Carnegie Mellon University
University of Pennsylvania

Automatic Recognition and Understanding of the Driving Environment for Driver Feedback

Jifu Zhou, Luis E. Navarro-Serment, and Martial Hebert

The Robotics Institute, Carnegie Mellon University

DISCLAIMER

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the U.S. Department of Transportation's University Transportation Centers Program, in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof.

Motivation

A smart driving system must consider two key elements to be able to generate recommendations and make driving decisions that are effective and accurate: The environment of the car and the behavior of the driver. Our long-term goal is to develop techniques for building internal models of the vehicle's static environment (objects, features, terrain) and of the vehicle's dynamic environment (people and vehicle moving in the vehicle's environment) from sensor data, which can operate online and can be used to provide the information necessary to make recommendations, to generate alarms, or to take emergency action. Our overall approach is to combine recent progress in machine perception with the rapid advent of onboard sensors, and the availability of external data sources, such as maps.

A comprehensive approach

Understanding the environment of a vehicle can be envisioned at different levels of details from low-level signals that characterize the location of potential hazards to high-level descriptions that include semantic information such as recognizing specific types of objects, e.g., traffic signs, or specific patterns of motion, e.g., erratic pedestrian motion. Current safety systems already include systems in which sensors are able to produce a coarse map of obstacles in regions around the vehicle. These capabilities are limited to fairly coarse descriptions of the environment. A notable exception is in the area of people and car detection (e.g., the MobileEye[1] system), in which a commercial product is already available. However, even in this case, interpretation of the sensor data is limited to the location and motion of the object but does not include higher-level predictive information about pattern of motion and future actions.

We believe that now, given the availability of sensors that provide rich data, there is an opportunity to develop techniques that can generate far more complete and higher-level

[1] www.us.mobileye.com

descriptions of the vehicle's environment than was ever possible before. Given input (images and 3D) from these sensors, the first component of our approach relies on recent development in the general area of scene understanding. Specifically, our approach is to extend state-of-the-art machine perception techniques in three areas: 1) Scene understanding from images in which objects, regions, and features are identified based on image input; 2) Scene understanding from the type of 3D point clouds acquired from, for example, stereo of LIDAR systems; and 3) analysis of moving objects which includes the ability to predict likely future motions in addition to modeling the current trajectory.

These machine perception elements have shown promising results, but they have major shortcomings that prevent their immediate applications to driving applications. Chief among them is the fact that their accuracy and computational efficiency are not at the level needed for use in driving applications. We believe that one major reason is that these tools have been developed in the context of unconstrained problems. However, this application is severely constrained by rules of the road, architectural and civil engineering constraints, and even by the knowledge of normal patterns of behaviors in the case of pedestrians.

Accordingly, the second key component of our approach is to extend the machine perception techniques to incorporate a complete ensemble of constraints from this application and environments. The technical challenge is to combine data of a statistical and "continuous" nature such as sensor signals and low-level features with knowledge of a symbolic and discrete nature. We expect that the systematic use of context and domain constraints will boost the performance of machine perception approaches and bring it closer to the desired level of performance.

Another key development in the automotive industry is the availability of massive amounts of data from a variety of external sources. For starters, some level of map data is continuously available to the system. Even the relatively coarse level of detail used in current systems (Garmin, Tomtom, etc.) would be able to focus the work of machine perception modules by

providing regions of interest, predicting occurrence of classes of objects (e.g., expected density of buildings), and prior information about likely flow patterns of pedestrians. Direct and continuous access to Web sources is also on the horizon as a common feature of cars, which opens access to other sources of information, such as weather or traffic conditions. V2V communication and infrastructure communication would provide additional sources.

From the point of view of machine perception, the availability of external sources of data is a unique opportunity to further constrain the perception process to increase performance. Accordingly the third component of our approach is to develop techniques to maximize the use of external sources of information. We start by using current map data from navigation systems to generate priors on distribution of features and objects in the environments, and to generate priors of pedestrian and tracking activity. The key technical challenge in this area is to design formal methods to meld the priors distribution predicted from the maps with the current machine perception algorithms.

During the course of the work reported here, we have made progress towards the implementation of this approach. In particular, we have focused on the problem of improving the performance of algorithms for the semantic labeling of images through the use of information from street maps (Figure 1). This is an important step toward a realistic design of a perception system that is truly integrated with all the sources of information.

Introduction

With the rapid growth in computer vision research, algorithms for scene parsing, object recognition or pedestrian detection have become increasingly applicable in commercial products. However, due to adverse conditions such as poor illumination, bad weather, etc., the performance of these algorithms decreases when operating in the real world. In these cases, external knowledge about the scene--such as a street map--can provide a context that contributes to disambiguate pure visual perception and leverage the performance. Fortunately,

intelligent transportation systems with access to external information such as maps, weather or traffic conditions can provide infrastructure support for various vision tasks.

Algorithms for scene parsing, or semantic labeling, assign a class label (e.g., tree, sky, building, etc.) to each pixel of an image based on a set of visual and contextual features. In this project, we explore the use of external information such as a map to improve the accuracy of scene parsing algorithms. In particular, we propose a system that retrieves map information from an external database, systematically handles the uncertainty of external information, and incorporates such information into semantic labeling algorithms.

This report is organized as follows: the following section reviews the related work. A subsequent section describes our system's pipeline. The next section presents the evaluation of our system performance from experiments using images from real world scenes. Finally, the last section summarizes the results and discusses the lessons learned.

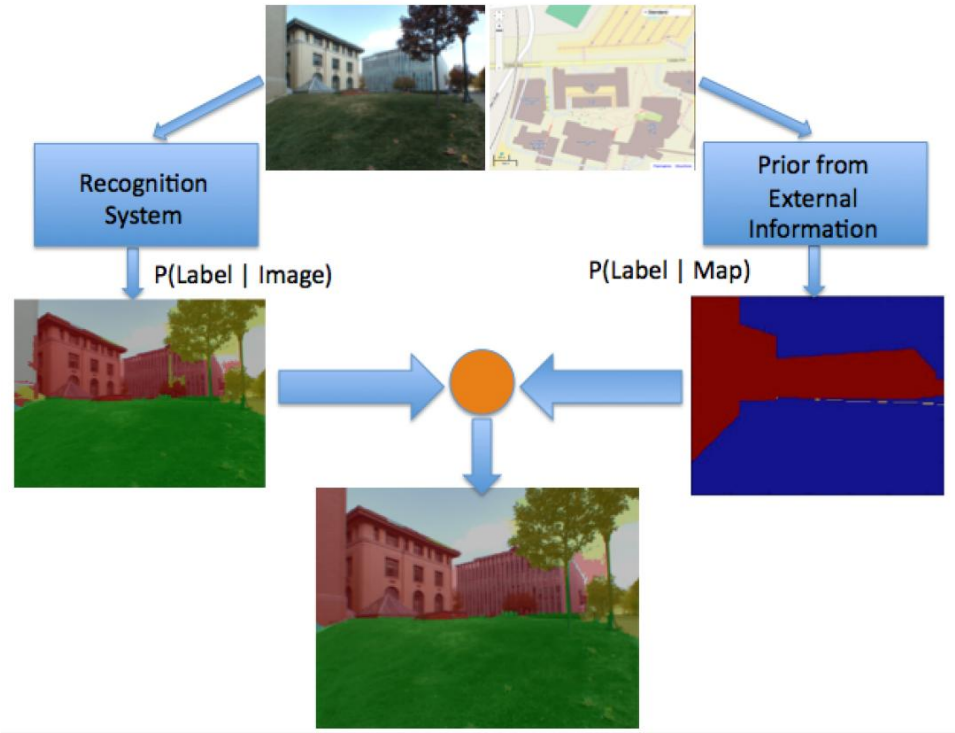

Figure 1. Incorporating information from a street map to improve scene understanding.

Related Work

The use of different types of external knowledge has been explored in prior work on scene parsing, object detection, etc. Information in terms of context proves to be effective in priming object detection [8], [27]. In semantic segmentation problems, information in the form of motion/geometric features [25], [14], temporal consistency [10], [3], [22], or scene type with jointly inference [29] also proved beneficial.

A series of work in label transfer [20], [30], [26], [7] uses an external image database as transferable knowledge, and demonstrate its effectiveness for scene parsing tasks. In our project, we explored using map information in the form of geo-contextual prior for scene semantic labeling.

Related work using external knowledge such as geo-information in vision mainly addresses landmark recognition or visual localization problems. Researchers in [6] proposed a system that recognizes city-scale landmark from mobile phone images where geo-information such as GPS is used to constrain landmark image retrieval from a database. In [2], GPS is used to access geo-contextual information such as the location of a point of interest, traffic and road information; such geo-contextual information then provides local context to select meaningful object hypotheses and prime recognition of landmarks. Different from their work, we do not attempt to recognize urban landmarks. Instead, we are interested in semantic labeling of outdoor scenes.

