

Multiple Toddler Tracking in Indoor Videos

Somaieh Amraee^{1,2}, Bishoy Galoaa¹, Matthew Goodwin³,
Elaheh Hatamimajoumerd^{1,2}, Sarah Ostadabbas^{1*}

¹Department of Electrical & Computer Engineering, Northeastern University, MA, USA

² Roux Institute at Northeastern University, ME, USA

³ Bouve College of Health Sciences & Khoury College of Computer Sciences, Northeastern University, MA, USA

*Corresponding author's email: Ostadabbas@ece.neu.edu

Abstract

Multiple toddler tracking (MTT) involves identifying and differentiating toddlers in video footage. While conventional multi-object tracking (MOT) algorithms are adept at tracking diverse objects, toddlers pose unique challenges due to their unpredictable movements, various poses, and similar appearance. Tracking toddlers in indoor environments introduces additional complexities such as occlusions and limited fields of view. In this paper, we address the challenges of MTT and propose MTTSort, a customized method built upon the DeepSort algorithm. MTTSort is designed to track multiple toddlers in indoor videos accurately. Our contributions include discussing the primary challenges in MTT, introducing a genetic algorithm to optimize hyperparameters, proposing an accurate tracking algorithm, and curating the MTTrack dataset using unbiased AI co-labeling techniques. We quantitatively compare MTTSort to state-of-the-art MOT methods on MTTrack, DanceTrack, and MOT15 datasets. In our evaluation, the proposed method outperformed other MOT methods, achieving 0.98, 0.68, and 0.98 in multiple object tracking accuracy (MOTA), higher order tracking accuracy (HOTA), and iterative and discriminative framework 1 (IDF1) metrics, respectively¹.

1. Introduction

Multiple toddler tracking (MTT) involves the detection of toddlers in video footage and continuous tracking with a unique identification number. According to the American Academy of Pediatrics (AAP), toddlers are children aged 1-3 years characterized by their active engagement in activities such as climbing, running, and jumping [8]. Ensuring the safety of toddlers during their daily routines, both at

home and in care facilities, is a paramount concern for parents and their assigned caregivers. The identification and monitoring of toddlers' movements is of significant interest for researchers in child development [10, 11], nursing [36], and various health-related fields [28], particularly early detection of motor-related abnormalities [6, 14, 20].

Multiple toddler tracking falls under the specialized category of multiple object tracking (MOT) algorithms. MOT is a crucial component of various scene-understanding tasks, including surveillance [27], robotics [1, 34], and autonomous [7]. These algorithms track multiple moving objects while assigning a unique identifier to each [2, 19, 31]. Typically, various objects are present in each frame and may belong to different classes. MOT algorithms are comprised of three key steps: detection (finding all objects in a frame), localization (determining detected object positions in a frame), and association (matching objects across frames to maintain consistent identifiers) [22]. Particularly in human tracking and monitoring systems, ensuring each person has unique and persistent identifier numbers throughout the video is a critical requirement.

While some toddler detection methods have been proposed to measure the distance of toddlers from dangerous objects and prevent potentially hazardous situations [8, 9], it is evident that developing an algorithm capable of simultaneously tracking multiple children with unique identifiers is significantly more challenging. Multiple toddler tracking (MTT) presents three primary obstacles. First, many existing face and bounding box detection methods are trained primarily on adult samples, leading to numerous errors when applied to child detection (i.e., detection challenge) [29]. Second, young children exhibit unpredictable movement patterns involving rapidly changing directions and positions, such as walking, sitting, and crawling, making it challenging to establish an accurate tracking model (i.e., localization challenge) [32]. Third, when multiple young children are present in a scene, distinguishing them can be problematic due to their similar appearances (i.e., as-

¹The MTTSort code is available at <https://github.com/ostadabbas/Multiple-Toddler-Tracking>.

sociation challenge) [29, 32]. Furthermore, to address the demand for intelligent systems for toddler tracking, they should be designed for indoor environments, such as homes and daily care centers. In these applications, the limitation of field of view, results in some challenges such as a high rate of occlusion [21]. These challenges underscore the essential need for a customized and accurate method for tracking multiple toddlers in indoor videos.

This paper discusses the main challenges and potential solutions of MOT methods for multiple toddlers tracking in indoor videos. Then a customized and accurate method, MTTSort, is proposed to resolve these challenges and achieve high efficiency for toddler tracking. This method builds upon the DeepSort algorithm [31] which is a state-of-the-art MOT approach that has demonstrated significant potential for customization in indoor applications in our experiments. In the first step of the proposed method, a genetic algorithm is proposed to optimize the hyperparameters. Then a new extension of the DeepSort algorithm is developed for multiple toddler tracking in indoor videos. This paper also provides a quantitative comparison of state-of-the-art methods including DeepSort [31], StrongSort [5], HybridSort [33], and Bytetrack [35]. Since there is no publicly available dataset for multiple toddler tracking, we built a video set, called the MTTrack dataset. Comprehensive evaluation and comparison have been conducted on MTTrack and two other public tracking datasets, DanceTrack [26], and MOT15 [3]. In summary, the paper introduces several significant contributions:

- Discussing the main challenges of MOT methods for multiple toddler tracking applications in indoor videos.
- Providing a genetic algorithmic method to make sure that the optimum hyperparameters are used for tracking.
- Proposing an accurate tracking algorithm that is customized for multiple subject tracking in indoor videos.
- Building and annotating the MTTrack dataset using the AI co-labeling techniques, ensuring no algorithmic biases.
- A quantitative comparison of state-of-the-art MOT methods on two public datasets as well as the MTTrack dataset.

