Espinosa, J.E., Velastin, S.A., y Branch, J.W. (2019). Detection and Tracking of Motorcycles in Congested Urban Environments Using Deep Learning and Markov Decision Processes. In *Mexican Conference on Pattern Recognition 2019*, 11524, pp. 139-148.

DOI: https://doi.org/10.1007/978-3-030-21077-9_13

Detection and Tracking of Motorcycles in Congested Urban Environments Using Deep Learning and Markov Decision Processes

Jorge E. Espinosa[1(✉)] , Sergio A. Velastin[2,3,4] , and John W. Branch[5]

[1] Politécnico Colombiano Jaime Isaza Cadavid,
Carrera 48 No. 7-151 El Poblado, Medellín, Colombia
jeespinosa@elpoli.edu.co
[2] Cortexica Vision Systems Ltd., London, UK
sergio.velastin@ieee.org
[3] Queen Mary University of London, London, UK
[4] University Carlos III Madrid, Madrid, Spain
[5] Universidad Nacional de Colombia – Sede Medellín,
Calle 59 A N 63-20, Medellín, Colombia
jwbranch@unal.edu.co

Abstract. This research describes "EspiNet", a Deep Learning Convolutional Neural Network model, in conjunction with a Markov Decision Process (MDP) tracker for detection and tracking of occluded motorcycles in urban environments. The model is trained and evaluated, using a new public dataset with up to 10,000 annotated images, created for this research, and captured in real urban traffic scenes. Images were captured using a moving camera mounted in a drone, where more than 60% of the motorcycles are affected by occlusions. The network design involves many tests, where a promising result of 88.84% in average precision (AP) is achieved, despite the considerable number of occluded vehicles, the movement of the camera and the low angle used for capture. The model predictions are used as input to an MDP tracker, reaching results up to 85.2% in Multiple Object Tracking Accuracy (MOTA). The proposed network architecture outperforms state of the art YOLO (You Look Only Once) v3.0 and Faster R-CNN (VGG16 based) detection models, producing also better tracking results in comparison with the use of the other two models as detector base for the MDP tracker.

Keywords: Motorcycle detection; Motorcycle tracking; Faster R-CNN; Region based detector; CNN; Deep learning; Occluded images; Markov Decision Process

1 Introduction

1.1 Motorcycles as Part of Urban Traffic

Motorcycles are currently one of the most popular means of transport in emerging countries, which results in important fatality rates [4] and a significant environmental impact due to emissions (e.g. P.M. 2.5) [19].

As an example of conditions in emerging countries, the annual report of Traffic Accidents of the Andean Community 2007–2016 [1], indicates that in 2017, for Bolivia, Colombia, Ecuador and Perú, of the 347,642 road accidents 88% correspond to urban occurrences. Colombian road users most affected by traffic accidents are motorcyclists, representing 49.82% of reported deaths and 56.36% of non-fatal injuries. Of the total drivers, motorcyclists represented 78.81% of the dead and 80.51% of the injured and for their passengers the figures were 50.69% and 48.99%, respectively [2].

Therefore, it is important to implement traffic management techniques or strategies starting from the detection and tracking of motorcycles to reduce accidents. The use of Intelligent Transportation Systems (ITS) and video analysis in particular, could be one way of helping address issues affecting road safety.

In this research we introduce EspiNet, a CNN model inspired on Faster R-CNN [15] combining it with a Markov Decision Process (MDP) tracker for the tasks of detecting and tracking motorcycles in urban traffic video sequences, especially under occlusion, characteristic of road conditions in emerging countries. General vehicle detection, under this condition, has been studied by many authors, benchmarking their results mainly using the KITTI dataset [10], which unfortunately lacks a motorcycle category. For this reason we have created and used a new public motorbike dataset of 7,500 and 10,000 annotated images, captured on a public road using a camera mounted in a drone.

The main contributions of this work are

1. The publication of a realistic annotated video dataset of motorcycle traffic, presenting realistic occlusion conditions;
2. The proposal of a new convolutional model, EspiNet, inspired by Faster R-CNN to detect motorcycles;
3. The combination of EspiNet and an MDP tracker that obtains competitive results and that provides a baseline for other researchers to improve upon.

