Predicting Geo-informative Attributes in Large-scale Image Collections using Convolutional Neural Networks

Stefan Lee Haipeng Zhang David J. Crandall
School of Informatics and Computing
Indiana University
Bloomington, IN USA
{steflee,zhanhaip,djcran}@indiana.edu

Abstract

Geographic location is a powerful property for organizing large-scale photo collections, but only a small fraction of online photos are geo-tagged. Most work in automatically estimating geo-tags from image content is based on comparison against models of buildings or landmarks, or on matching to large reference collections of geo-tagged images. These approaches work well for frequently-photographed places like major cities and tourist destinations, but fail for photos taken in sparsely photographed places where few reference photos exist. Here we consider how to recognize general geo-informative attributes of a photo, e.g. *the elevation gradient, population density, demographics, etc. of where it was taken, instead of trying to estimate a precise geo-tag. We learn models for these attributes using a large (noisy) set of geo-tagged images from Flickr by training deep convolutional neural networks (CNNs). We evaluate on over a dozen attributes, showing that while automatically recognizing some attributes is very difficult, others can be automatically estimated with about the same accuracy as a human.*

1. Introduction

Automatically organizing image collections is an important problem, especially with the recent explosive growth of online social photo-sharing websites — there are hundreds of billions of images on Facebook alone [3]. A natural way of organizing photos is based on the geospatial location of where they were taken. In fact, most modern online and offline photo organization tools, including Google's Picassa, Apple's iPhoto, and Yahoo's Flickr support browsing and search based on geo-tags. Of course, these features require photos to be geo-tagged; while GPS receivers are increasingly common on modern cameras (especially smartphones), for now only a small percentage of photos on the web are geo-tagged (*e.g*. less than 5% of Flickr [14]). We thus need content-based techniques to estimate geospatial annotations for unlabeled images.

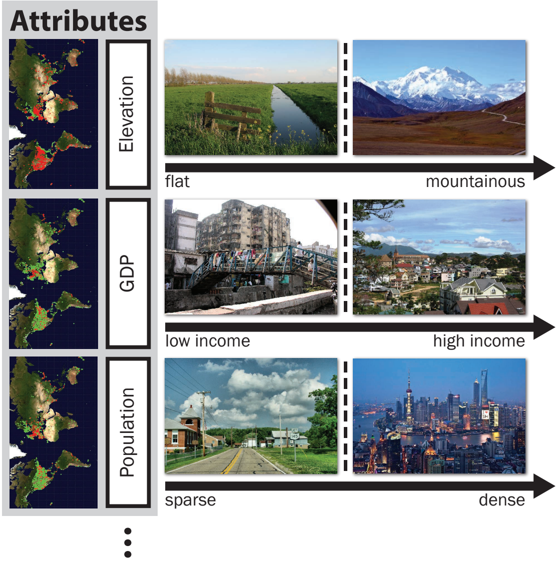

Figure 1: You probably can't identify exactly where these photos were taken, but you probably *can* estimate properties of the places, and then narrow down the possible locations accordingly based on your prior knowledge of the world. In this paper we estimate geo-informative attributes of images using deep classification networks trained on noisy but automatically-labeled datasets generated by combining public GIS maps and geo-tagged Flickr images.

Most recent work on geolocalizing consumer images takes one of two general approaches. The first is to build discrete models for a set of specific places, like buildings and other landmarks, and then to match unlabeled images against these models. The models vary in complexity from full 3D [6, 23, 36], to bags of feature points [22, 45], to hybrid approaches that use geometric verification of feature points [13, 21, 29, 30]. These approaches can produce precise geo-tags but require many training images for each place, and thus are only practical for major cities and landmarks. The second approach views geo-localization as an image matching problem: compare a query image to a large set of reference images having known GPS coordinates, and estimate a geo-tag based on the visual matches [15, 18]. These approaches avoid building discrete models and thus handle photos from a larger portion of the world, but require expensive searches against huge image libraries and the geo-tags they estimate are less precise.

These two types of techniques work well for photos of distinctive and highly-photographed places for which many reference images are available online. However, the geospatial distribution of online consumer photos follows a long-tailed distribution, such that a surprisingly large percentage of photos are taken in a few popular places, but the majority of photos are taken in places that are not often photographed [22]. Thus most images cannot be geo-located by either approach — there are simply not enough training images to densely cover the surface of the world. Moreover, many photos simply do not have enough information; a photo of a corn field could be taken almost anywhere in the midwest U.S., for instance. It is not practical to estimate a precise geo-tag for these images, and simply inferring the properties of the place where a photo was taken may be sufficient (*e.g.* rural, temperate, flat, not affluent, agricultural).

