Landmark Classification in Large-scale Image Collections

Yunpeng Li David J. Crandall Daniel P. Huttenlocher
Department of Computer Science, Cornell University
Ithaca, NY 14853 USA
{yuli,crandall,dph}@cs.cornell.edu

Abstract

With the rise of photo-sharing websites such as Facebook and Flickr has come dramatic growth in the number of photographs online. Recent research in object recognition has used such sites as a source of image data, but the test images have been selected and labeled by hand, yielding relatively small validation sets. In this paper we study image classification on a much larger dataset of 30 million images, including nearly 2 million of which have been labeled into one of 500 categories. The dataset and categories are formed automatically from geotagged photos from Flickr, by looking for peaks in the spatial geotag distribution corresponding to frequently-photographed landmarks. We learn models for these landmarks with a multiclass support vector machine, using vector-quantized interest point descriptors as features. We also explore the non-visual information available on modern photo-sharing sites, showing that using textual tags and temporal constraints leads to significant improvements in classification rate. We find that in some cases image features alone yield comparable classification accuracy to using text tags as well as to the performance of human observers.

1. Introduction

The billions of photographs in Internet-scale photo collections offer both exciting opportunities and significant challenges for computer vision, and for the area of object recognition in particular. Achieving Internet-scale object recognition and image classification is currently limited by the relatively small-scale datasets for which ground truth information is available. For instance, the widely-used PASCAL VOC 2008 dataset [7] has about 10,000 images and 20 categories, while the LabelMe dataset [13] is of similar size, with a larger hierarchically-organized label set. Bigger datasets such as Tiny Images [16] have millions of images but do not include category labels, whereas other datasets make use of visual features during image selection which may bias towards certain methods (e.g., [2, 14]). Recent work on scaling classification algorithms to Internet-sized datasets with millions of images (such as [17]) has thus been limited to evaluating classification performance on relatively small datasets such as LabelMe.

In this paper we consider image classification on much larger datasets featuring millions of images and hundreds of categories. First we develop a collection of over 30 million photos with ground-truth category labels for nearly 2 million of those images. The ground-truth labeling is done automatically based on geolocation information that is separate from the image content and the text tags that we use for classification. The key observations underlying our approach is that photos taken very near one another are likely to be of similar things. Moreover, if many people have taken photos at a given location, there is a high likelihood that they are photographing some common area of interest, or what we call a *landmark*. Thus we use a mean shift [3] procedure to find peaks in the spatial distribution of geotagged photos, and then use large peaks to define the category labels. The photographs taken at these landmarks are typically quite diverse (see Figure 1 for some examples), so that the labeled test datasets are challenging, with significant amounts of visual variation and a large fraction of outliers. In most cases, a landmark does not consist of any one prominent object; for example, many of the landmarks are museums, in which the photos are distributed among hundreds of exhibits. Our landmark classification problem can thus be thought of as more similar to an object category recognition problem than to a specific object recognition problem. In Section 3 we discuss the details of our dataset collection approach and compare it to some alternative techniques.

We use multiclass support vector machines [4] to learn models for various classification tasks on this labeled dataset of nearly two million images. We use visual features based on clustering local interest point descriptors [11] into a visual vocabulary that is used to characterize the descriptors found in each image. We also explore using the textual tags that Flickr users assign to photos as additional features. The learning and classification methods and the feature ex-

traction are discussed in more detail in Section 4.

Internet photo collections also include rich sources of relational information that can be helpful for classification. For instance, social ties have been found to improve face recognition performance on Facebook [15]. In this paper we consider the *photo stream* of a given photographer, using features from photos taken nearby in time to aid in classification decisions. In particular we use the structured support vector machine [18] to predict the sequence of category labels for a photo stream rather than classifying a single photo at a time. Feature extraction, learning, and classification methods using temporal context are discussed further in Section 5.

In Section 6 we present a set of large-scale classification experiments involving between 10 and 500 categories and tens to hundreds of thousands of photos (in contrast to other recent image recognition work which use large datasets but small test subsets). We find that the combination of image and text features performs better than either alone, even when we remove untagged photos from the dataset. We also describe a small study of human performance on landmark classification which suggests that a multiclass SVM using both image and text features performs nearly as well as people can. Finally we show that using temporal context from photos taken by the same photographer nearby in time yields a striking improvement compared to using visual features alone — around 10 percentage points in most cases. On the other hand, the improvement using the textual tags from those same nearby photos is small.

