Visual Census: Using Cars to Study People and Society

Timnit Gebru[1] Jonathan Krause[1] Yilun Wang[1] Duyun Chen[1] Jia Deng[2] Li Fei-Fei[1]
[1]Stanford University [2]University of Michigan
{tgebru,jkrause,yilunw,duchen,feifeili}@cs.stanford.edu jiadeng@umich.edu

1. Introduction

The cars people own can provide significant personal information: by knowing that the person in Fig. 1 drives a Prius we can guess that he or she is probably from San Francisco and earns an income of approximately $72k/year. A few pioneering works by Zhou *et al*., Ordonez *et al*., and Naik *et al*. have started to apply visual scene analysis techniques to infer characteristics of neighborhoods and cities [8, 4, 6, 7]. In this work, we are also interested in using images to understand cities, neighborhoods and the demographic makeup of their inhabitants. However, instead of using global image statistics, we achieve this goal by detecting and classifying cars on the street (Fig. 2). 95% of American households own cars [1], and as seen in Fig. 1 cars give a lot of information about individuals as well as neighborhoods.

Our contributions are two-fold. **First**, we offer the largest fine-grained dataset to date, composed of 2,657 car classes, nearly all car types produced in the world after 1990, with a total of 712,430 images from websites such as edmunds.com, cars.com, craigslist.com and Google Street View. We use our dataset to train a large scale fine-grained detection system, detecting cars in more than 45,000,000 Google Street View images collected from 200 of the largest American cities. **Second**, we present a suite of interesting social analyses of American people and cities using our car detections. In Sec. 4, we show that from a single source of data, Google Street View images, we are able to predict diverse sets of important societal information such as voting patterns, crime rates, median household income, race and education levels.

2. A Large Scale Fine-Grained Car Dataset

In order to analyze society via the cars visible in it, one must first create a dataset of all possible cars we would like to study. However, this poses a challenge. What are all of the cars in the world, and how can we possibly collect data for each one of them? We leverage existing knowledge bases of cars, downloading images and data for roughly 18,000 different types of cars from the website edmunds.com. We then grouped these car types into sets of indistinguishable classes and collected more images for each class from both craigslist.com and cars.com as well as one bounding box for each image from Amazon Mechanical Turk (AMT). This yielded 2,657 visual groups of cars for us to perform an analysis with, which to our knowledge represents the largest fine-grained dataset in the community.

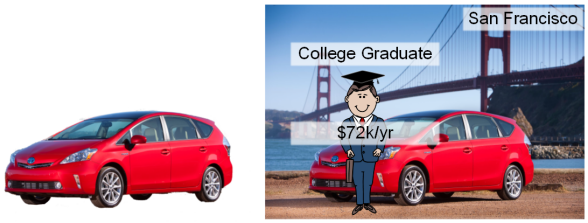

Figure 1: Look at the car on the *left*. It is a Prius 2012. Can you guess where the owner lives? What his or her household income and highest education level is? Whether he or she votes republican or democrat? In this paper, we use fine-grained car detection algorithms with census data provided by the U.S. government, and are able to infer that the Prius is the most popular car in the city of San Francisco, its owner often holds a college degree and earns a household income of $72,000 (*right*).

Google Street View images, though, our source of imagery on the street level, may contain multiple cars per image, each of which can be occluded and low resolution, presenting a challenge for recognition. Thus, as part of our fine-grained car dataset we also collected bounding boxes for nearly 400,000 cars in Street View images, of which 69,562 were annotated by expert human labelers with one of the 2,657 fine-grained classes, allowing us to train detectors and classifiers effective on such real-world images. Finally, to perform our visual census, we collected 45 million Google Street View images, sampling 8 million GPS points in 200 large US cities every 25m, with images from 6 camera rotations captured per point.

3. Fine-Grained Detection and Classification

What is the best way to accurately detect and classify fine-grained car classes in 45 million Street View images? R-CNN [3] has seen impressive performance in detecting and classifying fine-grained categories. However, by our estimates, an R-CNN takes roughly 20 seconds to run on a