A closely related work in the context of street scenes is [1], in which the author demonstrates an application of road detection using an external map database. In particular, the author uses GPS to retrieve road maps from an external database and combines them with visual cues to detect roads, showing the effectiveness of using external information to cope with difficult vision problems in real world scenarios. Similarly, we also use external information for the interpretation of scenes. Yet different from their work, we propose a general framework that can encode different types of prior information from external sources, e.g., not only roads but can also include high level knowledge and structures of the scenes. This allows us to answer

more general scene understanding questions. One direct application is to provide the prior probability distribution for scene parsing.

Further application can include improving object detection, pedestrian prediction, traffic forecast, etc. Another closely related work in the context of outdoor scene understanding using scene structure is from Geiger et al. [13], in which they jointly infer object location and 3D scene layout from a monocular camera. Their previous work [11] estimates 3D scene layout from dynamic traffic objects. In both cases, they are able to estimate scene layout using information such as traffic objects. Different from their work, we use external information to aid scene understanding problems such as scene parsing or object detection, where the scene layout can be directly modeled from an external database.

System Description

There are three major components in the system. The first component retrieves external information from an existing map database, and handles the uncertainty of camera pose given the GPS trace of the camera in a Kalman Filter manner. The second component encodes the external information in the form of semantic label probability distribution using a maximum entropy framework. The last one incorporates this prior information into scene parsing algorithm to improve recognition performance.

Retrieving Map

Vision systems for intelligent transportation have the advantage that the image sequences taken by on-board cameras can be GPS/time stamped. For each image, one can estimate the camera pose and use GPS index to directly access a map of the surrounding area from an external database, which can help understand the scene in the image. Thus, the first component of our pipeline retrieves map information from an existing Geographical Information Systems (GIS) database.

We use OpenStreetMap[2], an open sourced world-scale map database maintained by an active community. Unlike other GIS database such as Google Map/Latitude, from which the user can mostly access high level static information such as images of the map, OpenStreetMap database provides lower level and editable map information. For example, users can access 2D bird-view geometry of the map elements such as outlines of buildings and roads, as well as semantic tags of the map elements such as scene types or road conditions.

Fig.2 illustrates the processing pipeline for integrating the map prior. First, a GPS stamp with the format of a bounding box (top-left-latitude, bottom-right-latitude, top-left-longitude, bottom-right-longitude) is used to index OpenStreetMap via its API, as illustrated in Fig.2.A; the returned data includes the geometry of the map in a 2D bird view in the format of points, lines, and polygons, each with its own Lat/Lon stamps. Then the conversion from Lat/Lon to Universal Transverse Mercator (UTM) coordinate system is performed to obtain the flat 2D map of bird view geometry, as shown in Fig. 2.B. For the purpose of generalization, our system has an interface to allow users to configure customized camera parameters. With the 2D map geometry and camera parameters, the system can then compute the perspective projection of the 2D map onto the image plane, as shown in Fig. 2.C. Finally, since OpenStreetMap objects and polygons include standard tags such as road type, building, scene type, our system can parse the projected map with semantic masks, as illustrated in Fig. 2.E. As a reference of the actual scene, Fig. 2.D is an example image of the same location obtained from Google Street View.

[2] www.**openstreetmap**.org

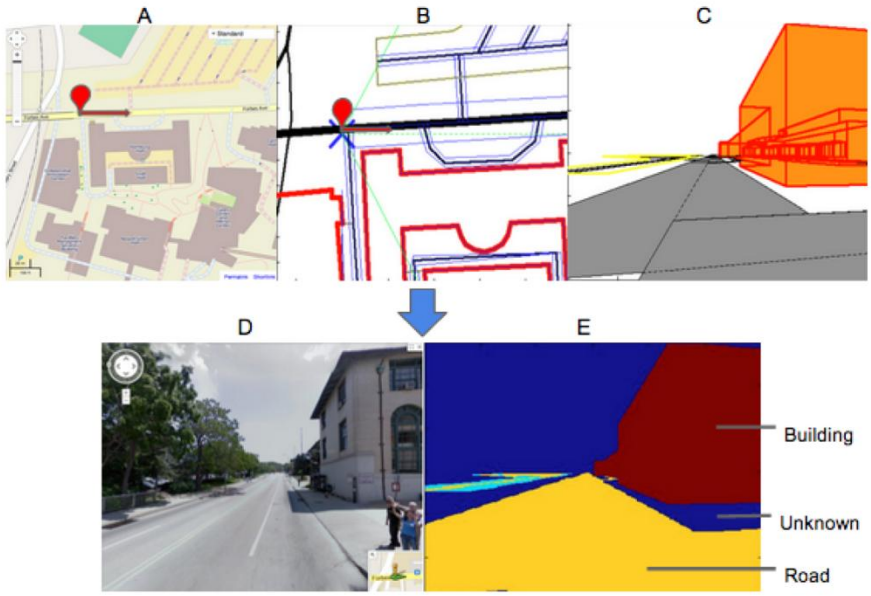

Figure 2. OpenStreetMap indexed by latitude/longitude (A); parsed map geometry in 2D, top view (B); perspective projection of map onto image plane (with estimated building height) (C); scene reference of the example (image obtained from Google Street View) (D); projection image masked with semantic labels from map (E).

Note that although external maps can provide accurate geometry and some underlying semantic knowledge of the surrounding environments, it only models limited information of the scene, e.g., mostly static manmade structures, while actual scene images may contain much richer information such as natural elements (trees, grass) or dynamic objects (cars, pedestrians). However, we believe that accurate underlying contexts of the static scenes is the first step, and can be helpful to overall scene understanding.

Depending on the accuracy of the GPS receiver, the recorded GPS trace can have inherent localization uncertainty, and it is likely that the projected map priors misalign with the actual objects in the image in the case of inaccurate localization. To minimize this effect, our system handles the localization uncertainty using a Kalman filter. In particular, we first apply the filter on a sequence of GPS stamps to estimate the uncertainty of each camera pose (location and heading direction); then we sample and aggregate multiple instances of the camera poses to model the belief of the location. With multiple samples projected in the image plane, we can obtain a "blurred" map prior that encodes the confidence. Fig. 3 shows such an example of the projected map prior with uncertainty, as well as the actual scene image overlaid with the prior.

Figure 3. Map prior with position uncertainty (top); original image overlaid with prior (bottom). Buildings are indicated in red, and the roads in green. A higher color saturation indicates higher certainty, while lower saturation indicates higher uncertainty.

Encoding External information as Prior

Our goal is to infer the most likely probability distribution of semantic labels for each segment in the images. For a generic scene parsing algorithm that infers semantic labels using image features, our system provides additional cues and contexts using information from external sources, e.g., maps. In order to aid the semantic label inference, the system needs to encode various sources of external information into some prior probability distribution of semantic labels. For example, given that the external information indicates that an image segment is a street in a business district, the system needs to model the prior probability of the segment being the semantic label of road vs. vehicle.

The maximum entropy framework is widely used in machine learning tasks for modeling posterior distribution discriminatively, with applications in natural language processing, inverse reinforcement learning, optimal control, etc.

In computer vision it has been successfully applied as a discriminative model for classification and inference in object recognition of multi-class from multiple types of features [23], [31], [21], [18], and in modeling prior probability [17]. Intuitively, the maximum entropy principle states that one exploits all that is known, and assumes nothing that is unknown. In other words, given some training data, the model learned using the framework must be consistent with all empirical facts and constraints, but should be as uniform as possible otherwise. The property is particularly useful when modeling prior knowledge, where one should claim least commitment.

In our case, we are modeling information from external sources as prior knowledge for semantic labeling. The maximum entropy model is used to obtain a prior distribution of semantic labels in the images. Mathematically, we denote y_s the random variable representing the label c of the pixel s in images, $f_E \in R^n$ the n features computed from external information E, e.g., map evidence. For example, one such feature of pixel s could be the map evidence indicating the label being road. Finally, $p(y_s|f_E)$ is the max entropy model we want to learn, which outputs the confidence of label y_s given map evidence.

The maximum entropy distribution $\mathbf{p(y_s = c|f_E)}$ that satisfies the empirical constrains has the exponential form

$$\mathbf{p(y_s = c|f_E) = \frac{\exp \lambda_c f_E}{\sum_{c'} \exp \lambda_{c'} f_E}}$$

where $\lambda_\mathbf{c}$ is the vector of scalar weights for external evidence features $\mathbf{f_E}$, and the vector of scalar weights $\Lambda = \{\lambda_{i,c}\}$ contains the parameters we want to learn.

The learning of Λ is done by maximizing the entropy

$$H(p) = \sum_{y_s, f_E} \tilde{p}(y_s)\, p(y_s|f_E) \log p(y_s|f_E)$$

where $\tilde{p}(y_s)$ is the empirical distribution computed from the training data.

