Dense Semantic Stereo Labelling Architecture for In-Campus Navigation

Jorge Beltrán, Carlos Jaraquemada, Basam Musleh, Arturo de la Escalera and Jose María Armingol
Intelligent Systems Lab (LSI) Research Group, Universidad Carlos III de Madrid (UC3M), Leganés, Madrid, Spain
jbeltran@ing.uc3m.es, carlosborja.jaraquemada@uc3m.es, {bmusleh, escalera, armingol}@ing.uc3m.es

Keywords: Dense Labelling, Semantic Labelling, Stereo Vision, Off-road Navigation, ROS.

Abstract: Interest on autonomous vehicles has rapidly increased in the last few years, due to recent advances in the field and the appearance of semi-autonomous solutions in the market. In order to reach fully autonomous navigation, a precise understanding of the vehicle surroundings is required. This paper presents a novel ROS-based architecture for stereo-vision-based semantic scene labelling. The objective is to provide the necessary information to a path planner in order to perform autonomous navigation around the university campus. The output of the algorithm contains the classification of the obstacles in the scene into four different categories: traversable areas, garden, static obstacles, and pedestrians. Validation of the labelling method is accomplished by means of a hand-labelled ground truth, generated from a stereo sequence captured in the university campus. The experimental results show the high performance of the proposed approach.

1 INTRODUCTION

The interest on autonomous vehicles has undergone a significant growth in the last 10 years due to its rapid development and the arrival of the first semi-autonomous commercial solutions. As a consequence, both companies and the university community are putting much effort into doing research in this field driven by its potential advantages in a wide number of areas like traffic management, road safety or disabled-passengers mobility.

However, this upcoming horizon may require dealing with complex tasks such as localization, navigation or inter-vehicle cooperation, depending on the vehicle desired level of automation.

A standardized scale for automation degrees is the SAE International's table (SAE On-Road Automated Vehicle Standards Committee, 2014). It describes the different automation levels which range from 0 (no automation) to 5 (full automation). A human driver is considered to be in charge of monitoring the environment in the first three levels. The big turning point comes at level 3 (Conditional Automation), where the system progressively replaces the driver intervention as the automation level increases. Therefore, one of the main challenges for automation involves acquiring a detailed knowledge of the vehicle surroundings.

There are several methods for environment information retrieval based on the kind of used sensor. On the one hand, those based on laser (Urmson et al., 2008) (Broggi et al., 2008) obtain very high precision data although they only provide distance information. Therefore, these approaches are usually suitable for detection tasks and map generation. On the contrary, they don't provide enough information from the environment to classify the different scene elements.

Figure 1: Autonomous vehicle iCab.

On the other hand, computer vision based systems obtain rich information of the vehicle surroundings at the expense of less precise distance measurements. Concretely, stereo vision systems allow depth estimation for all pixels in the image by computing the disparity map. After this process, the *uv-disparity* maps (Labayrade and Aubert, 2003) (Hu et al., 2005) can be obtained so that it can be used to detect both

BeltrÃąn J., Jaraquemada C., Musleh B., De La Escalera A. and Armingol J.
Dense Semantic Stereo Labelling Architecture for In-Campus Navigation.
DOI: 10.5220/0006131602660273
In *Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2017)*, pages 266-273
ISBN: 978-989-758-226-4

obstacles and free space within the scene (Bernini et al., 2014), thus obtaining the so-called obstacles map and free map (Soquet et al., 2007) (Guo et al., 2009) (Musleh et al., 2011). In addition, there are also many related works providing a more advanced scene labelling, both for urban (Sengupta et al., 2012) (Sengupta et al., 2013) (Long et al., 2015) and indoor environments (Golodetz et al., 2015).

Apart from stereo rigs, monocular lenses are also commonly used for scene understanding. Despite the handicap of not having precise depth information out-of-the-box, they are very suitable for classification tasks as they provide rich information at low cost.

In contrast with stereo approaches, monocular algorithms for scene classification and labelling does not rely on previously segmented Regions of Interest (ROIs) (Yao et al., 2012) (Mottaghi et al., 2014) (Ren et al., 2015) (Arnab et al., 2016).

