Using fine-grain detection to understad society

Anonymous CVPR submission

Paper ID ****

Abstract

Detecting a large number of BMWs in images informs us that those images may be of a wealthy area. Conversely, knowing that our images were obtained from a wealthy neighborhood increases the likelihood of detecting expensive cars. We explore this relationship between demographic factors and fine-grain classes by performing large scale detection of over 2600 car classes and conducting a social analysis of unprecedented scale in computer vision. Using 45 million images from 200 of the biggest cities in the United States,we predict demographic factors such as neighborhood wealth and crime statistics. Finally we show that just as fine-grain classes provide demographic information, societal cues can assist in fine-grain classification and improve accuracy. To facilitate our work, we have collected the largest and most challenging fine-grain dataset reported to date consisting of 3147 classes of cars comprised of images from google street view and other web sources and classified by car experts to account for even the most subtle of visual differences. We hope our work ushers in a new research area fusing fine-grained object detection and societal analysis.

1. Introduction

The ubiquity of street view images has jumpstarted a new line of computer vision research focused on understanding cities through images [6] [4] [5]. For example, [6] showed that crime predictions can be improved by incorporating human perceptions of neighborhoods' images rather than using census data such as income alone [6] and [5] and [4] learn these perceptions using computer vision techniques. However, in order to extend these methods to other cities,extensive annotations of millions of images from each city would be required since, as [5] showed, algorithms trained on images of Boston, for example, cannot predict safety or wealth on images from San Francisco. We explore the question of learning social priors using large scale fine-grain classifications of cars and show that many neighborhood statistics such as income and crime rate can be predicted from car detections. Furthermore, using our detections in conjunction with census data, we can answer questions like what types of cars do rich/poor people drive?

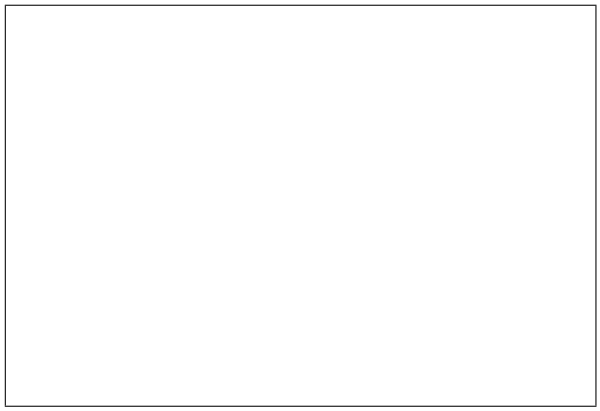

Figure 1. some pull figure

Finally we show that we can use the answers to these questions to help improve fine-grained classification. Although an increasing number of images that we interact with daily are associated with GPS tags, there are very few computer vision algorithms that take advantage of location based metadata. This metadata can be especially important in fine grain classification. For example, just as detecting a large number of expensive cars in one area can give us a hint that we are in the vicinity of a wealthy neighborhood, knowing that we are in a wealthy neighborhood can also increase our likelihood of detecting expensive cars. Similarly, knowing that we are in a farm area increases our likelihood of detecting farm related cars and seeing many family households with young children increases our likelihood of detecting SUVs. We show that this information can be leveraged to improve fine-grain classification. Although there has been previous work on learning spatio-temporal priors for fine-grain classification [2] and exploiting street view geometry and GIS systems to improve object detection [3, 1] to our knowledge this is the first time census data and other social cues have been used to assist in fine-grain classification.

Summarizing our contributions:

1. We perform a large scale analysis of cities using our car detections and present intuitive as well as interesting insights
2. We show that using social cues extracted from census data can improve fine-grain classification accuracy
3. We present the largest fine-grain car dataset reported to our knowledge, complete with geotags and class as well as geography metadata
4. We include a larger set of 45 million street view images with car detections and fine-grain class predictions

2. Related Work

Analysis of cities using images.

1. Plos one journal from MIT asking people to predict whether an area is safe/wealthy etc... after looking at the images
2. streetscore MIT paper predicting safety wealth scores etc.. just from images [4]
3. tamara Berg's paper on safety on ECCV [8] [7]

   Using GPS data to improve object detection.
4. Amir's work in GIS assisted object detections. For objects like streetlamps and trashcans, uses GIS to reproject objects to a plane and reduce the search space for object detection .
5. NYC 3D uses geographic elevation data to create view-point aware detectors and extract ground planes for them
6. Birdsnap fine-grain dataset with spatio-temporal prior for birds

3. Cars and Cities dataset

Our dataset consists of W number of images annotated with bounding boxes and fine-grained classes for training and validating fine-grained car detectors and 45 million street view images for societal analysis.

