all previous frame features of tracks to find the closest track. Treating each frame separately introduces the possibility that two different tracks will appear similar. We mitigate this risk by using a smoothed global appearance feature F^i for each track i considering the current frame track feature f_i, given by the following formula.

$$F^i = \alpha * f_i + (1 - \alpha) * F^i$$

Reducing the set of observations of a track to a single smoothed version reduces the danger of a pair of frames from different tracks inadvertently appearing similar. For our experiments, we set α as 0.1, weighing heavily towards past observations and changing the representation only slowly over time.

3.2.4 Detectron

To identify bounding boxes, F-RCNN or YOLO based models have been shown to be very effective, which are also used in DeepSORT. Detectron (Wu et al., 2019) is an object detection model that is able to detect more concise human-based bounding boxes but with higher accuracy which is essential in our cases because the appearance features might confuse with the other people or objects if the bounding box is not very accurate in person position. We call this model DeepDSORT^{+}.

4 Experiments

4.1 Dataset

In order to evaluate our multimodal approach in small group activities and social processes, we collected and annotated an exploratory video corpus from a summer course conducted at Carnegie Mellon University. During the course, groups of 2-3 students participated in a robotic arm instruction task. The activity occurred over two collaborative sessions, each lasting around several hours: a robotic construction session in which students built their mechatronic arms and a robotic arm learning activity session in which students operated their robot. Each group collected video and audio data during each session using a Kodak Orbit 360 4K VR Camera with its 197° 4K Ultra Wide View Front Lens. Students were instructed to place each camera on a small tripod at the end of their table to capture every member of the group and the robotic arm.

	Dataset	MOTA ↑	MOTP ↑	IDF1 ↑	MT ↑	ML ↓	IDsw ↓	FP ↓	FN ↓
DeepSORT	Group 1	24.6	62.1	43.1	18.2	9.1	77	2229	2795
DeepSORT	Group 2	5.2	56.4	38.7	25	25	24	2274	2718
DeepSORT	Combined	15.4	59.7	41	21.6	17.1	101	4503	5513
DeepSORT$^+$	Group 1	31.2	62.8	52.7	18.2	18.2	22	1135	2613
DeepSORT$^+$	Group 2	6.3	56.5	46.1	25	25	16	1813	2463
DeepSORT$^+$	Combined	18.8	59.6	49.4	21.6	21.6	**19**	1474	2538
DeepDSORT$^+$	Group 1	78.1	89.9	85.9	63.6	9.1	36	165	1033
DeepDSORT$^+$	Group 2	93.1	90.3	96.5	100	0	20	92	204
DeepDSORT$^+$	Total	**85.6**	**90.1**	**91.2**	**81.8**	**4.6**	56	**257**	**1237**

Table 1: Combined results on our videos for DeepSORT, DeepSORT$^+$ and DeepDSORT$^+$ models. The arrow indicates whether higher value indicates a good (↑) or a bad (↓) result.

	MOTA ↑	MOTP ↑	IDF1 ↑	MT ↑	ML ↓	IDsw ↓	FP ↓	FN ↓
DeepDSORT$^+$	**85.6**	**90.1**	**91.2**	**81.8**	4.6	56	**257**	**1237**
DeepDSORT$^+$ − smooth	84.3	90.1	84.5	81.8	**0**	77	795	1824
DeepDSORT$^+$ − #tracks	83.6	90	85.6	77.3	0	78	957	1581
DeepDSORT$^+$ + max age	83.6	90.1	73.9	21.1	74.9	**15**	1842	2636

Table 2: Different ablations for our DeepDSORT$^+$ model on both groups of videos by removing each component that we introduce specifically for tracking group social processes.

These videos help study tracking in confined spaces with limited number of people in social processes. The videos feature many interactions between students and the robotic arm, as well as movement in different locations. Other complex scenarios include occasional off-camera movement, irrelevant background activity from other groups, and intermittent occlusions of students. Moreover, by using a video corpus collected via a small portable camera, social processes such as these may be collected and given support in real-time. Consequently, this corpus provides key scenarios that are essential to be able to track people consistently across frames for downstream automatic analysis of individual and group traits and outcomes.

