International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 22 (2018) pp. 16060-16063

Semantic Segmentation for Applications in Autonomous Vehicles

Luis Alfredo Rodríguez Umaña
Professor, Faculty of Basic Sciences and Engineering. University of the Llanos, Villavicencio, Colombia.

Javier Eduardo Martínez Baquero
Professor, Faculty of Basic Sciences and Engineering. University of the Llanos, Villavicencio, Colombia.

Robinson Jiménez Moreno
Professor, Department of Mechatronics Engineering, Nueva Granada Military University, Bogotá, Colombia.

Abstract
This paper presents a navigable area identifier for autonomous terrestrial vehicles, based on artificial intelligence, implementing Semantic Segmentation techniques. In the first place, an exploratory study is carried out in the field of artificial intelligence, which through the training of a network with SegNet architecture and the implementation of 3 datasets (2 of training and 1 of validation), and the implementation of 3 datasets, composed of images in RGB format and resolution of 320x180 pixels, where the training datasets have been manually segmented for 6 categories which are: "Car", "Bike", "Person", "Ts", "Road" and "Background", it was possible to be used as a means for autonomous driving of an unmanned vehicle, making use of a Dash-cam or web-cam as an input source, in an average time of 0.25s together with a global accuracy of 76.69%.
Keywords: SegNet, Mask, Deep Learning, Road detection, Semantic Segmentation.

INTRODUCTION

The different techniques of artificial intelligence [1], since its inception, have evolved to the degree of generating a revolution in machine learning systems [2]. These systems are integrated with sensors such as cameras, in order to develop what is known today as artificial vision systems, which presents multiple applications to industry [3]. Within these systems, techniques such as semantic segmentation [4] are presented for the identification of objects in images. This technique has several applications in the recognition of urban scenes as streets [5] [6], which can be applied to safe and autonomous driving environments.

One of the most recent techniques of artificial intelligence for machine learning, focuses on deep learning [7], which already presents developments in applications of semantic segmentation [8]. Within these developments is the SegNet network [9], which is used in this work for applications oriented to autonomous driving, subject of strong interest within the research area [10].

It is proposed the implementation of a navigable area identifier through semantic segmentation for vehicles, pedestrians, traffic signals, general scene (background) and available lane, recognition aspects necessary for the displacement of land-based vehicles [10].

The article is structured in 3 sections, the first one corresponds to the methods and materials, where the database, the network architecture and its training options are presented. The second section corresponds to the application and results obtained regarding training, detection of semantic segmentation and evaluation of it. Finally, the conclusions obtained are presented.

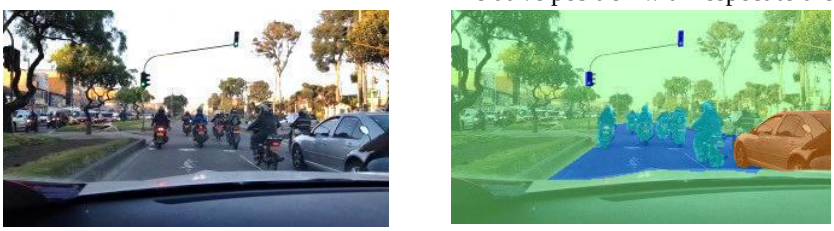
METHODS AND MATERIALS

The base of the application focuses on the training of the network, for which, this section is divided into 4 subsections that allow to guide such training, which are: *Datasets*, *Architecture*, and *Application*, as illustrated in Fig. 1.

Fig. 1. Development of the general network structure

A. Dataset

For the identification of the navigable area of the system, 3 training datasets were used (see Table 1), which are distributed 2 for training and one for validation, and which in turn were segmented manually, as shown in Fig. 4. These datasets were created by combining video material from an own Dash-cam and Google Street View and YouTube platforms, where the relative location of the Dash-cam inside the vehicle is seen in Fig. 2, while Fig. 3 shows the view of the camera (2.a) and the relative position with respect to the view (2.b) of the Dash-cam.



a) Original frame. **b)** Segmented frame.