3.1. Images with Labeled fine-grained classes

Out of the images annotated with fine-grained classes and bounding boxes, X were obtained from google street view, Y from craigslist.com and Z from cars.com. A of our images have bounding boxes for cars and B are annotated with fine-grain labels. The bounding boxes were obtained through a series of AMT tasks. The fine-grain labels were created by first coming up with a class list of 18000 cars comprising of all cars listed on edmunds.com and grouping them into visually indistinguishable sets of groups using a series of amazon mechanical turk tasks as well as manual labor by the authors. After creating an exhaustive class list of 3147 classes, images from craigslist and cars.com were labeled by parsing the posting titles while cars from google street view were labeled using 100 hired car experts.

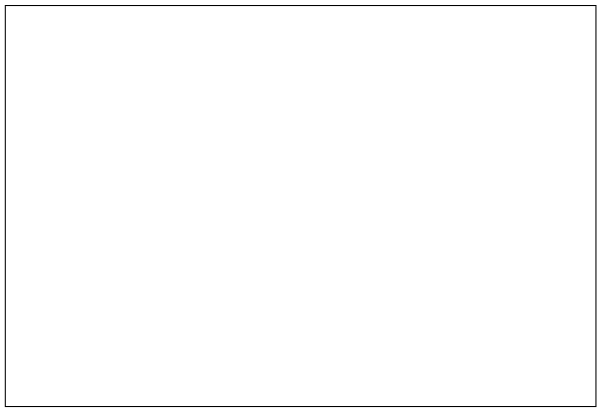

Figure 2. Example images from our data

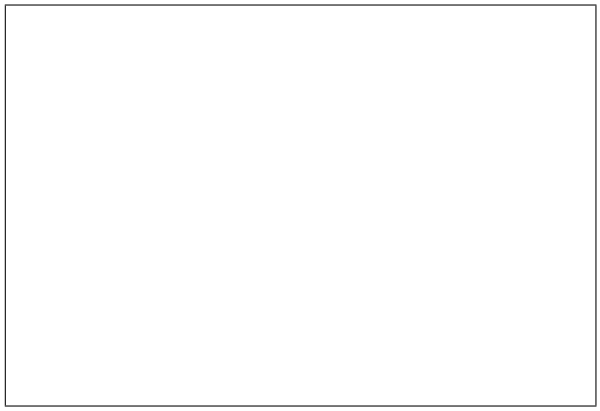

Figure 3. Image statistics from our data

Fig. 2 shows example images from our data while Fig. 3 shows some image statistics. Images from craigslist.com and cars.com have one large bounding box whereas google street view images have multiple small boxes with cars that are blurred and occluded. As shown in Fig. 4 the different fine-grain classes are very difficult to distinguish

3.2. Images with no labeled fine-grained classes

In order to perform societal analysis, we collected 45 million google street view images from 8 million points in 200 of the biggest cities in the United States. These images were collected by sampling latitude,longitude points on roads, spaced 25 meters apart. Fig. 5 shows maps from two cities with the samples that were collected and fig. 6 shows the number of samples for the 10 biggest cities. For