We conduct our experiments across 2 student groups. We divided each video session into short 8 minute sections and extract about 500 frames with 1 fps for tracking annotation from every section. Each video has a gold standard tracking annotation created by extracting person class-based bounding boxes for each frame using F-RCNN and labeling each box with person IDs. In sum, we experimented with 4 8-minute videos from each group (8 videos in total), which comprises of 4148 annotated frames with a maximum of 3 students in a particular frame. People in the background of the frames uninvolved in the activity are not annotated because they are not part of the group collaboration.

4.2 Metrics

We evaluate our videos on metrics that have been commonly used for MOT benchmark, particularly we focus on the following values:

MOTA: Combines three error sources: false positives, missed targets and identity switches
MOTP: Misalignment between the predicted and ground-truth bounding boxes
IDF1: Ratio of correctly identified detections over the average of computed and ground-truth detections
MT: Mostly tracked targets that are tracked atleast 80% of their life span
ML: Mostly lost targets that are tracked atmost 20% of their life span
IDsw: Total no of identity switches
FP: Total no of false positives
FN: Total no of false negatives / missed targets

We highlight these metrics because they are cru-

cial for further downstream applications concerning individual and group activity in social processes. That is, if models do not perform well on these metrics, they cannot perform an essential goal in multi-modal video understanding: identifying key roles and salient interactions during social processes. It is most important for models in this domain to accurately identify tracks and consistently maintain tracks without error. Additionally, we are most concerned with a model's ability to never lose or miss tracks in an activity, highlighting an emphasis on reducing false positives.

4.3 Experiments and Ablation

For both Group 1 and Group 2, we conduct the following experiments across variations of the DeepSORT model as described in 3:

- **DeepSORT** : original DeepSORT model as implemented by (Bewley et al., 2016) which uses YOLO to identify bounding boxes [1]
- **DeepSORT$^+$** : modified DeepSORT with YOLO and all additions mentioned in 3.2
- **DeepDSORT$^+$** : modified DeepSORT with Detectron [2] and all additions mentioned in 3.2

We also perform an ablation over the modified DeepDSORT$^+$ model through the removal of modified individual components:

- **DeepDSORT$^+$ − smooth** : modified Detectron DeepSORT without smoothing appearance feature
- **DeepDSORT$^+$ − # tracks** : modified Detectron DeepSORT without restriction of number of tracks
- **DeepDSORT$^+$ + max age** : modified Detectron DeepSORT with max age for tracking

For our experiments, we run the model over tracking for all the frames of the original video at 30 fps but the model is evaluated only on the gold-standard annotated frames extracted at 1 fps. We used 1 NVIDIA GTX 1080 GPU to run tracking over each video. Each frame takes a processing time of 0.18 s yielding a total of 5 fps. Running this online, in real time, would process 5 frames per second which is quite efficient for a tracking algorithm. This is another reason for choosing DeepSORT as the baseline because it is an ONLINE algorithm which is suitable for our purposes. For the CNN based Person REID model, we pretrain the model on the market1501 (Zheng et al., 2015) dataset, which is also used in the original DeepSORT implementation.