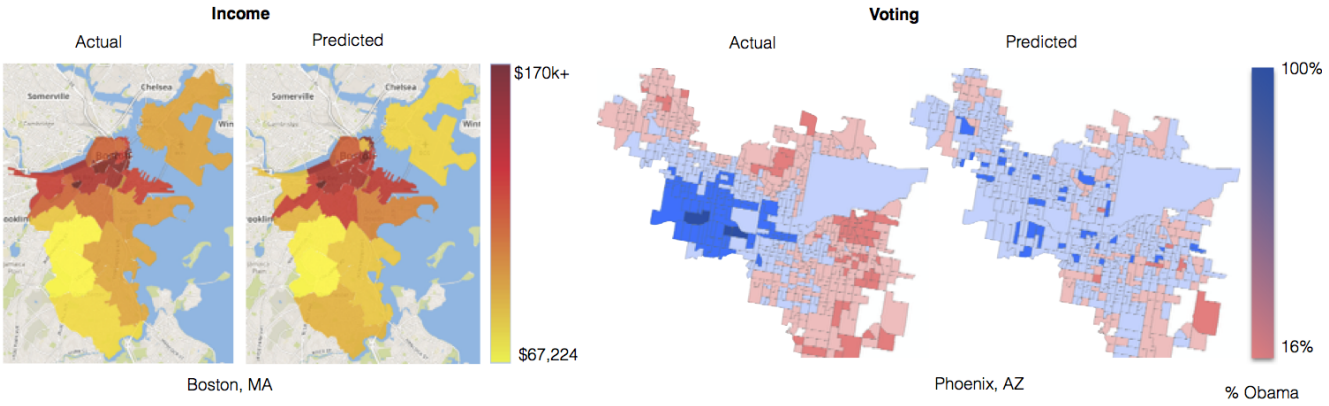

Figure 2: Examples of predicted demographic variables using attributes of cars detected from Google Street View images. Left: Actual and predicted median household income values in Boston, MA. Right: Actual and predicted voting results in Phoenix, AZ for the 2008 presidential election.

single Street View image using a high-performance GPU, so running on all of our Street View images would take 28 GPU *years*. Instead, we use the efficient and CPU-based deformable part model (DPM) [2] for detecting cars, and then evaluate the top detections with a convolutional neural network (CNN) [5].

Even within the scope of deformable part models, there is a wide range of parameters that can have large effects on both speed and accuracy. After an extensive analysis, varying the number of parts and components, we decided upon a single component DPM with 8 parts, achieving an AP of 64.2% at 5 seconds per Street View image. For comparison, the highest DPM AP was 68.7%, achieved with 5 components and 8 parts, but took 22 seconds per image. On 200 2.1 GHz cores, we can run detection on our entire dataset in less than two weeks.

To classify our car detections into one of the 2,657 fine-grained car classes, we use a convolutional neural network with an architecture following [5]. Since a majority of our training images take the form of product shots rather than Street View images, we apply deformations to the product shot images, such as blurring, to make them appear similar to the more challenging images from Street View. At test time we use the top 10% scoring DPM bounding boxes to decrease the amount of time needed to classify detections, which drops detection AP by 2, but also speeds up classification by an order of magnitude. On ground truth bounding boxes, this CNN achieves a remarkable classification accuracy of 31.27%, and on true positive DPM detections, which are generally of better image quality than average Street View boxes, performance is 33.27%.

4. Visual Census

We use the attributes of cars detected from Google Street View images to train regressors and predict demographic variables such as median household income, race, education level, voting patterns and crime rates (Fig. 2). Our predictions correlate well with ground truth data (Pearson r=0.70 for median household income, r>=0.76 for all races and r=0.67 for voting patterns). We also quantify the relationship between demographic variables and cars. For example, we learned that vans are the most highly correlated car attribute with crime, a 1% increase in the percentage of Cadillacs in a zip code predicts a 23% increase in the number of black people and those who drive pick up trucks with crew cabs are 6% less likely to vote for Obama than those who do not.

References

[1] R. Chase. Does everyone in america own a car? @ONLINE, June 2009. 1

[2] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010. 2

[3] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *arXiv preprint arXiv:1311.2524*, 2013. 2

[4] A. Khosla, B. An, J. J. Lim, and A. Torralba. Looking beyond the visible scene. 1

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 2

[6] N. Naik, J. Philipoom, R. Raskar, and C. Hidalgo. Streetscore–predicting the perceived safety of one million streetscapes. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on*, pages 793–799. IEEE, 2014. 1

[7] V. Ordonez and T. L. Berg. Learning high-level judgments of urban perception. In *Computer Vision–ECCV 2014*, pages 494–510. Springer, 2014. 1

[8] B. Zhou, L. Liu, A. Oliva, and A. Torralba. Recognizing city identity via attribute analysis of geo-tagged images. In *Computer Vision–ECCV 2014*, pages 519–534. Springer, 2014. 1