Thus we find that bag-of-word models using structured SVM classifiers with vector-quantized SIFT features can in many cases yield classification results nearly as good as or better than those obtained using text features, while also performing nearly as well as humans. Moreover the running time is dominated by the feature extraction, with classification taking just a few milliseconds per photo even for hundreds of categories. These experiments demonstrate the power of large labeled datasets, even when a substantial fraction of the training images is mislabeled, suggesting that for certain kinds of problems visual matching of Internet-scale datasets could be quite feasible with today's techniques.

2. Related Work

Image classification using bag-of-features models has been studied extensively (see [6] or [20] for recent surveys), however such previous work has been carried out only at much smaller scales. The work we report here uses two orders of magnitude more labeled photos – nearly two million photos as opposed to a few thousand in previous work – and one to two orders of magnitude more categories – up to 500 compared to tens in most previous work. This larger scale allows us to study how performance is affected by the number of categories and the number of training images available. Our investigation also evaluates text tags versus image features, and considers the use of temporal context which has not received much attention in the literature.

Some recent work has used large datasets, but the number of *labeled* photos available for evaluating performance has usually been quite small. For instance [12] uses one million photos but only 5,000 of them have ground truth labels. The recent work of [17] considers a dataset with tens of millions of images, but only at thumbnail resolutions and again without labels for assessing classification accuracy. Another line of research uses small training sets to automatically label larger image sets (e.g., [2, 14, 21]), however such approaches generally make use of image features and machine learning techniques, and thus the resulting datasets are not independent of the kinds of features and methods that one wants to test. This raises the possibility that methods related to the ones used to create the dataset might be at an unfair advantage.

We also investigate how the visual vocabulary size affects classification performance. Although [19] presents a technique for finding the optimal visual vocabulary size for their task, it is not clear that their method can scale to large datasets because the running time is linear in the number of images and quadratic in the number of categories.

The paper of [8] is related to our work in that it studies geolocating photographs, but their goal is quite different from ours, as we do not try to predict location but rather just use location to derive category labels. (For instance, in our problem formulation a misclassification with a geographically proximate category is just as bad as with one that is far away.) Our experiments use a standard classification paradigm and thus are comparable with many other studies. Moreover, the test set in [8] contains only 237 images that were partially selected by hand, making it difficult to generalize the results beyond that set. In contrast we use automatically-generated test sets that contain tens or hundreds of thousands of photos, providing highly reliable estimates of performance accuracy.

Some very recent papers have considered landmark classification tasks similar to the one we study here, but again have done so at a much smaller scale. For example, [10] studies how to build a model of a landmark by extracting a small set of iconic views from a large set of photographs. The paper tests on just three hand-chosen categories, making it unclear how well the method would scale to more realistic classification tasks. The very recent work of [21] is similar to our approach in that it finds highly-photographed landmarks automatically from a large collection of geotagged photos. However the test set they use is hand-selected and very small — 728 total images for a 124-category problem, or fewer than 6 test images per category — and their approach is based on nearest-neighbor search,

which is unlikely to scale to the millions of test images we consider here. Our recent paper on organizing large photo collections [5] uses a dataset of geotagged photos similar to the one we describe here, however the focus of that work is on geographic embedding and organization of photos instead of image classification.

3. Building Internet-Scale Datasets

Our long-term goal is to create large publicly-available, labeled datasets that are representative of photos found on photo-sharing sites on the web. In constructing such datasets, it is critical to avoid potential biases either in selecting the images to include in the dataset or in assigning ground-truth labels. For instance, methods based on searching for photos tagged with hand-selected keywords (e.g., [8, 12]) are prone to bias because one might inadvertently choose keywords corresponding to objects that are amenable to a particular image classification algorithm. A number of previous collection efforts also use unspecified criteria to discard certain photos from the dataset, again introducing the potential for bias towards a particular algorithm. Also problematic is using the same kinds of features to produce ground-truth labels as are used by the classification algorithm (e.g., as in [2, 14, 21]). We thus advocate automatic techniques for creating datasets based on features that are independent from those used by the algorithms being tested. In our case, we avoid using textual tags or visual features to label or select images, instead using a completely separate source of information: geotags.