[1] https://github.com/mikel-brostrom/Yolov3_DeepSort_Pytorch

[2] https://github.com/facebookresearch/detectron2

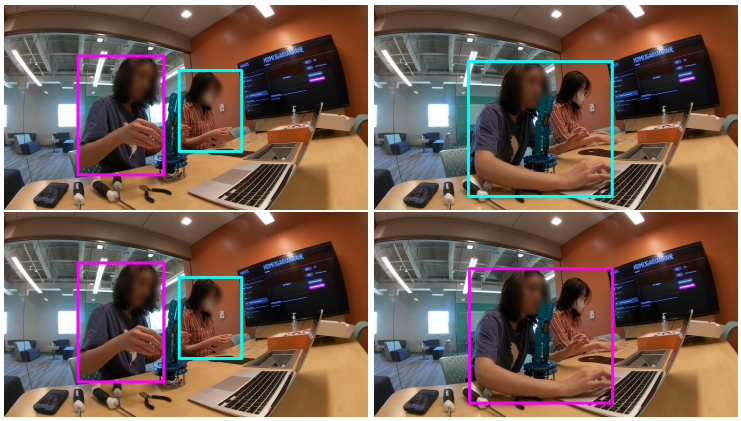
5 Results

Table 1 shows the results of tracking over two sets of videos collected for two different groups, across DeepSORT, DeepSORT$^+$ and DeepDSORT$^+$ models. We can see that introducing the required components discussed in Section 3.2, to just the DeepSORT model, leads to a decrease in false positives and false negatives in DeepSORT$^+$. Further, we see improvements in almost all the metrics in DeepDSORT$^+$ showing that Detectron, in general, is a better model than YOLO, and our additional extensions lead to further improvement. Better bounding boxes implies better appearance features that make the appearance REID model less confused, leading to a drop in false positives and false negatives, thereby increasing IDF1. MOTA and MOTP metric also improve because the detected bounding boxes are closer to the ground-truth ones. We see that for Group2 videos the performance improvement is larger due to the Detectron model detecting people in the videos more accurately. There are more ID switches in the DeepDSORT$^+$ model than in DeepSORT$^+$, but significantly less than in DeepSORT. These ID switches account for a count of the frames in which the IDs are switched. The increased performance of DeepDSORT$^+$ implies that in the face of ID switches, it is able to recover.

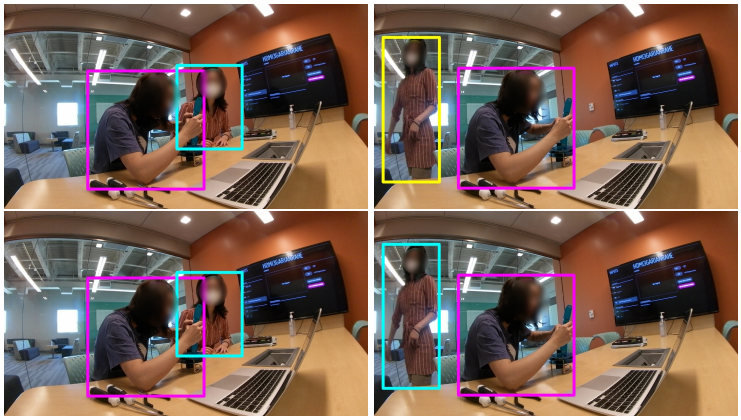
6 Analysis

We perform ablation of our model to assess the impact of each of our extensions. Table 2 shows the results of removing each component. Figure 2 shows examples of qualitative errors introduced by removing each component compared to the complete model.

Smooth vs Non Smooth: For the smoothing ablation, instead of adding the new appearance feature as discussed in 3.2, we average all appearance features so that each frame feature receives equal importance such that smoothing is not applied. Without smoothing, abrupt changes in appearance or slight movement will likely change inference more drastically. We expect that without smoothing, the



(a) Tracking results when smoothing is not done and the features are averaged out for all the past frames, showing that it gives rise to false positives and ID switches. The upper results are DeepDSORT^{+} results without smoothing and the lower ones are DeepDSORT^{+} results.



(b) Tracking results when there is no cap on the number of tracks, showing that it gives rise to new tracks getting created when there is a slight mismatch in features (like when the person is in motion). The upper results are DeepDSORT^{+} results without limit on number of tracks and the lower ones are DeepDSORT^{+} results.

(c) Tracking results when maximum age is included as 100 frames, showing that when bounding box detections are missed in between frames or people move in and out of the frame, new IDs are created, (cyan to yellow). The upper results are DeepDSORT^{+} results with maximum age and the lower ones are DeepDSORT^{+} results.

