Scene parsing and fusion-based continuous traversable region formation

Xuhong Xiao, Gee Wah Ng, Yuan Sin Tan, Yeo Ye Chuan

20 Science Park Drive, DSO national Laboratories, Singapore 118230

Abstract. Determining the categories of different parts of a scene and generating a continuous traversable region map in the physical coordinate system are crucial for autonomous vehicle navigation. This paper presents our efforts in these two aspects for an autonomous vehicle operating in open terrain environment. Driven by the ideas that have been proposed in our Cognitive Architecture, we have designed novel strategies for the top-down facilitation process to explicitly interpret spatial relationship between objects in the scene, and have incorporated a visual attention mechanism into the image-based scene parsing module. The scene parsing module is able to process images fast enough for real-time vehicle navigation applications. To alleviate the challenges in using sparse 3D occupancy grids for path planning, we are proposing an approach to interpolate the category of occupancy grids not hit by 3D LIDAR, with reference to the aligned image-based scene parsing result, so that a continuous $2\frac{1}{2}D$ traversable region map can be formed.

1 Introduction

It is widely accepted that humans possess two distinct visual pathways, the ventral stream (known as "what pathway") involved with object identification and recognition, and the dorsal stream (or, "where pathway") involved with processing the object spatial location relevant to the viewer [1, 2]. Similarly, for autonomous vehicle to successfully drive on road, our scene understanding systems must tell the vehicles not only what are in the environment, but also where the roads and obstacles are in a physical coordinate system, other than in the image plane.

In the last two decades, the study on computer vision systems mainly focus on solving the problem about what are in a scene, taking images as input. Many prototypes detect/recognize individual objects in the images. Effective features, such as SIFT [3] and HOG [4], have been proposed and widely used in these systems. The HMAX [5, 6] tried to model objects via a hierarchical representation involving increasingly complex features. In [7], Latent SVM was proposed to learn deformable part models. One common feature of these systems is that they all apply sliding windows to exhaustively search for object locations. To reduce unnecessary computation and speed-up the detection process, strategies like combining multiple, increasingly complex classifiers in a "cascade" have been

proposed [8] and adapted in [9]. Some other systems recognize overall scene categories (e.g., beach) from the image, usually making use of features reflecting the "gist" of scenes [10–14]. To facilitate autonomous systems' response to environment, it is insufficient to recognize only a certain category of objects. Instead, scene parsing, which conducts classification for all parts in the scene including road and vegetation, is required to form a continuous traversable map for navigation. While early scene parsing systems tried to classify each pixel separately, recent scene parsing systems perform the task on the superpixel level. In the superParsing system [15], features like shape, SIFT, color and appearance are extracted for each superpixel. Total scene understanding, which conducts scene parsing and object detection simultaneously, has also been studied, as in [16, 17]. Inspired by the findings in cognitive science, the strategies taken for object recognition and scene understanding also evolves from pure bottom-up processing to combination of both bottom-up processing and top-down facilitation [18, 19]. Graph models have been proposed to control the top-down process, among which, CRF (Conditional Random Field) has been the most popular framework [20–22].

Another line of research in perception for autonomous vehicles is to understand "where" the objects are in a physical coordinate system, which is necessary for autonomous vehicles to navigate or to response to the environment. 3D LIDAR sensors have been widely used to achieve it, as demonstrated in DARPA Grand Challenges [23]. 3D LIDAR-based scene classification is usually conducted on occupancy grids, whereby the physical region is divided into 3D cubes or 2D grids of pre-defined size. Authors in [24] have made attempt to classify, from LIDAR responses, classes including surface (ground bare terrain surface, solid object, large tree trunk), linear structures (wires, thin branches) and scatter (tree canopy, grass), based on eigen-values of the covariance matrix of the local pointcloud. In [25], an SVM classifier is trained using features extracted from LIDAR points, such as intensities of LIDAR points, scatterness, linearness, and surfaceness. One disadvantage of 3D LIDAR based classification is that the classified grids are not continuous due to the sparseness of pointcloud. Furthermore, the classified grids will get even sparser with the increase of distance from the vehicle, resulting in extra challenges to the path planning and navigation modules.

With the progress in the above two directions, it is natural to compensate LIDAR with image based scene understanding. For example, Stanley [26], the 1st winner of DARPA Grand Challenge 2004, made use of 3D LIDAR data to conduct terrain labelling, and image-based vision analysis for early warning of obstacles in the distance beyond the range of LIDAR. Little fusion of image and LIDAR was involved in the Stanley system. There are also work to fuse LIDAR and camera for road boundary detection [27, 28], but such tracking based approaches are constrained to roads with nearly parallel borders or when the road model is known *a prior*. They do not work well on complex terrains, e.g., open fields without clear road boundary.

Our work follows the line of fusion of image and LIDAR for scene parsing. The scene parsing module is an integral part of the Cognitive Architecture, for which we have proposed a computational infrastructure that defines the various regions and functions working as a whole to produce human-like intelligence [2, 31]. In particular, for scene parsing, the architecture has specified sub-functions including initial scene classification based on bottom-up, low-level features, top-down facilitation, visual attention-based priming and object-level fine grained classification. In this paper, we will first give some details about our strategy for initial classification and context-based top-down facilitation. We will then focus on reporting our effort to combine the LIDAR and scene parsing result to create a continuous traversable region map for vehicle navigation in outdoor, off-road environment. In the same time, we will also exemplify how the visual attention mechanism is implemented and applied to enhance the capability of man-made obstacle detection from images.

