Tracking using Human Pose Matching with Deep Association Metric

Atishay Jain, Abhishek Dhiman, Balakrishna Pailla

Vellore Institute of Technology, Reliance Jio, Reliance Jio, India
atishaysjain@gmail.com, Abhishek1.dhiman@ril.com, Balakrishna.Pailla@ril.com

Abstract

This paper proposes a novel approach to track multiple people utilizing skeletal information combined with visual appearance features to improve the accuracy of tracking people across different frames of a video. We extracted the appearance feature vectors and skeletal feature vectors for each detected person in every frame. Each individual was tracked by considering the cosine distance between the skeletal feature vectors along with the euclidean distance between the appearance feature vectors across different frames of a video. This reduces the dependency of the tracker over appearances of people thus making it more consistent, especially in videos with people having similar appearances such as sports videos with players wearing similar jerseys. The stance of an individual in continuing frames is expected to be similar considering the high frame rate of modern camera devices. Therefore it is befitting to consider skeletal features along with appearance features for tracking. Our paper is an incremental paper demonstrating improvement over SORT with a deep association metric approach(Wojke et al., 2017). Our approach utilizing skeletal information combined with visual appearance information returns better MOT results on the MOT17 dataset using the yolov3 detector.

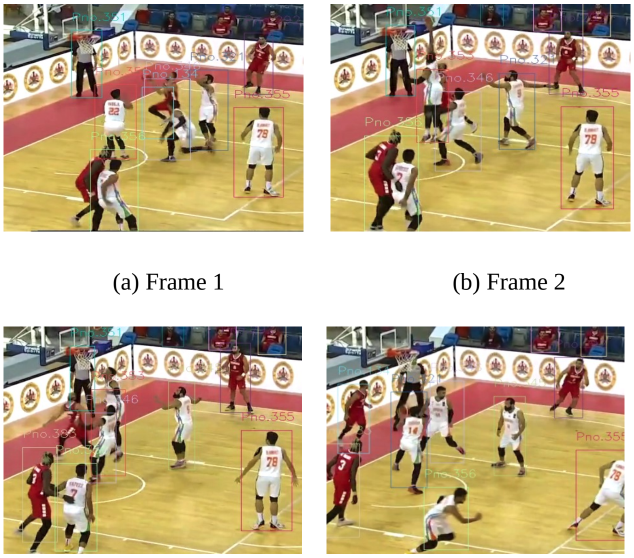
Introduction

The problem statement of Multiple Object Tracking(MOT) constitutes assigning unique identifiers to each object (Human person in our case) and preserving their consistency throughout the frames of a given video.

MOT is one of the primary topics in the field of computer vision which finds its application in domains like surveillance, trajectory analysis, and autonomous driving. MOT continues to be a challenging task. Despite some serious advancements in recent years, MOT has substantial potential for improvement and innovation.

MOT is dominantly seen as a streak of 2 tasks, detection and tracking. This is known as tracking-by-detection paradigm. The first step i.e. detection involves a detector that creates bounding boxes containing locations of each detected person in that frame. The second step i.e. tracking involves associating these detected people across frames to track individuals throughout a video.

This approach involves different modules with specific jobs working with each other. As a result, optimization efforts are concentrated on strengthening the targeted disciplines of various modules and their compatibility with one another to provide more consistent results.