In this paper, we introduce the problem of classifying *geo-informative attributes* of where a photo was taken, including geographic properties like elevation and land type as well as demographics like wealth and population. These classifiers produce useful information about any image, even one taken in a place that has never been photographed before. The estimated geo-informative attributes could be used to add tags to photos, or they could (coarsely) estimate geolocation: once an image's attributes are recognized, a GIS map could be consulted to identify the set of places matching that specific combination of attributes. This use of mid-level features that are both visually distinctive and semantically meaningful in order to overcome insufficient training data is similar to the use of attributes in fine-grained object classification [2, 10, 27].

We learn classification models for over a dozen geo-informative attributes using millions of geo-tagged Flickr images, referencing the geo-tags against GIS maps to automatically produce (noisy) ground truth attribute labels. We apply deep convolutional neural networks (CNNs) to this problem, and compare their performance to standard techniques including both scene-level (*e.g.* GIST [25]) as well as local (vector-quantized HOG [7]) image features. We find that despite the large amount of noise in our training and test sets, the geo-spatial attribute classifiers perform nearly as well as humans on some attributes. Moreover, the CNNs perform significantly better than any of the other methods.

While some have considered specific scene classification tasks that could be considered geo-informative, including land use type [20], elevation gradient [15], and urbanicity [16], we believe ours is the first paper to propose general geospatial attribute recognition as an important task, and to evaluate the feasibility of geospatial attribute recognition on over a dozen attributes on a worldwide scale. Moreover while deep learning has been found to give impressive results on other classification problems, we are not aware of other work that has applied it to image geolocation.

To summarize our contributions, in this paper we:

1. propose attribute recognition as a means of estimating geo-informative annotations, even for photos taken in geographic areas with very sparse training data;
2. apply convolutional neural networks to recognize geo-informative attributes;
3. characterize and compare the effectiveness of classification techniques on this difficult new problem; and
4. introduce large-scale labeled datasets for geospatial attributes.

2. Related work

Our work connects to several lines of recent research.

Visual geo-localization. Recent papers have studied image geo-localization, typically using geo-tagged photos from photo-sharing sites like Flickr as (noisy) reference images [13,15,21–23,29,30,36,45]. Among those most related to ours, Hays and Efros [15] geo-locate photos by matching against a huge collection of geo-tagged images, and also infer population and elevation by looking up the estimated geo-tag on a GIS map. In this paper, we also estimate population and elevation (in addition to many other attributes), but by classifying attributes directly instead of geo-locating and then looking up the corresponding attribute values.

Model-based techniques like Li *et al* [22] avoid the high cost of explicit image matching by building models for each of thousands of discrete highly-photographed places on Earth. These techniques work well for popular landmarks, but cannot infer geo-tags for photos taken outside these places. Other work computes full 3D models of landmarks [36] which allow photos to be very precisely geo-tagged, sometimes within centimeters [6]. But these approaches require thousands of training images per place so are only feasible around popular landmarks, and 3d models cannot be built reliably for highly dynamic scenes.

Geo-informative attributes. Work in scene classification has considered categories that could be geo-informative, like urban versus rural [40]. Similarly, recent work by Zhou *et al.* [46] learns a suite of hand-selected scene classifiers as composites of many categories from the SUN attribute dataset [28]. The attributes capture different facets of a city including architecture, greenery, and transportation. They apply these classifiers on a dense corpus of social geotagged images to analyze the role of different attributes in city recognition and similarity tasks. These approaches and others like them use hand-selected categories and carefully-labeled training data, whereas we take a data-driven approach, learning over a dozen attribute classifiers using social images annotated with noisy training labels.

Leung and Newsam [20] and Xie and Newsam [41] reconstruct maps of land use type and "scenicness" by pooling visual features from images taken in a particular location, while Zhang *et al.* [44] similarly try to infer maps of weather patterns. We study a related but distinct problem: we try to classify geo-informative properties of individual images to help determine where they were taken, whereas they try to classify sets of images with known geo-tags to estimate land use properties of the physical world. Our scope is also broader: we test over a dozen attributes at a worldwide scale, whereas they study one attribute for part of the United Kingdom. Doersch *et al.* [9] discover local image patches characteristic of the architecture of particular cities by mining Google Streetview data. These patches can be thought of as a specific type of geo-informative feature, but are only applicable to street scenes in cities. Finally, identifying land features from aerial imagery is studied extensively in remote sensing [33]; we study the distinct (and arguably more difficult) problem of detecting attributes in unconstrained ground-based consumer images.

Visual attributes. Our challenge of training data sparsity is similar to that of fine-grained object recognition, like classifying between different bird species. *Attributes*, or mid-level features that are visually discriminative yet have semantic meanings (like "red beak" or "cluttered space") [2, 10,27], can alleviate problems of limited training data by allowing human experts to specify at least some portion of an object model, providing a connective "language" between computational models and human semantics. We propose a similar technique here: our goal is to produce a connection between worldwide GIS maps and visual features of individual images, through mid-level geo-informative attributes (elevation gradient, population density, etc.). In our context, an advantage of learning mid-level attributes (as opposed to directly learning low-level visual features to distinguish between places [9]) is that the estimated attributes may be useful for automatic annotation applications, even if precise geolocations cannot be estimated.

