Beyond fine-grained classification: 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 infromation, 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 streetview 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 streetview images has jumpstarted a new line of computer vision research focused on understanding cities through images [6] [4] [5]. For example, Hedalgo et al show that crime predictions can be improved by incorporating human perceptions of neighborhoods' images rather than using census data such as income alone [6] and Tamara's guy et al. and MIT people 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 tamara 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?

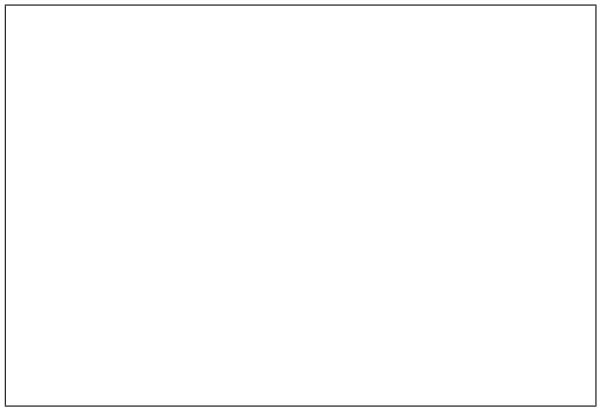


Figure 1. Example of caption. It is set in Roman so that mathematics (always set in Roman: $B \sin A = A \sin B$) may be included without an ugly clash.

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 likelyhood of detecting expensive cars. Similarly, knowing that we are in a farm area increases our liklihood 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. Al-

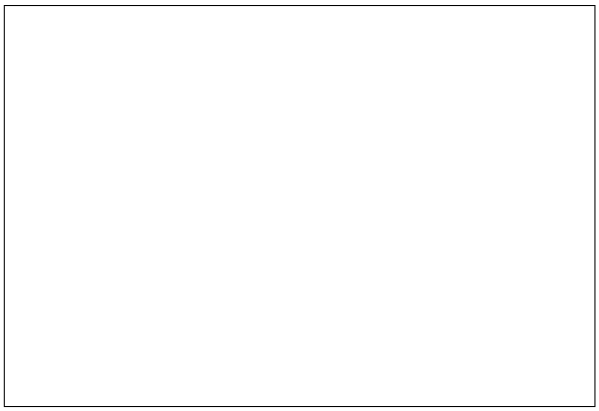


Figure 2. Example of caption. It is set in Roman so that mathematics (always set in Roman: $B \sin A = A \sin B$) may be included without an ugly clash.

though there has been previous work on learning spatio-temporal priors for fine-grain classification [2] and exploiting streetview 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 classificaiton 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 streetview 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 .

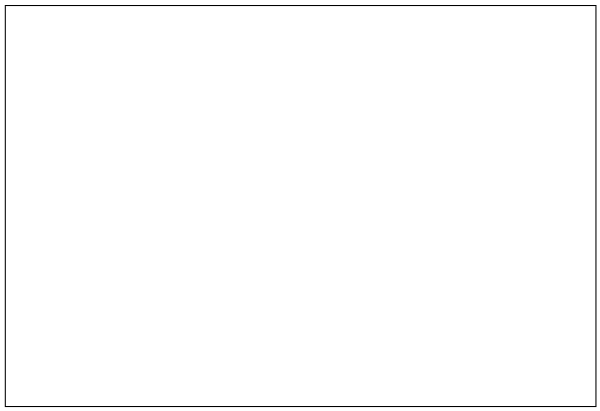


Figure 3. Example of caption. It is set in Roman so that mathematics (always set in Roman: $B \sin A = A \sin B$) may be included without an ugly clash.

5. NYC 3D uses geographic elevation data to create view-point aware detectors and extract ground planes for them

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 streetview 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 streetview, 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 classlist of 18000 cars comprising of all cars listed on edmunds.com and grouping them into visualy 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 streetview were labeled using 100 hired car experts.

Fig. 3 shows example images from our data while Fig. 4 shows some image statistics. Images from craigslist.com and cars.com have one large bounding box whereas google streetview images have multiple small boxes with cars that are blurred and occluded. As shown in Fig. 5 the different finegrained classes are very difficult to distinguish

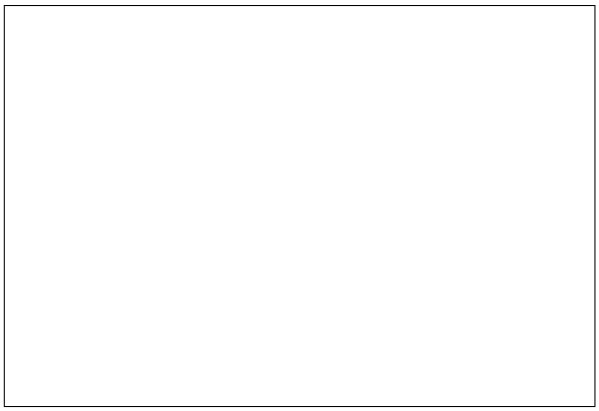


Figure 4. Example of caption. It is set in Roman so that mathematics (always set in Roman: $B \sin A = A \sin B$) may be included without an ugly clash.