(a) Frame 1

(b) Frame 2

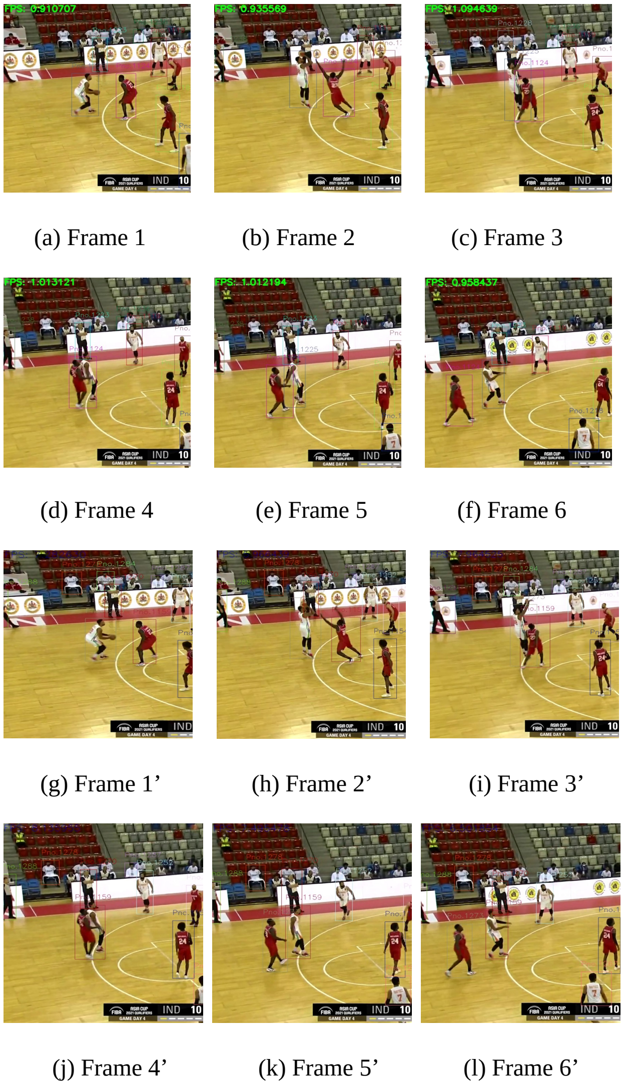
(c) Frame 3

(d) Frame 4

Figure 1: The person with track id ‘134’ (player in red jersey trying to score a basket) is detected in Frame 1 but got occluded and was not detected in Frame 2 and Frame 3. Subsequently, he is once again detected in frame 4 and reassigned the track id ‘134’, thus preserving his identity across frames.

The goal of our research is to enhance SORT with a deep association metric(Wojke et al., 2017) approach of MOT by adding a module that extracts key-points to depict a skeleton-based representation of a person. This skeleton-based representation fundamentally represents their pose. We have used this pose along with a person’s visual appearance features for data association and hence to associate people across frames.

The upside of our method is that it mitigates the reliance on visual appearance elements, making it more suitable for MOT on videos with individuals having similar facial appearances and clothing. Refer to Figure 1 and Figure 2 for illustration. We demonstrated that considering additional modules that give data of detected people for data association across frames can assist improve tracking accuracy, and further study in this area should be pursued.



(a) Frame 1 (b) Frame 2 (c) Frame 3

(d) Frame 4 (e) Frame 5 (f) Frame 6

(g) Frame 1’ (h) Frame 2’ (i) Frame 3’

(j) Frame 4’ (k) Frame 5’ (l) Frame 6’

Figure 2: In the first sequence of frames, the player with track id 1225 (wearing white jersey) shoots the ball and gets occluded by the player with track id 1124 (wearing red jersey) while doing so. In frame 4 the shooter is not detected, yet in frame 5 correct ids have been assigned to both the shooter(1225) and the occluder(1224). While in the second sequence of frames, there is an id switch in frame 5’ between the red jersey wearing shooter(1273) and the occluder(1159) in white jersey. In the first sequence, we have used our proposed tracker while in the second sequence of frames, we have used the tracker proposed by (Wojke et al., 2017).

In summary, the paper contributes the following salient features to the existing approach to further improve the accuracy :

1) Incorporation of a module that extracts the skeletal representation of a person. We use this skeletal representation in conjunction with visual features for data association across frames.

2) Visual features, Skeletal features, Kalman Filtering, Hungarian Assignment Algorithm, and Siamese Graph Convolution Network were used for object tracking.

3) Outperformed (Wojke et al., 2017) on MOT17 dataset(Milan et al., 2016) using yolov3(Redmon and Farhadi, 2018) detector.

4) Put forward an idea that tracking accuracy could be improved by extracting more attributes of an individual that describes detections for associating data across frames.

