Unsupervised Learning of Visual Object Recognition Models

Dulce J. Navarrete, Eduardo F. Morales, and Luis Enrique Sucar

Computer Science Department,
Instituto Nacional de Astrofísica, Óptica y Electrónica.
Luis Enrique Erro 1, 72840 Tonantzintla, México
{dj.navarrete,emorales,esucar}@inaoep.mx

Abstract. Object recognition from images is traditionally based on a large training set of previously annotated images which is impractical for some applications. Also, most methods use only local or global features. Due to the nature of objects some features are better suited for some objects, so researchers have recently combined both types of features to improve the recognition performance. This approach, however, is not sufficient for the recognition of generic objects which can take a wide variety of appearances. In this paper, we propose a novel object recognition system that: (i) uses a small set of images obtained from the Web, (ii) induces a set of models for each object to deal with polymorphism, and (iii) optimizes the contribution of local and global features to deal with different types of objects. We performed tests with both generic and specific objects, and compared the proposed approach against base classifiers and state-of-the-art systems with very promising results.

Keywords: object recognition, global features, local features, few images, multiple-classifiers.

1 Introduction

Consider a service robot that helps elderly people. Mary just acquired such a robot, and will like the robot to fetch a medicine she left in the kitchen and bring it to the bedroom. The robot, however, does not have a visual model of this particular medicine, so it searches for images in the Web, builds a model from a few images and then use this model to recognize it in an image.

Visual object recognition has been an area of research for several decades, in which important advances have been achieved in the last years. However, most object recognition systems: (i) require a large sample of annotated images of the object to be recognized [4], (ii) are usually focused on recognizing a particular class of object (i.e., faces) [6,9,15], (iii) many are based on local features which are good for recognizing specific objects (e.g., my cup) but are not so reliable for object categories (e.g., any cup), and (iv) fail with objects classes that have a high variability (e.g., recognizing apples - different colors, a single apple or a bunch of apples).

J. Pavón et al. (Eds.): IBERAMIA 2012, LNAI 7637, pp. 511–520, 2012.

In this work we have developed a general visual object recognition system that overcomes some of the previous limitations. It incorporates several novel features: (i) It starts with a small set of training images, obtained from the Web, and autonomously expand this initial set with artificial transformations for robustness against changes in scale, rotation and noise. (ii) It builds several classifiers to deal with polymorphism of objects. (iii) It automatically obtains an optimal combination of local and global features to recognize different types of objects.

A classifier ensemble is built for each object based on 12 images obtained with *Google Images*. We compared two selection strategies: (a) unsupervised –the first images returned by the search engine, (b) semi-supervised –a user selects a subset of images from the images returned by the search engine. Then each classifier is evaluated using a different set of positive and negative sample. We evaluated our object recognition system with 10 different objects from past editions of the Semantic Robot Challenge competition (SRVC) [1], that include specific objects and object categories, and with a set of images extracted from Google images used by other authors. The results are promising, with an average recognition rate of approximately 89.5% for specific objects and 78% for object categories. We also compared our method against other state-of-the-art systems trained with Web images. We obtained a clear superior performance, in terms of F-measure, using their same set of training images.

The rest of the paper is organized as follows. The next section summarizes related work. Section 3 describes the proposed method for object recognition and Sect. 4 presents the experiments and results. Finally, we conclude with directions for future work.

2 Related Work

The field of object recognition is very wide, so we focus on the most closely related work in 3 areas: (i) techniques that combine global and local features, (ii) methods based on multiple classifiers, and (iii) systems that make use of training images obtained from the Web.

Most object recognition methods are based either on global or local features (we do not consider in this review *structured* techniques). Global methods usually include color, texture and shape features, and are traditionally used for recognizing object categories. Local or key-point based methods model an object based on local descriptors and are more appropriate for recognizing specific objects [12], although recently there have been some extensions to recognize object categories [3]. There are few works that combine both types of features. In [10] they use global shape features, Fourier Descriptors combined with SIFT descriptors to help in improving the performance of object class recognition. In [18], the authors combine the advantages of appearance-based methods and key-point descriptors. First key-points are used for eliminating false matches, and then Local Binary Patterns confirm the match.

