DOI: http://dx.doi.org/10.26483/ijarcs.v13i4.6882

Volume 13, No. 4, July-August 2022

ISSN No. 0976-5697

International Journal of Advanced Research in Computer Science

SURVEY REPORT

Available Online at www.ijarcs.info

REVIEW ON SCENE SEMANTICS EXTRACTION FOR DECISION MAKING SYSTEM IN AUTONOMOUS VEHICLES

Yuvraj Hembade
E&TC Engineering, ZCOER,
Pune, India

Dr. Suresh Shirbahadurkar
E&TC Engineering, ZCOER,
Pune, India

Dr. Arun Gaikwad
E&TC Engineering, ZCOER,
Pune, India

Abstract: It is a worldwide witnessed fact that traditional manual driving mechanism will be superseded by Autonomous Vehicles [AVs] in coming years. Autonomous vehicles are going to be most foreseen development in the automotive industry. That would require Decision Making System which will enable AVs to intuitively interpret the real-time situations around. Most importantly scene recognition on streets & extracting relevant semantics from the scene is challenging task. So, image classification & object detection techniques using Deep Convolutional Neural Networks [DCNN] are going to play vital role in every other methodology designed for scene semantics extraction. As per the extracted scene semantics DMS actuates the necessary devices which control the speed of vehicle & steering angel. So for that matter information extraction from road scene images covering all aspects to take intuitive decisions has huge concern with overall performance of the AV's.

Keywords: Autonomous Vehicles, Deep Convolutional Neural Networks, Scene Semantics, Object Detection.

I. INTRODUCTION

Recent trend in automotive industry is more aligned towards development of Autonomous Vehicles. Significant progress has been witnessed in obstacle detection & drivers state recognition but due to certain limitations scene centric recognition of on road situations is not yet effectively addressed. The main reasons behind this: 1) the lack of shared large datasets with comprehensively annotated road scene information, and 2) the difficulty to find effective ways to train networks associated with issues like image resolutions scene dynamics and scene capturing conditions, etc

Self-driving has received vast capital inflow and tremendous research interest in both academia and industry in recent years. Researchers from various globally esteemed institutes and top-tier mobile manufactures, join together to push the boundary of self-driving challenges. Self-driving technology can be divided into several parts, including localization and mapping [2], [3], [4], motion planning [5], [6], behavioral decision [7], and scene understanding [8], etc.

Generally, scene understanding involves a number of subtasks such as scene categorization, object detection/tracking, and semantic segmentation, etc. Above listed tasks describe a specific aspect of scene. It's not that easy to jointly model these varied aspects to understand the relations between elements of scene.

Innovative research and novel solutions are in great demand. Deep convolutional neural networks (DCNNs) have been extensively useful in a numerous computer-vision tasks such as image classification [9], [10], object detection [7], [11]. Some recently-presented architecture even allows for per-pixel predictions like semantic segmentation [12], [13] or scene-flow estimation [14], [15].

By deploying a deep data integration method, we re-balance class priors while emphasizing the cost for misclassifications based on a combination of two classic schemes – data resampling and cost-sensitive learning. Another problem is how to handle images in multiple resolutions. A typical approach is the multi-scale cropping, adopted by the VggNet [32] and then inherited by following works, such as ResNet [33] and Inception V3 [30]. Aside from that, in other works, the multi-scale representation is preferred. However, the cropped patch may lose the label information associated with other image regions and additional training work is required to learn networks for multiple resolutions or scales.

All manuscripts must be in English. These guidelines include complete descriptions of the fonts, spacing, and related information for producing your proceedings manuscripts.

This template provides authors with most of the formatting specifications needed for preparing electronic versions of their papers. Margins, column widths, line spacing, and type styles are built-in; examples of the type styles are provided throughout this document and are identified in italic type, within parentheses, following the example. PLEASE DO NOT RE-ADJUST THESE MARGINS. Some components, such as multi-leveled equations and graphics, are not prescribed, although the various table text styles are provided. The formatter will need to create these components, incorporating the applicable criteria that follow.

