Large-scale Real-time Object Identification Based on Analytic Features

Stephan Hasler, Heiko Wersing, Stephan Kirstein, and Edgar Körner

Honda Research Institute Europe GmbH
D-63073 Offenbach/Germany
stephan.hasler@honda-ri.de

Abstract. Inspired by biological findings, we present a system that is able to robustly identify a large number of pre-trained objects in real-time. In contrast to related work, we do not restrict the objects' pose to characteristic views but rotate them freely in hand in front of a cluttered background. We describe the essential system's ingredients, like prototype-based figure-ground segmentation, extraction of brain-like analytic features, and a simple classifier on top. Finally we analyze the performance of the system using databases of varying difficulty.

1 Introduction

The recognition of objects under real-world conditions is a difficult problem. Because of this, most approaches limit the complexity by using only few objects, restricting the pose to canonical views, or by providing controlled background conditions. In contrast to this, we freely rotated the objects in hand in front of a cluttered background. For this unconstrained setting, we describe a system that can robustly identify a large number of objects in real-time.

The recognition task and the given setting define the generalization capabilities the system requires. These have to be achieved by the interplay of the system components, but most strongly by the chosen type of object representation. On the one hand, the representation must be specific, i.e contain enough details to distinguish the objects. On the other hand, it must be general to yield invariance to the expected variations.

A main distinction with regard to representations can be made between holistic and parts-based approaches. Both types differ in the way they handle spatial information. Holistic approaches look at the whole image and represent global patterns in fixed relation to the image frame. All features are bound to a certain image location. Such representations are very specific and break down if the constellation of features changes strongly as it is the case for occlusion and 3D rotation. A simple holistic method might use the images directly as templates, or learn simple global features [1]. A more advanced processing is described in [2]. Here a hierarchical processing related to the ventral visual pathway is used, where stages of local spatial pooling soften the rigid coding of patterns. We will use this approach to provide a baseline for our results.

In contrast to holistic processing, parts-based methods have in common that they detect the presence of features or parts independent of their position in the image. The relative position between the parts of an object can be handled differently.
Some approaches store the constellation of parts on a reduced resolution [3] or by explicitly modeling there position by means of a Gaussian distribution. If multiple objects are in an image, this information is necessary to bind features to the corresponding object models. The handling of spatial information is less specific than for holistic coding but still leads to problems when the constellation undergoes strong changes as it is the case for 3D rotation. Additionally, these approaches often extract features at so-called keypoints only. Keypoints are determined by saliency detectors that favor parts whose position is not ambiguous (like vertices or highly textured regions, but not parallel lines or shadings). This is a limitation since meaningful information might be neglected.
Other parts-based approaches, like the one we use here, leave out spatial information by determining only the maximum response of an alphabet of features to an image [4, 5]. The use of such an alphabet is motivated by biological findings. The experiments in [6] revealed that columns in inferotemporal cortex represent a large set of complex features that can be recognized invariant to position and other transformations. Combinations of activated columns then code for the presence of an object [7]. The maximum step can also be interpreted by means of neural latency coding where the highest activations provoke the fastest response and non-optimal local responses are delayed and usually do not contribute to further feed-forward processing.
Leaving out spatial information, these approaches implicitly assume that only a single object is in view so that no binding is necessary. To balance this more general type of representation the parts themselves have to be more specific and meaningful. This can be achieved in different ways. The work in [5] uses a similar hierarchical processing like the holistic framework in [2]. But on the highest feature layer a maximum step is performed using an alphabet with millions of local features that were randomly selected. Finally a support vector machine (SVM) is trained to separate the classes in this high-dimensional space. Here the final SVM learns which parts of the large set are meaningful. In contrast to this, we use a much smaller alphabet of so-called analytic features, which are optimized using the supervised selection method described in [4], which will be explained later. Because of this smaller subset our system runs in real-time.

Besides feature selection and handling of spatial information, also the coding of the parts is important. For our analytic approach we describe the parts by means of SIFT descriptors [3]. A SIFT descriptor is made up of a grid of local gradient histograms. Thus it shows similarity to the response properties of neurons in primary visual cortex. Gray-scale gradients are a very simple form of edge detectors found in the so-called simple cells, while building of local histograms is comparable to spatial pooling which is attributed to so-called complex cells. Simple cells of a higher visual area respond to activation patterns of these complex cells. Such patterns can be interpreted as a grid.

