# Learning Visual Object Categories with Global Descriptors and Local Features

Rui Pereira[1] and Luís Seabra Lopes[1,2]

[1] Transverse Activity on Intelligent Robotics, IEETA, Universidade de Aveiro,
[2] DETI, Universidade de Aveiro,

**Abstract.** Different types of visual object categories can be found in real-world applications. Some categories are very heterogeneous in terms of local features (broad categories) while others are consistently characterized by some highly distinctive local features (narrow categories). The work described in this paper was motivated by the need to develop representations and categorization mechanisms that can be applied to domains involving different types of categories. A second concern of the paper is that these representations and mechanisms have potential for scaling up to large numbers of categories. The approach is based on combinining global shape descriptors with local features. A new shape representation is proposed. Two additional representations are used, one also capturing the object's shape and another based on sets of highly distinctive local features. Basic classifiers following the nearest-neighbor rule were implemented for each representation. A meta-level classifier, based on a voting strategy, was also implemented. The relevance of each representation and classifier to both broad and narrow categories is evaluated on two datasets with a combined total of 114 categories.

## 1 Introduction

Category learning is a core problem in cognitive science and artificial intelligence. Visual category learning and recognition are capabilities relevant for a wide range of applications, from digital libraries and the semantic web to manufacturing systems and robotics. Different types of visual categories can be found in real-world applications. Some are very heterogeneous in terms of local features while others are consistently characterized by several highly distinctive local features.

Given the different characteristics of categories, different representations can be more suited to some categories than to others. Based on this fact, Wardhani and Thomson [19] divide images into several very abstract categories, such as natural scenes, people, etc.., and use different methods according to this division. This way, they obtain better results than using a single method. However, it appears that most successful results reported in the literature come, not from using multiple classifiers according to the type of images, but from their combination, be it by using voting [5][18][16] or other approaches [1][6][16].

This paper also explores multiple representations and classification mechanisms to address domains where different types of categories must be processed.

For that purpose, visual object categories are divided into two main groups: if a category exhibits a high variation among its members, then it is said to be a *broad* (or *general*) category; on the other hand, if a category exhibits small variations between its members and they display a high number of unique local features, the category is said to be *narrow* (or *specific*). Broad categories usually are named by common nouns like "cup", "chair", "apple", etc., while narrow categories are named by proper nouns, including brand names, book titles, etc..

Previous work in our group [16] already explored the geometrical analysis of objects to compute multiple object representations, based on divisions of the object in layers and slices. In that work, the combination of the different representations is achieved through a multi-classifier architecture. The final categorization of an object is delivered by a classifier combination. With this approach, it was possible to learn over 70 categories. However, the approach was only tested with broad categories, in the sense explained above. Sarfraz and Rhida [14] also use multiple representations derived from divisions of the objects in slices and layers: one based on three concentric layers and another based on eight slices around the centre of the object. In addition, several other different types of features, such as moment invariants and the so-called "simple shape descriptors", were also used. All features of an object were gathered in a feature vector and objects were compared with Euclidean distance. The approach was applied to the problem of learning 12 broad object categories.

To enhance performance, several projects have been exploring representations combining global and local features. Murphy et al. [9] used global and local features to detect and localize objects in images. Patches around interest points are converted to codewords. Because this type of features is sometimes ambiguous, the coarse texture and spatial layout of the image are also used to help overcoming this difficulty. Once global features are computed, an approximate location and scale of the object are predicted and a local detector is applied to refine this prediction.

Lisin et al. [7] developed a system for object classification and inquired on two different types of combinations of local and global features: stacking, in which outputs of the individual classifiers are used as input features for a meta-classifier; and hierarchical classification, in which classes that are not separable by global features, are clustered and a local feature classifier determines to which class they belong. SIFT [8] is used for computing local features and three shape properties (area, perimeter and compactness) are used as global features. The authors concluded that stacking is superior to hierarchical classification, using images of plankton from a total of 14 categories.

Neumann et al. [10] researched the use of local and global features for logo classification. Logos are analyzed as sets of connected components. Local shape information is extracted from these components, such as eccentricity, circularity, etc.. Global shape information is retrieved by computing the horizontal and vertical projections of the logo's binary image (i.e., counting the number of white pixels for each row and column). A vector describing the logo (its signature) is then obtained using a wavelet transformation. A combined classifier works by

adjusting the weights of individual classifiers while classifying a logo. Tests were carried out on the UMD-Logo-Database, containing 123 logos.