The Improved Iterative Scaling (IIS) technique [4] was initially explored to learn the parameters Λ, but we found that using optimization techniques such as Broyden-Fletcher-Goldfarb-Shanno

(BFGS) method [5] performed better in terms of convergence and efficiency. Fig.4 shows the comparison of BFGS approach and initial IIS approach in modeling the prior distribution. The Euclidean distance between the prior distribution and the ground-truth distribution is used to measure how close the two approaches model the true label distribution.

From Fig.4 one can see that the BFGS approach models the distribution closer to the true label distribution.

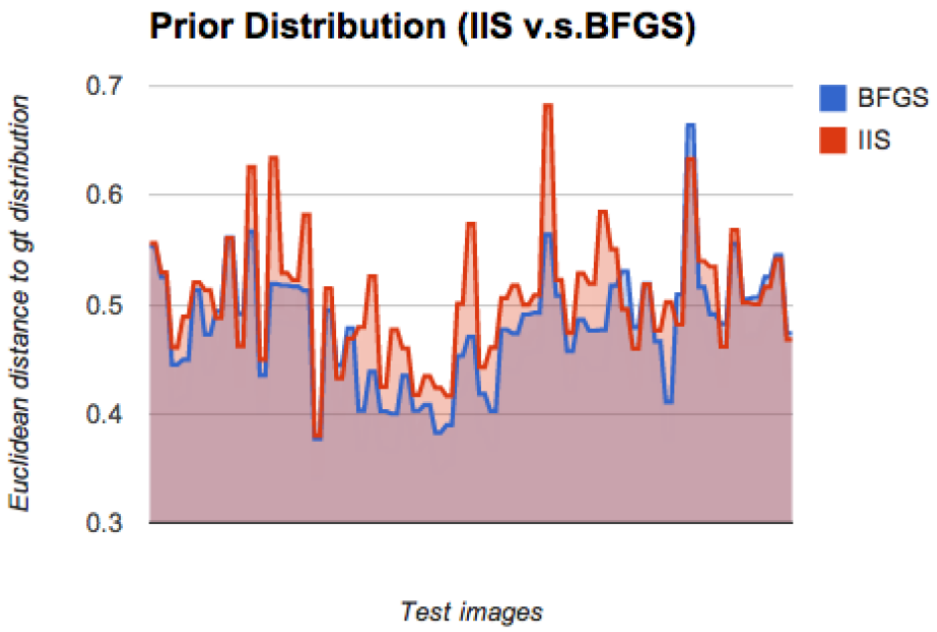

Figure 4. The *y*-axis measures the Euclidean distance of labels' prior distribution to the ground-truth distribution. The *x*-axis represents the test image's number. The red line indicates the prior learned using the IIS approach, while the blue line indicates the prior learned using the BFGS approach. Note that BFGS approach models the prior distribution more accurately, in terms of Euclidean distance.

Incorporating Prior in Scene Parsing

We explored several approaches to incorporating the prior distribution into the scene parsing algorithm as discussed below.

1.- Simple log-linear discriminative model:

Our first approach is to use a simple log-linear discriminative model to directly combine two probability distributions. Originally proposed in [16] as Product-of-Expert, it was also used in [24], [15] to combine the probability distributions of different components. In particular, it is of the form

$$P(y_s = c|f_I, f_E) = \frac{1}{Z} P(y_s = c|f_I) P(y_s = c|f_E)^{\gamma}$$

Or equivalently in log-linear form:

$$\log P(y_s = c|f_I, f_E) = \\ \log P(y_s = c|f_I) + \gamma \log P(y_s = c|f_E) - \log Z$$

where γ is the relative confidence, $P(y_s = c|f_I)$ is the output of a generic scene parsing algorithm with image feature f_I, $P(y_s = c|f_E)$ is the output of the prior system given external context features f_E computed from map information, and Z is the normalization term.

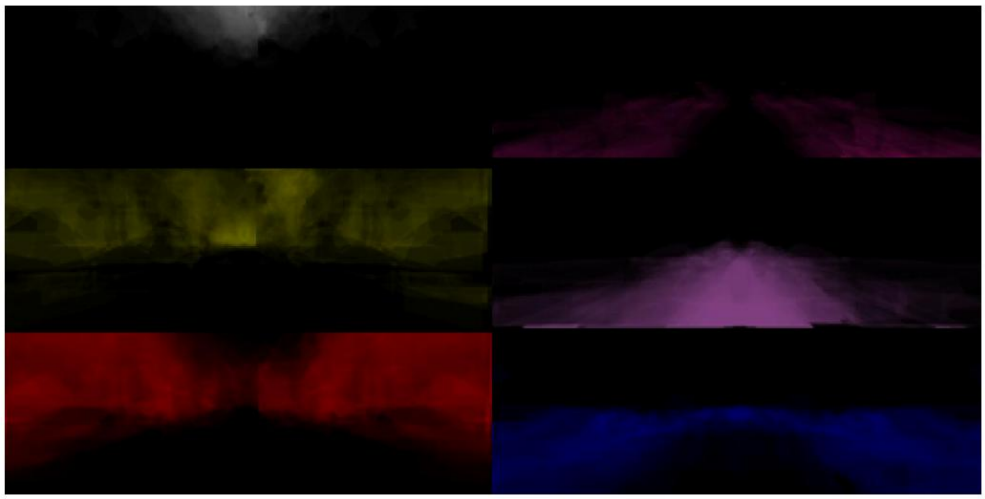

Figure 5. Sample label priors learned from data, mostly urban street scenes. Left, from top to bottom: sky, tree, building; right, from top to bottom: sidewalk, drive road, vehicles.

This is a discriminative framework since it models conditionals $P(y_s = c|f_E, f_I)$, and that $\log P(y_s = c|f_I)$ and $\log P(y_s = c|f_E)$ can also be treated as potential terms in Markov Random Field framework, feature terms in traditional log-linear classifier, or weak predictors in an

ensemble classifiers setting. One advantage is that $P(y_s = c|f_I)$ and $P(y_s = c|f_E)$, also discriminative components, can be learned independently. Thus, our map prior system which outputs $P(y_s = c|f_E)$ is compatible with an arbitrary vision system that is trained separately and outputs the discriminative label likelihood$P(y_s = c|f_I)$.

We tested this approach with a small data set. Fig. 6 shows an example of the scene parsing results before and after the map prior is combined. The approach is simple and fast, yet one drawback is that the relative confidence is set offline and is fixed when operating online; another drawback is that the direct combination is too simple in that it only considers $P(y_s = c|f_I)$ and $P(y_s = c|f_E)$ when inferring class $y_s = c$, while belief of other classes $P(y_s = c'|f_I)$ and $P(y_s = c'|f_E)$ where $c \neq c'$ may also contribute to the inference as context. As shown in Fig. 7, with this simple model, by increasing the relative confidence γ on the prior, the accuracy increases and then decreases. It is because when the weight on map prior is small, building regions are recovered and the overall accuracy increases; however as the weight on the map prior increases, foreground objects such as trees tend to be incorrectly labeled as buildings, resulting in a decrease in accuracy. The underlying cause is that the map prior only models manmade structures such as roads and buildings, but does not model natural/foreground objects such as trees or vehicles. Thus, directly combining the map prior with the semantic label probability is not ideal. Also, the effect on the performance is minimal (within 1%), due to the limited presence of manmade structures in the images. Thus, we further explored other approaches to incorporate the map priors, as described subsequently.

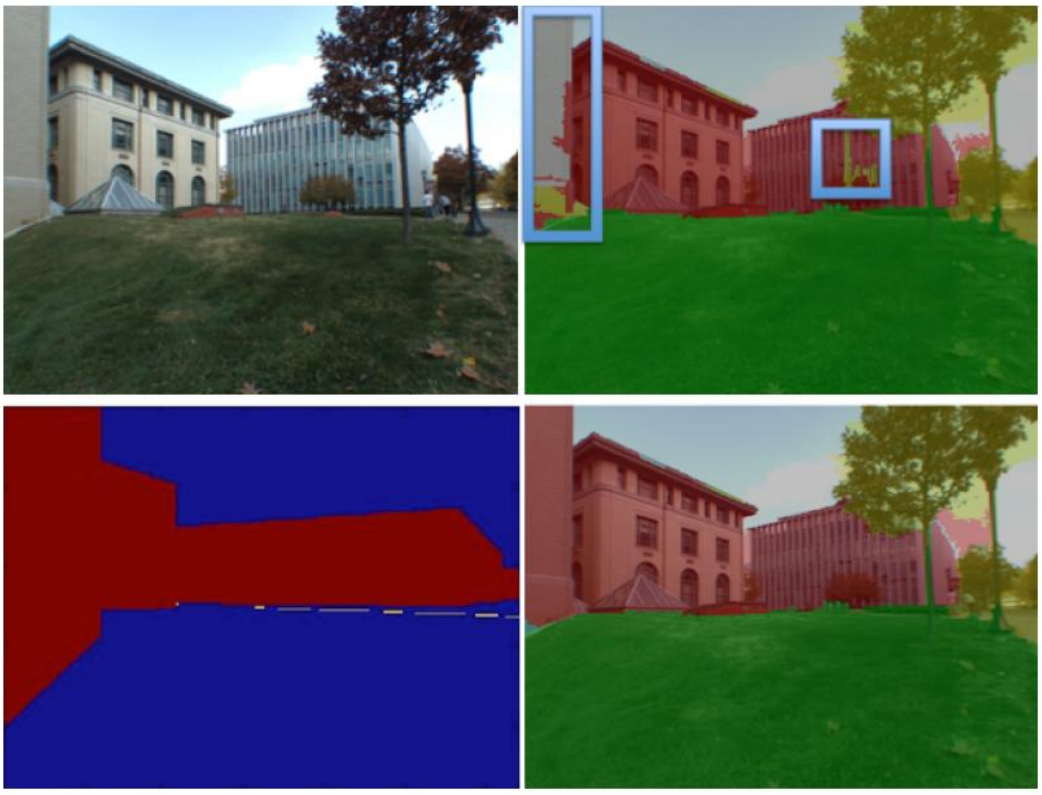

Figure 6. Example results of log linear model for combining probability; top-left: original image; top-right: original scene parsing results with per-pixel accuracy 89.80%; bottom-left: retrieved map prior (blue-unknown, red-building, yellow-road); bottom-right: scene parsing results combined with map prior using log linear model with accuracy 94.35%.

