Traffic Intersection Vehicle Movement Counts with Temporal and Visual Similarity based Re-Identification

Sahil Jaiswal[1], Pranamesh Chakraborty[2], Tingting Huang[3], Anuj Sharma[4]

Abstract—**Vehicle movement counting and classification is one of the critical components for traffic intersection monitoring and management. Cameras can be used to determine the vehicle movement counts (left-turning, right-turning, and through movements). Typically, cameras are installed with the view focused towards a particular approach and there is no overlap or very low overlap between the different camera views. Therefore, vehicles need to be re-identified across multiple cameras to detect the complete movement trajectory of the vehicle. In this study, we proposed combining visual similarity obtained using Convolutional Neural Networks (CNN) and temporal similarity (vehicle re-appearance time in cameras) using logistic regression (LR) model to perform vehicle re-identification. The logistic regression model has been used in two stages (without and with hard-negative mining) combining visual and temporal similarity. The results showed that using the hard-negative mining based LR model, the Top@1 results improved by 22% and Top@5 results improved by 8.48%, compared to the results obtained using only visual similarity measure for generating the rankings.**

Index Terms—**traffic intersection monitoring, traffic intersection movement count, vehicle re-identification**

I. Introduction

Traffic intersections are one of the critical components of roadway infrastructure, in terms of traffic operations and traffic safety due to complex interactions between vehicle to vehicle and vehicle to pedestrians. Therefore, it is essential to develop smart traffic intersection monitoring and management for improving traffic safety and mobility.

Signalized intersections are controlled using traffic signals to maintain a steady flow by assigning the right of way to every approach in a systematic manner. To efficiently handle signalized intersections, it is critical to estimate the vehicle movement counts (left-turning, right-turning, and through movement) across different intersection approach legs such that the green time can be provided to all approaches as per the demand. Further, turning movement counts are also required to design facilities, maintenance, and planning at an intersection.

Typically, vehicle movement counts in a traffic intersection can be monitored using loop detectors, radar sensors, or bluetooth devices. However, cameras are typically also installed in traffic intersections, which can be directly used for vehicle movement counts, thereby providing a cheaper solution. In recent times, with advancements in computer vision techniques based Intelligent Transportation Systems (ITS), cameras can be used to estimate the traffic characteristics apart from being used for surveillance alone. Typically, the cameras are provided with its view focused towards a particular approach/intersection leg and and there is no overlap or very low overlap between the different camera views. Hence, a single camera view does not provide the complete view of the intersection. Therefore, when an intersection is monitored using multiple cameras, vehicle re-identification can prove a good way to monitor and manage the traffic.

Vehicle re-identification technique has been an integral part of ITS-based solutions and it can be used to find same vehicle appearing across various cameras at different times. The features obtained from the visual appearance of the vehicles can be used to re-identify the same vehicle in different cameras. Typically, Convolutional Neural Networks (CNNs) have been used as the state-of-the-art approach for image feature extraction and re-identification purposes. Coarse-to-fine features matching along with license plate verification though a Siamese Neural Network and spatio-temporal similarity metric has developed to rank similar looking vehicles [1]. The visual, space and time information are collectively leveraged to improve the re-identification results [2]. Probability model has also been used to determine the chances of reappearance based on visual appearance and spatial-temporal constraints [3].

In this study, we have focused on vehicle re-identification based traffic intersection movement counting. While state-of-the-art re-identification approaches based on CNN visual similarity can be directly used to determine movement counts across different cameras in a traffic intersection, their performance are typically impacted due to high traffic volume with visually similar looking images. Therefore, we propose to use temporal and visual similarity to detect the vehicles efficiently across different cameras in a traffic intersection. We

[1]Sahil Jaiswal is a M. Tech student in the Department of Civil Engineering, Indian Institute of Technology Kanpur, Kanpur, India

[2]Pranamesh Chakraborty is an Assistant Professor in the Department of Civil Engineering, Indian Institute of Technology Kanpur, Kanpur, India

[3]Tingting Huang is a Data Scientist in ETALYC Inc., Ames, Iowa, USA

[4]Anuj Sharma is a Professor in the Department of Civil, Construction, and Environmental Engineering, Iowa State University, Ames, Iowa, USA pranames@iitk.ac.in

The research results are based upon work supported by the Science and Engineering Research Board (SERB) under Sanction Order No. SRG/2020/001708. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the SERB. The authors would also like to thank Dave Ness and Duane Nitcher, City of Dubuque for providing the traffic video data.

have used two different stages of logistic regression with and without hard-negative mining to show the importance of hard-negative mining and benefits of adding temporal similarity along with visual appearance for vehicle movement count generation using traffic cameras.

