Dataset Issues in Object Recognition

J. Ponce[1,2], T.L. Berg[3], M. Everingham[4], D.A. Forsyth[1], M. Hebert[5], S. Lazebnik[1], M. Marszalek[6], C. Schmid[6], B.C. Russell[7], A. Torralba[7], C.K.I. Williams[8], J. Zhang[6], and A. Zisserman[4]

[1] University of Illinois at Urbana-Champaign, USA
[2] Ecole Normale Supérieure, Paris, France
[3] University of California at Berkeley, USA
[4] Oxford University, UK
[5] Carnegie Mellon University, Pittsburgh, USA
[6] INRIA Rhône-Alpes, Grenoble, France
[7] MIT, Cambridge, USA
[8] University of Edinburgh, Edinburgh, UK

Abstract. Appropriate datasets are required at all stages of object recognition research, including learning visual models of object and scene categories, detecting and localizing instances of these models in images, and evaluating the performance of recognition algorithms. Current datasets are lacking in several respects, and this paper discusses some of the lessons learned from existing efforts, as well as innovative ways to obtain very large and diverse annotated datasets. It also suggests a few criteria for gathering future datasets.

1 Introduction

Image databases are an essential element of object recognition research. They are required for learning visual object models and for testing the performance of classification, detection, and localization algorithms. In fact, publicly available image collections such as UIUC [1], Caltech 4 [10], and Caltech 101 [9] have played a key role in the recent resurgence of category-level recognition research, driving the field by providing a common ground for algorithm development and evaluation. Current datasets, however, offer a somewhat limited range of image variability: Although the appearance (and to some extent, the shape) of objects does indeed vary within each class (e.g., among the airplanes, cars, faces, and motorbikes of Caltech 4), the viewpoints and orientations of different instances in each category tend to be similar (e.g., side views of cars taken by a horizontal camera in UIUC); their sizes and image positions are normalized (e.g., the objects of interest take up most of the image and are approximately centered in Caltech 101); there is only one instance of an object per image; finally, there is little or no occlusion and background clutter. This is illustrated by Figures 1 and 3 for the Caltech 101 database, but remains true of most datasets available today.

The problems with such restrictions are two fold: (i) some algorithms may exploit them (for example near-global descriptors with no scale or rotation invariance may perform well on such images), yet will fail when the restrictions

J. Ponce et al. (Eds.): Toward Category-Level Object Recognition, LNCS 4170, pp. 29–48, 2006.

Fig. 1. Sample images from the Caltech 101 dataset [9], courtesy of Fei-Fei Li

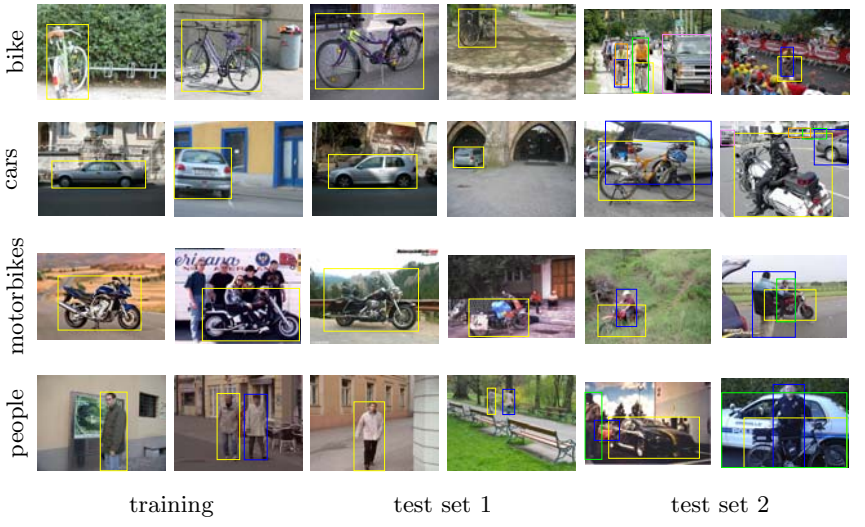

Fig. 2. Image examples with ground truth object annotation for different categories of the PASCAL 2005 challenge. The dataset may be obtained from http://www.pascal-network.org/challenges/VOC.

do not apply; and, related to this, (ii) the images are not sufficiently challenging for the benefits of more sophisticated algorithms (e.g., scale invariance) to make a difference. This means that progress in algorithm capability cannot be assessed. For example, multiple algorithms currently achieve close to 100% object vs. background classification accuracy on Caltech 4. There is a clear need for new datasets with more realistic and less restrictive image conditions: multiple object class instances within a single image, with partial occlusion (e.g., by other objects) and truncation (e.g., by the image edge), with size and orientation variations, etc. A first step in that direction has been taken with the datasets gathered for the PASCAL challenge, as illustrated by Figure 2. The rest of this chapter discusses some of the lessons learned from existing datasets such as Caltech 101 and those available under the PASCAL challenge. It also presents innovative ways to gather very large, annotated datasets from the World Wide Web, and concludes with some recommendations for future datasets, including a brief discussion of evaluation procedures.

2 Lessons Learned from Existing Datasets

2.1 The Caltech 101 Dataset

Most of the currently available datasets only contain a small number of classes, such as faces, pedestrians, and cars. A notable exception is the Caltech 101 database [9], with 101 object classes (Figure 1), which has become a de facto standard for evaluating algorithms for multi-class category-level recognition, and can be