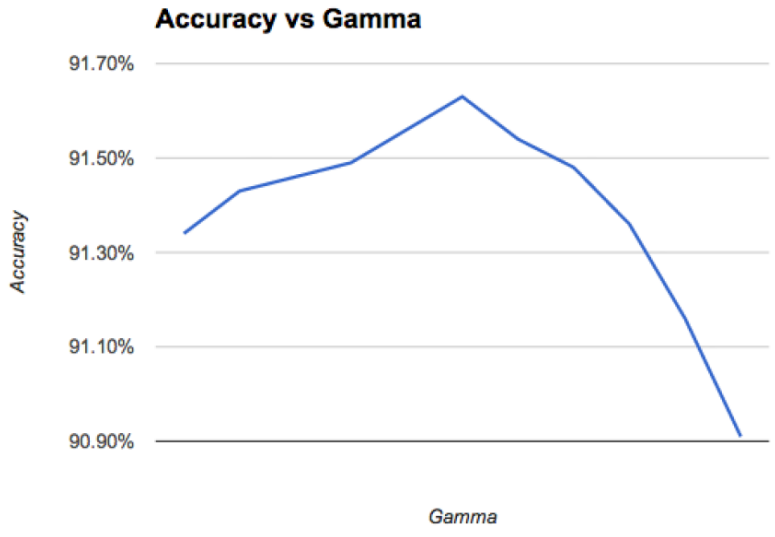

Figure 7. Example results of log linear model for combining probability; note the accuracy improves and then decreases by increasing γ the relative confidence from 0 to 1. Also note that the influence of the prior probability is within 1%, due to the amount prior information that can be applied (mostly manmade structures such as buildings) and their presence in the images.

2.- Re-weighting label distribution with prior context:

Instead of directly combining probability distribution, we also attempted to post-process the output of a generic scene parsing algorithm with prior context. In particular, we implemented a simple procedure that re-weights the label probability distribution using prior contexts. Formally, taking the label likelihood, the output of a generic scene parsing algorithm, and other contexts as input, we train a classifier to re-weight the label probability distribution.

Algorithm 1 below outlines the procedure. $P_{I,r}$ is the label distribution of segment r in image I computed from a generic scene parsing algorithm; $P_{I,N(r)}$ is the label distribution of r's neighboring regions $N(r)$ weighted by sizes, $P_{E,r}$ is a vector of other context features including the map prior probability computed from external sources, relative location and size of the segment r in the image. $Y_{I,r}$ is the ground-truth label distribution. We tried logistic regression, and non-linear classifiers (ensembles of trees) for the re-weighting classifier q used to update the semantic label probability distribution given prior and contextual features. Results are presented in the Experiments section.

Algorithm 1: Re-weighting with Prior Context

Data:	$Y_{I,r}, P_{I,r}, P_{I,N(r)}, P_{E,r}$
Result:	Re-weighted label probability distribution $\tilde{P}_{I,E,r}$
Training:	$X = \{P_{I,r}, P_{I,N(r)}, P_{E,r}\}$ $Construct\ data\ set$
	$D = \{Y_{I,r}, X_{I,r}\}$
	classifier q $\leftarrow$ train(D)
Testing:	$X = \{P_{I,r}, P_{I,N(r)}, P_{E,r}\}$ $Construct\ data\ set$
	$\tilde{P}_{I,E,r} \leftarrow$ q.apply(X)

3.- Appending prior as context feature in Scene Parsing algorithm:

We also attempted to incorporate the prior distribution by appending the prior probability onto the feature stacks of Hierarchical Inference Machine [23], which in our experiments is the main scene parsing algorithm.

The inference machine performs structure prediction on the semantic labels for each region in the image using visual and contextual features at multiple scales, and it achieves state-of-the-art performance on many datasets. In our experiment, the priors are used as additional context features at each scale.

We ran experiments in which the prior context features are used in training and testing, and compared it with the results that do not use the prior features. The KITTI dataset is used for evaluation. The results and discussions are presented subsequently.

Experiments

A. Dataset used

The dataset we used in the experiment is from several image sequences of the KITTI benchmark raw data [12], taken from cameras mounted on top of a vehicle driving in urban and suburban area. Each image in the sequence is associated with a GPS stamp, allowing our system to retrieve corresponding map geo-information from a GIS database. We sparsely sampled the image sequence once every 50 frames to obtain training and testing image sets.

B. Metrics used

For evaluation, we used the semantic segmentation accuracy (SA) from the PASCAL Challenge [9], ”intersection over union”:

$$SA = \frac{true\ positive}{true\ positive + false\ positive + false\ negative}$$

The standard precision and recall metrics are also used. Since we have models trained with prior information and models trained without prior information, we also evaluated the pixels whose labels changed before and after prior is applied.

C. Evaluation on appending prior as context feature in scene parsing algorithm

As described in the previous section, we attempted to incorporate features by appending prior probability computed from external information onto the context feature stack of the Inference Machine [23]. Two sets of experiments were conducted: one with a small dataset of urban street scenes, and the other with a large dataset of natural/suburban scenes. In both experiments, we split the images into a training set and a testing set. We train a model that uses the prior map information and one that does not use the prior map information on the training set, and we apply both models on testing set to produce semantic label inference.

Figure 8. Sample urban street scene.

Small Set - Urban Street Scene: We first experimented with a small set of images that are mostly canonical urban street scenes with various illumination conditions as shown in Fig.8. Quantitatively, Fig.10 shows a comparison of the inference accuracy between the model trained with prior information and the one trained without prior information.

The accuracies of most classes improve with prior information. Fig.9 shows examples of street scene semantic labeling using both the model without prior information (top of each pair) and the model with prior information (bottom of each pair). Note that regions such as building or sky are misclassified originally, but are corrected when prior information is applied. The prior to some extent imposes the spatial consistency of the label distribution in the street scene image, e.g. sky is above horizon, cars cannot be same height as building. This helps disambiguate the original inference, especially under adverse lighting conditions where textures of different objects are visually ambiguous.