Our dataset was formed by using the Flickr API to retrieve metadata for over 60 million publicly-accessible geotagged photos. We eliminate photos for which the precision of the geotags (as reported in Flickr metadata) is worse than about a city block. For each of the remaining 30 million photos we consider the latitude-longitude coordinates as a point in the plane, and then perform a mean shift clustering procedure [3] on the resulting set of points to identify local peaks in the photo density distribution, as in [5]. The radius of the disc used in mean shift is about 100m. Since our goal is to identify locations where many different people took pictures, we count at most 5 photos from any given Flickr user towards any given peak. We currently use the top 500 such peaks as categories; the number of photos becomes small for lower-ranked categories (e.g. the 500th largest peak has 585 photos whereas the 1000th largest peak has 284 photos). Figure 1 illustrates the top 10 categories in our dataset, corresponding to the ten most photographed landmarks.

We downloaded the image data for all 1.9 million photos known to our crawler that were geotagged within one of these 500 landmarks. For the experiments on classifying temporal photo streams, we also downloaded all images taken within 48 hours of any photo taken in a landmark, bringing the total number of images to about 6.5 million. The images were downloaded at Flickr's medium resolution level, which is about 1/4-megapixel. The total size of the dataset is just over one terabyte.

4. Single Image Classification

To perform image classification we adopt the bag-of-features model of [6]. As in that paper, we build a visual vocabulary by clustering SIFT descriptors from photos in the training set using the k-means algorithm. To make k-means clustering tractable on this quantity of data we use the approximate nearest neighbor (ANN) technique of [1] to efficiently assign points to cluster centers. The advantage of this technique is that it guarantees an upper bound on the approximation error, unlike other techniques that have recently been used for clustering such as randomized k-d trees [12]. In our implementation we set the bound such that the cluster center found by ANN is no further away than 110% of the distance between the point and the optimal cluster center.

Once a visual vocabulary of size k has been generated, a k-dimensional feature vector is constructed for each image by using SIFT to find local interest points and assigning each interest point to the visual word with the closest descriptor. We then form a frequency vector which counts the number of occurrences of each visual word in the image. For textual features we use a similar vector space model in which any tag used by at least three different users is a dimension in the feature space, so that the feature vector for a photo is a binary vector indicating presence or absence of each text tag. Both types of feature vectors are normalized to have L2-norm of 1. We also study combinations of image and textual features, in which case the image and text feature vectors are simply concatenated.

We learn a linear model that scores a given photo for each category and assigns it to the class with the highest score. More formally, let m be the number of classes and $\mathbf{x}$ be the feature vector of a photo. Then the predicted label is

$$\hat{y} = \underset{y \in \{1, \cdots, m\}}{\arg\max}\ s(\mathbf{x}, y; \mathbf{w}), \tag{1}$$

where $\mathbf{w} = (\mathbf{w}_1^T, \cdots, \mathbf{w}_m^T)^T$ is the model and $s(\mathbf{x}, y; \mathbf{w}) = \langle \mathbf{w}_y, \mathbf{x} \rangle$ is the score for class y under the model. Note that in our settings, the photo is always assumed to belong to one of the m categories. Since this is by nature a multi-way (as opposed to binary) classification problem, we utilize the multiclass SVM [4] to learn the model $\mathbf{w}$, using the SVM^{multiclass} software package [9]. For a set of training examples $\{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_N, y_N)\}$ the multiclass SVM opti-

Landmark (most distinctive tag)	**Random tags**	**Random images**
1. eiffeltower	eiffel city travel night street	
2. trafalgarsquare	london summer july trafalgar londra	
3. bigben	westminster london ben night unitedkingdom	
4. londoneye	stone cross london day2 building	
5. notredame	2000 portrait iglesia france notredamecathedral	
6. tatemodern	england greatbritian thames streetart vacation	
7. empirestatebuilding	manhattan newyork travel scanned evening	
8. venice	tourists slide venecia vacation carnival	
9. colosseum	roma england stadium building italy	
10. louvre	places muséedulouvre eau paris canon	