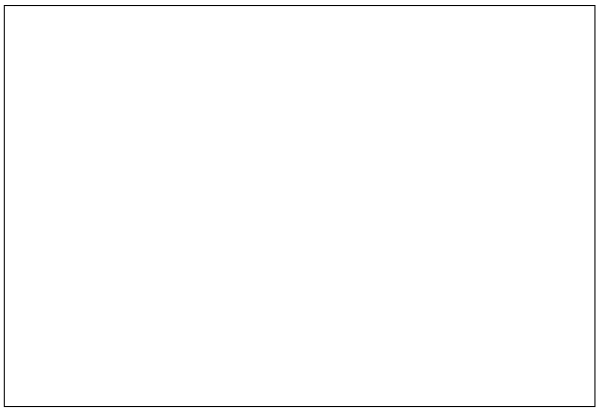

Figure 4. Some fine-grain car classes are very visually similar

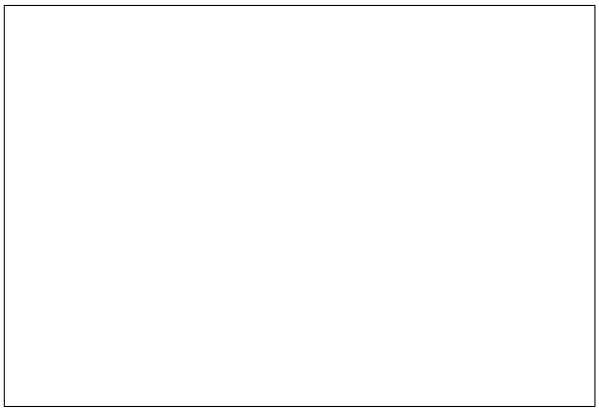

Figure 5. Maps of Boston and San Francisco, each dot represents a sample image in our data

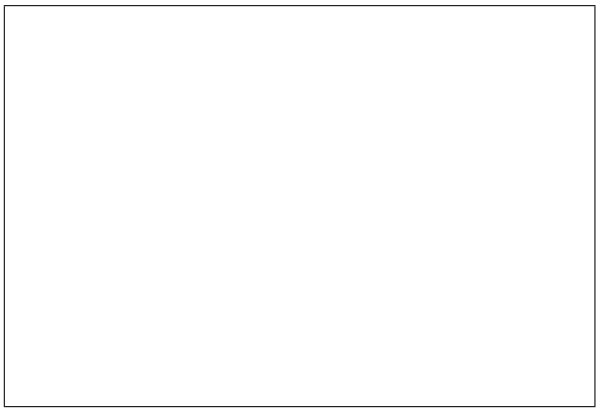

Figure 6. The number of latitude, longitude samples for 10 cities with the most number of samples

each sample, we collect images at 0,60,120,180,240,300 degrees.

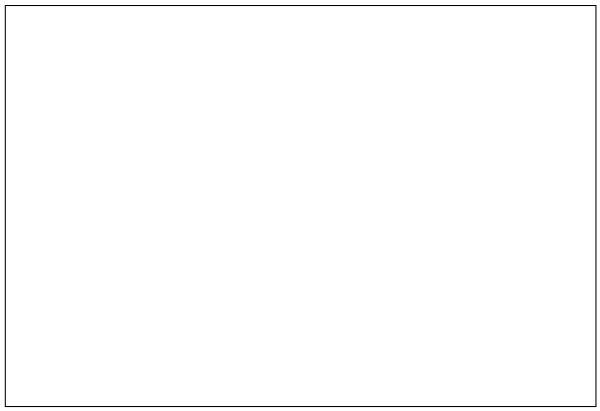

Figure 7. Average precision Vs. number of components and number of parts for DPM

4. Car detection and fine-grained classification

In order to detect and classify 45 million images, we need to use an efficient car detection and classification algorithm. Although RCNN [**?**] has been shown to be state of the art in object detection, it's memory and computation requirement make it impossible for use in a large scale detection problem such as ours. Specifically training with RCNN would require XXXXGB of memory and XXXX GPUs and car detection on 45 million images would take XXminutes per bounding box on XXX machine. We therefore used a simple DPM model with 0 components to detect cars and a standard CNN from [**?**] to perform classification on the detected bounding boxes. As shown in Fig. 7 there is only an XX% drop in average precision between using a dpm with 0 components and 0 parts and one with XXX components and YYY parts.

In order to classify the detected cars, we train a standard CNN from [**?**] using caffe [**?**]. Although our aim is to train fine-grain detections for street view images, many of our training images are obtained from other web sources due to the fact that annotating street view images is expensive. Thus we apply various deformations such as blurring in an attempt to make the web images more similar to street view images. During test time, we classify the top 10% scoring dpm bounding boxes using our CNN. This speeds up classification time by 10X while only resulting in a drop of .5 AP.

5. Societal analysis

After collecting data, training fine-grained car detectors and classifying cars in all of our images, we now have all the components to perform social analysis. We show some general results from the entire united states as well as case studies from cities we could obtain ground truth data for. We divide our analysis into different sections below.