Related Work

Initially, MOT was achieved by using Multiple Hypothesis Tracking(MHT)(Reid, 1979). Probability is calculated for every detection of being either an already present object, a newly introduced object, or just a false detection. Kalman filter is used for estimating the target state from each such data association hypothesis. As more detections are received, the probabilities are again calculated. This gives us the information we need to correlate detections across states. Unlikely hypotheses are eliminated with the help of Kalman filtering and hypotheses with similar target state estimates are combined. The technique described above is repeated recursively. However, as the number of objects being tracked grows, the combinatorial complexity grows exponentially. As a result, it becomes unsuitable for usage in dense and dynamic systems with a high number of objects. This approach was used in ocean surveillance, air traffic control, systems to defend against ballistic missiles, and battlefield surveillance.

Joint Probabilistic Data Association(JPDA)(Fortmann et al., 1983) algorithm calculates joint posterior association probabilities for multiple targets in Poisson clutter. Joint association probabilities are used for estimating target state and hence tracking targets. When a high number of objects appear in the system, this approach, like MHT, becomes inefficient since it gets exponentially more computationally expensive to run with an increase in the objects detected in the system. (Rezatofighi et al., 2015) makes the JPDA algorithm feasible by decreasing the computational time significantly. This is achieved by reformulating the calculation of individual JPDA assignment scores and approximating the joint score by the m-best solutions using a binary tree partition method.

(Kim et al., 2015) further improves MHT by introducing a way for which each track hypothesis trains an online appearance model. To improve the object tracking accuracy, they have taken into account both the appearance and the motion information.

(Geiger et al., 2013) presents a probabilistic model for comprehending multi-object traffic scenarios from mobile platforms. Tracklets are defined as detections seen from a bird’s eye view. Hungarian algorithm(Kuhn, 1955) is used to associate data across two consecutive frames. Affinity matrix contains appearance and geometric cues of objects. The geometric cue is the detected bounding box’s intersection over union(IOU) score. The appearance information is the correlation of these detected bounding boxes. To make up for any localization ambiguity, 20% additional area is considered when comparing bounding boxes.

“Simple Online And Realtime Tracking (SORT)”(Bewley et al., 2016) simplifies data association into a single-step task. Here the bounding box’s coordinates for each target are predicted for the present frame using Kalman filtering. The affinity matrix is formed from the IOU distances between the detected bounding box’s coordinates and the predicted bounding box’s coordinates. Any assignments between targets and detections with less than the minimum threshold IOU distance intersection are rejected using a minimum IOU distance as a gate. The Hungarian algorithm(Kuhn, 1955) is then used to allocate targets to tracks in the most efficient way possible.

“SORT with a deep association metric”(Wojke et al., 2017) incorporated appearance data to SORT(Bewley et al., 2016) to improve its tracking performance. It mitigates the high dependence of SORT on state estimation by Kalman filtering thus reducing the quantity of identity switches on videos with non-linear object motion.

We have taken “SORT with a deep association metric”(Wojke et al., 2017) one step further by taking skeletal information into account along with the appearance information while constructing an affinity matrix for data association. In the following section, we will discuss our approach in greater detail.

Lighttrack(Ning et al., 2020) only uses skeletal information for tracking. The Lighttrack tracker gives too much importance to spatial consistency which may result in inaccurate tracking while handling videos with a lot of movement, relatively low fps, and multiple people having similar poses.

(Braso and Leal-Taix´e, 2020) also applied learning to the data association step unlike all the approaches discussed till now where the learning process was confined to the phase where features are extracted. They propose a message passing network(MPN) for feature learning and final solution prediction.

Proposed Methodology

Skeletal Feature Matching

The pose estimator that we have deployed to detect pose is based on the MobileNet v1(Howard et al., 2017) architecture and has been taken from (Ning et al., 2020) which prioritizes speed over accuracy so that our tracker doesn’t become slow.