Recently, classifier ensembles have shown superior performance in several domains. In [14] a meta-level classifier produced the best results from nearest-neighbor classifiers that combined local and global features. In [8] the authors describe two alternative techniques for recognizing marine objects, one based on stacking and other on a hierarchical classifier, combining global and local features. In [7] the authors show a method based on hierarchical Dirichlet processes to generate intermediate mixture components for recognition and categorization.

Unsupervised object categorization relies on samples collected from the Web, some representative examples are [4,11,17]. Fergus et al. [4] apply probabilistic clustering methods to discover a mixture of visual topics in a collection of semi-organized image data.

Curious George [11] is a robot develop for the SRVC competition. Its object recognition component uses images retrieved from Computer Vision databases in the Web to build several classifiers based on SIFT features, shape and deformable parts models; which are combined for object recognition.

In [17] the authors propose an unsupervised learning algorithm for visual categories using potential images obtained from the web building on the work by Bunescu and Mooney [2]. The idea is to obtain several images by translating the category name into different languages and searching the web for images using those translated terms. The negative examples are collected from random images with obtained from different categories. They consider the fact that at least one of the positive images is truly a positive instance of the target concept while all the negative examples can be safely considered as negative instances of the target concept.

A multi-modal approach using text, metadata and visual features is used in [16]. Candidate images are obtained by a text-based web search querying on the object identier. The task is then to remove irrelevant images and re-rank the remainder. First, the images are re-ranked using a Bayes posterior estimator trained on the text surrounding the image and meta data features, and then the top-ranked images are improved by eliminating drawings and symbolic images using an SVM classifier previously trained with a hand labeled dataset.

Our unsupervised object recognition approach differs in three main aspects from previous work. It uses a smaller sample set of images which is expanded via transformations; this reduces the risk of introducing irrelevant images. It includes a learning mechanism that creates various visual models of the same category to deal with intraclass variability. It automatically weights the contribution of local and global features according to the object characteristics, as well as the rest of the model parameters via cross-validation.

3 Methodology

The general outline of the procedure for building a classification model for an object (specific or category) is the following (see Fig. 1):

1. A set of C training images for the *concept* to be learned are retrieved from the Web using Google Images.

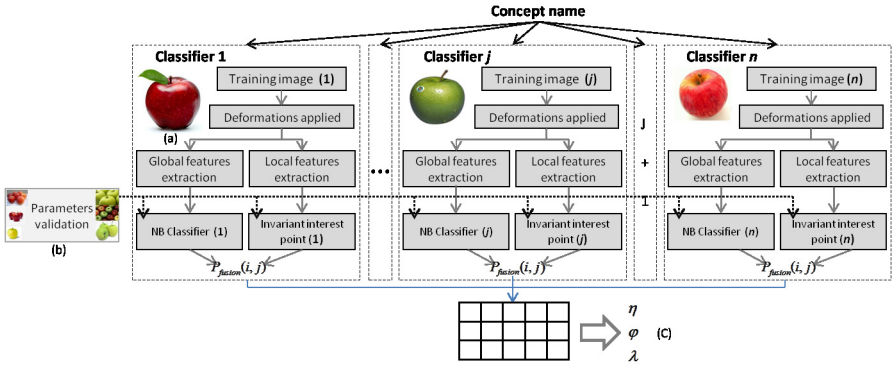

Fig. 1. Schematic representation of the object recognition system. A set of n classifiers is built, one for each sample image, generating d new images from each sample by applying different transformations. Global and local features are extracted from each image, and based on these two classifiers are trained, and combined. The parameters of the model are obtained via cross validation.