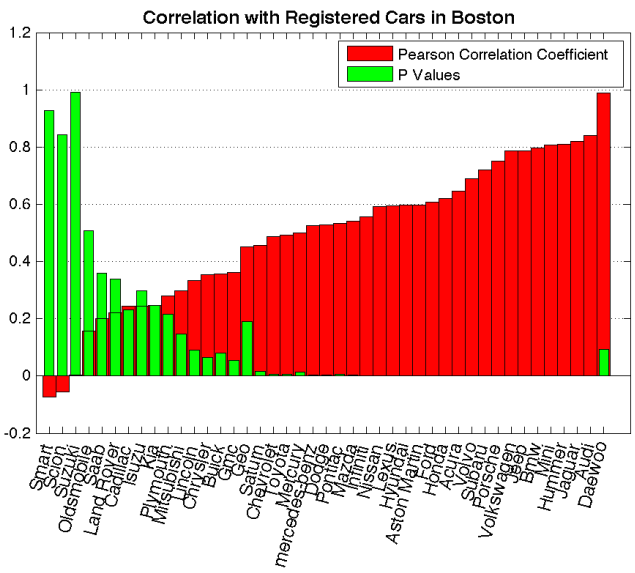

Figure 8. Pearson correlation coefficient and p values between the number of detected and registered cars in Boston for each make.

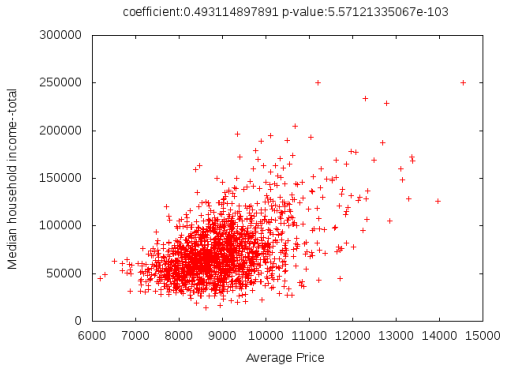

Figure 9. Average car price w.r.t. median household income.

Method	Frobnability
Theirs	Frumpy
Yours	Frobbly
Ours	Makes one's heart Frob

Table 1. Pearson correlation coefficient between various census variables and detected car attributes. All p values are ~ 0

5.1. What cars on the street tell us about people

5.1.1 Cars on the street versus cars people drive

The first question we asked is how do cars on the street relate to the cars that people drive? Specifically, can we learn about the registered cars in a zipcode from our street view detections? In order to answer this question we downloaded the vehicle census from Massachusettes, which is the only state to release extensive vehicle registration data. Surprisingly, we found an extremely high pearson correlation coefficient of 0.9 (p value=XXX) between the number of cars we detected per zipcode and the number of cars that are registered per zipcode in the three cities we analysed in Masachusettes: Boston, Worcester and Springfield. The high correlation was only obtained after aggregating cars at the zip code level which tells us that most people in these cities must drive within their zip code. After establishing a high correlation between the number of cars we detect and the number of registred cars, we also measured the correlation between the make of the detected and registered cars per zip code. As we can see in fig. 8 there is a high correlation for most of the makes. This shows us that the cars we detect from street view images can actually tell us a lot about the types of cars that people in a particular zipcode drive.

5.1.2 What do rich/poor people drive?

We gathered zip code level as well as census tract level 2007-2012 American Community Survey data for the 200 cities in our dataset and analyzed how the census data relates to statistics from our detected cars.

Table 1 shows correlation values between various attributes of the detected cars and median household income per zipcode. Fig. 9 shows a plot of median household income vs. average car price in a zipcode. As expected, the highest correlation is between median household income and the average car price per zipcode (r=0.59, p $\sim$ 0). This makes sense since rich people mostly live around places with expensive cars and tend to drive expensive cars. Our results also indicate that rich people prefer to drive foreign, especially German cars (r=0.59). This observation is also consistent with our expectations since expensive cars such as BMWs are German. What is perhaps surprising is that there is a very high negative correlation (r=-0.55, p $\sim$ 0) between the percentage of American cars in a zip code and median household income. So poor people live in places with many American cars.

Poor people also live near very old cars where as rich people live near newer ones. As table 1 shows, the correlation between median household income and the number of cars in 1990-1994 is very negative and increases to a high positive 0.59 for cars in the 2005-2009 range. Finally, a perhaps not surprising result is that poor people live near cars with low miles per gallon (MPG). This corroborates [**?**] study showing that poor people are more exposed to car pollution than rich people.

