Graph neural network based human detection in videos during occlusion environments

Kusuma Sriram[1,2], Kiran Purushotham[2]
[1]Department of Information Science and Engineering, M S Ramaiah Institute of Technology, Bangalore, India
[2]Department of Computer Science and Engineering, RNS Institute of Technology, Bengaluru, Affiliated to Visvesvaraya Technological University, Belagavi, India

Article Info

Article history:

Received Jul 11, 2023
Revised Feb 19, 2024
Accepted Feb 20, 2024

Keywords:

Computer vision
Graph neural network
Multi object tracking
Multiple object detection
Occlusion

ABSTRACT

One of the most difficult perceptual problems for many applications is accurately recognizing the human object in a variety of circumstances. This can be difficult due to obstructions, weather, complex backdrops, cast shadows, and occlusions. Occlusion is a challenging open problem where a detector can only perceive a portion of the target human because of obstacles in the surrounding. In this research, an experimental investigation was conducted using the multi object tracking (MOT17) datasets to construct a graph neural network-based solution for the detection of humans in videos while considering the possibility of occlusion. Graph neural network (GNN) is used for the construction of neural solver model for detecting human object in occlusion scenario. The results obtained shows that this proposed method offers a considerable improvement in efficiency in comparison to the ways that have been used in the past. The values obtained for the standard performance metrics are higher than the state-of-the-art methods.

Corresponding Author:

Kusuma Sriram
Department of Information Science and Engineering, M S Ramaiah Institute of Technology
MSR Nagar, Mathhikere, Bengaluru, Karnataka State-560054, India
Email: kusumas.phd@gmail.com

1. INTRODUCTION

Humans are incredibly quick and precise in spotting and identifying things in our environs, even under situations when the object is just poorly visible. Our brains may make up for missing information by making associations between the visible portions of an item [1]. The technology behind computers is still not advanced enough to do this. Therefore, object recognition is a major topic in the area of computer vision. in recent years, it has made great strides because of the advent of deep neural systems and the accessibility of massive amounts of information. Many industries have taken notice because of potential uses in areas including advanced driver assistance systems [2], surveillance [3], and scene interpretation [4]. Accurately detecting the object in varied contexts is the greatest difficulty and is the first step in tracking, although this can be challenging because of factors including complicated backdrops, weather, cast shadows, and occlusions. There are several computer vision techniques that fail when obstructed items are present in the scene. It is important to note that the camera’s perspective determines the obscured area. The minimization technique, temporal selection, graph cut technique, and squared sum distances are all used to deal with the same obstruction challenge [5]. In other cases, the camera’s orientation determines if any section is obscured or not.

When trying to keep tabs on individuals in a variety of settings, occlusion may provide some serious difficulties. Video surveillance and intelligent vehicles, including unmanned ground vehicles (UGVs) and unmanned aerial vehicles (UAVs), rely significantly on human detection technology. Human identification is difficult in both images and videos because of the wide variety of stances that individuals may adopt. Due to the proximity of other objects, a significant degree of partial occlusion happens in the actual world when people walk around. Using a combination of images from many cameras to calculate the depth and then estimate the obstructed section is one approach to this problem [6]. Due to the difficulty in estimating depth from a single location using just local information, a polynuclear stereo technique is utilized for occlusion management [7]. Temporal and trajectory prediction was first proposed by Rosales and Sclaroff for the same issue [8]. Geiger *et al*. [9] employed a method including epipolar lines and a disparity map to handle obstruction, which relies heavily on the geometry of the picture. Suppression by use of a limit on the sequence of operations is a method designed by Little and Gillett [10]. The displacing field between the pictures plays a crucial role in this competence, and it is used to locate the occluded areas in the images by establishing two occlusion maps [11].

The original goal of scenario comprehension was to create robots that can perceive like humans, allowing them to infer broad concepts and immediate context from visuals. Unanticipated uses for scene understanding technologies, such as picture search engines, computational imaging, vision for infographics, and human-machine interface (HMI) and autonomous vehicles are gaining traction. The optimization challenges and obstruction problems can be addressed with the use of cost aggregation based on energy functions [12]. Cross-checking and extrapolating is a straightforward approach to obstruction detection [13]. Partially occluded events can be processed with an additional normalized cross-correlation procedure [14]. Particle filters with information association can be used to detect and follow occluded objects in a scene [15]. Tracing the paths taken by each individual circumstance of an entity in a video is referred to as multi-object tracking (MOT). Among its many potential uses are driverless vehicles, biological research, and video surveillance systems. The challenge of associating data in this type of tracking under occlusion conditions is typically posed in the form of a graph partitioning topic.