II. RELATED WORK

Here brief survey of contributions of present researches related to scene interpretation from two aspects: the dataset of self-driving scenes and the classification with biased data.

A. Dataset of self-driving scenes:

Large amounts of labelled dataset are often used as fuel to deep learning rocket, without which tremendous progress of vision based research can't be achieved. Below context emphasizes on study of different datasets proposed by several researchers.

1. J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Feifei, "Imagenet: A large-scale hierarchical image database": Taking object recognition for example, ILSVRC uses a subset of ImageNet [9] with roughly 1,000 images in each of 1,000 categories, and the COCO dataset [31] provides nearly 120,000 training samples with a total of 80 categories. Both of them provide a large number of training samples to guide the convolutional network to achieve high recognition rate.
2. J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo": Another examples for large training set-benefitted scene recognition can also be found, e.g., the SUN dataset [19] provides a wide coverage of scene categories containing 397 categories with more than 100 images per category.
3. B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition": The Places [10] contains 1.8 million images from 365 scene categories, with at most 5,000 images per category, etc. However, there is no such dataset in the autonomous-driving field. Some traffic-scene datasets mainly focus on environmental perception, with the self-driving scene recognition almost neglected. For example, KITTI [20] comprises a wide range of challenges like stereo vision, odometry, object detection and tracking, etc. CompCar dataset [21] specifically focuses on fine-grained car classification/verification and attribute prediction.
4. M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding": CityScapes [8], provides a large-scale dataset derived from stereo sequences, aiming at both pixel- and instance level semantic labelling. The self-driving scene recognition is simple for human brains but extremely difficult for computers to address.
5. L. Chen, W. Zhan, W. Tian, Y. He and Q. Zou, "Deep Integration: A Multi-Label Architecture for Road Scene Recognition": To attract and motivate more research on self-driving scene recognition Long Chen et al. have introduced a new large scale dataset of over 110k scene images cutting across 52 categories, which is currently, to my best knowledge, the largest dataset in terms of scene recognition in self-driving domain. FM2 dataset [22] is also quite relevant to DrivingScene dataset, yet it contains a total number of 6,237 images from eight scene classes.

B. Classification with biased data:

In scene recognition, a single image usually associates with multiple scene labels. Thus, most prior works train a deep neural network to assign the multi-class label to the query image [23]–[26]. Although some deep structures such as VggNet [32], Inception V3 [33] and ResNet [30] have demonstrated higher performance with deeper layers in classification, the training still suffers from the negative impact of data imbalance. Attempting to solve this problem, two approaches are well studied in past years.

1. M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks": The first approach is re-sampling, adopted by Oquab et al. [27] to rebalance class priors during training through under- and oversampling.
2. C. Huang, Y. Li, C. C. Loy, and X. Tang, "Learning deep representation for imbalanced classification": The second is the cost-sensitive learning, utilized by Huang et al. [34] along with the Triple-Header Hinge Loss to assign different costs for misclassification on the majority and minority classes.

C. Research Gap

Despite a good performance, both methods are proposed mainly for single-label classification. Moreover, the over-and under-sampling may introduce undesired noise or remove valuable sample information while the cost-sensitive learning usually requires utilization of additional features. The above works can be seen as natural extensions to the existing imbalanced learning techniques, while neglecting the underlying data structure for discriminating imbalanced data. To explore a more effective way to deal with data imbalance in the context of deep representation, we incorporate both single and multi-label training by a multi-level loss function, making our architecture more flexible and compact. By deploying a deep data integration method, we re-balance class priors while emphasizing the cost for misclassifications based on a combination of two classic schemes – data resampling and cost-sensitive learning. Another problem is how to handle images in multiple resolutions. A typical approach is the multi-scale cropping, adopted by the VggNet [32] and then inherited by following works, such as ResNet [33] and Inception V3 [30]. Aside from that, in other works, the multi-scale representation is preferred. For instance, Oquab *et al.* [27] propose a new re-sampling method about multi-scale part proposals for fine-grained categorization. Wang *et al.* [28] combine multi-resolution architecture with a confusion matrix for scene classification. However, the cropped patch may lose the label information associated with other image regions and additional training work is required to learn networks for multiple resolutions or scales.

