Unsupervised Detection of Artificial Objects in Outdoor Environments

Luciano Spinello[1] and Roland Siegwart[2]

[1] ASL, Swiss Federal Institute of Technology Zurich
luciano.spinello@mavt.ethz.ch
[2] ASL, Swiss Federal Institute of Technology Zurich
roland.siegwart@mavt.ethz.ch

Summary. This paper presents a novel unsupervised sensor fusion method to detect artificial objects in outdoor environments. We define artificial objects present in outdoor environments using structure and appearance information: an artificial object is composed of several smooth surfaces with sufficiently extended area and distinctive colors with respect to the environment. Structure information is obtained extracting smooth linked surfaces from 3D range data. Appearance information is computed on image data through a color processing. The problem of fusing these two different kinds of information is addressed through the use of a Bayesian sensor fusion approach. A probability map is built and then clustered with a 2.5D unsupervised self organizing network. This method defines objects according to the distribution of the probability values in the fusion map. The resulting clusters are labeled with their mean probability value representing the confidence the detected regions have of being artificial objects. In order to show the validity of the proposed algorithm some experiments have been performed in real outdoor environments showing promising results.

1 Introduction

Scene understanding and object detection in outdoor environments play an important role in a number of mobile robotics problems, including outdoor robot navigation, object tracking and prediction.

Several methods are used in computer vision to achieve object detection. Classifier based methods are often successfully used: cascade of boosted classifiers [1] (or SVMs) as well as template based methods. Due to its insensitivity to illumination, range data is well suited for object detection but a dense range measurement is required for a reliable object detection [2]. In this paper, we propose an approach to address the problem of artificial object detection in outdoor environments using a method that relies on range and image data.

We define artificial objects present in outdoor environments using structure and appearance information: an artificial object is composed of several

smooth surfaces with sufficiently extended area and distinctive colors with respect to the environment. Smooth surfaces and distinctive colors are key characteristics of mostly every kind of human-made object. Thus, these features can be representative of the presence of artificial objects due to unstructured nature of outdoor environments. Range data processing is achieved to obtain structure information and image data processing is executed to obtain appearance information. Nevertheless it's possible to have cases in which structural information is more relevant than appearance information (or viceversa). It's therefore necessary to manage the fusion of the information in a probabilistic manner using a Bayesian modeling approach.

The technique presented here aims to segment the space in labeled regions of interest and it's conceived to be fast and deployable in field robots. This technique can be a valid hypothesis generator for further reasoning (i.e. object classification) in the salient 3D space.

The novelty of this article is to present a multiple cues approach to unsupervisely detect artificial objects in outdoor environment based on 3D smooth surface extraction and salient color blobs. Moreover, object segmentation is achieved through a fast self organizing network in 2.5D.

2 Structure information processing

Salient structure information for artificial objects in outdoor environments is defined by a metric on the smoothness and flatness of sampled 3D points. An outdoor scenario is an highly unstructured environment: natural objects are generally constituted by irregular 3D shapes with complex surface profiles (i.e. trees, leaves, rocks etc). Artificial human-made objects are instead usually symmetric and smooth surfaced because of common design/object production phases. It is therefore acceptable, for the aim of this paper, to consider surfaces with peculiar properties of smoothness and area as interesting geometrical cues of an artificial object.

In order to define regions of common smoothness a process to compute local tangent planes is needed. The fitted plane of the point p_i, in the k-neighborhood subset, is defined by the centroid o_i and the normal n_i computed using a PCA plane fitting approach.

The normals of the planes can be orientated only in the opposite direction of the laser ray, thus the sign of the resulting normal is therefore adjusted.

A further geometrical processing is needed to define common smoothness regions. The normal orientation of each centroid is compared with the others in the neighborhood. When the angle between two consecutive normals is smaller than the threshold α_{max} then the region is grown with that vector. A new region is created when it's not more possible to find unvisited neighbors.

The point cloud is segmented in smooth regions when every normal has been analyzed . A value h_i^l is stored for each region related to its inverse area measurement. $H^l = h_1^l \dots h_w^l$ defines the set of w smooth patches. A

penalty weight is assigned to each horizontal surface found proportional to its position in the Z axis. This procedure aims to give a low weight value to smooth ground surfaces.

Furthermore, an adjacency graph is built among the regions to explain the geometric topology. A graph exploration algorithm is thus applied to link contiguous smooth surfaces.