Overall, these contributions enhance our understanding of multiple object tracking and provide valuable resources for further research in this domain. The rest of this paper is structured as follows: State-of-the-art MOT methods and their shortcomings for multiple toddler tracking in indoor videos are described in Section 2. Then, in Section 3, the proposed customized method for multiple toddler tracking

is described. The experimental results are shown in Section 4. Finally, Section 5 is dedicated to conclusions and recommendations for future research.

2. Related Works

In this section, the widely used existing MOT methods are briefly described, and then the main shortcomings of these methods for multiple toddler tracking especially in indoor videos are discussed.

DeepSort [31] is designed to track multiple objects in real-time applications by combining object detection with deep neural networks, specifically using YOLO [23] and Deep Association Network (DAN) [30]. DeepSort extends the original simple online and real-time tracking [2] (SORT) algorithm by improving the re-identification of objects after an occlusion. In this method, the object detection module detects objects in each frame. YOLOv8 has been used as a deep object detection model in this paper. DeepSort uses the Kalman filter to estimate the state of each track (i.e., position, velocity, size, and age) and predict its future location. The Kalman filter is used in combination with the data association module to update the state of each track based on the detected objects in the current frame and predict the state of each track in the next frame. The track management module in DeepSort is responsible for updating the state of each track over time and removing tracks that are no longer valid or reliable. This module keeps track of each object’s position, velocity, size, and age, and uses this information to predict the object’s future location and update its state. In addition to updating the state of each track, the track management module also performs track maintenance tasks such as track initiation and track termination. While DeepSort exhibits the capability to track multiple moving objects effectively in straightforward scenarios where the objects are widely spaced and follow uncomplicated trajectories, it tends to produce numerous errors in more intricate situations, particularly in indoor video environments characterized by prolonged occlusions.

StrongSort [5] revisits the classic DeepSort tracker with the appearance-free (AFLink) model and Gaussian-smoothed interpolation (GSI). StrongSort first uses an enhanced correlation coefficient maximization (ECC) to account for motion noise caused by movements. Then, a modified Kalman filter that emphasizes non-linear motions is used to calculate the weightings during each update across frames. Lastly, for object association, StrongSort directly includes the motion information in addition to appearance for more accurate tracking. While StrongSort generally outperforms DeepSort as a more accurate tracker when evaluated on publicly available outdoor datasets, our experimental findings indicated an unexpected outcome. Specifically, in our experiments, the performance of StrongSort was notably inferior to DeepSort, particularly in indoor video sce-

narios. Consequently, we introduced our tracking method, based on DeepSort, with the primary objective of achieving precise tracking of multiple toddlers within indoor videos.

HybridSort [33] improves tracking when long-standing failure cases are caused by heavy occlusion and clustering. In this situation, strong cues such as spatial and appearance information become unreliable simultaneously. In this research, they demonstrated an important finding that previously overlooked weak cues, such as confidence state, height state, and velocity direction, can compensate for the limitations of strong cues. Then, they proposed HybridSort by introducing simple modeling for the newly incorporated weak cues and leveraging both strong and weak cues. The design effectively and efficiently resolves ambiguous matches generated by strong cues and significantly improves association performance. A critical limitation we observed in our experiments regarding HybridSort is its performance inconsistency. Specifically, when toddlers are in motion, this method yields highly satisfactory results. However, it encounters significant challenges when toddlers are either stationary or exhibiting minimal movement, leading to a notable increase in tracking errors.

ByteTrack [35] has been proposed to fix missing predictions by using low-confidence candidates in association, achieving good performance by balancing the detection quality and tracking confidence. It focuses on associating almost every detection box, including low-score ones, to recover true objects and filter out background detections. ByteTrack associates every detection box and uses similarities between tracklets to reduce false positives and negatives for low-score detection bounding boxes. It proposes a second matching stage in which low-confident detections are associated with unmatched tracks from the first stage. The low-confident detections are not used to start new tracks, ensuring no ghost tracks from low-confident false positive detections are introduced. The authors of ByteTrack show that this two-stage matching (TSM) improves the tracking performance when integrated into various frameworks. In our initial experiments, ByteTrack showed promising results. However, it became clear that it had limited potential for customization and parameter adjustment, particularly when dealing with indoor videos, in contrast to the DeepSort algorithm. Consequently, we made the decision to enhance and tailor DeepSort with our modifications to create a highly accurate tracker for the purpose of tracking multiple toddlers in indoor video settings.

Public Datasets for multi-object tracking have been organized into two main resources, MOT [3] and DanceTrack [26]. The MOT datasets are designed for the task of multiple object tracking. There are several variants of the dataset released each year, such as MOT15, MOT17, and MOT20. DanceTrack is a large-scale multi-object tracking dataset for human tracking in occlusion, frequent crossover, uni-

form appearance, and diverse body gestures. It is proposed to emphasize the importance of motion analysis in multiple object tracking instead of mainly appearance-matching-based diagrams.