Figure 1. The world's most photographed landmarks, and the first 10 categories of our dataset. We show the highest-frequency tag relative to the background distribution, 5 random tags, and 5 random images. The landmark tagged "venice" is Piazza San Marco.

mizes the objective function

$$\min_{\mathbf{w},\boldsymbol{\xi}} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i \qquad (2)$$
$$\text{s.t. } \forall i, y \neq y_i : \langle \mathbf{w}_{y_i}, \mathbf{x}_i \rangle - \langle \mathbf{w}_y, \mathbf{x}_i \rangle \geq 1 - \xi_i$$

where C is the trade-off between training performance and margin in SVM formulations.[1] Hence for each training example, the learned model is encouraged to give higher score

[1]For all our experiments, we simply set C to $1/\bar{x}^2$ where $\bar{x}$ is the average L2-norm of the training feature vectors.

to the correct class label than to the incorrect ones. In fact, by simply rearranging terms it can be shown that the objective function is an upper bound on the training error.

In contrast, many previous approaches to object recognition using bag-of-parts models (such as [6]) train a set of binary SVMs (one for each category) and classify an image by comparing scores from the individual SVMs. Such approaches are problematic for n-way forced-choice problems, however, because the scores produced by a collection of independently-trained binary SVMs may not be comparable, and thus such approaches lack any performance guarantee. It is possible to alleviate this problem by using a different C value for each binary SVM (as is done in [6]), but this introduces additional parameters that need to be tuned, either manually or via a process such as cross validation.

Note that while the categories in this single-photo classification problem correspond to geographic locations, there is no geographical information used in the learning or classification. For example, unlike [8] we are not concerned with pinpointing a photo on a map, but rather with classifying images into discrete categories.

5. Temporal Information

Modern photo-sharing sites collect a rich set of metadata which is potentially useful for image classification tasks. For example, photos taken by the same photographer at nearly the same time are quite likely to be related. In the specific case of classifying landmarks, practical and physical constraints on human movement mean that certain sequences of category labels are much more likely than others. To learn the patterns created by such constraints, we view temporal sequences of photos taken by the same user as a single entity and label them jointly as a structured output.

5.1. Temporal Model for Joint Classification

We model a temporal sequence of photos as a graphical model with a chain topology, where the nodes represent photos and edges connect nodes that are consecutive in time. The set of possible labels for each node is simply the set of m landmarks, indexed from 1 to m. The task is to label the entire sequence of photos with category labels, however we score correctness only for a single selected photo in the middle of the sequence, with the remaining photos serving as temporal context for that photo. Denote an input sequence of length n as $X = ((\mathbf{x}_1, t_1), \cdots, (\mathbf{x}_n, t_n))$, where $\mathbf{x}_v$ is a feature vector for node v (encoding evidence about the photo such as textual tags or visual information) and t_v is the corresponding timestamp. Let $Y = (y_1, \cdots, y_n)$ be a labeling of the sequence. We would like to express the scoring function $S(X, Y; \mathbf{w})$ as the inner product of some *feature map* $\Psi(X, Y)$ and the model parameters $\mathbf{w}$, so that the model can be learned efficiently using the structured SVM.

Node Features To this end, we define the feature map for a single node v under the labeling as,

$$\Psi_V(\mathbf{x}_v, y_v) = (I(y_v = 1)\mathbf{x}^T, \cdots, I(y_v = m)\mathbf{x}^T)^T, \quad (3)$$

where $I(\cdot)$ is an indicator function. Let $\mathbf{w}_V = (\mathbf{w}_1^T, \cdots, \mathbf{w}_m^T)$ be the corresponding model parameters with $\mathbf{w}_y$ being the weight vector for class y. Then the node score $s_V(\mathbf{x}_v, y_v; \mathbf{w}_V)$ is the inner product of the $\Psi_V(\mathbf{x}_v, y_v)$ and $\mathbf{w}_V$,

$$s_V(\mathbf{x}_v, y_v; \mathbf{w}_V) = \langle \mathbf{w}_V, \Psi_V(\mathbf{x}_v, y_v) \rangle . \quad (4)$$

