Intelligent traffic city management from surveillance systems (CERTH-ITI)

Konstantinos Avgerinakis Panagiotis Giannakeris Alexia Briassouli

Anastasios Karakostas Stefanos Vrochidis

Ioannis Kompatsiaris

Centre for Research and Technology Hellas (CERTH) - Information Technologies Institute (ITI)

koafgeri@iti.gr

giannakeris@iti.gr abria@iti.gr stefanos@iti.gr akarakos@iti.gr ikom@iti.gr

Abstract

Surveillance and more specifically traffic management technologies constitute one of the most intriguing aspects of smart city applications. In this work we investigate the applicability of an object detector for vehicle detection and propose a novel hybrid shallow-deep representation to surpass its limits. Furthermore, we leverage the detector's output, so as to localize new vehicles and track them throughout the whole duration that they exist in the video scene. The detection and tracking system is then evaluated and compared with other State-of-the-Art algorithms on the new developed NVIDIA AI city datasets.

1. Introduction

Smart city technologies and more specifically assistive transportation and safe driving make up one of the most intriguing domains of computer science and have attracted significant attention during the last decade. Close-Circuit Television (CCTV) systems and other types of visual monitoring infrastructure provide vast amounts of potentially useful data for optimal management of traffic, safety in crowded urban environments, mitigation of traffic issues in adverse weather conditions and numerous other applications of traffic monitoring. Moreover, increasing industry trends towards autonomous driving, vehicles, and transportation in general, is changing the landscape of traffic management. The data from traffic cameras (static, mobile, drone-based and others, depending on the application) will, in the near future, also be used to manage autonomous vehicle navigation, by sending information about events elsewhere in the city, traffic conditions, pedestrian congestion, to optimally guide vehicles [15].

The wealth of information in these videos remains inaccessible without their automated analysis, as the manual extraction of information from them is very time consuming and cumbersome, making automated video analysis methods a necessity. To this end, numerous computer vision and machine learning solutions have been developed for automatic object detection and tracking in traffic, abnormal event detection in videos of crowded scenes (e.g. pedestrians, traffic), human activity recognition and others. The improving accuracy of visual analysis of traffic videos and its decreasing computational cost, in combination with the increasing use of GPUs, is facilitating the automation of traffic video analysis in recent years. However, the large amounts of surveillance data, in addition to the lack of annotations of these datasets, create obstacles for the development of automated analysis methods. For this reason, large scale annotation efforts and real-world benchmarking challenges and competitions are necessary, to help researchers develop novel, highly efficient and useful solutions in this domain. In this work, CERTH-ITI investigates the performance of State-of-the-Art (SoA) object detector [13] and proposes a novel hybrid one, namely **DeepHOG**, that combines shallow with deep representation schemes in order to improve the detection of the former. Furthermore, CERTH-ITI introduces a tracking algorithm that uses deepHOG's bounding boxes to localize new vehicles in the video scene and monitor them throughout the whole duration that exist inside it. It is of our greatest belief that shallow descriptors and more specifically the relation amogst them(i.e. bag-of-word, fisher vectors) can be combined with Deep Learning SoA techniques, such as Faster R-CNN, so as to introduce a great boost in the representation framework. CERTH-ITI hopes that the proposed framework will help tackling traffic congestion, safety and security issues related to traffic and urban areas and participate in the development of safer,

more liveable and enriching smart urban environments.

2. Related work

Vehicle detection constitutes an essential subcategory of object detection and generally it follows the same framework in order to accomplish its purposes; (a) Object localization is initially performed so as to find the regions of interest (i.e. multi-scale bounding boxes) that exist inside each image or video frame, (b) Object representation uses these areas in order to describe the information that exist inside and machine learning is finally deployed to discriminate between classes.

As far as localization is concerned, earlier techniques followed the computationally expensive sliding window paradigm, but have been recently replaced by selective search [17] and techniques that deploy multi-scale bounding box proposals [5, 22] instead of exhaustive dense searching in the image scene. Similar results have accomplished the objectness measure [1] and its computationally efficient counterpart, named BING [2]. While, geodesic object proposal [8] achieved among the highest detection performance (i.e. recall) even when the requested number of candidate proposals was small.

Vehicle representation uses the localization outcome in order to describe the objects that exist inside these areas, such as cars, trucks and pedestrians. Until recently, this would require the extraction of local based histograms (i.e. HOG, LBP, HOGles etc.) that encode light intensity [16], texture [19] and the existence of specific shapes or other image features [18]. More recently, SoA techniques of this domain turned their attention into deep convolutional neural networks to represent vehicles inside images. They train the parameters of their models on large datasets, like COCO [10] or ImageNet 1000-class [14] and then retrain the weights and parameters of the model in vehicle-tailored datasets, such as UA-Detrac [20].