Convolutional neural networks. Following the amazing success of deep convolutional neural networks in the 2012 ImageNet [8] visual recognition challenge [19], CNNs have been applied to a variety of computer vision tasks with similar improvements over the state of the art [26, 37, 38]. The resurgence of neural networks in computer vision is thanks in no small part to powerful GPUs and large annotated datasets. The capability to process hundreds of thousands of images has been shown to be crucial to these networks. However, recent work suggests that training CNNs on large-scale supervised problems produces networks capable of richly modeling generic imagery [12,26,32,35,43]. It has been shown that starting from these pretrained models allows CNN-based techniques to be applied to a diverse set of target domains without the need for massive training sets. In this paper, we follow this methodology by beginning with a network trained on ImageNet and then specializing it for our attribute recognition on our own large-scale dataset.

3. Geo-informative attributes

Our main goal in this paper is to recognize geo-informative attributes, especially for photos without distinctive geo-spatial features or taken in sparsely-photographed places, where there is little hope of estimating an accurate geo-tag. We propose to learn classifiers for these attributes with large sets of geo-tagged image data from Flickr, generating ground truth labels automatically through GIS maps.

3.1. Datasets

Image data. We assembled a large collection of about 40 million geo-tagged images from Flickr, downloaded using the public API. From this set, we filtered out photos having imprecise geo-tags (with Flickr precision value less than 13, implying less accurate than about a city block). This collection of course exhibits the long-tailed spatial distribution discussed in Section 1, such that a large fraction of images come from a relatively small number of places. If we simply use the whole collection (or sample uniformly from it), we risk producing classifiers that simply memorize the appearance of a few key landmarks without abstracting general visual properties of places that exhibit various attributes.

We thus attempt to bias the sampling as if we were drawing uniformly at random over the surface of the globe, instead of sampling directly from the geo-spatial distribution of Flickr photos. To do this, we discretize the world into $0.01^\circ \times 0.01^\circ$ latitude-longitude bins (roughly 1 km × 1 km at the middle latitudes). We randomly sample photos one-by-one, but ignore samples from bins from which we already have 100 photos. To prevent individual highly-active users from introducing bias, we avoid sampling more than five photos from any single user. Finally, we partition the data into training and testing sets; to help prevent (nearly) identical photos from leaking across the partitions, we divide on a per-user basis (so that all photos from a single photographer are placed in one set or the other).

Geospatial attributes. In order to test geospatial attribute recognition, we collected public gridded GIS data for 15 at-

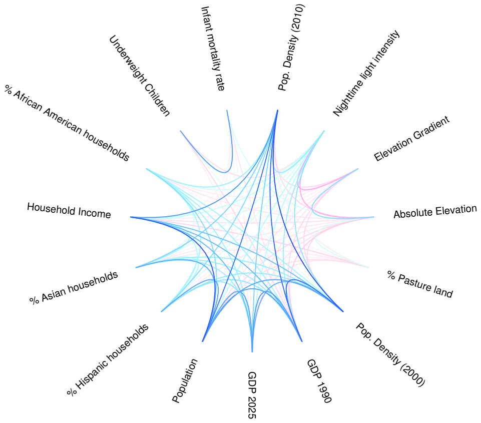

Figure 2: *Correlations between attributes* estimated from 100,000 samples. Color intensity is proportional to magnitude, with positive correlation in blue and negative in pink.

tributes including geographic features (*e.g.* elevation, elevation gradient and land-use) and demographic features (*e.g.* population density, wealth, ethnic composition). The data came from a variety of public sources including NASA and USGS, and the granularity of the gridded data ranged from about 30 arc-seconds to up to about 15 arc-minutes (see Table 1 for details). We used global data when available, although some of the attributes were available only for the U.S. We avoided time-varying attributes (*e.g.* temporal climatic attributes like daily temperature or rainfall); these attributes could be useful if accurate timestamps are known but we do not assume that here. Many of these attributes are correlated to some degree, as visualized in Figure 2, and thus we can expect some structure in the recognition accuracies across different attributes.

Automatically labeled datasets. We automatically produce labeled datasets for each attribute, by simply examining each photo's geo-tag, looking up the value of the attribute at that location in the gridded GIS map, and then assigning that label to the photo. Of course, this process is noisy: many geo-tags on Flickr are incorrect [14], and our worldwide attribute maps are coarse enough (about 1–10km square) that attribute values may vary dramatically even within a single bin. Attribute values are most geo-informative if they are relatively extreme — *i.e.* quite high or quite low. Thus we consider a restricted classification problem in which the goal is to label an image as having a high or low value for an attribute (*e.g.* high population or low population). We label images as high or low by thresholding at the highest and lowest quartile (25% and 75%) of the worldwide value of each attribute.