However, two main downsides are present in most of the methods mentioned above: the high hardware specifications requirements to guarantee real-time execution, as in deep learning approaches, as well as the need of massive datasets for training classifiers. Regarding the availability issue of datasets, many research laboratories are publicly releasing their own, so there already exist some large annotated datasets for scene labelling (Cordts et al., 2016).

Nevertheless, both the need of task-specific datasets and the time required in annotation process represent a bottleneck for widening the application scope of this technology. As a result, recent work (Richter et al., 2016) is taking advantage of videogames calls to GPU interface to fetch labels for pixels of the different objects in the scene, so that labelling stage can be partially automated in order to easily build large datasets.

The main contribution of this work is a ROS-based architecture for dense image labelling able to obtain rich understanding of vehicle surroundings for autonomous navigation tasks. The presented approach takes advantage of stereo information for scene segmentation. The organization of the algorithm into loosely coupled stages provides the proposed architecture with the capability of being extended with ease so that other classifiers can be easily integrated.

The rest of the paper is organized as follows. Next section, focuses on the architecture description. Section 3 gives a description of the proposed algorithm. Afterwards, Section 4 presents a novel database for algorithm validation and a detailed description of the experimental results. Finally in Section 5, conclusions and future work are presented.

2 ARCHITECTURE DESIGN

The proposed architecture has been designed to run in our research platform called iCab (Hussein et al., 2016), a vehicle for autonomous in-campus navigation (see Fig. 1). Therefore, it has been fully integrated with the iCab framework previously developed by the authors (Marın-Plaza et al., 2016). As a result, the architecture is built on top of ROS (Robot Operating System) (Garage, 2010).

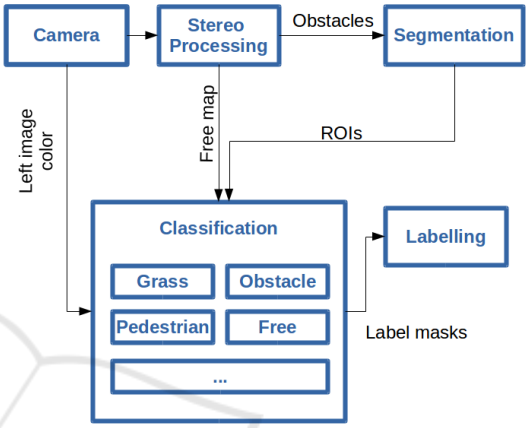


Figure 2: Architecture scheme.

The architecture scheme presented in this work is shown in Fig 2. As can be observed, it is composed by 5 different nodes corresponding to each of the stages of the algorithm. First of all, in the acquisition stage (*camera*), images are captured from the sensors both in colour and gray-scale. Afterwards, the *stereo processing* node receives the synchronized stereo images and builds the disparity map. Therefore depth information can be retrieved. Then, *uv-disparity* maps are computed so that they can be used for calculating the obstacle and free map. Later, *segmentation* stage separates the different obstacles present in its input map into ROIs based on their disparity information. In *classification* phase, different classifiers make use of the previously generated images and ROIs in order to determine the class which each pixel belongs to, thus producing per-class masks. Finally, at the *labelling* stage the produced masks are fused in order to obtain the final label of the pixels.

3 ALGORITHM

The algorithm for dense semantic stereo labelling introduced in this paper is based on different methods of computer vision. A large proportion of the algorithm uses the stereo information in order to detect the ob-

stacles and the free space in front of the vehicle (Section 3.1), whereas the visible information is used to classify obstacles as pedestrians and non-traversable areas, such as gardens (Section 3.3). The different stages of the proposed algorithm will be explained in this section.

3.1 Obstacle and Free Space Estimation

As commented above, the stereo images supplied by the vision system can be used in order to obtain 3D information of the vehicle's environment. This 3D information is usually represented by the disparity map (see Fig. 3a), where the value of the each pixel is proportional to its depth. An useful method to depict the stereo information of the vehicle's environment is the *uv-disparity* which is obtained from the disparity map (Hu and Uchimura, 2005). The *uv-disparity* contains information of the location both of the obstacles and ground ahead the vehicle; being able to distinguish between them. A previous work (Musleh et al., 2011) is then used to separate the disparity map into to different disparity maps: the obstacle map (see Fig. 3c), which contains the pixels belonging to the obstacles, and the free map (see Fig. 3d), which contains the pixels of the ground.