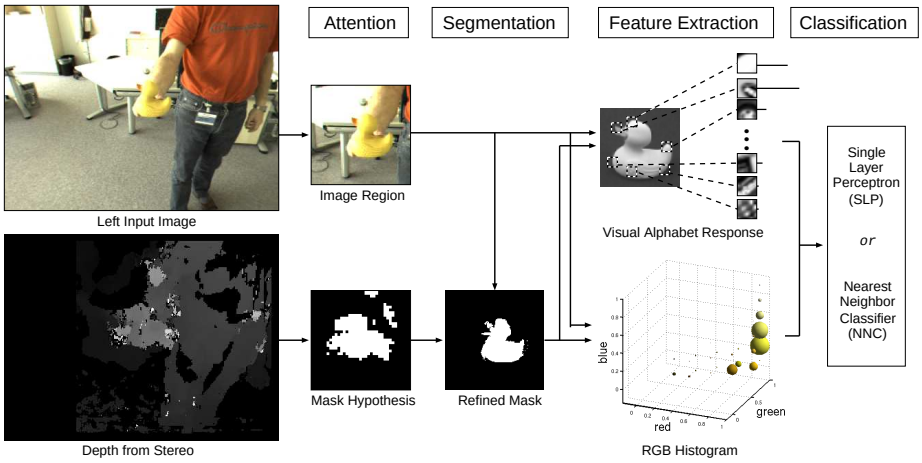

Fig. 1. Basic system architecture. Using a depth criterion a region of interest is cropped from the input image. After computing an improved mask, the response to an alphabet of parts is calculated together with a histogram in RGB color-space. The resulting activations are presented to the final classifier.

Previous experiments in [4] revealed that SIFT descriptors outperform the use of gray-scale patches, which are too specific, and also patches from the output of the hierarchy in [2], which are too general. Please note, that instead of using the whole framework usually associated with SIFT, we use the simple maximum step as outlined before. Additionally, we omit the use of keypoint detectors to avoid restrictions on the parts that can be learned.

With regard to the recognition task and the system architecture, the work of [8] is quite similar to ours. But they use a rather large alphabet of features which is trained in an unsupervised fashion and they represent spatial relations. This more complex and slower processing is not reflected in a gain in performance as we report a similar performance for an even higher number of objects.

We describe the building blocks of our system in Sect. 2 with a special focus on the learning and use of the analytic features. Later we investigate and discuss its performance in Sect. 3 and present our conclusions in Sect. 4.

2 System

In this section we describe the essential building blocks of the system, whose overall architecture is shown in Fig. 1.

Attention. When performing recognition tasks in unconstrained environments (presence of background clutter and variation in object position), the system has to decide which part of the input image should be processed. Here we use the concept of peri-personal space [9] to generate such a hypothesis. This concept defines an image region in close distance range to the camera as being relevant. In

each input image a square region of interest (ROI) is defined around the current hypothesis, whose size depends on the estimated distance. This ROI is scaled to a fixed output resolution of 144x144 pixels. In this way, we normalize object size variation caused by different viewing distances. We obtain the necessary depth information from stereo disparity and employ a pan-tilt unit to actively track the hypothesis until it violates the peri-personal constraints.

Segmentation. The size-normalized region contains the object, but also a substantial amount of background clutter. Since we do not represent spatial information, features detected on the background would be wrongly associated to the object. This would harden the task of the classifier and therefore we need to segment the object from its background. Following the peri-personal concept, a first foreground hypothesis can be derived by binarizing the depth image. Since the depth information based on stereo disparity usually wears out at the object's border and cannot be estimated for non-textured regions, we apply the segmentation method proposed by [10]. As pre-processing, this method removes all skin-colored pixels from the foreground hypothesis, because otherwise the hand holding the object would have a systematic influence on the result. Second, based on color and position information a prototype-based model for foreground (i.e. everything activated in the initial hypothesis) is learned and a model for background correspondingly. Finally, these models are used to classify each pixel as being figure or ground, where the learned prototype-specific distance metrics leads to a good generalization performance at the border of an object. In the following, features are extracted only at locations marked as foreground.