Edge Features The feature map for an edge (u, v) under labeling Y is defined in terms of the labels y_u and y_v, the time elapsed between the two photos $\delta t = |t_u - t_v|$, and the speed required to travel from landmark y_u to landmark y_v within that amount of time, $speed(\delta t, y_u, y_v) = distance(y_u, y_v)/\delta t$. Since the strength of the relation between two photos decreases with the elapsed time between them, we divide the full range of δt into M intervals $\Omega_1, \cdots, \Omega_M$. For δt in interval Ω_τ, we define feature vector

$$\psi_\tau(\delta t, y_u, y_v) = (I(y_u = y_v), I(speed(\delta t, y_u, y_v) > \lambda_\tau))^T, \quad (5)$$

where λ_τ is a speed threshold. This feature vector encodes whether the two consecutive photos are assigned the same label and, if not, whether the transition requires a person to travel at an unreasonably high speed (i.e. greater than λ_τ). The exact choices of the time intervals and the speed thresholds are not crucial, so long as they are sensible. We also take into consideration the fact that some photos have invalid timestamps (e.g. a date in the 22nd century) and define the feature vector for edges involving such photos as,

$$\psi_0(t_u, t_v, y_u, y_v) = I(y_u = y_v)(I(z = 1), I(z = 2))^T, \quad (6)$$

where z is 1 if exactly one of t_u and t_v is invalid and 2 if both are. Here we no longer consider the speed, since it is not meaningful due to invalid timestamps. The complete feature map for an edge is thus,

$$\begin{aligned} \Psi_E(t_u, t_v, y_u, y_v) \quad = \quad & (I(\delta t \in \Omega_1)\psi_1(\delta t, y_u, y_v)^T, \cdots, \\ & I(\delta t \in \Omega_M)\psi_M(\delta t, y_u, y_v)^T, \\ & \psi_0(t_u, t_v, y_u, y_v)^T)^T \end{aligned} \quad (7)$$

and the edge score is,

$$s_E(t_u, t_v, y_u, y_v; \mathbf{w}_E) = \langle \mathbf{w}_E, \Psi_E(t_u, t_v, y_u, y_v) \rangle , \quad (8)$$

where $\mathbf{w}_E$ is the vector of edge parameters.

Overall Feature Map The total score of input sequence X under labeling Y and model $\mathbf{w} = (\mathbf{w}_V^T, \mathbf{w}_E^T)^T$ is simply

the sum of individual scores over all the nodes and edges. Therefore, by defining the overall feature map as,

$$\Psi(X,Y) = (\sum_{v=1}^{n} \Psi_V(\mathbf{x}_v, y_v)^T, \sum_{v=1}^{n-1} \Psi_E(t_v, t_{v+1}, y_v, y_{v+1})^T)^T,$$

the total score becomes an inner product with $\mathbf{w}$,

$$S(X,Y;\mathbf{w}) = \langle \mathbf{w}, \Psi(X,Y) \rangle . \quad (9)$$

The predicted labeling for sequence X by model $\mathbf{w}$ is one that maximizes the score,

$$\hat{Y} = \arg\max_{Y \in \mathcal{Y}_X} S(X,Y;\mathbf{w}), \quad (10)$$

where $\mathcal{Y}_X = \{1, \cdots, m\}^n$ is the the label space for sequence X of length n. This can be obtained efficiently using Viterbi decoding because the graph is acyclic.

5.2. Parameter Learning

The model parameters are learned using structured SVMs [18]. Let $((X_1, Y_1), \cdots, (X_N, Y_N))$ be the training examples. The structured SVM optimizes for parameters $\mathbf{w}$ by minimizing a quadratic objective function subject to a set of linear soft margin constraints,

$$\min_{\mathbf{w},\boldsymbol{\xi}} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^{N} \xi_i \quad (11)$$
$$\text{s.t. } \forall i, Y \in \mathcal{Y}_{X_i} : \langle \mathbf{w}, \delta\Psi_i(Y) \rangle \geq \Delta(Y_i, Y) - \xi_i,$$

where $\delta\Psi_i(Y)$ denotes $\Psi(X_i, Y_i) - \Psi(X_i, Y)$ (thus $\langle \mathbf{w}, \delta\Psi_i(Y) \rangle = S(X_i, Y_i; \mathbf{w}) - S(X_i, Y; \mathbf{w})$) and the loss function $\Delta(Y_i, Y)$ in this case is simply the number of mislabeled nodes (photos) in the sequence. It is easy to see that the structured SVM degenerates into a multiclass SVM if every example has only a single node.