5.1.3 How does education relate to cars on the street?

One would probably guess that there is a high negative correlation between the number of people with only a high school education and the average price of a car in a zipcode. Indeed, we found that this is the case (r=0.3 p $\sim$ 0). As expected, we also found a high correlation between the number of college educated people in a zip code and the average car price. What is perhaps surprising is that although there is a large increase in correlation coefficient as we go from high school to college educated, the jump from college to graduate school is very low (r=0.31 for college educated and 0.39 for graduate school). This tells us that there is a very low difference in the price of cars driven by people who only hold bachealors as opposed to graduate degrees.

5.2. What cars on the street say about neighborhoods

5.2.1 Which neighborhoods are wealthy/poor?

We ask the question: what can our street view car detections tell us about the wealth of a neighborhood? Specifically, can we predict which neighborhoods are wealthy/unwealthy using our detections? Intuitively, if we see many expensive cars on the street, we suspect that we are in a rich neighborhood and vice versa. However, the correlation between car prices and neighborhood wealth is not going to be perfect because we are not necessarily detecting the cars that are registered by residents. Figure 11 A shows a heat map of the average price of detected cars within a zip code and median household income in a zip code for Boston and figure 11 B shows the same visualization for San Francisco. We can see that in both cities, the average car price in a zipcode is a very good predictor of wealthy/unwealthy neighborhoods.

5.2.2 Which neighborhoods have high car pollution?

Can our street view detections tell us anything about which neighborhoods are affected by highly polluting cars? To answer this question, we plotted a heatmap of the expected number of cars per sample inversely weighted by the expected MPG of that sample. Using this simple measure we should be able to have a rough idea of the location of highly polluting neighborhoods. For the same density of cars, areas with high MPG result in lower numbers than those with low MPG. For different densities of cars, the relative magnitude of the measure depends on both the density of cars and how efficient they are. Fig. 10A shows the density of cars in San Francisco and B shows the pollution heatmap that we created. Although we could not find ground truth data of car pollution, Fig. 10C is a map of San Francisco air quality measuring annual average particulate matter concentration (MPG) from all sources. Our maps seems to aggree with their data in most cases.

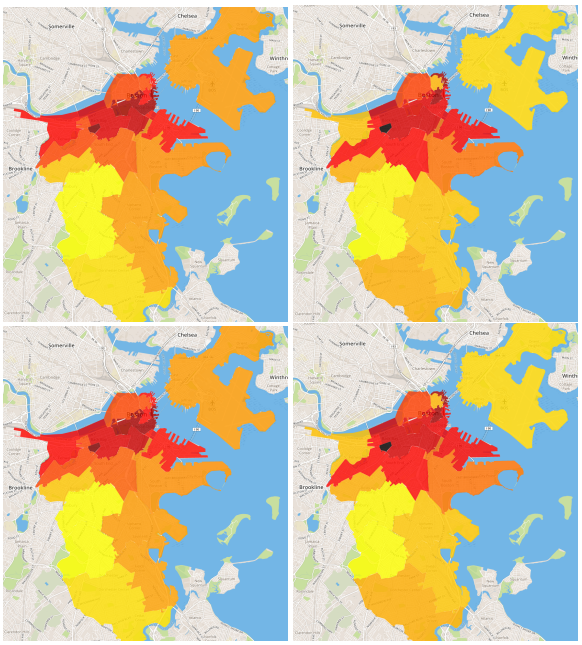

Figure 11. (A) Heatmap of average car price in Boston. (B) Heatmap of median household income in Boston

5.3. What cars on the street tell us about cities

5.3.1 Which cities are more segregated?

Following the analysis of [6] we use Getis Gi* statistic to produce statistically significant clusters of expensive and cheap cars and use Moran's I statistic [**?**] to measure the spatial segregation of car price with values ranging from -1 to 1. A value of -1 indicates perfect anti-correlation like a checkerboard where as a value of 1 indicates that similar values are perfectly clustered. The null hypothesis of complete spatial randomness produces values near 0. After creating a z-score for Moran's statistic we see that Boston, for example has a higher z-score than **XXX** showing that Boston has more segregated neighborhoods than **XXX**. After measuring the length of the spatial autocorrelation function of the average car price per zip code and calculating the correlation length, we find that *XXXX* is found to be the city with the highest correlation length, and therefore the city with the most segregated neighborhoods by income which is corroborated by census data [**?**]. Fig. 12 A shows statistically significant clusters of high and low car prices in Boston and Fig. **??** shows spatial correlograms for 5 cities displaying the decay of spatial autocorrelation as a function of distance.