II. Methodology

Re-identifying and counting vehicle movements across different cameras consists of three broad parts: (a) representative image generation, (b) ranking of visually similar images, and (c) re-ranking based on temporal and visual similarity. Each of these sections are discussed next in details.

A. Representative image generation

The first step for vehicle re-identification is to generate representative vehicle images from the trajectories. These images can then be used to re-identify vehicles across two different cameras in a traffic intersection, thereby providing the directional movement of vehicles in the intersection (left, right or through movements).

Vehicle movement in a camera can be divided into two categories: (1) vehicles approaching the intersection (i.e., moving towards the camera), and (2) vehicles leaving the intersection (i.e., moving away from the camera), as shown in Figure 1. The vehicles approaching the intersection will take a turn at the intersection or go through according to their destination and hence these vehicles will act as queries that need to be identified in another camera of the intersection to determine the vehicle movement class (left, right, or through movement). The vehicles that are leaving the intersection as seen in the camera will be the final appearance of the vehicle at the intersection and these will act as the gallery dataset in which we will search our queries. Therefore, query images need to be extracted from trajectories approaching the intersection, while gallery images extraction will be from vehicle trajectories leaving the intersection.

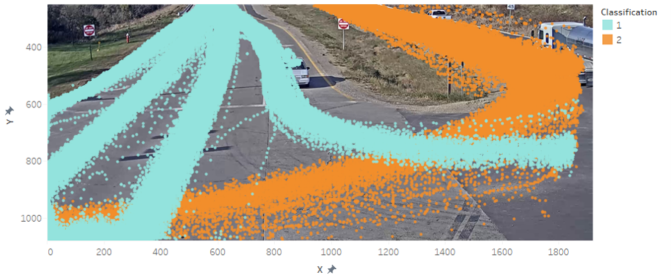

Fig. 1. Two different vehicle movement categories in a camera: Class 1: vehicle approaching the intersection, Class 2: vehicles leaving the intersection

1) Vehicle trajectory selection handling identity switches: The representative images for query and gallery dataset need to be extracted from each vehicle trajectory. However, frequent stopping and heavy occlusion in traffic intersections leads to identity switches while tracking vehicles, thereby leading to multiple track ids of a single vehicle. These cases of identity switches can lead to generation of multiple representative images of same vehicle with different IDs. This problem can be handled by selecting a virtual representative line (VRL) through which all vehicles pass through. This will ensure that only one trajectory passing through that line for each vehicle is selected for representative image generation.

The concept of virtual stop bar, as described in [4], has been used to determine the location of the VRL for the vehicles of query set. A stopped location can be considered as the location where no displacement of vehicles is observed in consecutive video frames. All such locations are extracted using the trajectory data of vehicles and a horizontal line passing through the 50^{th} percentile value of y-coordinates of stopped locations has been marked as the virtual stop bar. Figure 2 shows a sample approach with the stop location points and the corresponding stop bar, generated for that approach. This stop bar line is used as the VRL and all trajectories passing through this are used for generation of representative images of the query set. The gallery images consists of vehicles that are leaving the intersection so the vehicles don't have to stop. This line is drawn manually in the FOV of camera for the vehicles of gallery image set.