3.2 Pedestrian Classification

The university campus environment contains different kind of areas and obstacles (buildings, trees, lamppost pedestrians, garden, etc.) that have to be avoided while navigating. A basic classification for in-campus navigation requires at least identifying garden areas and dynamic obstacles, mainly pedestrians.

In the approach presented in this paper, obstacle classification is based on the determination of Regions of Interest in the visible image (Llorca et al., 2012). These ROIs isolate obstacles so that they can be processed by classifiers in subsequent stages. However, due to the characteristics of the campus environment, most of the obstacles arise in the proximity of the vehicle, so obstacles may be fragmented into different ROIs as they may get different disparity values due to high depth precision in the short distance, making classification process much harder. To address this issue, an algorithm for ROI grouping has been designed. Looking at the *u-disparity map*, obstacles are represented as continuous blobs in white. These blobs are analysed to compute the maximum and minimum disparity levels as well as the max and min horizontal coordinates of each obstacle. Afterwards, both depth and width data is used to group previously computed ROIs, therefore fixing obstacle fragmentation issue.

Once ROIs have been computed for each obstacle in the image, a HOG classifier is used to determine the probability of each region to be a pedestrian. In case of a positive classification, obstacle map is threshed to get a binary mask of the pixels within the previously obtained region's disparity range. Finally, pixels of the mask inside the ROI area are labelled as pedestrians.

This architecture design, where segmentation and classification stages are loosely coupled, makes it possible to use different classifiers for multiple classes and thus, provides great versatility.

3.3 Determination of Traversable Areas

Taking the free map as the starting point, *backprojection* algorithm is used to obtain the probability of a pixel belonging to garden class. For this stage, the visible image is converted into HSV colour-space as this kind of colour representation is more robust against light condition changes than RGB. After the conversion is performed, a synthetic histogram is built encompassing the HUE range corresponding to green values usually taken by garden areas. Then, it is used as input to the *backprojection* algorithm.

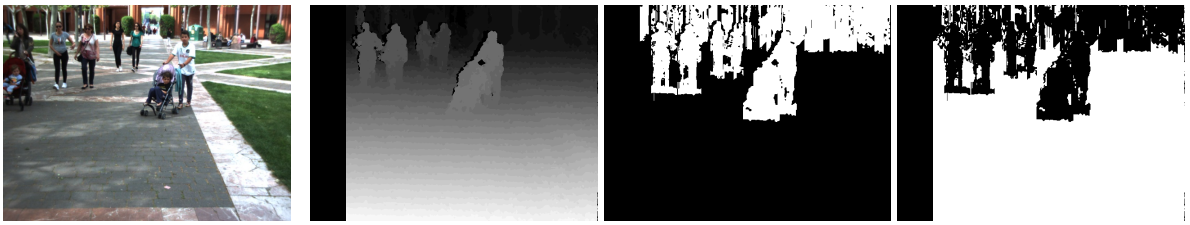
Once the probabilities of belonging to garden class are obtained, an empirically tuned threshold is used to label pixels as garden or traversable area.

4 RESULTS

In order to test the performance of the proposed method for dense image labelling, an annotated database is used. Since the research platform is aimed to navigate inside the university campus area in harmony with the university community, a specific ground truth is required for evaluation. Thus, a novel database (Beltrán et al., 2016) has been generated considering the particular needs for this task.

4.1 Dataset Description

The developed dataset is composed of a set of 30 manually-annotated images with a 640x480 resolution (see Fig. 4). Frames are captured by a Bumblebee 2 stereo rig sensor with a focal length of 6mm, a baseline of 0.12m, and a HFOV angle of 43°. The camera is mounted on the forepart of the iCab research platform. The small resolution of the dataset is due to the actual restrictions of the platform: limited computing resources as well as the demanding computing times required by real-time applications, like



(a) Visible image. (b) Disparity map. (c) Obstacle map. (d) Free map.

