Filtering Internet Image Search Results
Towards Keyword Based Category Recognition

Kamil Wnuk Stefano Soatto
Computer Science Department
University of California, Los Angeles
Los Angeles, CA 90095, USA
{kwnuk,soatto}@cs.ucla.edu

Abstract

In this work we aim to capitalize on the availability of Internet image search engines to automatically create image training sets from user provided queries. This problem is particularly difficult due to the low precision of image search results. Unlike many existing dataset gathering approaches, we do not assume a category model based on a small subset of the noisy data or an ad-hoc validation set. Instead we use a nonparametric measure of strangeness *[8] in the space of holistic image representations, and perform an iterative feature elimination algorithm to remove the most* strange *examples from the category. This is the equivalent of keeping only features that are found to be consistent with others in the class. We show that applying our method to image search data before training improves average recognition performance, and demonstrate that we obtain comparative precision and recall results to the current state of the art, all the while maintaining a significantly simpler approach. In the process we also extend the strangeness-based feature elimination algorithm to automatically select good threshold values and perform filtering of a single class when the background is given.*

1. Introduction

The field of object and category recognition is inherently dependent on image datasets for classifier training and evaluation. Until very recently, such training and testing datasets were gathered and annotated manually. In addition to the high labor cost that goes into an endeavor like manual collection and annotation, other now known problems [14], such as artificially introduced artifacts, have resulted from manual image dataset collection.

Since a large majority of current recognition algorithms require at least a weak level of supervision, the quantity of available training and testing data must continue to grow with increasing algorithm capabilities. This is necessary in order for the field to continue making progress towards the lofty goal of someday scaling to human-level recognition.

To avoid manual collection and annotation, a number of vision researchers have recently turned to the Internet as a source of loosely annotated image data [2, 5, 6, 9, 15, 17]. Image search engines can provide up to thousands of images in response to a particular query, thus they seem like a promising starting point for automated dataset acquisition. The problem is that the precision of image search results is low, and retrieved results are therefore typically polluted with large amounts of noise. If this noise can be consistently overcome, the vision community would have a powerful tool to automatically collect representative image data for any desired keyword.

There are two clear starting points from where one can approach this noise reduction issue. The first approach is to design models and learning procedures to be robust to noisy training data. A classifier can then either be learned directly for recognition, or it can be applied back to the noisy data for the purpose of filtering or re-ranking, generating a dataset with improved precision for other algorithms to use.

A second approach is to first filter the data using discriminative methods, and then use the new noise reduced set of images to train any desired classifier for recognition.

It is important to note that capturing diversity is just as critical as obtaining high precision, since category objects may appear in distinct but common poses, and may vary significantly in appearance. It may also be the case that a single keyword may have multiple visual meanings, a situation known as *polysemy*. We argue that at the lowest level of observation, it is important to capture all consistent visual definitions corresponding to a particular keyword query. For example, one could imagine requesting that a hypothetical recognition system learn to detect the visual interpretation of the keyword “bat”. A classification decision made by such a system would be considered correct whether it positively classified a baseball bat or a winged rodent. If desired, further disambiguation should either be up to the user or a higher level reasoning component.

Until now, the majority of automatic dataset gathering approaches have fallen into the paradigm of first learning a model or classification boundary from the noisy training data, and then applying a classifier back to the data to refine the image search results. Unfortunately, such a procedure typically encourages the selection of a dominant appearance model, which is likely to noticeably reduce diversity in filtered data.

In this work we outline a very simple and computationally inexpensive method for consistency based filtering. Because it is a nearest neighbor based approach, it is robust to a discontinuous category appearance by design. We show that it performs sufficiently well to generate an improvement of 15.2% in the average recognition performance on a 21 category subset of Caltech-101 [4], using filtered versus unfiltered training data gathered from the web. We also compare recall and precision performance with other, more complex, recent approaches and demonstrate comparable results.

In the process of defining our filtering approach we also contribute a simple heuristic for automatically choosing a good threshold for the strangeness-based feature selection method developed by Li *et al*. [8]. Finally, we also adapt the Li *et al*. approach to the case where only one category must be filtered, but a known background category is given.

1.1. Related work

Several very recent works have addressed the related problems of image relevancy ranking and large image dataset acquisition. Here we provide a quick overview of existing approaches and explain what makes ours distinct.