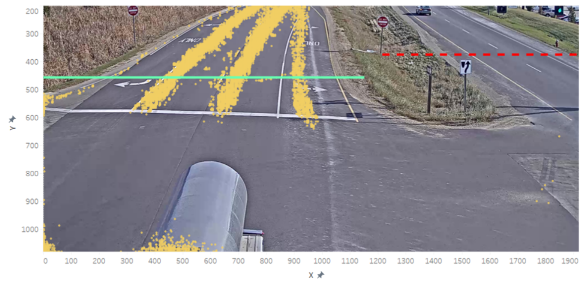

Fig. 2. Virtual representative line generation example: Yellow points indicate no vehicle displacement points, Green line represents the VRL for vehicles of the query dataset, Red-dashed line represents the VRL for vehicles of the gallery dataset

2) Image extraction: Each selected trajectory passing through the VRL contains multiple instances of the same vehicle (i.e., trajectory points). Therefore, we can select more than one instance of a particular vehicle to increase the representativeness of the same vehicle with varying illumination and/or appearance features. This can help to generate a more robust re-identification model. In this study, we have chosen a maximum of 3 representative images for each vehicle (i.e., 3 trajectory points at suitable time-gap) such that the images has at least some amount of variation in appearance features.

Each vehicle trajectory is generated by the tracking-by-detection framework, where the vehicle is detected and localized in a bounding box in every video frame. These bounding box regions need to be cropped out from the video frame to be used as the representative images of the vehicles.

The best representative image of a vehicle can be extracted when it is near to the camera. The area of the bounding boxes starts reducing as the vehicle moves away from the camera. Hence, the area of the bounding box is used to generate the best possible representative image set as well as to avoid the cases of the partial vehicle occurrences. As discussed before, the vehicles in the query dataset are approaching the intersection, so they move closer to the camera and then leave the camera Field of View (FOV). On the other hand, in case of gallery dataset, the vehicles are leaving the intersection, so they move away from the camera and leave the camera FOV. The area vs time (in terms of video frames) is analysed for both query and gallery image sets. For query dataset images, the area of vehicles increases as it moves closer to the camera at the intersection, reaches a peak, and starts decreasing as the vehicle leaves the camera FOV (partial vehicle is typically visible in this region). Figure 3 shows area vs frame plot for such a sample trajectory. To remove partial vehicles appearing in the border regions of the frame, a region of 20 pixels from each edge of the frame is first trimmed. The best representative image set can be obtained from the trajectory points near the peak area. Therefore, in this study, 30% of the trajectory points near the peak area is chosen, as shown in Figure 3. Out of these trajectory points, a maximum of 3 representative image set for each vehicle need to be extracted then using a suitable time gap. In this study, we tested a time gap of 5 frames and 10 frames to determine the best suitable time gap.

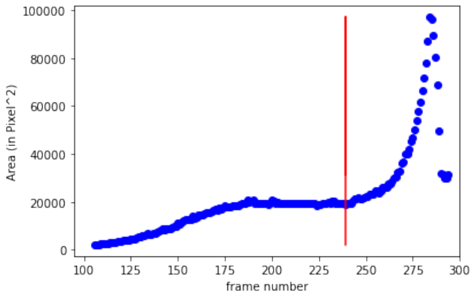

Fig. 3. Area vs. frames plot for a sample trajectory of a query vehicle

Although ideally, we would like to extract 3 representative images for each vehicle in query and gallery dataset, however, taking the time gap of 5 or 10 frames between the images results in vehicle images with very small bounding box areas. To handle this, we filtered out and removed vehicle bounding boxes with height less than 40 pixels [5], and therefore a maximum of 3 representative images has been generated for each vehicle, both in query and gallery dataset.

B. Ranking of visually similar images

Once the representative images of the query and gallery dataset is generated, the first step is to determine the visually similar image pairs and rank them based on the visual similarity distance metric. This is the backbone of standard vehicle re-identification model and consists of two parts: (a) feature extraction and (b) similarity distance metric using the extracted feature vectors.

1) Feature extraction from representative images: To perform vehicle re-identification, vehicle features need to be generated first. In recent years, Convolutional Neural Networks (CNNs) have produced state-of-the-art results for feature extraction from images for different computer vision related tasks such as image classification, object detection, and also vehicle re-identification.