This paper is organized as follows: Sect. 1.2 describes works related with vehicle and motorbike detection. Section 2 describes the motorcycles dataset created for this research. Section 3 shows the EspiNet model, describing its main improvements w.r.t to Faster R-CNN, and providing an insight about the advantages of its architecture. Section 4 the MDP tracking strategy used in this research. Section 5 shows the experimentation done for detection and tracking, and a comparative study with state of the art detectors trained end-to-end for this purpose. Finally, Sect. 6 presents the conclusions around the proposed model and the directions for future research.

1.2 Motorcycle Detection

Traditionally, video technique analysis requires reliable methods for object feature extraction to obtain accurate classification results. Video detection systems in the last decade are implemented through discrimination capabilities on appearance features. Motorcycle detection and classification using appearance features include the construction of 3D models [5], dimensions of the vehicles [8], there is also used colour, symmetry, shadows, texture and geometrical fea-tures (e.g. circles) as wheel contours [7]. Description of features as histogram of oriented gradients (HOG) used for detection of helmet in motorcycle riders [16]. There are also variations of HOG [6] and the use of scale-invariant feature transform (SIFT), DSIFT and speeded up robust features (SURF) [17].

Deep learning theory (DL) has emerged as an important breakthrough in the field of computer vision in the last nine years, with astonishing results in image processing. This theory has been successfully used in vehicle detection, mainly based on DL general object detectors. These detectors can be divided into region based stage detectors and single stage detectors. Region based detectors involve two general components, the region proposal step (RPN) and the classifi-cation step. R-CNN [11], combines selective search algorithm for region proposal (RPN) and CNN features to perform object detection. This model was used in [3] to classify motorcycles according to the USA Federal Highway Administration (FHWA) scheme. There is also work based on single stage detectors, where a single convolutional architecture simultaneously predicts bounding boxes and class scores associated. Huynh et al. [13] designed a network for this purpose, work-ing on top-view captured images, that significantly reduces occlusion between urban objects. Other approaches based their detection of moving objects on background subtraction methods using a pre-trained network (AlexNet) for fea-ture extraction for helmet detection [18]. However there are no reports using an open dataset of motorcycles on congested urban environments.

2 The Dataset

Real urban traffic present occlusions between vehicles or with regular urban furniture (Fig. 1). Despite that there had been effort to construct a dataset such as KITTI [10], where occluded vehicles are annotated, and state of the art algorithms benchmark their performance, there is not motorcycle category created on this dataset or any dataset public available explicitly oriented to occluded motorcycles in urban environments.

This is why we have created an annotated motorcycle public dataset, which contains images taken from a camera mounted in a drone, subject to subtle unstable conditions. To speed up processing analysis, images were resized to one-third of their original size, reaching 56,975 ROI (Region of Interest) annotated motorcycles, and 25 pixels as the minimal height size. 60% of the annotated motorcycles has a level of occlusion. Objects partially occluded smaller than 25 pixels were discarded. The ground truth generated is specified in an XML file that describes the class, frames covered by the object, Name, Id, height

Fig. 1. Original dataset image vs. annotated image. Note the small object size of some annotated objects and the variate level of occlusions.

Table 1. Details of the new dataset

Dataset	Classes	Objects	Min. vertical size	Occlusion	Format
MotorBikes7,500	1	221	25 px	>60%	XML
MotorBikes10,000	1	317	25 px	>60%	XML

and width of the bbox surrounding the object. Table 1 describes these datasets, available on the internet[1].

3 EspiNet

The EspiNet model is based on Faster R-CNN. The difference here is that we create a more compact model, hence with faster inference, with just 4 layers of convolution. EspiNet is the evolution of the architecture described in [9]. Now, the number of convolutional layers has been increased to four, to capture more discriminating features. All the convolutional filters implemented are of size [3 3]. The first convolutional layer includes 64 filters, used to work with the three image channels and capturing primitive features. These primitive features are fed to a second convolutional layer, with 32 filters to create more complex features, which are aggregated even more in the third (64 filters) and fourth (128 filters) layer. As in [9], this architecture is used simultaneously as a region proposal network (RPN) and detection network. Figure 2 shows the described model able to identify motorcycles even under occluded scenarios.

