ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

Neural network based 2D/3D fusion for robotic object recognition

Louis-Charles Caron, Yang Song, David Filliat and Alexander Gepperth

ENSTA ParisTech - UIIS Lab
828 Blvd des Maréchaux, 91762 Palaiseau - France

Abstract. We present a neural network based fusion approach for real-time robotic object recognition which integrates 2D and 3D descriptors in a flexible way. The presented recognition architecture is coupled to a real-time segmentation step based on 3D data, since a focus of our investigations is real-world operation on a mobile robot. As recognition must operate on imperfect segmentation results, we conduct tests of recognition performance using complex everyday objects in order to quantify the overall gain of performing 2D/3D fusion, and to discover where it is particularly useful. We find that the fusion approach is most powerful when generalization is required, for example to significant viewpoint changes and a large number of object categories, and that a perfect segmentation is apparently not a necessary prerequisite for successful discrimination.

1 Introduction

From the point of view of a mobile robot, the world is a confusing place bombarding the robot with a huge amount of information avery second. Only after an appropriate decomposition or *segmentation* of scenes and the *categorization* of detected obstacles into a possibly large number of classes can any meaningful "intelligent" behavior take place, while it remains crucial to meet real-time execution constraints on usually limited hardware. Here, we describe a computationally efficient robotic segmentation/recognition architecture as visualized in Fig. 1 and study how the fusion of several information sources from 2D (structure and color) and 3D (holistic object shape) sensors improves categorization under realistic conditions, which means in particular working with less-than-perfect object hypotheses coming from a computational segmentation step. As the number of classes, N, is large, we choose a neural network-based fusion approach (which has practical advantages compared to training N support vector machines) and validate it on a publicly available benchmark database[1].

1.1 Related work

The most commonly used 2D object descriptors are SIFT[1] and SURF[2] which offer convenient invariance properties at a reasonable (SIFT) or optimized (SURF) computational cost. There is a large number of other approaches which we will not review here as our focus is on the fusion of multimodal information. Object

[1] `www.gepperth.net/alexander/downloads/2014_fusion.tar.gz`

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

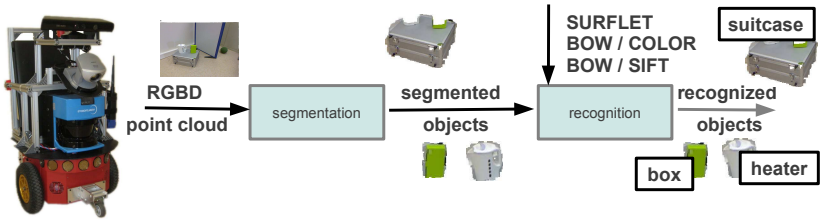

Fig. 1: Implemented processing architecture as a block diagram.

recognition from 3D data is equally well-studied in the literature, see [3] for surveys. Proposed methods can be grouped into holistic and local approaches. For the former group, an important technique is iterative closest point estimation (ICP)[4] where point clouds are stored "as is" and matched to observations in an efficient way. If object classes are restricted in advance, descriptors based on the Generalized Hough Transform can be useful[5] as well. In addition, histogram-based descriptors are very common, which may be based on surface normal histograms [6] or other global object features[7, 8]. In contrast to this, local approaches , e.g., [9, 10] are similar to SIFT/SURF for RGB images, trying to find a small number of distinctive "key points" in a point cloud whose associated descriptors can be matched against stored templates.

1.2 Article structure and key questions

We start out with a description of segmentation, feature computation and fusion methods in Sec. 2, then goe on to describe the used database and its creation in Sec. 2.3, and finally present and discuss experimental results in Secs. 3, 4. The key questions we are seeking to answer are, firstly, whether there is an **efficient 3D-based segmentation** step that avoid RANSAC. Secondly, we wish to investigate which are the **concrete benefits of multisensory fusion** by comparing the individual and combined performance of all available modalities.

2 Methods

2.1 Segmentation using 3D data

For the segmentation, only the distance information from the RGBD point clouds is used. All processing is implemented using the free Point Cloud Library library (PCL, [11]). The 3D point cloud is first filtered to remove distant, noisy points and the surface normals to each point [12] are calculated.

The first major computational step is the removal of the ground plane. For this step, the Kinect is assumed to be at a known distance and orientation with regard to the ground plane, as happens when it is mounted on a wheeled robot. All points sufficiently close to the theoretical floor plane, and having a normal approximately perpendicular to the plane, are recognized as part of the ground and removed from the cloud. The exact floor plane coefficients for this point cloud are calculated from these points. This method robustly detects the floor even if the robot accelerates brusquely, bumps into obstacles or faces a slightly inclined floor sections.