In this study, we have determined the visual similarity distance based on the methodology adopted by [6]. ResNet-50 [7] architecture is used to train the feature extractor based on transfer learning technique. The backbone model is pretrained on ImageNet dataset [8] for an image recognition task. The classification layer of the final model was removed [6] and an adaptive pooling layer was used to output the mean of input feature map in terms of width and height channels. The feature dimensions were reduced to 512-dimensions using a fully connected layer of length 512-d along with a batch normalization layer. This 512-d output vector for every image has been considered as the extracted features for that image. The details of the training and feature extraction can be found in [6].

2) Calculation of similarity between representative images: The features extracted from the CNN model are the mathematical representation of the query and gallery images. The similarity between the images can be determined using these feature vectors to generate ranking between each query image dataset and all gallery images. A suitable distance metric has to be used to convert the similarity of images to a single numerical value. In this study, Mahalanobis distance [9] has been used to obtain the similarity between the features of a query image (q_i) and a gallery image (g_i). The mathematical expression is shown in Equation 1:

$$d(q, g_i) = (x_q - x_{g_i})^T M (x_q - x_{g_i}) \tag{1}$$

where $G = g_i \mid i = 1, 2, 3, .., n$ are the gallery image set or search space of size n for a particular query image q. x_q and x_{g_i} denotes the appearance feature vector of a query image q and a gallery image g_i respectively. M denotes a positive semi-definite matrix.

Each vehicle in query and gallery dataset has a maximum of 3 representative images, the number can be lesser than 3 images, if bounding box height is less than 40 pixels [5]. Therefore, it generates a maximum of 9-dimensional (3x3) distance matrix for each vehicle pair in the query and gallery dataset. To proceed further, we choose the minimum distance obtained in each 3x3 matrix and use them to generate the rank of each query dataset to the gallery dataset images.

However, ranking using visual similarity distance metric only can lead to significant false calls due to large query and gallery dataset with different visually similar images. Therefore, we need to re-rank using both visual and temporal similarity to determine the final re-identification query and gallery image pair. Next, we discuss our re-ranking strategy using visual and temporal similarity.

C. Re-ranking based on temporal and visual similarity

The query and gallery dataset typically consists of a large number of visually similar looking images, thereby resulting in poor performance of re-identification using visual similarity measure alone. The search space (gallery dataset) can be reduced by considering vehicles passing the intersection in a certain time interval after the query vehicle has left from the FOV of query camera and is due for reappearance in one of the camera present at other intersection approach legs. This time constraint can help to reduce the number of vehicles in the search space which has to be matched with the query vehicles thus reducing the search time as well as false positive cases. Therefore, we propose in this study to re-rank the gallery dataset images using temporal similarity, along with the visual features using a logistic regression model, described next.

1) Logistic regression model based re-ranking: Logistic regression (LR) model is used for performing predictive analysis to estimate the probability of any event occurrence determined by a set of features, also called as independent variables. The probability outcome (of the dependent variable) is bounded between 0 and 1. In a discriminative learning, if y is our dependent variable, x is the set of independent variables (or features), then the probability of outcome will be $P(y|x)$. Similar to a linear regression, the estimated value of y ($\hat{y}$) in a logistic regression can be modelled as: $w^T \cdot x$, where w is the vector containing weight (or coefficients) of each independent variable and x is the set of independent variables. The predicted probability values ($\hat{y}$) are mapped to probabilities using sigmoid function, to map the values between 0 and 1.

In this study, we have used the visual similarity distance metric and the reappearance time as the two independent variables in the logistic regression model to re-rank gallery images to query image. Reappearance time is the difference between the last occurrence of query vehicle in the query camera and the first occurrence of any gallery vehicle in the gallery camera. The mathematical expression is shown in Equation 2:

$$T_{diff} = T_{q,L} - T_{g_i,F} \tag{2}$$

Where T_{diff} is the reappearance time, $T_{q,L}$ is the last occurrence of query vehicle in query cam, $T_{g_i,F}$ is the first occurrence of gallery vehicle in gallery cam.

The LR model is used to rank the vehicles in the search space of a query vehicle based on the probability outcome obtained for the query-gallery image pairs through the model. The model is trained as a binary classifier where a true query-gallery image match is labelled as 1 and false query-gallery image match is labelled as 0. The LR model is trained to learn the difference between a true and false query-gallery image match. The search space of a query vehicle contains many vehicles among which there will be only one gallery vehicle that is the same query vehicle which has reappeared in a different camera. Therefore, the model is used generate a probability outcome for every query-gallery image pair in the search space. The pairs will be ranked based on this probability in a descending order.