The optimization algorithm used for training the model is Stochastic Gradient Descent with Momentum (SGDM). The training comprises four steps learning shared parameters for the RPN and detector networks. The RPN and the detector network are trained separately. EspiNet uses a learning rate of 1e-5 for these two steps since they require a quicker convergence. In the last two steps, the shared convolutional layers are fixed, fine-tuning the layers unique to RPN and detector network. In these last steps, the learning rate is set to 1e-6 for a smooth fine-tuning process.

[1] http://videodatasets.org/.

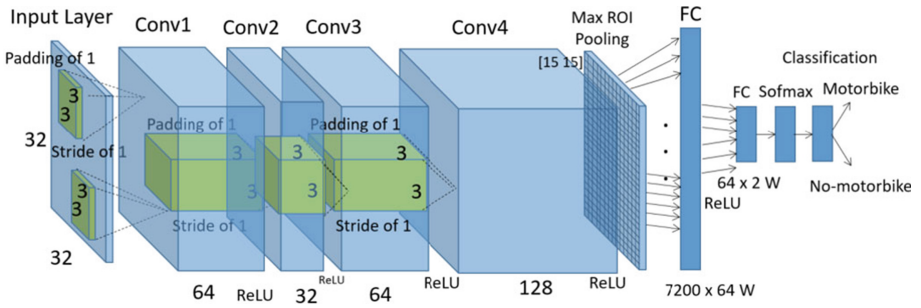

Fig. 2. The EspiNet CNN model. Used simultaneously for RPN and for classification

For RPN training, positives image examples patches are those which have 0.6 to 1.0 overlapping with the ground truth boxes. Negative ones have overlapping of 0 to 0.3. The overlapping criteria used is IoU (Intersection over Union) or Jaccard coefficient.

4 Multi-Object Tracking (MOT)

Multiple object tracking (MOT) is a challenging task, since it implies localizing and following multiple objects in the same scene during their life in the sequence, even under occlusion.

4.1 Online Multi-Object Tracking by Decision Making

We adopt the work proposed for online multi-object tracking based on Markov Decision Processes (MDP) [20], where the object tracked lifetime is modeled by using a MDP with four state subspaces: Active, Tracked, Lost and Inactive. With the state transitions the appearance/disappearance of objects is modeled in the MDP tracking process. Data association is performed by learning (reinforcement learning) a similarity function within the MDP policy scheme. The modular architecture of the framework can be combined with different object detection approaches, single object tracking and data association techniques which could be used for learning the MDP policy. We tested this tracking framework on the challenging new dataset (Sect. 2), and evaluated tracking performance for Faster R-CNN (VGG16 based), YOLO v3, and EspiNet, all models are trained end to end.

The framework is categorized as a hybrid learning model, implementing off-line learning using supervision from ground truth trajectories, and online-learning method while tracking objects on training examples, making the MDP able to decide supported on the history and the current status of the target.

Table 2. Comparative detection results EspiNet model against a state of the art trained network Faster-R-CNN (based on VGG16) [12] and YOLO [14]

Metrics	EspiNet	Faster R-CNN	YOLO
Precision (%)	**93.7**	57.3	93.0
Recall (%)	**90.0**	76.3	81.0
F1 score (%)	**91.8**	65.4	86.6
AP (Average Precision)	**88.84**	68.75	80.72

5 Experiments

5.1 Detection

For comparative purposes, we chose state of the arts models, representative of single stage detectors (YOLO [14]) and the region based Faster-R-CNN (VGG16 based) [12]. For a fair comparison, all these models were trained end-to-end using the challenging 10,000 examples dataset (Sect. 2). Evaluating the model we achieved an Average Precision (AP) of 88.84% and an F1-score of 91.8%, outperforming results of Faster-R-CNN (based on VGG16) and YOLO (Table 2).

In all metrics, EspiNet outperforms the other two detectors, the closet best performance is achieved by YOLO that showed almost the same precision but poorer recall because the single stage model lacks RPN and thus it can fail to detect objects that appear too close or too small. Video results can be viewed in the following link[2].