Feature extraction. The feature extraction is the most important part of the system. In this work we extract features for texture and color. Texture is represented by means of analytic features as proposed in [4] which are a preselected alphabet of SIFT-descriptors. These descriptors are widely used for coding local texture with invariance to lighting and planar rotation [3]. For a given input image i the response of a feature $\mathbf{w}_m$ is determined by $r_{mi} = \max_n (\mathbf{w}_m \cdot \mathbf{p}_{in})$, where the $\mathbf{p}_{in}$ are SIFT-descriptors from all image locations n, and $\cdot$ denotes the dot product. Keeping only the maximum response of each analytic feature over the image, we measure the pure presence of a certain object part and do not represent their spatial constellation. This yields invariance to translation of parts together with a strong reduction of dimensionality. In contrast to this, the standard SIFT framework calculates descriptors at interesting keypoints and their constellation is then matched to those of the training images. In Sect. 3 we show that this has shortcomings for several reasons.
The alphabet of analytic features is optimized for the scenario at hand using the selection method proposed in [4]. Starting from a large set of candidate SIFT-descriptors, this method first evaluates how well each element m can separate views from a single class. This is done by assigning scores s_{mi} for each combination of feature and image as shown in Fig. 2. After that, out of the candidates a subset M is selected that can separate most of the views among the training images. This subset has a predefined cardinality (usually several hundreds) and

Feature $\mathbf{w}_m$ 	Image i 								
	Response r_{mi}	0.43	0.45	0.48	0.49	0.54	0.56	0.60	0.85	0.90
	Score s_{mi}	0	0	0	0	0	0	0	1	1

Threshold t_m

Fig. 2. Score table of single feature. For visualization the images are sorted on their response r_{mi}. The threshold t_m separates views of a single class (here smiley cup) from all other images. To these views a score $s_{mi} = 1$ is assigned.

maximizes $\sum_i f\left(\sum_{m \in M} s_{mi}\right)$ with $f(z) = \frac{1}{1+e^{-kz}}$ and $k = 3$. Because trying all possible subsets M is intractable, a greedy iterative selection method is used instead. The described method is dynamic in the way that it selects more features for objects with strong variation in appearance.
To represent color we calculate a histogram in RGB color-space ($6x6x6 = 216$ bins) and normalize it by dividing by the highest entry. Histograms combine robustness against view and scale changes with computational efficiency [11]. Before the calculation, we apply the color constancy method proposed by [12]. The activation of the RGB histogram bins are combined with the responses of the analytic features to form the final feature vector.

Classification. To associate an object label to the current input image we use simple classifiers as a nearest neighbor classifier (NNC) or a single layer perceptron (SLP). The NNC stores the feature vectors of the training images as representatives and determines the object label for a test image based on closest Euclidean distance. The SLP has a neuron for each object. Using the training data, the weights of one neuron are adapted to produce a strong response for the corresponding object and a low response for views of other objects. The object label of a test image is determined by the highest activated neuron.
For the real-time system we use the SLP because it consumes drastically less memory and CPU time, and also has a slightly higher performance for the combined use of analytic and color features.

As outlined before, the usual platform is a stereo camera head mounted on a pan-tilt unit. When using our humanoid robot ASIMO instead, its degrees of freedom are used to track but also follow the current peri-personal hypothesis [13]. The proposed system runs in real-time with a frame-rate of 6Hz. The limiting factor is the calculation of the analytic feature response for each possible location in the size-normalized region of interest.

3 Results

In this section we first present results for an object database which has been acquired to train and optimize the final real-time recognition system. Using a

Fig. 3. HRI126 database. Database contains 126 objects with 1200 views each. Objects were rotated in hand in front of a cluttered background.

simpler database, we later distinguish the analytic feature approach from the standard SIFT framework.

The HRI126 database which corresponds to the scenario for the real-time system is shown in Fig. 3. It contains 126 objects with 1200 views each. The objects were freely rotated in hand in front of a cluttered background. Because of this unconstrained setting the database is very difficult compared to ones used in related work. In the following we evaluate the recognition performance and scalability of the proposed approach and test the necessity of the segmentation step. All training was done on the first 1000 views per object while the offline performance was evaluated on the remaining 200 views.