In the context of human-robot interaction, some recent approaches explore the combination of incremental learning and interaction with teachers to ground vocabulary about physical objects [16][15][13][17]. This type of approach to category learning naturally raises the problem of scalability: How many visual categories can an artificial system learn? What are the main features of an incremental and open-ended category learning process?

The work described below in this paper was motivated by two of the main problems identified above. First, the need to combine different types of representations to support the acquisition of different types of categories. Second, the need to develop powerful representations and categorization mechanisms enabling to scale up to larger numbers of categories. The work was carried out in the framework of the development of UA@SRVC, a team for participation in the *2nd Semantic Robot Vision Challenge, SRVC'2008* (Anchorage, Alaska, June 2008), an event sponsored by NSF and Google. Other aspects of UA@SRVC are presented in a separate paper[11].

Figure 1 provides an overview of the developed category learning and recognition system. Three basic classifiers, using different representations, are the base of the categorization system. Two of the representations are global shape representations and the third is based on SIFT local features. A meta-level classifier is also included. This classifier combines the decisions of three basic classifiers through voting. The approach is tested on a large number of categories (68 broad categories and 46 narrow categories, for a total of 114 categories).

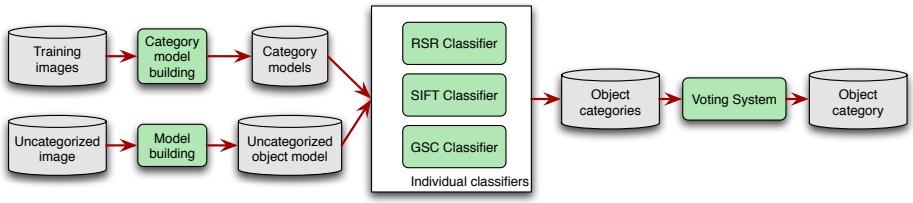


**Fig. 1.** System overview.

The paper is organized as follows: Section 2 describes the representations used for objects and categories. Section 3 describes the categorization mechanisms. Section 4 describes the performance evaluation approach and obtained results. Section 5 summarizes conclusions and future work.

## 2 Representations

As mentioned before, different representations can be more suited to some categories than to others. This section presents the alternative object and category

representations used in this work. A global shape context was designed and implemented. For comparison, the shape representation proposed by Roy [13] was also implemented and used. For handling the specificities of narrow categories, local features extracted with SIFT [8] are also used.

**Global shape context.** The edges of an object are usually representative of its shape. For a human, it is usually easy to associate a "line drawing" with the object that it is supposed to represent. One of the shape representations used in this work is a polar histogram of edge pixels that we call a "global shape context" (GSC). A polar frame of reference is located at the geometric centre of the object. Then, the space around the centre up to the most eccentric pixel is divided into $a$ slices (angle bins) and $d$ layers (distance bins) [3]. The intersection of slices and layers results in a polar matrix (Fig. 2)(a) that will be mapped to a 2D histogram counting the number of edge pixels in each cell. This histogram is finally normalized by dividing the counts by the total number of edge pixels.

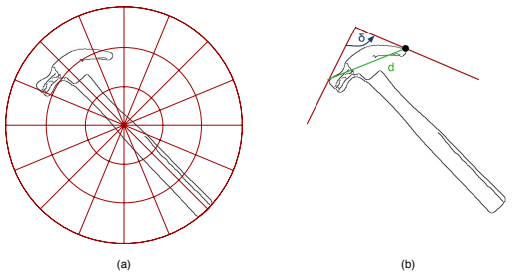


**Fig. 2.** Shape representations: (a) Global shape context (GSC). (b) Roy's shape representation (RSR)

In most real-world applications, rotation-, translation- and scale-invariance are necessary. The proposed GSC is translation invariant, since it is computed in a frame of reference centered in the geometric centre of the object. It is also scale invariant because the image region mapped to the histogram is delimited by the minimal circle enclosing the object. The histogram itself is not invariant to rotation, but similarity between any two objects can be computed in a rotation-invariant way, as explained in section 3.