In this study, the logistic model is trained in two stages. As discussed before, there will be many gallery vehicles in the search space of a single query vehicle and there will be only one true query-gallery match. Therefore, the number of positive and negative samples are not equal. A balanced dataset needs to be created before training the model. In the first stage training, the balanced training dataset contains randomly picked negative samples equal to the number of positive samples present, the logistic regression model is initially trained on this dataset.

In the second stage, the logistic regression model trained in the first stage is used to generate the probability outcomes of every query-gallery image pair in the training data of the first stage. The query-gallery image pairs are ranked within the search space of the query vehicles using these probability outcomes. Now, the training dataset for the second stage is recreated by using the set of negative samples (N) generated using two parts: (1) The hard negative samples which ranked in the top 5 with a probability outcome higher than 0.5, these are the samples among which a model can get confused to make a decision, and (2) The negative samples ranking below 5 with a probability outcome lower than 0.5. The mathematical representation is shown in the Equation 3.

$$N \in \begin{Bmatrix} (rank \leq 5 \ \cap \ probability \geq 0.5) \cup \\ (rank > 5 \ \cap \ probability < 0.5) \end{Bmatrix} \tag{3}$$

III. Dataset

The videos used in this study were obtained from a traffic intersection in Dubuque, Iowa, USA. Every approach leg of the intersection is covered by a different camera. The video data consists of 4 time-synchronized videos with a 10 fps (frames per second) pace, each lasting 1 hour. Figure 4 shows the camera views of the 4 approach legs of the study intersection. We chose this intersection because the camera views helps to determine the vehicle movement direction, which can be used for generating the groundtruth query-gallery image pair. The unmasked regions shown in the Figure 4 has been used to generate the representative images of the query and gallery dataset such that there is very low overlap between multiple camera views.

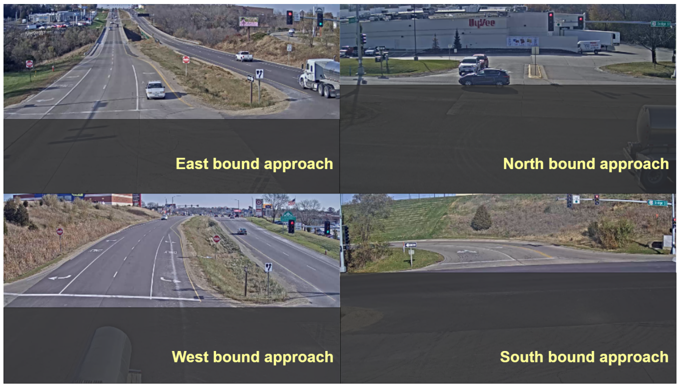

Fig. 4. Camera view of the 4 approaches of study intersection

A. Ground truth generation for re-identification

A vehicle passing an intersection is expected to appear in at least 2 cameras, as per the camera setup considered in this study. The inter camera ground truths are the pairs of query-gallery images of the same vehicle appearing in different cameras at the intersection. The ground truths represent the turns taken by the vehicle as it contains a pair of representative images taken from different approach leg of an intersection at different time instants hence depicts the complete movement across the intersection.

The inter camera ground truths generated for the re-identification analysis contains the vehicles occurring in all the cameras in the video dataset for a duration of 1 hour. The total re-identification dataset contains 2756 query vehicles in a 1-hour duration. Out of this, the Eastbound, Westbound, Northbound, and Southbound approaches contains 1283, 1203, 148, and 122 query vehicles, respectively. This dataset is divided into two parts for training and testing of the re-identification analysis. The training and testing dataset contains vehicle occurring in 30 minutes duration. The first 30 minutes duration out of the 1-hour data is taken as training and the later 30 minutes is taken as testing. The training and testing dataset contains 1313 and 1443 vehicles respectively.