The difficulty of this formulation is that the label space $\mathcal{Y}_{X_i}$ grows exponentially with the length of the sequence X_i. Structured SVMs address this problem by iteratively minimizing the objective function using a cutting-plane algorithm, which requires finding the *most violated constraint* for every training exemplar at each iteration. Since the loss function $\Delta(Y_i, Y)$ decomposes into a sum over individual nodes, the most violated constraint,

$$\hat{Y_i} = \arg\max_{Y \in \mathcal{Y}_{X_i}} S(X_i, Y; \mathbf{w}) + \Delta(Y_i, Y), \quad (12)$$

can be obtained efficiently via Viterbi decoding in the same manner as making a prediction using the model.

6. Experiments

Figure 2 presents results for various classification experiments on our dataset of nearly 2 million images. For each of these experiments we evenly divided the dataset into test and training image sets that are disjoint by photographer, so that duplicate photos taken by the same user could not appear during both training and testing. To make classification results easier to interpret across different categories with differing numbers of images, we constructed the test and training datasets by sampling the same number of images from each category. In practice this means that the number of images used in an m-way classification experiment is equal to m times the number of photos in the least popular of the m landmarks, and the baseline probability of a correct random guess is $1/m$.

We see from Figure 2 that in classifying single images (as described in Section 4), the visual features are less accurate than textual tags but nevertheless significantly better than random baseline — four to six times higher for the 10 category problems and nearly 50 times better for the 500-way classification. The combination of textual tags and visual tags performs significantly higher than either alone, increasing performance by about 10 percentage points in most cases. This performance improvement is partially because about 15% of photos in the dataset do not have any textual tags. However even when such photos are excluded from the evaluation, adding visual features still gives a significant improvement over using text tags alone, increasing accuracy from 79.2% to 85.47% in the top-10 category case, for example.

The figure also shows a dramatic improvement in visual classification performance when photo streams are classified jointly using a structured SVM (as described in Section 5) — nearly 12 percentage points for the top-10 category problem, for example. In contrast, the temporal information provides little improvement for the textual tags, suggesting that tags from contemporaneous images contain largely redundant information. In fact, the classification performance using temporal and visual features is actually slightly higher than using temporal and textual features for the top-20 and top-50 classification problems. For all of the experiments, the best performance is achieved using the full combination of visual, textual and temporal features, which gives for example 82.54% correct classification for the 10-way problem and 45.34% for the 500-way problem — more than 220 times better than the baseline! For these experiments, the maximum length of a photo stream was limited to 11, or five photos before and after a photo of interest.

Figure 2 shows classification experiments for different numbers of categories and also for categories of different rank. For the textual features, problems involving higher-ranked categories are more difficult; for example, the performance on classifying landmarks ranked 1 through 10

Categories	Baseline	Single images			Photo streams		
		visual	textual	combined	visual	textual	combined
Top 10 landmarks	10.00	57.55	69.25	80.91	68.82	70.67	82.54
Landmarks 200-209	10.00	51.39	79.47	86.53	60.83	79.49	87.60
Landmarks 400-409	10.00	41.97	78.37	82.78	50.28	78.68	82.83
Top 20 landmarks	5.00	48.51	57.36	70.47	62.22	58.84	72.91
Landmarks 200-219	5.00	40.48	71.13	78.34	52.59	72.10	79.59
Landmarks 400-419	5.00	29.43	71.56	75.71	38.73	72.70	75.87
Top 50 landmarks	2.00	39.71	52.65	64.82	54.34	53.77	65.60
Landmarks 200-249	2.00	27.45	65.62	72.63	37.22	67.26	74.09
Landmarks 400-449	2.00	21.70	64.91	69.77	29.65	66.90	71.62
Top 100 landmarks	1.00	29.35	50.44	61.41	41.28	51.32	62.93
Top 200 landmarks	0.50	18.48	47.02	55.12	25.81	47.73	55.67
Top 500 landmarks	0.20	9.55	40.58	45.13	13.87	41.02	45.34