Fig. 4. Manual segmentation

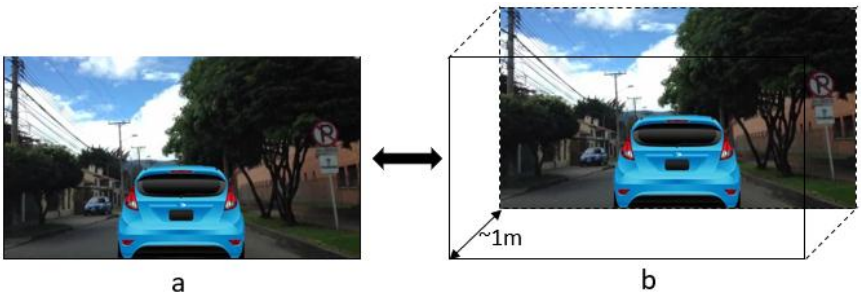


Fig. 3. (a) Camera/frame view. **(b)** Relative Camera/frame position.

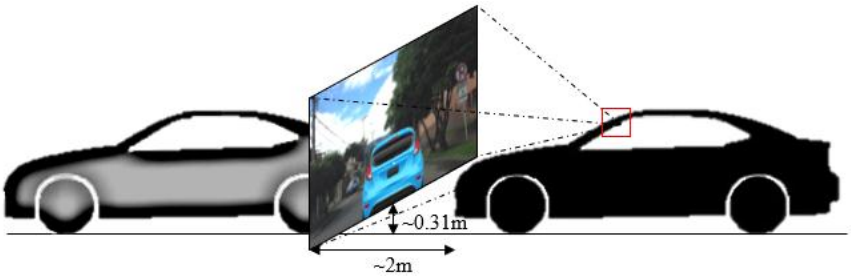


Fig. 2. Relative location of Dash-cam inside vehicle.

Table 1. Dataset Parameters

Dataset	CAT	TI	CHL	RES	MS	SC
Train 1	6	300	RGB	320x180	Yes	Y+G
Train 2	6	500	RGB	320x180	Yes	Y+G+DC
Validation	6	50	RGB	320x180	No	Y+G+DC

As shown in Table 1, *CAT* corresponds to the total categories of the dataset, *TI* refers to the total images of the dataset, *CHL* is the number of channels of the images and *RES* is its resolution, being 320x180 pixels with the aim of guaranteeing a resolution compatible with common resolutions of the current Dash-cam, *MS* refers to manual segmentation and finally, *SC* is used to define the origin of the images: Y (YouTube), G (Google Street View) and DC (Dash-Cam own).

B. Architecture

A SegNet network architecture [11] was implemented with an Encoding depth parameter of 4, which was selected after an iterative process, until a network performance of more than 85% accuracy was obtained. Within the different deep learning algorithms, it was decided to implement this network taking into account the high performance it has shown when performing segmentation tasks related to this same topic [12] [13].

Taking into account the characteristics of the training datasets, a computer with NVIDIA GeForce 940MX graphics card is used, selecting a batch size of 2, with 50 epochs of training and a Learning Rate of 0.01 (see Table 2), where, in addition, a drop factor of the Learn Rate is used which is defined in this case in 0.1, allowing to have a better learning in the training.