The histogram here proposed is similar to the shape context of Belongie et al. [2]. However, in their approach, the shape context itself is used as a local feature, and thus a logarithmic distance scale is used. For each edge pixel, a different shape context is computed. In contrast, the histogram here proposed is used as a global shape descriptor. In terms of computational complexity, building the GSC takes $O(n)$ time, $n$ being the number of edge pixels, whereas building the

[3] a=40 and d=10 were used.

shape representation proposed by Belongie et al. involves building shape contexts for all pixels, which takes $O(n^2)$ time. The GSC can also be related to the binary shape matrix of Goshtasby [4], derived from a polar raster sampling also centered in the object's geometric center. The shape matrix, whose cells represent the points of intersection between the circles and radial lines in the polar raster are mapped to the cells in the shape matrix. The value in a cell is 1 if the corresponding intersection point is inside the object and 0 otherwise. Similarity between two objects is given by the percentage of matrix cells with the same value for both objects. While the shape matrix is a very light representation, the proposed global shape context contains much more information on shape-related details of an object, such as internal edges. The GSC can be seen as a compromise between the computational lightness of the shape matrix and the high expressivity of the Belongie et al.'s approach.

**Roy's shape representation.** Roy [13] proposed a shape representation based on tangents to object edges. It will be refered here as "Roy's shape representation" (RSR). A Canny edge detector is used to obtain the edges from an object. Then, for all pairs of edge pixels, the distances between them, $D$, and the angles between the respective tangents, $\delta$, are computed (Fig. 2)(b). The tangent at a pixel is approximated by the linear regression line of the neighboring pixels and its orientation is estimated from the regression line [12]. All this information is summarized through a two-dimensional histogram with $a$ angle bins on one dimension and $d$ distance bins on the other [4]. In each cell, the histogram counts the number of edge pixel pairs in the corresponding distance and angle bins.

To obtain scale invariance, distances are normalized by dividing each of them by the maximum distance between any two edge points. Since the angle calculated between the tangents at two edge points is a relative measurement, rotation invariance is also achieved.

For an object with $n$ edge points, $\frac{n(n-1)}{2}$ distances and angles have to be determined, giving a complexity of $O(n^2)$ for the process of building a RSR.

**SIFT.** SIFT (Scale Invariant Feature Transform [8]) produces highly distinctive features that can be used for matching objects with different scales, positions and orientations, as well as with some variations in illumination. A Difference-of-Gaussians function is used to detect keypoints, based on scale-space extrema. Local gradients sampled in a grid in the vicinity of a keypoint are summarized through a histogram that becomes the keypoint descriptor, as illustrated in Figure 3. From an $8\times8$ grid centered in the keypoint, gradients are grouped $4$ by $4$ forming a $2\times2$ descriptor with $8$ directions. The length of the arrows in the right represents the sum of the magnitudes of the gradients, which were nearer to each of the $8$ directions, weighted by a Gaussian window. In the work

[4] a=32 and d=32 were used in the implementation [12]. This was decided based on experimentation showing it was better for our objectives than the 8×8 suggested by Roy [13].

of this paper, a *4*×*4* descriptor with also *8* directions is computed from a grid of *16*×*16*, where gradients are grouped *4* by *4*, forming a vector of 128 features per keypoint.

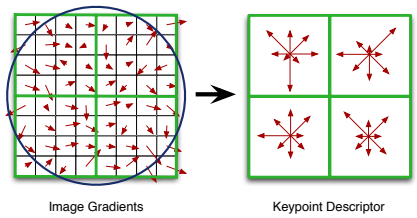


**Fig. 3.** SIFT keypoint descriptor creation.

**Category models.** An instance-based approach was adopted for category representation. Three alternative category models are used: (1) The category is represented by the set of global shape contexts of the respective training images; (2) The category is represented by the set of Roy's shape representations of the respective training images; (3) The category is represented by the concatenation of the lists of SIFT keypoints extracted from the respective training images.

## 3 Classification

The goal of the classification (or categorization) module is to categorize objects in the context of an application. This module is composed of three basic classifiers, each based on one of the object representations presented above, and a meta-level classifier. In the basic classifiers, categorization is done by computing a representation of the target object and comparing it to the category models. Categories are ranked in descending order of similarity to the target object. A nearest neighbor strategy is adopted, in which the similarity of the target object to a category is given by the highest similarity between the target object and one instance of the category. The highest ranked category will be the category predicted by the classifier. The ranking itself is used by the meta-level classifier.