Compared with existing work, our system has some special features. First, we have incorporated not only the contextual-based top-down processing in scene parsing, but also the visual attention mechanism to enhance the capability of obstacle (man-made objects) detection. Although strategies of applying different features to detect various objects have been widely used [33], traditional super-pixel level scene parsing techniques treat all parts of an image equally, ignoring the function of visual attention in finding objects of interest [34]. Secondly, although it is not new anymore to utilize the top-down process to enhance scene parsing, most existing systems only make use of the co-occurrence or neighboring relations between objects in the process, while our approach explicitly interpret the relative spatial relations between irregular-shape components in a semantic way (for example, in a front view image, road regions cannot be above the sky, but can be below the sky). It is no doubt that the interpretation of specific spatial relation will solve uncertainties in the initial classification more efficiently than simple neighboring relations. Thirdly, to our knowledge, this is the first work to generate continuous road regions from combination of scene classification and LIDAR detections, making it possible to provide continuous traversable region map under complex situations, e.g., when there are obstacles on road and the road boundary is cluttered or in irregular shapes.

2 Image-based scene understanding in the Cognitive Architecture

Fig. 1 presents the high-level structure of the visual perception module, showing the biological regions where the sub-functions are accomplished in human brains and the interaction between the perception module and other modules, such as the Reasoner module, in our Cognitive Architecture. Total scene understanding is achieved via two processes:

1. An interactive process among initial classification and top-down facilitation, which emulates the typical bottom-up and top-down interaction [31,

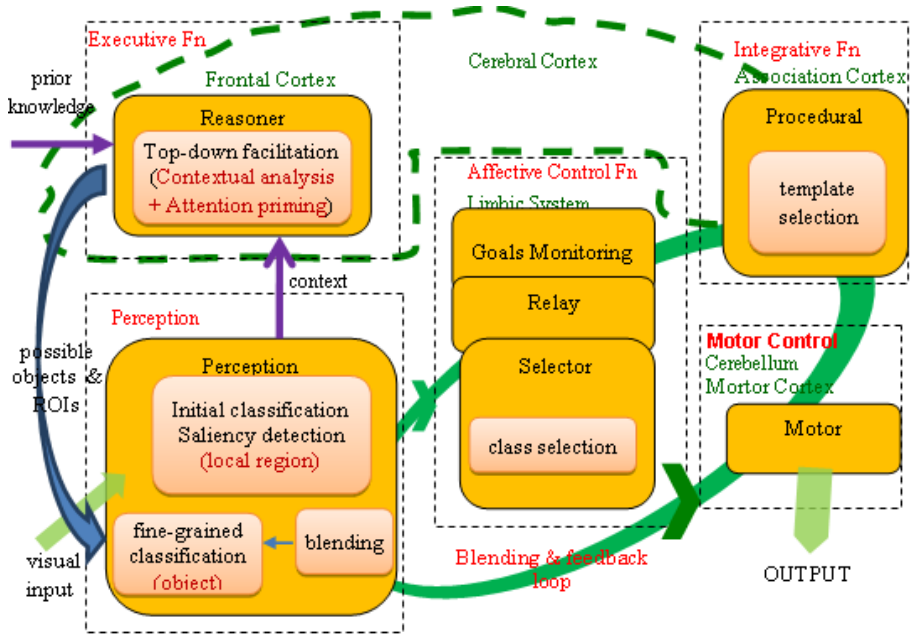

Fig. 1. Building block of the visual perception module in DSO-CA

32], and the visual attention process[34]. Initial classification refers to the early, coarse-level classification (e.g., road, vegetation) of scene parts based on low-level features. The visual attention mechanism is responsible for identifying potential objects of interest, some of which may have been successfully detected by the coarse-level initial classification sub-module. Furthermore, it will also pick up some new objects of interest which are unknown to the learning-based initial classification sub-module. Top-down facilitation is not only responsible for making use of contextual knowledge to resolve uncertainties in initial classification through a contextual analysis process, but also responsible for suggesting regions of interest worthy of further attention via attention priming.

2. Fine-grained object classification which determines more specific categories of objects of interest (e.g., pedestrians, bus, cars, etc), identified by attention priming. Known object classification techniques, such as HOG-based classification[4], and the deformable part-based models[7], can be adapted for the purpose. In particular, a special blending mechanism, as proposed in [29] has been adapted in the Cognitive Architecture, as shown in Fig. 1.

In comparison, most state-of-the-art scene parsing systems treat every part of the image equally, in that they classify each part (an object may be segmented into several parts), while individual object detection systems search for specific categories of objects via sliding window. There is barely a system that conducts both functions simultaneously. Furthermore, although sliding window

based search of objects is effective in engineering systems, it is not consistent with the biological way of object detection, and it takes unnecessary longer time to process irrelevant information. According to cognitive findings, visual search is conducted via the visual attention mechanism: individual objects pop-out due to its saliency, and task-based intention help to achieve selective attention[34], resulting in detections of potential objects of interest. Fine-grained object classification is conducted only to verify these potential object in the identified regions. The integration of attention mechanism in our framework makes it more similar to biological system in quickly switching the focus of process to the regions potentially containing objects of interest.

A detailed breakdown of the scene understanding module is illustrated in Fig. 2. The scene part parsing mechanism will classify each superpixel via a bottom-up initial classification process and a top-down contextual analysis process. The visual attention mechanism involves bottom up saliency map generation and binding, as well as top-down attention priming.