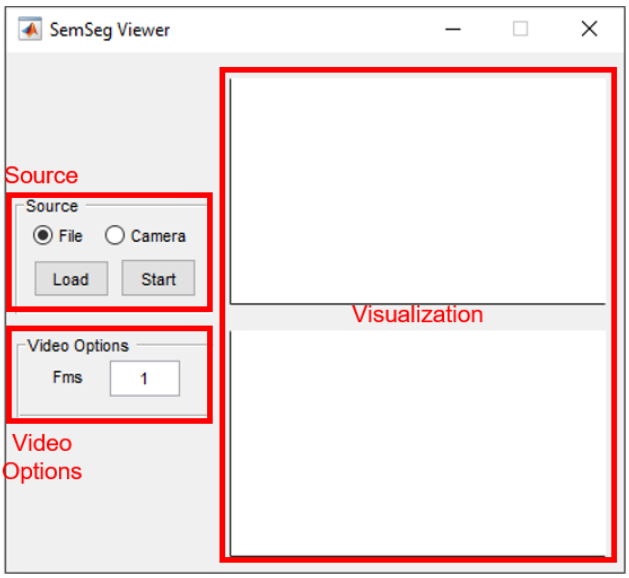


Fig. 5. Matlab Graphic User Interface.

Table 2. Training options used for SegNet

Training Options	
Batch Size	2
Epochs	50
Initial Learn Rate	0.01

C. Application

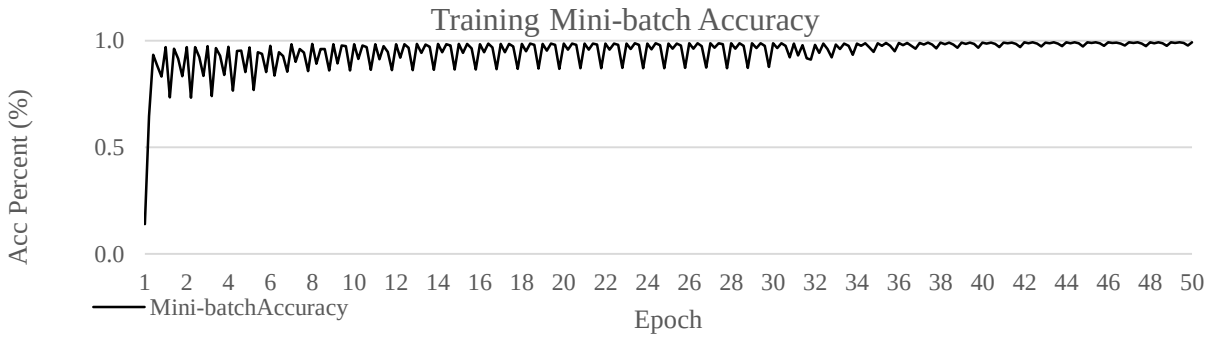


Fig. 6. SegNet training accuracy.

A graphic user interface was developed in Matlab®, as shown in Fig. 5, through which the automatic segmentation of the current scene is generated, referring to the 6 categories for which the SegNet was trained. This application consists of 3 main sections, which are: "Source", where it is selected if the webcam will be used directly as an image source or if a video will be loaded in .mp4 format for revision, "Video Options", where the parameter "Fms" refers to the number of frames that the video will advance during its revision and finally, "Visualization", an area that is responsible for showing both the original image and the image with the superimposed layer.

RESULTS AND DISCUSSIONS

The training with the final network lasted approximately 2.25 hours, presenting a percentage of accuracy in training of 99.3% (see Fig. 6) with total losses of 0.0208 in the last iteration of epoch 50.

Table 3. Training results for 2, 3 and 4 depth Encoding

Depth Encoding	TRN Acc	TRN Losses
2	87.91%	0.3295
3	99.28%	0.0318
4	**99.30%**	**0.0208**

As shown in Table 3, another 2 trainings are presented, only modifying the depth of the encoder using, in this case, encoding 2 and 3. Finding this way, improvement of the network to the amount of present parameters, allowing then to extract more features, without getting to incur an over training of this.

When performing an analysis on the training datasets "Train 1 and Train 2", based on the frequency of the pixels present by category, there was an absence of information related to the categories "Bike", "Person" and "Ts", as shown in Fig. 7.

