HEAD BASED TRACKING

Mohib Ullah[1], Maqsood Mahmud[2], Habib Ullah[3], Kashif Ahmad[4],
Ali Shariq Imran[1], Faouzi Alaya Cheikh[1]

[1] Department of Computer Science (IDI), Norwegian University of Science and Technology, Norway.
[2] Department of Management Information System,
University of Imam Abdulrahman Bin Faisal, Dammam, Saudi Arabia.
[3] College of Computer Science and Engineering, University of Ha'il, Saudi Arabia.
[4] College of Science and Engineering, Hamad Bin Khalifa University (HBKU), Doha, Qatar.

ABSTRACT

For tracking multiple targets in a scene, the most common approach is to represent the target in a bounding box and track the whole box as a single entity. However, in the case of humans, the body goes through complex articulation and occlusion that severely deteriorate the tracking performance. In this paper, we argue that instead of tracking the whole body of a target, if we focus on a relatively rigid body organ, better tracking results can be achieved. Based on this assumption, we followed the tracking-by-detection paradigm and generated the target hypothesis of only the spatial locations of heads in every frame. After the localization of head location, a constant velocity motion model is used for the temporal evolution of the targets in the visual scene. For associating the targets in the consecutive frames, combinatorial optimization is used that associates the corresponding targets in a greedy fashion. Qualitative results are evaluated on four challenging video surveillance dataset and promising results has been achieved.

Index Terms— combinatorial optimization, tracking-by-detection, video surveillance.

1. INTRODUCTION

One of the primary tasks of machine learning is to enable computers to learn from the data and automatically do thoughtful predictions. Such capabilities has applications in airline scheduling [1], cybersecurity [2], and fraud detection [3]. For the visual data, it helps to analyze a visual scene [4]. In the realm of visual scene analysis, multi-target tracking is one of the most important low-level computer vision problems that provides a backbone to many high level tasks like autonomous driving [5, 6], action recognition [7, 8], behavior analysis [9, 10], anomaly detection [8, 11], crowd management [12, 12], and sports players analysis [13, 14], to name a few. Even though tracking on its own is a low-level computer vision problem, intrinsically, it consists of other low-level tasks like object segmentation, object detection, and motion modeling. With the advancement in object detection algorithms [15, 16], the tracking-by-detection paradigm becomes the most suitable for tracking multiple objects in a visual scene. However, the biggest question that arises is which object part to track or track the whole object mass. Until now, almost all the tracking algorithms use the whole body detection [17–28]. For example, Milan et al. [21] proposed a highly non-convex cost function for multi-target tracking where different components like appearance, detection, motion, target mutual osculation, etc. are combined in a weighted average function. A gradient descent based optimization is used to optimize the cost and transdimentional jumps are used to avoid the local minima. Ullah et al. [20] proposed a bag of Bayesian filters to track multiple targets in the scene. Additionally, sparse coded deep features are incorporated to model the appearance of the targets. Schulter et al. [19] modeled multi-target tracking as a network flow graph. Instead of calculating the edge cost of the graph manually through handcrafted features, they learned the edge of the graph through back propagation. Similarly, Ullah et al. [28] also generated a directed acyclic graph for the multi-target tracking but used deep features for calculating the edge cost of the graph. Dehghan et al. [29] formulated the tracking problem as a multi clique problem. Initially, a graph is generated from a batch of frames and later each target trajectory is found as the maximum clique of the graph. Chu et al. [27] proposed a spatial-temporal attention mechanism for occulation handling and the interaction among different targets. They extract features from the different layers of CNN for modeling the appearance of the target. Compare to that, a Siamese neural network was proposed in [18] for modeling the appearance of targets and establishing the association between two targets.