6. Using social priors to improve classification

As shown in section 5 there is a very high correlation between some census variables such as income, and car

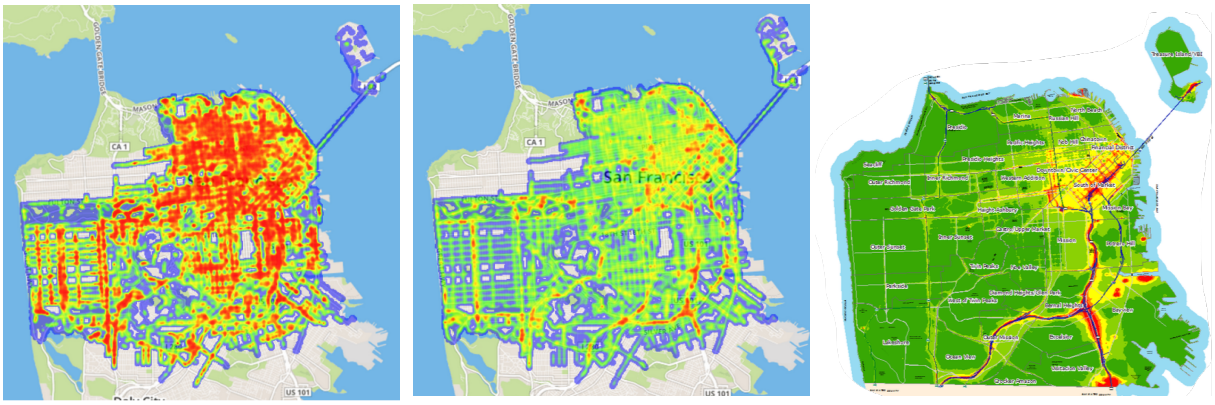

Figure 10. A. Density of cars in San francisco, B. Our measure of car pollution in San Francisco, C. Ground Truth for Air quality (measured in annual particulate matter) in San Francisco.

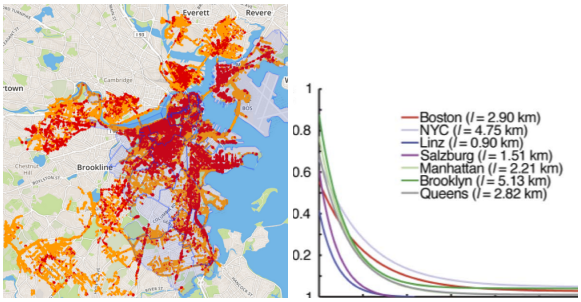

Figure 12. (A) Map of Boston showing statistically significant clusters of high -and low- car parices according to Getis Gi* statistic. Red shows clusters of high prices and orange shows clusters of low prices. (B) Spatial correlograms showing the decay of spatial autocorrelation as a function of distance.

attributes such as price and year. Given this relationship, we explore the use of census data to improve our fine-grain classification.

6.1. Analyzing classification accuracy

Since census data is most highly correlated with aggregate car attributes, one question is how much knowing ground truth car attributes would help in classification accuracy. This gives us an upper bound for the gain in accuracy that can be obtained by using census variables as a prior. Table 2 lists the classification accuracy after using ground truth car attributes. Surprisingly, knowing the manufacturing country of the car gives very little gain in accuracy ($\sim$ 0.5%). However, localizing the car price to within one of two bins of expensive vs cheap cars provides a gain in accuracy of 3%. Looking at the confusion matrices in fig. 13 we can see that after dividing the car price into 5 bins using quantiles, some expensive cars are confused with cheap cars where as most car countries of origin are not confused with other countries, except for one case of confusion between

Method	Frobnability
Theirs	Frumpy
Yours	Frobbly
Ours	Makes one's heart Frob

Table 2. Classification accuracy with ground truth attributes

South Korean and Japanese cars.