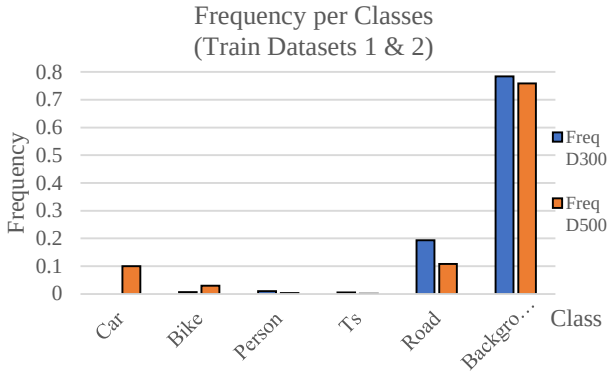


Fig. 7. Pixels frequency for 6 classes in Train Datasets.

Similarly, it can be seen the wide difference between the amount of information in the category "Background" with respect to the other categories. Therefore, a general evaluation was made on the validation dataset (see Table 4), finding a global accuracy of 76.69% and the accuracy per class, as indicated in table 5. Which shows the loss of these categories, which is due to the small number of pixels they occupy in the image.

Table 4. Dataset Metrics

Metric	Value
Global Accuracy	0.7669
Mean Accuracy	0.2091
Mean IoU	0.1693
Weighted IoU	0.5951
Mean BF Score	0.2598

Table 5. Class Metrics

Category	Accuracy	IoU	Mean BF Score
Car	0.0379	0.0362	0.0333
Bike	0.0526	0.0505	0.0490
Ts	0	0	-
Person	0	0	-
Road	0.17236	0.1681	0.0989
Background	0.99162	0.7613	0.6437

Taking into account the objective application, a review is made in the minimum and maximum detection process time, using

the same training hardware, finding an average time of 0.25s in this task (see Table 6). Additionally, in Fig. 8 to 10, the performance of the network is evident, determining in blue the possible navigable area of the vehicle.

Table 6. Timing detection required

Hardware	Min time (s)	Max time (s)
NVIDIA - PC	0.18	0.32

CONCLUSIONS

It is evident that this type of networks (encoder-decoder) are a good alternative solution for semantic segmentation tasks by Deep Leaning methods, since taking into account the characteristics of the dataset used for training, in addition to the short number of epochs with the training hyperparameters used, they unexpectedly offer opportune results.

After reviewing the training datasets together with the results obtained, such as the global accuracy of 76.69%, the great problem represented by the construction of data at the time of the training of models for intelligent semantic segmentation is evident, where the relationship between categories presents a wide gap in terms of quantity of information, negatively affecting the result of the models by allowing the possibility of overfitting occurring for some categories in particular, as well as the possibility of generating under-fitting for the categories with the least amount of information available.

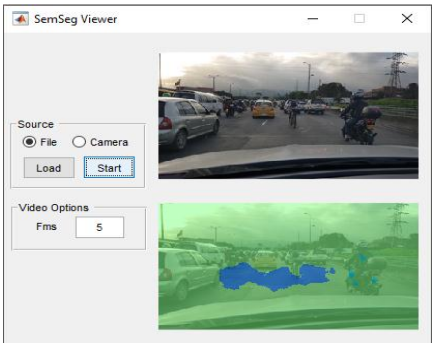


Fig. 8. Graphic User Interface – 5 Frames.

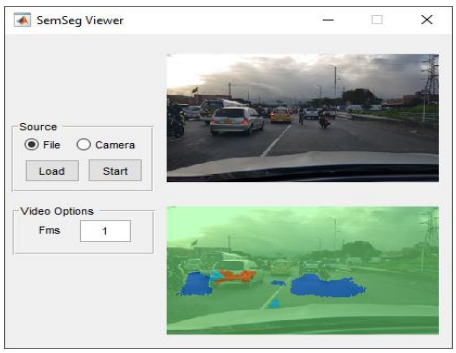


Fig. 9. Graphic User Interface – 1 Frame

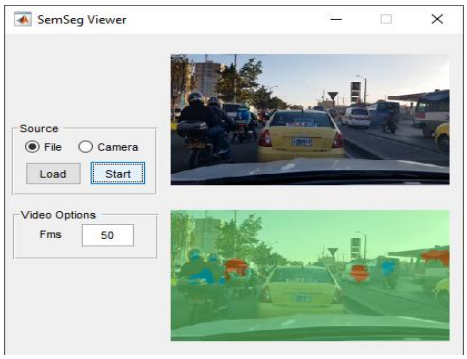


Fig. 10. Graphic User Interface – 50 Frames.