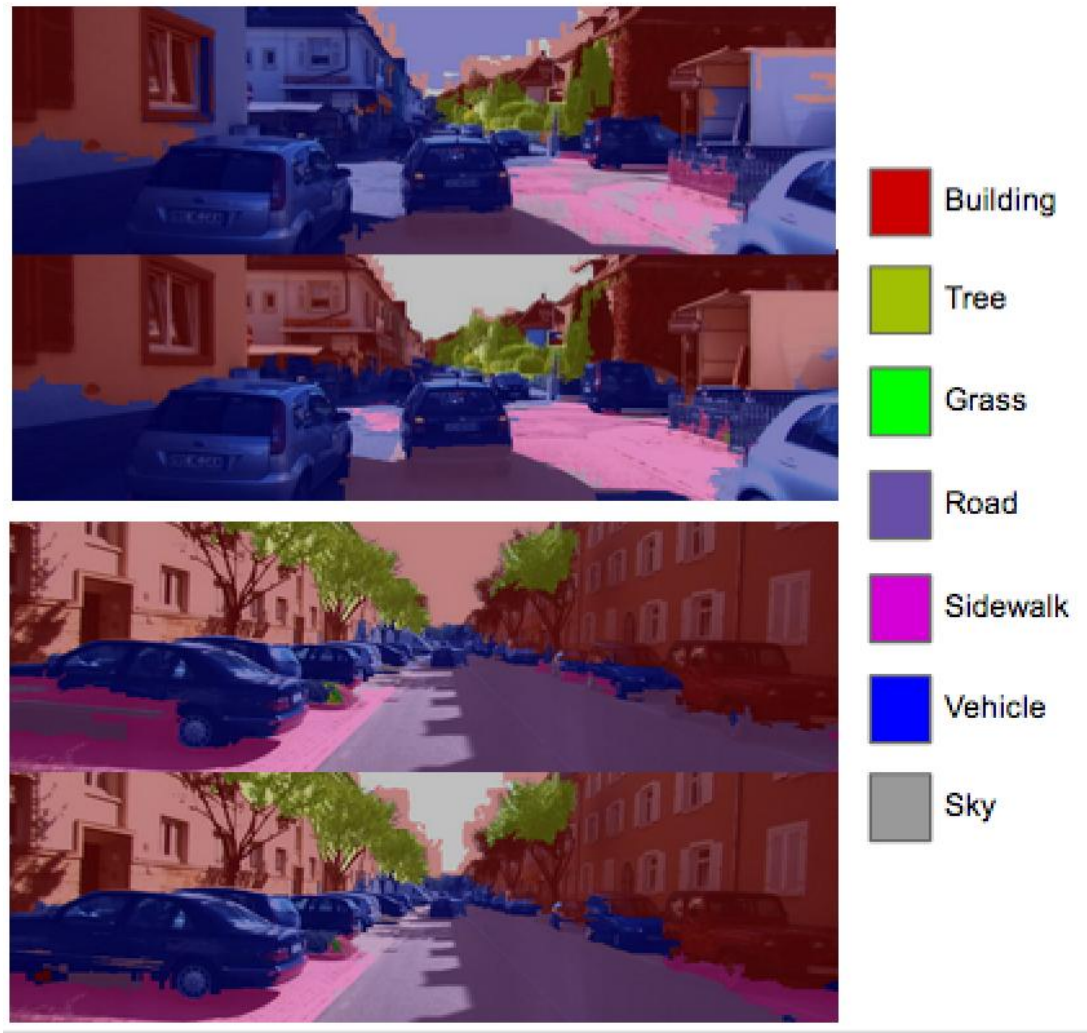

Figure 9. Example results for experiment in urban street scenes; for each pair: top - infer without priors, bottom - infer with priors; note the corrected segments such as building and sky when prior is applied.

We also evaluated only the segments whose labels changed when prior information is used. Specifically, we counted the number of pixels in segments whose labels changed, and divided these changed pixels into three categories: corrected from incorrect class, misclassified from correct class, and remain incorrect in a different class.

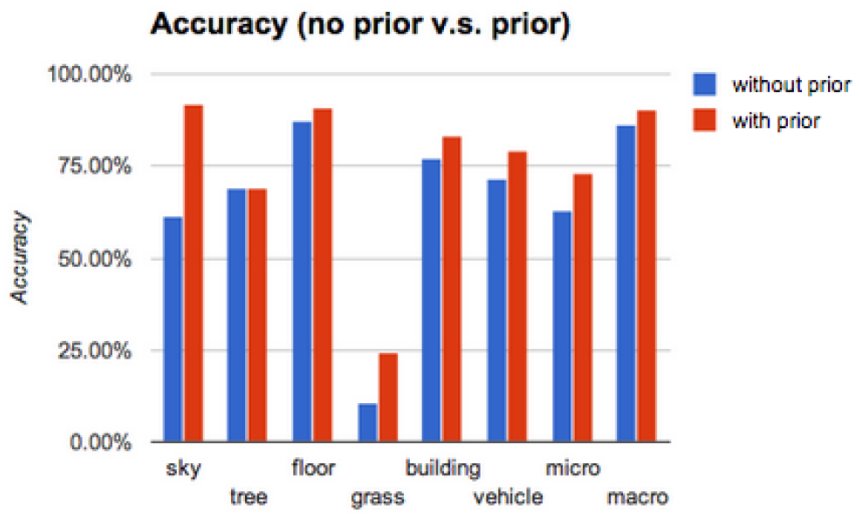

Figure 10. Label Accuracy for urban scene experiment; blue-without prior, red-with prior.

As shown in Fig.11, segments of sky and building are mostly corrected with the prior information, and segments of tree or foreground objects such as vehicles, which are not modeled by the map or not representative in the dataset, are not as much corrected.

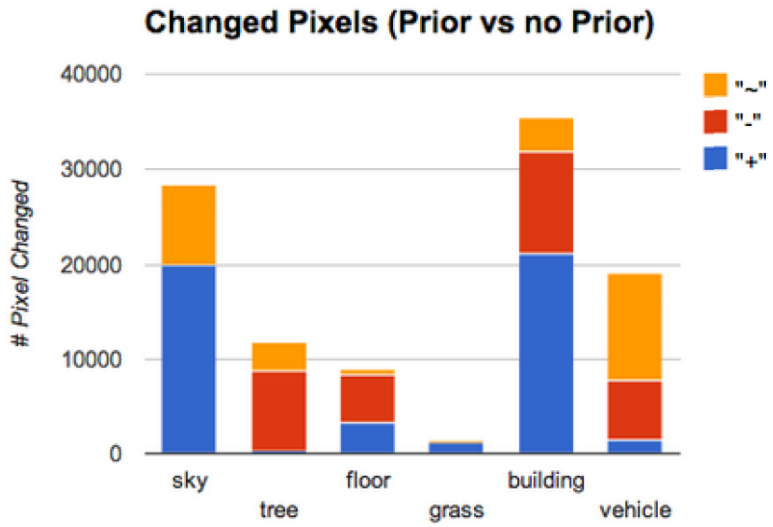

Figure 11. Urban Scene Experiment - Number of pixels where labels changed after prior is applied; blue (”+”) refers to pixels corrected from incorrect labels; red (”-”) refers to pixels changed from correct labels to incorrect ones; yellow (”~”) refers to pixels changed from incorrect labels to another incorrect ones.

Large Set - Suburban/Residential Street Scene: We further evaluated the approach using a larger set of 200 sparsely sampled images. This set contains mostly suburban/residential scenes with unstructured and natural foreground objects such as trees occluding the underlying man-made structures, as shown in Fig.12. A sample prior learned from this data set is shown in Fig.13. Note that compare to the prior learned from the urban scenes in Fig.5, trees are much more visible in the suburban scenes, while buildings are more visible in urban street scenes.

Figure 12. Residential/Suburban Street Scene; note most underlying manmade structures such as buildings are fully occluded by foreground unstructured natural objects.

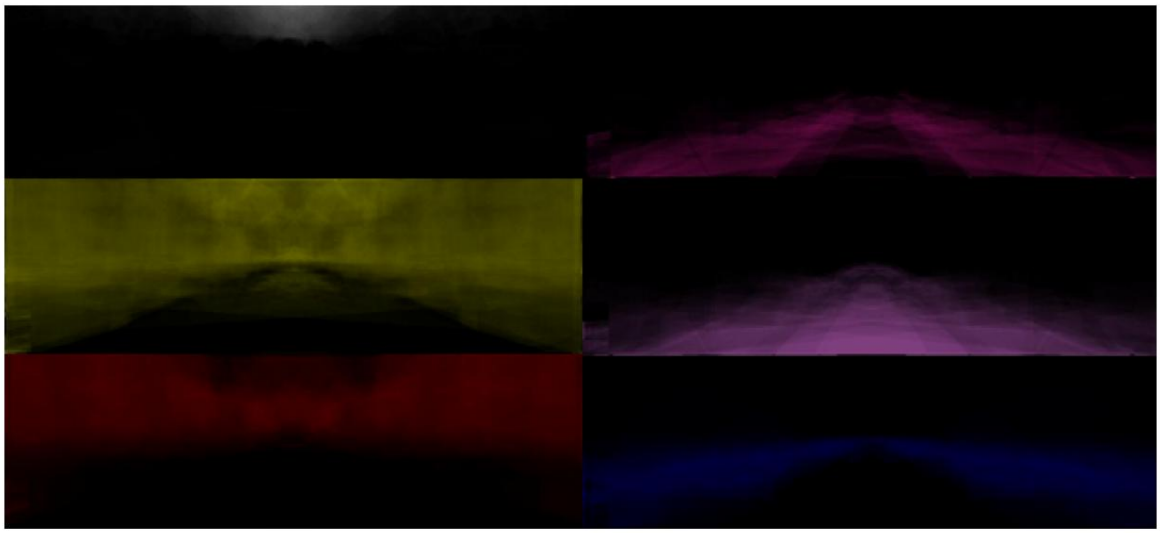

Figure 13. Sample label priors learned from suburban/residential scenes. Left, from top to bottom: sky, tree, building; Right, from top to bottom: sidewalk, drive road, vehicles. Note trees are much more salient compared to priors of urban scenes in Fig.5.

Table I and Fig.14 show the quantitative results with and without prior information. First of all, we notice that spatially consistent regions such as sky and floor are mostly improved with prior information. Vertical parts of the scenes usually consist of trees, buildings, and vehicles. We found in these segments that the model with prior information has more tendencies to label trees instead of buildings and vehicles, compared to the model without prior information, as illustrated in Fig.15, resulting in decrease in accuracy for these classes. Quantitatively, the bias to tree is shown with the increasing recall and decreasing precision of tree labeling in Table I. Also shown in the difference of Confusion Matrices (CM) Table II (computed by CM prior −CM no−prior), increasing number of building/vehicle segments are labeled as trees with prior information.