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

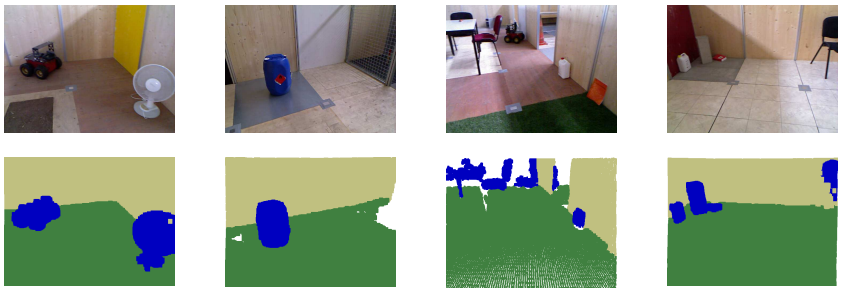

Fig. 2: Examples of segmentation results. Upper row: original images. Lower row: point cloud segmentation results back-projected into the image plane. Color code: green = floor, beige = wall, blue = object, white = ignored.

The remaining points are projected onto the ground plane and grouped according to their (projected) distance. For this purpose, we use a region-growing type of algorithm to form groups of 3D points that may correspond to objects. Fig. 2 shows images of scenes taken with the Kinect along with the color-coded masks for floor, wall and objects. Segmentation results do not always correspond to objects: in particular, objects that are too close together, or which are placed on top of each other are often segmented as a single object.

2.2 Fusion approach for recognition

We use texture, color and shape feature vectors as input to a multi-class neural network classifier and compare the results when using these vectors either separately or jointly (by simple concatenation). The neural network has a topology of $K - 50 - N$ (where K, N denote the dimensionality of input and the number of classes, respectively) using sigmoid activation functions for the hidden and a softmax one for the output layer, with bias units in all layers, and is implemented using the PyBrain library[13]. For obtaining a category decision, we find the output neuron with the highest activity which exceeds a rejection threshold. If no such neuron exists, we *reject* the input and take no decision.

2.2.1 Bag of words (SIFT)

Texture description is performed using a bag of words (BOW) approach [14] applied to visual recognition. SIFT features are computed from each image of the learning database to extract a visual description of all objects. These descriptors are then clustered using k-means, each found cluster center forming a *word* in the *dictionary*. To describe an unknown object, the SIFT descriptors extracted from its RGB image mask are matched to the dictionary, and a histogram of occurrence of the dictionary descriptors is built. The resulting feature vector has the size of the dictionary, which is 100.

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

Fig. 3: Example image of the objects used in the experiments.

2.2.2 Bag of words (color)

Due the value of color information in indoor scenes, we implemented an additional color-based descriptor. The processing is similar to what is described in Sec. 2.2.1, but this time the features are local hue histograms that are computed on regularly spaced windows as detailed in [15]. The resulting histogram has again dimensionality 100.

2.2.3 Surflet histograms

Following [6], we obtain the descriptor of a point cloud as follows: we randomly choose a set $P = \{(p_i, p_j)\}_{i \neq j}$ containing N point pairs and calculate the angles between the points' normal vectors, as well as the angles between the normal vectors and the difference vector between the points. Additionally, we compute the distance between the points and normalize it by the largest distance between any two point pairs sampled in this point cloud. These four numbers are discretized into 5 bins, and used to compute a histogram of 625 entries.

2.3 Database

Using a Kinect mounted on a pioneer robot, a custom database of 53 common objects was made. The objects were shot from 6 different angles of view, taking one hundred point clouds per angle while the robot moved back and forth in front of the objects, at a distance of 1 to 4 meters. This database is segmented using the algorithm described in Sec.2.1 to get cropped images and point clouds of the objects. As the data is taken by an moving robot, overall data quality is mediocre as some objects are badly segmented. Figure 3 shows examples of the

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

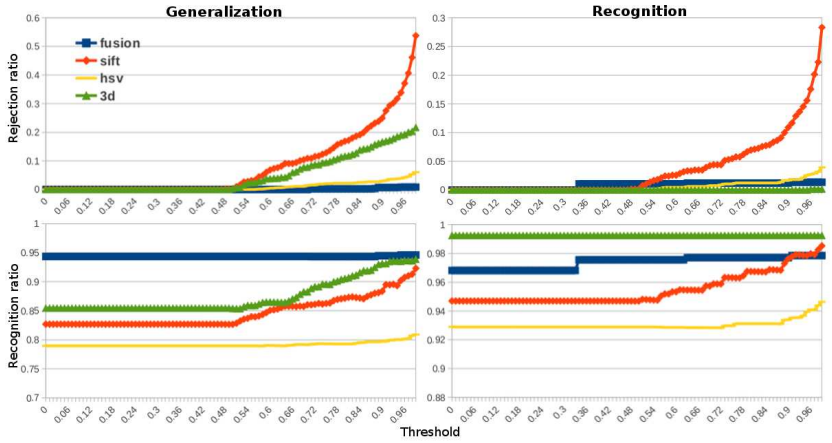

Fig. 4: Results of the 3-classes experiments. Upper row: rejection ratio (rejected good classifications / good classifications). The rejection ratio is 1 for a threshold value of 1. Lower row: recognition ratio (accepted good classifications / accepted classifications). The recognition ratio is undefined for a threshold of 1.

objects in the database,showing also some common segmentation errors. The red chair's mask is misaligned, probably because the robot was turning when the point cloud was captured, and half of the right-hand side of the computer is missing because its surface is too reflective for the Kinect to perceive it.