Figure 3: Example of the obstacle and ground estimation.

navigation in unstructured environments such as the University campus.

All images in the collection are part of a single sequence recorded by the iCab moving around the University campus. The original video rate is 20 fps with a total length of 60 seconds. However, although the labelling algorithm is working in real-time, only one out of every 40 frames of the video is used for performance evaluation.

Four different classes are used for labelling the dataset. The chosen categories correspond to the most popular instances found around the University campus and compose the minimum set required for in-campus navigation. The four categories are: traversable area, garden, obstacles and pedestrian. Segmentation and classification of the selected classes provide the necessary knowledge about the platform surroundings in order to detect traversable areas avoiding collision with static obstacles such as buildings and urban furniture and most common dynamic obstacles like pedestrians, being able to guarantee that the platform will not navigate over green areas and parks.

The selected dataset is publicly accessible and is composed of two sets of images: the original pair of colour images and their corresponding annotated ground truth containing the labels for the aforementioned classes. Fig. 4 shows an example of an original image and its annotations from the dataset. As can be observed, the ground truth is made up of one fully annotated RGB image *.png* for each original frame where each pixel takes the colour associated to the category it belongs to: *traversable area* in blue, *garden* in green, *obstacles* in red and *pedestrian* in yellow. Additionally, *unknown* pixels will be assigned black colour during classification stage.

4.2 Metrics

The computer used to perform the experiments to assess the performance of the proposed labelling method is an on-board embedded computer having an Intel Core i7 processor with 8 cores at 4.0 GHz and 16 GB RAM. All developed algorithms are running over ROS Kinetic on an Ubuntu 16.04 environment.

$$ACC = \frac{TP}{\text{num. of pixels}} \tag{1}$$

$$IoU_{class} = \frac{TP_{class}}{TP_{class} + FP_{class} + FN_{class}} \tag{2}$$

In order to assess the per-pixel semantic labelling performance of our method two different metrics are used. On the one hand, the global accuracy of the algorithm is measured as in (1), allowing to determine the proportion of properly-classified pixels. On the other hand, for the purpose of identifying the per-class classification accuracy, the Jaccard Index, commonly known as PASCAL VOC intersection-over-union, is used. The Jaccard Index is computed as in (2), where TP, FP, and FN stand for true positive, false positive and false negative pixels, respectively. The reported results correspond to the mean performance computed over the whole set of images.

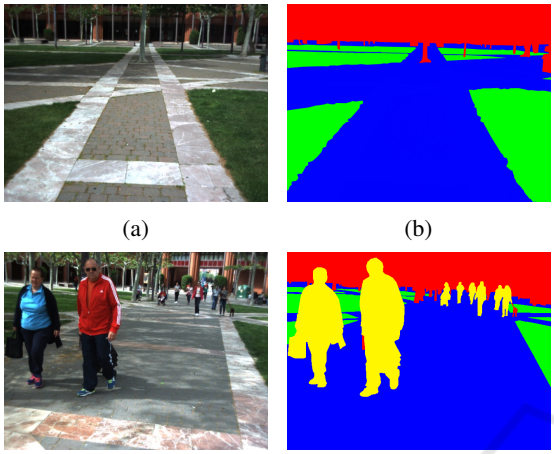
In order to measure the suitability of our method for the task of real-time image labelling for autonomous navigation, two well-known disparity algorithms are tested. Thus, it is possible to compare which one provides the best trade-off between per-pixel accuracy and computation time, provided that our method strongly relies on the quality of the disparity map at its first stages. The disparity methods considered in our experiments are Block Matching (BM) and Semi Global Block Matching (SGBM) (Hirschmuller, 2008).

4.3 Experimental Results

The results of applying our labelling algorithm to the dataset images are collected in Table 1 (see Fig. 5). As can be observed, there are no significant differences between the results of the two disparity methods, being the SGBM slightly more accurate in both per-pixel and per-class classification. These findings meet the expected outcomes provided that the semi-global algorithm take into account a greater amount of information in order to compute the depth estimation of each of the pixels.