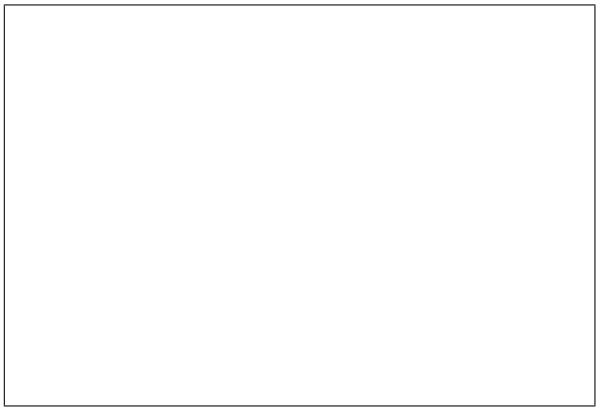


Figure 5. Example of caption. It is set in Roman so that mathematics (always set in Roman: $B \sin A = A \sin B$) may be included without an ugly clash.

3.2. Images with no labeled fine-grained classes

In order to perform societal analysis, we collected 45 million google streetview 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. 6 shows maps from two cities with the samples that were collected and fig. 8 shows the number of samples for the 10 biggest cities. For each sample, we collect images at 0,60,120,180,240,300 degrees.

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 Alex etal [**?**] to perform classification on the detected bounding boxes. As shown in Fig. **??** 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 streetview images, many of our training images are obtained from other web sources due to the fact that annotating streetview images is expensive. Thus we apply various deformations such as blurring and aspect ratio distortion in an attempt to deform the web images into streetview images. Table 1 shows the obtained

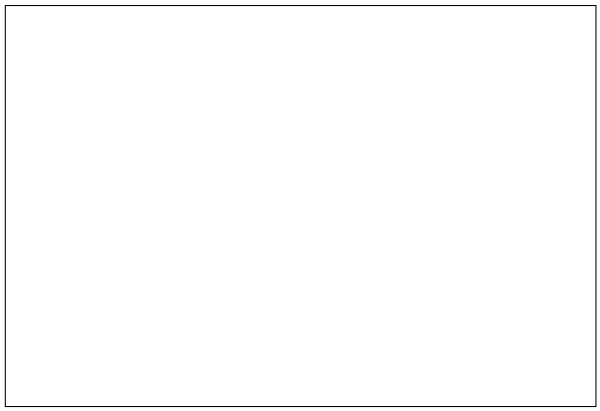


Figure 6. Example of caption. It is set in Roman so that mathematics (always set in Roman: $B \sin A = A \sin B$) may be included without an ugly clash.

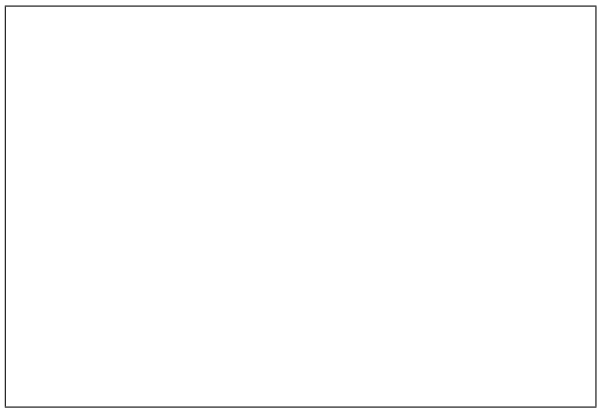


Figure 7. Example of caption. It is set in Roman so that mathematics (always set in Roman: $B \sin A = A \sin B$) may be included without an ugly clash.

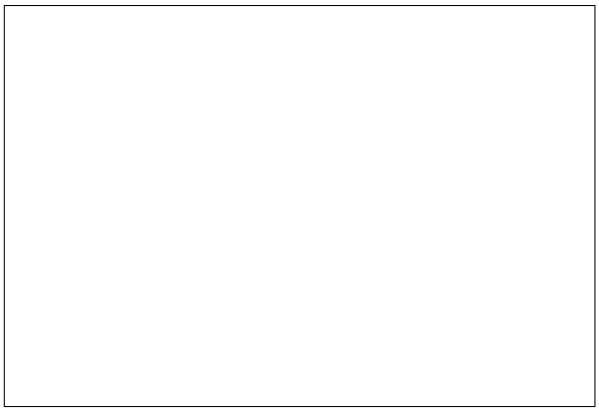


Figure 8. Example of caption. It is set in Roman so that mathematics (always set in Roman: $B \sin A = A \sin B$) may be included without an ugly clash.