From the domain of model based filtering approaches, there exist three recent notable works [6, 5, 9].

In [6] a best constellation model was selected using a RANSAC approach based on validation set classification performance. Similarly [5] trained a multiple-topic pLSA model directly on the full unprocessed image search result set. The topic most representative of the category was selected according to individual topic performance on a validation set classification task. The number of topics used during model learning was a critical design parameter. Since the number of visual clusters in a category will vary depending on dataset and category characteristics, including unpredictable factors such as polysemy, statically fixing the number of topics may limit the robustness of this approach.

In addition to limiting diversity by selecting only the dominant model, these methods also used an ad-hoc validation set, gathered for each category by translating a query into 7 languages, and keeping the top 5 image search results returned for each translation. Though this may work in many cases, it introduces an additional point of failure into both algorithms, since the wrong models are likely to be selected if a keyword is mistranslated, resulting in a bad validation set.

Li *et al*. [9] took a different approach, relying on a small initial set of seed images to bootstrap an incremental learning procedure. The first 15 images returned by a search engine were assumed to be sufficiently precise to serve as the seed set for a category. The seed image set was used to learn a Hierarchical Dirichlet Process (HDP) category model. Using the HDP model allowed Li *et al*. to overcome the issue of selecting the proper number of topics to represent a category in advance. Once an initial model was learned based on the seed set, it was used to classify subsequently obtained images. New positively classified images were then added to the dataset, and the HDP model was updated. Since such a learning approach is subject to becoming increasingly more specialized with each iteration, a special training set selection step was necessary. This was accomplished by only updating the HDP model with new images that were positively classified but had high entropy relative to those already incorporated into the model. This concept is very similar in spirit to the automatic query expansion technique used by Chum *et al*. [3] to obtain robust models for individual instances of objects.

Despite the query expansion style procedure to obtain a more robust model, the OPTIMOL system of [9] still assumes that as the classifier continues to expand its definition of the category, it will be able to encompass the majority of a class's visual appearance characteristics. However, if a category consists of two or more objects that occur consistently within that data but are visually very different, it is uncertain whether OPTIMOL would be able to learn both category appearances, unless an instance of each variety appears in the 15 initial training examples.

Taking a different perspective, a recent discriminative approach to image re-ranking ordered image search results based on consistency of color blobs within the data [1]. However, this approach was limited by its color-only representation and explicit selection of a dominant visual cluster.

Some other related approaches [2, 15], which set out to construct datasets that are as large as possible, have gone beyond the limitations of image search (whose results are sometimes limited to the first 1000) by using a hybrid of visual features and their own text-based ranking to retrieve images of interest directly from web pages. In comparison, our method avoids reimplementing text-based web search and relies on publicly available image search engines, whose results are constantly improving. We focus only on visual consistency within the image search results.

A final distinction of our approach from all of the above methods is that our aim is exclusively the creation of a good training set. Unlike some of the earlier discussed approaches, we are not trying to create the largest dataset possible, nor are we trying to capture difficult situations such as occlusions for the purpose of creating a challenging test set. We specifically seek easy images that are likely to contain the most information about a particular class.

2. Approach

Underlying our method are two simple assumptions. The first is the fundamental assumption of the category recognition problem; that objects belonging to the same category share some common visual properties. Therefore, we can expect that in an image search result for a particular query, images relevant to the search will demonstrate some within category visual consistency. Irrelevant images, on the other hand, are much less likely to be related to each other, and thus are not expected to express a similar degree of visual consistency. Due to varying viewpoints, poses, and within category appearance variations, relevant images may lie on a discontinuous subset. This observation is critical because it indicates that it is incorrect to select a single "most representative" cluster of a category and rank all search results with respect to its mean. Any consistency or model-based method should anticipate such potential discontinuity in order to retrieve a fair variety of representative category images.

Our method capitalizes on the visual consistency assumption by iteratively eliminating strongly inconsistent or *strange* images as determined by the *k-Nearest-Neighbor (k-NN) strangeness* measure [8]. The strangeness metric and its associated selection algorithm are outlined in sections 2.1 and 2.2.