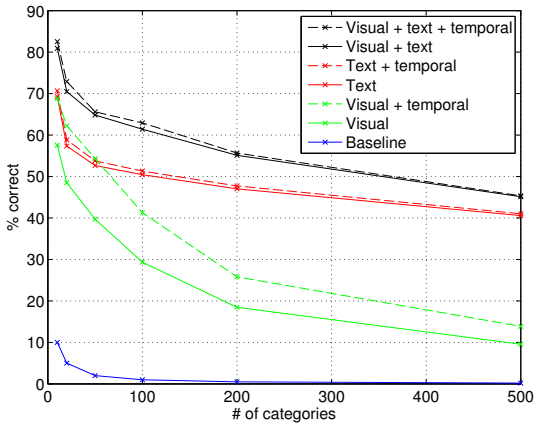

Figure 2. Percentage of images correctly classified for varying numbers of categories and combinations of features.

is about 10 percentage points worse than for landmarks 200 through 209. This is because the top landmarks are mostly located in a small set of cities including Paris, London, and New York, so that textual tags like "london" are relatively uninformative. On the other hand, classification using visual cues is significantly better for higher-ranked landmarks, probably because higher-ranked categories have more training images (e.g., 1,829 per category for the top 20 categories vs. 542 per category for 400-419).

A substantial number of Flickr photos are mislabeled or inherently ambiguous — a close-up photo of a dog or a sidewalk could have been taken at almost any landmark. To try to gauge the frequency of such difficult images, we conducted a small-scale human subject study. We asked 20 well-traveled people to each label 50 photos taken at the world's top ten landmarks. Textual tags were also shown for a random subset of the photos. We found that the average human classification accuracy was 68.0% without textual tags and 76.4% when both the image and tags were shown (with standard deviations of 11.61 and 11.91, respectively). Thus the humans performed better than the automatic classifier when using visual features alone (68.0% versus 57.55%) but about the same when both text and visual features were available (76.4% versus 80.91%).

For most of the experiments shown in Figure 2, the visual vocabulary size was set to 20,000. This size was computationally prohibitive for our (single-threaded) structured SVM learning code for the 200- and 500-class problems, so for those tasks we instead used 10,000 and 5,000, respectively. An interesting question is how the vocabulary size impacts classification performance on large-scale image sets. To study this we repeated a subset of the experiments for several different vocabulary sizes. As Table 1 shows, classification performance improves as the vocabulary size increases, but the relative effect is more pronounced as the number of categories increases. For example, when the vocabulary size is increased from 1,000 to 20,000, the relative performance of the 10-way classifier improves by about 20% (10.05 percentage points, or about one baseline) while the accuracy of the 100-way classifier increases by more than 50% (10.65 percentage points, or nearly 11 times the baseline). We found that performance on the 10-way problem asymptotes by about 80,000 clusters at about 59.3%. Unfortunately we could not try such large numbers of clusters for the other tasks because the learning becomes intractable; studying how to efficiently learn structured SVMs with such large feature vectors would be an interesting area for future work.

# of categories	Single images				
	1,000	2,000	5,000	10,000	20,000
10	47.51	50.78	52.81	55.32	57.55
20	39.88	41.65	45.02	46.22	48.51
50	29.19	32.58	36.01	38.24	39.71
100	19.77	24.05	27.53	29.35	30.42

Table 1. Visual classification rates for different vocabulary sizes.

In the experiments presented so far we sampled from the test and training sets to produce equal numbers of photos for each category, in order to make the empirical results easier to interpret. However our approach and results do not depend on this property of the experimental setup; when we sample from the actual photo distribution our techniques still perform dramatically better than the baseline (which is to guess the most frequent category). For example, in the top-10 category classification problem using the actual photo distribution we achieve 53.58% accuracy with visual features and 79.40% when tags are also used, versus a baseline of 14.86%; the 20-way classifier produces 44.78% and 69.28% respectively, versus a baseline of 8.72%.

The experimental results we report here are highly precise because of the large size of our test dataset. Even the smallest of the experiments, the top-10 classification, involves about 35,000 test images. To give a sense of the

variation across runs due to differences in sampling, we ran 10 trials of the top-10 classification task with different samples of photos and found the standard deviation to be about 0.15 percentage points. Due to computational constraints we did not run multiple trials for the experiments with large numbers of categories, but the variation is likely even less due to the larger numbers of images involved.