One of the common attributes among all the tracking techniques is that they represent the whole human body as a rectangular rigid object, even though the human body goes through severe articulations. Compared to the standard approaches, in this paper, we focused on the relatively rigid or-
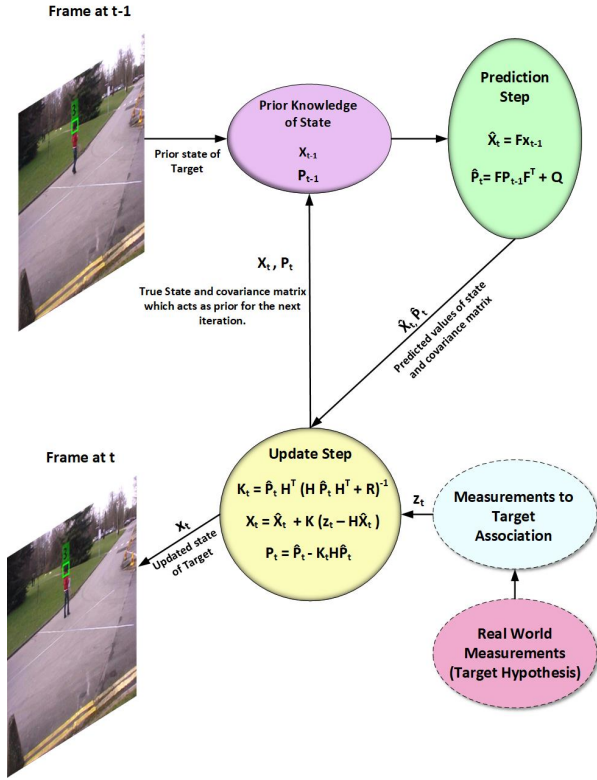
IS&T International Symposium on Electronic Imaging 2020
Intelligent Robotics and Industrial Applications using Computer Vision

074-1

Fig. 1: Prediction and the update step of the Kalman filter. It not only keeps track of the target state x_t but also the uncertainty P_t of the state. Initially, a prediction is made for the state of the target using the previously known state and the state transition model. In the update step, the measurement z_t is used to correct the predicted state. In an iterative process, the target is tracked between two consecutive frames.

gan of the human body i.e. head of a person and first generated the head hypothesis in all the frames and then used combinational optimization to associate the target head in the consecutive frames. The organization of the paper is the following: In section 2, the proposed approach is briefly explained. The tracker details are given in section 3. The qualitative and quantitative results are given in section 4 and section 5 concludes the paper with future directions.

2. PROPOSED APPROACH

The proposed approach is based on the tracking-by-detection paradigm. Initially, the target hypothesis is generated in every frame. In our case, the target hypothesis is the spatial position of the target head. In theory, any rigid body part can be used as the target hypothesis. But due to the most important and the most visible position of the head, we select it as the key location for tracking. Our tracker is based on the Kalman filter. The block diagram of the Kalman filter is given in Fig. 1. We assumed a smooth and constant velocity model for the

targets in the scene. This is a reasonable assumption because once a target appear in the visual scene, it can not disappear abruptly. Similarly, the motion of the target is smooth as long as it stays in the scene. We modeled the target association as a combinatorial optimization problem. Association is important as with every time step, we have N numbers of target hypothesis and M number of tracks. In order to track the targets as accurately as possible, the correct hypothesis should be assigned to the corresponding tracks. Hence, at every time instance, we produce a $N \times M$ matrix and used the Hungarian assignment algorithm to get the correct associations. A brief description of the Kalman Filter and the assignment algorithm is given in section 3 and 3.1, respectively.

3. KALMAN FILTER

Kalman filter is an Online filtering algorithm. Its graphical model is similar to a hidden Markov model. However, it assumes that the process v_{k-1} and measurement n_k noises as well as the posterior pdf $p(x_k|z_{1:k})$ are normally distributed (Fig. 1). Moreover, the function f_k and h_k are linear. Based on these assumptions, the following state transition and measurement equations are conceived.

$$x_k = F_k x_{k-1} + v_{k-1} \qquad (1)$$

$$z_k = H_k x_{k-1} + n_k \qquad (2)$$

The matrix F_k is called the state transition matrix and it helps to predict the current state of the target based on its previous state. Similarly, the matrix H_k associate the observation z_k to the target state x. The random variables v_{k-1}, n_k show the process and measurement noise. They are zero mean, normally distributed with covariance matrices Q_{k-1} and R_k respectively. A detailed description of the Kalman filter is beyond the scope of this paper. For details, readers may refer to [30, 31]. In our problem, we instantiated an instance of the Kalman filter for each target in the visual scene.