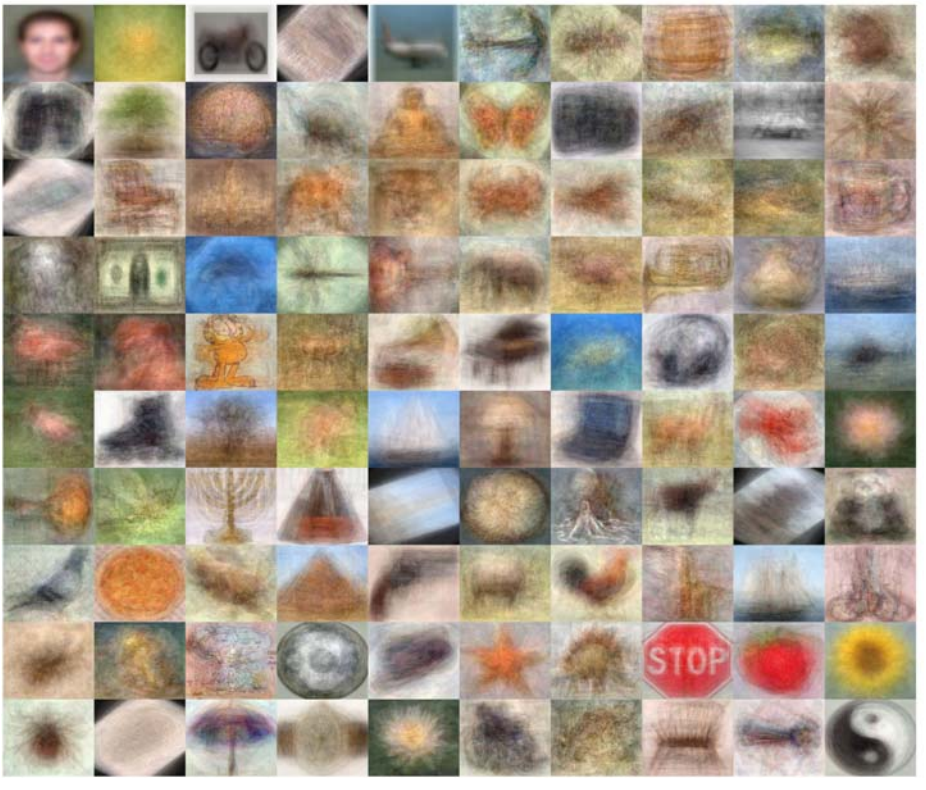

Fig. 3. The Caltech 101 average image

credited for a recent increase in efforts in this fundamental area of computer vision. Even though Caltech 101 is one of the most diverse datasets available today in terms of the amount of *inter-class* variability that it encompasses, it is unfortunately lacking in several important sources of *intra-class* variability. Namely, most Caltech 101 objects are of uniform size and orientation within their class, and lack rich backgrounds: This is demonstrated by the composite image shown in Figure 3, which was constructed by A. Torralba by averaging the RGB values of all the images for 100 of the object classes in the Caltech 101 dataset. The averaged images are computed by first resizing all the images to be 150×128 pixels and the intensity values of the final average are scaled to cover the range [0, 255]. They reveal regularities in the intensity patterns among all the images for each object category. If the images had a wide range of variations in object pose and object location, the resulting averages (before scaling the intensity values) would result in a (roughly) homogeneous field. This is clearly not the case, and many of the object classes are still easily recognizable by a human. Some of the characteristics of the dataset that are revealed by this experiment are that most images have little or no clutter, the objects tend to be centered in each image, and most objects are presented in a stereotypical pose.

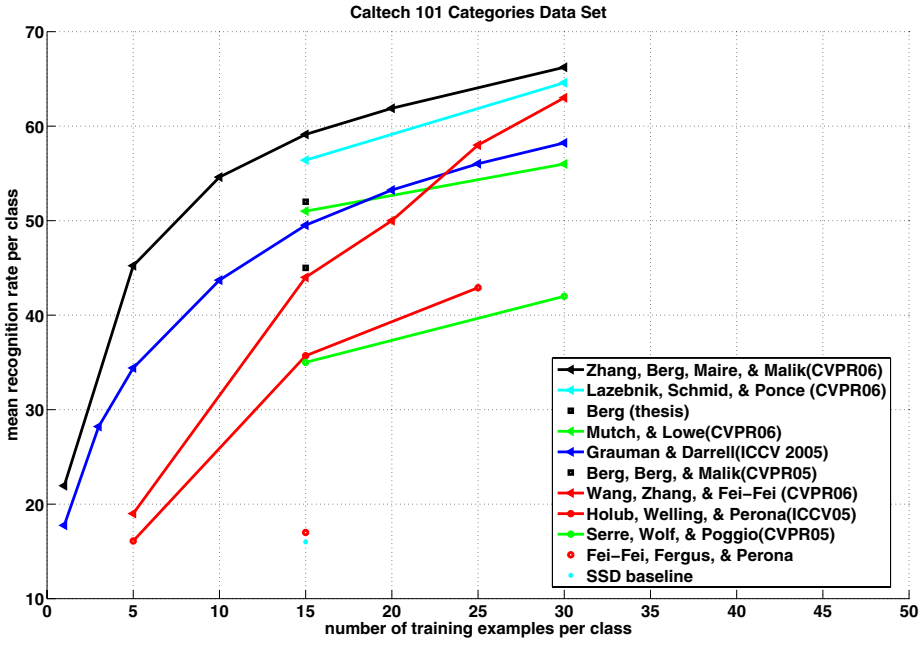

Fig. 4. A comparison of several algorithms on the Caltech 101 dataset [33], courtesy of H. Zhang

As noted earlier, despite its limitations, the Caltech 101 dataset has essentially become a de facto standard for multi-class recognition algorithms. Figure 4 shows the results of a comparative evaluation of several recent recognition algorithms on the Caltech 101 dataset [33], including those proposed by Fei-Fei *et al.* [9], Berg *et al.* [4], Grauman and Darrell [14], Holub *et al.* [17], Serre *et al.* [26] in 2005, and Berg [5], Lazebnik *et al.* [18], Mutch and Lowe [21], Ommer and Buhmann [22], Wang *et al.* [31], H. Zhang *et al.* [33] in 2006. The comparison also includes a baseline method comparing size-normalized greyscale images using correlation and nearest-neighbor classification [4].

We will not try to assess the merits of the different algorithms here. Instead, it is worth discussing what this comparison reveals about Caltech 101 as an evaluation tool. There are three clear trends: First, performance improves, as expected, with the number of training samples. Second, algorithms using SVMs as classifiers tend to do well, and include the two top performers [18,33]. Third, the classification rate steadily improves with time, from 17% in 2004 [9][1] to about 60% in 2006 [18,31,33]. None of these conclusions is very surprising, nor very telling about object recognition technology: The three methods achieving the best (and very close) performances for 30 training sample use totally different

[1] Very close to the 16% achieved by the baseline method — a reminder of the constant need for baseline comparisons.

models for image categories: a bag of features [33], a spatial pyramid [18], and a Bayesian model encoding both the appearance of individual features and the co-occurrence of feature pairs [31].

The steady improvement of classification rates over time apparent in this study is probably a sign that computer vision researchers are more and more adept at using and perhaps improving upon methods borrowed from statistical machine learning. It is not quite clear, however, that this performance increase *by itself* is a sign of progress toward better models for object categories and the recognition process. The (relatively) long-time public availability of image databases makes it possible for researchers to fine-tune the parameters of their recognition algorithms to improve their performance. Caltech 101 may, like any other dataset, be reaching the end of its useful shelf life.

2.2 The PASCAL Visual Object Classes Challenge

The first PASCAL[2] VOC (visual object classes) challenge ran from February to March 2005. The goal of the challenge was to recognize objects from a number of visual object classes in realistic scenes (i.e., not pre-segmented objects). Four object classes were selected: motorbikes, bicycles, cars, and people.

