Object-based Place Recognition for Mobile Robots Using Panoramas

Arturo RIBES [a,1], Arnau RAMISA [a] and Ramon LOPEZ DE MANTARAS [a] and Ricardo TOLEDO [b]

[a] *Artificial Intelligence Research Institute (IIIA-CSIC), Campus UAB, 08193 Bellaterra, Spain*

[b] *Computer Vision Center (CVC), Campus UAB, 08193 Bellaterra, Spain*

Abstract. Object recognition has been widely researched for several decades and in the recent years new methods capable of general object classification have appeared. However very few work has been done to adapt these methods to the challenges raised by mobile robotics. In this article we discuss the data sources (appearence information, temporal context, etc.) that such methods could use and we review several state of the art object recognition methods that build in one or more of these sources. Finally we run an object based robot localization experiment using an state of the art object recognition method and we show that good results are obtained even with a naïve place descriptor.

Keywords. Object recognition, Localization, Mobile robots

Introduction

Object recognition[2] is one of the most promising topics of Computer Vision. Having the ability to learn and detect hundreds of arbitrary objects in images from uncontrolled environments would be a major breakthrough for many artificial intelligence applications. One of the areas of research that would benefit the most of that innovation is intelligent robotics. The capacity to perceive and understand the environment is an important limitation for designing robots suitable for being deployed in houses to perform in domestic tasks or help the disabled. Also, being able to communicate at human-level semantics with its owners would make this robots much easier to control for a non-technical end user. In this work we review some of the most relevant state-of-the-art object detection and classification methods. Also, we discuss which sources of information can be exploited in order to have a fast and robust object recognition method, and which of the existing techniques could be applied in such domain. An area of research that recently started to profit from object recognition capabilities is topological localization. In [16] an object-based robot navigation and place categorization method is proposed. To illustrate our interest for object recognition methods for topological localization, we perform

[1] Corresponding Author: Arturo Ribes, Artificial Intelligence Research Institute (IIIA-CSIC), Campus UAB, 08193 Bellaterra, Spain; E-mail: aribes@iiia.csic.es.

[2] Object recognition is used loosely to embrace the concepts of detection, identification and eventually classification.

an object-based localization experiment using a publicly available dataset of stitched panoramic images acquired with a mobile robot. For this experiment, a custom implementation of the Lowe object recognition algorithm [3] has been used to learn and detect the instances of forty different objects in the panoramic images.

The paper is divided as follows. In Section 1 the state of the art in visual object recognition is reviewed, paying special attention to those methods and strategies that adapt well to the domain of mobile robotics. In Section 2 the Lowe object recogniton algorithm is explained. In Section 3 the experiments performed are described and the results presented. Finally, in Section 4 the conclusions are drawn.

1. State of the Art

Recently significant work has been done in visual object **classification**, with many methods making analogies to document retrieval literature. Visual vocabularies have been extensively used to relate local feature regions to visual words by clustering its descriptors with algorithms like k-means [14] or hierarchical k-means [8], among others. This reduces its dimensionality to a fixed number, so visual-word statistics can be used to classify the input image. Local features are widely used in computer vision due to its relation with research in primate vision. Those features are regions in an image providing relevant information used as building blocks of the models that caracterize an object. This bottom-up approach to vision enables the models to be stable against background clutter and occlusions, while properties of local descriptors account for illumination and viewpoint invariance. Feature sampling strategies commonly used are based in interest region detectors, which finds regions in a well-founded mathematical way that is stable to local and global perturbations. Using this techniques texture-rich regions can be detected at repeatable positions and scales.