Table 1: Metrics for Priors vs. No Priors

Without Prior	Sky	Tree	Floor	Building	Vehicle
Accuracy	13.13%	55.67%	64.64%	33.68%	33.47%
Precision	52.97%	64.97%	74.45%	55.07%	66.81%
Recall	14.87%	79.54%	83.08%	46.45%	40.15%
With Prior	Sky	Tree	Floor	Building	Vehicle
Accuracy	16.69%	54.08%	65.11%	27.56%	27.58%
Precision	41.15%	61.64%	73.24%	56.05%	61.69%
Recall	21.93%	81.51%	85.43%	35.16%	33.28%

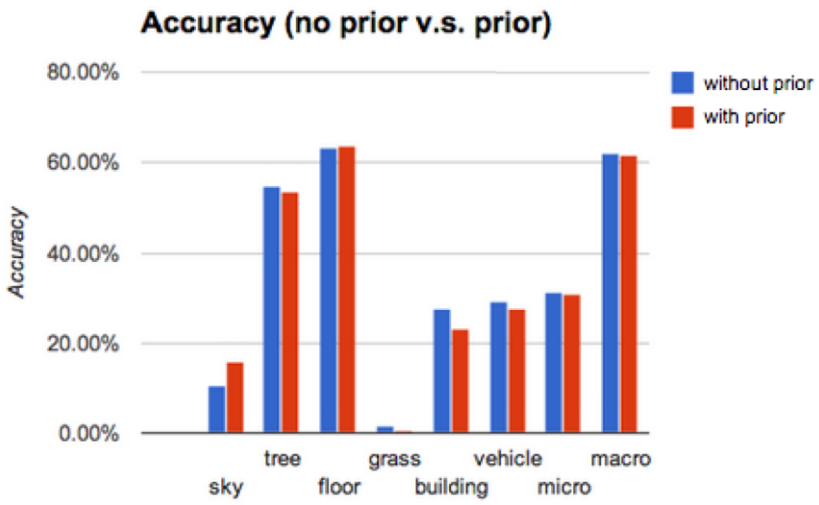

Figure 14. Label Accuracy for residential scene experiment; blue-without prior, red-with prior

This reveals the limitation of the extent of external information one can use for scene parsing tasks. The current map information mostly models underlying manmade structures such as roads and buildings, etc., but does not model natural elements or foreground objects such as trees or vehicles, etc. However, in the task of scene parsing, an inference system makes decisions for each segment in the image considering all label classes. Therefore, a prior framework that does not model certain label classes should comply with the principle of least commitment for these label classes, and model these label classes as uniform as possible based

on empirical distribution. The maximum entropy framework that we use to model the prior is one which optimally finds such label distributions. In this case, the empirical probability distribution of training data will heavily influence the resulting prior models, especially for those label classes not modeled in maps. We further investigate the specific image set used for training, and found that the ground truth label distribution is biased towards trees as shown in Fig.17.

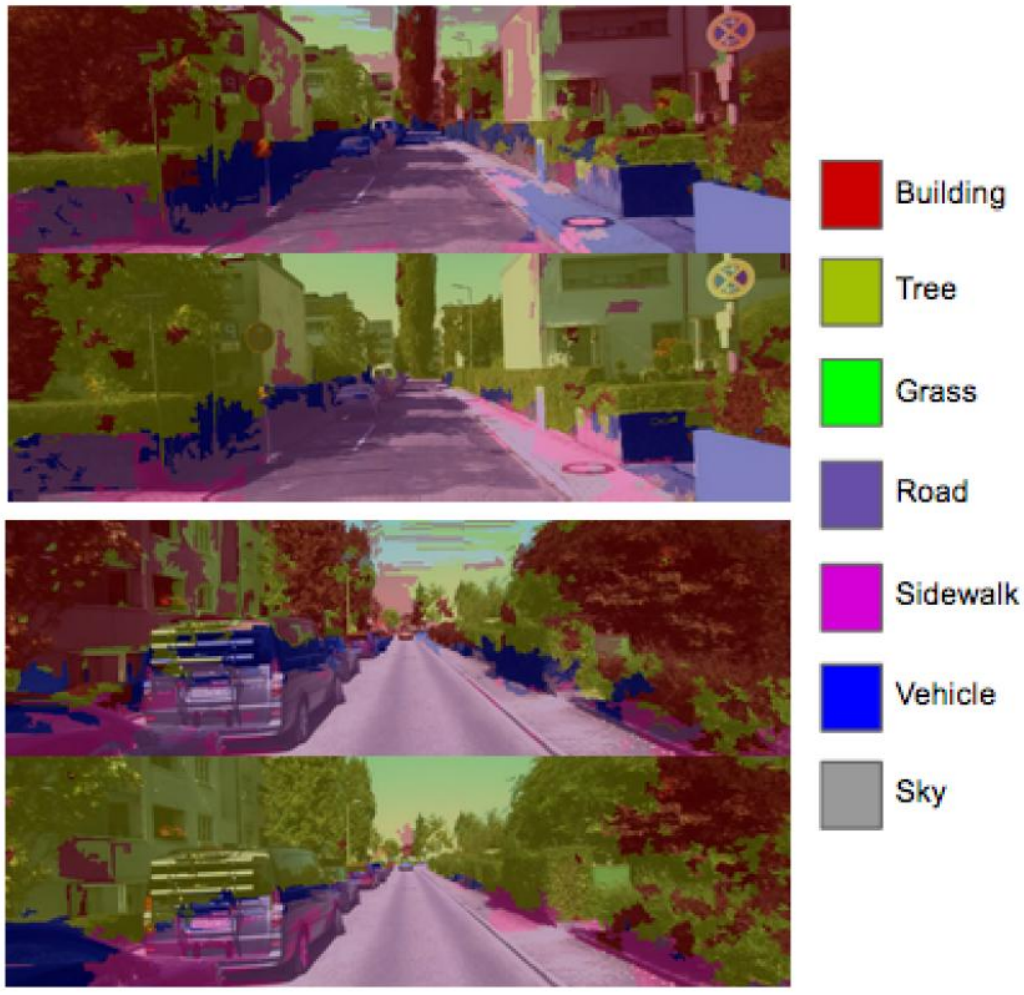

Figure 15. Example results for experiment in residential/suburban area where prior is biased to tree; for each pair: top - infer without priors, bottom - infer with priors; note that most vertical parts of the scenes are labeled as trees when prior is applied.

Table 2: Difference of Prior Model Confusion Matrix and Non-Prior Model Confusion Matrix

Infer → True ↓	Sky	Tree	Floor	Building	Vehicle
Sky	**54966**	689	0	-55655	0
Tree	103039	**46038**	62981	-207918	8429
Floor	0	-39705	**114069**	6588	56942
Building	14351	308097	-1957	**-291403**	-27671
Vehicle	0	113811	50897	-14850	**-37356**

Furthermore, due to the large number of images that are suburban scenes, most underlying manmade structures are occluded by foreground objects such as trees. As a result, the prior model learned from the training set is biased towards trees especially for vertical regions of the street scenes.

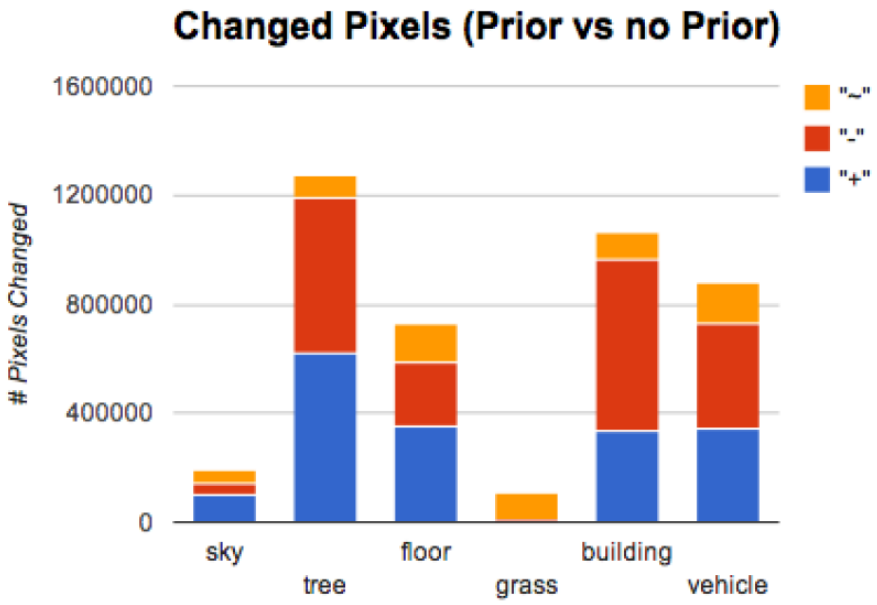

Figure 16. Residential Scene Experiment - Number of pixels where labels changed after prior is applied; blue ("+") refers to pixels corrected from incorrect labels; red ("-") refers to pixels changed from correct labels to incorrect ones; yellow ("~") refers to pixels changed from incorrect labels to another incorrect ones. Note bias in tree labels lead to misclassification in building and vehicle.