Both classification (where for each of the classes, the presence or absence of an object is predicted in the test image) and localization (where the algorithm must predict the bounding box and class of each object in the test image) were evaluated. A particular feature of the challenge was that two test sets were provided. For the first, images were assembled from a number of standard sources (e.g., the Caltech sets) and split randomly into training and test subsets with the same distribution of variability. Many algorithms already achieve very good performance on images of this difficulty; they have almost reached their peak performance. The second test set was designed to address this problem. It was assembled from new sources (Google image search, local photographs, etc.) with the intention of providing a harder test set with greater variability of scale, pose, background clutter and degree of occlusion, and assess the generalization ability of current algorithms. Needless to say, performance was inferior on the second test set. Twelve teams entered the challenge. Participants were provided with a development kit consisting of training and validation images, baseline algorithms, and evaluation software.[3] Figure 5 shows ROC curves for classification on the first and second test sets. The difference in performance is evident.

For the classification task, most participants used "global" methods in which a descriptor of the overall image content is extracted (such as a bag of words representation and a SVM classifier), which leaves the task of deciding which elements of the descriptor are relevant to the object of interest to the classifier.

[2] PASCAL stands for pattern analysis, statistical modelling and computational learning. It is the name of an EU Network of Excellence funded under the IST Program of the European Union.

[3] The development kit and test images are available from `http://www.pascal-network.org/challenges/VOC/`

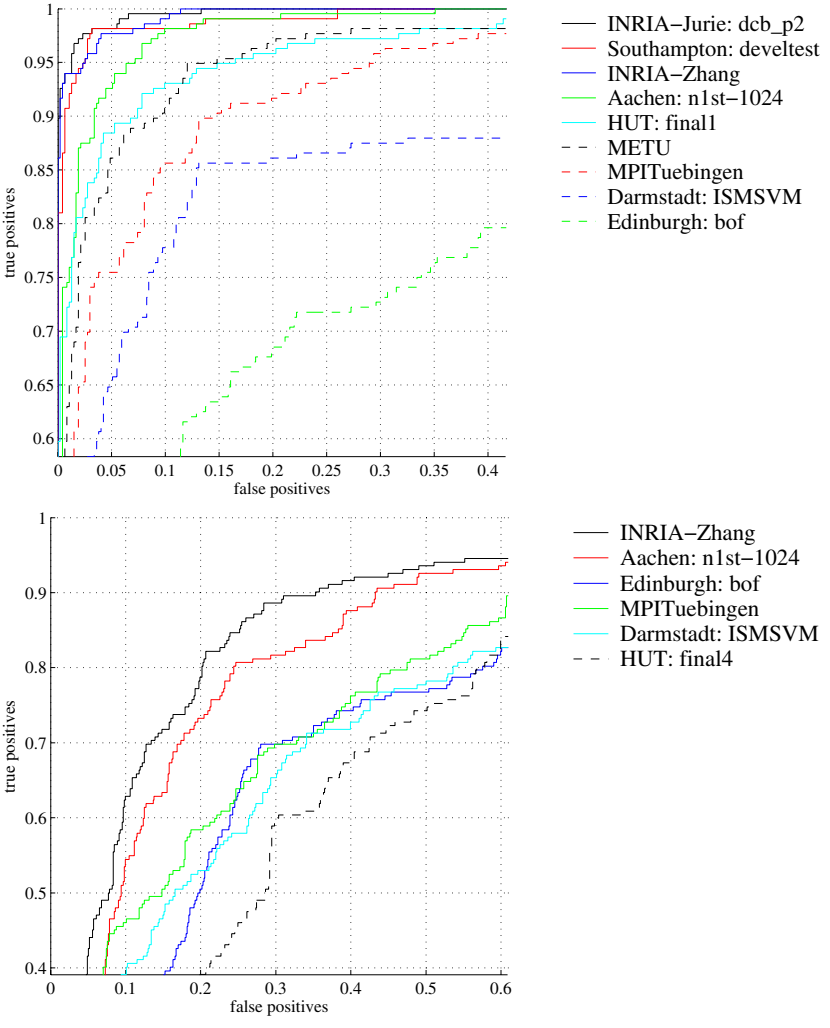

Fig. 5. PASCAL 2005 results. Top: ROC curves for classifying *motorbikes* images for test set 1 (where images taken from the same distribution of images as the training data). The best result in terms of EER (equal error rate) from each participant is shown, with curves ranked by decreasing EER. The axes cover a range equal to two times the maximum EER of the submitted results. Bottom: ROC curves for classifying *motorbikes* images for test set 2 (where images had more variability than the training data). The performance is inferior to that for test set 1.

All of these participants used only the class label attached to an image for training, ignoring additional annotation such as the bounding boxes of objects in the image. One possible advantage of "global" methods is that the image description captures information not only about the object of interest, e.g.,

a car, but also its context, e.g., the road. This contextual information might prove useful in recognizing some object classes; however, the risk is that the system may fail to distinguish the object from the context and thus show poor generalization to other environments, for example recognizing a car in a street vs. in a field. We return to this issue in the next section. In contrast, one participant (Technical University of Darmstadt) used a "classification by detection" approach which explicitly ignores all but the object, using bounding boxes or segmentation masks for training, and looking at local evidence for testing; this ensures that the method is modelling the object class of interest rather than statistical regularities in the image background, but may also fail to take advantage of contextual information.

Further details on the challenge, the tested algorithms, and the results of the evaluation can be found in [8].[4] The challenge is running again in 2006 with more classes (10) and a greater number and variability (in pose, partial occlusion) of images for each class. In the 2006 challenge, the classes are: bicycle, bus, car, motorbike cat, cow, dog, horse, sheep and people.[5] Figure 6 shows an average image for the 10 classes constructed in a similar manner to that shown for the Caltech 101 database in Figure 3. The color patterns are much more homogeneous in this case and the categories barely visible – providing some evidence of a greater image variability within each category of this set.

Images are obtained from three main sources: `flickr.com`, Microsoft Research Cambridge, and personal photographs. In the case of the flickr images the examples for each class are obtained by text search on the annotations, followed by manual inspection and annotation. In total there are 5,304 images, containing 9,507 annotated objects – each image may contain multiple objects from multiple classes, but all instances of the 10 classes are annotated. The data has been split into 50% for training/validation and 50% for testing. The distributions of images and objects by class are approximately equal across the training/validation and test sets. However, the level of difficulty is closer to that of test set 2 in the 2005 challenge.

2.3 The Importance of Context in Object Recognition Databases

In the PASCAL VOC challenge, several of the competing teams did quite well on full-image classification tasks. On the other hand, localization results were poor. As noted earlier, this suggests that background and contextual information may have played an important role in detection results. In other words, is it the object, or its background, which is recognized? J. Zhang *et al.* [34] have conducted a detailed study of this issue on the PASCAL dataset. We present in the rest of this section a summary of their findings.

A bag-of-features algorithm is used in the study. Like most modern approaches to category-level object detection, this algorithm does not attempt segmentation,

[4] Available at `http://www.pascal-network.org/challenges/VOC/voc2005/chapter.pdf`

[5] See `http://www.pascal-network.org/challenges/VOC/voc2006/` for additional details.