**GSC-based classifier.** This classifier follows the general nearest-neighbor scheme just outlined. The $\chi^2$ distance is the base for assessing similarity between any two objects.

$$D_{pq} = \frac{1}{2} \sum_{i=1}^{a} \sum_{j=1}^{d} \frac{[h_p(i,j) - h_q(i,j)]^2}{[h_p(i,j) + h_q(i,j)]} \tag{1}$$

where $p$ and $q$ are two objects, $h_p$ and $h_q$ are the respective histograms (GSCs), $a$ is the number of angle bins, $d$ is the number of distance bins, $i$ is an angle bin and $j$ is a distance bin.

Finally, the similarity between $p$ and $q$ is given by $S_{pq} = 1/D_{pq}$.

As mentioned before, the global shape context is not invariant to rotation. To make rotation invariant matching possible, the histograms are rotated and compared $a$ times (where $a$ is the number of angle bins). The angle displacement that results in the lowest distance is the one used to calculate the similarity between the two shapes.

For $a \times d$ histograms, computing the similarity between two histograms has a complexity of $O(a \times (a \times d))$, since we calculate the $\chi^2$ distance $a$ times (for each histogram rotation). Being independent of the number of edge pixels, this is far more efficient than object matching with the shape representation of Belongie et al. [2], which is based on local shape contexts centred in edge pixels. In this case, most of the steps of the similarity computation algorithm run in time quadratic to cubic in the number of edge pixels.

**RSR-based classifier.** Follows the general nearest-neighbor scheme just outlined and also uses the $\chi^2$ distance as the base for assessing similarity. Since RSR is rotation-invariant, a single comparison is enough. So, for $a \times d$ histograms, computing the similarity between two histograms has a complexity of $O(a \times d)$.

**SIFT-based classifier.** In this case, a category is represented by the concatenation of the lists of SIFT keypoints of all training images that belong to the category. When comparing two objects, the features (keypoints) in each of them are paired according to a nearest-neighbor criterion. Instead of using a global threshold to discard matches, the distances to the two nearest neighbors are computed and compared. If the ratio between them is greater than 0.35, the pair of features is rejected. Finally, similarity between the two objects is given by the number of accepted pairs of features.

**Voting system.** The performance of a combination of classifiers is often superior to their individual performance [3][6]. A voting approach makes the system very scalable, because it is possible to insert any number of additional classifiers, as long as they vote according to the same rules. In this work, the following voting scheme was implemented. Each classifier determines the 3 most likely categories of an object, i.e., the categories with the higher similarity scores to the object, and casts a variable number of votes to these top ranked categories. For the shape-based classifiers, the first category in the ranking gets 3 votes, the second gets 2 and finally the third gets 1 vote. In the case of the SIFT-based classifier, the three top categories get 6, 4 or 2 votes, respectively. This is done to balance the weight of global (shape-based) and local (SIFT) features in the final voting. In the end of the voting process, the category with more votes is selected. The maximum number of votes a category can have is 12, meaning that all classifiers agree on the best category for the image.

## 4 Performance evaluation

**Datasets.** To evaluate the performance of the classifiers, two sets of images are used. For the evaluation with broad categories, a set of 6914 images of 109 objects belonging to 68 different categories[5] was used. These images have been collected in experiments in the framework of the LANGG project[16] and do not have background noise, occlusion or deformation. However, images were captured under different illumination conditions (artificial / natural, morning / afternoon / evening, etc.) and therefore there is some variability between them.

In case of narrow categories, a set of 459 images of 46 different categories[6] was used. Many of these images were obtained from the *SRVC* (*Semantic Robot Vision Challenge*) datasets, while others were mannualy collected on the Internet. Objects can have any orientation or localization inside the image and can have a small quantity of noise, variations in 3D viewpoint and occlusion. For fair comparison between shape-based and SIFT-based classifiers, care was taken to not include images with significant clutter, occlusion or background noise, since this can ruin the performance of shape-based classification.

**Experimental approach.** The evaluation of the developed system on a particular dataset is carried out as a series of training and testing experiments, for increasing number of categories. For a given number of categories, $N$, the same basic experiment is repeated 5 times. In each basic experiment, a set of $N$ categories is randomly generated (from the categories in the dataset) and then training and test sets are created as follows:

- Training set: A set of randomly chosen images is used to build the model of the category. In the performed experiments, 4 training images were used for each category.
- Test set: Another set of randomly chosen images is used for testing the category models. In the experiments, the test set contains 3 images from each category.