3.1. Hungarian Algorithm

The Hungarian algorithm is a greedy combinatorial optimization algorithm and solves the assignment problem in polynomial time. The tracking problem is modeled as a bipartite graph matching problem where the first set of nodes corresponds to the established trajectories and the seconds set of nodes corresponds to the target hypothesis measurements from the real world. In our case, the measurements correspond to the spatial locations of the head in every frame. The input to the algorithm is a cost matrix with N number of rows and M number of columns. N corresponds to the established trajectories where the M corresponds to the number of measurements at time step t. There are a variety of ways to obtain

the cost matrix [32]. In our work, we mainly used the special constraints for calculating the cost matrix. Specially, we measured the Euclidean distance between the targets head location in the current and previous frame and treat it as the cost. The nearer are the targets in the consecutive frames, the smaller will be the cost and most probably, the targets with the least distance correspond to the same targets in the temporal domain. Similarly, the targets that are far from each other in the consecutive frames would yield the highest cost and corresponds to different targets in the temporal domain. Once the cost matrix is obtained, the Hungarian algorithm [33] works in three steps as the following:

- **Row reduction operation:** Find the minimum cost of each row. Then subtract the corresponding minimum from each row entry to ensure at least one zero-entry in each row.

- **Column reduction operation:** Repeat the same procedure for each column. It will ensure at least on zero entry in each column.

- **Optimally test:** Find the minimum number of straight lines to cover all the zeros in the cost matrix. If the number of lines covering all the zeros equal to the number of rows and columns, optimality is achieved. However, if the number of lines covering all the zeros is not equal to the number of rows and columns, shift zeros such as to achieve the optimal assignment.

4. EXPERIMENT

The proposed algorithm is implemented in Matlab on a Core i7 system with 16 GB RAM. To ensure real-time performance, we excluded the deep feature based appearance model [20] but it could easily be incorporated in the cost matrix. To evaluate the network, we have chosen four datasets [34, 35] that are commonly used for pedestrian tracking. It is also worth noticing that all the standard datasets have annotation available but that is for the whole body which is not useful for our case. Therefore, we annotated the datasets to generate the target hypothesis in every frame. The qualitative results of the proposed method are given in Fig. 2. It is interesting to observe that the head based tracking works well when the heads are not covered with anything. Additionally, due to the most visible part of the body, it also helps in accurate tracking for the partially occluded regions. The proposed algorithm fails when the people use an umbrella or cover the head with an opaque material. However, in the majority of surveillance scenarios where the head is visible, the proposed algorithm works well.

5. CONCLUSION

We proposed a multi-target tracking algorithm of tracking the heads of multiple humans in the visual scene. The tracking-by-detection paradigm is followed where the spatial locations of the head are generated in every frame and a combinatorial optimization is used to establish the association between the corresponding targets. Especially, the Hungarian algorithm is used to associate the head of the corresponding targets in the consecutive frames in a greedy fashion. In the future, we are aiming to extend our approach to other Keypoint of the body part and rather than tracking the head of a person, track different body parts. Tracking the individual body parts would be a direction for the research for pose estimation and high-level behavior analysis.

6. REFERENCES

[1] Massoud Bazargan, *Airline operations and scheduling*, Routledge, 2016.

[2] Muhammad Mudassar Yamin, Basel Katt, and Vasileios Gkioulos, "Cyber ranges and security testbeds: Scenarios, functions, tools and architecture," *Computers & Security*, p. 101636, 2019.

