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

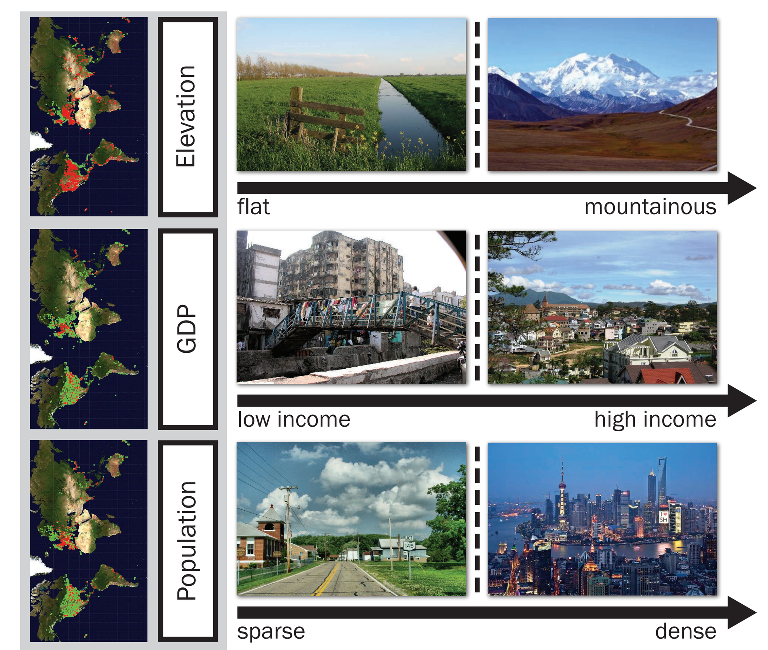
Stefan Lee, Haipeng Zhang, David Crandall
School of Informatics and Computing, Indiana University, Bloomington, IN

1. Overview

- Geographic position is useful to organize photos, but **most photos (~95% of Flickr) lack geo-tags.**
- Others have studied automatic geo-tagging using huge collections of geo-tagged reference images (e.g. [1],[2],[3],…).
- But most photos **are not from distinctive landmarks** or **densely photographed areas,** so matching may be hopeless.

*Can you figure out where these random Flickr photos were taken?**

- Instead, **we estimate** geo-informative properties **of the scene,** that could narrow down position using GIS maps,
  - letting us potentially geo-locate images even in places that have never been photographed before!



- Specifically, we:
  - build large-scale geo-informative attribute datasets **combining Flickr images and public GIS maps;**
  - learn models for **geo-informative attributes** with CNNs; and
  - evaluate on realistic, large-scale image collections.

2. An automatically labeled dataset

- From 200 million public geotagged Flickr photos, we sampled **~50,000 images attempting to avoid biases:**
  - Sampling is spatially uniform (i.e. not biased towards cities)
  - Limit contribution of any single photographer
  - No manual filtering based on content, position, etc.
- Also collected **publicly-available GIS attribute maps.**
  - Global or continent (North America) scale
  - Includes binned geographic, demographic, economic, agricultural attributes
- For each Flickr image, we look up its attribute in the GIS map, to produce **a labeled geo-informative attribute dataset.**