A modified MOT solver technique under occlusion conditions which is based on a graph partitioning technique is proposed here. The proposed network architecture consists of a convolutional-based transfer learning network for feature learning. The construction employs a message-passing neural network to enhance the learning procedure. As it develops, this network uncovers how to merge profound characteristics into high-order data throughout the graph. Even though our technique is based on a very straightforward graph-based concept, it is nonetheless capable of taking into consideration global interconnections between detection methods. Proposed approach outperforms the contemporary by a wide margin, is significantly quicker than several classic graph partitioning approaches, and does so without needing any specially formulated attributes.

In this work, the tracking scenario is splitted into two parts, single and double camera obstruction. So, three different cameras were used by Chang *et al*. [6] is an effort to overcome the occlusion. Optical coherence tomography (OCT) employs geometrical and recognition-based methodologies with the use of signals, including epipolar-lines, landmarks, perceived color, and elevation.

Hu *et al*. [16] employ a Bayesian network strategy to address the problem of occlusion when following individuals. The subject's expected joint characteristics were followed using a condensing algorithm. There are three possible values for *t*, with zero indicating no occlusion, one indicating that A occludes B, and two indicating that B occludes A [16]. Here the histogram of gradient (HOG) with the local binary pattern is used to create an innovative and influential approach for human detecting and handling partial occlusion (LBP). A global window detector is used to scan, whereas local area detectors are utilized to pinpoint specific features. Results on the INRIA and Pascal datasets show that with the enhanced HOG LBP feature & the global part occlusion management approach, they obtained 91.3% rate of detection [17]. Before anything else, they generate an occlusion map that details where people and fixed objects block the view. Second, when the item reappears in view, the odds of locating it are improved by using an extended Kalman filter (EKF) [18].

Enzweiler *et al*. [19] weights that are proportional to the level of visibility to account for partial occlusion. Classifying pedestrians in photographs is the focus of this work. Intensity, depth, and motion characteristics comprise the framework employed for training the component-based expert classifiers [19]. Although stereo vision is preferable for dealing with occlusions, Wojek *et al*. [20] have claimed success with on the ETH-PedCross2 dataset based on scene interpretation based on monocular vision. The proposed method is effective for dealing with long-term occlusions [20]. When dealing with web-based programs, data association can be performed either track-by-track [21] or on a frame-by-frame basis [22]. Because batch approaches are more resistant to being affected by occlusions, they are the approach of choice for video processing tasks that can be completed offline. The use of a graph as the basis for this kind of model is considered to be the industry standard [23]. In this model, a node represent single detection, and the presence of edges shows the existence of potential connections between the nodes [24]. The association of data can then be phrased as flow with maximum value [25] or, correspondingly, as a minimal cost issue with either fixed costs depending on the distance [26], incorporating motion models [27], or learning costs. This can be done in one of two ways. Alternate approaches often result in optimization issues that are more complex, such as minimal cliques [28], general-purpose solutions, such as multi-cuts [29]. The building of ever-more-complex models that take into account additional types of visual input, like reconstructing for multi-camera frames [30], activity identification [31], segmentation [32], keypoint trajectory [33] or joint identification, has been more popular as of late.

2. METHOD

2.1. Data collection and preprocessing

2.1.1. Dataset

The MOT Challenge is a standard dataset for multiple objects tracking that comprises various challenging walker monitoring sessions. These sessions include significant occlusions and they are set in congested environments. For this proposed approach, MOT17 dataset is being used [34], [35]. All MOT16 sequences are used in MOT17, along with an improved ground truth that has better precision. For every session, three different recognition types are available: DPM, Faster-RCNN, and SDP. All six subsets of this particular dataset were used in the training and testing stages.

Originally MOT17 consisted of 42 distinct video sessions. For training half of the data set is used, and the rest are used for the testing of model implementation. To make data set and training more efficient, all the six subsets of the main data set are used for this experimental study instead of restricting any of the single subset. Figure 1 shows three frames from each of the 6 subsets. These sub-datasets are: MOT17-04 – DPM, MOT17-04 – FRCNN, MOT17-04 – SDP, MOT17-11 – DPM, MOT17-11–FRCNN, MOT17-11 – SDP. The MOT17 dataset series includes various video data sets which are captured specifically with different challenges. As an example, a video of the pedestrian at night on street was taken from an elevated viewpoint in the MOT17-04 dataset and MOT17-11 data set series consists of a video recording from a forward-facing camera inside a crowded retail mal.