[3] Muhammd Mudassar Yamin and Basel Katt, "Detecting malicious windows commands using natural language processing techniques," in *International Conference on Security for Information Technology and Communications*. Springer, 2018, pp. 157–169.

[4] Andrew Owens and Alexei A Efros, "Audio-visual scene analysis with self-supervised multisensory features," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 631–648.

[5] Sultan Daud Khan and Habib Ullah, "A survey of advances in vision-based vehicle re-identification," *Computer Vision and Image Understanding*, vol. 182, pp. 50–63, 2019.

[6] Jiachen Li, Wei Zhan, Yeping Hu, and Masayoshi Tomizuka, "Generic tracking and probabilistic prediction framework and its application in autonomous driving," *IEEE Transactions on Intelligent Transportation Systems*, 2019.

[7] Mohib Ullah, Habib Ullah, and Ibrahim M Alseadonn, "Human action recognition in videos using stable features," 2017.

[8] Tian Wang, Meina Qiao, Zhiwei Lin, Ce Li, Hichem Snoussi, Zhe Liu, and Chang Choi, "Generative neural networks for anomaly detection in crowded scenes," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 5, pp. 1390–1399, 2018.

[9] Mark Marsden, Kevin McGuinness, Suzanne Little, and Noel E O'Connor, "Resnetcrowd: A residual deep learning architecture for crowd counting, violent behaviour detection and crowd density level classification," in *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 2017, pp. 1–7.
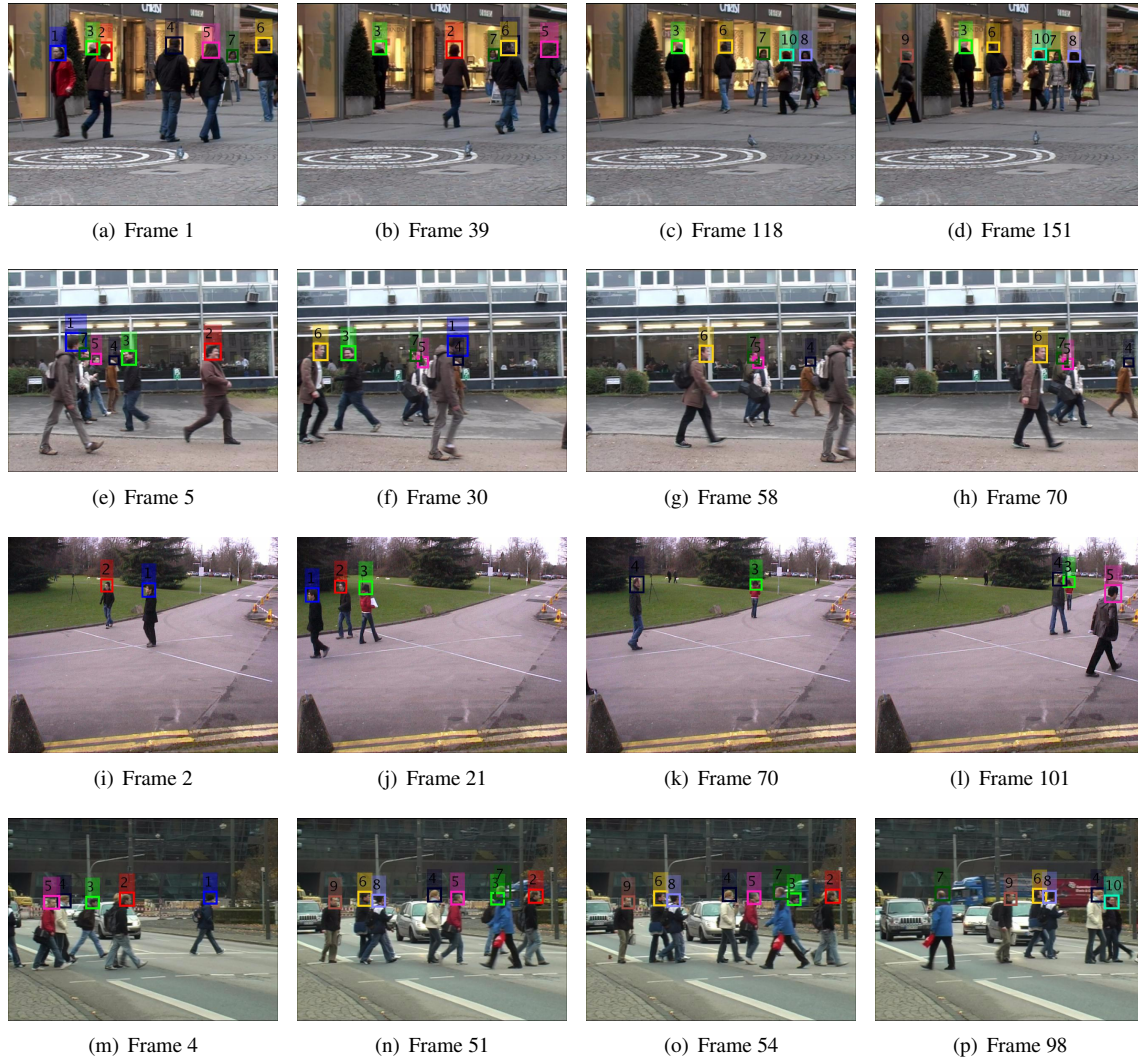
IS&T International Symposium on Electronic Imaging 2020
Intelligent Robotics and Industrial Applications using Computer Vision