We partitioned the data into training, validation, and test sets (about 60%, 10%, 30% respectively) on a per-user basis to prevent leakage between the sets. We also ensured the class distribution in the test sets remained equal such that the random baselines are 50% for ease of interpretation.

3.2. Attribute classifiers

Inspired by the impressive results of convolutional deep learning on a variety of recognition problems (*e.g.* [19, 26, 37, 38] among others), we apply them to our problem of geo-informative attribute recognition. We start from the neural network architecture proposed by Krizhevsky *et al.* [19], with five convolutional layers (C1 to C5) followed by three fully connected layers (FC6 to FC8), and the same mechanisms for contrast normalization and max pooling. In total, this model has about 60 million parameters. Our training dataset is not sufficiently large to train such a large number of parameters from scratch, so we follow Oquab *et al.* [26] and others and initialize from a model pretrained on ImageNet. We modify the final fully connected layer (FC8) for each of our attribute classification problems such that it has two outputs (rather than the 1,000 of the original model) to account for our binary classification problem (estimating whether the attribute value is high or low). Additionally, the initial weights for FC8 are randomly sampled from a zero-mean normal distribution.

We used Caffe [1] for training and testing our networks, using the pretrained ImageNet network packaged with Caffe as initialization. The network for each classifier was trained independently using stochastic gradient descent with a batch size of 128 images. The learning rate was set at 0.001 and decreased by an order of magnitude every 2500 batches. The training process was allowed to continue for a maximum of 25,000 batches, but generally converged much earlier for our problems. To avoid overfitting, the validation set was evaluated every 500 batches and the weights with the lowest validation error were used for testing.

To put the CNN results in context, we also tested classifiers using several other recognition approaches. We constructed a bag-of-words vocabulary using Histogram of Oriented Gradients (HOG) [7]; our hypothesis was that local evidence such as particular types of objects might be helpful to predict geospatial attributes. Given an image, we sample 5×5 blocks of HOG cells to produce local feature vectors in overlapping square sub-images. We use the 31-dimensional variant of HOG [11], so that the feature dimensionality of each patch is $5 \times 5 \times 31 = 775$. We represent the image in the standard bag-of-words fashion as a histogram over these features quantized to a vocabulary. In this case, we constructed a 100,000 codeword vocabulary by clustering (with k-means) over 10 million HOG features sampled from random Flickr images. We then learned a linear SVM [17] for each of our 15 attributes, to predict high/low labels.

We also tested simpler global scene-level features, under the hypothesis that some geo-informative attributes could be inferred based on the overall appearance of a scene. We specifically used GIST [25] and spatially-pooled color histograms. For the histograms, we computed 8-bin histograms over each RGB plane within spatial regions of dif-

Attribute	**Source**	**Year(s)**	**Grid size**	**Area**	**Description**
Elevation	USGS GTOPO30 [39]	1996	30 arcsec	Global	Elevation according to USGS's global digital elevation model.
Elevation gradient	USGS GTOPO30 [39]	1996	30 arcsec	Global	Elevation gradient according to USGS's global digital elevation model.
GDP 1990	NASA SEDAC [42]	1990	15 arcmin	Global	GDP in millions of USD.
GDP 2025 (predicted)	NASA SEDAC [42]	2025	15 arcmin	Global	GDP in millions of USD.
Infant mortality rate	NASA SEDAC [4]	2000	2.5 arcmin	Global	Infant mortality rate.
Night light intensity	NOAA NGDC [24]	2009	30 arcsec	Global	Nighttime lights as seen from space, composited across the year.
Population density (2000)	NASA SEDAC [5]	2000	2.5 arcmin	Global	Population density.
Population density (2010)	NASA SEDAC [5]	2010	2.5 arcmin	Global	Population density.
% Underweight children	NASA SEDAC [4]	1990-2002	2.5 arcmin	Global	Percentage of children who are underweight.
% Pasture land	NASA SEDAC [31]	2000	0.5 arcmin	Global	Proportion of land areas used as pasture land.
% African American households	NASA SEDAC [34]	2000	30 arcsec	U.S.	Percentage of households identifying as African-American.
% Asian households	NASA SEDAC [34]	2000	30 arcsec	U.S.	Percentage of households identifying as Asian.
% Hispanic households	NASA SEDAC [34]	2000	30 arcsec	U.S.	Percentage of households identifying as Hispanic.
U.S. household income	NASA SEDAC [42]	2000	30 arcsec	U.S.	Aggregated household income in 2000, according to U.S. census.
U.S. population	NASA SEDAC [42]	2000	30 arcsec	U.S.	U.S. population according to 2000 census.