B. Vehicle Trajectory Generation

The vehicle's trajectory information is crucial for the task of vehicle re-identification. In this study, we have used a tracking-by-detection framework to solve the multiple object tracking (MOT) problem of trajectory extraction from the videos. In this study, the object detection task was carried out using YOLOv5 [10], a more recent version of the most popular real-time object detection architecture, YOLO (You Only Look Once) [11]. The YOLOv5 model, pretrained on Microsoft COCO dataset [12] was used for detecting the object classes motorcycle, car, bus, and truck. A tracking algorithm is then employed to generate the trajectory by using detections obtained by the YOLOv5 model on the video dataset's frames. Simple Online Realtime Tracking (SORT) [13] is used as the tracking algorithm to generate the trajectory using the detections of the objects obtained from the YOLOv5 model.

IV. Results and Discussions

A. Representative image generation

As discussed in the Section II-A2, there will be a maximum of 3 representative images for any vehicle having a suitable time gap between every image extracted. A higher time gap can result in generating representative images with significant variation in visual appearances, thereby generating a robust re-identification. However, the concern about generating sequential representative images with a higher time frame gap is that the size of images decreases as vehicle move away from the camera with time. KITTI dataset [5] considers a minimum of 40 pixels height as the cut off for their analysis. In this study, we tested, 5 frames and 10 frames gap for generating representative images. It was observed that 25% of vehicles generated with 10 frames gap have bounding box height less than the critical threshold of 40 pixels, while the number was only 5% using 5 frames gap. Therefore, we have chosen 5 frames gap for generating representative images.

B. Search space analysis

The logistic regression model was trained in two stages based on the criteria for picking negative samples to create a balanced dataset, as discussed in Section II-C1. The results obtained from the two stages LR models (without hard negatives and with hard negatives) is shown in the Table I. The model results have been compared with the ranks generated by using only the visual similarity value, shown in Table I.

TABLE I
Results on test data

Metrics	**Logistic Regression model**		**Using similarity measure only**
	1st Stage	2nd Stage	
Top @ 1	53.08 %	70.76 %	58 %
Top @ 5	98.13 %	98.27 %	90.58 %

The Top@1 result is initially better by using visual similarity measure only compared to the first stage logistic regression model due to the fact that the model is trained on randomly picked negative samples. Therefore it is not able to differentiate between true positives and hard negatives effectively. However, when the model is trained for the second stage using the hard negatives (as per Equation 3), a significant improvement in both the Top@1 and Top@5 accuracy is observed compared to only using the visual similarity measure only. The Top@1 results improved by 22% and Top@5 results are improved by 8.48%, compared to the results obtained by using the visual similarity measure only for ranking.

Table II shows the rank improvement and rank degradation when the finally trained LR model (second stage) is used over the visual similarity measure method. It can be seen that the LR model (second stage) helped to accurately detect 271 additional vehicles as Rank 1, compared to using visual similarity measure alone. Only 80 vehicles rank, on the other hand got downgraded to rank 2 compared to rank 1 obtained using visual similarity. Further, the Cumulative Matching Characteristics (CMC) curve is also plotted against the ranks to visualize the performance of all the methods at various ranks, shown in Figure 5. Figure 5 and Table II shows the efficacy of using a logistic regression based visual temporal distance metric for vehicle re-identification and counting in traffic intersections, compared to using visual similarity alone.

C. Turning movement count

The application of vehicle re-identification for this study is to count the turning movement at an intersection where each leg of the intersection is monitored by a different camera rather than a single camera covering the whole intersection. The turning movement undertaken by a query vehicle is determined by using the Top@1 query-gallery vehicle pair within the search space of query vehicle because the actual

TABLE II
RANKS IMPROVED AND RANKS DOWNGRADED BY USING LR MODEL (SECOND STAGE) OVER USING VISUAL SIMILARITY METRIC ONLY

Logistic Regression (Second stage) over using similarity measure only					
Ranks improved			**Ranks downgraded**		
Initial rank	Final rank	Count	Initial rank	Final rank	count
2	1	154	1	2	80
3		59		3	10
4		24		4	0
5		14		5	0
>5		20		>5	0
Total		271	Total		90