5.2 Tracking

The tracking algorithm is evaluated also in the new dataset described in Sect. 2. For comparative purposes, we used the detection results of EspiNet, YOLO and Faster R-CNN (VGG-16 based) described in Sect. 5.1.

Tracking Performance Metrics. Based on the metrics defined in the MOT challenge[3], we evaluated the results of the modified MDP, tracking detections from the new dataset. Two main performance metrics were used:
Multiple Object Tracking Precision (MOTP) $MOTP = \frac{\sum i, td_t^i}{\sum t, C_t}$ which calculates the total error for matched pairs of ground truth-tracker pairs over all frames, averaged by the total number of matches found.
Multiple Object Tracking Accuracy (MOTA) $MOTA = 1 - \frac{\sum_t (m_t + fp_t + mme_t)}{\sum_t g_t}$ where m_t, fp_t and mme_t are the number of misses, of false positives, and of mismatches, respectively, for time t.

[2] https://goo.gl/RSZQGz.
[3] https://motchallenge.net/.

Additionally, other metrics are used that allow understanding tracking behaviour: Mostly Track targets (MT), percentage of ground truth tracks covered by the tracking mechanism for at least 80%. Mostly Lost targets (ML, percentage of ground truth tracks covered by the tracking mechanism less than 20%. ID Switches (IDS) ID of the tracks that are erroneously changed by the algorithm. Fragmentation (Frag) corresponds to the total number of times a trajectory appears fragmented.

Tracking Results. For training proposes, the 7,500 section of the dataset 2 was used. For data association in the Lost state, we use $K = 10$ as trade off as suggested in [20]. The tracking results described in Table 3 show the importance of using a quality detector for the MDP tracking process. According to the results presented 5.1, the use of EspiNet as a detector for the MDP tracker offers the best results achieving a MOTA of 85.2% or tracking accuracy, which evaluates how many mistakes the tracker made in terms of misses, false positives, mismatches and failures to recover tracks. This result is consistent with a good recall result (90.0), reflected in a low false alarm rate (0.24) and a relative low number of ID switches (89). The use of EspiNet as a base detector reports also a better MOTP (82.5%), or tracking precision which expresses how well the exact position of the motorcycles is estimated. These results are consistent with a slightly better precision compared to YOLO and Faster R-CNN, which allows the MDP tracker to obtain less fragmentations (341). Video tracking results can be viewed via the following link[4].

Figure 3 shows an annotated frame and the results of the MDP tracker using the above mentioned detectors. In this comparative figure, it is possible to see the difference between the detectors and the ground truth, where the motorcycle tracking is annotated with a different count number that the one generated by different detectors. The more approximate numbers of the tracker objects to the ground truth are in this order Espinet: (d), YOLO (c) and Faster R-CNN (b).

An Nvidia Titan X (Pascal) 1531 Mhz GPU is used for training the EspiNet model and the Faster R-CNN model (VGG 16 based), both installed on a Windows 10 Machine with a CPU core i7 7th generation 4.7 GHz, with 32 GB of RAM. The training process on the dataset in EspiNet model took 32 h for training, and 57 h for Faster R-CNN (VGG 16). For YOLO training a Titan Xp 1582 Mhz GPU was used, running with Ubuntu 16.04.3, with a Xeon E5-2683 v4 2.10 GHz CPU, and 64 GB of RAM, taking 18 h for training. All models were trained from scratch. Meanwhile the training of the MDP tracker took 18 h on the Windows environment.

6 Discussion

This research has combined EspiNet and an MDP Tracker for motorcycle detection and Tracking in urban scenarios. The model can deal with highly occluded

[4] https://goo.gl/rLL2Le.