Fig. 6. The PASCAL 2006 average image. Each cell is an average over all images containing a particular object category (of 10). Figure courtesy of T. Malisiewicz and A. Efros. Other averages are available at `http://www.cs.cmu.edu/~tmalisie/pascal/means_trainval.html`

and uses both foreground and background features as input in both training and testing tasks. Briefly, an image is characterized by its scale-invariant Harris and Laplacian regions, along with their SIFT descriptors. Clustering is used to construct the image's *signature* formed by the centers of its clusters and their relative sizes. Support vector machines (SVMs) using the Earth Mover's Distance [24] as a kernel are then trained on each object category, and used for image classification [34].

PASCAL images are annotated with ground truth object regions, as shown in Figure 2. Foreground features (FF) can thus be identified as those located within the object region, while background features (BF) are those located outside. Many object categories have fairly characteristic backgrounds. For example, most of the car images contain a street, a parking lot, or a building. To determine whether this information provides additional cues for classification, let us examine the change in classification performance when the original background features from an image are replaced by two specially constructed alternative sets: *random* and *constant natural scene* backgrounds (referred to as *BF-RAND* and *BF-CONST*, respectively). BF-RAND samples are obtained by randomly shuffling background features among all of the images in the PASCAL dataset. For example, the background of a face image may be replaced by the background of a car image. Note that the total number of features and the relative amount of clutter in an image may be altered as a result of this procedure. BF-CONST examples consist of background features extracted from images captured by a fixed camera observing a natural scene over an extended period of time, so they include continuous lighting changes and the movement of trees and clouds (Figure 7).

Figure 8 (a)–(b) shows ROC curves obtained by training and testing on only the background features (BF) for test sets 1 and 2. In the case of test 1, it is clear that background features alone are often sufficient to determine the category of the image. This is not quite the case for test set 2. For example, BF features perform close to chance level for bicycles. Thus, one of the reasons why test set 2 is considered more difficult than test set 1, is the fact that its background

Fig. 7. Image examples of the constant natural scene background. They are captured with lighting changes and the movement of clouds and trees.

features are much less correlated with the foreground. The performance of the BF-RAND and BF-CONST feature sets (not shown on the figure) is at chance level as one would expect, since they do not contain any information about the foreground object class by construction.

Figure 8 (c)–(e) evaluates combinations of foreground features with different types of background features. Due to space limitations only results for the people test set 1 are presented. Results for the other test sets are similar. AF denotes all the features extracted from the original image, AF-RAND denotes the combination of FF and BF-RAND and AF-CONST denotes the combination of FF and BF-CONST. Figure 8 (c) shows ROC curves for a situation where training and testing are performed on the same feature combination. FF gives the highest results, indicating that object features play the key role for recognition, and recognition with segmented images achieves better performance than without segmentation. Mixing background features with foreground features *does not* give higher recognition rates than FF alone. For images with roughly constant backgrounds (AF-CONST), the performance is almost the same as for images with foreground features only. It is intuitively obvious that classifying images with fixed backgrounds is as easy as classifying images with no background clutter at all. Finally, the ROC curves for AF-RAND are the lowest, which shows that objects with uncorrelated backgrounds are harder to recognize.

Figure 8 (d) shows ROC curves for a setup where the training set has different types of backgrounds and the test set has its original background (AF). We can observe that training on AF or AF-RAND while testing on AF gives the highest results. Thus, even under randomly changed training backgrounds, the SVM can find decision boundaries that generalize well to the original training set. Training on FF or AF-CONST and testing on AF gives lower results, most likely because the lack of clutter in FF set and the monotonous backgrounds in AF-CONST cause the SVM to overfit the training set. By contrast, varying the object background during training, even by random shuffling, tends to increase the robustness of the learned classifier.

Finally, Figure 8 (e) shows ROC curves for a situation where the training set has the original backgrounds and the test set has different types of backgrounds. When the test set is "easier" than the training one, performance improves, and when it is "harder," the performance drastically drops. This is consistent with the results of Figure 8 (d), where training on the "harder" sets AF or AF-RAND gave much better results than training on the "easier" sets FF and AF-CONST.