Examples can be seen in actual scene images Fig.12 or sample prior Fig.13. The issue of data bias has also been studied in the context of object recognition recently [19], [28]. Note that in

our experiments it does not influence label classes such as sky and road significantly, which are distributed in relatively fixed spatial locations in the images. In fact, one can see from Fig.16 that with priors the number of sky and floor segments that are corrected exceeds the number of misclassified ones. However, the vertical parts of the scenes suffer from the data bias issue since trees, buildings, and vehicles are mostly distributed in overlapping regions with occlusion. As analyzed before, with the biased prior, the inference is more likely to label tree over building and vehicle for such regions.

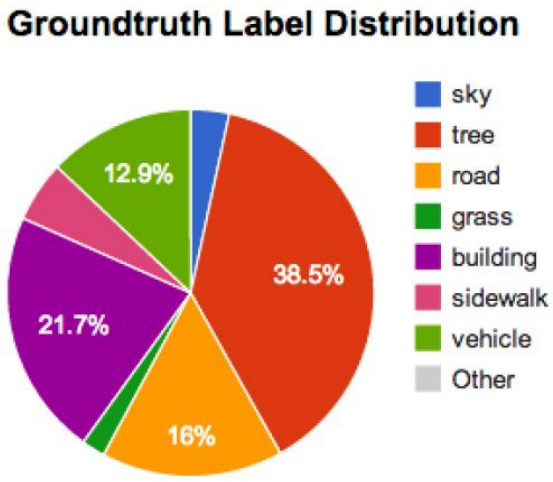

Figure 17. Ground-truth label percentage (residential scene dataset); Note the large percentage of tree labels.

It is reasonable because the inference machine is highly capable of empirical inference and fits data very well. Given relatively small (200 images) data set, it is not surprising that, by adding biased prior, the inference becomes more biased as well.

D. Evaluation on re-weighting label distribution with prior context

Fig 18 shows the performance of two approaches on incorporating priors: re-weighting using logistic regression and appending prior into the inference machine. The small data set of urban street scenes is used for this experiment. Note that the inference machine fits data by taking prior information as input features, while the re-weighting approach combines the output of scene parsing algorithm with prior information. Thus, appending prior approach in the

inference machine has the advantage of aggressively using prior information in the structure prediction and fitting training data well in experiments, but on the downside might lead to bias and over fitting in practice. While the re-weighting approach has the advantage of generality since it can work with any generic scene parsing algorithms in practice, but on the downside might not match the performance of appending prior in the inference machine in experiments.

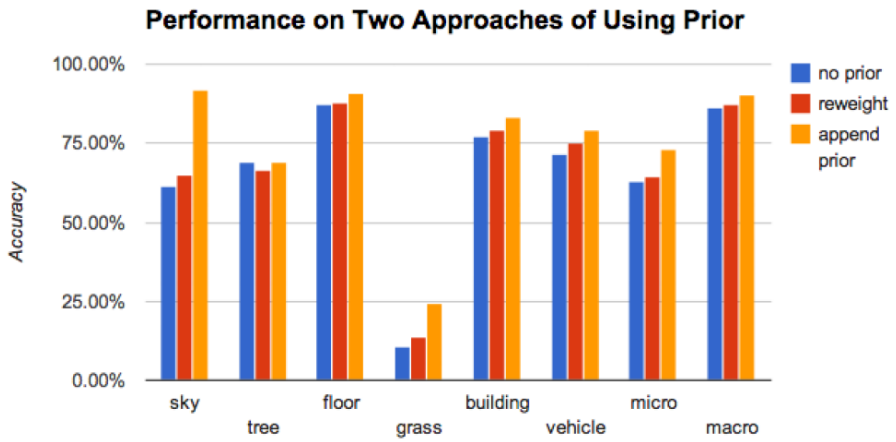

Figure 18. Performances on two approaches of incorporating priors: re-weighting using logistic regression, and appending prior into inference machine. Dataset used: urban scenes.

We also attempted to re-weight using a non linear classifier, boosted trees, and experimented on the relatively large data set of suburban/natural scenes. Fig 19 shows the performance, and not surprisingly, using prior also suffers from map limitation and data set bias issue, resulting in decreasing performance in several testing images.

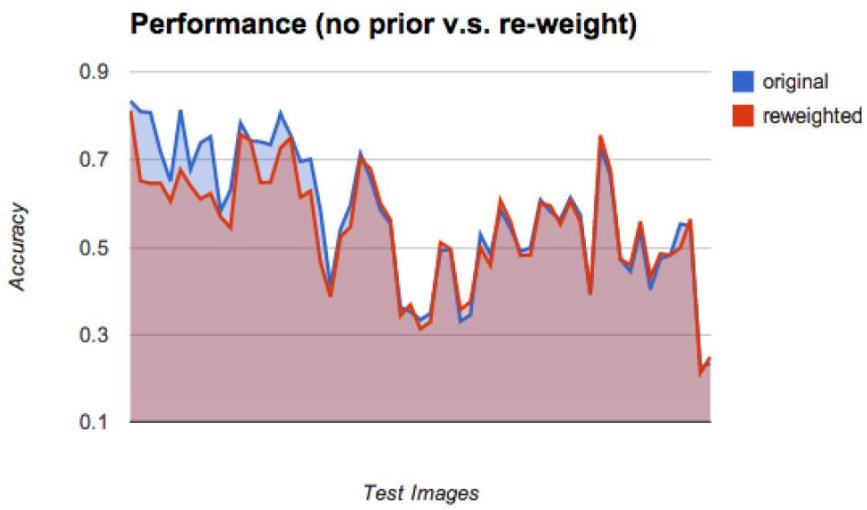

Figure 19. Performance of incorporating priors by re-weighting using boosting trees. The experiment is run on the relatively large data set of suburban/natural scenes. Note that using prior decreases performance on average given the dataset bias and limitation of map prior. Dataset used: residential scenes.

Discussion

A. Current Use of External Information

The current use of map information could not model the whole complexity and richness of visual information in the images. It is shown from small experiments that in certain cases such as urban street scenes where the map information models the scene structures well, the prior information can improve the accuracy and enforce the invariance of inference at spatially consistent regions such as sky and roads.

However, in cases such as natural/unstructured scenes, the current map prior has limitation in modeling and relies on empirical statistics, which leads to decrease in robustness depending on training data size. Given the limitation of the extent of map information, we believe richer representation of the scene prior such as 3D geometric features, or more extensive use of higher level knowledge can further help scene parsing.

B. Label Confidence

In most evaluations from the experiments, we used the label accuracy as the metric to analyze system performance. The accuracy is computed such that the label of highest inferred probability for each region is used to compare with ground-truth label. While the highest probability label is the one we used to label each region, the underlying label distributions may be a richer representation of the label confidence and belief for each region. This suggests that further metrics on probability distribution instead of label accuracy can potentially be used to evaluate the system performance and the effect of using prior. As an example, we attempted relative confidence to test how discriminative an inferred label is for each segment. It is defined as the difference between the highest probability and the second highest probability in the label distribution for each segment.

As visualized in the Fig.20, C is the relative confidence of inferred label distribution from a model that does not use priors and D the one that uses the priors. Note in C without prior, the incorrect sky segment is uncertain, but in D with prior, it is corrected also becomes certain. The correct building segment also becomes more certain with prior applied.