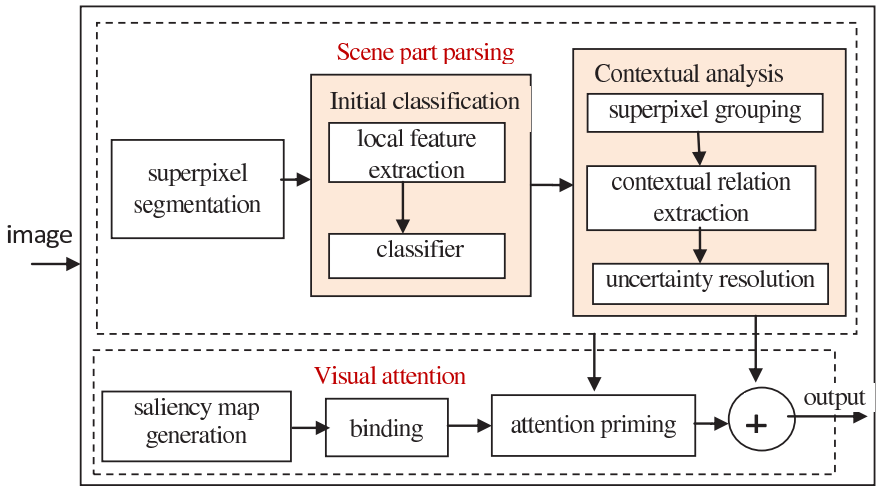

Fig. 2. Detailed diagram of the implementation of scene understanding

2.1 Initial classification

The initial classification sub-module consists of algorithms to achieve early, coarse-level classification of local image regions. Each image is over-segmented into superpixels as described in [35]. Color and texture features are extracted to describe each superpxiel. In detail, histograms of RGB/HSV color features, anisotropic Gauss filtering responses[36], Gabor filter responses[37] and Local Binary Patterns[38], can be extracted for each superpixel.

Popular classifiers, such as Support Vector Machine (SVM) and Multiple Layer Perceptron (MLP) are included in the Perception module of the Cognitive Architecture. According to our experience, the MLP classifier can consistently achieve similar accuracy to that achieved by SVM with RBF kernel, which is much better than linear SVM for natural scene parsing. Besides, MLP can run much faster than SVM with RBF kernel when there are thousands of support vectors generated, which is always the case for natural scene parsing.

2.2 Top-down facilitation via contextual analysis

The design of the top-down facilitation strategy in the Cognitive Architecture is mainly motivated by the cognitive findings as reported in [39]. According to [39], there are two factors contributing to top-down facilitation: the object-based and context-based facilitation. The object-based mechanism refers to the case that "initial guess" of object type based on low level information triggers a more fine-grained object classification process, which is emulated by the attention priming mechanism in the framework. The context-based mechanism triggers top down facilitation through contextual association between objects in scenes. The contextual association activates predictive information about which objects are likely to appear together, and can influence the "initial guesses" about an object's identity. It has been widely accepted[40] that contexts affects classification in two aspects: (1). The presence of objects that have a unique interpretation improves recognition of ambiguous object in a scene; (2). Proper spatial relations among objects decreases error rate in the recognition.

Spatial relations have been applied to improve object detection, which confines objects by bounding boxes [41]. However, most superpixel-based scene parsing systems limit contextual analysis to the co-occurrence of neighboring objects. As a matter of fact, the spatial relation is very useful for natural scene classification. For example, we know that "road" cannot be on top of a "tree", but "road" is possible to appear beside a "tree". In this case, simple co-occurrence, or adjacency relation between "road" and "tree" will not help much in resolving uncertainties in the classification of either "road" or "tree". However, the spatial relation, one is on top of the other, will clearly verify the situation. To efficiently interpret such semantic relations, we first group the superpixels into connected components based on their initial classification. For example, as shown in Fig. 3(b), all superpixels falling on the circled component, which are initially classified as "sky", are grouped together. The contexts about whether its top, bottom and sides are tree, sky, etc., are extracted and represented as soft evidence passed to a learned Nave Bayes structure illustrated in Fig. 4. Each node in the structure, except the node "category", represents a spatial relation. For example, the node "isTopSky" has two values: 0 if the top of the component under consideration is not "sky", 1 otherwise. Inference based on soft evidence updates the probability of the values of node "category", i.e., the classification of the component under consideration. For example, based on the evidence that the top and two sides of the circled component in Fig. 3(b) are all "road", the classification of the component will be updated to "road", as shown in Fig. 3(c).

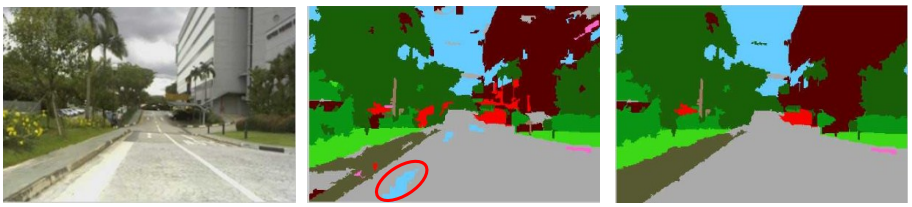

Fig. 3. Illustration of the contextual analysis process. (a). Original image; (b). Result of initial classification, with misclassifications for some parts; (c). Result after contextual analysis.