We have employed the Siamese Graph Convolution Network(SGCN) proposed by (Ning et al., 2020) for comparing key points which represent the estimated spatial location of body joints. SGCN takes a vector of normalized key points as input where each key-point corresponds to a body joint coordinate(2D coordinate positioned at a body joint of an individual). The returned output is a 128-dimensional skeletal feature vector encoding the spatial connection among human key points (joints). We have estimated the closeness between two feature vectors by calculating the euclidean distance between them. The value of the euclidean distance between two skeletal feature vectors is inversely proportional to the similarity between the poses of two individuals encoded by them. The Euclidean distance between 2 vectors is calculated as :

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (y_i - x_i)^2} \qquad (1)$$

Where x_i and y_i are the i^{th} elements of their respective vectors. Figure 3 illustrates the above-explained process.

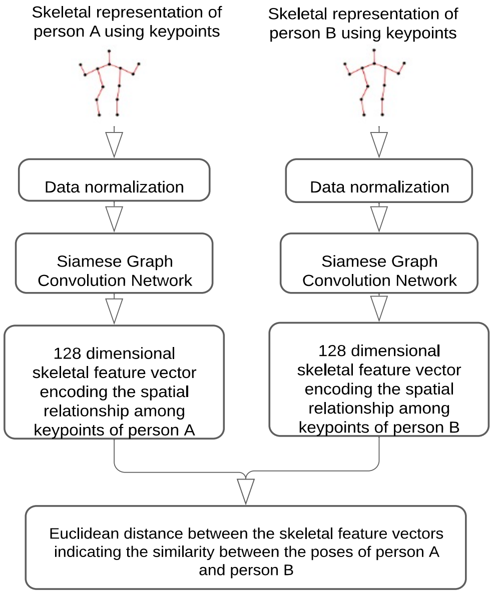

Skeletal representation of person A using keypoints
Skeletal representation of person B using keypoints
Data normalization
Data normalization
Siamese Graph Convolution Network
Siamese Graph Convolution Network
128 dimensional skeletal feature vector encoding the spatial relationship among keypoints of person A
128 dimensional skeletal feature vector encoding the spatial relationship among keypoints of person B
Euclidean distance between the skeletal feature vectors indicating the similarity between the poses of person A and person B

Figure 3: The process to compare two skeletal feature vectors.

Appearance Feature Matching

We have utilized a deep appearance descriptor from (Wojke et al., 2017) as an encoder to return a 128-dimensional appearance feature vector encoding the visual appearance of a detected individual. The architecture of the employed encoder is shown in Figure 4.

The cosine distance between two appearance feature vectors is then calculated. The cosine distance indicates the dissimilarity between the visual appearances of two detections. The formulae for computing the cosine distance is as follows :

$$\textit{cosine distance} = 1 - \textit{cosine similarity} \qquad (2)$$

$$cosine\ similarity(X,Y) = \frac{\sum_{i=1}^{n}(x_i * y_i)}{\sqrt{\sum_{i=1}^{n}(x_i)^2} * \sqrt{\sum_{i=1}^{n}(y_i)^2}} \qquad (3)$$

Where x_i and y_i are the ith elements of vectors X and Y respectively.

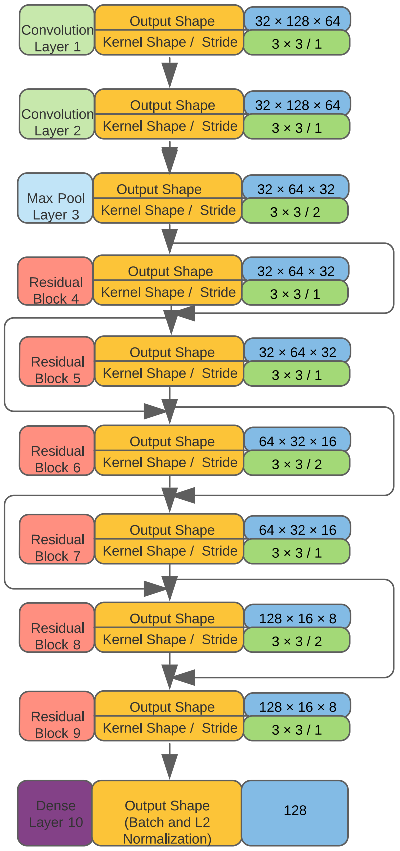
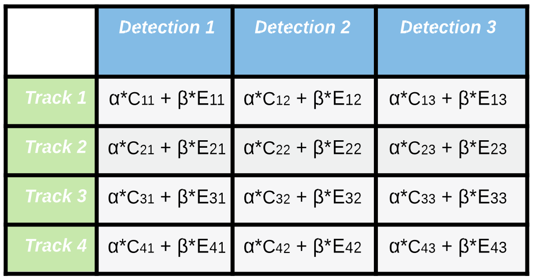


Figure 4: The architecture of the encoder employed to return a 128-dimensional appearance feature vector.