Table 1: Four-classes classification performance (%).

Disparity Method	Pixel-wise Accuracy	IoU_{class}			
		Free	Garden	Obstacle	Pedestrian
BM	86.81	87.42	69.01	66.41	47.05
SGBM	87.79	88.34	70.31	67.75	43.43



(a) (b) (c) (d)

Figure 4: Image - Ground truth pair from dataset a) and c) Visible images. b) and d) Labelled ground truth. Best viewed in colour.

Despite of the minor differences, the proposed method behaves similarly for each of the selected metrics independently of the chosen disparity algorithm, thus indicating an existing trend in terms of per-pixel accuracy and per-class classification. Taking into consideration the results provided for each of the existing classes, it can be observed that there is an important difference between traversable area classification and the other categories. This is explained by the fact that the stereo sensor provides better depth estimation in the nearest environment as the textures are sharper. Thus, pixels corresponding to the areas which are closer to the camera will be more likely to get its disparity properly computed. Consequently, as the area in front of the vehicle is usually ground, best classification corresponds to *traversable* category. As a result, pixels belonging to obstacles usually lay on farther areas and tend to be more likely misclassified.

The performance of the garden classifier behaves as expected. As it is based on ground detection, its results are bounded by the free map segmentation phase. Moreover, garden recognition through *back-projection* (Swain and Ballard, 1992) algorithm is only based on HUE channel, therefore suffering from drastic contrast or light changes in the scene.

Finally, pedestrian classifier provides the worst per-class classification output. This situation is a consequence of using a standard HOG classifier from external libraries, which is not adjusting well to multi-scale classification and does not work on semi-occluded persons that may appear at the edges of the images or cropped at the segmentation phase. These issues can be observed in detail in Fig. 6.

Regarding the overall per-pixel accuracy, the algorithm labels properly more than 85% of the pixels in each frame, overcoming by far the average IoU_{class} and getting really close to the best classified category, as it makes up most of the pixels in the sequence.

Table 2: Three-classes classification performance (%).

Disparity Method	Pixel Accuracy	IoU_{class}		
		Free	Garden	Obst.
BM	88.35	87.42	69.01	72.31
SGBM	89.53	88.34	70.31	74.74

For the purpose of analysing the effect of the pedestrian classification results in the method accuracy, an experiment has been carried out considering only the other three categories. Table 2 comprises the performance of the proposed method for this case. As can be observed, both the $IoU_{obstacle}$ and the overall accuracy get higher, being more significant the obstacle class increment. These variations shows how much pedestrian classifier deteriorates overall accuracy. Furthermore, the big growth in obstacle classification performance indicates that not only the classifier produces many false positive and false negatives, but also that some pixels labelled as pedestrian, really belong to obstacle category. However, as they are wrongly grouped as part of the person shape in the segmentation phase, they end up being considered as part of the pedestrian class.

Considering traversable and garden categories, they both provides the same performance as in the 4-class case, as long as they are not affected by the pedestrian classifier.

4.4 Computational Time

As time is a key factor in real-time algorithms, measured times for the performed experiments are presented in Table 3 for the purpose of determining which configuration is more suitable for the task of scene labelling for autonomous navigation.

As can be appreciated, the Block Matching algorithm is considerably less time consuming than