**Broad categories.** Experiments with the broad categories dataset were carried out for the number of categories varying between 5 and 40, with a step of 5. This results in a total of $5 \times 180 \times 3 = 2700$ object classifications with each method. Figure 4 shows the average percentages of correct classifications for each

[5] letter a; battery; bottle top; box; box2; boy; cd; cigbox; circle; circuit board; coffee cup; coffee mug; coffee spoon; cup; duster; floppy; glove; glue bottle; horse; icetea can; ink remover bottle; key; key1; lighter; mobile; mouse; nail; number one; passbook; pen; pencil; penguin; penguin sitting; postit; remote; screw; sd; sf; stapler1; stapler2; stapler3; staple remover; sugar packet; table fork; table knife; table spoon; tape; teddy bear; number three; tilted coffee mug; tilted cup; tilted tractor; toy bike; toy car; toy jeep; toy mobile; toy saw; toy scissor; toy screw; toy sd; toy tractor; toy train; train top; twenty cent; ubuntu cd cover; usb pen; water bottle; water cup

[6] Aerius Allergy; book "Artificial Intelligence: a Modern Approach"; book "Big Book of Concepts"; book "Harry Potter and the Deathly Hallows"; book "Lonely Planet Walking in Italy"; book "Paris to the Moon" by Adam Gopnik; Butterfinger logo; Cadbury Caramilk; CD "Begin to Hope" by Regina Spektor; CD "Hey Eugene" by Pink Martini; CD "Introducing Joss Stone"; CD "Look-Alikes Jr." by Joan Steiner; CD "Retrospective" by Django Reinhardt; Cheerios cereal box; "Coca-cola" can; Colgate Fresh; Crayola Crayons box; Dasani bottle; Doritos Blazin' Buffalo Ranch; DVD "Gladiator"; DVD "I, Robot"; DVD "Madagascar"; DVD "Shrek"; Fanta orange can; "Gears of War" box; Gillette Mach3; Head & Shoulders shampoo; Ikea logo; Kellogg's Corn Flakes; "Lucky charms"; Nescafe Tasters Choice; Nintendo Wii box; "Pirates of the Caribbean" dvd; Portugal flag; Pringles; Red Bull can; Ritter Sport Marzipan; Ritz crackers; Snickers wrapper; Star Wars logo; Tide detergent; Toblerone; Twinings Earl Grey Tea; Twix Candy Bar; Vodafone logo; Whiskas logo was used.

number of categories and each method. While for a small number of categories there isn't a big gap in the performance of the classifiers, this is not true when more categories are added. The SIFT classifier has a faster and more evident degradation in the number of correct classifications. Shape-based classifiers, on the other hand, are more suitable for these types of objects. Both shape-based classifiers clearly outperform the SIFT-based classifier. In fact, while the SIFT-based classifier had an average accuracy of 53.9% in the experiments, the RSR-based classifier had an average accuracy of 68.0% and GSC was even better, with an average accuracy of 78.3%. The difference between these strategies becomes more noticeable with the increase on the number of categories.

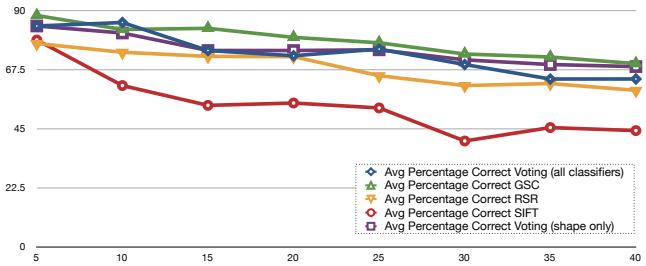


**Fig. 4.** Classifier performance on broad categories versus number of categories

Two voting strategies were tested, one combining only the two shape-based classifiers and the other combining all classifiers. No significant difference on their performance was noticed. In fact, the first had an average accuracy of 74.8% and the second 73.6%. They both performed worse than the best basic classifier (GSC), but closer to this one than to any of the others.

We can therefore conclude that, for broad categories, because of high intra-category variance and few distinctive keypoints, shape-based classification provides much better results than SIFT-based classification. We can also conclude that including the SIFT-based classifier in the voting system doesn't degrade performance significantly.