In [14] the authors use k-means to cluster local feature descriptors into a vocabulary of visual words and the TF-IDF (Term Frequency - Inverse Document Frequency) scheme to priorize distinctive feature descriptors. Local features are detected in an image with an affine-invariant adaptation of Harris corner detector [5] and MSER detector [4] and a histogram is built from visual-word counts. Experiments are done in scene matching - where query descriptor vectors are compared to the ones in database - and object retrieval throughout a movie, where a user-specified query region is used to re-rank the results using spatial consistency, accomplished by matching the region visual-words to the retrieved frames. Similarly [8] uses the same TF-IDF scoring scheme but this time the vocabulary is constructed with a hierarchical k-means applied to MSER features extracted from the set of training images. A tree with a branch factor of k, where at each child node k-means is applied to the contents of the parent node. The step is repeated until a desired depth is reached. This allows an efficient lookup of visual words in logarithmic time, enabling the use of larger vocabularies. The results show how well the system scales to large databases of images in terms of search speed, and the larger vocabulary describes much better the images so, in contrast with [14], geometric information is not really needed to obtain good performance. The work of [9] proposes combining multiple feature types in a single model making use of boosting to form each classifier. This enables each category to choose the best features, regardless of its type, that describes it. Local features used are both discontinuity and homogeneity regions. The former are appearance features detected and described with various methods that capture local in-

tensity changes and the later are regions extracted with wavelets or segmentation that contain repetitive textures or stable intensity distributions respectively. Results show that different feature types perform much better in some classes than others, so combining them in a single model greatly improves classification results.

Although good results have been obtained using *bag of words* type methods, when the object does not occupy the majority of the image or detection and pose estimation is necessary, information on the **relative position between object features** is essential. This is usually a requirement when dealing with real world images, and several methods have been developed in order to take advantage of this positional information. Leibe et al. present in [2] the Implicit Shape Model, an object detection method that combines detection and segmentation in a probabilistic framework. First a codebook of local patches around Harris corner points is built using agglomerative clustering For every resulting cluster center, the Normalized Grey-scale Correlation distance is computed to all the training patches and, for those that are over a certain threshold, the relative object center position is stored. For detection a 2D Generalized Hough Transform (with continuous space to avoid discretization) is used to cluster probabilistic votes for object centers. Finally, the image is segmented assigning to each pixel the most probable object class with respect to the matched patches that include it. A Minimal Description Length procedure is used to reject overlaping hypotheses based on segmentation results. Tests have been done with the UIUC car database and with the individual frames of a walking cows video database, reporting very good performance. However, results show that the method only is able to deal with very small scale changes (10% to 15%). To avoid this, authors suggest using scale-invariant interest point detectors or rescaled versions of the codebook.

In [10], Opelt et al. present the Boundary-Fragment Model (BFM). This strategy is similar to the one of [2], but instead of local patches, it uses boundary fragments. A codebook of fragments for detecting a particular class is built by first computing Canny edges of the training images and finding edge segments that match well in a validation set (and bad in the negative examples set) using an optimization procedure. Next the candidate edge segments are clustered using agglomerative clustering on medoids to reduce redundacy, storing also the centroid of the object. Groups of two or three segments that estimate well the centroid of the object are used as weak detectors in a boosting framework. A strong detector is trained selecting weak detectors with good centroid prediction capabilities in positive images and that do not fire in negative images. For detection, weak detectors are matched to image Canny edges, with each one voting for one or various centroid positions in a 2D Hough voting space. Votes are accumulated on a circular window around candidate points, taking those above a threshold as object instances. Finally approximate segmentation can be obtained backprojecting the segments that voted for a centroid back into the image. In order to make the method robust to scale and in-plane rotation different scaled and rotated of the codebook are used simultaneously. The authors extended this method for multi-class object detection in [11] using a shared codebook. The method can be trained both jointly or incrementally. In [17], authors propose a shape-based method for object detection. The method builds on the ISM and uses k-Adjacent Segments with k=3 (TAS) as shape-based feature detector. A codebook of TAS is generated by first building a full connected graph of all the TAS in the training set with distance in the edges between every pair of TAS and then applying Normalized Graph Cuts to obtain the cluster centers. Similarly to the ISM model, probabilistic votes are casted in a 2D Hough Transform, and Parzen Windows are used

to find the most plausible object centers. Finally Gradient Vector Flow is used to find an approximate segmentation of the objects. The proposed method is interesting and has a similar performance to BFM in cows and motorbikes dataset used in [10] although using simpler features, specially in the case of few training images and small codebooks (200 features).