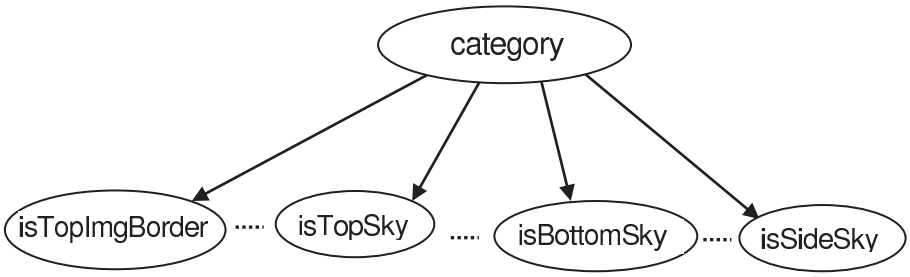

Fig. 4. Nave Bayes structure for contextual analysis

2.3 Visual attention

In a computer vision system, the purpose of visual attention module is to determine the region of objects of interest. The visual attention mechanism in the framework consists of a saliency map generation submodule, a binding submodule and a top-down priming submodule. In the following, we will use the image in Fig. 5(a) to illustrate how the mechanism works.

First, an initial saliency map is computed for the image based on the algorithm in [42]. This initial saliency map is further refined by suppressing the effect of large patches of background, whereby the background region is determined based on the consistency of color/intensity of the image. The refined saliency map is shown in Fig. 5(b). The task of binding is to group the saliency regions into proto-objects, which are believed to be the form of output in visual attention [43]. The MSER algorithm [44] is modified to accomplish the binding task. Major modifications to the MSER algorithm lie in the change of maximally stable criteria. Besides the ratio of region change as defined in [45], we also consider the orientation consistency when two components are to be merged, as well as the contour completeness. The bounding boxes of the initial proto-objects are shown in Fig. 5(c). There are false alarms of proto-objects in Fig. 5(c) because only information from the bottom-up saliency map and the contour is made use of at this stage. The top-down process - attention priming, steps in to reduce false alarms. It works as follows: First, it makes use of domain knowledge about

the task requirement to determine potential object size and object types. For example, for a driving vehicle, the potential objects of interest may be cars and other obstacles on the road that the vehicle should avoid colliding. This inference will further activate the use of other knowledge such as "obstacles should be on the road, not on the top of trees". Combining such domain knowledge with the image context derived from the scene parsing process (as shown in Fig. 5(e)), the priming process will switch the attention to the true object of interest, removing the false alarms that does not fit for our domain knowledge, leading to the result as in Fig. 5(d), with only one potential obstacle. This visual attention result is then applied to update the scene part parsing result on superpixel level, and the final scene understanding result is shown in Fig. 5(f).

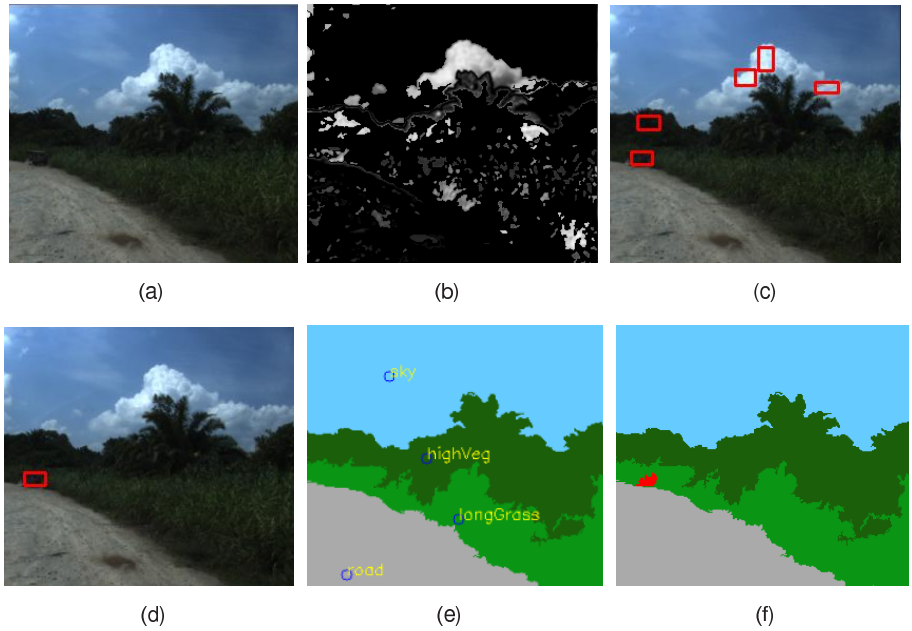

Fig. 5. : Illustration of the Visual attention process. (a). Original image; (b). Saliency map; (c). Bounding boxes for initial proto-objects output by binding process, (d). Final bounding box for a proto-object after attention priming; (e). Output of the scene part parsing mechanism; (f). Final output combing scene part parsing and visual attention.

The example shown in Fig. 5 clearly demonstrates one function of the visual attention mechanism - it can discover some potential objects of interest that the scene part parsing mechanism misses. As shown in Fig. 5(e), the obstacle (in this case, the truck) is misclassified as "longGrass" by the scene part parsing mechanism. There are two possible reasons for the misclassification. It is possible that the superpixel level classification, which is based on only partial structure of objects, is not effective in detection of obstacles. It is possible too that the

object is "new" to the training-based scene part parsing process. However, the visual attention mechanism, which is based on contrast and contours, successfully makes the obstacle "pop out", and complement the scene part parsing mechanism in detection of obstacles. On the other hand, the information from scene part classification provides scene contexts for attention priming.