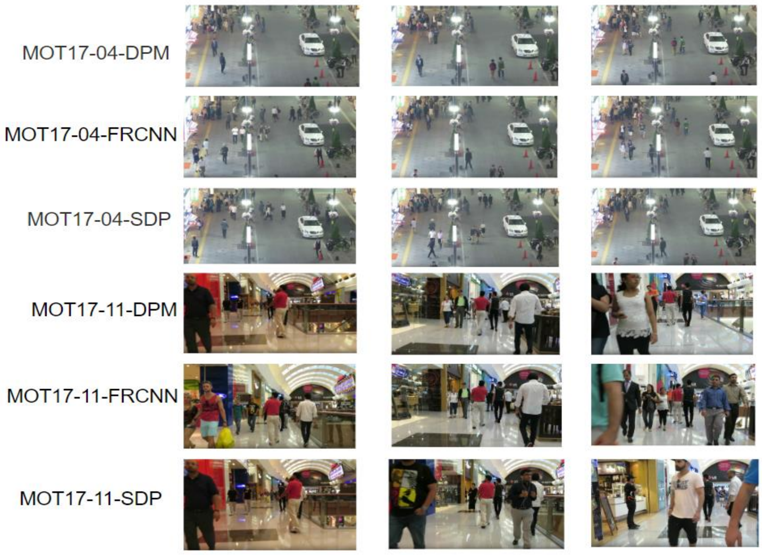

Figure 1. Sample frames from the six-input data subset

2.1.2. Augmentation

Our network is trained using random samples from pools of 8 graphs. For stationary scenes, the sampling rate is six frames per second, while for dynamic scenes, it is nine fps. Each graph represents a series of 15 frames. As part of our data preprocessing, arbitrarily removal of nodes from the network is done to mimic skipped detection systems and also arbitrarily moved the bounding frames. The percentage of detections that might be randomly dropped is kept in the range of 0-0.3. In CNN usual image frame size is kept at 128×128. A frame size of 256×256 is used in this model for bounding box image generation. The frame sampling rate probability, set at 0.5, will vary randomly during the learning phase. The batch size for future evaluations remains constant at 5,000.

2.2. Network architecture

The aim of this proposed neural solver is to detect people under occlusion in a crowded environment. The problem will be more challenging as the crowd is in motion in the input dataset. For this model execution, a hardware setup RAM 16 GB, graphics 8-10 GB and 24 core CPU of is made use. Using a collection of dynamic images, this technique builds a graph by considering the frames as nodes and joining frame pairs with edges to form a coherent structure. To encode the geometric information of nodes and edges in the provided set of dynamic images, a CNN is first used in the graph embedding process. An MLP is then used. These hidden layers are used as input to a neural message-passing algorithm, which will iteratively spread the data contained inside them throughout the graph for a certain number of times. Later embedding's are utilized to label edges as active or inactive. The cross-entropy loss is calculated for our projections relative to the ground truth labels throughout the learning period. A straightforward rounding strategy is used to convert our categorization scores to bins to get our final outcomes. Figure 2 presents the basic design of the proposed neural solver for multiple human detections under occlusion conditions.

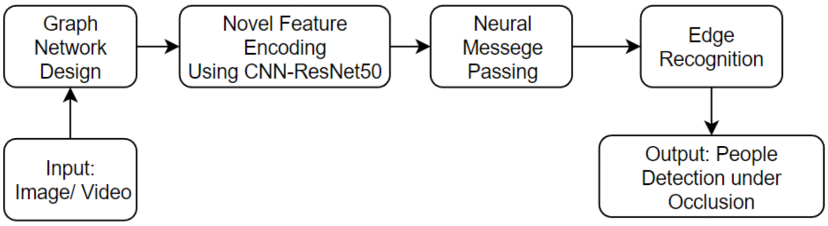
Graph Network Design
Novel Feature Encoding Using CNN-ResNet50
Neural Messege Passing
Edge Recognition
Input: Image/ Video
Output: People Detection under Occlusion

Figure 2. The basic design of our proposed neural solver for human detection (MOT) under occlusion conditions

2.3. Graph neural network

Graph neural networks (GNNs) are a subset of deep learning techniques. It is optimized for inferential tasks containing graph-based information. Applying these neural networks explicitly to graphical structures simplifies node-stage, edge-stage, and graph-stage prediction problems. Consider p as the collection of vertices and E is the link between them. Which is represented as:

$$G = (p, E) \tag{1}$$

The '*sacred MOT metrics logger*' is employed to compute all MOT metrics from a set of output tracking files. Also, the node embedding is set as $l_i^{(0)}$where $(i \in p)$ and the edge embedding can be presented as $l_{(i,j)}^{(0)}$, where $(i, j) \in E$.

In this proposed method, a graph model is built with nodes depicting individual detections and edges representing the links between them. This is used to see the relationships between the many detected items in the video input data. The highest number of frames that can be found in each sampled graph is set to 15. The graph that has been built can identify between 25 and 500 different human figures in the scene. A minimum visibility score of 0.2 is required for the ground truth (GT) boxes during the learning phase. Each of the graph's charts shows fifteen distinct picture frames. The GNN is constructed by connecting nodes in the graph using the k-nearest neighbor (KNN) method. The parameter 'k', which denotes the number of neighbors in a KNN, is essential to the GNN architecture. 'A1' is regarded as the target position for label prediction in this case.