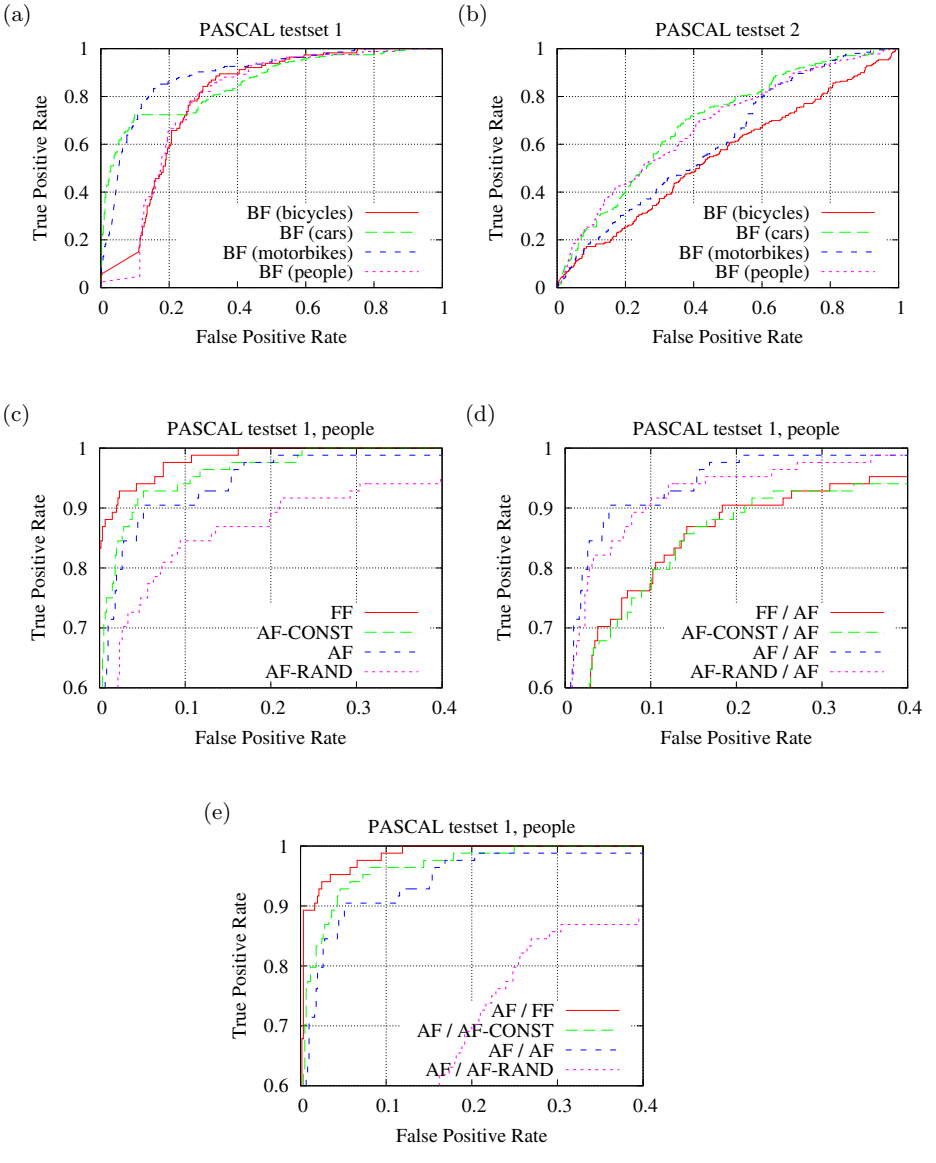

Fig. 8. ROC curves for the bag-of-features method of Zhang *et al.* [34] trained and tested on the PASCAL 2005 challenge dataset for different combinations of foreground and background features. (a)–(b): Training and testing on background features only. The left part of the figure corresponds to test set 1, and the right one to test set 2. (c)–(e): Training and testing using four combinations of foreground features with different types of background.

In conclusion, the evaluation of the role of background features in bag-of-keypoints classification highlights two important facts: First, while the backgrounds in most available datasets have non-negligible correlations with the foreground objects, using both foreground and background features for learning and recognition does not result in better performance, at least for the basic bag-of-features method evaluated by J. Zhang *et al.* This illustrates the limitations as evaluation platforms of datasets with simple backgrounds, such as CogVis [35], COIL-100 [36], and to some extent, Caltech 101 [9]: Based on the evaluation presented in this section, high performance on these datasets do not necessarily mean high performance on real images with varying backgrounds. Second, when the training set has different image statistics than the test set, it is usually beneficial to train on the most difficult dataset available, since the presence of varied backgrounds during training improves the generalization ability of the classifier. Note that these conclusions have been reached for the particular classifier used in the experiments, but similar trends are expected to hold for other bag-of-features methods that do not explicitly separate foreground from background features but use both for recognition at the same time. However, it is probable that such methods do not make the most effective use of the context provided by background features. The presence of background correlations may well improve the performance of methods that use contextual information to *prime* subsequent object detection and recognition stages [16,28].

3 Innovative Methods for Acquiring New Datasets

3.1 Web-Based Annotation

Web-based annotation tools provide a new way of building large annotated databases by relying on the collaborative effort of a large population of users [25,27,29,37]. Two examples are the ESP and Peekaboom internet games. ESP is an online game in which players enter labels describing the content of images [29]. ESP has been used with over 10 million labels for images collected from the Web. In a similar vein, the Internet game Peekaboom is designed to use "bored human intelligence" to label large image datasets with object, material, and geometry labels [30]. Peekaboom has been released to a general audience and it has already collected millions of data points. Its first task will be to label entire databases, such as Corel, which will be an enormous help to the object recognition community.

LabelMe is another online annotation tool that allows sharing of images and annotations [25]. The tool provides many functionalities such as drawing polygons, querying images, and browsing the database. Both the image database and all of the annotations are freely available. The tool runs on almost any Web browser, and includes a standard Javascript drawing interface that is easy to use (see Figure 9 for a screenshot). The resulting labels are stored in XML file format, which makes the annotations portable and easy to extend. A Matlab toolbox is available that provides functionalities for manipulating the database

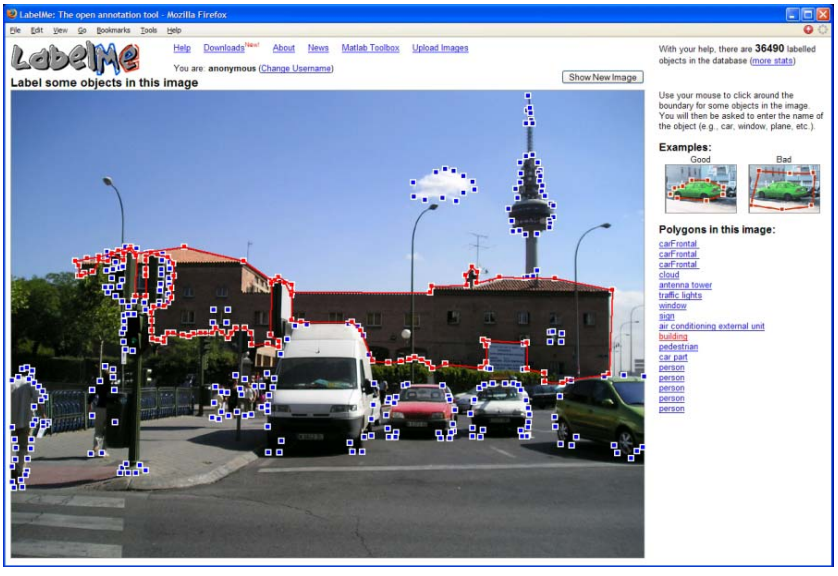

Fig. 9. Screenshot from the LabelMe labeling tool in use [25]. The user is presented with an image, possibly with one or more existing annotations in the image. The user has the option of annotating a new object, by clicking around the boundary of the object, or editing an existing annotation. The user can annotate an arbitrary number of objects in the image. Once finished, the user then clicks the "Show New Image" button to see a new image.

(database queries, communication with the online tool, image transformations, etc.). The database is also searchable online.

Currently the database contains more than 36,000 objects labeled within 6,000 images covering a large range of environments and several hundred object categories (Figure 10, left). The images are high resolution and cover a wide field of view, providing rich contextual information. Pose information is also available for a large number of objects. Since the annotation tool has been made available online there has been a constant increase in the size of the database, with about 5,000 new labels added every month, on average.

One important concern when data is collected using Web-based tools is quality control. Currently quality control is provided by the users themselves. Polygons can be deleted and object names can be corrected using the annotation tool online. Despite the lack of a more direct mechanism of control, the annotations are of quite good quality (Figure 10). Another issue is the complexity of the polygons provided by the users – do users provide simple or complex polygon boundaries? Figure 10 (right) illustrates the average number of points used to define each polygon for four object classes that were introduced using the Web annotation tool. These object classes are among the most complicated. These polygons provide a good idea of the outline of the object, which is sufficient for most object detection and segmentation algorithms.