We take this experiment further and plot accuracy Vs. price bin in fig. 14 for various numbers of price bins, all generated using quantiles, localized to various degrees of accuracy. For example, we can see that if we localize the price of the car to one of 4 bins we would get an 8% increase in classification accuracy. However, even localizing the price to within 3 out of those 4 bins would result in a 1% increase.

6.2. Using census priors to predict attributes

As shown in [2] [**?**] using contextual priors can improve object classification accuracy. In order to directly use census information as a prior, we would like to find $P(C|I, Sk)$ where C is the fine-grain class,I is an image and $Sk \in \{S1 \ldots Sn\}$ is a particular zip code level census variable such as median household income. Using Bayes' rule:

$$P(C|I, Sk) = \frac{P(I, Sk|C)P(C)}{P(I, Sk)} \quad (1)$$

If we assume that the image and census data are conditionally independent given the fine-grained class label, the above equation can be written as

$$P(C|I, Sk) = \frac{P(I|C)P(Sk|C)P(C)}{P(I, Sk)} \quad (2)$$

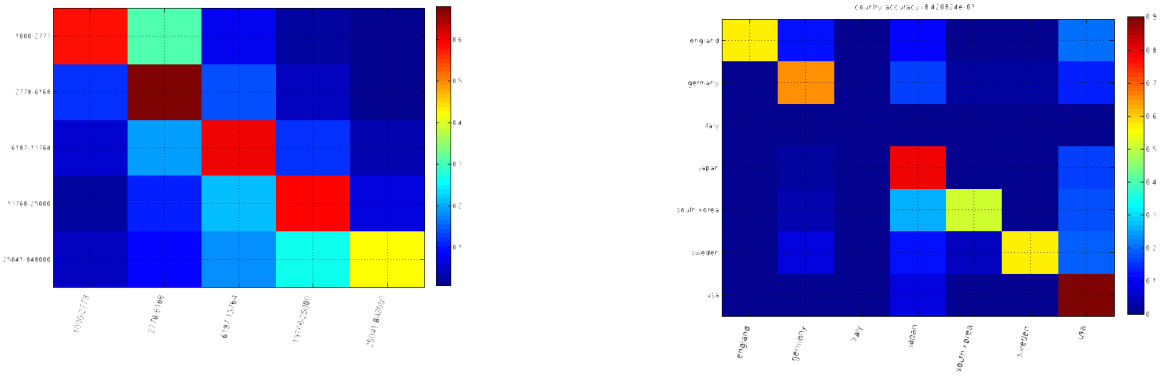

Figure 13. (A) Confusion matrix between different price bins. Cheap cars are mistaken for expensive cars (B) Between cars made in different countries

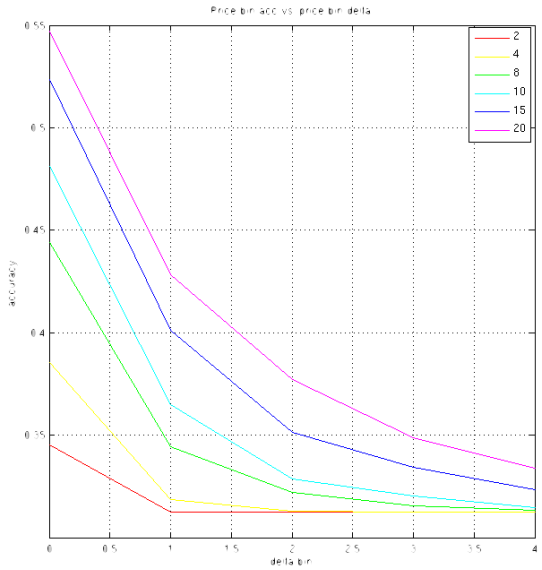

Figure 14. Fine-grain classification accuracy with ground truth price bin. X axis is the number of bins to which the price is localized and Y axis is classification accuracy. For example, localizing the price to 3 out of 8 bins results in a classification accuracy of 33% which is 2% higher than the baseline, 31.27%