III. PROPOSED METHODOLOGY

Publicly available image datasets are task dependent. And it is difficult to fairly compare them in terms of whether a particular dataset is better or worse, although the image number and category coverage range are often regarded as two import metrics. Here we argue that dataset diversity and density are two indicators for large-scale

image dataset evaluation (i.e., the Places dataset [10]), especially for deep feature learning tasks.

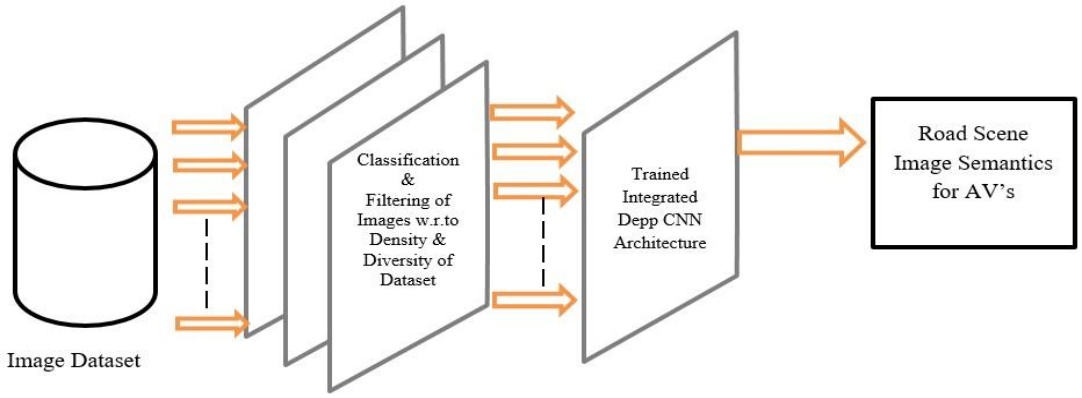

Fig. 1. Extraction of Image Semantics

Density is equivalent to data concentration, measuring the similarity level of an image with its neighbors. A dataset with higher density often guarantees to learn powerful representative features through deep convolutional neural networks (CNNs). The high density alone can deteriorate the dataset quality. An extreme situation is that all the images regarding a scene category are taken within the same viewpoint or with less, camera pose variability. This high overlap of image appearance leads to large dataset redundancy, inevitably jeopardizing the algorithm's performance. Thus, a good dataset should have strong generalization capability and involve as many diverse images as possible regarding a scene category. We maximize Driving Scene dataset diversity from three aspects. First, we insist to capture images of the same scene category at different locations. Second, images of the same scene category are captured under at different environmental conditions like sunlight, night view, evening, or shadow.

Further Images in data set needed to be classified and filtered as per the density & diversity factor in the dataset. In order to get equally sized & properly labelled images for training convolutional neural networks some cropping & resizing of the images is supposed to be done in the same filtration layer. So finely Deep CNN could get images with the desired attributes.

IV. EXPECTED RESULTS

Till date, there are few methods proposed for effectively training network among imbalanced categories and most of them still transfer the multi label task in to training a single-label model. As no unified measurement is available to evaluate the quality of classifiers trained by imbalanced dataset, we will follow the protocol of mean average precision (mAP) [41].

It compares two baseline approaches baseline1 [27] & baseline2 [28] with ground reality & proposed model, where road scene recognition with respect to traffic scenario, road type, area, weather condition, time, road structure etc. is represented. Here we would like to explore the leveraged image information by the network through visualization of utilized deep features. The technique of convolution visualization will be progressively developed in recent years and related researches can be roughly divided into two categories: the dataset-centric and the network-centric approach. The former one requires to train a DNN and afterwards to feed the data into the network; the latter one, however, only requires the trained network itself. Although the latter procedure is a relative simple, the former is generally accepted in most works because it has a clearer visual effect. In this experiment, we will utilize the test set of DrvingScene (i.e, 33k images) as the input for the network.