Our second assumption is that a better visual model can be obtained from training images where the object of interest is clearly visible, fills the majority of the given image, and appears in one of several common poses. Note that some of these assumptions contradict those desired in a good testing set [14]. However, these conditions provide the most information about the appearance of the category, and allow for maximum flexibility in the types of models that could be learned. The driving intuition is that more informed classifiers will generally be more robust to noisy testing cases that may include occlusions, illumination changes or other factors.

This second assumption partially motivates our choice of descriptor. Since we wish to find images where the object fills the majority of the view, that means that we seek appearance consistency on the scale of the entire image. Holistic features are thus most appropriate for roughly comparing structure and appearance of whole images. In our work we use the holistic features of Torralba [16] to encode the *gist* of images. Even though these features were initially used for scene and context recognition, they become object descriptors in cases where an object fills the majority of an image. This was acknowledged by both Torralba [16] and Lazebnik *et al*. [7]. We give a brief overview of these holistic features in the latter portion of this section.

Considered in concert, the above assumptions suggest that a simple discriminative framework with the ability to capitalize on visual consistency of image search results may be able to provide sufficiently good training data to enable classification of previously unknown categories. In the remainder of this section we describe our pursuit of such a framework based on k-NN strangeness and image gist.

2.1. k-NN strangeness

The main workhorse of our filtering procedure is a feature elimination algorithm based on the *k-Nearest-Neighbor strangeness* measure. The k-NN strangeness of a data point, as defined by [8], is the ratio of the sum of distances to the k nearest within class neighbors to the sum of distances to the k nearest members of a different (closest) class. Formally, the strangeness, α, of a data point j belonging to class c is given by:

$$\alpha_j = \frac{\sum_{l=1}^{k} d_{jl}^{c}}{min_{n, n \neq c} \sum_{l=1}^{k} d_{jl}^{n}}, \tag{1}$$

where d_{jl}^{n} is defined as the distance from point j to the lth closest point that is a member of class n. In our application the distance metric is simply the L_1 distance between holistic image descriptors.

When used as a classification boundary, Li *et al*. [8] show that k-NN produces a smoother boundary than the standard Nearest Neighbors classifier, thus giving better overall generalization. In addition to its generalization benefits, this particular strangeness metric is a natural choice for our problem because it allows us to measure consistency without assuming any explicit category model or specifying a concrete number of feature clusters.

We performed experiments for a number of values of k, and found that filtering results tend to stabilize and show little change after $k > 3$. Thus, all experimental results were obtained with $k = 5$.

2.2. Strangeness based feature selection

In [8], k-NN strangeness was used in an iterative feature elimination framework to simultaneously remove non-discriminative feature instances from multiple classes. The algorithm typically eliminated features that appeared in more than one class, producing the net effect of eliminating features belonging to the background and leaving features that were most descriptive of the classes of interest.

The standard feature elimination as explained in [8] requires that there initially be at least two classes of data, and that an initial strangeness threshold, γ, be established. Strangeness is then computed for each feature instance with respect to its preliminary class label, as in equation 1. All feature instances for which $\alpha_j > \gamma$ are then eliminated and strangeness values are updated appropriately. The process of feature elimination and re-estimation of strangeness continues until $\alpha_j \leq \gamma$ for all j in the remaining set of features.

2.3. Feature selection with known background

The above feature selection framework relies on the existence of at least two classes for feature selection to be possible. As a result of refining the features in each class with respect to the other classes, the background is effectively eliminated.

If instead samples from the background were given, we could perform strangeness-based feature selection from a different perspective. This time instead of refining the definition of multiple classes with respect to each other, we are able to refine a single class with respect to the background. Note that we are not interested in refining our concept of the background. We also prefer to be extra cautious in refining the class of interest to obtain high precision for the feature instances that remain after elimination. To accomplish this, features are only eliminated from the class being filtered, and the background is kept unchanged throughout the iterative filtering procedure. The algorithm operates the same as described above, but terminates when all strangeness values in the class being filtered are below γ.

If a background dataset is not available but multiple categories exist, an initial background dataset can be constructed from all features filtered out by the procedure in section 2.2 applied to the multiple bootstrap classes.

2.4. Strangeness threshold selection

As noted earlier, strangeness-based feature selection requires a maximum threshold, γ, above which feature instances are eliminated. For the single category versus background filtering procedure, a threshold estimate can be made based on the initial distribution of strangeness values within the class to be filtered. We want to select a threshold that will eliminate many strange images, but simultaneously leave enough images to have diversity in the results when the procedure converges.