**Narrow categories.** In narrow categories, a similar set of experiments was carried out, i.e. from 5 to 40 categories, with a step of 5. Figure 5 shows the average classification accuracies for each method. According to these experiments, the best method for this type of categories is the one based on SIFT, with an average accuracy of 92.1%. In contrast, RSR has an average accuracy of 64.4% and GSC has an average accuracy of only 45.1%.

SIFT is a very good method for this type of categories. Their richness in local features makes it possible to achieve good classification results with SIFT, while at the same time (together with variations in 3D viewpoint, occlusion, etc.) confuses the classifiers based on shape.

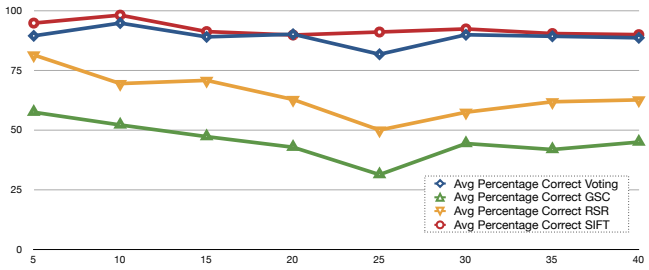


**Fig. 5.** Classifier performance on narrow categories versus number of categories

The voting system, combining the three basic classifiers, reached an average accuracy of 89.0%, i.e. performed only slightly worse than the best basic classifier.

**Mixed categories.** In the previous experiments, it could be observed that the voting system performed only slightly worse than the basic classifiers more suited for the type of categories (either broad or narrow) present in the dataset. Therefore, it becomes interesting to find out how the voting system behaves in a dataset where both types of categories are present. For that purpose, the two mentioned datasets were united producing a new dataset with 7373 images of 114 different categories. Experiments on this united dataset were carried out starting with 10 categories and going up to 110 categories with a step of 10. Experiments were also conducted for the complete set of 114 categories. This results in a total of $5 \times 774 \times 3 = 9288$ classifications with each method.

Figure 6 shows the results. Given the presence of both broad and narrow categories, no major differences in performance were observed between the three basic classifiers: 60.3% for RSR, 57% for GSC and 50.9% for SIFT. The superior performance of GSC reflects, at least in part, the fact that there are more broad categories (for which GSC is the most suited) than narrow categories. The average accuracy of the voting system was 65.3%. The fact that the voting system improves on the individual classifiers is inline with previous observations [3][5][18][16][6].

With respect to scalability, the obtained results seem promising. In fact, while the average accuracy of the best classifier (voting) degradates visibly (from 90% to 67.5%) between 10 and 40 categories, it nearly stabilizes between 40 and 114 categories. We therefore have reasons to expect that the approach will easily scale up to larger numbers of categories.

## 5 Conclusions

The work described in this paper was motivated by the need to develop representations and categorization mechanisms that can be applied to domains involving different types of categories. The approach is based on combinining global shape descriptors with local features. A new shape representation was proposed, the

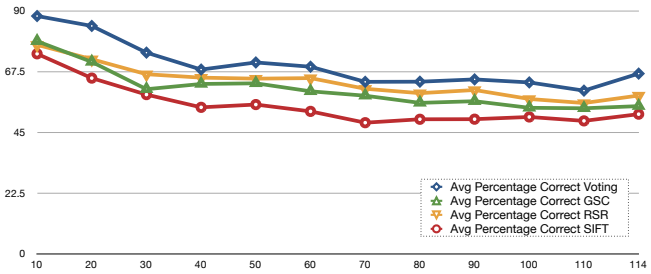


**Fig. 6.** Classifier performance on mixed categories versus number of categories

global shape context (GSC). Two additional representations were used. One is Roy's shape representation (RSR). The other is based on sets of SIFT local features. Basic classifiers following the nearest-neighbor rule were implemented for each representation. In tests with up to 40 broad categories, GSC performed clearly above the other two classifiers, and SIFT delivered the worse results. By contrast, in similar tests with narrow categories, SIFT delivered the best results, far above the other two classifiers. RSR had intermediate performance on both domains. In a mixed domain combining 68 broad categories and 46 narrow categories (114 categories in total), the three classifiers had a more balanced performance. In this case, GSC was the best, which also reflect the fact that broad categories were in majority.