The scene part parsing mechanism has been tested with many scenarios of different terrains. In a recent exercise, we have collected a variety of data covering diversified terrains like cluttered unstructured fields, narrow tracks with water bodies and wide open areas with ponds. From the data collected, we selectively labelled 356 images, with effort to ensure that the selected images are sufficient to reflect the diversities in terrain and illumination changes. From these labelled images, we have randomly chosen 178 images to train the classifier for initial classification and the Naive-Bayes model for contextual analysis, while the remaining 178 images for test. The F-measure (measured on superpixel level) for the major categories is shown in Table 1.

Table 1. F-measure for different categories

processing stage	F-measure (%)					
	road	highVeg	longGrass	sky	water	obstacle
initial classification	0.961	0.910	0.767	0.991	0.726	0.624
after contextual analysis	0.965	0.922	0.784	1.0	0.731	0.622

As can be easily observed from Table 1, the top-down contextual analysis improves the classification performance for most categories, and the scene part parsing submodule performs constantly well for categories like "road", "highVeg" (representing high vegetation) and "sky", which are the major categories in off-road natural scenes. Considering that the vehicle has to traverse over long grass on narrow tracks, a category "longGrass" is added to the classification module. Although "longGrass" is quite confusing with "highVeg", the scene part parsing mechanism works reasonably well for the category.

Another feature of the module is its high efficiency in computation. It takes only about 0.08 seconds to process an image of 400X300 in a 64-bit Windows system with Inter i7-3520 Dural Core and 8GB RAM.

The scene part parsing submodule is not very successful in classification of "obstacle", which includes all kinds of man-made structures/objects in our case. It is reasonable considering the diversity of obstacles and that the scene parsing is conducted on superpixel level, which may not extract complete topological features to interpret the structure of obstacles as a whole. Its weakness in obstacle detection can be partially complemented by the visual attention mechanism, as shown in Fig. 5. It will also be complemented by the LIDAR obstacle detector, which is known to be good for obstacle detection.

3 Fusion-based traversable region formation

Fig. 6 presents our framework to fuse the image-based scene parsing module and LIDAR detections. Both monocular camera and LIDAR are installed on top of an autonomous vehicle. The LIDAR applied is Velodyne 64-HDL with 64 sensors to generate pointcloud. A LIDAR based detector classifies the 3D points into ground points or obstacle points mainly based on their z-coordinates (the height of objects). It futher groups the obstacle points into clusters based on the distance between the points. A data registration process is conducted to align the image frames and LIDAR scans in time and estimate the homography transform so as to acquire the point-level correspondence between image pixels and 3D LIDAR points.

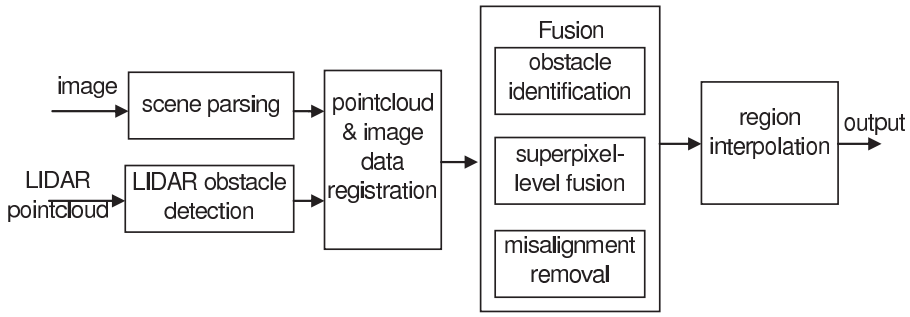

Fig. 6. Framework for image and LIDAR fusion

Fig. 7 illustrates an example of data registration. The original image is shown in Fig. 7.(a), and the aligned LIDAR occupancy grid with cell size 20cm x 20cm) is shown in Fig. 7.(e). Fig. 7(f) shows the clusters (on occupancy grid) provided by the LIDAR detector, while Fig. 7(b) shows the correspondence between the LIDAR points and the image after data registration. In both Fig. 7(b) and Fig. 7(f), ground points are represented in gray, while each other color corresponds to an individual cluster output by the LIDAR detector.

3.1 Image and LIDAR fusion

The fusion module first identifies the LIDAR clusters belonging to man-made "obstacle" and "longGrass" based on features including the height of the cluster, its position relative to the ground plane, depth difference within the cluster, and the image-based scene parsing result. It then makes use of LIDAR information to achieve more reliable scene parsing. There are several cases that LIDAR information will help to resolve ambiguities in image-based scene parsing. For example, if all 3D points projected to a superpixel are on ground level, the superpixel is unlikely to be part of an obstacle. On the other hand, if most of the 3D