In [3] the author proposes a single-view object recognition method along with the well known SIFT features. First SIFT features of the training images are compared with Euclidean distance to the set of descriptors from the test image using a K-D tree and the Best Bin First algorithm to speed up the matching process. Matches in which the distance ratio between first and second nearest neighbors is greater than 0.8 are rejected. This eliminates 90% of the false matches while discarding less than 5% of the correct matches. Matches remaining after this pruning stage are clustered using its geometrical information in a 4D Generalized Hough Transformation with broad bin sizes, which gives the initial set of object hypotheses. To avoid the problem of boundary effects in bin assignment, each keypoint match votes for the 2 closest bins in each dimension. Bins of the Hough Transform containing more than three matches constitute an object location hypothesis. Finally a least-squares method (IRLS) is used to estimate the affine transformation of the detected object.

Context has been argued to provide very relevant information in computer vision. In [9], weak classifiers are selected with a boosting method to form the final classifiers; it was found in the experimental results that a high percentage of weak classifiers with higher weights selected local features corresponding to background. This can be interpreted as that context in some classes is more discriminative than foreground parts, although it greatly depends on the training set used. Also, context can be introduced in an ontology-based manner between objects as it is done in [12], where object co-ocurrence probability is computed by extracting related terms to each class from Google Sets or directly computing it from the ground truth. The system first segments query images into regions, which are in turn classified in a Bag-of-Words fashion and results obtained are re-ranked based in object co-ocurrences. In [15] context is provided using global features. These are computed from local features extracted with a collection of steerable filters applied over the image, then taking the average magnitude of the responses over a coarse spatial grid in order to obtain a global descriptor of 384 dimensions, which is finally projected to its 80 PCs. Knowing that video frames captured in the same place have a strong correlation, the authors use a Hidden Markov Model to compute the probabilities of being in a specific place given global features from some frames ago. Place recognition results are also proven to provide a strong prior for object detection.

We are aware of few works that consider the case of general object recognition in the domain of **mobile robots**. In [1] the authors integrate object recognition and SLAM. The proposed object recognition method works with few training images, and consists of two stages (Hypotheses generation and verification) that use active vision to focus on interesting parts of the image. The first stage uses Receptive Field Coocurrence Histograms (RFCH) and, in the verification stage, an area of interest that contains a maximal number of hypotheses is zoomed in using the optical zoom of the camera. Again RFCH are used to verify each hypothesis and finally SIFT features are extracted and matched to the model image. No further geometrical constraints are aplied to discard outliers, and if an object dependent number of SIFT features are matched the object is considered detected. Background substraction is used during the training stage to precisely segment

the objects. In [7] the authors present a system that learns and recognizes objects in a semi-autonomous fashion. Their system acquires learning samples of new objects by analyzing a video sequence of a teacher showing the object in a variety of poses and scales. Then it automatically segments and builds an internal representation with a shared vocabulary of visual features. The vocabulary is made using k-means over features based on Gabor-jets and hue-saturation values extracted at Harris corner points. Each visual word stores the relative position of the objects where it appears. Recognition is done by accumulating votes on a Hough Transform in a similar way as in [2]. In the experiments, the system runs at about 1-2Hz, recognizing objects from a 100 object database with 78% of mean accuracy Additionally, the database included difficult objects, like untextured, flexible and very similar objects (wires, cans, water bottles). Pose and scale invariance is obtained by superimposing multiple learning samples, so it is only invariant in the range that the object was learned.

2. Object Recognition in Panoramas

The method used for object recognition is the one proposed by Lowe [3] with some improvements (Figure 1 shows an outline of the method) in order to obtain better results given that we use panoramic images from our mobile robot (see Figure 2). We use one training image per object in the library, storing its descriptors along with their object ID in a descriptor database. All training images are taken from a 1Mpx digital camera. Testing images are panoramas built from 36 images acquired with the pan-tilt camera mounted on top of our mobile robot, storing the set of SIFT descriptors extracted from image regions used to construct the panorama. The position of each descriptor is relative to the total panorama size, not to the individual images. The panoramic image is used only to show recognition results, as the internal representation used is the descriptor set. Our implementation of the Lowe SIFT object recognition method differs from the original in the keypoint matching method, where instead of one single K-D tree and the BBF algorithm we use the improved method proposed in [6]. This method gives an improvement in search time of two orders of magnitude with respect to linear search and retrieves 90% of true nearest neighbors. Hough clustering stage is very tolerant to outliers, but only provides a similarity transform as a starting point for the more accurate affine transform estimation. Given that IRLS is very sensitive to outliers, we evaluate the impact of adding an intermediate RANSAC (RANdom SAmple Consensus) stage that provides a good initialization of affine parameters and is also tolerant to outliers. Furthermore, given our mobile robot setup, we can discard hypotheses with some heuristics applied to the affine solution bounding box related to the panorama size.