074-3

(a) Frame 1 (b) Frame 39 (c) Frame 118 (d) Frame 151

(e) Frame 5 (f) Frame 30 (g) Frame 58 (h) Frame 70

(i) Frame 2 (j) Frame 21 (k) Frame 70 (l) Frame 101

(m) Frame 4 (n) Frame 51 (o) Frame 54 (p) Frame 98

Fig. 2: Tracking results of the proposed algorithm. Top to left: datasets Stadmitt (a-d), TUD-campus (e-h), Pets2009 (i-l), and TUD-crossing (m-p). Targets are given a unique integer ID.

[10] Mohib Ullah, Habib Ullah, Nicola Conci, and Francesco GB De Natale, "Crowd behavior identification," in *IEEE International Conference on Image Processing (ICIP)*, 2016, pp. 1195–1199.

[11] Habib Ullah, Ahmed B Altamimi, Muhammad Uzair, and Mohib Ullah, "Anomalous entities detection and localization in pedestrian flows," *Neurocomputing*, vol. 290, pp. 74–86, 2018.

[12] Habib Ullah, Muhammad Uzair, Mohib Ullah, Asif Khan, Ayaz Ahmad, and Wilayat Khan, "Density independent hydrodynamics model for crowd coherency detection," *Neurocomputing*, vol. 242, pp. 28–39, 2017.

[13] Graham Thomas, Rikke Gade, Thomas B Moeslund, Peter Carr, and Adrian Hilton, "Computer vision for sports: Current applications and research topics," *Computer Vision and Image Understanding*, vol. 159, pp. 3–18, 2017.

[14] Sultan Daud Khan, Habib Ullah, Mohib Ullah, Nicola Conci,

Faouzi Alaya Cheikh, and Azeddine Beghdadi, "Person head detection based deep model for people counting in sports videos," in *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 2019, pp. 1–8.

[15] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi, "You only look once: Unified, real-time object detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.

[16] Sultan Daud Khan, Habib Ullah, Mohib Ullah, Faouzi Alaya Cheikh, and Azeddine Beghdadi, "Dimension invariant model for human head detection," in *2019 8th European Workshop on Visual Information Processing (EUVIP)*. IEEE, 2019, pp. 99–104.