As far as deep convolutional techniques are concerned, we can encounter several works in the literature that deal with object detection. First and foremost Fast R-CNN [4] and Faster R-CNN [13], which leverage deep convolutional representation schemes to lead to robust and highly accurate object detection. An interesting modification of Fast R-CNN was proposed in [21] and applied a MultiPath network to predict segmentation masks in addition to bounding boxes. Position-sensitive score maps in a fully convolutional network [3], on the other hand, enabled the fully adoption of the ResNet architecture for the purposes of object detection. Current SoA techniques has currently turned its attention to developing faster, rather than more accurate techniques, such as YOLO [12], SSD [11] and the relief R-CNN [9] which generates proposals from convolutional features by simple rules.

3. Methodology

Traffic management in our methodology is accomplished by combining vehicle detection with multi-target tracking algorithm. We investigate a SoA object detection scheme, namely Faster R-CNN, and based on its limitations we propose a novel hybrid representation (i.e. DeepHOG) that leverages both shallow, mid-level and deep representation to overcome and introduce a better recognition performance. Bounding boxes with a detected vehicle or object are then used by our multi-target tracker so as to localize and monitor them throughout the whole duration that exist in the video scene. Fig. 1 shows the framework that is followed so as to tackle the detection and tracking issues.

3.1. Vehicle detection

Vehicle detection in this work is deployed by following two approaches: (a) CERTH-RCNN and (b) DeepHOG. An ensemble framework is also investigated as a complementary approach so as to study the impact that the fusion of the the two. Object localization in both cases is performed by using the Region Proposal Network (RPN) that CERTH-CNN uses.

CERTH-RCNN uses a modification of the original Faster R-CNN [13], the Faster-RCNN-Resnet101 architecture. We chose to implement this technique as it was found as one of the most accurate and computational efficient models amongst the current SoA [7]. This is attributed to the fact that it uses a single feed-forward convolutional network to localize object proposals and predict classes, without requiring a second stage per-proposal classification operation. We used the Faster-RCNN-Resnet101 model that is pretrained on the COCO dataset and tuned it on NVIDIA AI city dataset, so as to be able to detect vehicles in video frames.

For **DeepHOG** vehicle detection, CERTH deployed a novel hybrid representation scheme that combines shallow with deep features. More specifically, we used Histograms of Oriented Gradients(HOG) [16] as a local appearance features to represent pedestrians and vehicle objects and encode them into a Fisher vector. This resulted in a powerful mid-level representation vector that maintain the relation difference of each object to the most discriminant ones and provided to a Neural Network in order to train a highly accurate hybrid feature representation.

The ensemble model decides about the class for each box based on the most confident score given. This way we hope to expose and take advantage of a possible complementary nature of the two models.

3.2. Veicle tracking

CERTH-KCF is based on the tracking algorithm that was proposed in [6]. Vehicle detection is deployed every

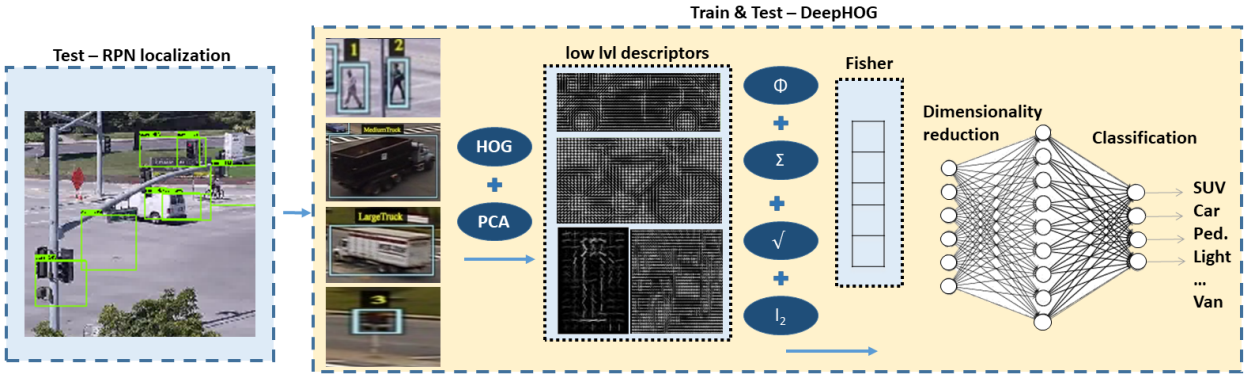

Figure 1: Block diagram of deepHOG, our novel representation scheme. Representation entails the extraction of HOG descriptors in groundtruth bounding boxes of vehicle objects and pedestrians, Fisher encodes the shallow descriptors and fed a Neural Network so as to encode the mid-level descriptor and train the detection model. When testing is concerned bounding boxes are given from a Region Proposal Network(RPN), represented by using DeepHOG representation and classified by the trained detection model.