- A box that spans more than 90° of field-of-view (25% the width of panorama) is rejected: this would correspond to an object being in touch with the robot, thus not recognizable by its huge scale and viewing conditions.
- A box that is not at least 10% inside of view is rejected because there cannot be sufficient distinguishable features to estimate the affine transform between the object and panorama.
- A box in which the ratio of minor to major axis is less than 0.05 is rejected, corresponding to objects in viewing conditions that are proven not to be recognizable with this method.

Although this heuristics come from human knowledge of the method limitations or the robotic setup used, they can be supported with detector experiments or inferred from localization results and incorporated in a probabilistic model, which is left for further research. We use this method to build a signature for a panorama consisting in the presence or absence of database objects.

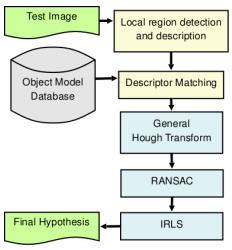

Figure 1. Block diagram of the Lowe object recognition method

Figure 2. Mobile robot platform used for the experiments.

3. Experimental Results

In this section we explain the experimental setup used for this paper and the results obtained in tests. Training images are close-ups of forty objects appearing in some panoramas, taken with a digital camera at a resolution of 1 megapixel. There is one training image per object, corresponding to its frontal view, except in some cases where the object can be found in two different views, i.e. fire extinguisher. This later case is treated as having two different objects. Descriptor vectors are extracted from training images and stored as a pre-processing step. The testing dataset consists of five sequences of 11-22 panoramas taken with our mobile robot in our facilities. Panoramic images have a size about 6000x500 pixels. Figure 3 shows panoramas used to test our approach with the detected objects. Four experiments are done in order to test the importance of using

(a) Panorama from *lab02* sequence

(b) Panorama from *actes03* sequence

Figure 3. Final results obtained by our method in random panoramas of sequences. Only hypotheses consistent with ground truth are shown. Ground truth is shown in red.

RANSAC as hypothesis outlier filtering stage prior to IRLS, and also to test the performance loss and speedup of using aproximate nearest neighbour over exhaustive search in

	ACC	PRE	REC
RANSAC, no ANN	0,81	0,37	0,53
no RANSAC, no ANN	0,81	0,37	0,53
RANSAC, ANN	0,81	0,36	0,53
no RANSAC, ANN	0,81	0,35	0,52

Table 1. Object recognition results.

Sequence	Hit	Miss	Rate (%)
lab02	10	1	91
actes03	13	1	93
actes04	11	3	79
lab08	18	1	95
passadis04	19	3	86

Table 2. Place recognition results.

matching stage. We also provide timing results of matching, Hough clustering and affine estimation stages. As can be observed in Table 1, RANSAC provides no improvement over precision or recall. It provides better affine parameters initialization than the similarity transform obtained from Hough clustering step, so IRLS converges quicker than if RANSAC is not used. Although it has to be noted that neither performance nor speed are altered in a significant way, usually RANSAC filters more outliers than IRLS, resulting in more accurate pose estimation. The processing time for each panorama, ignoring panorama acquisition time and feature extraction step, is about 1 second for matching using FLANN library set at 90% of retrieval precision, 5 ms for Hough Transform and 1-2 ms for the affine stage (RANSAC+IRLS or IRLS alone), so the processing can go near real-time. Tests have been done in a 3 Ghz Pentium IV with 1 Gb of RAM. In order to test the applicability of object recognition in place classification task we set up the following test. Mean vectors are computed for each panorama sequence, except in the sequence we are testing, where one panorama is used to test and the rest are used to compute the mean vector, resulting in a leave-on-out evaluation strategy and a minimum distance classifier. The results of this test are shown in Table 2. As can be seen, localization rate is good despite that object recognition only recalls 50% of objects and there is a significant amount of false positives.