Affinity Matrix Construction

The novelty of our approach lies in the construction of the affinity matrix. An affinity matrix is constructed for every frame and is used to link tracks with bounding box detections where detections are the individuals detected in the current frame. Thus affinity matrix is used to assign track identities to the individuals detected in the current frame. The elements of the affinity matrix are calculated as a_{ij} = α * (cosine distance between appearance feature vectors of tracks and detections) + β * (euclidean distance between skeletal feature vectors of tracks and detections) where α and β are constants, i is the ith row corresponding to the ith track and j is the jth column corresponding to the jth detection. An example of an affinity matrix is shown in Figure 5. Intuitively the value of α is proportional to the weightage given to visual appearance features and the value of β is proportional to the weightage given to human pose (skeletal features) for associating data.



	Detection 1	Detection 2	Detection 3
Track 1	$\alpha*C_{11} + \beta*E_{11}$	$\alpha*C_{12} + \beta*E_{12}$	$\alpha*C_{13} + \beta*E_{13}$
Track 2	$\alpha*C_{21} + \beta*E_{21}$	$\alpha*C_{22} + \beta*E_{22}$	$\alpha*C_{23} + \beta*E_{23}$
Track 3	$\alpha*C_{31} + \beta*E_{31}$	$\alpha*C_{32} + \beta*E_{32}$	$\alpha*C_{33} + \beta*E_{33}$
Track 4	$\alpha*C_{41} + \beta*E_{41}$	$\alpha*C_{42} + \beta*E_{42}$	$\alpha*C_{43} + \beta*E_{43}$

Figure 5: An example of an affinity matrix utilized to assign track ids by linking tracks and detections. In this example, we have 4 tracks and 3 bounding box detections. Here Cij is the cosine distance between the appearance feature vector of track i and detection j, while Eij is the euclidean distance between the skeletal feature vectors of track i and detection j.

Kalman Filtering

Our approach of Kalman filtering has been inspired by (Wojke et al., 2017). We use Kalman filtering to reject the association between tracks and bounding box detections with large differences between their respective states. We define the state by the following 8-dimensional vector :

$$[x, y, r, h, x^0, y^0, r^0, h^0]$$

where x and y are the coordinates of the bounding box center, r is the ratio between the width and height of the bounding box, h is the bounding box height, (x',y',r' and h' are the respective velocities of x,y,r and h). The difference between tracks and bounding box detection states is computed as the square of Mahalanobis distance between (mean, covariance) of tracks and a 4-dimensional vector [x,y,r,h] of detections. Here :
i) mean is an 8 dimensional mean vector of the track state elements ii) covariance is an 8*8 dimensional matrix that represents the covariance of a track's state elements iii) [x,y,r,h] is [(x coordinate of the bounding box's center),(y coordinate of the bounding box's center),(ratio between the bounding box's width and height),(bounding box's height)].

If the calculated squared Mahalanobis distance doesn't lie within a 95% confidence interval range calculated from the inverse Chi-squared distribution, the association between the track and the detection is rejected. Therefore if the squared Mahalanobis distance is greater than a minimum threshold gating distance of 9.4877, the association between the track and the detection is rejected.

Linking Detections and Tracks using the Hungarian Algorithm

We apply the Hungarian algorithm(Kuhn, 1955) over the affinity matrix for track-detection assignment and thus for linking tracks with detections.