Image classification on a single 2.66 GHz processor takes about 2.4 seconds, most of which is consumed by SIFT interest point detection. Once the SIFT features are extracted, classification requires only approximately 3.06 ms for 200 categories and 0.15 ms for 20 categories. SVM training times varied by the number of categories and the number of features, ranging from less than a minute on the 10-way problems to about 72 hours for the 500-way structured SVM on a single CPU. We conducted our experiments on a small cluster of 60 nodes running the Hadoop open source map-reduce framework.

7. Summary

We have presented a means of creating large labeled image datasets from geotagged image collections, and experimented with a set of over 30 million images of which nearly 2 million are labeled. Our experiments demonstrate that multiclass SVM classifiers using SIFT-based bag-of-word features achieve quite good classification rates for large-scale problems, with accuracy that in some cases is comparable to that of humans on the same task. We also show that using a structured SVM to classify the stream of photos taken by a photographer, rather than classifying individual photos, yields dramatic improvement in the classification rate. Such temporal context is just one kind of potential contextual information provided by photo sharing sites. When these image-based classification results are combined with text features from tagging, the accuracy can be hundreds of times the random guessing baseline. Together these results demonstrate the power of large labeled datasets and the potential for classification of Internet-scale image collections.

Acknowledgments

The authors would like to thank Jacob Bank for his assistance with the experiments. This work was supported in part by NSF grants BCS-0537606, IIS-0705774, and IIS-0713185, and used the resources of the Cornell University Center for Advanced Computing, which receives funding from Cornell, New York State, NSF, and other agencies, foundations, and corporations.

References

[1] S. Arya and D. M. Mount. Approximate nearest neighbor queries in fixed dimensions. In *ACM-SIAM Symposium on Discrete Algorithms*, 1993.

[2] B. Collins, J. Deng, K. Li, and L. Fei-Fei. Towards scalable dataset construction: An active learning approach. In *ECCV*, 2008.

[3] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *PAMI*, 2002.

[4] K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *JMLR*, 2001.

[5] D. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. Mapping the world's photos. In *WWW*, 2009.

[6] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *ECCV Workshop on Statistical Learning in Computer Vision*, 2004.

[7] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL VOC 2008. http://www.pascal-network.org/challenges/VOC/voc2008/workshop/.

[8] J. Hays and A. A. Efros. IM2GPS: Estimating geographic information from a single image. In *CVPR*, 2008.

[9] T. Joachims. Making large-scale SVM learning practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*. 1999.

[10] X. Li, C. Wu, C. Zach, S. Lazebnik, and J.-M. Frahm. Modeling and recognition of landmark image collections using iconic scene graphs. In *ECCV*, 2008.

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

[12] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *CVPR*, 2008.

[13] B. Russel, A. Torralba, K. Murphy, and W. Freeman. LabelMe: A database and web-based tool for image annotation. *IJCV*, 77(1-3):157–173, 2008.

[14] F. Schroff, A. Criminisi, and A. Zisserman. Harvesting image databases from the web. In *ICCV*, 2007.

[15] Z. Stone, T. Zickler, and T. Darrell. Autotagging Facebook: Social network context improves photo annotation. In *1st IEEE Workshop on Internet Vision*, 2008.

[16] A. Torralba, R. Fergus, and W. T. Freeman. 80 million tiny images: A large dataset for non-parametric object and scene recognition. *PAMI*, 30(11):1958–1970, 2008.

[17] A. Torralba, R. Fergus, and Y. Weiss. Small codes and large databases for recognition. In *CVPR*, 2008.

[18] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support vector machine learning for interdependent and structured output spaces. In *ICML*, 2004.

[19] J. Winn, A. Criminisi, and T. Minka. Object categorization by learned universal visual dictionary. In *ICCV*, 2005.

[20] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *IJCV*, 2007.

[21] Y.-T. Zheng, M. Zhao, Y. Song, H. Adam, U. Buddemeier, A. Bissacco, F. Brucher, T.-S. Chua, and H. Neven. Tour the world: building a web-scale landmark recognition engine. In *CVPR*, 2009.