3. Our MTTSort

Here, we outline the primary difficulties faced by MOT techniques when tasked with tracking multiple toddlers in indoor video environments. To address these challenges, we introduce our novel approach. Our method (MTTSort), which builds upon the DeepSort algorithm, consists of two pivotal phases. In the initial phase, we utilize a genetic algorithm to determine the optimal parameters for DeepSort. This crucial step ensures the use of optimized parameters prior to any further adjustments. Following this, in the second phase, we implement specific modifications to DeepSort, thereby improving its precision in tracking multiple toddlers in indoor video scenarios.

3.1. MTT Main Challenges

Utilizing a conventional MOT method for tracking toddlers in indoor videos presents several noteworthy challenges, as outlined below:

Adult-Centric Models: Existing detection and tracking models have predominantly been trained on adult samples, resulting in a significant number of errors when detecting children.

Unpredictable Movements: Young children exhibit unpredictable and rapid changes in their movements and positions, making it challenging to develop an accurate tracking model for them.

Activity Variability: Toddlers engage in diverse activities, including walking, sitting, lying, and crawling, all within a single video sequence. This variability introduces higher error rates in both detection and tracking processes.

Similar Appearances: Distinguishing between toddlers can be problematic due to their similar physical appearances. This challenge becomes more pronounced, particularly when tracking twins, given their resemblance.

Toy Confusion: In scenarios where a child interacts with humanoid toys or action figures, the detection stage may mistakenly identify the toy as a real subject, leading to tracking errors.

Limited Camera View: Indoor video setups often employ stationary cameras with restricted fields of view, resulting in frequent occlusions, even with a small number of subjects being tracked.

Extended Occlusion: Unlike outdoor environments where occlusion might occur over a few frames, indoor scenarios involve prolonged occlusion extending over consecutive frames. This extended occlusion poses a significant challenge in maintaining individual subjects' identities.

Data Challenges: Developing a novel model for tracking multiple toddlers necessitates a substantial dataset for training. However, data collection and labeling for children’s research are expensive, individualized, and subject to stringent privacy and classification laws. Capturing multiple video clips from various children can also be time-consuming due to their unpredictable movements.

In summary, the importance of “SmallData” is evident, emphasizing the challenge of obtaining sufficient labeled data for toddler tracking. Customizing existing methods is crucial to adapting to the unique demands and complexities associated with toddler tracking, enabling the development of more precise tracking solutions tailored to their distinctive characteristics and requirements.

3.2. Parameter Optimization Using Genetic Algorithm

In the realm of optimization problems, the Genetic Algorithm (GA) stands out as a bio-inspired heuristic, rooted in the process of natural selection. GAs operate by simulating the process of evolution found in nature. Beginning with a population of potential solutions (analogous to individuals or organisms), the algorithm iteratively evaluates, selects, mates, and mutates these solutions. The key principle is that over successive generations, better and fitter solutions emerge, closely resembling the evolutionary concept of “survival of the fittest.” The algorithm’s capacity to explore a vast solution space by intelligently combining and modifying solutions makes it especially effective for problems where the optimal solution is elusive or computationally intensive to ascertain directly [13].

In our pursuit of tackling the complexities of multi-toddler tracking, we embarked on a comprehensive exploration of various configurations encompassing diverse aspects of the detection and tracking challenges. Following a rigorous evaluation, we concluded that DeepSort emerged as the most suitable solution to meet our tracking needs. Nonetheless, optimizing performance necessitated a meticulous fine-tuning of DeepSort hyperparameters using a genetic algorithm [17]. We pinpointed specific hyperparameters, each associated with a defined range derived from empirical insights and domain expertise. These critical hyperparameters are shown in Table 1.

To discover the optimal values within these ranges, we adapted a mainstream genetic algorithm. The fitness function for our genetic algorithm was designed to maximize the aggregate score, *Score*, defined as:

$$Score = HOTA + MOTA + IDF1, \quad (1)$$

where, HOTA (higher order tracking accuracy) [16], MOTA (multiple object tracking accuracy) [12], and IDF1 (iterative and discriminative framework 1) [24] are the three main accuracy criteria in MOT algorithms. Each metric was given

an equal weight. This approach allows us to automatically and efficiently search the hyperparameter space, ultimately leading to enhanced tracking performance in our MTT system. Algorithm 1 shows the main body of the proposed genetic algorithm for optimizing the parameters.

Table 1. MOT hyperparameters with descriptions.

Parameter	Range	Description
MAX_DIST	[0.1, 1.0]	Maximum distance for matching
MIN_CONFIDENCE	[0.2, 0.7]	Minimum confidence for detection
NMS_MAX_OVERLAP	[0.2, 0.8]	Max overlap for non-max suppression
MAX_IOU_DISTANCE	[0.2, 0.9]	Max IoU distance for tracking
MAX_AGE	[10, 200]	Max age of a track without detection
N_INIT	[2, 15]	Number of initial detections
NN_BUDGET	[20, 120]	Nearest neighbors’ budget

Algorithm 1 Genetic algorithm for hyperparameter optimization in MOT

- 1: **Define** `config_template` with hyperparameters of Table 1.
 - 2: **Initialize** a population of individuals using the `config_template`.
 - 3: **Evaluate** the fitness of each individual in the population using the fitness function:
 $Score = HOTA + MOTA + IDF1$.
 - 4: **while** standard deviation of scores in the population is greater than tolerance OR maximum number of generations not reached **do**
 - 5: **Select** parents from the current population based on their scores.
 - 6: **Perform** crossover (recombination) on pairs of parents to produce offspring.
 - 7: **Mutate** offspring based on mutation rate.
 - 8: **Evaluate** the score of the offspring using the fitness function.
 - 9: **Replace** the current population with the offspring.
 - 10: **end while**
 - 11: **return** The solution (individual) with the best score from the population.
-