Table 1: *Details of the 15 attributes,* showing data sources and year(s), GIS grid size, and area covered. (There are 60 arc-minutes or 3600 arc-seconds in a degree, so 30 arc-seconds correspond to about 1 km at the middle latitudes.)

	# images	**CNN**	**HOG**	**Color**	**GIST**
Global attributes					
Elevation	14,230	**61.11**	56.50	53.34	52.92
Elevation gradient	13,266	**60.63**	55.86	53.76	52.44
GDP, 1990 actual	14,940	**71.33**	64.83	60.45	58.02
GDP, 2025 predicted	14,906	**73.58**	66.05	61.79	59.26
Infant mortality	14,634	**55.88**	52.88	52.63	50.92
Night light intensity	15,004	**73.61**	68.03	62.58	59.98
Population density, 2010	14,840	**74.43**	67.48	62.05	60.11
Population density, 2000	14,892	**72.38**	65.75	61.62	58.71
Underweight children	1,896	**62.88**	51.72	51.72	51.55
% Pasture land	14,972	**58.50**	54.62	54.54	52.99
U.S.-only attributes					
% African American	7,190	**65.42**	62.17	57.79	57.79
% Asian	7,006	**63.02**	58.99	57.48	56.54
% Hispanic	6,898	**65.65**	60.88	58.56	56.92
U.S. household income	6,900	**67.60**	64.55	58.12	57.61
U.S. population	6,866	**68.10**	64.76	61.17	59.22
Average		**66.27**	61.00	57.84	56.33

Table 2: *Classification accuracies* for geo-spatial attributes. Random baseline is 50%; see text for human baselines.

ferent sizes (in a spatial pyramid with three levels of 1 × 1, 2 × 2, and 4 × 4, yielding 502 dimensional feature vectors). As with HOG features, we then learned linear binary SVMs.

4. Experimental results

We tested our attribute classification techniques on the large-scale image and attribute datasets described in Section 3.1, using Convolutional Neural Networks as well as the baseline techniques discussed in Section 3.2.

4.1. Automatic attribute classification

The results of applying our classifiers on the 15 geospatial attributes are shown in Table 2, where again the task is to determine whether each image was taken in a place with a high or low value of the attribute — e.g. for the first row of the table, whether a given photo was taken at a low or high elevation. For each attribute, we normalized the test dataset such that a random baseline achieves 50% accuracy. We find that the correct classification rates vary significantly, from close to random guessing for infant mortality to nearly 75% correct classification for population density. This range reflects the difficulty of the geo-informative attribute tasks we have proposed: a photo full of buildings and people is obviously probably taken in a high-population area, whereas inferring infant mortality (which is a good correlate for poverty rate) requires more subtle analysis (*e.g.* examining architectural features, or the clothes people are wearing). Some of these attributes are correlated and thus show similar performance, although we do see interesting differences amongst them: we can predict estimated GDP for 2025 more accurately than in 1990, presumably because Flickr images were mostly taken in the last 5 years, whereas the worldwide wealth distribution has changed dramatically since 1990 (*e.g.* China's GDP has increased by an order of magnitude). Figure 3 shows randomly-sampled correctly and incorrectly classified images for each attribute.

For all of the attributes, we found that the deep learning CNNs beat the other techniques by a decisive margin. GIST and color features had an average accuracy of 56.33% and 57.84%, respectively, compared to a 50% random baseline. This confirms the hypothesis that some attributes can be (weakly) estimated based only on the overall properties of the scene. Using HOG features improved results significantly to 61.0%, suggesting that local object-level features help, while the CNNs yielded a substantial further improvement to 66.3%. Our results thus add to the growing evidence that deep learning can yield large improvements over traditional techniques on many vision problems.

We are not aware of other work that has studied geo-informative attribute classification, so we cannot compare against published results. Perhaps the closest paper is Leung *et al* [20], who try to reconstruct land use maps by analyzing pools of geo-tagged photos from Flickr — a very different task than our goal of labeling images. Though not directly comparable, as a weak comparison we note that we achieve greater accuracy relative to our baseline: they report