In the first training step we selected 441 analytic features (see Fig. 4a) using the algorithm proposed in [4]. Combined with the 216 RGB histogram bins, this yields a 657 dimensional feature vector for each view. Although in the final system an SLP is used, Fig. 4b gives the result of some NNC experiments using different representations and varying the number of representatives. Here especially the result of two holistic approaches GRAY and C2 is important to judge the difficulty of the database. GRAY simply uses the holistic gray-scale images as representatives while the so-called C2-activation is the output of the biologically inspired, edge-based, feed-forward hierarchy proposed in [2]. Both holistic methods show a very weak performance. With many training views C2 outperforms GRAY but still does not generalize as well as the analytic approach using few training views. One reason for this is that the coding of spatial information is too rigid. Additionally, the local features underlying C2 are too coarse to separate certain objects in the database, e.g. individual mobile phones. In contrast to this ANALYTIC uses very specific features while neglecting spatial information completely. ANALYTIC also outperform the color histograms while the concatenation of both complementary feature types yields the best result.

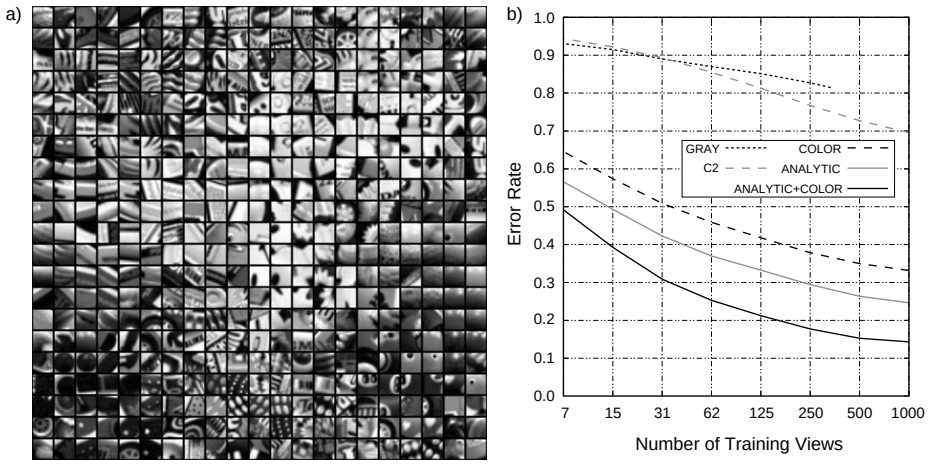

Fig. 4. a) Selected 441 analytic features. Most of the features are quite object specific. E.g., different versions of keypads and wheels are necessary to distinguish individual mobile phones and cars respectively. The set contains features that would not have passed usual keypoint criteria (e.g. several versions of parallel lines). For visualization the features are arranged using a self organizing map. b) Results of NNC experiments. Error rate over number of training views. For GRAY not all views could be used because of the required memory.

Similar conclusions can be drawn from the SLP experiments shown in Fig. 5a. Here the error rates of recognition are given depending on the number of used objects. This should help to predict the scalability of the approach towards larger number of objects. The value for 126 objects is directly the performance of the SLP on the test images. The performance for less objects was determined by choosing a random subset of objects and removing their test-views and SLP neurons from the experiment. The SLP was not retrained on the remaining objects. Interestingly, this yields better results because the SLP profits from a high number of negative training examples. The curves show the average of 100 runs. In general, the order from the NNC experiment is preserved. Only for ANALYTIC+COLOR the SLP is better than the NNC, because it finds a better weighting between both feature types than the simple concatenation used for the NNC experiment. The selective use of ANALYTIC and especially COLOR prevents the SLP from finding a good separation because of the low input dimensionality. In contrast to this we observed some over-fitting for C2.
For a given error rate much more objects can be distinguished by means of analytic features than by C2. The combination ANALYTIC+COLOR again provides the best result with an error rate of only 10.35% for 126 objects. Taken the difficulty of the database into account this is a very high performance and a big step towards invariant 3D object recognition. In the real-time system we accumulate the classification results over 10 successive frames and only output