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

Table 1. Results.Ours is better.

accruacy after adding each deformation. 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 are now ready to perform some social analysis. We show some general results from the entire united states as well as a case study from Massachussettes.

We gathered zipcode 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. As shown in fig. 9, median household income is highly correlated with the average price of cars in a given zip code (Pearson correlation 0.493, P-value of T-Test $<< 0.001$. The correlation coefficient varies city by city from 0.8 for city XXXX to 0.4 for city YYY.

5.1. What do poor people drive and what do rich people drive?

We found a strong positive correlation (Pearson Correlation 0.592, P-value of T-Test $<< 0.001$ between the percentage of cars made outside of the United States and the median household income per zip code. As shown in Figure 10 (A), rich people seem to drive German cars (Pearson correlation 0.57) where as there is a surprising strong negative correlation of -0.59 between the number of American cars per zip code and median household income even though the number of American cars in our dataset is higher than the number of German cars.

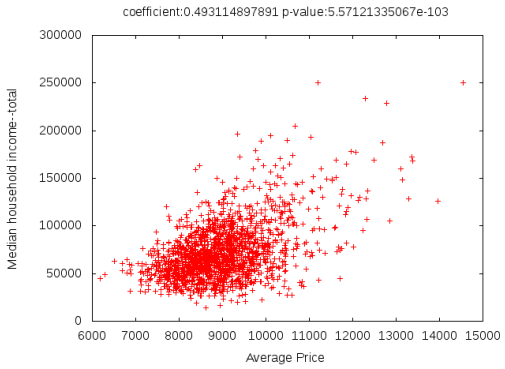


Figure 9. Average car price w.r.t. median household income.

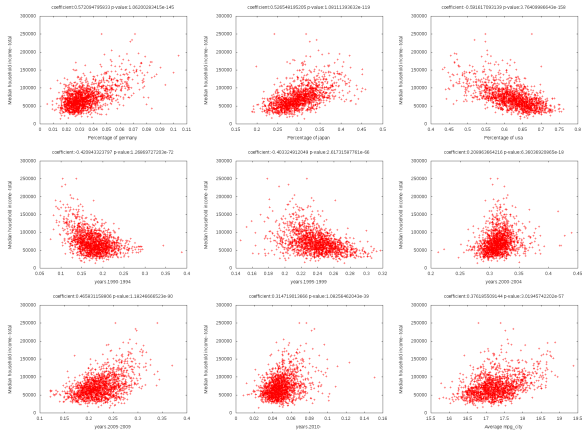


Figure 10. (A) Countries of cars vs. median household income, (B) years of cars vs. median household income, (C) years of cars vs. median household income, last column is avg. city MPG vs. median household income

Not surprisingly, rich people also dirve newer cars. Figure 10 B shows the correlations between the number of cars in different year bins vs. median household income. Starting with a strong negative correlation of -0.4 between the number of cars from 1990-1994 in a given zipcode and median household income, the correlation keeps on increasing for newer cars reaching a small positive value for cars in 2000-2004 and a maximum value of 0.46 for cars in 2000-2013.

Finally,as shown in 10 C poor people live in zipcodes with cars that have low mileage while rich people live in zipcodes having cars with high MPG. As shown in [?] poor people are exposed to higher levels of pollution because they live near areas with higher traffic levels and cars

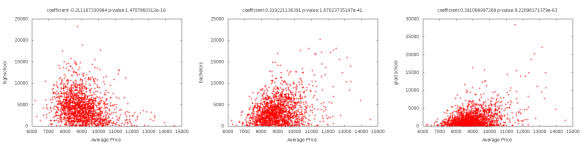


Figure 11. Number of people with different education level vs. average car price per zipcode

with high C02 emissions. Finally Figure 11, shows the relationship between education and the average car price of detected cars in a zipcode. There is a moderate negative correlation of -0.211 between highschool educated people and the price of the car in their zipcode as opposed to a positive correlation of 0.319 between the number of graduate school educated people in a zipcode vs. the average car price. As expected, the jump in correlation coefficient from high school educated to college educated (-0.21 to 0.32) is higher than the jump from college educated to graduate school educated (0.32-0.39). Although we have found many interesting results, we will cover them in the supplemental material for the sake of brevety.