TABLE I: mAP value over all classes for different architectures & proposed Deep CNN

Sr. No.	*Model*	*mAP(%)*
1	Google Net	76.5
2	ResNet-50	76.1
3	VGG Net	74.9
4	Google Net + Data Integration	81.3
5	VGG Net + Data Integration	81.0
6	ResNet-50 + Data Integration	81.1
7	Proposed Deep CNN	76 - 82

V. CONCLUSION

In studied literature it is found that, variety of datasets plays an important role in getting good accuracy of the models to be trained for getting classified set of images with several classes which are taken in to consideration. Even in case of object detection & tracking of road scene images has huge dependence over accurate semantics extraction from road scene images. So before training models for image semantics extraction varied set of road scene images needs to gather & validated as per requirements.

VI. REFERENCES

[1] L. Chen, W. Zhan, W. Tian, Y. He and Q. Zou, "Deep Integration: A Multi-Label Architecture for Road Scene Recognition," in IEEE Transactions on Image Processing, vol.

28, no. 10, pp. 4883-4898, Oct. 2019. doi: 10.1109/TIP.2019.2913079

[2] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, "Mono SLAM: Real-time single camera SLAM," IEEE Transactions on Pattern Analysis & Machine Intelligence, no. 6, pp. 1052–1067, 2007.

[3] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. J. Leonard, "Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age," IEEE Transactions on Robotics, vol. 32, no. 6, pp. 1309–1332, 2016.

[4] Q. Li, L. Chen, M. Li, S. Shaw, and A. Nuchter, "A sensor-fusion drivable-region and lane-detection system for autonomous vehicle navigation in challenging road scenarios," IEEE Transactions on Vehicular Technology, vol. 63, no. 2, pp. 540–555, 2014.

[5] D. Gonzlez, J. Prez, V. Milans, and F. Nashashibi, "A review of motion planning techniques for automated vehicles," IEEE Transactions on Intelligent Transportation Systems, vol. 17, no. 4, pp. 1135–1145, 2016.

[6] L. Chen, L. Fan, G. Xie, K. Huang, and A. Nuchter, "Moving-object detection from consecutive stereo pairs using slanted plane smoothing," IEEE Transactions on Intelligent Transportation Systems,vol.18,no.11, pp. 3093–3102, 2017.

[7] L. Chen, X. Hu, T. Xu, H. Kuang, and Q. Li, "Turn signal detection during night time by cnn detector and perceptual hashing tracking," IEEE Transactions on Intelligent Transportation Systems, vol. 18, no. 12, pp. 3303–3314, 2017.

[8] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," IEEE Conference on Computer Vision and Pattern Recognition, pp. 3213–3223, 2016.

[9] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Feifei, "Imagenet: A large-scale hierarchical image database," European Conference on Computer Vision, pp. 248–255, 2009.

[10] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 6, pp. 1452– 1464, 2018.

[11] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster r-cnn: Towards real time object detection with region proposal networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137– 1149, 2017.

[12] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," computer vision and pattern recognition, pp. 3431–3440, 2015.

[13] Q. Zou, Z. Zhang, Q. Li, X. Qi, Q. Wang, and S. Wang, "Deepcrack: Learning hierarchical convolutional features for crack detection," IEEE Transactions on Image Processing, vol. 28, no. 3, pp. 1498–1512, 2019.

[14] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," computer vision and pattern recognition, pp. 4040–4048, 2016.

[15] L. Chen, M. Cui, F. Zhang, B. Hu, and K. Huang, "High speed scene flow on embedded commercial-off-the-shelf systems," IEEE Transactions on Industrial Informatics, pp. 1–1, 2018.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in neural information processing systems, 2012, pp. 1097–1105.

[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," International Conference on Learning Representations, 2015.

[18] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9, 2015.

[19] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," IEEE Conference on Computer Vision and Pattern Recognition, pp. 3485– 3492, 2010.