The process involves the following steps:

- K nearest approach: The first step is to locate the k points that are on the graph which are closest to 'A1'. The proximity metric is used to find these points; Euclidean distance measure is used.
- Vote-based classification: Following the identification of the K nearest points, the following stage involves classifying these points based on the votes cast by their closest neighbors. Votes are cast for each point according to its object type.
- Category prediction: Based on the votes, the most popular category is chosen to make the final prediction. The forecast for the object in the target location 'A1' is derived from the category that received the most votes.
- At first, 100 was chosen as the number of k-closest points (k). The model's performance was found to be improved more effectively when the value of 'k' was increased to 50.

2.4. Feature encoding

Feature encoding is a significant step in this whole model construction. This step transforms all the required features into a standard format so that the model can be implemented on the extracted features from the dataset. ResNet50, a novel convolutional transfer learning model is utilized for the embedding of the features. The ResNet50 model extracts features from the RGB image frames.

Hence, we can be able to retrieve q_i's corresponding node embedding by calculating. $l_i^{(0)} := CNN$, and this is done for each detection of q_i where $q_i \in Q$ and each image patch that corresponds to q_i. Next, encoding the features with geometric information is achieved. The considered parameters for the feature encoding step are Time distance, 2-D co-ordinate distance (d), Bounding box height (a) and width (b) and ReID score. This distance d in between the i^{th} and j^{th} timed detection can be calculated using the left top corner image coordinate of the bounding box (m_i, n_i). Equation (2) shows the mathematical formula to calculate the distance d.

$$d = \left(\frac{2(m_j-m_i)}{a_i+a_j}, \frac{2(n_j-n_i)}{a_i+a_j}, \log\frac{a_i}{a_j}, \log\frac{b_i}{b_j}\right) \tag{2}$$

2.5. Neural message passing

Following the feature encoding phase, the message passing stage is the next. Several iterations of message transfer are carried out on the graph. Throughout each iteration of the message passing process, nodes impart visual details about themselves to their connected edges, and likewise, edges impart geometric details about themselves to the nodes upon which they make contact. As a result, improved embedding's for nodes and edges are obtained that account for the graph's topology at an advanced level. Edge projections can be improved by recurrent refinement, if the number of message-passing steps are increased, an additional detail can be encoded in each node and also in edge encoding. Therefore, larger values are likely to result in more efficient networks. In our approach, message-passing steps are set to 24. After the message is passed, the number of steps during which feature vectors are categorized is set to 23. The initially encoded feature nodes are also included during the update process of the nodes. Figure 3 shows the development stage of nodes inside the model architecture for message passing. Edges with arrows point in the direction of time. The picture depicts the initial state following an edge update and the computation of the intermediate node upgrade embedding's.

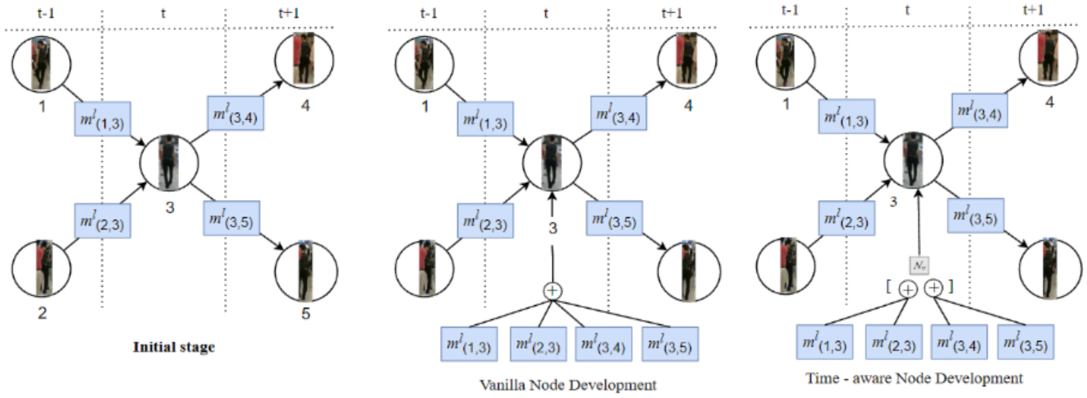

Figure. 3. Node development due to message passing in the network

3. RESULTS AND DISCUSSION

3.1. Results of model implementation

Results obtained from the model implemented on six different subsets of the MOT17 dataset are discussed here in this section. While training, During the experimental investigation, the program '*MOTMetricsLogger*' is used to compute all MOT metrics using a collection of output tracking data. GNN is used for the basic construction of this neural solver for detecting an object (here, target object = human) under an obstruction. Next, CNN-based ResNet50 is applied for feature encoding. In the network, CNN downsizes the image frames, which eventually compresses the MOT dataset. GNNs are able to accomplish goals that CNNs were unable to achieve. CNNs are based on a fundamental idea that incorporates concealed convolution and pooling levels to detect locally specific characteristics using a kernel-based collection of visual fields. The window of the convolutional controller is moved along a 2-D picture, and a function is then computed across a certain window, which goes through many filters., because graphs can have any size and any structure and the lack of physical proximity, CNN is difficult. The sequencing of nodes is likewise not set in stone. Since graphs are order-invariant, the goal is to get the outcome irrespective of the way the nodes are arranged. Figure 4 depicts the output images with detected human objects shown in bounded boxes.