Naively, we could assume that the distribution of preliminary strangeness values of all within class features is best separated at the mean. The problem is that the mean incorporates highly strange data points that are practically guaranteed to be out of class features. Therefore, we choose to ignore outliers with very large strangeness by only considering a certain percentage of features with the lowest α_j. We have found empirically that computing the mean within class strangeness based on the lowest 80% of the strangeness values consistently provides a good γ.

The above gives a conservative approach that always biases the threshold closer to the consistent data than the noisy samples. When the distribution of features is such that the good samples appear in well defined clusters, the preliminary α_j values of the good samples will be low. In this scenario even some noise samples may have $\alpha_j < 1$. In such a case, our approach will select a $\gamma < 1$. When good features are distributed in a more loosely structured fashion but still distinguishable from the background, our heuristic will choose $\gamma \approx 1$. In the third case, when features are difficult to distinguish from the background and strangeness values of even some good samples may be greater than 1, our heuristic selects $\gamma > 1$. This ensures that not all features are eliminated but only the most consistent ones are kept.

All experiments performed use this threshold selection technique, with the exception of the precision and recall comparison, which is explained later.

2.5. A holistic image representation: image gist

In order to apply the strangeness procedure to whole images, we need a lower dimensional description of image content such that the difference between descriptors is representative of image similarity. The majority of related works have used bag-of-features or other part-based models to represent image content. Instead of identifying an image by a histogram of visual words, we select a holistic representation that directly encodes global statistics of an entire image.

The rationale for our feature choice comes from our earlier assumption that we seek to find good training data, where the object of interest fills the majority of the image. Thus we are interested in computing visual consistency on the scale of an entire image.

We selected Torralba's holistic representation which captures the *gist* of an image by encoding spectral scene components and their spatial layout at low resolution [16]. In color images gist is computed independently for each color channel, thus also incorporating color information.

Before computing gist on our data, all images were resized to 128×128. This size was selected to approximately match average search engine result thumbnail sizes, so that only thumbnails needed to be harvested from image search results. This allowed for a simpler harvesting procedure and required less storage space than downloading all search results in their original sizes. After resizing, we used Oliva and Torralba's implementation [13] to compute the gist using Gabor filters quantized into 8, 8, and 4 orientations, respectively, over three scales. We then projected the resulting descriptors onto 32 principal components, found by performing PCA on 12,651 images selected from a subset of the Caltech-101 and Web-23 datasets (described below). We also duplicated a number of our experiments in this paper for a 64 principal component gist projection, but found the results to be nearly identical. Thus 32 principal components were used for all reported results.

2.6. Automatic duplicate removal

Because our method relies on visual consistency, results can be thrown off by the presence of duplicate images in web gathered data. Since duplicate images tend to occur sufficiently often in Internet image search results to affect our filtering procedure, we devised a very simple automatic duplicate removal scheme based on the gist descriptors before projection onto their principal components.

To overcome differences in compression, size, and other small but insignificant inconsistencies between otherwise visually identical images, we identified duplicates by establishing a threshold on the L_1 distance between gist descriptors. The threshold was selected based on duplicates in the Fergus [5] dataset. Visually exact duplicate images in several categories were manually identified, and the lowest threshold that eliminated all duplicates was selected.

We found that this simple method consistently removed exact duplicates from other datasets as well, even with varying levels of compression, and kept any images with visually detectable differences.

3. Results

3.1. Datasets

Three datasets were used in the experiments described in this paper. The first was *Caltech-101* [4], which includes annotated images of 101 different object categories, with 31 to 800 images per category. Caltech-101 was selected for initial experiments because it provided good ground truth labels, allowing for a quantitative evaluation of the filtering procedure proposed by this work.

The second dataset used was *Web-23* [9]. This dataset includes a set of 23 object categories, downloaded from Internet image search engines in response to 21 query words randomly selected from labels in the Caltech-101 dataset, as well as two additional categories. The number of images per category in Web-23 ranges from 577 to 12414. However, reflecting the nature of currently existing image search engines, each harvested category includes a high number of irrelevant images. Li *et al*. [9] noted that the accordion category, for example, contains only 352 out of 1659 correct accordion images.