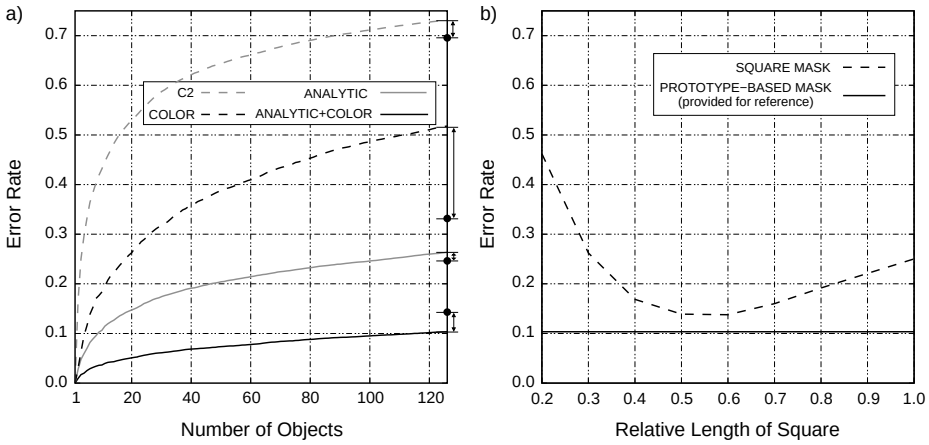

Fig. 5. a) Error rate of SLP over the number of objects. The symbol • denotes the performance of the corresponding NNC experiments. b) Error rates for differently sized square masks. For 1.0 the whole region is used for feature extraction (no masking) while for decreasing values only a smaller square inner part contributes.

the most voted object label. This removes outliers and thus leads to further improvement and stability.

After having investigated the contribution of the feature extraction, Fig. 5b sheds light on the importance of the segmentation step. The horizontal line is the reference performance for 126 objects when using the prototype-based segmentation proposed in [10], while the other curve gives the results when simply placing differently sized, square masks in the center of the region. Even in the best case such a simple mask has a 4% higher error rate. Using no mask gives a 15% higher error rate. For too small masks there is an even higher loss in performance. These results confirm that our position invariant object representation requires a good segmentation to counteract the binding problem.

In Sect. 2 we shortly compared the basics of the analytic feature approach with the standard SIFT framework. The effect of the differences become clear in the results in Fig. 6. For this experiment we used the simple COIL100 database [14]. Because of the non-cluttered background we did not use a mask and we also abandoned the color features to get a fair comparison. Fig. 6b shows the result of different nearest neighbor classifications where we varied the number of stored representatives (out of 72 available ones per object) for different approaches. For the analytic approach we used the same set of 441 features that was selected for the HRI126 database. This was done because of the low number of available training views in the COIL100 database and the strong similarity in the types of objects in both databases. For the SIFT framework we applied the visual pattern recognition system (ViPR) by Evolution Robotics (see [15], www.evolution.com) which is claimed to be the "gold standard" implementation of the SIFT approach. We optimized the parameters of this software for best

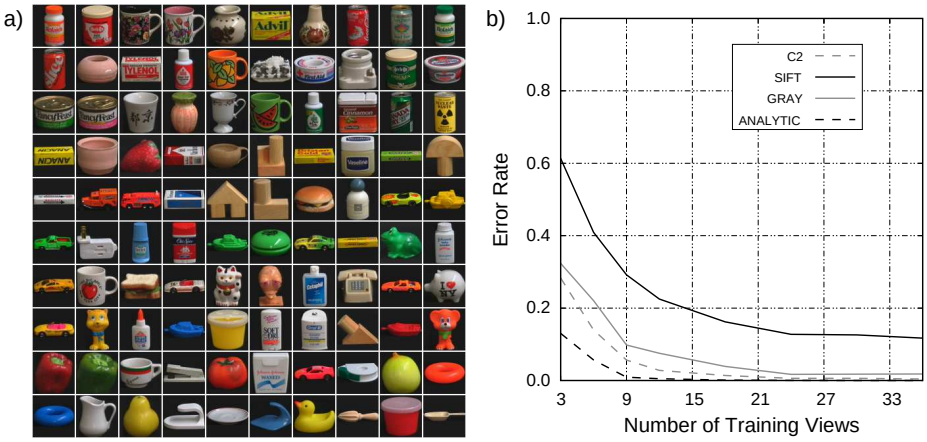

Fig. 6. a) COIL100 database. b) Error rates of NNC over number of training views. For 9 representatives the analytic approach has an error rate of 1% while the standard SIFT approach has 28%. The objects in a) are sorted in ascending order on their individual probability of being misclassified using the SIFT approach with 9 representatives.