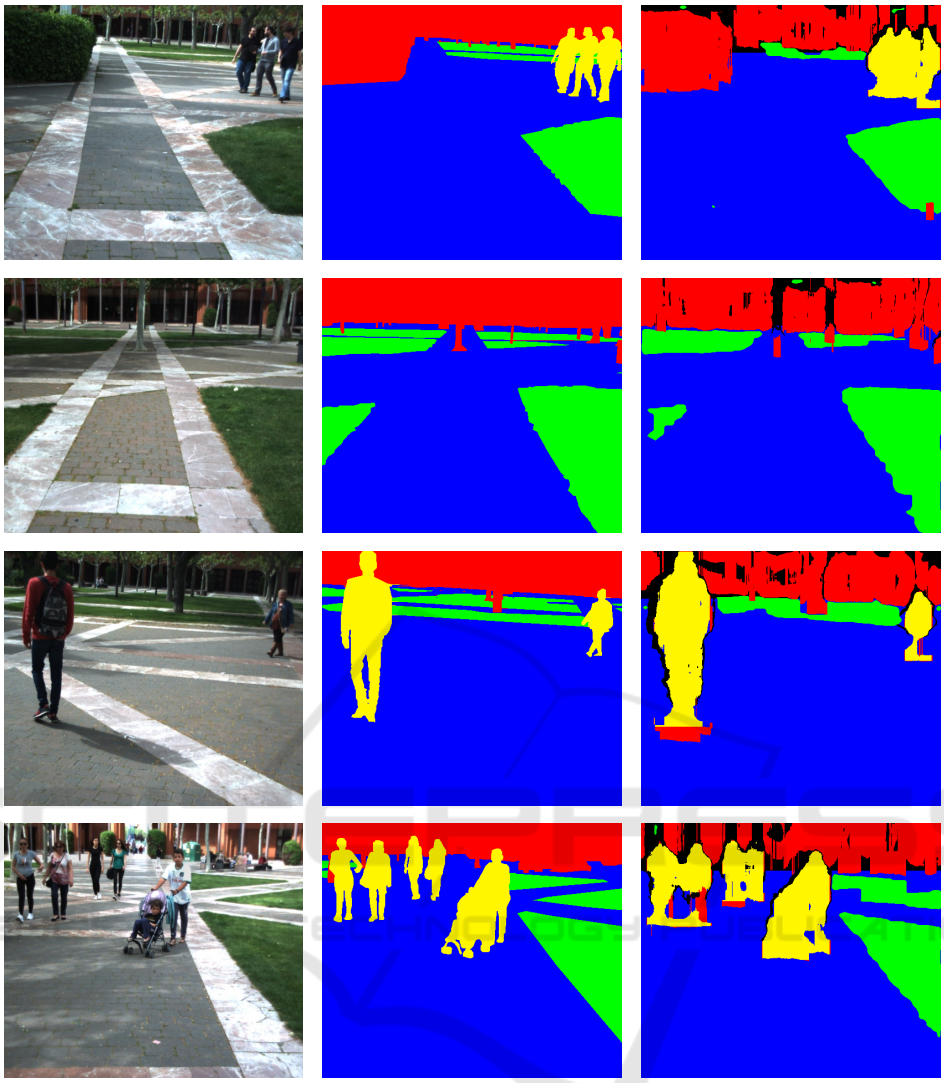


Figure 5: Labelling results. From left to right: visible image, ground truth and labelled image. Best viewed in colour.

Table 3: Computing times (ms).

Disparity Method	# Classes	
	4 classes	3 classes
BM	120.75	65.10
SGBM	187.98	134.32

SGBM at the expense of depth estimation accuracy. In addition, there is a big difference between 4 and 3-class cases, indicating that pedestrian classifier is not only the worst at performance, but also one of most time costly. However, taking into consideration both the real-time requirements and the perception needs for in-campus navigation task, it can be asserted that the best configuration for the proposed method is composed of Block Matching disparity method and 4-class classifier, as long as it gives the best trade-off between the labelling accuracy and the needed operating frame rate, since camera images come at 20Hz and labelled images are available at 16Hz due to parallel processing. Therefore, this setup is suitable to work in real-time on the described platform since iCab's top speed is 10km/h, thus covering 0.18m between two consecutive frames when driving at maximum speed.