For the purpose of recognition performance comparison, throughout this work we used only the 21 category subset of Caltech-101 for which corresponding categories exist in Web-23.

Our final dataset was that of Fergus *et al*. [5], which consists of 7 categories of images gathered from Google's image search in 2005. Each image in the data set includes a ground truth label indicating whether it is a good representation of its category, a mediocre one (termed *okay*), or completely unrelated (called *junk*). This dataset was useful for evaluating recall and precision on real web gathered data and comparison to past approaches.

Category (C)	$\lvert C\rvert$	$\lvert +BG\rvert$	γ	$\lvert C^*_{good}\rvert$	$\lvert C^*_{BG}\rvert$
accordion	55	55	0.98	54	0
bonsai	128	128	0.995	85	1
euphonium	64	64	1.03	37	0
Faces	435	200	0.78	199	0
grand piano	99	99	0.949	62	0
inline skate	31	31	1.07	27	0
laptop	81	81	1.06	63	0
menorah	87	87	1.02	54	0
nautilus	55	55	1.13	23	0
pagoda	47	47	0.968	47	0
panda	38	38	1.14	29	0
pyramid	57	57	1.09	43	1
revolver	82	82	0.957	56	0

Table 1. Results of filtering polluted Caltech-101 categories. $|C|$ is the size of each category dataset, $|+BG|$ is the number of random images mixed in from the Caltech-101 background dataset, γ is the strangeness threshold used, and $|C^*_{good}|$ and $|C^*_{BG}|$ are the number of good images and background images remaining after filtering, respectively. The γs are selected automatically via our described heuristic.

3.2. Filtering polluted Caltech-101 categories

Several categories from the Caltech-101 dataset were selected and polluted with random images from the Caltech-101 background category. The background dataset was split so that the background images used in the filtering procedure were disjoint from the set used to pollute individual categories.

The results of filtering each polluted category with our method are outlined in Table 1. One can see that the algorithm proved to be fairly effective in these simple test cases. Thus this experiment confirmed that the concept of our approach is reasonable for the desired task.

3.3. Training on polluted Caltech-101 data

The following three experiments serve to demonstrate the critical importance of good training data and put a quantitative measure on the performance of our filtering approach, in terms of percentage improvement on the category classification task.

To reflect current trends in computer vision, we used a simple bag-of-features approach, as proposed by [11, 12], the implementation of which has been made available at [18]. The algorithm constructs a hierarchical dictionary of SIFT [10] features with which each image is represented, and performs simple classification using the k-Nearest Neighbors method with $k = 5$. It has been found to give an average recognition rate of 46% when evaluated on the full Caltech-101 dataset. The key contribution of this paper, however, is the dataset filtering procedure; thus we are interested in the algorithm's sensitivity to noisy training data, rather than the absolute recognition performance. Since the bag-of-features model is currently prevalent in the field, the effect of bad training data on the algorithm's performance will be reflective of a large number of recently published recognition algorithms.

We first trained the bag-of-features algorithm in the canonical fashion to provide a baseline: splitting each category in the Caltech-101 dataset, and using one fraction for training and the remainder as a test set. Throughout all experiments, 30 random images were taken from each category as good images for the training set. With only good images used for training, the recognition algorithm achieved an average recognition rate of $77.4 \pm 1.8\%$ for the 21 selected categories.

To investigate the effect of noisy training data on recognition algorithm performance, each good training set was polluted with an equal amount of randomly selected images from the Caltech-101 background dataset. 30 training images were then sampled uniformly from the 60 images constituting a polluted training set. With the noisy training set, the algorithm experienced a significant reduction in performance, yielding a $66.5 \pm 1.1\%$ average recognition rate on the 21 categories.

In order to avoid any overlap with the training set and the filtering algorithm's background data, the background dataset was split randomly into 200 training images used for category pollution, and the remaining 267 images, used as the known background dataset for the filtering procedure.

In the third experiment, each category training set was polluted as in the second experiment, but the full polluted training sets of 60 images each were then immediately filtered with the algorithm presented in this paper. If more than 30 images remained in a category after the filtering procedure, then 30 were uniformly sampled from the filtered set to create each training set. Otherwise, all remaining images were used as the training set. The experiment was repeated 10 times. Training from noisy data after applying our filtering approach yielded a $73.9 \pm 1.6\%$ average recognition rate, an improvement over the $66.5 \pm 1.1\%$ obtained without filtering. Confusion matrices from one experiment sequence are illustrated in Figure 1.