3.3. Indoor MTT

In the realm of tracking applications, particularly those involving subjects like toddlers with highly similar appearances and unpredictable movements, conventional track-

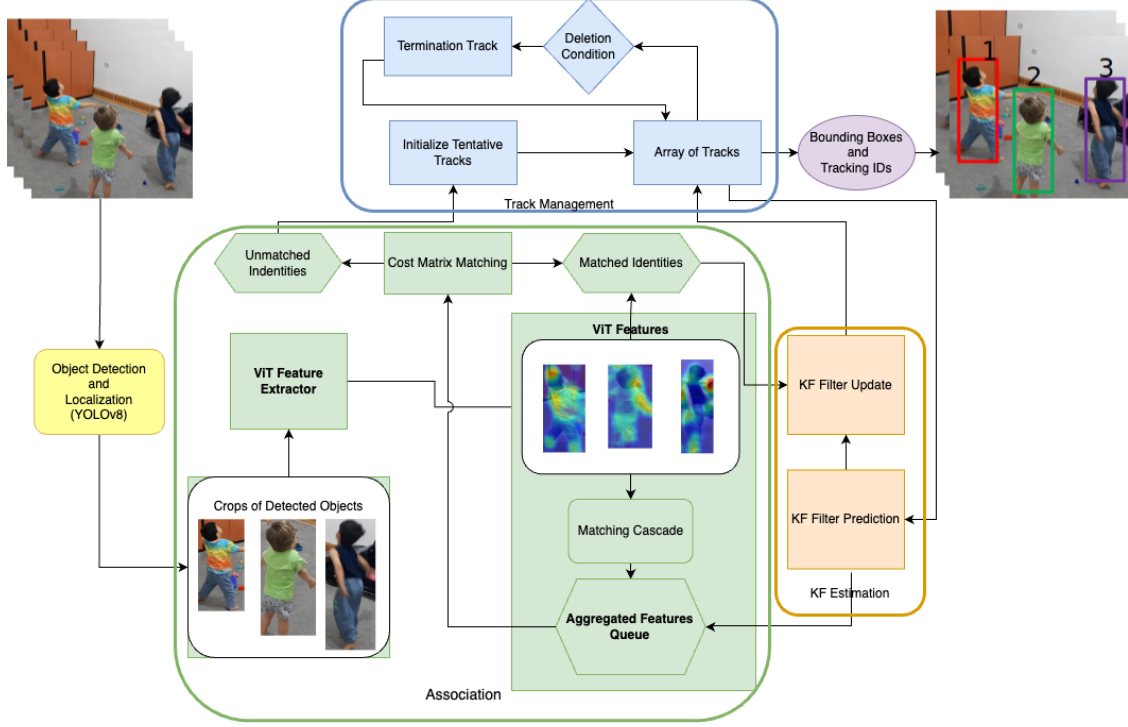


Figure 1. Our proposed method, MTTSort for multiple toddler tracking in indoor videos. This diagram illustrates two significant enhancements to the traditional DeepSort framework: (1) Pooled Aggregated Feature Association with a Custom Buffer, a mechanism that accumulates and consolidates features across consecutive frames in a user-defined buffer, and (2) Attention-Based Feature Extraction with ViT, which replaces conventional CNNs with the Vision Transformer for a more refined and attention-focused feature extraction process. Both modifications are designed to tackle the challenges posed by subjects like toddlers, characterized by their similar appearances and unpredictable movements.

ing algorithms such as DeepSort often grapple with maintaining consistent identities. Recognizing these challenges, we have introduced significant modifications, integrating a state-of-the-art feature association mechanism into the DeepSort framework. **Figure 1** illustrates the essence of our proposed method. Our method enhances the DeepSort algorithm by adding two parts to it: (1) pooled aggregated feature association with custom buffer, and (2) attention-based feature extraction with the vision transformer (ViT).

Pooled aggregated feature association with custom buffer: While traditional tracking methods, including DeepSort, predominantly rely on the immediate features from the current frame, our approach takes a leap forward. We’ve introduced a custom-sized feature buffer that aggregates features over a series of frames. Specifically, in our experiment, the buffer size was fixed to store up to 5 features extracted per object. This choice of a buffer size of 5 was deliberate; we found that increasing the buffer size further led to an accumulation of more divergent features over time. As a result, the matching process could become less accurate, since features could deviate significantly from the object’s most recent appearance. Hence, a size of 5 strikes a balance between retaining recent appearance information and ensuring effective feature matching.

The features buffer was designed as a queue, and the queuing mechanism operates under two distinct conditions: (1) When the Kalman filter terminates a track, leading to the removal of associated features, or (2) when the buffer reaches its full capacity, implying that 5 features have already been buffered. In such a scenario, the oldest feature is dequeued, ensuring that only the 5 “last seen” features are stored. Essentially, this buffer acts as a temporal sliding window for each object, capturing the most recent and relevant appearance data over time.

This buffer retains the “last seen” features for each subject across multiple frames, which, when subjected to an average pooling operation, produces aggregated features that capture the historical appearance nuances of each subject [18]. This innovation not only ensures that the most recent and pertinent features are always in play but also amplifies the reliability of associations. By pooling these features, our algorithm achieves a holistic representation, adeptly handling transient appearance changes, momentary occlusions, or drastic appearance shifts — a marked enhancement over traditional methodologies.