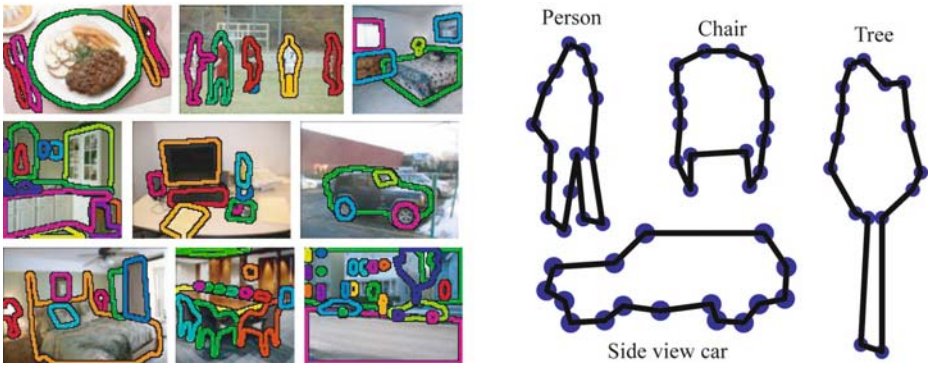

Fig. 10. Left: Examples of annotated images in the LabelMe database. The images cover a large range of scenes and object categories. Right: These polygons correspond to the average quality of the annotations for four object categories.

Another issue is what to label. For example, should you label a whole pedestrian, just the head, or just the face? What about a crowd of people – should you label all of them? Currently such decisions are left to each user, with the hope that the annotations will reflect what various people think are "natural" ways to segment an image. A third issue is the label itself. For example, should you call this object a "person", "pedestrian", or "man/woman"? An obvious solution is to provide a drop-down menu of standard object category names. Currently, however, people use their own descriptions, since these may capture some nuances that will be useful in the future. The Matlab toolbox allows querying the database using a list of possible synonyms.

3.2 Data Collection as Recognition

There is a plentiful supply of images available at the typing of a single word using Internet image search engines such as Google, and we discuss now two methods for obtaining object class images from this source. The first method uses Google image search as its source, the second uses web pages directly.

Starting from Image Search. Internet image search engines currently do not search directly on image visual content but instead use the image name and surrounding text. Consequently, this is not a source of pure images without filtering: for example, a Google image search for "monkey" yields only 30 actual monkey pictures in the first 100 results. Many of the returned images are visually unrelated to the intended category, perhaps arising from polysemes (e.g. "kite" can be kite-bird or kite-flying-toy). Even the small proportion of retrieved images that do correspond to the category are substantially more demanding than images in typical training sets (such as Caltech) – the number of objects in each image is unknown and variable, and the pose (visual aspect) and scale are uncontrolled.

Fergus *et al* [12] have proposed an unsupervised clustering method for extracting the "good" images from raw Google output. Each image is first described by a bag of keypoints/visual words. The clustering into visually coherent components is then achieved by applying probabilistic Latent Semantic Analysis (pLSA) [15] – a technique from the field of textual analysis originally developed for topic discovery in text corpus.

There then remains the problem of determining which of the clusters corresponds to the true object images. This problem could be solved by manual intervention, but Fergus *et al* [12] instead build a validation set automatically by noting that the first few images returned by Google tend to contain more good images than those returned later on. Using Google's automatic translation, the user's keyword is translated into a number of languages; the first few images are automatically downloaded, and combined to give a validation set of a reasonable size without degradation in quality.

In this manner, starting with Google image search, followed by clustering of visual words and automatic selection of the correct cluster using a generated validation set, image datasets can be generated using just the object's name. The main limitation of this approach is the effectiveness of the original image search engine. This limitation can be overcome by a semi-automatic method, as presented next.

Starting from Text Search. There are currently more than 8,168,684,336 Web pages on the Internet.[6] A search for the term "monkey" yields 36,800,000 results using Google text search. There must be a large number of images portraying "monkeys" within these pages, but retrieving them using Google image search is not successful, as described above. It has been known for a while that textual and visual information could effectively be combined in tasks such as clustering art [2], labeling images [3,13,19], or identifying faces in news photographs [6]. In these cases, an explicit relationship between words and pictures is given by image annotations, or photograph and video captions. On Web pages, however, the link between words and pictures is less clear. Berg and Forsyth consider in [7] the problem of combining the text and image information stored on Web pages to re-rank Google search results for a set of queries (see [11,32] for related work). They focus on *animal* categories because these provide rich and difficult data, often taking on a wide variety of appearances, depictions, and aspects, thus providing a good yardstick for demonstrating the benefits of exploiting the visual and textual information contained in Web pages.

In the method described in [7], a set of images is first obtained by text search. For example, 9,320 Web pages are collected using Google text search on 13 queries related to monkeys. From these pages, 12,866 distinct images of sufficiently large size (at least 120×120 pixels) are selected for further consideration. Of these images 2,569 are actual monkey images. The algorithm proceeds in two stages: first a set of visual exemplars (exemplars for short) is selected using only text-based information. Then, in the second stage, visual and textual cues are combined to rank the downloaded images, with monkey images being highly ranked.

[6] Google's last released number of indexed Web pages.

In the first stage Latent Dirichlet Allocation (LDA) is applied to the words contained in the Web pages to discover a set of latent topics for each category. These latent topics provide a distribution over words and are used to select highly likely words for each topic. Images are ranked according to their nearby word likelihoods and a set of 30 exemplars selected for each topic. As mentioned earlier, words and images can be ambiguous (e.g. "alligator" could refer to "alligator boots" or "alligator clips" as well as the animal). Currently there is no known method for breaking this polysemy-like phenomenon automatically. Therefore, at this point the user is asked to label each topic as relevant or background, depending on whether the associated images and words illustrate the category well. Using this labeling all selected topics are merged into a relevant topic and all unselected topics are merged into a background topic (and their exemplars and likely words are similarly pooled).

In the second stage, each image in the downloaded dataset is ranked according to a voting method using the knowledge base collected in the training stage. Voting uses image information in the form of shape, texture and color features as well as word information based on words on the associated pages that are located near the image.

Because exemplar-based voting incorporates multiple templates per category, it allows image retrieval across different poses, aspects, and even species. The top results returned by the composite classifier are in general quite good [7]: 81% of the top 500 images, and 69% of the top 1000 are correct (Figure 11). Most importantly, the images classified using both textual and visual features make up a very large, high-quality dataset for further object recognition research.

4 Recommendations

Research on object detection and recognition in cluttered scenes requires large image and video collections with ground truth labels. The labels should provide information about the object classes present in each image, as well as their shape and locations, and possibly other attributes such as pose. Such data is useful for testing, as well as for supervised learning. Even algorithms that require little supervision need large databases with ground truth to validate the results.[7] New algorithms that exploit context for object recognition [16,28] require databases with many labeled object classes embedded in complex scenes. Such databases should contain a wide variety of environments with annotated objects that co-occur in the same images. Future databases should exercise the ability of recognition systems to handle intra-class variability, varying size and pose, partial occlusion, and contextual cues. They should also display different levels of difficulty, including restricted viewpoints (e.g., cars will only be seen more or less from the side), reasonable levels of occlusion and viewpoint variation (e.g., cars from all viewing angles), a higher degree of intra-class variability (chairs,

[7] This is not to say that the annotations of future databases will be perfect: We expect that segmentations may be inaccurate, and labellings questionable, but scale will probably rescue us.