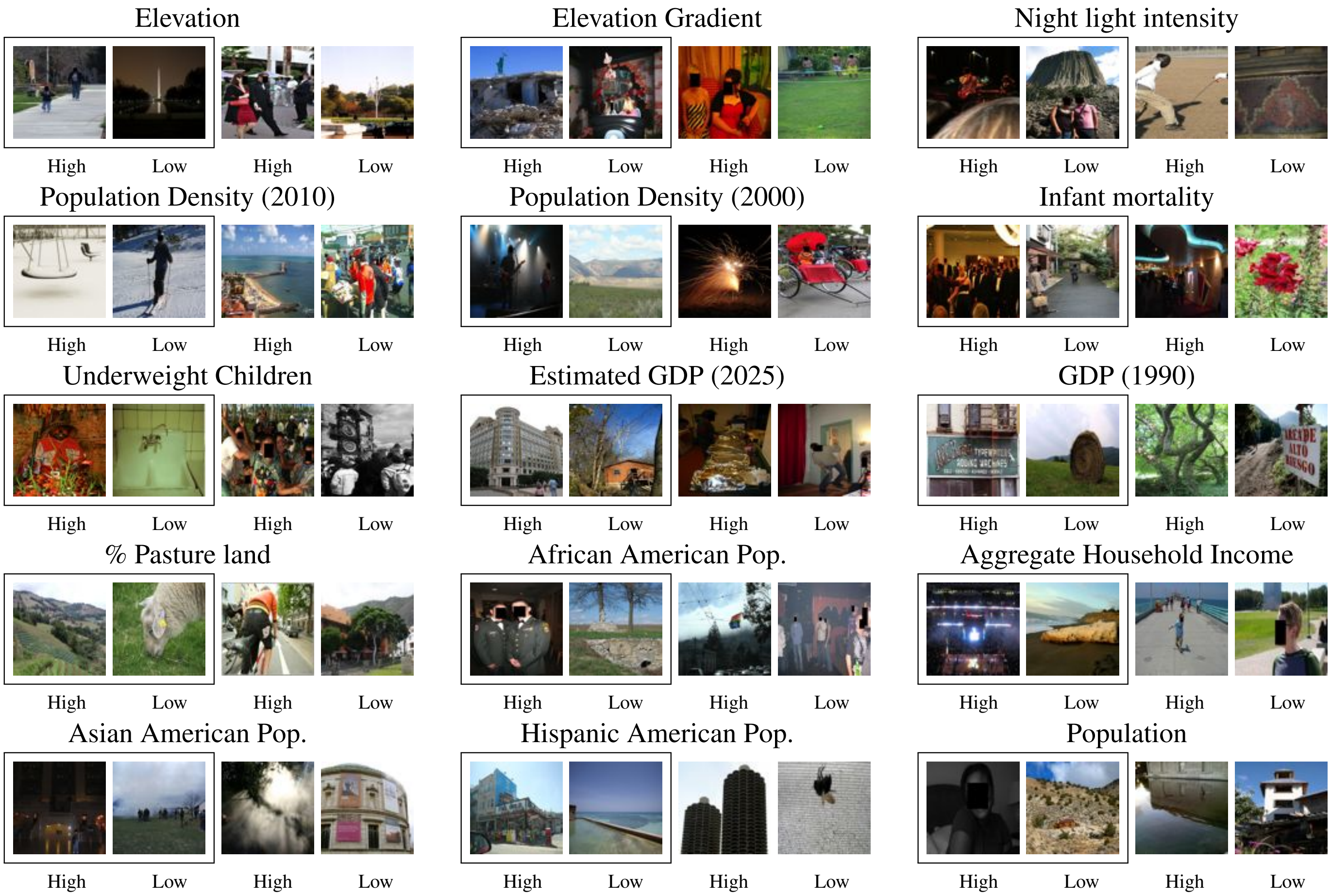

Figure 3: *Some correctly- and incorrectly-classified images*. For each attribute, we show a correctly-predicted high- and low-valued image inside the box, and an incorrectly-predicted high- and low-valued image outside the box to the right.

64% accuracy versus a 61.1% random baseline for urban development classification, while we achieve 73.6% accuracy on our similar "night light intensity" attribute versus 50%. Again, our tasks are very different so a direct comparison is not meaningful, but this at least suggests that our improvement over baseline is state-of-the-art.

4.2. Human baselines

Although the automatic classifiers beat the random baseline by substantial margins, our accuracies are not near the 100% performance we might aspire to. However it is important to note that our test dataset is extremely difficult, consisting of a raw set of Flickr photos; we have deliberately made no attempt to filter out difficult or noisy images (because doing so could inevitably inject biases into the dataset). Thus many of our test set photos, including indoor images and close-ups of objects, are difficult or impossible to geo-locate because there is simply not enough visual evidence to detect any geo-informative attributes. Moreover, the ground truth labels themselves are noisy, as a significant fraction of Flickr geo-tags are wrong [14].

The sample of correctly and incorrectly classified images for each attribute shown in Figure 3 gives a sense for the difficulty of our dataset, and the limited amount of evidence that some images contain. For instance, in the upper right of the figure, the classifier correctly estimates that the photo of a concert probably occurs in a city and the photo of a mountain is in a rural area. But it decides that the fencers are in the country and the art is in the city. These are very reasonable decisions based on the visual evidence at hand, but turn out to be incorrect.