Figure 2: Qualitative ablation results showing the removal of each component as discussed in Table 2, by removing smoothing in a), removing cap on number of tracks in b) and adding maximum age of tracks in c). Colors around bounding box indicate the track associated with the person, where a change in the color indicates an error made by the model.

model is more likely to confuse IDs when there is motion. Quantitatively, Table 2 supports this finding by revealing smoothing decreases ID switches and false positives.

In qualitative analysis of the smoothing ablation, we find errors that align with these expectations. Note in Figure 2a, the track in the purple bounding box is incorrectly switched when the individual leans forward in the non-smoothing model. However, with smoothing, the model correctly maintains their track. From this ablation, we conclude that smoothing helps decrease noise in abrupt changes of appearance in cases of obstruction and motion.

Limited vs Unlimited Number of Tracks: We also examine the ablation that removed the limit of the number of tracks that can be created. By limiting the maximum number of tracks to the number of participants within the activity, we hypothesized that the model would maintain tracks more consistently with less likelihood of creating irrelevant tracks during motion. The results in Table 2 support this hypothesis, as allowing the model to infer an unlimited number of tracks increased the rate of false positives.

This can be seen qualitatively in Figure 2b. Due to motion by the individual in the cyan bounding box, the ablation model mistakes motion for a new person and incorrectly creates a new yellow bounding box track around the individual. When tracks are limited, the model does not have the ability to create a new track and correctly maintains the identity of the moving individual.

No Max Age vs Max Age: In the maximum age ablation, we limit the maximum age of tracks to 100 frames. Originally, this threshold was used to remove unnecessary tracks that leave and never return in frame, commonly experienced in the benchmark MOT dataset. In collaborative social processes, the maximum age assumption was no longer appropriate. We noted that often individuals returned to the field of view after being occluded or out of frame for long periods of time, or they remain undetected by the model. By removing the max age threshold, we suspected the model would correctly maintain relevant tracks rather than discarding them.

This is quantitatively supported by the large increase in false positives and the decrease in IDF1 when a maximum age threshold of 100 frames was introduced. This can also be observed in Figure 2c, as the cyan bounding box individual is incorrectly discarded after leaving the frame and labeled as a new yellow bounding box individual when returning. This identity is correctly maintained without the maximum age threshold.

We note that keeping a higher maximum age threshold above 100 frames may also be a solution to this issue. However, it is impossible to define a generalizable amount of time for which people within a given activity will be out of frame. Hence, we conclude removing the maximum age threshold so that relevant tracks are never removed is the best approach for this modification.

7 Limitation

This paper targets tightly coupled group work, which is a closed setting with a fixed finite number of participants. If this assumption were required to be lifted, then people in the background might introduce the potential for false positives. A direction that would be valuable to explore in that case would be taking depth-perception into account in order to properly distinguish those engaged in the task from people in the background. As people move, their appearance changes, which introduces challenges for the matching process. One possible direction would be to tune the REID model over the first few frames when a new track appears. In order to further extend capabilities to participants who are easily confused, for example because of wearing similar clothing, more sophisticated REID models might be used that treat different body parts of individuals separately.

8 Conclusion

This paper presents a successful proof of concept for the transfer of models trained to track pedestrians to a scenario that features tightly coupled group work. With a small change to the original DeepSORT algorithm, using Detectron instead of YOLO, we are already able to achieve substantial improvement. Additional extensions motivated by the characteristics of tightly coupled group work add further improvement. In future work we play to explore more sophisticated REID models for this purpose. While this study lays the foundation for joint activity understanding, much is left to be done to explore aspects other than participant trajectories, such as the interplay of participant emotions and joint eye gaze.

Acknowledgements

This research was funded in part by NSF grants 2100401 and 1917955.

References

Karan Ahuja, Dohyun Kim, Franceska Xhakaj, Virag Varga, Anne Xie, Stanley Zhang, Jay Eric Townsend, Chris Harrison, Amy Ogan, and Yuvraj Agarwal. 2019. Edusense: Practical classroom sensing at scale. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 3(3).