$W_{detect} = 3$ video frames and in conjunction with KCF tracking algorithm are responsible to monitor the vehicles that exist in the video scene. For that purposes, a vehicle database is built so as to record the vehicle ids: veh_{id} by creating space for new detected vehicles, and retaining it until they disappear from the scene (i.e. when the algorithm could not detect the veh_{id} for more than $W_{track} = 3$ sequential video frames). Along with the veh_{id}, the appropriate vehicle class: veh_{label} and its detection score: veh_{score} are also retained in this structure. Overlapped bounding boxes are also tackled by CERTH-KCF by allowing the creation of a new veh_{id} only when the intersection over union score with the already existing bounding boxes is larger than 70%. To tackle occlusions between different veh_{id}, CERTH-KCF merge the two ids at the current frame, keep the oldest veh_{id} and throw the other.

4. Experimental work

Our detection algorithm was evaluated on NVIDIA AI city datasets $aic480$, $aic540$ and $aic1080$. The datasets acquired video frames from surveillance cameras that depict intersections in urban areas under diverse weather conditions, daytime and nighttime.

4.1. Parameter selection

As far as **CERTH-RCNN** is concerned, a convolutional feature extractor (Resnet101) is initially applied on the input image so as to obtain high level features, which are then given as input to the R-CNN to detect what exist inside it. The model is initially trained on the COCO dataset and then finetuned to the NVIDIA AI city $aic1080$ and $aic480$ dataset, while $aic540$ shared the same model with $aic1080$. Localization and classification losses were weighted equally, allowing a maximum of 500 detections and online training with a learning rate schedule initialized at $5e^{-5}$ and decreasing it progressively. We concluded that training at 40000 steps was the optimal solution to train our model. $Argmax$ was used for classification purposes and $SmoothL_1$ to compute location loss function.

For the training of the **DeepHOG's** neural network, we used RELU activation functions on a 3 Full Convolutional(FC) layer with a width of 512 for each one, using 0.8 dropout chance after every layer and performing Adam optimization with default parameters.

4.2. Results

Observing Tables **??** we can safely say that $DeepHOG$ improved against $CERTH - RCNN$ (RCNN in the table to save space), almost in all classes except from car, which in the case of $DeepHOG$ is usually confused with category SUV. The ensemble algorithm though fused the two outcomes efficiently and lead to even higher accuracy rates.

Table 4 shows the great difference between the mean Average Precision (mAP) scores that our algorithm (i.e. Team2) achieved comparing to the other teams of the challenge. It is clear that our best algorithm is far from the penultimate teams (i.e. Team23, Team14) in all 3 datasets and failed to detect vehicles in most cases. This needs further investigation since our localization and recognition approaches were similar to other teams and followed the same procedure as Deep Network is concerned (CERTH-RCNN).

5. Concluding

The CERTH team proposed a novel detection algorithm that in most cases performed better than SoA CERTH-RCNN and we believe that it can be considered a fair con-

aic480	Car	SUV	Van	Bus	Bicycle	Motorcycle
RCNN	0.54	0.12	0.01	0	0	0
DeepHOG	0.41	0.21	0.03	0.05	0	0
Ensemble	0.50	0.17	0.03	0.01	0	0
	Small Truck	**Medium Truck**	**Large Truck**	**Pedestrian**	**Localization**	**Overall**
RCNN	0.03	0.03	0.02	0	0.84	0.15
DeepHOG	0.10	0.10	0.02	0	0.64	0.14
Ensemble	0.08	0.07	0.02	0	0.74	0.15

Table 1: Test results on aic480 dataset

aic540	Car	SUV	Van	Bus	Bicycle	Motorcycle	Small Truck	Medium Truck
RCNN	0.49	0.09	0.01	0	0.02	0.01	0.02	0.02
DeepHOG	0.26	0.13	0.05	0	0.15	0.05	0.14	0.09
Ensemble	0.49	0.1	0.02	0	0.15	0.04	0.08	0.05
	Large Truck	**Pedestrian**	**Group Of People**	**Red Signal**	**Green Signal**	**Yellow Signal**	**Localization**	**Overall**
RCNN	0.01	0	0	0	0	0	0.73	0.09
DeepHOG	0.01	0	0.01	0	0	0	0.56	0.10
Ensemble	0.01	0	0.01	0	0	0	0.70	0.11