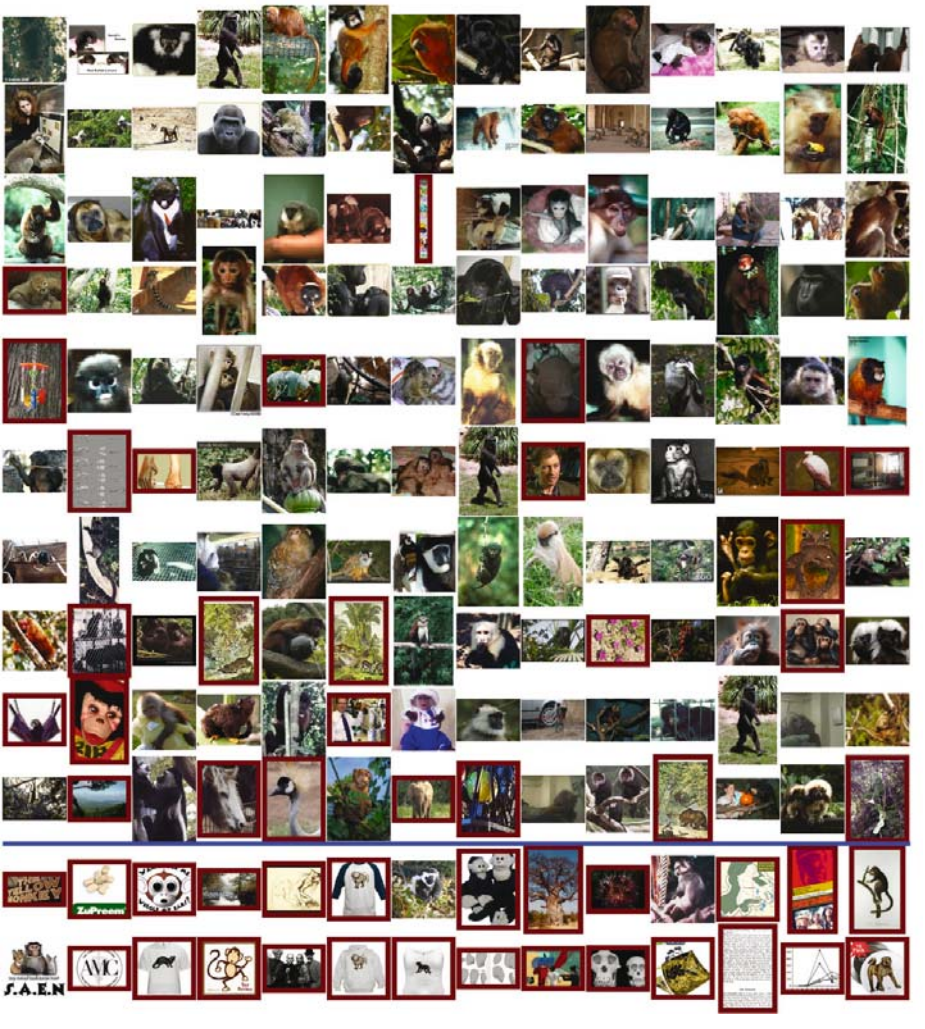

Fig. 11. Example of a "monkey" dataset generated semi-automatically, starting from a keyword search [7]. False positives are shown with a heavy border (dark red). The first 10 rows are sampled every 4^{th} image from the top 560 results, while the last two rows are sampled every 250^{th} image from the last 5,000-12,866 results. The dataset is quite accurate, with a precision of 81% for the top 500 images, and a precision of 69% for the top 1000 images. Deciding what images are relevant to a query doesn't have a single interpretation. Here, primates like apes, lemurs, chimps and gibbons have been included, but monkey figurines, people, monkey masks and monkey drawings have been excluded. The results include a huge range of aspects and poses as well as a depictions in different settings (e.g. trees, cages and indoor settings). The animal image classifiers inherently take advantage of the fact that objects are often correlated with their backgrounds ("monkeys" are often in trees and other greenery).

churches, clocks, etc.), and classes which share "parts" and might thus be confused (e.g., bikes/motorbikes, cars/lorries, etc.). Constructing large, annotated datasets featuring this type of variability will be a difficult and time-consuming task. It should also be one of the priorities of the object recognition community. We believe that innovative approaches to data collection such as those discussed in the previous section will play a major role in fulfilling this objective. A serious notion of object and/or scene category would help the data collection/organization process, since good/bad choices make problems easier/harder, and we do not know how to model this effect.

A very important issue that has not not been addressed in this chapter is the need of rigorous evaluation protocols for recognition algorithms on standard datasets. Standard performance measures for information retrieval, such as *interpolated average precision*, have been defined by the Text REtrieval Conference (TREC), and the object recognition community would probably be well advised to follow that example.[8] Further, the (rather typical) restriction of experiments to selected parts (e.g., "easy" or "hard" pictures) of the training and/or test set may bias the evaluation of a given method. See Müller, Marchand-Maillet, and Pun [20] for a discussion of this problem in the context of image retrieval.[9] In multi-class recognition tasks, gathering statistics over all test images instead of averaging them over categories may also bias the results when there are many more pictures for some "easy" classes than for some "hard" ones (this is the case for the Caltech 101 dataset for example). As discussed by Philips and Newton [23] in the face recognition domain, it is actually possible for some datasets to *predict* the performance of new algorithms from that of simple baseline methods (e.g., PCA plus nearest-neighbor classification, a.k.a. "eigenfaces"). This indicates that face recognition experiments often test the difficulty of a dataset instead of the effectiveness of new techniques. Conducting such a *meta-analysis* of category-level object recognition algorithms could prove to be fruitful. In this context, the "hardness" of different datasets is not well understood, and a good pool of baseline methods would help. Designing and implementing tools for testing specific aspects of recognition algorithms (e.g., robustness to viewpoint or illumination changes, or to within-class shape or texture variations), and correlating evaluation results across different standard datasets would also be extremely useful.

Acknowledgments. Many thanks to Fei-Fei Li for providing Figure 1, H. Zhang for providing Figure 4, and T. Malisiewicz and A. Efros for providing Figure 6. This work was supported in part by the National Science Foundation under grants IIS-0308087 and IIS-0535152; the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778; the French ACI project MoViStaR; the UIUC-CNRS-INRIA collaboration agreement; and

[8] Initial efforts are under way. See `http://trec.nist.gov/pubs/trec10/appendices/measures.pdf` and `http://www-nlpir.nist.gov/projects/trecvid/`.

[9] The discussion in this paper focuses on the Corel database, which is widely (and perhaps unwisely—see the comments in the article) used in the image retrieval community.

the Lava European project. M. Marszalek was supported by the INRIA student exchange program and a grant from the European Community under the Marie Curie Visitor project. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Science Foundation or European Community.