When carrying out the validation by means of the application developed, it is evident that, although the detection times are small, with an average of 0.25s per scene, it is still a considerable time of execution, so it is recommended in the first instance, to improve the processing hardware in order to reduce execution times.

ACKNOWLEDGMENT

The research for this paper was supported by Davinci research Group of Nueva Granada Military University.

REFERENCES

[1] D. Marr, “Artificial Intelligence: A Personal View”, 1977, Vol. 9, no 1, p. 37-48.

[2] R.S. Michalski, J.G. Carbonell and T.M. Mitchell, “Machine learning: An artificial intelligence approach”. Springer Science & Business Media, 2013.

[3] P. Constante, A. Gordon, O. Chang, E. Pruna, F. Acuna and I. Escobar, “Artificial Vision Techniques to Optimize Strawberry's Industrial Classification”, IEEE Latin America Transactions, 2016, vol 14, no 6, p. 2576-2581. doi: 10.1109/TLA.2016.7555221.

[4] R.M. Haralick and L.G. Shapiro, “Image Segmentation Techniques”, In Applications of Artificial Intelligence II, International Society for Optics and Photonics, 1985, vol. 548, pp. 2-10. doi: 10.1117/12.948400.

[5] J. Xiao and L. Quan, "Multiple view semantic segmentation for street view images," 2009 IEEE 12th International Conference on Computer Vision, Kyoto, 2009, pp. 686-693. doi: 10.1109/ICCV.2009.5459249

[6] C. Zhang, L. Wang and R. Yang, “Semantic segmentation of urban scenes using dense depth maps”, InEuropean Conference on Computer Vision, 2010, pp. 708-721, Springer, Berlin, Heidelberg.

[7] Y. LeCun, Y. Bengio and G. Hinton, “Deep learning”, In Nature, 2015, vol. 521, no 7553, p. 436.

[8] A. Paszke, A. Chaurasia, S. Kim and E. Culurciello, “Enet: A deep neural network architecture for real-time semantic segmentation”, 2016, arXiv preprint arXiv:1606.02147.

[9] V. Badrinarayanan, A. Handa, and R. Cipolla, “SegNet: A Deep Convolutional Encoder-Decoder Architecture for Robust Semantic Pixel-Wise Labelling”, 2015, arXiv preprint arXiv:1511.00561.

[10] T. Lozano-Pérez, “Autonomous robot vehicles”. Springer Science & Business Media, 2012.

[11] Alberto Garcia-Garcia, Sergio Orts-Escolano, Sergiu Oprea, Victor Villena-Martinez, Pablo Martinez-Gonzalez, Jose Garcia-Rodriguez, A survey on deep learning techniques for image and video semantic segmentation, Applied Soft Computing, Volume 70, 2018, Pages 41-65, ISSN 1568-4946, https://doi.org/10.1016/j.asoc.2018.05.018.

[12] Yongcheng Liu, Bin Fan, Lingfeng Wang, Jun Bai, Shiming Xiang, Chunhong Pan, Semantic labeling in very high resolution images via a self-cascaded convolutional neural network, ISPRS Journal of Photogrammetry and Remote Sensing, Volume 145, Part A, 2018, Pages 78-95, ISSN 0924-2716, https://doi.org/10.1016/j.isprsjprs.2017.12.007.

[13] Nicolas Audebert, Bertrand Le Saux, Sébastien Lefèvre, Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks, ISPRS Journal of Photogrammetry and Remote Sensing, Volume 140, 2018, Pages 20-32.