3.4. Filtering Web-23

To demonstrate how our approach improves recognition performance when training on real web-gathered images, we performed similar classification experiments to those above, but replaced artificially polluted Caltech-101 training data with sampled images from Web-23. In these experiments, we also augmented the Caltech-101 background dataset with 1000 images gathered from Yahoo Image Search for the query "things". Duplicates were eliminated from the combined dataset using the approach described in section 2.6.

For each of the 21 Web-23 categories that corresponded to categories in Caltech-101, we randomly sampled 30 images before and after filtering, to respectively serve as the noisy and filtered training sets. Over 15 experiments, average classification performance on 21 Caltech-101 categories when trained on unfiltered Web-23 data was $25.8 \pm 2.8\%$. When filtered data was used for training, performance improved to $41.0 \pm 2.3\%$. Sample confusion matrices from one run are shown in Figure 2.

3.5. Precision-recall in search engine gathered data

To further analyze filtering performance on web gathered data and compare to other approaches, the filter was tested on the dataset of Fergus *et al*. [5].

Fergus *et al*. [5] and Schroff *et al*. [15] compared the precision of their image re-ranking methods at 15% recall. Since our approach eliminates images until stable strangeness values are achieved with respect to a specified threshold, we cannot provide precision results at exactly 15% recall. To compare as fairly as possible, we generated results for a large range of strangeness thresholds and provided our precision and recall values for our closest available datapoints. This comparison is shown in Table 2.

We found that the categories for which filtering performance was the lowest were "polluted" in a consistent manner. For example, the airplane category contained over 170 screen captures of software, all of which were similar enough to emerge as a part of the airplane category. With such a high number of a particular type of image appearing in the search results, one could argue that according to visual consistency, the keyword "airplane" visually corresponds to not only a familiar mode of transportation, but a software application as well. In our case, the resulting filtered training set consisted of a mixture of both of these visual sub-categories.

As a secondary experiment, to compare more closely with the results of Schroff *et al*., we trained a χ^2 kernel SVM with the same training set as [15], in order to duplicate the drawing filter used in that work. For the more difficult categories such as airplane, this filter helped significantly because it eliminated some consistent noise images, such as screenshots, which were previously being incorporated into the category. For simpler categories, such as wristwatch, however, many good images were eliminated by the drawing filter due to their simplicity. For the wristwatch category in particular, over half of the best images were removed by the drawing filter. Our precision and recall results after use of the drawing filter are reported in Table 2.

To see how use of the drawing filter impacts a larger

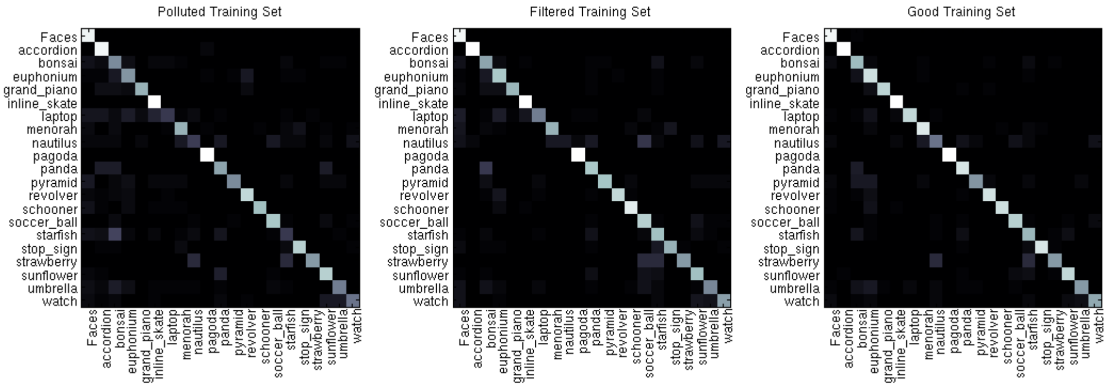

Figure 1. Confusion matrices demonstrating results on the category recognition task for a 21 category subset of Caltech-101. Lighter colors indicate higher percentage accuracy. The rightmost confusion matrix shows performance when trained on 30 good images per category ($77.4 \pm 1.8\%$ average recognition rate). The leftmost matrix shows the degrade in performance when 30 training images are sampled from an image set where half the samples are background ($66.5 \pm 1.1\%$ average recognition). Finally, the center matrix shows the performance improvement when each noisy dataset is filtered with the method described in this work before training ($73.9 \pm 1.6\%$ average recognition).