Linking Cascade

The above-explained process for a given frame occurs over batches of tracks as proposed in (Wojke et al., 2017). Tracks are grouped into batches based on the number of frames since they were most recently linked with a detection. Thus the first batch will contain the tracks that were linked with detections in the frame before the current frame. The second batch will contain the tracks that were most recently linked with detections two frames before the current frame, and so on. Therefore the tracks that were linked more recently will have a higher chance of being linked with a detection of the current frame. As the inaccuracy in the state estimation by Kalman filtering will be higher for tracks that are less recently linked with a detection, it is appropriate that less recently linked tracks are given a lower preference for being linked with a detection of the current frame.

Assignment using Intersection over Union scores

We then associate the remaining unlinked tracks and detections by running the Hungarian algorithm over an assignment cost matrix with intersection over union scores between the current frame bounding box detections and predicted bounding boxes of unlinked tracks.

Experimentation And Results

Experimentation

We have used yolov3(Redmon and Farhadi, 2018) as the detector. As explained above, the coefficients of the feature vectors in an affinity matrix are α and β. The values of α and β are proportional to the importance given to visual appearance and skeletal features respectively for associating tracks with current frame bounding box detections, effectively tracking individuals. To determine the most appropriate values of α and β .i.e. to determine how much weightage should be given to visual appearance and skeletal features respectively, we experimented by employing our tracker over the MOT17 dataset(Milan et al., 2016) and varying the values of α and β, thus summarily varying the importance given to visual and skeletal features for tracking. Table 1 shows the results of our experimentation.

We have considered HOTA(Luiten et al., 2021) i.e. Higher Order Tracking Accuracy as the primary metric to rank a tracker.

α	β	HOTA	MOTA	MOTP	IDF1	IDs
0.0[a]	1.0[a]	29.783	29.655	76.451	35.705	1023
0.1	0.9	30.127	29.681	76.445	36.008	1023
0.2	0.8	30.121	29.682	76.444	35.964	1020
0.3	0.7	30.353	29.695	76.457	36.155	1008
0.4	0.6	30.454	29.709	76.457	36.347	1005
0.5	0.5	30.441	29.711	76.456	36.165	993
0.6	0.4	30.430	29.686	76.446	36.197	1008
0.7	0.3	30.314	29.682	76.452	36.013	1011
0.8	0.2	30.199	29.671	76.443	35.718	1002
0.9	0.1	30.159	29.670	76.427	35.769	1008
1.0[b]	0.0[b]	30.302	29.681	76.441	35.895	1029

[a] $\alpha = 0.0$ and $\beta = 1.0$(first row) does not take visual appearance features into account for tracking. [b] $\alpha = 1.0$ and $\beta = 0.0$(last row) does not take human pose features into account for tracking.

Table 1: MOT accuracy metric results over different values of α (proportionate to the importance given to appearance features for tracking) and β (proportionate to the importance given to skeletal features for tracking).

HOTA has been chosen as the primary metric because it measures how consistently the same id has been assigned to the detection corresponding to a particular individual across frames against the ground truth identity links. HOTA also considers the spatial alignment between each predicted detection and each ground truth detection, and the overall consistency between the set of all predicted and ground truth detections. Our tracker returns the best HOTA score with $\alpha = 0.4$ and $\beta = 0.6$, thus we have fixed these values.

Results

The evaluation results of our tracker as compared to Deep Sort(Wojke et al., 2017) are shown in Table 2. Table 2 shows that our tracker outperforms (Wojke et al., 2017) with an increase in HOTA score by 0.152% and a decrease in identity switches by 2.33%. Our tracker returns better results than (Wojke et al., 2017) when evaluated by HOTA, MOTA, MOTP, IDF1, PT, ML metrics.

Tracker	HOTA	MOTA	MOTP	IDF1	IDs
Our Tracker	30.454	29.709	76.457	36.347	1005
(Wojke et al., 2017)	30.302	29.681	76.441	35.895	1029

Table 2: Comparing tracking results of our tracker with (Wojke et al., 2017) over MOT17 Dataset(Milan et al., 2016). Both the trackers use the yolov3(Redmon and Farhadi, 2018) detector.