5.2. Which neighborhoods are wealthy/poor?

We found that the results of our fine-grained car detections can be used to learn various attributes of neighborhoods. Intuitively, if we see many expensive cars on the street, we suspect that we are in a rich neighborhood and viceversa. However, the correlation between car prices and neighborhood wealth is not going to be perfect because we are not necessarily detecting the cars that residents drive. Figure 12 shows a heatmap of the average price of detected cars within a zipcode and median household income in a zipcode. We can see that just by looking at the average price of cars within a zipcode we can correctly predict most of the high and low income zipcodes. We show these viusalizations for all 200 cities in our supplemental material and only present two cities here.

5.3. 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 seggregation 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 seggreaged neighborhoods than **XXX**. After measuring the length of the spatial autocorrelation function of the average car price per zipcode and

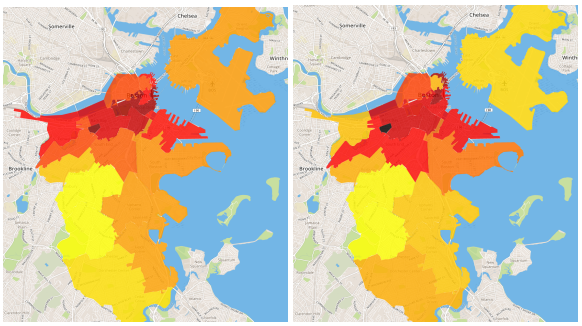


Figure 12. (A) Heatmap of average car price in Boston. (B) Heatmap of median household income in Boston

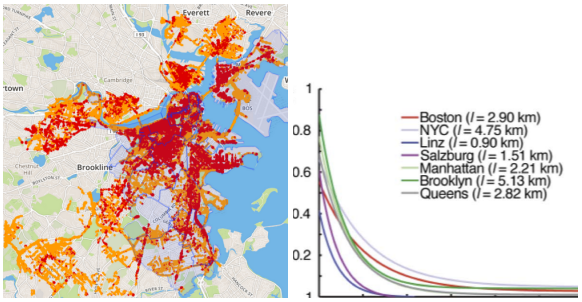


Figure 13. (A) Map of Boston showing statistically significant clusters of high -and low- car parices according to Getis Gi* statistic. Red shows clusters of high prices and orange shows clusters of low prices. (B) Spatial correlograms showing the decay of spatial autocorrelation as a function of distance.

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 seggregated neighborhoods by income which is corroborated by census data [**?**]. Fig. 13A 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 attributes such as price and year. Given this relationship, we explore the use of census data to improve our fine-grain classification.

6.1. Analysing classification accuracy

Although we are classifying 2657 classes,it is important to know whether our classification errors are between highly similar or dissimilar classes. Fig. 14 A shows one of a few possible hierarchies of car classes. As shown in Fig. 14 B, most of the errors are lower in the hierarchy and are at the level of trims and years.

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

Table 2. Results.Ours is better.

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. Table 2 lists the classification accuracy after using ground truth car attributes. Surprisingly, knowing the manufactured country of the car gives very little gain in accuracy (0.5%). However, localizing the car price within one of two bins of expensive vs cheap cars provides a gain in accruacy of 3%. Looking at the confusion matricies in fig. **??** we can see that after dividing the 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 South Korean and Japanese cars.

We take this experiment further and plot accuracy Vs. price bin in fig. 15 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 $P(C|I,S)$ where I is image and S is census information. Using Bayes' rule:

$$\Pr(C|I,S) = \Pr(I,S|C)\Pr(C)/\Pr(I,S) \quad (1)$$

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

$$\Pr(C|I,S) = \Pr(I|C)\Pr(S|C)\Pr(C)/\Pr(I,S) \quad (2)$$

After applying Baye's rule again to $P(I|C)$ and $P(S|C)$, we get

$$\Pr(C|I,S) = \frac{\Pr(S|I)\Pr(I)}{\Pr(C)}\frac{\Pr(C|S)\Pr(S)}{\Pr(C)}\frac{\Pr(C)}{\Pr(I,S)} \quad (3)$$

$$\propto \frac{\Pr(C|I)}{\Pr(C)}\Pr(C|S) \quad (4)$$

We experimented with different methods of using census variables as priors. The first method was to quantize