Attention-based feature extraction with ViT: DeepSort, like many tracking algorithms, has conventionally leaned on CNNs for feature extraction. While CNNs have

been instrumental in many computer vision tasks, in our experiments they occasionally missed the mark in scenarios demanding meticulous attention to minute details. Addressing this gap, we have integrated the Vision Transformer (ViT), an attention-centric model [15], supplanting the conventional CNN in DeepSort. The ViT, renowned for its self-attention mechanisms, shines in pinpointing subtle differences by zeroing in on vital image regions. This capability is paramount for our toddler tracking application, ensuring that even the most nuanced appearance variations are meticulously captured, offering a richer and more detailed feature set for association.

4. Experimentation Results

In the conducted experiments, multiple configuration setups were systematically assessed to understand the robustness and efficiency of the object tracking model. Each configuration was run across five different sub-scenes, generating individual results per sub-scene. To aggregate the results, the metrics from each configuration were averaged across all the sub-scenes, providing a comprehensive view of the model’s performance under varying conditions. This approach ensures that the derived insights and the comparative analysis are based on consistent and averaged data points, mitigating the impact of outlier sub-scenes on the overall evaluation.

The experimentation was structured around several focused scenarios, each emphasizing different aspects of the model’s parameters. One configuration concentrated on detection confidence, another on distance measures, the third on overlapping and intersection over union (IoU), and the last on age and budget of the tracks. These configurations were crafted to observe the impact of selective variation of parameters on the model’s outcomes and to deduce which parameters are crucial for optimizing performance. Interestingly, during the experimentation, it was observed that varying only the IoU or the confidence did not significantly alter the results, implying a degree of robustness in the model against these parameters. Most configurations yielded similar performance metrics, indicating that the model’s effectiveness is less sensitive to alterations in IoU and confidence values. This insight is instrumental in understanding the inherent stability of the model and guides further refinement and tuning of the model parameters.

4.1. Evaluation Criteria

Accurate evaluation of MOT algorithms has proven to be very difficult, because MOT is a complex task, requiring accurate detection, localization, and association over time. Generally, there are five types of errors in an MOT method: 1- false negative or misses when ground truth exists but the prediction is missed, 2-false positive when tracker prediction exists for no ground truth tracker, 3- merge or ID switch

when two or more object tracks are swapped, 4- deviation which measures the average distance between the predicted location of an object and its true location over time, and 5- fragmentation which shows a track suddenly stops getting tracked but the ground truth track still exists. It causes a false increment of identifier numbers. Figure 2 shows an example of ID switches and fragmentation, which are the most challenging errors, in MOT algorithms. In this figure, there is an ID switch between toddler 1 and toddler 3. Also, toddler 2 has gotten a new ID due to the fragmentation error.

Using the mentioned five types of errors, various evaluation metrics can be calculated. In this paper, HOTA [16] (higher order tracking accuracy) is considered as the primary metric. HOTA combines several sub-metrics that evaluate algorithms from different perspectives, providing a comprehensive assessment of algorithm performance. In addition to HOTA, we also include other well-established metrics, such as MOTA [12] (multiple object tracking accuracy) and IDF1 [24] (iterative and discriminative framework1). IDF1 reflects the association aspect of the tracker, while MOTA is primarily influenced by detection performance. However, HOTA explicitly measures both types of metrics and combines detection and association in a balanced way. It can be used as a single unified metric for ranking trackers.

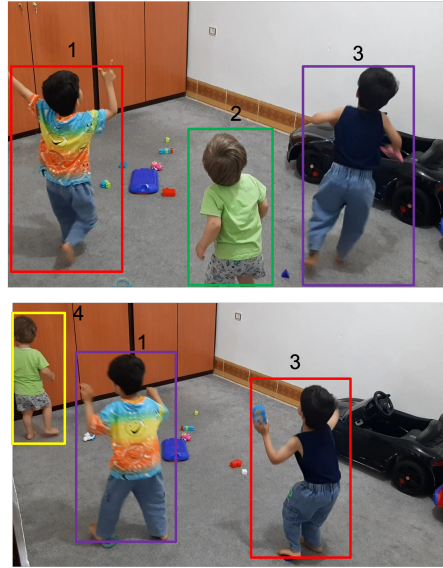


Figure 2. ID switch and fragmentation errors: toddler 1 and toddler 3 in the top image, have had their ID numbers swapped with each other in the bottom image, constituting an ID switch error. Toddler 2, present in the top image, is no longer tracked in the bottom image and is treated as a new toddler assigned the ID 4, indicative of a fragmentation error.

4.2. Building our MTTrack Dataset

The dataset building was a sophisticated and detailed endeavor, primarily concentrating on toddler videos, which

necessitated the precise and accurate labeling of the selected frames. To facilitate this intricate procedure, an innovative labeling technique was utilized, allowing for the efficient auditing and refinement of labels generated by an established MOT algorithm. This negated the need to initiate labeling from scratch for each frame, thus optimizing the process. To assure the integrity of the dataset and try to cancel biases towards specific algorithms, two distinguished MOT algorithms, StrongSort and DeepSort, were incorporated. The calculated average between the bounding boxes generated by these algorithms yielded unbiased and consistent labels, adding to the robustness of the dataset.