performance. As a baseline we again provide results for the use of holistic gray-scale images and for C2.
For few training views, the analytic approach generalizes well while the SIFT framework shows a very weak performance. A reason for this is the different handling of spatial information. SIFT tries to re-detect the rigid constellation of parts that was present in the training images, which usually changes strongly under rotation in depth. Additionally, there are several objects for which the SIFT framework completely fails, as shown by the bad convergence towards larger numbers of training views. A reason for this is the dependency on the (re-)detection of interesting keypoints. This breaks down for objects with little texture, which is underlined by the order of objects in Fig. 6a. Both holistic approaches show good convergence and an intermediate capability to generalize from few training views, as they also use a rigid spatial representation. This rigidness is a little softened by the hierarchical processing underlying C2.

To also compare our approach to that in [5], we trained a set of 300 analytic features for an animal vs. non-animal separation task. On the test data we reached an error rate of 20% compared to 18% reported in [5]. This small difference makes it questionable if the hierarchical processing and the use of millions of local features do provide a gain over our simpler and much faster method.

4 Conclusion

On the basis of a biologically motivated, parts-based representation, we developed a real-time system capable of robustly recognizing a large number of arbitrary objects under 3D rotation. We evaluated the scalability of the approach

and showed the necessity of a good object segmentation to deal with background clutter. The shown performance marks a major step towards invariant object recognition, especially in comparison to existing work where mostly more complex processing is used to solve easier tasks.

Using the presented pre-trained architecture as a starting point, we target at a flexible, life-long learning system. Therefore we investigate in hierarchical classifiers to deal with the increasing complexity of the scenario and in an incremental build-up of the visual alphabet.

References

1. Turk, M., Pentland, A.: Eigenfaces for Recognition. Journal of Cognitive Neuroscience **3**(1) (1991) 71–86
2. Wersing, H., Körner, E.: Learning Optimized Features for Hierarchical Models of Invariant Object Recognition. Neural Computation **15**(7) (2003) 1559–1588
3. Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision **60**(2) (2004) 91–110
4. Hasler, S., Wersing, H., Körner, E.: A comparison of features in parts-based object recognition hierarchies. In: Artificial Neural Networks – ICANN. (2007) 210–219
5. Serre, T., Oliva, A., Poggio, T.: A feedforward architecture accounts for rapid categorization. In: Proc. of the National Academy of Science. (2007) 6424–6429
6. Tanaka, K.: Inferotemporal Cortex And Object Vision. Annual Review of Neuroscience **19** (1996) 109–139
7. Tsunoda, K., Yamane, Y., Nishizaki, M., Tanifuji, M.: Complex objects are represented in inferotemporal cortex by the combination of feature columns. Nature Neuroscience **4**(8) (2001) 832–838
8. Kim, H., Chutorian, E.M., Triesch, J.: Semi-autonomous learning of objects. In: IEEE CVPR Workshop: Vision for Human-Computer Interaction. (2006) 145
9. Goerick, C., Mikhailova, I., Wersing, H., Kirstein, S.: Biologically motivated visual behaviours for humanoids: Learning to interact and learning in interaction. In: Proc. IEEE/RSJ Int. Conf. on Humanoid Robots, Tsukuba, Japan. (2006)
10. Denecke, A., Wersing, H., Steil, J.J., Körner, E.: Online figure-ground segmentation with adaptive metrics in generalized LVQ. Neurocomputing (2009) in press
11. Swain, M.J., Ballard, D.H.: Color indexing. International Journal of Computer Vision **7**(1) (1991) 11–32
12. Pomierski, T., Gross, H.M.: Biological neural architecture for chromatic adaptation resulting in constant color sensations. In: IEEE International Conference on Neural Networks. (1996) 734–739
13. Bolder, B., Dunn, M., Gienger, M., Janssen, H., Sugiura, H., Goerick, C.: Visually guided whole body interaction. In: IEEE International Conference on Robotics and Automation. (2007)
14. Nayar, S.K., Nene, S.A., Murase, H.: Real-time 100 object recognition system. In: Proc. IEEE Conference on Robotics and Automation. Volume 3. (1996) 2321–2325
15. Munich, M.E., Pirjanian, P., Bernardo, E.D., Goncalves, L., Karlsson, N., Lowe, D.G.: SIFT-ing through features with ViPR: Application of visual pattern recognition to robotics and automation. IEEE Robotics and Autom. Mag. (2006) 72–77