3 Appearance information processing

The purpose of camera image processing is to obtain the appearance information of artificial objects. In order to define salient color blobs in the picture a two-steps technique is applied: two axis color clustering and a color quantization.

Firstly, an image enhancement process is applied. Two morphological image operations are used to simplify color diversity: erosion and dilation. The color space selected for the image processing is the HSB (Hue, Saturation and Brightness) due the similarities of this color model to the way humans tend to perceive colors.

3.1 Unsupervised Color Clustering and Quantization

The purpose of any clustering method is to group entities on the basis of similarity of features. In our case we want to cluster in the color space the main color blob to detect the principal color palette present in the image. Furthermore we want to achieve an unsupervised color clustering.

The Self Organizing Network (SON or Self Organizing Map) is a neural network often used for unsupervised learning, but it can be also applied as an unsupervised clustering method [3]. It has the advantage of adaptively computing the number of clusters and it has also a low computational complexity. Due the color perception importance, only the data present on the plane Hue-Saturation is fed to the SON. Therefore the topology of the self organizing network is a 2D squared grid and the computational clustering performance is assured.

The main color cluster is detected calculating the biggest cluster area in the color plane. The remaining image color information, without the principal mean colors, is then HSB color quantized to discard the most common outdoor natural palette.

The coarse filtering of discarding *greens*, *browns*, and too dark colors is achieved using several not-admissible 3D color windows in the image color space. This process introduces some a-priori knowledge in the algorithm, but it is a reasonable assumption for natural environments. In order to take in account the mean luminosity present in the picture, B_{env} is computed as the mean environment brightness. Thus, it's possible to change the minimum value of the brightness for the coarse color filtering windows to $B_{min} = \frac{B_{env}}{2}$.

Therefore the palette quantization process is adaptive to the environment brightness changes.

The Sobel operator is applied on each channel of each color blob to compute an approximation of the gradient of the image intensity. Therefore, using the information theory, it's possible to compute the blob entropy value, considering each channel as an independent intensity image:

$$N_i = card(blob_i) \quad p = \frac{hist(blob_i)}{N_i}$$

$$h_i^c = [-\sum_{N_i} p\ log_2(p)]_H + [-\sum_{N_i} p\ log_2(p)]_S + [-\sum_{N_i} p\ log_2(p)]_B \tag{1}$$

$$h_z = \arg\max_{[1,w]}(h_i^c) \tag{2}$$

The set $H^c = (1 - \frac{h_1^c}{h_z}) \dots (1 - \frac{h_w^c}{h_z})$ defines the weighted set of w extracted color blobs. Image entropy is a quantity which is used to describe the amount of information richness present in the image blob. Following the principle of the artificial object appearance given, less entropy a color patch has, more its confidence value is high. A low energy patch is a more reasonable evidence of an artificial object as it encodes an almost uniform color patch.

4 Information Fusion

Let's resume the information retrieved using range and intensity data processing before explaining the information fusion. Structure and appearance characteristics of artificial objects are retrieved. Therefore a careful merging of these incomplete information it's necessary to handle the partial and noisy information.

Structure information is defined in $\Re^3$ space, appearance information lays in $\Re^2$ space. A dimension reduction is defined for the structure information space $\Re^3 \rightarrow \Re^2$ projecting the selected smooth surfaces onto the image plane[3]. Furthermore, the structure information that falls out of the boundaries of the viewing area is discarded.

3D range laser to omnidirectional camera rigid transformation is calibrated using a method based on [4].

Structure and appearance information are considered cues of same importance, thus the same confidence level of detection should be given in the fusion information process. The information fusion is addressed using a Bayesian modeling approach. The variables used to formalize the problem are:

- ϕ : it describes the existence of an artificial object
- θ_l : it encodes the structure information

[3] A previously calibrated camera-laser system is supposed.

- θ_c : it encodes the appearance information

Starting from the joint distribution and applying recursively the conjunction rule we obtain the decomposition:

$$P(\phi \wedge \theta_l \wedge \theta_c) = P(\phi)\ P(\theta_l|\phi)\ P(\theta_c|\phi) \tag{3}$$

In equation 3 the phenomenon ϕ is considered to be the main reason for the contingency of the structure and appearance information, thus knowing the cause ϕ of the readings the variables θ_l and θ_c are independent. In general, this hypothesis is not always satisfied, but it is often used in literature and it has the main advantage of considerably reducing the complexity of the computation.