[20] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," IEEE Conference on Computer Vision and Pattern Recognition, pp. 3354–3361, 2012.

[21] L. Yang, P. Luo, C. C. Loy, and X. Tang, "A large-scale car dataset for fine-grained categorization and verification," IEEE Conference on Computer Vision and Pattern Recognition, pp. 3973–3981, 2015.

[22] I. Sikiri´c, K. Brki´c, J. Krapac, and S. ˇSegvi´c, "Image representations on a budget: Traffic scene classification in a restricted bandwidth scenario," IEEE Intelligent Vehicles Symposium, 2014.

[23] Y. Luo, T. Liu, D. Tao, and C. Xu, "Multi view matrix completion for multi label image classification," IEEE Transactions on Image Processing, vol. 24, no. 8, pp. 2355–2368, 2015.

[24] X. Li, X. Zhao, Z. Zhang, F. Wu, Y. Zhuang, J. Wang, and X. Li, "Joint multi label classification with community-aware label graph learning," IEEE Transactions on Image Processing, vol. 25, no. 1, pp. 484–493, 2016.

[25] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu, "Cnnrnn: A unified framework for multi-label image classification," IEEE Conference on Computer Vision and Pattern Recognition, pp. 2285– 2294, 2016.

[26] H. Lai, P. Yan, X. Shu, Y. Wei, and S. Yan, "Instance-aware hashing for multi-label image retrieval," IEEE Transactions on Image Processing, vol. 25, no. 6, pp. 2469–2479, 2016.

[27] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," IEEE conference on computer vision and pattern recognition, pp. 1717– 1724, 2014.

[28] L. Wang, S. Guo, W. Huang, Y. Xiong, and Y. Qiao, "Knowledge guided disambiguation for large-scale scene classification with multi-resolution cnns," IEEE Transactions on Image Processing, 2017.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 9, pp. 1904– 1916, 2015.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778, 2016.

[31] T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick, "Microsoft coco: Common objects in context," IEEE Conference on European Conference on Computer Vision, pp. 740–755, 2014.

[32] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," International Conference on Learning Representations, 2015.

[33] C. Szegedy,V. Vanhoucke, S. Ioffe,J. Shlens, and Z.Wojna, "Rethinking the inception architecture for computer vision," IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826, 2016.

[34] C. Huang, Y. Li, C. C. Loy, and X. Tang, "Learning deep representation for imbalanced classification," IEEE conference on computer vision and pattern recognition, pp. 5375–5384, 2016.

[35] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi. (2016). "Inception-v4, inception-ResNet and the impact of residual connections on learning." [Online]. Available: https://arxiv.org/abs/1602.07261

[36] A. Oliva and A. Torralba, “Modeling the shape of the scene: A holistic representation of the spatial envelope,” Int. J. Comput. Vis., vol. 42, no. 3, pp. 145–175, 2001.

[37] A. Veit, N. Alldrin, G. Chechik, I. Krasin, A. Gupta, and S. Belongie. (2017). “Learning from noisy large-scale datasets with minimal supervision.” [Online]. Available: https://arxiv.org/abs/1701.01619

[38] L. Li, K. Ota and M. Dong, "Humanlike Driving: Empirical Decision-Making System for Autonomous Vehicles," in IEEE Transactions on Vehicular Technology, vol. 67, no. 8, pp. 6814-6823, Aug. 2018, doi: 10.1109/TVT.2018.2822762.

[39] Yuan, S.; Chen, Y.; Huo, H.; Zhu, L. Analysis and Synthesis of Traffic Scenes from Road Image Sequences. Sensors 2020, 20, 6939. https://doi.org/10.3390/s20236939

[40] https://idd.insaan.iiit.ac.in/dataset/download/

[41] W. Zhiqiang and L. Jun, "A review of object detection based on convolutional neural network," 2017 36th Chinese Control Conference (CCC), 2017, pp. 11104-11109, doi: 10.23919/ChiCC.2017.8029130.

[42] https://cloud.google.com/tpu/docs/inception-v3-advanced