2. A series of transformations are applied to each training image, obtaining $\mid C \mid$ sets of d images each.
3. Global (color and texture) and local (SIFT key-points) features are extracted from each image in the extended training set ($C \times d$ images).
4. Two classifiers are trained for each of the $\mid C \mid$ sets of d images, one with the global features and other based on SIFT descriptors, so in total $2 \times C$ classifiers are obtained.
5. The local and global classifiers for each of the $\mid C \mid$ subsets are integrated via a linear weighted combination.
6. The set of $\mid C \mid$ classifiers are combined using a voting scheme.
7. The model parameters –weights for the linear combination of local and global features, thresholds for each classifier, and threshold for the combination of classifiers– are determined via cross-validation for each object.

Image Transformations

Inspired by [13], we generate several images from each sample (model) image via different transformations. This provides a training set for building a classifier for each model image. It also allows to select robust local features that are invariant to these transformations. The transformations include: (i) additive Gaussian noise, (ii) salt and pepper noise, (iii) rotations, (iv) changes in scale, and (v) changes in intensity. Five different levels are used for each transformation, so in total $5 \times 5 = 25$ images are generated.

Object Representation – Global and Local Features

An object is represented by a combination of global and local features extracted from the training images.

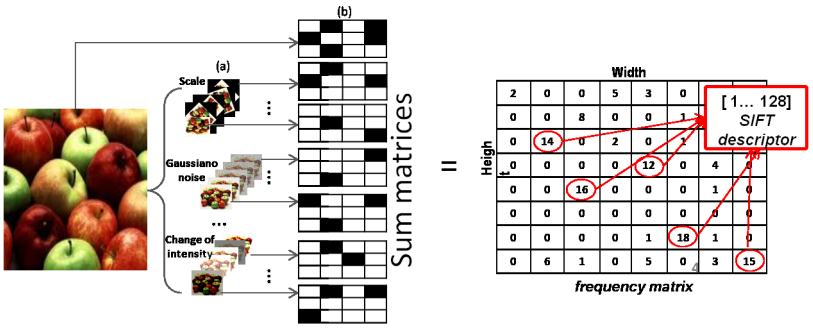

Fig. 2. SIFT key-points are obtained for the original image and the images obtained under the different transformations (scale, noise, intensity, etc.). The frequency of each point is collected in a *frequency matrix* and those points with higher frequency are selected as local features.

Global Features: Include color and texture information. To avoid the requirement of object segmentation we do not consider shape in this work. As color features we combine three different color models – RGB, HSV and CieLab–, each is represented via a normalized histogram with 20 bins per channel.

As texture descriptors we used a Grey Level Co-occurrence Matrix (GLCM) and Gabor Filters. A GLCM is obtained for 4 different angles (0 , $\pi/4$, $\pi/2$ and $3\pi/4$), and for each one several statistical descriptors are obtained –contrast, correlation, energy and homogeneity. This gives $4 \times 4 = 16$ features. Gabor filters [5] are applied at the same 4 angles and with two different wavelengths, so in total there are 8 filters; for each one we obtain the mean and variance. This gives another $8 \times 2 = 16$ features.

The global feature vector has $3 \times 3 \times 20 = 180$ color descriptors and $16 + 16 = 32$ texture descriptors, which gives a total of 212 features. Considering that in the sample images usually the object of interest is approximately centered, we restrict the calculation of the global features to a central window of the image.

Local Features: As local features we use the SIFT descriptor [9]. The SIFT features are obtained for a sample image and all the transformations (25 images). For each SIFT point we count the number of repetitions in the set of images, and we select those that are preserved in at least υ transformed images (in the experiments we set $\upsilon = 5$). To detect the matching SIFT points across the modified images, we obtain the coordinates of each key-point in the original model image, and these coordinates are geometrically mapped to the other images according to the corresponding transformation. With this process we obtain a set of *robust* SIFT descriptors for each model image, stable against affine transformations, noise and illumination variations. This process is illustrated in Fig. 2.