The conditional probability that defines the information fusion is:

$$P(\phi|\theta_l \wedge \theta_c) = \frac{P(\phi)\ P(\theta_l|\phi)\ P(\theta_c|\phi)}{\sum_\phi (P(\phi)\ P(\theta_l|\phi)\ P(\theta_c|\phi))} \tag{4}$$

$$P(\phi = true) = p_s \tag{5}$$

If the method of artificial object detection is, for example, used as a part of a warning system, then false positive rates can be tolerated and p_s in equation 5 can be set to an high value. In contrast, if it's included as part of active vehicle control a more conservative choice is needed.

Equation 4 is evaluated for each point of structure and appearance information in the fusion plane. Thus $P(\theta_l|\phi)$ is defined for each point by its value h_i^l of the i-surface patch. $P(\theta_c|\phi)$ is defined for each point by its value h_i^c of the i-color blob.

The resulting probability map has to be clustered to segment objects. Local maxima in the fusion map are thus segmented with a 2.5D self organizing network.

The weights of the SON, during the learning phase, are updated according to the probability value of the fusion map. The proposed cluster algorithm, described in section 4.1, has the advantage to cluster in 2D space, maintaining a fast execution time, and it has also the advantage of shaping the clusters according to the amount of confidence present.

4.1 2.5D Self Organizing Network for clustering

The network is composed of $T = M \times N$ nodes connected each other with undirected edges [5] arranged in a circular or squared grid, with M rows and N columns. Every node is connected with four other nodes (excluding the nodes located on the network border) and it has two associated variables: its mean value $\mu_i(x_i, y_i)$ and the counter $c_i \in [0, D]$, where D represents the size of input data elements. The weight $e_{i,j}$ is stored for every arc that connects the node i with the node j. Without loss of generality, a circular topology

SON is described here that will be used in the experiments. The Cartesian coordinates of a node constituting the grid are computed iteratively:

$$n_{i_x} = (R + R_\Delta)cos(\alpha + \alpha_\Delta) \tag{6}$$

$$n_{i_y} = (R + R_\Delta)sin(\alpha + \alpha_\Delta) \tag{7}$$

where R is the current radius of the SON grid, R_Δ is the radius increment, α is the current angle, α_Δ is the angle increment. R_Δ and α_Δ are chosen beforehand.

During the initialization of the algorithm, every node is placed to obtain a regular circular grid and every c_i and $e_{i,j}$ are set to zero. The learning phase is processed every time an input data of the fusion plane is fed to the SON. Let's define

$$p(x, y, v)$$

as data point expressed by its Cartesian coordinates x, y in the fusion plane and its value v_i. The first step of the learning phase consists in the selection of the two network nodes closest to the input data. Only the Cartesian coordinates of the data input $p_{x,y}$ are considered in this process:

$$w_1 = \arg \min_{i \in [0,T]} \parallel p_{x,y} - \mu_i \parallel \tag{8}$$

$$w_2 = \arg \min_{i \in [0,T]/w_1} \parallel p_{x,y} - \mu_i \parallel \tag{9}$$

The update values for c_{w_1} and e_{w_1,w_2} are set:

$$e_{w_1,w_2} = e_{w_1,w_2} + 1 \tag{10}$$

$$c_{w_i} = 1 + c_{w_i} - (1 - p_v) \tag{11}$$

In equation 11 the counter dynamic c_{w_i} is changed according to the value present in the input data p_v. The mean of the closest node and its four neighbors is then modified:

$$\mu_{w_1} = \mu_{w_1} + \frac{e_w}{c_{w_1}}(p_{x,y} - \mu_{w_1}) \cdot p_v \tag{12}$$

$$\mu_i = \mu_i + \frac{e_n}{c_i}(p_{x,y} - \mu_i) \cdot p_v \quad \forall_i \in neigh(w_i) \tag{13}$$

The update equations 12, 13 and 11 are modified with respect to the theory proposed in [5] to keep in account the third coordinate that define the 2.5D space.

It's important to notice that the mean μ_i changes more if an high value v_i is present in the input data set. Viceversa, if p_v is a small value, μ_i reflects small changes. After the learning process phase, the cluster representation is then accomplished using a graph cutting approach: arcs having low weight values are cut, leaving connected components with high edge values[4].