Karan Ahuja, Deval Shah, Sujeath Pareddy, Franceska Xhakaj, Amy Ogan, Yuvraj Agarwal, and Chris Harrison. 2021. Classroom digital twins with instrumentation-free gaze tracking. CHI '21, New York, NY, USA. Association for Computing Machinery.

Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. 2016. Simple online and realtime tracking. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 3464–3468.

Mauro Cherubini, Marc-Antoine Nüssli, and Pierre Dillenbourg. 2008. Deixis and gaze in collaborative work at a distance (over a shared map): A computational model to detect misunderstandings. In *Proceedings of the 2008 Symposium on Eye Tracking Research Applications*, ETRA '08, page 173–180, New York, NY, USA. Association for Computing Machinery.

P. Dendorfer, H. Rezatofighi, A. Milan, J. Shi, D. Cremers, I. Reid, S. Roth, K. Schindler, and L. Leal-Taixé. 2020. Mot20: A benchmark for multi object tracking in crowded scenes. *arXiv:2003.09003[cs]*. ArXiv: 2003.09003.

Caglar Genc and Jonna Häkkilä. 2021. Using body tracking for involving museum visitors in digital storytelling. In *Augmented Humans Conference 2021*, AHs'21, page 304–306, New York, NY, USA. Association for Computing Machinery.

Rohit Kumar, Carolyn P. Rosé, Yi-Chia Wang, Mahesh Joshi, and Allen Robinson. 2007. Tutorial dialogue as adaptive collaborative learning support. In *Proceedings of the 2007 Conference on Artificial Intelligence in Education: Building Technology Rich Learning Contexts That Work*, page 383–390, NLD. IOS Press.

Chenxu Luo, Chang Ma, Chunyu Wang, and Yizhou Wang. 2017. Learning discriminative activated simplices for action recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).

Bruce M. McLaren, Oliver Scheuer, Maarten De Laat, Rakheli Hever, Reuma De Groot, and Carolyn P. Rosé. 2007. Using machine learning techniques to analyze and support mediation of student e-discussions. In *Proceedings of the 2007 Conference on Artificial Intelligence in Education: Building Technology Rich Learning Contexts That Work*, page 331–338, NLD. IOS Press.

Mauro Mezzini, Carla Limongelli, Giuseppe Sansonetti, and Carlo De Medio. 2020. Tracking museum visitors through convolutional object detectors. In *Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '20 Adjunct, page 352–355, New York, NY, USA. Association for Computing Machinery.

Joseph Redmon and Ali Farhadi. 2018. Yolov3: An incremental improvement. *ArXiv*, abs/1804.02767.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Carolyn Rosé, Yi-Chia Wang, Yue Cui, Jaime Arguello, Karsten Stegmann, Armin Weinberger, and Frank Fischer. 2008. Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning. *I. J. Computer-Supported Collaborative Learning*, 3:237–271.

Daniel Spikol, Emanuele Ruffaldi, Giacomo Dabisias, and Mutlu Cukurova. 2018. Supervised machine learning in multimodal learning analytics for estimating success in project-based learning. *Journal of computer assisted learning.*, 34(4).

Hideaki Uchiyama and Eric Marchand. 2012. Object Detection and Pose Tracking for Augmented Reality: Recent Approaches. In *18th Korea-Japan Joint Workshop on Frontiers of Computer Vision (FCV)*, Kawasaki, Japan.

Chunyu Wang, Yizhou Wang, and Alan L. Yuille. 2013. An approach to pose-based action recognition. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 915–922.

Nicolai Wojke, Alex Bewley, and Dietrich Paulus. 2017. Simple online and realtime tracking with a deep association metric. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3645–3649. IEEE.

Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. 2019. Detectron2. https://github.com/facebookresearch/detectron2.

Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. 2021. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *International Journal of Computer Vision*, 129:3069–3087.

Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. 2015. Scalable person re-identification: A benchmark. In *Computer Vision, IEEE International Conference on*.