The MTTrack Dataset consists of recorded videos capturing three toddlers engaged in various activities within a room. These toddlers, aged 2-4 years, can be observed performing actions such as jumping, walking, sitting, and playing with tablets and toys. We formatted the dataset frames into 10 subscenes, each comprising a maximum of 300 frames. This methodological division was instrumental in eliminating sudden changes in scenery and mitigating extreme ID switches, thereby ensuring the consistency and reliability of the labels [4]. The auditing and validation of the labels were meticulously executed using the Labelme [25] open-source tool, enabling the verification of each label’s accuracy, relevance, and compliance with established standards. This comprehensive approach to labeling, while extensive, was imperative in establishing a reliable and credible foundation for subsequent research phases, yielding a dataset of unparalleled accuracy and reliability.

4.3. Experimentation Configurations

To ensure a consistent evaluation, 10 toddler video clips, each with 300 frames, were used in our experiments using the MTTrack dataset. Also, we applied different MOT methods on two public datasets: MO15, and DanceTrack. To have a comprehensive evaluation, various configurations were tested to examine object tracking algorithms. Each configuration is tailored to address specific challenges and requirements in object tracking, ranging from high precision and reliability to robustness against occlusions and appearance changes. A brief description of each configuration is listed below:

Configuration 1: This default configuration serves as a balanced setup suitable for general-purpose object-tracking scenarios.

Configuration 2: With a heightened `MIN_CONFIDENCE` of 0.7, this configuration is optimized to minimize false positives by considering only high-confidence detections.

Configuration 3: This configuration, with an increased `MAX_DIST` of 0.4 and `MAX_AGE` of 80, is tailored for scenarios where objects may change appearance significantly or be temporarily occluded.

Configuration 4: The reduced `NMS_MAX_OVERLAP` and `MAX_IOU_DISTANCE` of 0.3 in this setup makes it suitable for tracking smaller or thin objects in densely populated scenes.

Configuration 5: By allowing more overlap between bounding boxes and between detections and trackers, this configuration is suitable for tracking larger or blob-like objects where overlap is expected.

Configuration 6: This configuration targets challenging conditions with substantial appearance changes or noise, allowing even low-confidence detections due to a `MIN_CONFIDENCE` of 0.3 and a `MAX_DIST` of 0.6.

Configuration 7: This configuration uses optimized parameters resulting from the proposed genetic algorithm. The GA algorithm’s configuration was number of generations is 50, the pop size is 10, the mutation rate is 0.1, and the crossover rate is 0.7.

4.4. Quantitative Comparison

To conduct a thorough assessment, we carried out a series of experiments from various perspectives. Firstly we examined different configurations of DeepSort and compared them with the configuration resulting from the proposed genetic algorithm. Table 2 shows the results of this experiment in terms of HOTA (higher order tracking accuracy), DetRe (detection recall), DetPr (detection precision), DetA (detection accuracy), and MOTA (multiple object tracking accuracy). As can be seen in this table, configuration 7 which is the result of parameter optimization of the genetic algorithm achieved the best results.

Table 2. Accuracy parameters in different configurations. Each configuration is designed to address a specific challenge. Configuration 7 with parameters resulting from the proposed genetic algorithm has achieved the best accuracy.

Config.	MOTA	DetRe	DetPr	DetA	HOTA
1	0.94	0.96	0.92	0.89	0.56
2	0.95	0.97	0.94	0.91	0.62
3	0.94	0.94	0.94	0.89	0.58
4	0.92	0.89	0.95	0.85	0.59
5	0.92	0.89	0.95	0.85	0.66
6	0.91	0.87	0.94	0.82	0.60
7	0.95	0.92	0.99	0.91	0.67

Table 3. A quantitative comparison between different algorithms on indoor and outdoor environments. The proposed method (MTTSort) has achieved the best results on the MTTrack dataset. DeepSort+GA shows a configuration of DeepSort resulting from the genetic algorithm. DeepSort+GA improves the HOTA and IDF1 on the MTTrack dataset significantly in comparison with the traditional DeepSort.

MOT Algorithms	Outdoor (MOT15)			Indoor (DanceTracker)			Indoor (MTTrack)		
	MOTA	HOTA	IDF1	MOTA	HOTA	IDF1	MOTA	HOTA	IDF1
DeepSort [31]	0.77	0.78	0.86	0.79	0.33	0.49	0.94	0.56	0.82
StrongSort [5]	0.74	0.76	0.85	0.75	0.31	0.47	0.91	0.47	0.78
HybridSort [33]	0.77	0.80	0.88	0.90	0.48	0.54	0.87	0.20	0.86
Bytetrack [35]	0.76	0.70	0.80	0.80	0.47	0.52	0.96	0.52	0.96
DeepSort+ GA	0.69	0.57	0.62	0.65	0.39	0.22	0.95	0.67	0.97
MTTSort (Ours)	0.66	0.59	0.67	0.72	0.43	0.23	0.98	0.68	0.98

Table 3 displays the results of benchmarked techniques when compared to our proposed method. Within this table, "DeepSort+GA" denotes an enhanced DeepSort configuration incorporating the proposed genetic algorithm. As depicted in the table, our method has delivered superior performance, particularly in the context of indoor videos, specifically excelling in the tracking of multiple toddlers, as evident in the MTTrack dataset. We also found out that the best performing parameters without the aggregated features were different from the ones we got after adding the ViT attention model and the aggregated features. Especially, the `N_INIT` which decreased, suggesting that tracks are now initialized with fewer frames. This could be because the new features provide more distinct and reliable information early on.