We would like the readers to note that our tracker obtained a HOTA score of 40.736 over the MOT17 benchmark using the officially provided detection coordinates. The officially provided detection coordinates were obtained using

SDP(Yang et al., 2016), Faster-RCNN(Ren et al., 2015), and DPM(Felzenszwalb et al., 2009) detectors. We decided to use the yolov3 detector to obtain detection coordinates instead of the already provided official coordinates because the yolov3 detector was returning more accurate detections.

Considering skeletal features in addition to appearance features for tracking reduces the dependency of our tracker on the appearances of people making it more reliable, especially in videos with people having similar appearances. Therefore we expect our tracker to significantly outperform (Wojke et al., 2017) on sports videos as players on the same team wear the same jerseys which results in a similar appearance. We tested our tracker on a six-and-half minute basketball match between India and Lebanon. The total number of track identities assigned by our tracker in the basketball match was 1945 while (Wojke et al., 2017) assigned 2029 track identities for the same basketball match. Our tracker assigned 4.14% fewer track identities indicating that the occurrence of identity switches is reduced by roughly 4% upon using our tracker.

Conclusion

We have put forward an extension of (Wojke et al., 2017) by considering human pose information along with appearance information for linking tracks and detections, making our tracker less sensitive towards visual appearance features. The human pose is an appropriate criterion to consider for data linkage as the human pose changes minimally across continual frames of a video, given the high frame rate of modern videos. Our tracker outperformed (Wojke et al., 2017) indicating that consideration of additional features for data association in tracking people across multiple frames results in improved accuracy.

References

Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In 2016 IEEE international conference on image processing (ICIP), pages 3464–3468. IEEE, 2016.

Andreas Geiger, Martin Lauer, Christian Wojek, Christoph Stiller, and Raquel Urtasun. 3d traffic scene understanding from movable platforms. IEEE transactions on pattern analysis and machine intelligence, 36(5):1012–1025, 2013.

Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861, 2017.

Anton Milan, Laura Leal-Taixe, Ian Reid, Stefan Roth, and´ Konrad Schindler. Mot16: A benchmark for multi-object tracking. arXiv preprint arXiv:1603.00831, 2016.

Chanho Kim, Fuxin Li, Arridhana Ciptadi, and James M Rehg. Multiple hypothesis tracking revisited. In Proceedings of the IEEE international conference on computer vision, pages 4696–4704, 2015.

Donald Reid. An algorithm for tracking multiple targets. IEEE transactions on Automatic Control, 24(6):843–854, 1979.

Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In European conference on computer vision, pages 17–35. Springer, 2016.

Fan Yang, Wongun Choi, and Yuanqing Lin. Exploit all the layers: Fast and accurate cnn object detector with scale dependent pooling and cascaded rejection classifiers. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2129–2137, 2016.

Guanghan Ning, Jian Pei, and Heng Huang. Lighttrack: A generic framework for online top-down human pose tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pages 1034–1035, 2020.

Guillem Braso and Laura Leal-Taix´ e.´ Learning a neural solver for multiple object tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6247–6257, 2020.

Harold W Kuhn. The hungarian method for the assignment problem. Naval research logistics quarterly, 2(1-2):83–97, 1955.

Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixe, and Bastian Leibe. Hota:´ A higher order metric for evaluating multi-object tracking. International journal of computer vision, 129(2):548–578, 2021.

Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767, 2018.

Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. EURASIP Journal on Image and Video Processing, 2008: 1–10, 2008.

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. IEEE, 2017.

Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. IEEE transactions on pattern analysis and machine intelligence, 32(9):1627–1645, 2009.

Seyed Hamid Rezatofighi, Anton Milan, Zhen Zhang, Qinfeng Shi, Anthony Dick, and Ian Reid. Joint probabilistic data association revisited. In Proceedings of the IEEE international conference on computer vision, pages 3047– 3055, 2015.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 28:91–99, 2015.

Thomas Fortmann, Yaakov Bar-Shalom, and Molly Scheffe. Sonar tracking of multiple targets using joint probabilistic data association. IEEE journal of Oceanic Engineering, 8 (3):173–184, 1983.