Table 1 shows the output performance metrics obtained from the model implementation on MOT17 datasets. The highest MOTA score is found for the MOT17-04-SDP subset. That same subset also shows the highest IDF1 score, with a value of 82.7%. The overall MOTA and IDF1 values for this experimental study are found to be 69% and 76.5%, respectively. These values are far higher than the state-of-the-art methods. Figures 5 compare precision and Figure 6 compares recall metrics for the outputs from the different data subsets. The precision score is really high for each subset of the input data, reaching an overall precision of 99.4%. The aggregate value for the recall is 69.5% here.

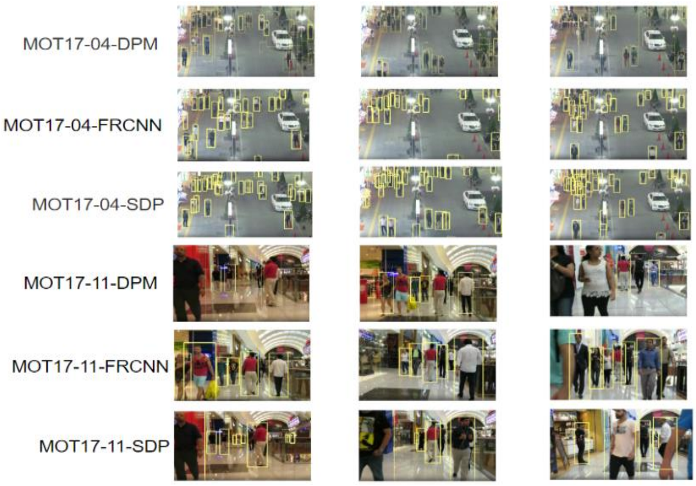

Figure 4. Output frames containing bounding boxes for multiple-human object detection

Table 1. Output performance for different subsets of MOT17

Dataset	MOTA	IDF1	IDP	MT	ML	FP	FN
MOT17-04-DPM	65.5%	74.4%	93.8	31	24	54	16341
MOT17-04-FRCNN	65.2%	75.3%	95.1	34	22	70	16479
MOT17-04-SDP	76.2%	82.7%	95.1	45	16	183	11127
MOT17-11-DPM	64.4%	67.5%	84.5	17	22	87	3258
MOT17-11-FRCNN	69.6%	74.1%	87.9	28	18	150	2703
MOT17-11-SDP	74.3%	71.7%	81.2	36	13	213	2192
Overall	69.0%	76.5%	92.9	191	115	757	52100

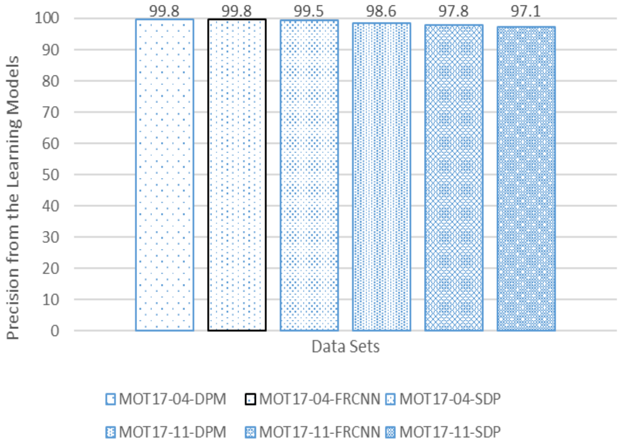

Figure 5. Comparison of precision values from different data subsets

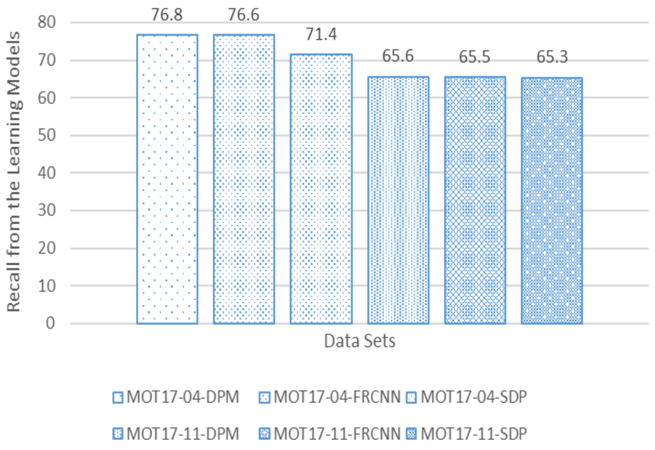

Figure 6. Comparison of recall values from different data subsets

3.2. Evaluation with other models

Table 2 presents the overall comparison of different model performances with our proposed model. This experimental study shows that this approach surpasses the state-of-the-art outcomes on all tasks by a significant margin. Our proposed technique is more efficient and much quicker than other graph partitioning algorithms.