5 CONCLUSIONS

Environment understanding plays a key role in any autonomous navigation task. Therefore, being able to classify the information obtained from sensors such as cameras and lasers is essential for the process of generating a reliable map that allows path planning. The main contribution of this paper is to provide an algorithm for dense scene labelling tested on

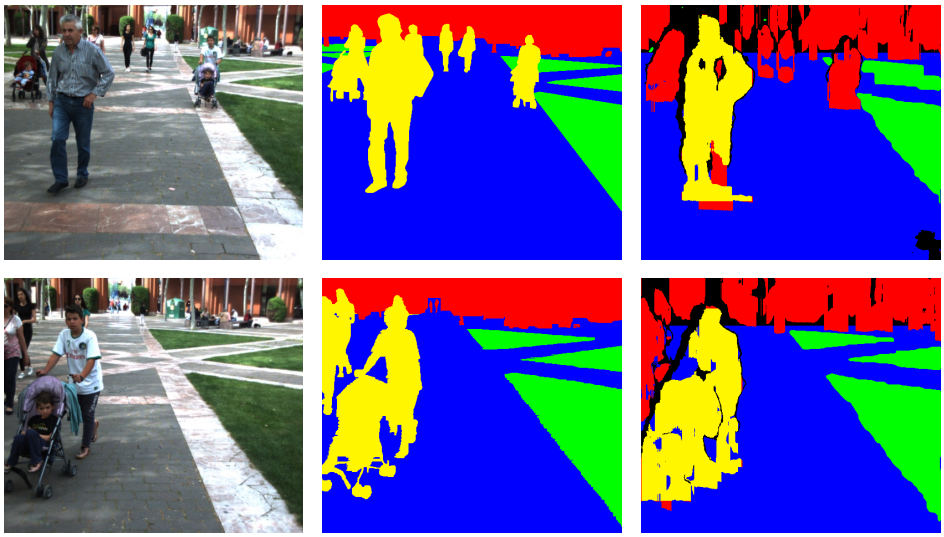


Figure 6: Labelling issues. From left to right: visible image, ground truth and labelled image. Best viewed in colour.

off-road environments based on a flexible and decoupled architecture built on ROS. Moreover, a novel manually-annotated dataset is presented for evaluation purposes.

The results of the proposed approach show a high performance at labelling in-campus scenarios. This accuracy, together with the low computational times, make the introduced algorithm a suitable and efficient solution for dense image classification within the environment under research.

In addition, the modularity of the proposed architecture provides the ability to adapt the algorithm to work well for other environments. The highly decoupled design gives the possibility of extending the number of categories by adding other classifiers in a plug-and-play manner. Similarly, the existing modules can be easily upgraded or replaced by new ones which might grant better performance.

In the future, pedestrian classifier will be replaced by a more advanced algorithm with better capabilities for multi-scale classification. Moreover, garden classifier will be tuned to combine information from all three channels (H, S and V) to improve results.

Finally, the released database will be extended with images from other sequences as well as non-used images of the actual video. In addition, database might be re-annotated to include more classes in order to achieve a better understanding of the vehicle surroundings.

ACKNOWLEDGEMENTS

This work was supported by the Spanish Government through the CICYTprojects (TRA2013-48314-C3-1-R and TRA2015-63708-R) and Comunidad de Madrid through SEGVAUTO-TRIES (S2013/MIT-2713).

REFERENCES

Arnab, A., Jayasumana, S., Zheng, S., and Torr, P. H. (2016). Higher order conditional random fields in deep neural networks. In *European Conference on Computer Vision*, pages 524–540. Springer.

Beltrán, J., Jaraquemada, C., Musleh, B., de la Escalera, A., and Armingol, J. M. (2016). SAUCE, Semantic Annotated University Campus Environment. Dataset. http://dx.doi.org/10.5281/ZENODO.167843.

Bernini, N., Bertozzi, M., Castangia, L., Patander, M., and Sabbatelli, M. (2014). Real-time obstacle detection using stereo vision for autonomous ground vehicles: A survey. In *Intelligent Transportation Systems (ITSC), 2014 IEEE 17th International Conference on*, pages 873–878. IEEE.

Broggi, A., Cappalunga, A., Caraffi, C., Cattani, S., Ghidoni, S., Grisleri, P., Porta, P., Posterli, M., Zani, P., and Beck, J. (2008). The passive sensing suite of the terramax autonomous vehicle. In *Intelligent Vehicles Symposium, 2008 IEEE*, pages 769–774. IEEE.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Garage, W. (2010). Ros. *ros. org*.