After applying Bayes' rule again to $P(I|C)$ and $P(Sk|C)$, we get

$$P(C|I,Sk) = \frac{P(C|I)P(I)}{P(C)}\frac{P(C|Sk)P(Sk)}{P(C)}\frac{P(C)}{P(I,Sk)} \quad (3)$$

$$\propto \frac{P(C|I)}{P(C)}P(C|Sk) \quad (4)$$

We experimented with different methods of using census variables as priors. The first method was to quantize them into varying numbers of bins and calculate $P(C|Sk)$ in equation 4 for each census variable.As shown in table 3 this method results in a reduction in accuracy because we do not have enough geotagged training data to gain informative knowledge from the census about one of 2657 fine-grain classes. However, although there are 2657 classes, the number of attribute classes is much lower. If $Aj \in \{A1 \ldots An\}$ represents a car attribute such as price, we can reformulate $P(C|Sk)$ in equation 4 as $P(C|Aj)P(Aj|Sk)$. This formulation comes from a naive bays generative model assuming that Sk are the observed variables from which we can calculate $P(Aj|Sk)$. After this modification, equation 4 can be written as

$$P(C|I,Sk) \propto \frac{P(C|I)}{P(C)}P(C|Aj)P(Aj|Sk) \quad (5)$$

We calculate $P(C|I,Sk)$ for all car attributes and 30 different census variables, quantizing them into bins ranging from 2-20. Table 3 shows the 3 highest accuracy numbers for various combinations of census variables and car attributes. It can be seen that using median household income and either car price or year result in the highest accuracy gain (although this gain is very slight). This result is to be expected given the social analysis results of section 5 showing high correlation between median household income and car price and year.

6.3. Multiple car attributes and census variables

In section 6.2 we used single car attributes and census variables to calculate prior probabilities. We also experimented with using multiple census variables to predict single car attributes as well as combining priors independently calculated from multiple census variables and car attributes. To obtain $P(Aj|S1...Sn)$ we first quantize attributes Aj into M bins where M ranges from 2 to 20. If Aj has a fixed number of classes (such as car make), M is just the

Method	Frobnability
Theirs	Frumpy
Yours	Frobbly
Ours	Makes one's heart Frob

Table 3. Results.Ours is better.

number of attribute classes. We then perform multi-class logistic regression along with feature selection to classify Aj into one of M bins using $S1 \dots Sn$. $P(Aj|S1 \dots Sn)$ is then the probability obtained through logistic regression.

Finally, we combine the priors learned for different attributes by using the fact that car attributes are independent from each other given the class. Using this assumption along with Bayes rule $P(C|A1 \dots An)$ can be written as

$$\frac{P(A1 \dots An|C)P(C)}{P(A1 \dots An)} \tag{6}$$

After making use of the conditional independence of $A1 \dots An$ given C this becomes

$$\frac{P(C)}{P(A1 \dots An)} \prod_{j=1}^{n} P(Aj|C) \tag{7}$$

And after applying Bayes rule to $P(Aj|C)$ we get

$$\propto \frac{P(C)}{P(A1 \dots An)} \prod_{j=1}^{n} P(C|Aj) \tag{8}$$

We assume that car attributes are independent from each other given census variables. $P(C|I, S1 \dots Sn)$ is then

$$\propto \frac{P(C)}{P(A1 \dots An)} \prod_{j=1}^{n} P(C|Aj)P(Aj|S1...Sn) \tag{9}$$

Where $P(Aj|S1...Sn)$ is given by the logistic regression probabilities.

Table XXX shows the accuracies obtained after combining multiple census attributes and multiple car attributes. ***Maybe more descussion after experiments***

7. Conclusion

References

[1] S. Ardeshir, A. R. Zamir, A. Torroella, and M. Shah. Gis-assisted object detection and geospatial localization. In *Computer Vision–ECCV 2014*, pages 602–617. Springer, 2014. 1

[2] T. Berg, J. Liu, S. W. Lee, M. L. Alexander, D. W. Jacobs, and P. N. Belhumeur. Birdsnap: Large-scale fine-grained visual categorization of birds. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 2019–2026. IEEE, 2014. 1, 6

[3] K. Matzen and N. Snavely. Nyc3dcars: A dataset of 3d vehicles in geographic context. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 761–768. IEEE, 2013. 1

[4] 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, 2

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

[6] P. Salesses, K. Schechtner, and C. A. Hidalgo. The collaborative image of the city: mapping the inequality of urban perception. *PloS one*, 8(7):e68400, 2013. 1, 5

[7] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 2

[8] N. Zhang, J. Donahue, R. Girshick, and T. Darrell. Part-based r-cnns for fine-grained category detection. In *Computer Vision–ECCV 2014*, pages 834–849. Springer, 2014. 2