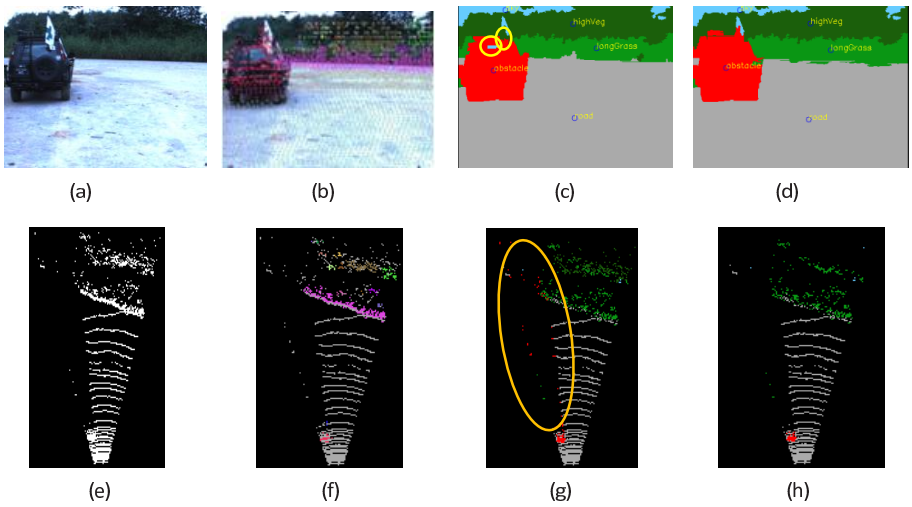

Fig. 7. Illustration of the fusion process. (a). Original image; (b). LIDAR clusters projected to the image; (c). scene classification without fusion, with errors in regions A and B; (d). scene classification after fusion; (e).Original LIDAR pointcloud (on occupancy grid); (f). LIDAR clusters output by the LIDAR detector; (g). Classification of LIDAR cells with presence of misalignment; (h). Classification of LIDAR cells with misalignment removed.

points projected to the superpixel belong to obstacle, the superpixel is unlikely to be classified as "road". The fusion strategy is similar to that applied in [29]. Once the classification of superpixels is updated, the categories of the individual 3D points are set to be the same as the corresponding image point. Accordingly, the classification of the cells in the occupancy grid that are occupied by LIDAR points can be determined based on that of points in it using majority voting rule. As shown in Fig.7(g), gray points correspond to ground cells, red for obstacle cells, green for long grass and dark green for high vegetation which is non-traversable.

Another issue in the fusion process is the imperfectness of data registration. For example, in Fig.7(g), there are some scattered false alarms of "obstacle" cells further away from the real obstacle. This is because, due to misalignment, some 3D points of road and vegetation in the distance are projected to the obstacle in the image, resulting in misclassification to these 3D points. The misalignment removal mechanism removes such misclassified points by referring to the classification of 3D clusters, utilizing ad-hoc rules. For example, If LIDAR points of a non-obstacle cluster is projected to an "obstacle" in the image, they are removed from the occupancy grid. After misalignment removal, we will get the updated, more reliable classification of occupancy grid , as shown in Fig.7(h).

3.2 Region interpolation

The classified cells as shown in Fig.7(h) are very sparse, resulting in extra challenge for autonomous vehicle navigation. We are intending to make the traversable region continuous to facilitate navigation.

The main technique used in continuous region formation is interpolation. To begin with, the physical regions corresponding to the field of view of the image are divided into an occupancy grid of 20cm x 20cm. The classification of the cells in the occupancy grid that are hit by LIDAR points can be determined based on that of points in it using majority voting rule. The task of interpolation is to infer the category of those unoccupied cells based on their occupied neighbors. The interpolation is conducted in order of categories. The first category to be processed is "road/ground", followed by other categories that are adjacent to the road or overlaid on the road, for example, "obstacle" and "longGrass".

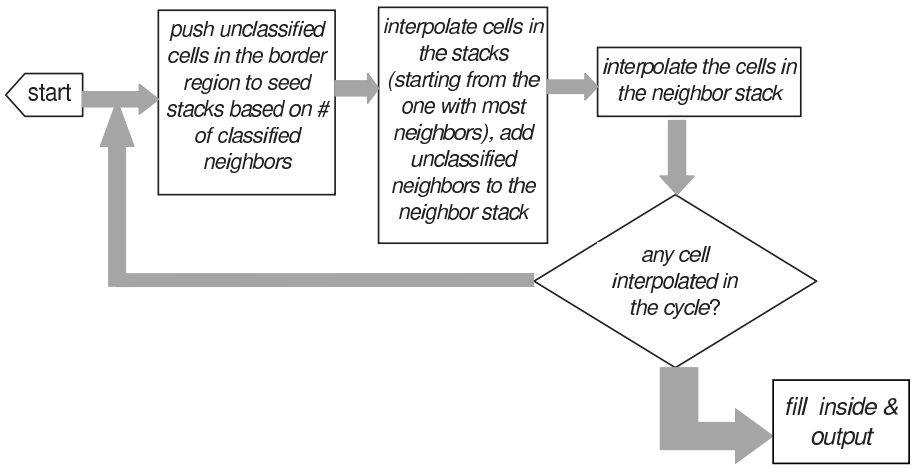

Fig. 8. Flowchart of the interpolation process along a border

For each category of interest (e.g., "road", "obstacles" and "longGrass"), the connected components are first identified in the image plane. For each connected component, the classified cells on the LIDAR occupancy grid are identified along the left, right, top and bottom borders of the component. The bordering cells along each direction are discontinuous, and they may not reflect the actual border position due to the sparseness of the classified cells. However, they form a proper estimation about the border regions of the component. For each direction, the search and interpolation of actual borders will be around the initial bordering region. As shown in Fig. 8, along each direction, the non-classified cells which have at least one neighboring cell classified, are pushed to respective seed stacks based on the number of classified neighbors. A classified neighbor is a cell on the occupancy grid, the category of which has been either inferred in the fusion

process, or interpolated before. The interpolation will start from the cells with the most classified neighbors: the z-coordinate of the cell is set to the averaged z-coordinates of its classified neighbors, the coordinates of the four corners of the cell (the x and y coordinates of the cell is known based on the location of the cell on the occupancy grid) is then projected to the image plane. If all the four corners fall in the connected component under consideration, the cell is considered to belong to the component. Therefore, its category is set to be the same category as the component. In the meantime, its unclassified neighbors are pushed to the neighbor stack. After all the cells in the seed stacks are interpolated, the cells in the neighbor stack will be interpolated. This interpolation process will be repeated until there is no new cell processed in the iteration.