Classifiers
We build two classifiers for each model image, one based on the global features and other based on the local features, which are then integrated via a linear weighted combination.

Global Classifier: For the global features we use a Naïve Bayes classifier (NB):

$$P_{Global}(V_i, C_j) = P(C_j) \prod_{k=1}^{z} P(F_{ik} = f_{ik} | C_j) \tag{1}$$

where f_{ik} is the k feature of image V_i. $P_{Global}(V_i, C_j)$ gives the probability of concept C_j in image V_i given the z global features.

Local Classifier: For the local features we estimate the probability of concept C_j in image V_i based on the number of matching key-points between the model and test images ($\#matches$). This probability, $P_{Local}(V_i, C_j)$, is estimated as:

$$P_{Local}(V_i, C_j) = \begin{cases} 1 - \frac{1}{\#matches+1}, & \#matches > 0 \\ 0.001, & \#matches = 0 \end{cases} \tag{2}$$

Classifier Combination: Local and global probabilities are combined via a weighted sum:

$$P_{fusion}(V_i, C_j) = \lambda P_{Global}(V_i, C_j) + (1 - \lambda) P_{Local}(V_i, C_j) \tag{3}$$

where λ is a parameter that gives different weight to the global or local features. A positive decision is obtained for classifier j if $P_{fusion}(V_i, C_j) > \eta$:

$$C_{decision}(V_i, C_j) = \begin{cases} 1, & P_{fusion}(V_i, C_j) \geqslant \eta \\ 0, & otherwise \end{cases} \tag{4}$$

This combined probability is obtained for each of the $\mid C \mid$ original training images. Thus, for a test image V_i, we obtain $P_{fusion}(V_i, C_j)$ for $j = 1.. \mid C \mid$. This process is depicted in Fig. 3. Finally, an object is recognized if at least φ classifiers give a positive classification:

$$R_{object}(V_i) = \begin{cases} 1, & \sum_{j=1}^{|C|} C_{decision}(V_i, C_j) \geqslant \varphi \\ 0, & otherwise \end{cases} \tag{5}$$

Parameter Adjustment: The model has 3 main parameters: (i) **Local-global Weight** ($\lambda : 0..1$) it determines the weight for the local vs. global features for each of the $\mid C \mid$ classifiers. (ii) **Classification Threshold** ($\eta : 0..1$) it sets the probability threshold so that each classifier gives a positive result if $P_{Global} > \eta$. (iii) **Recognition Threshold** ($\varphi : 1..|C|$) global threshold for combining the C classifiers, an object is recognized if at least φ classifiers give a positive classification.

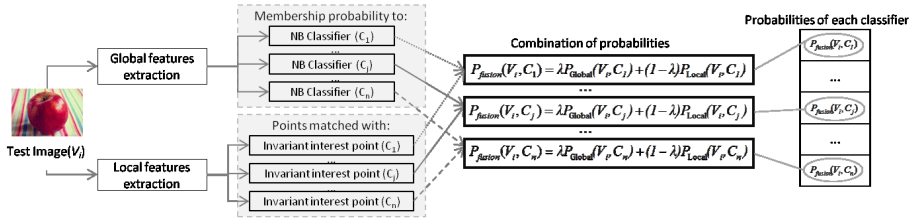

Fig. 3. For each test image the global and local features are extracted and from these the local and global probabilities are estimated based on the corresponding classifier, and then combined. This is repeated for each of the model images, obtaining $\mid C \mid$ probabilities.

The values of these parameters are automatically obtained by cross-validation in a set of validation images (also retrieved from the Web) to maximize the accuracy in the validation set.

4 Experiments and Results

We evaluated the proposed method in the recognition of some of the objects used in SRVC [1]. A robot is given several concepts that correspond to specific objects or object categories, and it has to search in the Web and build a model to recognize the objects in its environment. Thus, this competition provides a good testbed for our method. However, we can not make a direct comparison with competitors of SRVC, as they perform their tests in real environments[1] and we on images extracted from the Web.