4. Conclusions

In this work we have reviewed the different data sources that can be exploited for the object recognition task in the context of mobile robotics. Also, we tested a state-of-the-art method in images taken with a mobile robot to show one of the applications of object recognition in the robotics field. Although we have used a simple classification method for place recognition, it has produced good results despite the modest performance of the object recognition method. Furthermore, object recognition as a middle stage between image data and place recognition helps to reduce the complexity of the place descriptor while increasing its robustness. It also can help to reduce irrelevant information and noise from the background that could confuse other approaches to localization. For example in [13] the authors report false matches between individual features to be the most significant problem of the method. In this approach we only take into account *natural landmarks* formed with groups of localy coherent features, filtering potential false matches arising from spurious features. A possible extension to our approach would be incorporating position information to the detected objects along with a degree of mobility (for example a window would have mobility zero, while a chair would have a higher value). It has to be noted that not all object recognition methods are suitable for mobile robotics environment, as robots need to take decisions in a relatively short time and imag-

ing conditions and hardware used to take the pictures decrease the quality of input data. In future work we plan to review the main state-of-the-art object recognition methods to find insight to develop a robust object recognition method that can be applied in mobile robotics.

Acknowledgements

This work has been partially funded by the FI grant and the BE grant from the AGAUR, the 2005/SGR/00093 project, supported by the Generalitat de Catalunya , the MIDCBR project grant TIN 200615140C0301, TIN 200615308C0202 and FEDER funds. The authors would also like to thank Marius Muja and David Lowe for making available the FLANN library.

References

[1] S. Ekvall, P. Jensfelt, and D. Kragic. Integrating Active Mobile Robot Object Recognition and SLAM in Natural Environments. *Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on*, pages 5792–5797, 2006.

[2] B. Leibe, A. Leonardis, and B. Schiele. Combined object categorization and segmentation with an implicit shape model. *Workshop on Statistical Learning in Computer Vision, ECCV*, pages 17–32, 2004.

[3] D.G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[4] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing*, 22(10):761–767, 2004.

[5] K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. *Proc. ECCV*, 1:128–142, 2002.

[6] Marius Muja and D.G. Lowe. Fast Approximate Nearest Neighbors with Automatic Algorithm Configuration. *Preprint*, 2008.

[7] E. Murphy-Chutorian, S. Aboutalib, and J. Triesch. Analysis of a biologically-inspired system for real-time object recognition. *Cognitive Science Online*, 3(2):1–14, 2005.

[8] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. *Proc. CVPR*, pages 2161–2168, 2006.

[9] A. Opelt, A. Pinz, M. Fussenegger, and P. Auer. Generic object recognition with boosting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(3):416–431, 2006.

[10] A. Opelt, A. Pinz, and A. Zisserman. A boundary-fragment-model for object detection. *Proc. ECCV*, 2:575–588, 2006.

[11] A. Opelt, A. Pinz, and A. Zisserman. Incremental learning of object detectors using a visual shape alphabet. *Proc. CVPR*, 1(3-10):4, 2006.

[12] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in Context. *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8, 2007.

[13] A. Ramisa, A. Tapus, R. Lopez de Mantaras, and R. Toledo. Mobile Robot Localization using Panoramic Vision and Combination of Local Feature Region Detectors. *Proc. IEEE International Conference on Robotics and Automation, Pasadena, California*, pages 538–543, 2008.

[14] J. Sivic and A. Zisserman. Video Google: a text retrieval approach to object matching in videos. *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 1470–1477, 2003.

[15] A. Torralba, KP Murphy, WT Freeman, and MA Rubin. Context-based vision system for place and object recognition. *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 273–280, 2003.

[16] S. Vasudevan, S. Gächter, V. Nguyen, and R. Siegwart. Cognitive maps for mobile robots - an object based approach. *Robotics and Autonomous Systems*, 55(5):359–371, 2007.

[17] X. Yu, L. Yi, C. Fermuller, and D. Doermann. Object Detection Using A Shape Codebook. *Unpublished*.