Table 2. Comparison of different model performance for MOT17 dataset

Model	Paper	MOTA	IDF1	MT	ML	FP	FN
Tracker-aware	Bergmann *et al.* [36]	56.2%	54.9%	20.7	35.8	8866	235449
JBNOT	Henschel *et al.* [37]	52.6%	50.8%	19.7	35.8	31572	232659
FAMNet	Chu and Ling [38]	52.0%	48.7%	19.1	33.4	14138	253616
eHAF	Sheng *et al.* [39]	51.8%	54.7%	23.4	37.9	33212	236772
NOTA	Chen *et al.* [40]	51.3%	54.7%	17.1	35.4	20148	252,531
FWT	Henschel *et al.* [41]	51.3%	47.6%	21.4	35.2	24101	247921
Time-aware CNN	Brasó and Leal-Taixe [42]	64.0%	70.0%	648	362	6169	114509
Proposed	Our	69.0%	76.5%	191	115	757	52100

4. CONCLUSION

In conclusion, this research addressed the challenging problem of accurately recognizing human objects in various circumstances, particularly in the presence of occlusions. Using the MOT17 datasets, the study used an experimental methodology and suggested a GNN-based solution. The efficacy of the GNN-based neural solution in detecting humans, especially in situations when they are obscured, was proved by notable enhancements in efficiency when compared to traditional techniques. The acquired outcomes demonstrated better performance metrics in comparison to cutting-edge techniques, underscoring the possibility of the suggested strategy to improve human object recognition in difficult visual circumstances. It paves the way for future research that goes beyond extracting features for MOT jobs and instead focuses on incorporating learning into the process of total data correlation. This work is especially beneficial for detecting human objects in busy areas that are prone to frequent visual blockages due to obstacles or distracted environment.

REFERENCES

[1] R. A. Rensink and J. T. Enns, “Early completion of occluded objects,” *Vision Research*, vol. 38, no. 15–16, pp. 2489–2505, Aug. 1998, doi: 10.1016/S0042-6989(98)00051-0.

[2] H. T. Niknejad, T. Kawano, Y. Oishi, and S. Mita, “Occlusion handling using discriminative model of trained part templates and conditional random field,” in *IEEE Intelligent Vehicles Symposium, Proceedings*, Jun. 2013, pp. 750–755, doi: 10.1109/IVS.2013.6629557.

[3] A. M. Pinto, P. G. Costa, and A. P. Moreira, “An architecture for visual motion perception of a surveillance-based autonomous robot,” in *2014 IEEE International Conference on Autonomous Robot Systems and Competitions, ICARSC 2014*, May 2014, pp. 205–211, doi: 10.1109/ICARSC.2014.6849787.

[4] A. Karpathy, A. Joulin, and F. F. Li, “Deep fragment embeddings for bidirectional image sentence mapping,” *Advances in Neural Information Processing Systems*, vol. 3, no. January, pp. 1889–1897, Jun. 2014.

[5] S. B. Kang, R. Szeliski, and J. Chai, "Handling occlusions in dense multi-view stereo," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2001, vol. 1, pp. 103–110, doi: 10.1109/CVPR.2001.990462.

[6] T. H. Chang and S. Gong, "Tracking multiple people with a multi-camera system," *Proceedings - 2001 IEEE Workshop on Multi-Object Tracking, MOT 2001*, pp. 19–26, 2001, doi: 10.1109/MOT.2001.937977.

[7] Y. Sugaya and Y. Ohta, "Stereo by integration of two algorithms with/without occlusion handling," in *Proceedings - International Conference on Pattern Recognition*, 2000, vol. 15, no. 1, pp. 109–112, doi: 10.1109/icpr.2000.905286.

[8] R. Rosales and S. Sclaroff, "Improved tracking of multiple humans with trajectory prediction and occlusion modeling," *System*, Nov. 1998.

[9] D. Geiger, B. Ladendorf, and A. Yuille, "Occlusions and binocular stereo," *International Journal of Computer Vision*, vol. 14, no. 3, pp. 211–226, Apr. 1995, doi: 10.1007/BF01679683.

[10] J. J. Little and W. E. Gillett, "Direct evidence for occlusion in stereo and motion," *Image and Vision Computing*, vol. 8, no. 4, pp. 328–340, Nov. 1990, doi: 10.1016/0262-8856(90)80009-I.

[11] J. Weng, N. Ahuja, and T. S. Huang, "Two-view matching," in *[1988 Proceedings] Second International Conference on Computer Vision*, 1988, pp. 64–73, doi: 10.1109/ccv.1988.589972.