References

1. Agarwal, S., Roth, D.: Learning a sparse representation for object detection. In: Proc. European Conf. Comp. Vision. Volume LNCS 2353., Copenhagen, Denmark (2002) 113–127
2. Barnard, K., Duyguly, P., Forsyth, D.: Clustering art. In: Proc. IEEE Conf. Comp. Vision Patt. Recog. Vol. II (2001) 435–439
3. Barnard, K., Duygulu, P., Forsyth, D., de Freitas, N., Blei, D., Jordan, M.: Matching words and pictures. Journal of Machine Learning Research **3** (2003), 1107–1135
4. Berg, A., Berg, T.L., Malik, J.: Shape matching and object recognition using low distortion correspondence. In: Proc. IEEE Conf. Comp. Vision Patt. Recog. Vol. II (2005) 435–439
5. Berg, A.C.: Phd thesis. (To appear.)
6. Berg, T.L., Berg, A.C., Edwards, J., Forsyth, D.: Who's in the picture? In: Proc. Neural Inf. Proc. Syst. (2004)
7. Berg, T.L., Forsyth, D.: Animals on the Web. In: Proc. IEEE Conf. Comp. Vision Patt. Recog. (2006)
8. Everingham, M., Zisserman, A., Williams, C., Van Gool, L., Allan, M., Bishop, C., Chapelle, O., Dalal, N., Deselaers, T., Dorko, G., Duffner, S., Eichhorn, J., Farquhar, J., Fritz, M., Garcia, C., Griffiths, T., Jurie, F., Keysers, D., Koskela, M., Laaksonen, J., Larlus, D., Leibe, B., Meng, H., Ney, H., Schiele, B., Schmid, C., Seemann, E., Shawe-Taylor, J., Storkey, A., Szedmak, S., Triggs, B., Ulusoy, I., Viitaniemi, V., Zhang, J.: The 2005 PASCAL visual object classes challenge. In: Selected Proceedings of the First PASCAL Challenges Workshop. LNAI, Springer-Verlag (2006)
9. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. In: Proc. IEEE Conf. Comp. Vision Patt. Recog Workshop on Generative-Model Based Vision. (2004)
10. Fergus, R., Perona, P., Zisserman, A.: Object class recognition by unsupervised scale-invariant learning. In: Proc. IEEE Conf. Comp. Vision Patt. Recog. Vol. II (2003) 264–271
11. Fergus, R., Perona, P., Zisserman, A.: A visual category filter for Google images. In: Proc. Europ. Conf. Comp. Vision. (2004)
12. Fergus, R., Fei-Fei, L., Perona, P., Zisserman, A.: Learning object categories from Google's image search. In: Proc. Int. Conf. Comp. Vision. (2005)
13. Giridharan, I., Duygulu, P., Feng, S., Ircing, P., Khudanpur, S., Klakow, D., Krause, M., Manmatha, R., Nock, H., Petkova, D., Pytlik, B., Virga, P.: Joint visual-text modeling for automatic retrieval of multimedia documents. In: Proc. ACM Multimedia Conference. (2005)
14. Grauman, K., Darrell, T.: The pyramid match kernel: Discriminative classification with sets of image features. MIT-CSAIL-TR-2006-020 (2006). Updated version of the ICCV'05 paper with the same title, featuring the improved results shown in Fig. 4.

15. Hofmann, T.: Unsupervised learning by probabilistic latent semantic analysis. Machine Learning **43** (2001) 177–196
16. Hoiem, D., Efros, A., Hebert, M.: Geometric context from a single image. In: Proc. Int. Conf. Comp. Vision. (2005)
17. Holub, A., Welling, M., Perona, P.: Combining generative models and fisher kernels for object class recognition. In: Proc. Int. Conf. Comp. Vision. (2005)
18. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: Proc. IEEE Conf. Comp. Vision Patt. Recog. (2005)
19. Li, J., Wang, J.: Automatic linguistic indexing of pictures by a statistical modeling approach. PAMI (2003) **25**(9), 1075–1088
20. Müller, H., Marchand-Maillet, S., Pun, T.: The truth about Corel – Evaluation in image retrieval. In: The Challenge of Image and Video Retrieval (CIVR2002), London, UK (2002)
21. Mutch, J., Lowe, D.: Multiclass object recognition using sparse, localized features. In: Proc. IEEE Conf. Comp. Vision Patt. Recog. (2006)
22. Ommer, B., Buhmann, J.M.: Learning compositional categorization models. In: Proc. Europ. Conf. Comp. Vision. (2006)
23. Philips, P., Newton, E.: Meta-analysis of face recognition algorithms. In: Int. Conf. on Automatic Face and Gesture Recognition. (2002)
24. Rubner, Y., Tomasi, C., Guibas, L.: The Earth Mover's distance as a metric for image retrieval. International Journal of Computer Vision **40**(2) (2000) 99–121
25. B. C. Russell, A. Torralba, K.P.M., Freeman, W.T.: LabelMe: a database and web-based tool for image annotation. Technical report, MIT, AI Lab Memo AIM-2005-025 (2005)
26. Serre, T., Wolf, L., Poggio, T.: object recognition with features inspired by visual cortex. In: Proc. IEEE Conf. Comp. Vision Patt. Recog. (2005)
27. Stork, D.: The open mind initiative. IEEE Intelligent Systems and Their Applications **14**(3) (1999) 19–20
28. Torralba, A.: Contextual priming for object detection. International Journal of Computer Vision **53**(2) (2003) 153–167
29. von Ahn, L., Dabbish, L.: Labeling images with a computer game. In: Proc. ACM Conf. Hum. Factors Comp. Syst. (CHI). (2004)
30. von Ahn, L., Liu, R., Blum, M.: Peekaboom: A game for locating objects in images. In: Proc. ACM Conf. Hum. Factors Comp. Syst. (CHI). (2006)
31. Wang, G., Zhang, Y., Fei-Fei, L.: Using dependent regions for object categorization in a generative framework. In: Proc. IEEE Conf. Comp. Vision Patt. Recog. (2006)
32. Yanai, K., Barnard, K.: Probabilistic web image gathering. In: Workshop on MIR. (2005)
33. Zhang, H., Berg, A.C., Maire, M., Malik, J.: SVM-KNN: Discriminative nearest neighbor classification for visual category recognition. In: Proc. IEEE Conf. Comp. Vision Patt. Recog. (2006)
34. Zhang, J., Marszalek, M., Lazebnik, S., Schmid, C.: Local features and kernels for classification of texture and object categories: An in-depth study. Technical Report RR-5737, INRIA Rhône-Alpes (2005)
35. (http://www.vision.ethz.ch/projects/cogvis/CogVis-images/image-samples.html)
36. (http://www1.cs.columbia.edu/CAVE/research/softlib/coil-100.html)
37. (http://www.flickr.com)