A meta-level classifier, based on a voting strategy, was also implemented. In tests on domains with only one of the types of categories (either broad or narrow), the meta-level classifier performed only slightly worse than the best basic classifier for those domains. In tests with the mixed domain, the meta-level classifier produced the best results.

Another concern of this work was to assess the scalability of the developed/implemented representations and classification mechanisms. In this respect, the obtained results seem promising. In the largest tests (with 114 categories), the average accuracy of the best classifier (voting) degraded visibly between 10 and 40 categories, but stabilized between 40 and 114 categories, so it's possible that the approach will easily scale up to larger numbers of categories.

## 6 Acknowledgments

The first author is currently with a research grant funded by Aveiro University. The participation of the UA@SRVC team in SRVC'2008 was partially funded by Google. The used implementation of SIFT was developed by Rob Hess and is publicly available. The implementation also used OpenCV extensively.

## References

1. Al-Ani, A., and Deriche, M. A new technique for combining multiple classifiers

using the dempster-shafer theory of evidence. *Journal of Artificial Intelligence Research 17* (2002), 333–361.
2. Belongie, S., Malik, J., and Puzicha, J. Shape matching and object recognition using shape contexts. In *Pattern Analysis and Machine Intelligence, IEEE Transactions on* (2002), vol. 24, pp. 509–522.
3. Dietterich, T. G. Ensemble methods in machine learning. In *SpringerVerlag* (2000), pp. 1–15.
4. Goshtasby, A. Description and discrimination of planar shapes using shape matrices. *IEEE Trans. Pattern Anal. Mach. Intell. 7* (1985), 738–743.
5. Huang, Y. S., and Suen, C. Y. A method of combining multiple experts for the recognition of unconstrained handwritten numerals. *IEEE Trans. Pattern Anal. Mach. Intell. 17*, 1 (1995), 90–94.
6. Kittler, J., Hatef, M., Duin, R., and Matas, J. On combining classifiers. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 20*, 3 (Mar 1998), 226–239.
7. Lisin, D., Mattar, M., Blaschko, M., Learned-Miller, E., and Benfield, M. Combining local and global image features for object class recognition. In *Computer Vision and Pattern Recognition - Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on* (2005), pp. 47–47.
8. Lowe, D. G. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision 60* (2004), 91–110.
9. Murphy, K. P., Torralba, A. B., Eaton, D., and Freeman, W. T. Object detection and localization using local and global features. In *Toward Category-Level Object Recognition, Springer-Verlag LNCS* (2006), pp. 382–400.
10. Neumann, J., Samet, H., and Soffer, A. Integration of local and global shape analysis for logo classification. *Pattern Recognition Letters 23(12)* (2002), 1449–1457.
11. Pereira, R., Seabra Lopes, L., and Silva, A. Semantic image search and subset selection for classifier training in object recognition. *Progress in Artificial Intelligence: 14th Portuguese Conference on Artificial Intelligence - EPIA'2009, LNCS/LNAI, Springer* (2009), In Press.
12. Ribeiro, L. S. Object recognition for semantic robot vision. Master's thesis, Universidade de Aveiro, 2008.
13. Roy, D. K. *Learning Words from Sights and Sounds: A Computational Model*. PhD thesis, MIT, 2000.
14. Sarfraz, M., and Ridha, A. Content-based image retrieval using multiple shape descriptors. In *Computer Systems and Applications, 2007. AICCSA '07. IEEE/ACS International Conference on* (2007), pp. 730–737.
15. Seabra Lopes, L., and Chauhan, A. How many words can my robot learn? an approach and experiments with one-class learning. *Interaction Studies, 8(1)* (2007), 53–81.
16. Seabra Lopes, L., and Chauhan, A. Open-ended category learning for language acquisition. *Connection Science, 8(4)* (2008).
17. Steels, L., and Kaplan, F. Aibo's first words: the social learning of language and meaning. *Evolution of Communication,4(1)* (2002), 3–32.
18. van Erp, M., Vuurpijl, L., and Schomaker, L. An overview and comparison of voting methods for pattern recognition. In *8th International Workshop on Frontiers in Handwriting Recognition (IWFHR-8)* (2002), pp. 195–200.
19. Wardhani, A., and Thomson, T. Content based image retrieval using category-based indexing. In *Multimedia and Expo, 2004. ICME '04. 2004 IEEE International Conference on* (2004), vol. 2, pp. 783–786 Vol.2.