To try to quantify the fraction of these difficult images, we collected hand-labeled annotations for 3 attributes (population, income, and elevation gradient) to measure human performance on these classification tasks. We chose these attributes because they were the easiest to describe to users. For each attribute, we sampled 1,000 images from our dataset such that half had a high value of the attribute (e.g. high population density) and the other half had a low value according to the automatic labeling. We presented each image to two users on Mechanical Turk (restricting to "Masters" who have a long track record of quality work), asking them to classify the image into the low or high category and to provide some additional feedback.

We found that human performance ranged from 52.9% for poverty, to 60.0% for elevation gradient, to nearly 81% for population density. Our automatic classifiers actually beat human performance on poverty (taken as a proxy for infant mortality — 55.9% versus 52.9%), while achieving about the same performance on elevation gradient (60.6% versus 60.0%). However the human users performed significantly better on population density (80.8% versus 73.61%). Thus while our automatic classifiers do not get near 100% accuracy, neither do humans. One reason for this is that about 28% of our dataset is indoor images, which typically have very little evidence about geo-spatial attributes.

4.3. Discussion

The geo-informative attributes we detect could have a variety of uses in automatic image organization. For instance, instead of using geospatial organization techniques that require absolute GPS coordinates, consumer software like Picasa and iPhoto could arrange photos according to *relative* geo-spatial attributes, separating urban from rural images, mountains from plains, and so on. Another possibility would be to use the attributes to narrow down where on Earth a particular photo was taken by consulting GIS maps. For instance, knowing that a photo was taken in a high-population, low-income, high-elevation place already restricts the set of possible GPS locations dramatically, and using GIS maps we can determine this set even if we have no photos from the specific place in our training dataset. For instance, we have calculated that if our attribute classifiers could give 80% classification accuracy, then we could correctly narrow the geotag of about 10% of images in our dataset to a 100×100 km range, or about 65% of photos to within about a 500×500 km range.

We have so far posed our task as classifying whether an image has a high or low attribute value in order to avoid trying to draw precise boundaries between bins. However, this has the disadvantage that we are not able to predict attributes for images in about 50% of locations — the middle 50% of a given attribute. Using the same automatically collected dataset, we also trained versions of our classifiers modified to treat the problem as a ternary classification task, predicting an image as either low, average, or high (*i.e.* in the lower, middle two, or upper quartile). This is a much harder problem because images with 24th- and 26th-percentile attribute values have different ground truth labels, despite the fact that they could be taken in virtually identical places. On this ternary task, average accuracy across the 15 attributes was 44.08% relative to a 33.33% baseline. Despite this being a more difficult task, average accuracy relative to random chance remained the same as in the binary experiments (both having accuracies approximately 1.32 times baseline). Interestingly, simple binary experiments which split at the median performed well below either the binary or ternary experiments discussed, leading us to believe that a vast majority of images have similar characteristics until they reach the extreme values.

As discussed in Section 3.1, many of our attributes are correlated — areas with high infant mortality typically also have a high rate of underweight children, for example. An interesting question is whether these correlated attributes are predicted from similar visual features. To begin to form an answer, we trained a single Convolutional Neural Network (instead of 15 separate networks) to predict all 15 binary responses as a multi-label framework. Performance for this design was nearly identical to the independent classifiers despite the possibility for the network to learn and use these correlations. We take this as a weak indication that perhaps each attribute is identifying different visual cues despite being correlated geospatially. From a practical standpoint, using a single network has the advantage of requiring less computation during image classification.

5. Summary and Conclusion

We have proposed the problem of estimating geo-spatial attributes of the place where a photo was taken, based only on its visual content. We learned convolutional neural network classifiers for a wide variety of geo-spatial attributes by building large (albeit noisy) datasets by combining geotagged Flickr photos with attribute values from GIS maps. We evaluated the performance of the CNNs against more traditional scene-level and local features. While the CNNs give the best performance, we find that the local features outperform the simpler scene-level features by a significant degree, suggesting that the classifiers have discovered local features (like objects) that are predictive of attribute values. We believe that this is the first paper to propose general geo-spatial attribute recognition as an important task, to apply deep learning techniques to problems related to geolocalization, and to evaluate the feasibility of geo-spatial attribute recognition on over a dozen attributes and on a worldwide scale. We hope that this paper and our dataset will spark interest among other researchers into the problem of geo-spatial attribute classification.

Acknowledgments. The authors thank Alex Seewald and Dennis Chen for configuring and testing software for deep learning, Prashant Shiralkar for helping to collect GIS data, and Bingjing Zhang for implementing large-scale k-means clustering for building the HOG vocabulary. The research was supported in part by the Intelligence Advanced Research Projects Activity (IARPA) via Air Force Research Laboratory, contract FA8650-12-C-7212 and by the National Science Foundation through CAREER grant IIS-1253549. We used the high-performance computing resources of Indiana University which are supported in part by NSF (ACI-0910812 and CNS-0521433), the Lilly Endowment, Inc., and the

Indiana METACyt Initiative. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, AFRL, NSF, or the U.S. Government.