The Self Organizing Network here proposed has several advantages with respect to classic clustering techniques:

[4] Full explaination of the cluster representation is given in [5]

- The maximum cardinality of detectable clusters is not defined by the user.
- The algorithm complexity of the algorithm is $T\ N_t$, where N_t is the number of input points
- The clusters shapes are influenced by the third value present in each point of the map and thus the resulting clusters will more easily segment local maxima.

4.2 Region labeling

A label and a mean probability value are finally assigned to each detected cluster. The fusion plane is now completely segmented in regions where, with certain confidence, artificial objects can be found.

5 Experimental results

We demonstrate the performance of the proposed algorithm using a data set acquired in a real outdoor environment to test the detection of artificial objects.

The platform used to acquire data is composed of a rotating laser rangefinder and an omnidirectional camera mounted on the rooftop of a Daimler-Chrysler Smart vehicle. The SICK laser rangefinder is mounted with its scan line in vertical position; the rotation arounbd the Z axis is given by a stepper motor. The revolution frequency of the laser is $0.5Hz$. In a complete rotation the laser spans $\phi = [-50°, 50°]$ and $\theta = [0°, 360°]$ with a vertical resolution of $1°$. The omnidirectional camera is composed by a standard firewire camera with an hyperbolic lens. The images captured have the resolution of $640x480px$. The data is retrieved in a static environment.

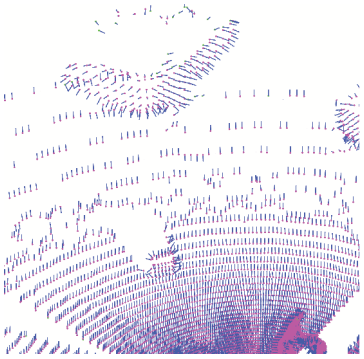

Fig. 1. Normals (in violet) and centroids (in blue) in the 3D laser point cloud scene.

The data retrieved from the 3D rotating laser is processed to compute local normals as presented in Fig. 1. Smooth linked surfaces are identified and are shown in Fig. 2. Note that the radial sectors missing in Fig. 1 and Fig. 2 are not caused by the algorithm of range data processing but they are caused by lack of data retrieved from the 3D laser. The data is then retrieved from

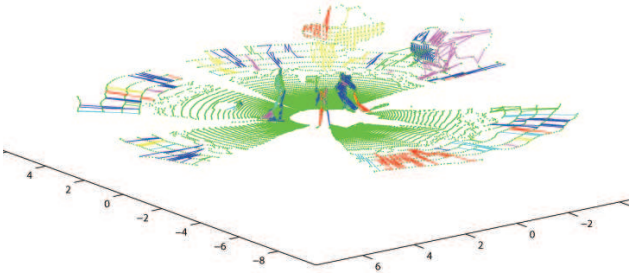

Fig. 2. The surfaces extracted from the 3D range data set are depicted in color.

the omnidirectional camera. In order to reduce computation time, a circular area of interest is set; the minimum bounding circle is set beforehand and represents the minimum radius of useful field of view. The most far away point from laser range data set is taken, then its 2D correspondence in the image plane is computed and the maximum radius for the image area of interest is set. The removal of the main color region is presented in Fig. 3.

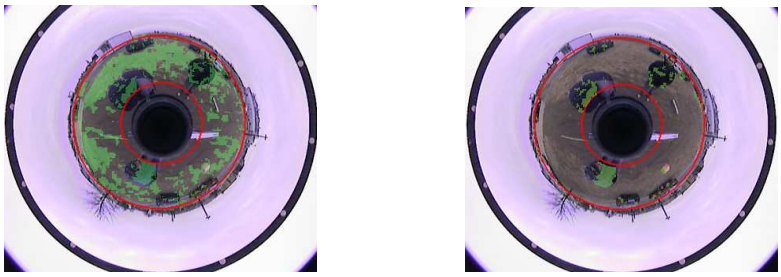

Fig. 3. Omnidirectional camera ROI is defined by the two red circles. Results of color clustering are shown in the left figure, final color blobs in the right figure

Therefore the color quantization is executed and the final image processing results are shown in Fig. 3.