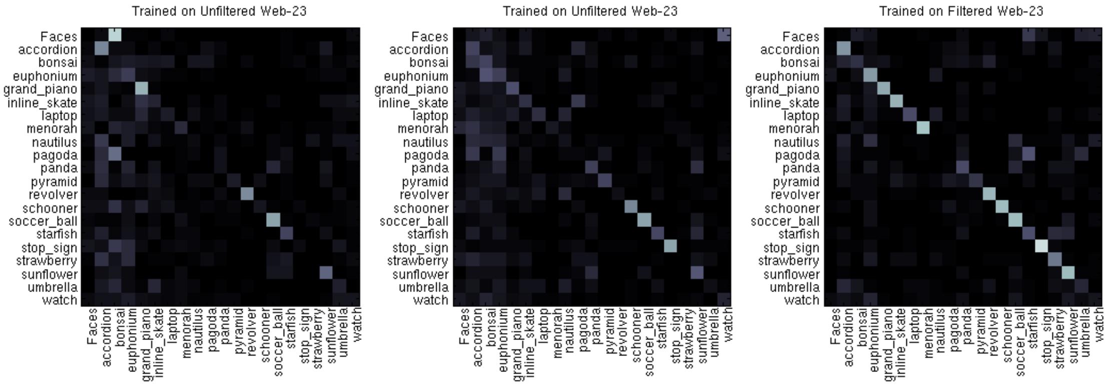

Figure 2. Confusion matrices demonstrating results on the 21 category recognition task when trained with Web-23 data and tested on Caltech-101. The two matrices on the left show sample results based on unfiltered training data, and the rightmost confusion matrix demonstrates sample results when trained from filtered data. Due to some polysemy effects in the web data and known artifacts in Caltech-101, recognition performance is not matched with the good data only training scenario. However, filtering with our approach was found to improve average recognition performance to $41.0 \pm 2.3\%$ from the $25.8 \pm 2.8\%$ achieved when using unfiltered Web-23 training data.

number of categories, we repeated our Web-23 recognition test, but applied the drawing filter to the data before using our procedure. We found, however, that over the 21 categories used, filtered recognition performance dropped from $41.0 \pm 2.3\%$ to only $32.1 \pm 3.0\%$. We thus conclude that for our method, filtering out simple images before applying our procedure is not practical in the general case.

4. Conclusions

We introduced a simple approach to filter Internet image search results based exclusively on visual content. We showed that this filtering procedure significantly improved performance of popular classifiers when trained on data obtained automatically from Internet image search engines. Specifically, we showed an improvement of 15.2% in the average recognition performance on a 21 category subset of Caltech-101, using filtered versus unfiltered web gathered data. We also compared our method against precision and recall results of previous model-based and text/image hybrid re-ranking algorithms, demonstrating comparable or improved performance with a significantly simpler approach.

In the process of defining our filtering procedure we also provided a simple heuristic for automatic strangeness threshold selection, and adapted the strangeness-based fea-

	airplane	guitar	leopard	motorbike	wristwatch
our-D(OK)	76.19 @ 13.73	80.39 @ 14.39	58.33 @ 14.4	84.09 @ 17.01	88.89 @ 15.15
our-D(G)	62.16 @ 14.56	28.79 @ 14.73	35.00 @ 20.79	61.22 @ 13.04	84.21 @ 14.75
our(OK)	34.27 @ 15.45	69.91 @ 27.72	75.76 @ 20.57	86.11 @ 14.25	100 @ 14.39
our(G)	74.19 @ 14.56	26.55 @ 23.26	41.67 @ 14.81	63.89 @ 20	100 @ 17.51
[15](OK)	45 ± 5	72 ± 11	72 ± 6	81 ± 9	97 ± 4
[15](G)	35 ± 4	29 ± 4	50 ± 5	63 ± 8	93 ± 7
[5](G)	57	50	59	71	88