[12] D. Min and K. Sohn, "Cost aggregation and occlusion handling with WLS in stereo matching," *IEEE Transactions on Image Processing*, vol. 17, no. 8, pp. 1431–1442, Aug. 2008, doi: 10.1109/TIP.2008.925372.

[13] G. Egnal and R. P. Wildes, "Detecting binocular half-occlusions: Empirical comparisons of five approaches," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 8, pp. 1127–1133, Aug. 2002, doi: 10.1109/TPAMI.2002.1023808.

[14] R. Mehmood, R. Nawaz, and N. I. Rao, "Occlusion handling in meanshift tracking using adaptive window Normalized Cross Correlation," in *Proceedings of 2014 11th International Bhurban Conference on Applied Sciences and Technology, IBCAST 2014*, Jan. 2014, pp. 126–129, doi: 10.1109/IBCAST.2014.6778134.

[15] Y. Guan, X. Chen, D. Yang, and Y. Wu, "Multi-person tracking-by-detection with local particle filtering and global occlusion handling," in *Proceedings - IEEE International Conference on Multimedia and Expo*, Jul. 2014, vol. 2014-September, no. Septmber, doi: 10.1109/ICME.2014.6890149.

[16] M. Hu, W. Hu, and T. Tan, "Tracking people through occlusions," in *Proceedings - International Conference on Pattern Recognition*, 2004, vol. 2, pp. 724–727, doi: 10.1109/ICPR.2004.1334361.

[17] X. Wang, T. X. Han, and S. Yan, "An HOG-LBP human detector with partial occlusion handling," in *2009 IEEE 12th International Conference on Computer Vision*, Sep. 2009, pp. 32–39, doi: 10.1109/ICCV.2009.5459207.

[18] A. Ess, K. Schindler, B. Leibe, and L. van Gool, "Improved multi-person tracking with active occlusion handling," *ICRA 2009 Workshop*, 2009.

[19] M. Enzweiler, A. Eigenstetter, B. Schiele, and D. M. Gavrila, "Multi-cue pedestrian classification with partial occlusion handling," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 2010, pp. 990–997, doi: 10.1109/CVPR.2010.5540111.

[20] C. Wojek, S. Walk, S. Roth, and B. Schiele, "Monocular 3D scene understanding with explicit occlusion reasoning," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 2011, pp. 1993–2000, doi: 10.1109/CVPR.2011.5995547.

[21] J. Berclaz, F. Fleuret, and P. Fua, "Robust people tracking with global trajectory optimization," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2006, vol. 1, pp. 744–750, doi: 10.1109/CVPR.2006.258.

[22] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool, "Robust tracking-by-detection using a detector confidence particle filter," in *Proceedings of the IEEE International Conference on Computer Vision*, Sep. 2009, pp. 1515–1522, doi: 10.1109/ICCV.2009.5459278.

[23] Y. Wang, K. Kitani, and X. Weng, "Joint object detection and multi-object tracking with graph neural networks," in *Proceedings - IEEE International Conference on Robotics and Automation*, May 2021, vol. 2021-May, pp. 13708–13715, doi: 10.1109/ICRA48506.2021.9561110.

[24] W. Shi and R. Rajkumar, "Point-GNN: Graph neural network for 3D object detection in a point cloud," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 2020, pp. 1708–1716, doi: 10.1109/CVPR42600.2020.00178.

[25] J. Berclaz, F. Fleuret, E. Türetken, and P. Fua, "Multiple object tracking using k-shortest paths optimization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 9, pp. 1806–1819, Sep. 2011, doi: 10.1109/TPAMI.2011.21.

[26] H. Pirsiavash, D. Ramanan, and C. C. Fowlkes, "Globally-optimal greedy algorithms for tracking a variable number of objects," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 2011, pp. 1201–1208, doi: 10.1109/CVPR.2011.5995604.

[27] L. Leal-Taixé, G. Pons-Moll, and B. Rosenhahn, "Everybody needs somebody: Modeling social and grouping behavior on a linear programming multiple people tracker," in *Proceedings of the IEEE International Conference on Computer Vision*, Nov. 2011, pp. 120–127, doi: 10.1109/ICCVW.2011.6130233.

[28] A. Roshan Zamir, A. Dehghan, and M. Shah, "GMCP-tracker: Global multi-object tracking using generalized minimum clique graphs," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7573 LNCS, no. PART 2, Springer Berlin Heidelberg, 2012, pp. 343–356.

[29] S. Tang, M. Andriluka, B. Andres, and B. Schiele, "Multiple people tracking by lifted multicut and person re-identification," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, Jul. 2017, vol. 2017-January, pp. 3701–3710, doi: 10.1109/CVPR.2017.394.

[30] Z. Wu, T. H. Kunz, and M. Betke, "Efficient track linking methods for track graphs using network-flow and set-cover techniques," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 2011, pp. 1185–1192, doi: 10.1109/CVPR.2011.5995515.