Table 3. Comparative results for MDP tracking using detectors EspiNet (Sect. 3), Faster-R-CNN (based on VGG16) [15] and YOLO [14]

Metrics	EspiNet	Faster R-CNN	YOLO
Recall	**90.0**	76.3	81.0
Precision	**93.7**	57.3	93.0
False Alarm Rate	**0.24**	2.04	0.38
GT Tracks	318	318	318
Mostly Tracked	**283**	113	142
Mostly Lost	2	7	2
ID Swiches	**89**	808	111
Fragmentations	**341**	1766	420
MOTA	**85.2**	34.1	71.4
MOTP	**82.5**	72.7	77.8

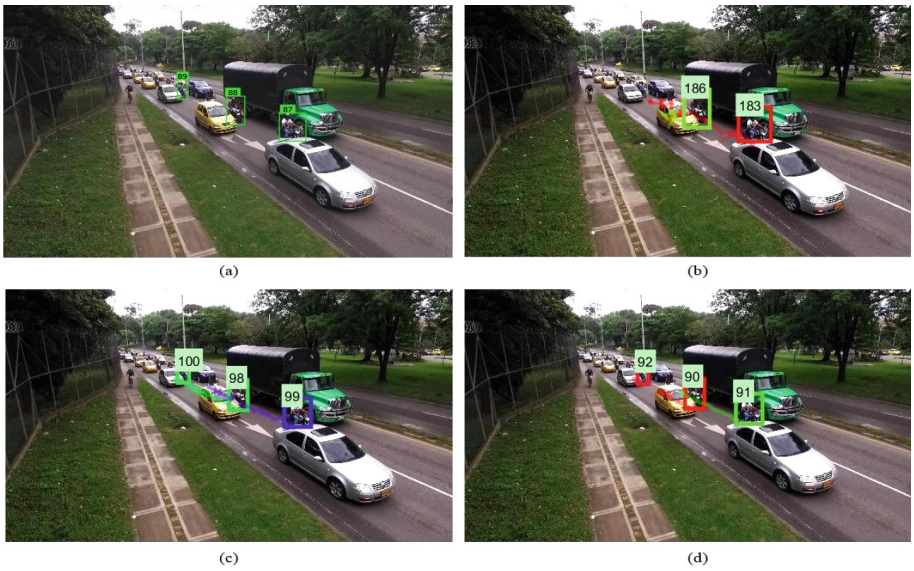

Fig. 3. Different detector effects on the MDP tracker. (a) Ground truth. (b) MDP tracker using Faster R-CNN as a base detector. (c) MDP tracker using YOLO as a base detector. (d) MDP tracker using EspiNet(Ours) as a base detector.

images and achieves in detection an Average Precision of nearly 90% and an F1 score of 91.8% for a newly annotated motorcycle urban dataset. Results can be compared with state of the art algorithms benchmarked in sites such as KITTI [10].

We have illustrated the importance of the use of an accurate detector in the implementation of the Multi-object tracking frameworks based on Markov decision process [20]. The tracker modeled the life time of the tracked object using four sub-space states (Active, Tracked, Lost and Inactive). The state transition handled by the MDP algorithm is now improved by the quality of the detections to handle incoming and leaving objects, as well as birth and death of the object being tracked.

Several tests were carried out to evaluate the influence of the detector in the MDP tracking. Lower quality of detections also produce poorer tracking, losing the tracking of some objects or forcing some tracks to drift.

Nevertheless the MDP tracking algorithm helped to preserve the identity of the detected object and the results showed performance close to the state of the art achieving a Multiple Object Tracking Accuracy of 85.2%, using a well trained EspiNet as base detector for the MDP tracker.

We have defined a model capable of detecting and following motorcycles in urban scenarios with a high level of occlusion, which represents an important aid to CCTV surveillance centers in emerging countries, where this type of vulnerable road users exceeds 50% of the vehicular park.

Results of the use of EspiNet + MDP tracker in CCTV surveillance center, can be retrieved from[5].

As future work we plain to move to an integrate deep learning model for simultaneously detect and track objects exploiting spatio-temporal features for improve detection and speed up tracking.

Acknowledgments. This work was partially supported by COLCIENCIAS project: Reduccion de Emisiones Vehiculares Mediante el Modelado y Gestion Optima de Trafico en Areas Metropolitanas - Caso Medellin - Area Metropolitana del Valle de Aburra, codigo 111874558167, CT 049-2017. Universidad Nacional de Colombia. Proyecto HERMES 25374. The authors gratefully acknowledge the support of NVIDIA Corporation with the donation of GPUs used for this research.