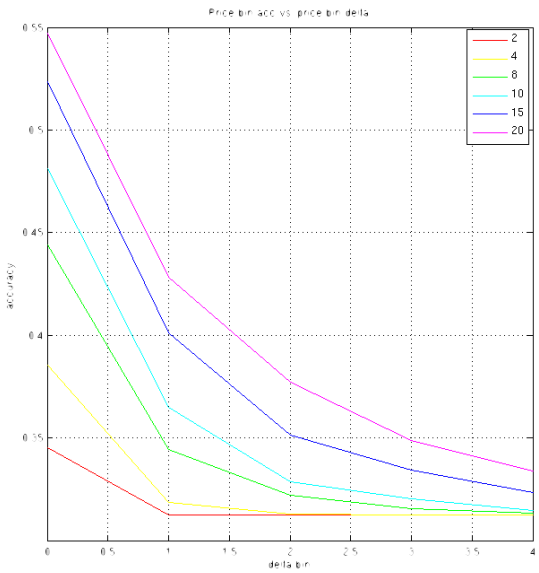
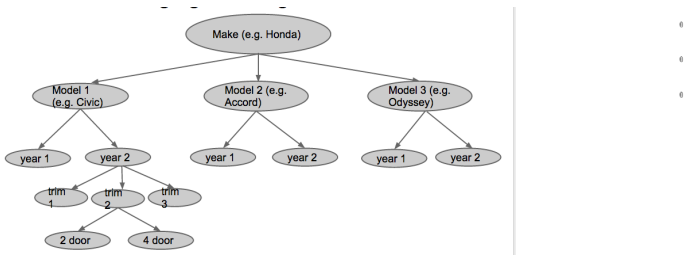


Figure 15. Fine-grain classification accuracy with ground truth price bin. X axis is the number of bins to which the price is localized and Y axis is classification accuracy. For example, localizing the price to 3 out of 8 bins results in a classification accuracy of 33% which is 2% higher than the baseline, 31.27%

census variables into variying numbers of bins and calculate $P(C|Sn)$ in equation **??** 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. Thus we reformulate $P(C|s)$ in equation **??** as $P(C|An)P(An|Sn)$ where An is a particular car attribute such as car price. This formulation comes from the following naive bays generative model assuming that $P(Sn)$ are the observed variables from which we can calculate $P(An|Sn)$. After this modification, equation **??** can be written as

$$\Pr(C|I,Sn) \propto \frac{\Pr(C|I)}{\Pr(C)}\Pr(C|An)\Pr(An|Sn) \quad (5)$$

We calculate $P(C|I,Sn)$ 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 of car 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 **??** showing high correlation between median household income and car price and car year.



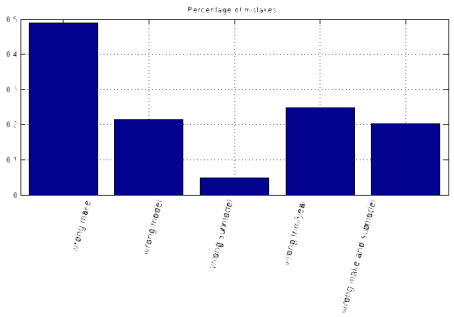


Figure 14. (A) Hierarchy of car classes. Images of cars become increasingly visually similar while travleing down the tree (B) Percentage of Error at each level of the hierarchy: for example trim/year error shows the percentage of errors that only occur at the trim year level (and classify make, model and submodel accurately)

6.3. Multiple car attributes and census variables

In the previous section we used single car attributes and census variables to calculate our prior probabilities. We also experimented with using multiple census variables to predict single car attributes as well as combining priors independently calculated from multiple cesnsus variables and car attributes. To obtain $P(An|S1...Sn)$ we first quantize continuous attributes An into M bins where M ranges from 2 to 20. If An is a descrete attribute (such as car make), M is just the number of attribute classes. We then perform logistic regression along with feature selection to classify An into one of M bins using $S1 \dots Sn$. $P(An|S1 \dots Sn)$ is then the probability obtained through logistic regression.

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

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

After the conditional independence assumption it becomes

$$\frac{\Pr(C)}{\Pr(A1 \dots An)} \prod_{i=1}^{n} \Pr(Ai|C) \tag{7}$$

And after applying Baye's rule to $P(Ai|C)$ we get

$$\propto \frac{\Pr(C)}{\Pr(A1 \dots An)} \prod_{i=1}^{n} \Pr(C|Ai) \tag{8}$$

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

$$\propto \frac{\Pr(C)}{\Pr(A1 \dots An)} \prod_{i=1}^{n} \Pr(C|Ai) \prod_{i=1}^{n} \Pr(Ai|S1...Sn) \tag{9}$$

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

Table 3. Results.Ours is better.

Where $P(Ai|S1...Sn)$ 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. 2

[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. 2, 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. 2

[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