Golodetz, S., Sapienza, M., Valentin, J. P., Vineet, V., Cheng, M.-M., Arnab, A., Prisacariu, V. A., Kähler, O., Ren, C. Y., Murray, D. W., et al. (2015). Semanticpaint: A framework for the interactive segmentation of 3d scenes. *arXiv preprint arXiv:1510.03727*.

Guo, C., Mita, S., and McAllester, D. (2009). Drivable road region detection using homography estimation and efficient belief propagation with coordinate descent optimization. In *Intelligent Vehicles Symposium, 2009 IEEE*, pages 317–323. IEEE.

Hirschmuller, H. (2008). Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on pattern analysis and machine intelligence*, 30(2):328–341.

Hu, Z., Lamosa, F., and Uchimura, K. (2005). A complete uv-disparity study for stereovision based 3d driving environment analysis. In *3-D Digital Imaging and Modeling, 2005. 3DIM 2005. Fifth International Conference on*, pages 204–211. IEEE.

Hu, Z. and Uchimura, K. (2005). Uv-disparity: an efficient algorithm for stereovision based scene analysis. In *IEEE Proceedings. Intelligent Vehicles Symposium, 2005.*, pages 48–54. IEEE.

Hussein, A., Marín-Plaza, P., Martín, D., de la Escalera, A., and Armingol, J. M. (2016). Autonomous off-road navigation using stereo-vision and laser-rangefinder fusion for outdoor obstacles detection. In *Intelligent Vehicles Symposium (IV), 2016 IEEE*, pages 104–109. IEEE.

Labayrade, R. and Aubert, D. (2003). In-vehicle obstacles detection and characterization by stereovision. *Proc. IEEE In-Vehicle Cognitive Comput. Vis. Syst*, pages 1–3.

Llorca, D., Sotelo, M., Hellín, A., Orellana, A., Gavilán, M., Daza, I., and Lorente, A. (2012). Stereo regions-of-interest selection for pedestrian protection: A survey. *Transportation research part C: emerging technologies*, 25:226–237.

Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440.

Marın-Plaza, P., Beltrán, J., Hussein, A., Musleh, B., Martın, D., de la Escalera, A., and Armingol, J. M. (2016). Stereo vision-based local occupancy grid map for autonomous navigation in ros.

Mottaghi, R., Chen, X., Liu, X., Cho, N.-G., Lee, S.-W., Fidler, S., Urtasun, R., and Yuille, A. (2014). The role of context for object detection and semantic segmentation in the wild. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Musleh, B., de la Escalera, A., and Armingol, J. M. (2011). Uv disparity analysis in urban environments. In *International Conference on Computer Aided Systems Theory*, pages 426–432. Springer Berlin Heidelberg.

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

Richter, S. R., Vineet, V., Roth, S., and Koltun, V. (2016). Playing for data: Ground truth from computer games. *arXiv preprint arXiv:1608.02192*.

SAE On-Road Automated Vehicle Standards Committee (2014). Taxonomy and definitions for terms related to on-road motor vehicle automated driving systems.

Sengupta, S., Greveson, E., Shahrokni, A., and Torr, P. H. (2013). Urban 3d semantic modelling using stereo vision. In *2013 IEEE International Conference on Robotics and Automation*. IEEE.

Sengupta, S., Sturgess, P., Torr, P. H., et al. (2012). Automatic dense visual semantic mapping from street-level imagery. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 857–862. IEEE.

Soquet, N., Perrollaz, M., Labayrade, R., Aubert, D., et al. (2007). Free space estimation for autonomous navigation. In *5th International Conference on Computer Vision Systems*.

Swain, M. J. and Ballard, D. H. (1992). Indexing via color histograms. In *Active Perception and Robot Vision*, pages 261–273. Springer.

Urmson, C., Anhalt, J., Bagnell, D., Baker, C., Bittner, R., Clark, M., Dolan, J., Duggins, D., Galatali, T., Geyer, C., et al. (2008). Autonomous driving in urban environments: Boss and the urban challenge. *Journal of Field Robotics*, 25(8):425–466.

Yao, J., Fidler, S., and Urtasun, R. (2012). Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 702–709. IEEE.