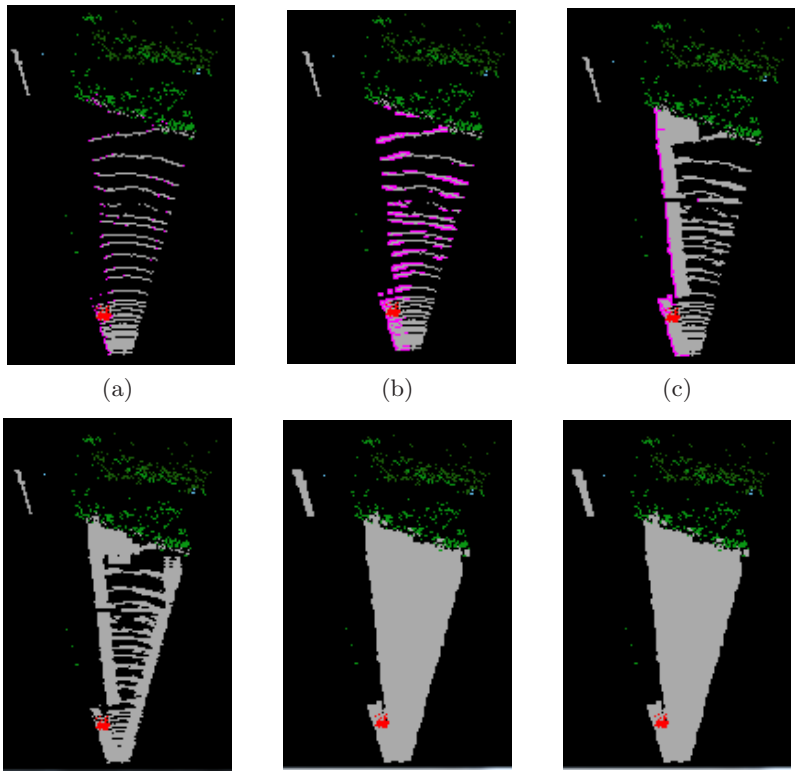

Fig. 9. Figure 9: Illustration of the interpolation process.(a). 1st iteration along left (seed cells are in purple); (b).3rd iteration along left; (c).10th iteration along left; (d). 4 borders of road are interpolated; (e). internal area of road are filled; (f). obstacle and long grass regions are interpolated

Fig. 9(a) to Fig. 9(c) illustrate the intermediate results of the 1st, 3rd and 10th iterations to interpolate the large road region in Fig. 7(a), along the left border. The points in cyan correspond to the cells in the seed stacks in the iteration. Fig. 9(d) illustrates the result after interpolation is conducted along left, right, top and bottom sides of the road region. As can be observed, most of the interpolated borders are continuous. The border regions are then smoothed and the internal holes are filled to achieve a continuous region as shown in Fig. 9(e). Fig. 9(f) presents the result after the interpolation process are also applied to category "longGrass" and "obstacle". The region for "longGrass" is not fully continuous even after the interpolation. One of the reasons is that due to irregular height of the long grass, the estimated coordinates of the four corners of unclassified cells covered by long grass are not accurate enough to get the correct projection to image pixels. The other reason is that there is no clear-cut boundary between "longGrass" regions and " "highVeg" regions, as shown in Fig. 9(a). Likewise, there are gaps between the road boundary and "longGrass" region. However, the continuous road region and the approximation of borders between "road" and other categories form a good 2 1/2 D map for navigation purpose.

4 Summary

In this paper, we have introduced the scene parsing mechanism in our Cognitive Architecture, which has successfully integrated a visual attention mechanism with the super-pixel level scene parsing mechanism. We have also proposed a novel way to explicitly interpret spatial relations between objects and applied them to the top-down facilitation process of the scene part parsing mechanism.The scene part parsing and visual attention mechanisms have been tested in many experiments and trials and proved to be effective. In the meantime, we have also proposed a new approach to acquire a continuous $2\frac{1}{2}D$ map of traversable regions via fusion of image and LIDAR detections. This algorithm is under test involving autonomous vehicle navigation in off-road environment.

References

1. Goodale, M. A. and Milner, A. D.: Separate visual Pathways for Perception and Action. Trends Neoroscience **15(1)** (1992) 20-25.
2. Ng, G. W.: *Brain-Mind Machinery*. World Scientific (2009).
3. Lowe, D. G.: Object recognition from local scale-invariant features. Proceedings of the International Conference on Computer Vision (1999) 1150-1157.
4. Dalal, N., Triggs, B.: Histograms of oriented Gradients for Human Detection. CVPR (2005).
5. Riesenhuber, M. and Poggio, T.: Hierarchical models of object recognition in Cortex. Nature Neuroscience(1999) 1019-1025.
6. Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., Poggio, T.: Robust object recognition with Cortex-like mechanisms. IEEE Transactions on Pattern Analysis and Machine Intelligence **29(3)** (2007) 411-426.