References

1. Accidentes de tránsito en la Comunidad Andina, 2007–2016. http://intranet.comunidadandina.org/Documentos/DEstadisticos/SGDE800.pdf
2. Mortalidad por accidentes de tránsito, June 2018. https://www.asivamosensalud.org/salud-para-ciudadanos/mortalidad-por-accidentes-de-transito
3. Adu-Gyamfi, Y.O., Asare, S.K., Sharma, A., Titus, T.: Automated vehicle recognition with deep convolutional neural networks. Transp. Res. Rec.: J. Transp. Res. Board **2645**, 113–122 (2017)
4. Bazargani, H.S., Vahidi, R.G., Abhari, A.A.: Predictors of survival in motor vehicle accidents among motorcyclists, bicyclists and pedestrians. Trauma Mon. **22**(2) (2017). https://doi.org/10.5812/traumamon.26019. http://traumamon.neoscriber.org/en/articles/13364.html. Accessed 26 Sept 2017

[5] https://goo.gl/8N5iZM.

5. Buch, N., Orwell, J., Velastin, S.A.: Urban road user detection and classification using 3D wire frame models. IET Comput. Vis. **4**(2), 105–116 (2010). https://doi.org/10.1049/iet-cvi.2008.0089
6. Chen, Z., Ellis, T., Velastin, S.A.: Vehicle detection, tracking and classification in urban traffic. In: 2012 15th International IEEE Conference on Intelligent Transportation Systems, September, pp. 951–956 (2012). https://doi.org/10.1109/ITSC.2012.6338852
7. Duan, B., Liu, W., Fu, P., Yang, C., Wen, X., Yuan, H.: Real-time on-road vehicle and motorcycle detection using a single camera, pp. 1–6. IEEE (2009). http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4939585
8. Dupuis, Y., Subirats, P., Vasseur, P.: Robust image segmentation for overhead real time motorbike counting. In: 2014 IEEE 17th International Conference on Intelligent Transportation Systems (ITSC), October, pp. 3070–3075 (2014). https://doi.org/10.1109/ITSC.2014.6958183
9. Espinosa, J.E., Velastin, S.A., Branch, J.W.: Motorcycle detection and classification in urban Scenarios using a model based on Faster R-CNN. arXiv preprint arXiv:1808.02299 (2018)
10. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? The KITTI vision benchmark suite, p. 3354–3361. IEEE (2012). http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6248074. Accessed 27 Oct 2016
11. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, June, pp. 580–587 (2014). https://doi.org/10.1109/CVPR.2014.81
12. Huang, J., et al.: Speed/accuracy trade-offs for modern convolutional object detectors. arXiv:1611.10012, November 2016
13. Huynh, C.K., Le, T.S., Hamamoto, K.: Convolutional neural network for motorbike detection in dense traffic. In: 2016 IEEE Sixth International Conference on Communications and Electronics (ICCE), July, pp. 369–374 (2016). https://doi.org/10.1109/CCE.2016.7562664
14. Redmon, J., Farhadi, A.: YOLOv3: an incremental improvement. arXiv:1804.02767, April 2018
15. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks, pp. 91–99 (2015). https://goo.gl/A3FHD7
16. Silva, R.R., Aires, K.R., Veras, R.M.S.: Detection of helmets on motorcyclists. Multimed. Tools Appl. **77**, 1–25 (2017)
17. Thai, N.D., Le, T.S., Thoai, N., Hamamoto, K.: Learning bag of visual words for motorbike detection. In: 2014 13th International Conference on Control Automation Robotics Vision (ICARCV), December, pp. 1045–1050 (2014). https://doi.org/10.1109/ICARCV.2014.7064450
18. Vishnu, C., Singh, D., Mohan, C.K., Babu, S.: Detection of motorcyclists without helmet in videos using convolutional neural network. In: 2017 International Joint Conference on Neural Networks (IJCNN), May, pp. 3036–3041 (2017)
19. Walsh, M.P.: PM 2.5: global progress in controlling the motor vehicle contribution. Front. Environ. Sci. Eng. **8**(1), 1–17 (2014)
20. Xiang, Y., Alahi, A., Savarese, S.: Learning to track: online multi-object tracking by decision making. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4705–4713 (2015)