We consider 10 objects from the SRVC, 5 specific (Colgate Total, Pepsi bottle, Coca-Cola can, Shrek DVD and Ritz crackers) and 5 categories (apple, banana, frying pan, white soccer ball, and eyeglasses). For each concept we used 12 positive examples for training obtained with Google Images, and 26 negative examples (images of indoor environments considering that a robot will search for the objects in this type of scenarios). Another 6 positive and negative examples (obtained from Web) are used as the validation set for tuning the parameters of the method. We used *precision, recall* and *accuracy* to quantitatively evaluate the performance of our method.

We performed two experiments. In Experiment 1 we used as training images the first 12 images returned by Google Images. For Experiment 2 we have a semisupervised scenario, where a person selects the training images among those 50 returned by the search engine. This could be a reasonable approach in some applications where a human can give some feedback to the robot (similar to intermediate relevance feedback). For testing we consider 40 new examples, 20 positive and 20 negative retrieved with Google Images (selected by a user).

[1] Their best reported results are in the order of 40% for object recognition in the competition.

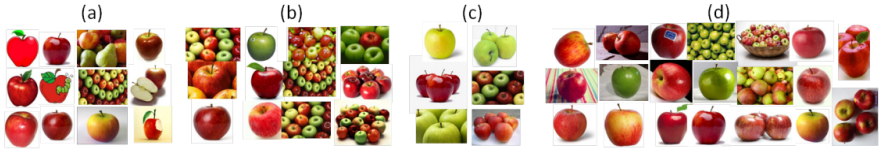

Fig. 4. Summary of apple model. (a) images used for the experiment 1: automatic selection. (b) images used for the experiment 2: semi-supervised. (c) examples used as validation. (d) images of test for model.

We compared our method against three different base classifiers. First, building a classifier using only the global information from the training set. Second, storing all the local information (SIFT features) from all the training set and checking if a testing image matches any of the stored features. Third, combining local and global information with equal weight to both features. The results are summarized in Table 1.

From this table, it can be seen that the proposed approach very clearly outperforms the other base classifiers in terms of precision and accuracy. So it is evident that combining local and global features with an adjusted weight, depending on the characteristics of the images, and generating artificial images proves to be beneficial.In terms of the results for recall, although our proposed approach presents lower results, it should be noted that this can be explained as some results from the other classifiers are close to 100% but with a precision close to 50%, which means that they are building trivial classifiers that accept every image as positive.

As expected, the results for specific objects are higher (89.5% in Accuracy) than for general objects (78% in Accuracy). Also, better results are obtained

Table 1. Experimental results for the testing set in both scenarios. Using only global information (G), using only local information (L), combining both with equal weight (GL), and the proposed approach (GLM).

Experiment 1: Automatic selection

	Apple				**Banana**				**Eyeglasses**				**Frying pan**				**White soccer ball**				**Average**			
	G	L	GL	GLM	G	L	GL	GLM	G	L	GL	GLM	G	L	GL	GLM	G	L	GL	GLM	G	L	GL	GLM
Precision	57	60	52	64	70	57	57	69	59	51	58	89	53	46	52	71	60	51	52	80	60	53	54	**75**
Recall	100	90	55	55	70	20	75	80	95	90	100	80	100	80	100	75	85	76	100	80	**90**	71	86	74
Accuracy	63	65	52	63	70	53	60	72	65	52	65	85	55	43	55	73	65	51	55	80	64	53	57	**75**
	Coca Cola can				**Colgate Total**				**DVD Shrek**				**Pepsi bottle**				**Ritz crackers**				**Average**			
	G	L	GL	GLM	G	L	GL	GLM	G	L	GL	GLM	G	L	GL	GLM	G	L	GL	GLM	G	L	GL	GLM
Precision	54	88	62	89	51	78	74	100	63	76	60	93	54	64	60	80	54	85	64	100	55	78	64	**92**
Recall	100	75	100	85	100	90	100	90	100	95	100	70	100	55	100	60	100	85	100	90	**100**	80	**100**	79
Accuracy	58	82	70	88	53	82	82	95	70	83	67	83	58	62	67	73	58	85	72	95	59	79	72	**87**