7. Felzenszwalb, P., McAllester,D., Ramanan, D.: A discriminatively trained multiscale deformable part model. CVPR (2008)
8. Viola, P., and Michael J. J.: Rapid object detection using a boosted cascade of simple features. CVPR (2001).
9. Felzenszwalb, P., Girshick, R. McAllester, D.: Cascade object detection with deformable part models. CVPR (2010).
10. Laxebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. CVPR (2006).
11. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. InternationalJournal of Computer Vision(2001), 42(3): 145-175.
12. Torralba, A., Murphy, K., P., Freeman, W. T., Rubin, M. A.: Context-based vision system for place and object recognition. ICCV (2003) 1023 - 1029.
13. Siagian, C., Itti, L.: Rapid biologically-inspired scene classication using features shared with visual attention. PAMI **29(2)** (2007) 300-312.
14. Renniger, L., Malik, J.: When is scene identification just texture recognition? Vision Research **44** (2004) 2301-2311.
15. Tighe, J., and Lazebnik. S.: SuperParsing: Scalable nonparametric image parsing with superpixels. ECCV (2010).
16. Li, L. J., Socher, R., Li, F. F.: Towards total scene understanding: classification, annotation and segmentation in an automatic framework. CVPR (2009).
17. Du, L., Ren, L., Dunson, D., B., Carin, L.: A Bayesian model for simultaneous image clustering, annotation and object segmentation. NIPS (2009).
18. Rabinovich, A., Vedaldi, A., Galleguillos, C.: Object in context. ICCV(2007).
19. Galleguillos, C., Belongie, S.: Context-based object categorization: a critical survey. Journal of Computer Vision and Image Understanding **114(6)** (2010) 712-722.
20. He, X., Zemel, R., Carreira-Perpindn, M. A.: Multiscale conditional random fields for image labelling. CVPR (2004) 695-702.
21. Kumar, S., Hebert, M.: A hierarchical field framework for unified context-based classification. ICCV (2005) 1284-1291.
22. Verbeek, J., Triggs, B.: Scene segmentation with conditional random fields learned from partially labeled images. NIPS (2008).
23. http://en.wikipedia.org/wiki/DARPA_Grand_Challenge
24. Vandapel, N., Huber, D., F., Kapuria, A., Hebert, M.: Natural terrain classification using three-dimensional Ladar data for ground robot mobility. Journal of Field Robotics **23(10)** (2006) 839-861.
25. Himmelsbach, M., Luettel, T., Wuensche, H., J.: Real-time object classification in 3D point clouds using point feature histograms. Proceedings of IEEE//RSJ International Conference on Intelligent Robots and Systems (2009) USA.
26. Thrun, S. et. al.: Stanley: the robot that won the DARPA grand challenge. Journal of Robotic Systems **23(9)**(2006) 661-692.
27. Rasmussen. C., A hybrid vision+Ladar rural road follower, Proceedings of the IEEE Conference on Robotics and Automation (2006) 156-161.
28. Manz, M., Himmelsbach, M., Luettel, T. Wuensche, H.: Detection and tracking of road networks in rural terrain by fusing vision and LIDAR. Proceedings IEEE//RSJ International Conf. on Intelligent Robots and Systems (2011) 4562-4568.
29. Ng, G. W., Xiao, X., Chan, R.Z., Tan, Y. S.: Scene understanding using DSO cognitive architecture. Proceedings of the 15th International Conference on Information Fusion (2012).

30. Zhao, G., Xiao, X., Yuan, J., Ng, G., W.: Fusion of 3D-LIDAR and camera data for scene parsing. Journal of Visual Communication and Image Representation **25(1)** (2013) 165-183.
31. Hochstein, S., Ahissar, M.: View from the top: hierarchies and reverse hierarchies in the visual system. Neuron **36** (2002) 791-804.
32. Bar, M.: A cortical mechanism for triggering top-down facilitation in visual object recognition. Journal of Cognitive Neuroscience **15(4)** (2003) 600-609.
33. Yao, J., Fidler, S., and Urtasun, R.: Describing the scene as a whole: joint object detection, scene classfication and semantic segmentation. CVPR (2012)
34. Kasther, S., Ungerleider, G.: Mechanisms of visual attention in the uuman cortex. Annual Rev. Neural Science **23** (2000) 315-341.
35. Felzenszwalb, P., Huttenlocker, D.: Efficient graph-Based imagesegmentation. IJCV **2** (2004).
36. http://www.robots.ox.ac.uk/ vgg/research/textclass/filters.html
37. http://www.mit.edu/ jmutch/fhlib
38. Ojala, T., Pietikainen, M., Maenpaa, T.: Multi-resolution gray-scaleand rotation invariant texture classification with local binary patterns.PAMI **24(7)** (2002).
39. Fenske, M., J., Aminoff, E., Gronau, N., Bar, M.: Top-down facilitation of visual object recognition: object-based and context-based contributions. Progress in Brain Research **155** (2006) 3-21.
40. Oliva, A., Torralba, A.: The role of context in object recognition. Trends in cognitive Sciences **11(2)** (2007).
41. Desai, C., Ramanan, D., Fowlkes C. C.: Discriminative models for multi-class object layout. IJCV **2** (2012).
42. Achanta, R., Hemami, S., Estrada, F., Susstrunk, S.: Frequency-tuned Salient Region Detection. CVPR (2009).
43. Rensink, R., A.: The dynamic representation of scenes. Visual Cognition **7(1/2/3)** (2000) 17-42.
44. Nister, D., Stewenius, H.: Linear time maximally stable extremal regions. Forsyth, D., Torr, P., Zisserman, A.(eds): *Part II: LNCS 5303*, ECCV (2008) 183-196.
45. Matas, J., Chum, O., Urban, M., Pajdla, T: Robust wide baseline stereo from maximally stable extremal regions. BMVC(2002).