To optimize the performance, the 3D laser and image processing are computed in parallel threads. The 3D smooth surfaces from the laser data processing are projected onto the omnidirectional camera plane and the information fusion takes place Fig. 4. The 2.5D self organizing network used is a neural

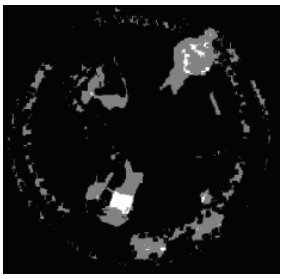

Fig. 4. Fusion plane is populated with structure information and appearance information probability.

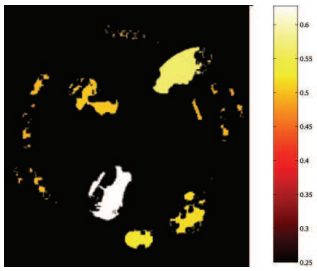

Fig. 5. Culsters expressing artificial objects are segmented from the fusion plane. The colorbar describes the clusters confidence.

network with circular topology in order to obtain a better data coverage in the fusion plane, which is the omnidirectional camera image plane. Therefore, the segmentation is performed. The labeled clusters are presented in Fig. 5. Three men, some cars, and some colored boxes are detected in the scene, every cluster with probability > 0.25 is shown in the 3D model figure presented in Fig. 6.

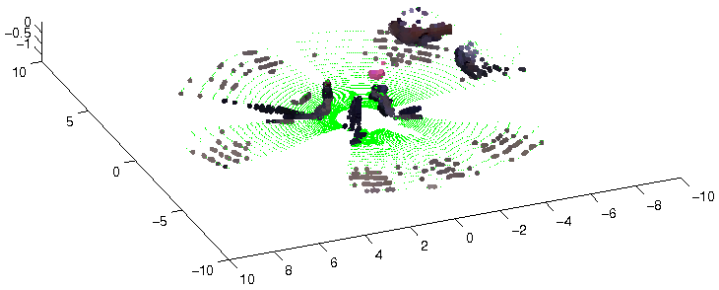

Fig. 6. 3D point cloud in which colored points represent detected artificial objects.

Even though the image processing involves a dynamic brightness technique, the algorithm still suffers from poor light conditions. In this case the resulting object detection has less probability and less spatial precision because the algorithm relies principally on range data processing. Similar results are obtained if reflective surfaces with low SNR are present in the laser data; in that case the algorithm relies more on appearance information. The distance also plays an important role in the detection phase. Distant surfaces have a low h_i^l due the big distance among far points with respect to the range

sensor position. In the experimental platform used the hyperbolic mirror of the omnidirectional camera gives a very small resolution of peripheral image, resulting a poor detection of far away objects.

The algorithm is completely implemented in C code, using ProBT lib [6] for Bayesian fusion, and it takes $1.2s$ to generate the results on a Intel Centrino 1.60 Ghz based system.

6 Conclusions and future works

This paper presented a novel unsupervised sensor fusion method to detect artificial objects in outdoor environments. We define artificial objects present in outdoor environments using structure and appearance information. The problem of fusing different kinds of information is addressed through the use of a Bayesian sensor fusion approach.

Further extensions of this work can be achieved maintaining the same probabilistic fusion method and adding other structure information (i.e. shape factors) and appearance information (i.e. texture analysis). Moreover the value p_s of equation 5 could be learned and tuned according to the type of outdoor environment in which the mobile platform is moving. Tracking of detected objects could also be used to improve the detection performance.

7 Acknowledgment

This work was conducted and funded within the EU Integrated Projects BACS (Bayesian Approach to Cognitive Systems) - FP6-IST-027140

References

1. P. Viola and Jones, "Robust real-time object detection," in *IEEE Workshop on Statistical and Theories of Computer Vision*, 2001.
2. J. Rodgers, D. Anguelov, H.-C. Pang, and D. Koller, "Object pose detection in range scan data," in *Int. Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.
3. J. Vesanto and E. Alhoniemi, "Clustering of the self-organizing map," *IEEE Transactions on Neural Networks*, May 2000.
4. D. Scaramuzza and A. Martinelli, "Flexible technique for accurate omnidirectional camera calibration and structure from motion," in *Fourth IEEE International Conference on Computer Vision Systems (ICVS 2006), Oxford, England*, 2006.
5. A. D. Vasquez and T. Fraichard, "A novel self organizing network to perform fast moving object extraction from video streams," *In Proceedings of the IEEE/RSJ Intenational Conference on Intelligent Robots and Systems (IROS), Beijing, China*, 2006.
6. Probayes, "Probt library," http://www.probayes.com/.