Experiment 2: semi-supervised

	Apple				**Banana**				**Eyeglasses**				**Frying pan**				**White soccer ball**				**Average**			
	G	L	GL	GLM	G	L	GL	GLM	G	L	GL	GLM	G	L	GL	GLM	G	L	GL	GLM	G	L	GL	GLM
Precision	50	57	52	77	54	48	54	71	51	75	55	100	53	57	54	74	50	55	51	81	52	58	53	**81**
Recall	85	100	100	85	100	85	100	100	100	75	100	80	100	100	100	70	95	90	100	86	96	90	**100**	84
Accuracy	50	63	55	80	58	47	57	80	53	75	60	90	55	63	57	72	50	58	52	83	53	61	56	**81**
	Coca Cola can				**Colgate Total**				**DVD Shrek**				**Pepsi bottle**				**Ritz crackers**				**Average**			
	G	L	GL	GLM	G	L	GL	GLM	G	L	GL	GLM	G	L	GL	GLM	G	L	GL	GLM	G	L	GL	GLM
Precision	51	76	66	100	54	75	83	100	56	83	76	89	57	61	62	89	54	86	68	95	54	76	71	**95**
Recall	100	80	100	80	100	90	100	95	100	95	95	85	100	40	100	80	100	95	100	100	**100**	80	99	88
Accuracy	53	77	75	90	58	80	90	98	60	88	82	88	63	57	70	85	57	90	78	98	58	78	79	**92**

Table 2. Experiments with the same objects reported in [16,17] in terms of F-measure. *sMIL* is the method reported in [17], and *Schoroff* is the method reported in [16].

	Airplane			**Guitar**			**Leopard**			**Motorbike**		
	GLM	sMIL	Schoroff	GLM	sMIL	Schoroff	GLM	sMIL	Schoroff	GLM	sMIL	Schoroff
F-measure	41	26	23	51	23	25	50	24	25	72	25	25
	Watch			**Car**			**Fase**					
	GLM	sMIL	Schoroff	GLM	sMIL	Schoroff	GLM	sMIL	Schoroff			
F-measure	60	26	25	52	25	n.a.	52	23	n.a.			

with the semi-supervised experiments (86.5%) than with the automatic selection (81%). We believe that these results (average global accuracy of 83.75%) are very promising as the system only receives the name of the object (with a possible selection of relevant images by the user) and has to recognize an unknown instance of that object in new images.

We should emphasize, that contrary to traditional datasets used in the many computer vision tasks, we are using images obtained directly from the Web.

We also performed experiments with the same dataset that was used in [17] and in [16] with images extracted as well from Google Images. Our approach was trained with the first 12 images (for each category) from Google Downloads [4] and tested with Caltech-7. We compared the performance of those two approaches with our own in terms of F-measure and using, as they did, 5-fold cross validation. For this experiment the 12 validation samples are obtained from Google Downloads. The results are shown in Table 2.

We can appreciate, from the results shown in the Table 2, that our approach is clearly superior to the other related work using their training images.

5 Conclusions and Future Work

Object recognition without the need of supervised training and considering different appearances is still a challenging problem. In this work we have presented an approach based on multiple-classifiers that can recognize an unknown instance of an object given only the name of the object. The proposed approach obtains a small set of images from the Web, creates variants of those images for robustness against changes in scale, rotation and noise, induces several classifiers to deal with polymorphism of objects, extracts global and local features and obtains an optimal combination to recognize different types of objects. We performed several experiments with specific and generic objects and compared the proposed approach against base classifiers and state-of-the-art systems with very promising results. As future work, we plan to implement the system on a mobile robot to test our approach in a real environment. We also plan to test the sensitivity of our method according to the number of model images.