References

[1] Caffe. `http://caffe.berkeleyvision.org/`.

[2] T. Berg, A. Berg, and J. Shih. Automatic attribute discovery and characterization from noisy web data. In *ECCV*, 2010.

[3] J. Bort. Facebook stores 240 billion photos and adds 350 million more a day. In *Business Insider*, Jan. 2013.

[4] CIESIN/Columbia University. Poverty Mapping Project, 2005.

[5] CIESIN/Columbia University, and Centro Internacional de Agricultura Tropical (CIAT). Gridded Population of the World, Version 3 (GPWv3): Population Density Grid, 2005.

[6] D. Crandall, A. Owens, N. Snavely, and D. Huttenlocher. SfM with MRFs: Discrete-continuous optimization for large-scale structure from motion. *PAMI*, 35(12), 2013.

[7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[9] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. Efros. What makes Paris look like Paris? In *SIGGRAPH*, 2012.

[10] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR*, 2009.

[11] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 32(9), 2010.

[12] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *arXiv preprint 1311.2524*, 2013.

[13] Q. Hao, R. Cai, Z. Li, L. Zhang, Y. Pang, and F. Wu. 3d visual phrases for landmark recognition. In *CVPR*, 2012.

[14] C. Hauff. A study on the accuracy of Flickr's geotag data. In *SIGIR*, 2013.

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

[16] G. Heitz, S. Gould, A. Saxena, and D. Koller. Cascaded classification models: Combining models for holistic scene understanding. In *NIPS*, 2008.

[17] T. Joachims. Making large-scale support vector machine learning practical. In *Advances in Kernel Methods*, 1999.

[18] E. Kalogerakis, O. Vesselova, J. Hays, A. A. Efros, and A. Hertzmann. Image sequence geolocation with human travel priors. In *ICCV*, 2009.

[19] A. Krizhevsky, I. Sutskever, and G. Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*, 2012.

[20] D. Leung and S. Newsam. Proximate sensing: Inferring what-is-where from georeferenced photo collections. In *CVPR*, 2010.

[21] 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.

[22] Y. Li, D. Crandall, and D. Huttenlocher. Landmark classification in large-scale image collection. In *ICCV*, 2009.

[23] Y. Li, N. Snavely, D. Huttenlocher, and P. Fua. Worldwide pose estimation using 3d point clouds. In *ECCV*, 2012.

[24] NOAA National Geophysical Data Center. Version 4 DMSP-OLS nighttime lights time series.

[25] A. Oliva and A. Torralba. Building the gist of a scene: the role of global image features in recognition. In *Progress in Brain Research*, 2006.

[26] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *CVPR*, 2014.

[27] D. Parikh and K. Grauman. Interactively building a discriminative vocabulary of nameable attributes. In *CVPR*, 2011.

[28] G. Patterson and J. Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *CVPR*, 2012.

[29] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *CVPR*, 2007.

[30] R. Raguram, J. Tighe, and J.-M. Frahm. Improved geometric verification for large scale landmark image collections. In *BMVC*, 2012.

[31] N. Ramankutty, A. Evan, C. Monfreda, and J. Foley. Global Agricultural Lands, 2010.

[32] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. CNN features off-the-shelf: an astounding baseline for recognition. *arXiv preprint 1403.6382*, 2014.

[33] O. Rozenstein and A. Karnieli. Comparison of methods for land-use classification incorporating remote sensing and GIS inputs. *Applied Geography*, 31(2):533–544, Apr. 2011.

[34] L. Seirup and G. Yetman. U.S. census grids, 2000.

[35] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint 1312.6229*, 2013.

[36] N. Snavely, S. M. Seitz, and R. Szeliski. Modeling the World from Internet Photo Collections. *IJCV*, 80(2), 2008.

[37] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, 2013.

[38] A. Toshev and C. Szegedy. DeepPose: Human pose estimation via deep neural networks. *arXiv preprint 1312.4659*, 2013.

[39] United States Geological Survey. Global 30 arc-second elevation (GTOPO30).

[40] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010.

[41] L. Xie and S. Newsam. IM2MAP: deriving maps from georeferenced community contributed photo collections. In *SIGMM International Workshop on Social Media*, 2011.

[42] G. Yetman, S. R. Gaffin, and X. Xing. Global 15x15 minute grids of GDP based on the SRES B2 scenario, 2004.

[43] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *ECCV*, 2014.

[44] H. Zhang, M. Korayem, D. Crandall, and G. LeBuhn. Mining photo-sharing websites to study ecological phenomena. In *WWW*, 2012.

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

[46] B. Zhou, L. Liu, A. Oliva, and A. Torralba. Recognizing city identity via attribute analysis of geo-tagged images. In *ECCV*, 2014.