[31] W. Choi and S. Savarese, "A unified framework for multi-target tracking and collective activity recognition," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7575 LNCS, no. PART 4, Springer Berlin Heidelberg, 2012, pp. 215–230.

[32] A. Milan, L. Leal-Taixé, K. Schindler, and I. Reid, "Joint tracking and segmentation of multiple targets," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 2015, vol. 07-12-June-2015, pp. 5397–5406, doi: 10.1109/CVPR.2015.7299178.

[33] W. Choi, "Near-online multi-target tracking with aggregated local flow descriptor," in *Proceedings of the IEEE International Conference on Computer Vision*, Dec. 2015, vol. 2015, pp. 3029–3037, doi: 10.1109/ICCV.2015.347.

[34] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler, "MOTChallenge 2015: Towards a benchmark for multi-target tracking," *arxiv.org/abs/1504.01942,* Apr. 2015.

[35] A. Milan, L. Leal-Taixe, I. Reid, S. Roth, and K. Schindler, “MOT16: A benchmark for multi-object tracking,” arxiv.org/abs/1603.00831, Mar. 2016.

[36] P. Bergmann, T. Meinhardt, and L. Leal-Taixe, “Tracking without bells and whistles,” in *Proceedings of the IEEE International Conference on Computer Vision*, Oct. 2019, vol. 2019-October, pp. 941–951, doi: 10.1109/ICCV.2019.00103.

[37] R. Henschel, Y. Zou, and B. Rosenhahn, “Multiple people tracking using body and joint detections,” in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, Jun. 2019, vol. 2019-June, pp. 770–779, doi: 10.1109/CVPRW.2019.00105.

[38] P. Chu and H. Ling, “FAMNet: Joint learning of feature, affinity and multi-dimensional assignment for online multiple object tracking,” in *Proceedings of the IEEE International Conference on Computer Vision*, Oct. 2019, vol. 2019-October, pp. 6171–6180, doi: 10.1109/ICCV.2019.00627.

[39] H. Sheng, Y. Zhang, J. Chen, Z. Xiong, and J. Zhang, “Heterogeneous association graph fusion for target association in multiple object tracking,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 11, pp. 3269–3280, Nov. 2019, doi: 10.1109/TCSVT.2018.2882192.

[40] L. Chen, H. Ai, R. Chen, and Z. Zhuang, “Aggregate tracklet appearance features for multi-object tracking,” *IEEE Signal Processing Letters*, vol. 26, no. 11, pp. 1613–1617, Nov. 2019, doi: 10.1109/LSP.2019.2940922.

[41] R. Henschel, L. Leal-Taixe, D. Cremers, and B. Rosenhahn, “Improvements to Frank-Wolfe optimization for multi-detector multi-object tracking,” *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, vol. 2018-June, pp. 1509–1518, May 2018.

[42] G. Braso and L. Leal-Taixe, “Learning a neural solver for multiple object tracking,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 2020, pp. 6246–6256, doi: 10.1109/CVPR42600.2020.00628.

BIOGRAPHIES OF AUTHORS

Kusuma Sriram is working as an assistant professor in the Information Science and Engineering Department, M S Ramaiah Institute of Technology, Bengaluru. She Pursued B. E degree under VTU in the year 2004. She completed M. Tech in the year 2010 from RVCE, Bengaluru. As an academician, her experience in the field is 19+ years. Her research area is video processing/machine learning, in which She is Pursuing her Ph.D. She has Handled versatile subjects of both UG and PG. Motivated students to do mini projects of subject relevance and guided many research projects. She worked as faculty trainer for the Infosys InfyTQ program, through which she supported students for Infosys job opportunities. Also, she has Worked for Infosys campus connects program as trainer. She can be contacted at email: kusumas.phd@gmail.com.

Kiran Purushotham currently working as Head of the department, CSE Department with total experience of 21+ years. His research is on privacy preserving data mining, where the focus is on detection of sensitivity patterns. He was awarded a Ph.D. from VTU in the year 2014. Cryptography, indexing techniques, anonymization methods in generalization and design patterns are his research interests. His academic experiences include research supervisor, BOE member and valuation coordinator for VTU examinations, training and placement officer at RNSIT and technical chair for International Conference ICDECS-2015 and ICDECS-2019 at RNSIT. He has published around 34 articles in reputed journals/conferences. His Industry interaction includes coordinator for Campus Connect program in collaboration with Infosys, trainer for Wipro Wase, BITS off campus Trainer, Trainer on Shell scripting for CDAC, Bangalore. His major achievement includes i) Recipient of Best Teacher Award in 2005 and 2006 at RNS Institute of Technology, ii) published a book on “Introduction to DBMS a simplified Approach”, and iii) he also has a patent published in his name. He has guided many M.Tech and B.E projects and Internships. Currently he is supervising four research candidates at VTU. He can be contacted at email: kiranpmys@gmail.com.