Table 2. A comparison of precision at 15% recall on a subset of the [5] dataset. Since specifying an exact precision is not possible in our approach, we compare the closest available datapoint, specified in the form "precision @ recall". The (G) indicates that recall and precision were computed using only data with ground truth labels of *good* as positive retrievals. (OK) indicates that both *good* and *ok* images were counted as positive. Further, the "-D" label indicates that the drawing removal SVM filter used in [15] was applied to the data before our filtering procedure. It is critical to note that the performance metrics reported by others may be influenced by the existence of duplicate images. The airplane category, for example, contains 24 exactly identical images of a particular airplane, all classified as *good*. Since our method eliminates duplicates before processing, all result images are guaranteed to be unique in our approach. However if duplicates are not eliminated, it may be possible that a high precision score is obtained for a low recall value by reporting duplicates of good images.

ture selection method to the case where a known background category is given.

In future work we wish to perform several additional experiments, including automatically learning all categories in Caltech-101 from keyword based image search results, and using our approach to gather training data for a window based classifier in order to evaluate classification performance improvements on more difficult data. We also plan to experiment with incorporating search engine ranking to improve the strangeness measure.

Acknowledgments

We would like to acknowledge the support of AFOSR FA9550-06-1-0138 and ONR 67F-1080868.

References

[1] N. Ben-Haim, B. Babenko, and S. Belongie. Improving web-based image search via content based clustering. In *Proceedings of the International Workshop on Semantic Learning Applications in Multimedia, New York City*, 2006. 2

[2] T. Berg and D. Forsyth. Animals on the web. In *Proc. CVPR*, 2006. 1, 2

[3] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman. Total recall: Automatic query expansion with a generative feature model for object retrieval. In *Proc. ICCV*, 2007. 2

[4] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *Proc. CVPR Workshop*, 2004. 2, 5

[5] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning object categories from google's image search. In *Proc. ICCV*, volume 2, pages 1816–1823, Oct. 2005. 1, 2, 5, 6, 8

[6] R. Fergus, P. Perona, and A. Zisserman. A visual category filter for google images. In *Proc. ECCV*, pages 242–256, May 2004. 1, 2

[7] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proc. CVPR*, 2006. 3

[8] F. Li and J. Kosecka. Strangeness based feature selection for part based recognition. In *Proc. CVPR, Beyond Patches Workshop*, 2006. 1, 2, 3

[9] L.-J. Li, G. Wang, and L. Fei-Fei. Optimol: automatic object picture collection via incremental model learning. In *Proc. CVPR*, 2007. 1, 2, 5

[10] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2(60):91–110, 2004. 5

[11] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *Proc. CVPR*, 2006. 5

[12] E. Nowak, F. Jurie, and B. Triggs. Sampling strategies for bag-of-features image classification. In *Proc. ECCV*, 2006. 5

[13] A. Oliva and A. Torralba. Spatial envelope. http://people.csail.mit.edu/torralba/code/spatialenvelope/, 2001. 4

[14] J. Ponce, T. L. Berg, M. Everingham, D. A. Forsyth, M. Hebert, S. Lazebnik, M. Marszalek, C. Schmid, B. C. Russell, A. Torralba, C. K. I. Williams, J. Zhang, and A. Zisserman. Dataset issues in object recognition. In J. Ponce, M. Hebert, C. Schmid, and A. Zisserman, editors, *Toward Category-Level Object Recognition*, volume 4170 of *LNCS*, pages 29–48. Springer, 2006. 1, 3

[15] F. Schroff, A. Criminisi, and A. Zisserman. Harvesting image databases from the web. In *Proc. ICCV*, 2007. 1, 2, 6, 8

[16] A. Torralba. Contextual priming for object detection. *IJCV*, 53(2):161–191, 2003. 3, 4

[17] A. Torralba, R. Fergus, and W. T. Freeman. Tiny images. Technical Report MIT-CSAIL-TR-2007-024, Computer Science and Artificial Intelligence Lab, Massachusetts Institute of Technology, 2007. 1

[18] A. Vedaldi. Bag of features. http://vision.ucla.edu/∼vedaldi/code/bag/bag.html, 2007. 5