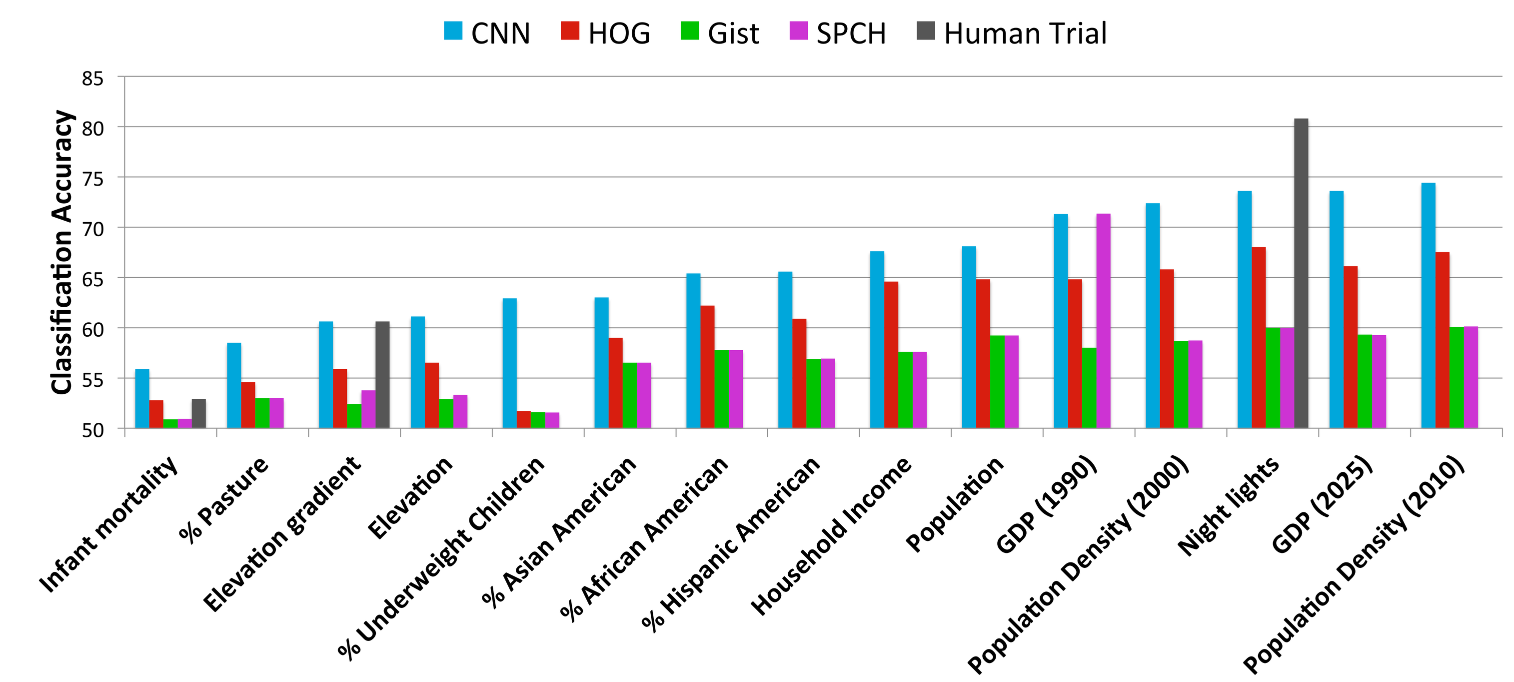
***Top row:** Istanbul, New Orleans, Kyoto, Portland; **Bottom row:** Carolina Beach (North Carolina), Berkeley, Kirkwood (Missouri), Surf City (North Carolina), Phang-Nga (Thailand), Solvang (California)

3. Estimating geo-informative attributes

- **Goal: Given an image, estimate its geo-informative attribute values,** using models built from training data.
  - Specifically a binary problem for each attribute: high vs low
- We train Convolutional Neural Networks for this task.
  - **Fine-trained from AlexNet** [4]
  - Training via stochastic gradient descent with **Caffe** [5]
  - Iterate until performance stagnated on validation set
- Compare against several baselines:
  - Multiple CNNs vs joint prediction with single multi-label net
  - BoW HOG, GIST, and spatially pooled color histograms (SPCH) with linear SVMs
  - Human (Mechanical Turker) performance

4. Results

- Accuracy on binary prediction (50% random baseline):



  - Individual and joint nets had about the same accuracy.
  - Also tested ternary (vs binary) labeling problem; mean accuracy was 44.08% (vs 33% random baseline)
- Sample correct (boxed) and incorrect results:

- Summary and conclusions:
  - **Propose geo-informative attributes** to help geolocate the (many) photos that cannot be matched.
  - **Build labeled datasets** using geo-tagged images and GIS maps.
  - **CNNs outperform other techniques**, sometimes even humans!

[1] J. Hays and A. Efros. IM2GPS: estimating geographic information from a single image. In *CVPR*, 2008.
[2] X. Li, C. Wu, C. Zach, S. Lazebnik, and J. Frahm. Modeling and recognition of landmark image collections using iconic scene graphs. In *ECCV*, 2008.
[3] Y. Li, N. Snavely, D. Huttenlocher, and P. Fua. Worldwide pose estimation using 3d point clouds. In *ECCV*, 2012.
[4] A. Krizhevsky, I. Sutskever, and G. Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*, 2012.
[5] Caffe. http://caffe.berkeleyvision.org/

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Air Force Research Laboratory, contract FA8650-12-C-7212. 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, or the U.S. Government.