Table 2: Test results on aic540 dataset

Team	aic480	aic540	aic1080
Team24	0.518	0.430	0.481
Team21	0.444	0.384	0.470
Team5	0.447	0.345	0.386
Team6	0.407	0.345	0.369
Team25	-	-	0.310
Team10	0.373	0.279	0.294
Team4	0.343	0.252	0.280
Team19	0.353	-	0.271
Team23	0.326	0.220	0.256
Team14	-	-	0.249
Team2	0.146	0.110	0.124

Table 4: Overall Comparison

tribution to the literature if the results get improved. As a future work, CERTH plans to extend its traffic management algorithm so as to enable traffic density classification and crossroad traffic analysis. Spatio-temporal information that leverage motion and appearance features is going to be deployed for the implementation of the former, while the second will leverage tracking results and vast log files to analyze the traffic behavioural patterns throughout large duration timelines (i.e. daily, weekly and monthly intersection analysis).

6. Aknowledgement

This work was supported by beAWARE[1] project partially funded by the European Commission under grant agreement No 700475.

References

[1] B. Alexe, T. Deselaers, and V. Ferrari. Measuring the objectness of image windows. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2189–2202, 2012.

[2] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. Torr. Bing: Binarized normed gradients for objectness estimation at 300fps.

[1] http://beaware-project.eu/

aic1080	Car	SUV	Van	Bus	Bicycle	Motorcycle	Small Truck	Medium Truck
RCNN	0.48	0.07	0.01	0	0.03	0.01	0.02	0.01
DeepHOG	0.27	0.18	0.08	0.01	0.23	0.12	0.14	0.11
Ensemble	0.48	0.11	0.06	0	0.22	0.10	0.09	0.08

	Large Truck	Pedestrian	Group Of People	Red Signal	Green Signal	Yellow Signal	Localization	Overall
RCNN	0	0	0	0.01	0	0	0.58	0.08
DeepHOG	0.02	0.13	0.05	0.05	0.01	0	0.38	0.12
Ensemble	0.01	0.13	0.05	0.05	0.01	0	0.47	0.12

Table 3: Test results on aic1080 dataset

In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3286–3293, 2014.

[3] J. Dai, Y. Li, K. He, and J. Sun. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in neural information processing systems*, pages 379–387, 2016.

[4] R. Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[5] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[6] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista. High-speed tracking with kernelized correlation filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(3):583–596, 2015.

[7] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, and K. Murphy. Speed/accuracy trade-offs for modern convolutional object detectors. *CoRR*, abs/1611.10012, 2016.

[8] P. Krähenbühl and V. Koltun. Geodesic object proposals. In *European Conference on Computer Vision*, pages 725–739. Springer, 2014.

[9] G. Li, J. Liu, C. Jiang, L. Zhang, K. Tang, and Z. Zhu. Relief r-cnn: Utilizing convolutional feature interrelationship for fast object detection deployment. *arXiv preprint arXiv:1601.06719*, 2016.

[10] 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*, pages 740–755. Springer, 2014.

[11] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. SSD: Single shot multibox detector. In *ECCV*, 2016.

[12] 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*, pages 779–788, 2016.

[13] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[14] 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*, 115(3):211–252, 2015.

[15] H. G. Seif and X. Hu. Autonomous driving in the icityhd maps as a key challenge of the automotive industry. *Engineering*, 2(2):159–162, 2016.

[16] F. Suard, A. Rakotomamonjy, A. Bensrhair, and A. Broggi. Pedestrian detection using infrared images and histograms of oriented gradients. In *Intelligent Vehicles Symposium, 2006 IEEE*, pages 206–212. IEEE, 2006.

[17] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013.

[18] C. Vondrick, A. Khosla, T. Malisiewicz, and A. Torralba. Hoggles: Visualizing object detection features. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1–8, 2013.

[19] X. Wang, T. X. Han, and S. Yan. An hog-lbp human detector with partial occlusion handling. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 32–39. IEEE, 2009.

[20] L. Wen, D. Du, Z. Cai, Z. Lei, M. Chang, H. Qi, J. Lim, M. Yang, and S. Lyu. UA-DETRAC: A new benchmark and protocol for multi-object tracking. *CoRR*, abs/1511.04136, 2015.

[21] S. Zagoruyko, A. Lerer, T.-Y. Lin, P. O. Pinheiro, S. Gross, S. Chintala, and P. Dollár. A multipath network for object detection. *arXiv preprint arXiv:1604.02135*, 2016.

[22] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *European Conference on Computer Vision*, pages 391–405. Springer, 2014.