In our experimental observations of HybridSort, a noteworthy drawback we encountered is its inconsistency in performance. As it can be seen in Table 3, this approach delivers exceptionally good results when toddlers are active and moving. However, it faces considerable difficulties when toddlers remain still or display minimal movement, resulting in tracking errors. Based on the experimental results and the flexibility for customization and parameter adjustment offered by various MOT methods, we have determined to adopt the DeepSort algorithm as the baseline framework for our proposed method designed for multiple toddler tracking in indoor video footage.

5. Conclusion

This paper discussed the primary challenges of multiple object tracking methods for tracking toddlers in indoor videos. We then introduced a new tracking method named "MTTSort," which is designed for multiple toddler tracking. In the initial phase of MTTSort, we employed a genetic algorithm to estimate optimized tracking parameters. By melding our custom feature buffer and the ViT-based feature extraction, we've re-engineered the foundational components of the DeepSort algorithm in order to capture the temporal features of each subject. This rejuvenated algorithm underwent rigorous evaluation and benchmarked against performance metrics like MOTA, HOTA, and IDF1

on the collected MTTrack dataset and two public tracking datasets, MOT15, and DanceTrack.

Looking ahead, our research will focus on addressing additional challenges in multiple toddler tracking, including scenarios involving action figures, crawling subjects, and twins detection and tracking. Furthermore, we plan to expand our research to encompass multi-view videos. This expansion will include work on multi-camera tracking and re-identification methods. Our ultimate goal is to implement our method in real-world multi-camera systems for tasks such as detection, tracking, action recognition of toddlers, and the prediction of potentially hazardous events in indoor videos.

References

- [1] Imran Ahmed, Sadia Din, Gwanggil Jeon, Francesco Piccialli, and Giancarlo Fortino. Towards collaborative robotics in top view surveillance: A framework for multiple object tracking by detection using deep learning. *IEEE/CAA Journal of Automatica Sinica*, 8(7):1253–1270, 2021. 1
- [2] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Uppcroft. Simple online and realtime tracking. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 3464–3468, 2016. 1, 2
- [3] Patrick Dendorfer, Aljosa Osep, Anton Milan, Konrad Schindler, Daniel Cremers, Ian Reid, Stefan Roth, and Laura Leal-Taixé. Motchallenge: A benchmark for single-camera multiple target tracking. *International Journal of Computer Vision*, 129:845–881, 2021. 2, 3
- [4] Zhaopeng Dou, Zhongdao Wang, Yali Li, and Shengjin Wang. Identity-seeking self-supervised representation learning for generalizable person re-identification. In *ICCV*, 2023. 7
- [5] Yunhao Du, Zhicheng Zhao, Yang Song, Yanyun Zhao, Fei Su, Tao Gong, and Hongying Meng. Strongsort: Make deepsort great again. *IEEE Transactions on Multimedia*, 2023. 2, 8
- [6] Xiaohui Gong, Xiao Li, Li Ma, Weilin Tong, Fangyu Shi, Menghan Hu, Xiao-Ping Zhang, Guangjun Yu, and Cheng Yang. Preterm infant general movements assessment via representation learning. *Displays*, 75:102308, 2022. 1
- [7] Diego Gragnaniello, Antonio Greco, Alessia Saggese, Mario Vento, and Antonio Vicinanza. Benchmarking 2d multi-object detection and tracking algorithms in autonomous vehicle driving scenarios. *Sensors*, 23(8), 2023. 1