References

1. The Semantic Robot Vision Challenge (2011), `http://www.semantic-robot-vision-challenge.org/`

2. Bunescu, R.C., Mooney, R.J.: Multiple Instance Learning for Sparse Positive Bags. In: 24th International Conference on Machine Learning (ICML 2007), pp. 105–112. ACM (2007)
3. Chang, L., Duarte, M., Sucar, L., Morales, E.: A Bayesian Approach for Object Classification Based on Clusters of SIFT Local Features. Expert Systems With Applications 39(2), 1679–1686 (2012)
4. Fergus, R., Fei-Fei, L., Perona, P., Zisserman, A.: Learning Object Categories from Google"s Image Search. In: 10th IEEE International Conference on Computer Vision (ICCV 2005), vol. 2, pp. 1816–1823. IEEE (2005)
5. Gabor, D.: Theory of Communication. Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering 93(26), 429–441 (1946)
6. Grimson, W., Huttenlocher, D.: On the Sensitivity of the Hough Transform for Object Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 12(3), 255–274 (1990)
7. Ji, Y., Idrissi, K., Baskurt, A.: Object Categorization Using Boosting Within Hierarchical Bayesian Model. In: 16th IEEE International Conference on Image Processing (ICIP 2009), pp. 317–320. IEEE (2009)
8. Lisin, D., Mattar, M., Blaschko, M., Learned-Miller, E., Benfield, M.: Combining Local and Global Image Features for Object Class Recognition. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005) - Workshops, vol. 03, pp. 47–54. IEEE (2005)
9. Lowe, D.: Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
10. Manshor, N., Rajeswari, M., Ramachandram, D.: Multi-Feature Based Object Class Recognition. In: International Conference on Digital Image Processing (ICDIP 2009), pp. 324–329. IEEE (2009)
11. Meger, D., Muja, M., Helmer, S., Gupta, A., Gamroth, C., Hoffman, T., Baumann, M., Southey, T., Fazli, P., Wohlkinger, W., Viswanathan, P., Little, J., Lowe, D., Orwell, J.: Curious George – An Integrated Visual Search Platform. In: Canadian Conf. on Computer and Robot Vision (CRV 2010), pp. 107–114. IEEE (2010)
12. Mikolajczyk, K., Schmid, C.: A Performance Evaluation of Local Descriptors. IEEE Transactions on Pattern Analysis & Machine Intelligence 27(10), 1615–1630 (2005)
13. Ozuysal, M., Calonder, M., Lepetit, V., Fua, P.: Fast Keypoint Recognition Using Random Ferns. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(3), 448–461 (2010)
14. Pereira, R., Lopes, L.S.: Learning Visual Object Categories with Global Descriptors and Local Features. In: Lopes, L.S., Lau, N., Mariano, P., Rocha, L.M. (eds.) EPIA 2009. LNCS, vol. 5816, pp. 225–236. Springer, Heidelberg (2009)
15. Rothganger, F., Lazebnik, S., Schmid, C., Ponce, J.: 3D Object Modeling and Recognition Using Affine-Invariant Patches and Multi-View Spatial Constraints. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2003), vol. 2, pp. 272–277. IEEE (2003)
16. Schroff, F., Criminisi, A., Zisserman, A.: Harvesting Image Databases from the Web. In: 11th International Conference on Computer Vision (ICCV 2007), pp. 1–8. IEEE (2007)
17. Vijayanarasimhan, S., Grauman, K.: Keywords to Visual Categories – Multiple-Instance Learning for Weakly Supervised Object Categorization. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2008), pp. 1–8. IEEE (2008)
18. Wang, Y., Hou, Z., Leman, K., Pham, N.T., Chua, T.W., Chang, R.: Combination of Local and Global Features for Near-Duplicate Detection. In: Lee, K.T., et al. (eds.) MMM 2011 Part I. LNCS, vol. 6523, pp. 328–338. Springer, Heidelberg (2011)