[17] Paul Voigtlaender, Michael Krause, Aljosa Osep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and

074-4

IS&T International Symposium on Electronic Imaging 2020
Intelligent Robotics and Industrial Applications using Computer Vision

Bastian Leibe, "Mots: Multi-object tracking and segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7942–7951.

[18] Mohib Ullah, Habib Ullah, and Faouzi Alaya Cheikh, "Single shot appearance model (ssam) for multi-target tracking," *Electronic Imaging*, vol. 2019, no. 7, pp. 466–1, 2019.

[19] Samuel Schulter, Paul Vernaza, Wongun Choi, and Manmohan Chandraker, "Deep network flow for multi-object tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6951–6960.

[20] Mohib Ullah, Ahmed Kedir Mohammed, Faouzi Alaya Cheikh, and Zhaohui Wang, "A hierarchical feature model for multi-target tracking," in *IEEE International Conference on Image Processing (ICIP)*, 2017, pp. 2612–2616.

[21] Anton Milan, Stefan Roth, and Konrad Schindler, "Continuous energy minimization for multitarget tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 1, pp. 58–72, 2014.

[22] ShiJie Sun, Naveed Akhtar, HuanSheng Song, Ajmal S Mian, and Mubarak Shah, "Deep affinity network for multiple object tracking," *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[23] Mohib Ullah and Faouzi Alaya Cheikh, "A directed sparse graphical model for multi-target tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 1816–1823.

[24] Seung-Hwan Bae and Kuk-Jin Yoon, "Confidence-based data association and discriminative deep appearance learning for robust online multi-object tracking," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 3, pp. 595–610, 2017.

[25] Mohib Ullah, Faouzi Alaya Cheikh, and Ali Shariq Imran, "Hog based real-time multi-target tracking in bayesian framework," in *2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 2016, pp. 416–422.

[26] Jeany Son, Mooyeol Baek, Minsu Cho, and Bohyung Han, "Multi-object tracking with quadruplet convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5620–5629.

[27] Qi Chu, Wanli Ouyang, Hongsheng Li, Xiaogang Wang, Bin Liu, and Nenghai Yu, "Online multi-object tracking using cnn-based single object tracker with spatial-temporal attention mechanism," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4836–4845.

[28] Mohib Ullah and Faouzi Alaya Cheikh, "Deep feature based end-to-end transportation network for multi-target tracking," in *IEEE International Conference on Image Processing (ICIP)*, 2018, pp. 3738–3742.

[29] Afshin Dehghan, Shayan Modiri Assari, and Mubarak Shah, "Gmmcp tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4091–4099.

[30] Simon Maskell and Neil Gordon, "A tutorial on particle filters for on-line nonlinear/non-gaussian bayesian tracking," *Target tracking: algorithms and applications (Ref. No. 2001/174), IEE*, vol. 2, pp. 21–215, 2001.

[31] Gary Bishop, Greg Welch, et al., "An introduction to the kalman filter," *Proc of SIGGRAPH, Course*, vol. 8, no. 27599-23175, pp. 41, 2001.

[32] Wenhan Luo, Junliang Xing, Anton Milan, Xiaoqin Zhang, Wei Liu, Xiaowei Zhao, and Tae-Kyun Kim, "Multiple object tracking: A literature review," *arXiv preprint arXiv:1409.7618*, 2014.

[33] Roy Jonker and Ton Volgenant, "Improving the hungarian assignment algorithm," *Operations Research Letters*, vol. 5, no. 4, pp. 171–175, 1986.

[34] James Ferryman and Ali Shahrokni, "Pets2009: Dataset and challenge," in *Twelfth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, 2009, pp. 1–6.

[35] Hayko Riemenschneider, Sabine Sternig, Michael Donoser, Peter M Roth, and Horst Bischof, "Hough regions for joining instance localization and segmentation," in *European Conference on Computer Vision*. Springer, 2012, pp. 258–271.

IS&T International Symposium on Electronic Imaging 2020
Intelligent Robotics and Industrial Applications using Computer Vision

074-5