- [8] Hanife Guney, Melek Aydin, Murat Taskiran, and Nihan Kahraman. A deep neural network based toddler tracking system. *Concurrency and Computation: Practice and Experience*, 34(14):e6636, 2022. 1
- [9] Hanife Guney, Melek Aydin, Murat Taşkıran, and Nihan Kahraman. Toddler tracking system with face recognition and object tracking using deep neural network. In *2020 International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*, pages 1–6, 2020. 1
- [10] Xiaofei Huang, Lingfei Luan, Elaheh Hatamimajoumerd, Michael Wan, Pooria Daneshvar Kakhaki, Rita Obeid, and Sarah Ostadabbas. Posture-based infant action recognition in the wild with very limited data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4912–4921, June 2023. 1
- [11] Xiaofei Huang, Michael Wan, Lingfei Luan, Bethany Tunik, and Sarah Ostadabbas. Computer vision to the rescue: Infant postural symmetry estimation from incongruent annotations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1909–1917, January 2023. 1
- [12] Rainer Stiefelhagen Keni Bernardin. Evaluating multiple object tracking performance: The clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008:246309,, 2008. 4, 6
- [13] Annu Lambora, Kunal Gupta, and Kriti Chopra. Genetic algorithm- a literature review. In *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, pages 380–384, 2019. 4
- [14] Marco Leo, Giuseppe Massimo Bernava, Pierluigi Carcagnì, and Cosimo Distanto. Video-based automatic baby motion analysis for early neurological disorder diagnosis: State of the art and future directions. *Sensors*, 22(3), 2022. 1
- [15] Rujia Li, Junya Liu, Zhen Yang, Xin Zhou, and Zhijian Yin. Attention-based multi-scale vit fine-grained visual classification. In *ACM International Conference Proceeding Series*, 2022. 6
- [16] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. Hota: A higher order metric for evaluating multi-object tracking. *International journal of computer vision*, 129:548–578, 2021. 4, 6
- [17] Aprinaldi Jasa Mantau, Irawan Widi Widayat, Yudhi Adhitya, S. W. Prakosa, Jenq-Shiou Leu, and M. Koppen. A ga-based learning strategy applied to yolov5 for human object detection in uav surveillance system. *IEEE International Conference on Automation and Computing (ICAC)*, 2022. URL: [Link](https://dx.doi.org/10.1109/icca54724.2022.9831954). 4
- [18] Dmitrii Marin, Jen-Hao Rick Chang, Anurag Ranjan, Anish Prabhu, Mohammad Rastegari, and Oncel Tuzel. Token pooling in vision transformers for image classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 12–21, 2023. 5
- [19] Dimitrios Meimetus, Ioannis Daramouskas, Isidoros Perikos, and Ioannis Hatzilygeroudis. Real-time multiple object tracking using deep learning methods. *Neural Computing and Applications*, 35:89–118, 2021. 1
- [20] Maria Eleonora Minissi, Lucía Gómez-Zaragozá, Javier Marín-Morales, Fabrizia Mantovani, Marian Sirera, Luis Abad, Sergio Cervera-Torres, Soledad Gómez-García, Irene Alice Chicchi Giglioli, and Mariano Alcañiz. The whole-body motor skills of children with autism spectrum disorder taking goal-directed actions in virtual reality. *Frontiers in Psychology*, 14, 2023. 1
- [21] Iason-Ioannis Panagos, Angelos P. Giotis, and Christophoros Nikou. Multi-object visual tracking for indoor images of retail consumers. In *2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*, pages 1–5, 2022. 2
- [22] Ricardo Pereira, Guilherme Carvalho, Luís Garrote, and Urbano J. Nunes. Sort and deep-sort based multi-object tracking for mobile robotics: Evaluation with new data association metrics. *Applied Sciences*, 12(3), 2022. 1
- [23] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2
- [24] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In Gang Hua and Hervé Jégou, editors, *Computer Vision – ECCV 2016 Workshops*, pages 17–35, Cham, 2016. Springer International Publishing. 4, 6
- [25] Bryan C Russell, Antonio Torralba, Kevin P Murphy, and William T Freeman. Labelme: a database and web-based tool for image annotation. *International journal of computer vision*, 77:157–173, 2008. 7
- [26] Peize Sun, Jinkun Cao, Yi Jiang, Zehuan Yuan, Song Bai, Kris Kitani, and Ping Luo. Dancetrack: Multi-object tracking in uniform appearance and diverse motion. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20961–20970, 2022. 2, 3
- [27] Thangaswamy Judi Vennila and Vanniappan Balamurugan. A rough set framework for multihuman tracking in surveillance video. *IEEE Sensors Journal*, 23(8):8753–8760, 2023. 1
- [28] Michael Wan, Xiaofei Huang, Bethany Tunik, and Sarah Ostadabbas. Automatic assessment of infant face and upper-body symmetry as early signs of torticollis. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–6, 2023. 1
- [29] Michael Wan, Shaotong Zhu, Lingfei Luan, Gulati Prateek, Xiaofei Huang, Rebecca Schwartz-Mette, Marie Hayes, Emily Zimmerman, and Sarah Ostadabbas. Infanface: Bridging the infant–adult domain gap in facial landmark estimation in the wild. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 4486–4492. IEEE, 2022. 1, 2
- [30] Xu-na Wang and Qing-mei Tan. Dan: a deep association neural network approach for personalization recommendation. *Frontiers of Information Technology & Electronic Engineering*, 21(7):963–980, 2020. 2

- [31] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3645–3649, 2017. [1](#), [2](#), [8](#)
- [32] Fan Yang, Shigeyuki Odashima, Shoichi Masui, and Shan Jiang. Hard to track objects with irregular motions and similar appearances? make it easier by buffering the matching space. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4788–4797, 2023. [1](#)
- [33] Mingzhan Yang, Guangxin Han, Bin Yan, Wenhua Zhang, Jinqing Qi, Huchuan Lu, and Dong Wang. Hybrid-sort: Weak cues matter for online multi-object tracking. *arXiv preprint arXiv:2308.00783*, 2023. [2](#), [3](#), [8](#)
- [34] Ilham Ari Elbaith Zaeni, Siti Sendari, Dyah Lestari, Yogi Dwi Mahandi, Mahfud Jiono, and M. Syaifuddin. An implementation of multi-object tracking using omnidirectional camera for trash picking robot. In *2018 Electrical Power, Electronics, Communications, Controls and Informatics Seminar (EECCIS)*, pages 154–158, 2018. [1](#)
- [35] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 1–21, Cham, 2022. Springer Nature Switzerland. [2](#), [3](#), [8](#)
- [36] Shaotong Zhu, Michael Wan, Elaheh Hatamimajoumerd, Kashish Jain, Samuel Zlota, Cholpady Vikram Kamath, Cassandra B. Rowan, Emma C. Grace, Matthew S. Goodwin, Marie J. Hayes, Rebecca A. Schwartz-Mette, Emily Zimmerman, and Sarah Ostadabbas. A video-based end-to-end pipeline for non-nutritive sucking action recognition and segmentation in young infants. In Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 586–595, Cham, 2023. Springer Nature Switzerland. [1](#)