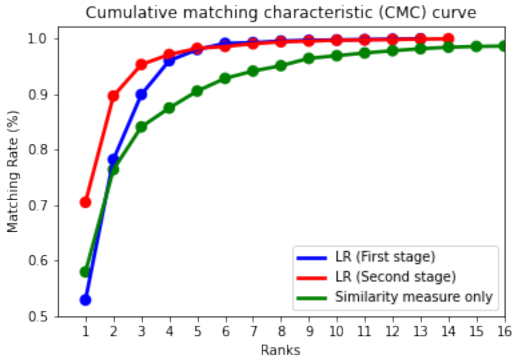

Fig. 5. CMC curve

turning movement cannot be determined by considering vehicles at multiple ranks. Table III shows the turning movement ground truths and the turning movement count that could be captured by using the Top@1 of the final logistic regression model (second stage). Total 1018 turning movements could be detected out of 1443 ground truths. Although this results in 70% accuracy in detecting vehicle movements, however the accuracy is significantly higher than the 58%, achieved using visual similarity alone.

TABLE III
TURNING MOVEMENT COUNTS

Approach	**Turning Movement Counts (Ground Truths)**			**Turning Movement Counts (Logistic Regression model)**		
	Left	Through	Right	Left	Through	Right
EB	35	528	87	31	378	60
NB	52	17	8	34	10	7
WB	5	620	27	3	422	27
SB	24	17	23	20	8	18
Total	116	1182	145	88	818	112

V. CONCLUSION

In this study, we have proposed vehicle re-identification using visual and temporal similarity to generate traffic movement counts in an intersection. A logistic regression based method have been used in two stages (without and with hard-negative mining) combining visual and temporal similarity. A CNN based visual feature extractor, specifically trained for extracting robust vehicle features has been used to extract visual features from the representative images and based on these features, the similarity for every query-gallery image pair was calculated using a distance metric. A two-stage progressive learning approach was used to train the logistic regression model using similarity measure and reappearance time as features. By using the proposed approach, the Top@1 results improved by 22% and Top@5 results improved by 8.48%, compared to the results obtained using the visual similarity measure only for generating the rankings.

REFERENCES

[1] X. Liu, W. Liu, T. Mei, and H. Ma, “A deep learning-based approach to progressive vehicle re-identification for urban surveillance,” in *European Conference on Computer Vision*. Springer, 2016, pp. 869–884.

[2] Y. Shen, T. Xiao, H. Li, S. Yi, and X. Wang, “Learning deep neural networks for vehicle re-id with visual-spatio-temporal path proposals,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1900–1909.

[3] W. Huang, R. Hu, C. Liang, Y. Yu, Z. Wang, X. Zhong, and C. Zhang, “Camera network based person re-identification by leveraging spatial-temporal constraint and multiple cameras relations,” in *International Conference on Multimedia Modeling*. Springer, 2016, pp. 174–186.

[4] K. R. Santiago-Chaparro, M. Chitturi, A. Bill, and D. A. Noyce, “Automated turning movement counts for shared lanes: leveraging vehicle detection data,” *Transportation Research Record*, vol. 2558, no. 1, pp. 30–40, 2016.

[5] A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for autonomous driving? the KITTI vision benchmark suite,” in *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 2012, pp. 3354–3361.

[6] Z. Zheng, T. Ruan, Y. Wei, Y. Yang, and T. Mei, “VehicleNet: Learning robust visual representation for vehicle re-identification,” *IEEE Transactions on Multimedia*, vol. 23, pp. 2683–2693, 2020.

[7] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2016, pp. 770–778.

[8] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, “ImageNet large scale visual recognition challenge,” *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[9] Z. Zhong, L. Zheng, D. Cao, and S. Li, “Re-ranking person re-identification with k-reciprocal encoding,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1318–1327.

[10] X. Zhu, S. Lyu, X. Wang, and Q. Zhao, “TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2778–2788.

[11] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[12] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft COCO: Common objects in context,” in *European Conference on Computer Vision*. Springer, 2014, pp. 740–755.

[13] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, “Simple online and realtime tracking,” in *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2016, pp. 3464–3468.