3 Experiments

Four types of experiments were conducted to demonstrate the capabilities of the proposed algorithm and the benefit of the 2D/3D fusion. In the recognition experiments, 0° and 60° angles of view are used to train and test the algorithm. In the generalization experiments, 0° and 60° angles of view are used for training and angles 120° and 300° for testing. In a first series of experiments, the 7 objects from the 3 classes cabinet, chair and sofa are used. In a second series, objects from all 7 classes shown in figure 3 are used. In all experiments, half of the data was used for training, and the other half for testing. Figure 4 shows the rejection ratio (rejected good classifications / good classifications) and recognition ratio (accepted good classifications / accepted classifications) for the first series of experiments when varying the threshold on decision making. Comparable results are obtained for the second series.

4 Discussion and Outlook

Several facts are revealed by the experiments. Firstly, the 3D features alone are excellent at distinguishing objects. For the recognition experiments, this technique always yielded better results than the 2D/3D fusion. The fact that this technique is oblivious to the color of the objects is however a severe drawback.

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

Secondly, the 3D recognition's maximum rejection ratio increases from 0.2 to 0.55 with the second series' larger number of objects. The performance of the fusion algorithm is more stable, and the rejection ratio remains low even with an increased number of object classes (always below 0.15). Lastly, the fusion algorithm produces much better results than the 3D alone when in the generalization setting. This is the most valuable asset of the fusion, as the algorithm is intended to be used online in a exploration experiment for semantic mapping [16]. In an online experiment, partial views, unseen angles, and occlusions will be commonplace and the generalization capabilities of the 2D/3D fusion will be important.

References

[1] D Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60, 2004.

[2] H Bay, T Tuytelaars, and L J Van Gool. Surf: Speeded up robust features. In *ECCV*, 2006.

[3] R. J. Campbell and P. J. Flynn. A survey of free-form object representation and recognition techniques. *Computer Vision and Image Understanding*, 81(2):166–210, 2001.

[4] Z Zhang. Iterative point matching for registration of free-form curves and surfaces. *Int. J. Comp. Vis*, 7(3):119–152, 1994.

[5] T Rabbani and F Van Den Heuvel. Efficient hough transform for automatic detection of cylinders in point clouds. In *Proceedings of the 11th Annual Conference of the Advanced School for Computing and Imaging*, volume 3, pages 60–65, 2004.

[6] E. Wahl, U. Hillenbrand, and G. Hirzinger. Surflet-pair-relation histograms: a statistical 3D-shape representation for rapid classification. In *Proceedings of the Fourth International Conference on 3-D Digital Imaging and Modeling (3DIM)*, 2010.

[7] W. Wohlkinger and M. Vincze. Ensemble of shape functions for 3d object classification. In *IEEE international Conference on Robotics and Biomimetics*. IEEE, 2011.

[8] I. K. Park, M. Germann, M. D. Breitenstein, and H. Pfister. Fast and automatic object pose estimation for range images on the GPU. *Machine Vision and Applications*, pages 1–18, 2009.

[9] Y Sun, J Paik, A Koschan, and MA Abidi. Point fingerprint: A new 3-D object representation scheme. *IEEE transaction on Systems, Man, and Cybernetics Part B: Cybernetics*, 33:712–717, 2003.

[10] R. B. Rusu, N. Blodow, and M. Beetz. Fast point feature histograms (FPFH) for 3D registration. 2009.

[11] RB Rusu and S Cousins. 3D is here: Point cloud library (PCL). 2011.

[12] K. Klasing, D. Althoff, D. Wollherr, and M. Buss. Comparison of surface normal estimation methods for range sensing applications. In *ICRA*, 2009.

[13] T Schaul, J Bayer, D Wierstra, Y Sun, M Felder, F Sehnke, T Rückstieß, and J Schmidhuber. PyBrain. *Journal of Machine Learning Research*, 11:743–746, 2010.

[14] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *ICCV*, volume 2, pages 1470–1477, October 2003.

[15] D. Filliat. A visual bag of words method for interactive qualitative localization and mapping. In *ICRA*, 2007.

[16] D Filliat, E Battesti, S Bazeille, G Duceux, A Gepperth, L Harrath, I Jebari, R Pereira, A Tapus, C Meyer, S Ieng, R Benosman, E Cizeron, J-C Mamanna, and B Pothier. Rgbd object recognition and visual texture classification for indoor semantic mapping. In *IEEE conference on Technologies for Practical Robot Aplications*, 2011.