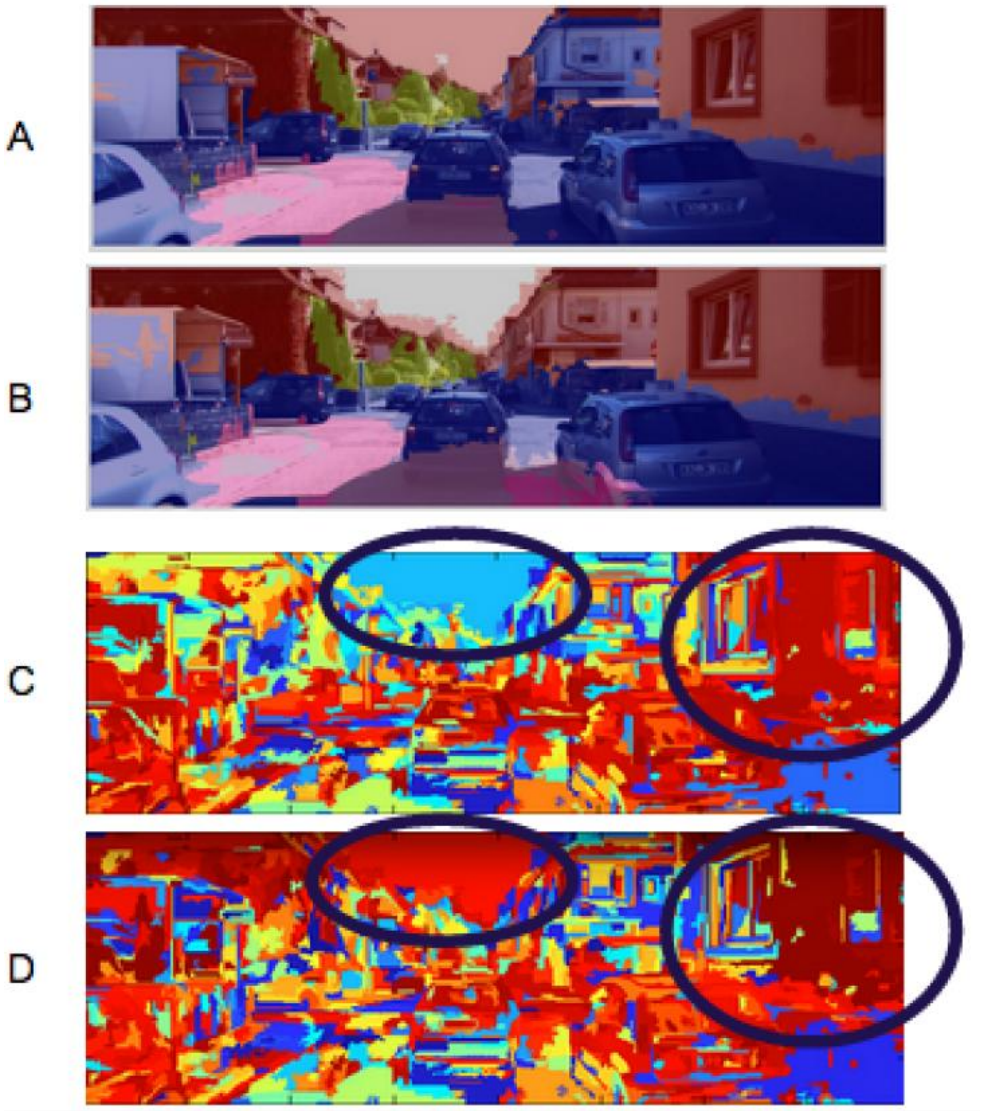

Figure 20. Label Confidence; top two are label inference results from models without prior (A) and with prior (B); bottom two are corresponding relative confidence (highest probability - 2nd highest probability) heat map (C-without prior, D-with prior); redder indicates more confident/discriminant about the labels, while bluer indicates more uncertain about the labels. Note the incorrect sky segment was uncertain without prior, but corrected with prior and was also certain; the correct building segment also becomes more certain with prior applied.

References

[1] ALVAREZ, J. M., LUMBRERAS, F., LOPEZ, A. M., AND GEVERS, T. Understanding road scenes using visual cues and gps information. In *Proceedings of the 12th international conference on Computer Vision* - Volume Part III (Berlin, Heidelberg, 2012), ECCV'12, Springer-Verlag, pp. 635–638.

[2] AMLACHER, K., FRITZ, G., LULEY, P., ALMER, A., AND PALETTA, L. Geo-contextual priors for attentive urban object recognition. In *Proc. of the 2009 IEEE International Conference on Robotics and Automation ICRA '09* (May 2009), pp. 1214 –1219.

[3] BADRINARAYANAN, V., GALASSO, F., AND CIPOLLA, R. Label propagation in video sequences. In *Proc. of the 2010 IEEE Conference on Computer Vision and Pattern Recognition CVPR*, June 2010, pp. 3265–3272.

[4] BERGER, A. L., PIETRA, S. A. D., AND PIETRA, V. J. D. A maximum entropy approach to natural language processing. *COMPUTATIONAL LINGUISTICS* 22 (1996), 39–71.

[5] BYRD, R. H., NOCEDAL, J., AND SCHNABEL, R. B. Representations of quasi-Newton matrices and their use in limited memory methods. *Math. Program*. 63, 2 (Jan. 1994), 129–156.

[6] CHEN, D. M., BAATZ, G., KOSER, K., TSAI, S. S., VEDANTHAM, R., PYLVANAINEN, T., ROIMELA, K., CHEN, X., BACH, J., POLLEFEYS, M., GIROD, B., AND GRZESZCZUK, R. City-scale landmark identification on mobile devices. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition* (Washington, DC, USA, 2011), CVPR '11, IEEE Computer Society, pp. 737–744.

[7] CHEN, X., LI, Q., SONG, Y., JIN, X., AND ZHAO, Q. Supervised geodesic propagation for semantic label transfer. In Computer Vision ECCV 2012, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds., vol. 7574 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2012, pp. 553–565.

[8] DIVVALA, S. K., HOIEM, D., HAYS, J., EFROS, A. A., AND HEBERT, M. An empirical study of context in object detection. In *Computer Vision and Pattern Recognition CVPR (2009)*, pp. 1271–1278.

[9] EVERINGHAM, M., VAN GOOL, L., WILLIAMS, C. K. I., WINN, J., AND ZISSERMAN, A. The Pascal visual object classes (VOC) challenge. *International Journal of Computer Vision* 88, 2 (June 2010), 303–338.

[10] FLOROS, G., AND LEIBE, B. Joint 2d-3d temporally consistent semantic segmentation of street scenes. In *Computer Vision and Pattern Recognition (CVPR),* 2012 IEEE Conference on (June 2012), pp. 2823 –2830.

[11] GEIGER, A., LAUER, M., AND URTASUN, R. A generative model for 3d urban scene understanding from movable platforms. In *Computer Vision and Pattern Recognition* (CVPR) (Colorado Springs, USA, June 2011).

[12] GEIGER, A., LENZ, P., AND URTASUN, R. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Computer Vision and Pattern Recognition* (CVPR) (Providence, USA, June 2012).

[13] GEIGER, A., WOJEK, C., AND URTASUN, R. Joint 3d estimation of objects and scene layout. In *Neural Information Processing Systems* (NIPS) (Granada, Spain, 2011).

[14] GOULD, S., FULTON, R., AND KOLLER, D. Decomposing a scene into geometric and semantically consistent regions. In *2009 IEEE 12th International Conference on Computer Vision* (Sep. 29-Oct. 2 2009), pp. 1 –8.

[15] HE, X., ZEMEL, R., AND CARREIRA-PERPINAN, M. Multiscale conditional random fields for image labeling. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR 2004*. Vol. 2, pp. II–695– II–702 Vol.2.

[16] HINTON, G. Training products of experts by minimizing contrastive divergence. *Neural Computation* 14 (2000), 2002.

[17] YUEN, J. C.LAWRENCE ZITNICK, C. L. A. T. A framework for encoding object-level image priors. *Microsoft Research Technical Report*, MSR-TR-2011-99, 2011.

[18] KEYSERS, D., OCH, F. J., AND NEY, H. Maximum entropy and Gaussian models for image object recognition. In *Pattern Recognition, 24th DAGM Symposium* (2002), Springer Verlag, pp. 498–506.

[19] KHOSLA, A., ZHOU, T., MALISIEWICZ, T., EFROS, A., AND TORRALBA, A. Undoing the damage of dataset bias. In *European Conference on Computer Vision (ECCV)* (Florence, Italy, October 2012).

[20] LIU, C., YUEN, J., AND TORRALBA, A. Nonparametric scene parsing via label transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33, 12 (dec. 2011), 2368 –2382.

[21] MAHAMUD, S., HEBERT, M., AND LAFFERTY, J. Combining simple discriminators for object discrimination. In *European Conf. on Computer Vision (ECCV)* (2002).

[22] MIKSIK, O., MUNOZ, D., BAGNELL, J. A., AND HEBERT, M. Efficient temporal consistency for streaming video scene analysis. In *IEEE International Conference on Robotics and Automation ICRA 2013*.

[23] MUNOZ, D., BAGNELL, J. A., AND HEBERT, M. Stacked hierarchical labeling. In *Proceedings of the 11th European conference on Computer Vision*: Part VI (Berlin, Heidelberg, 2010), ECCV'10, Springer-Verlag, pp. 57–70.

[24] MURPHY, K., TORRALBA, A., EATON, D., AND FREEMAN, W. Object detection and localization using local and global features. *Towards Category-Level Object Recognition*, Vol. 1 (2005).

[25] STURGESS, P., ALAHARI, K., LADICKY, L., AND TORR, P. H. S. Combining appearance and structure from motion features for road scene understanding. In *British Machine Vision Conference*, BMVC 2009, London, UK, September 7-10, 2009. Proceedings (2009), British Machine Vision Association.

[26] TIGHE, J., AND LAZEBNIK, S. Superparsing: scalable nonparametric image parsing with superpixels. In *Proceedings of the 11th European conference on Computer Vision*: Part V (Berlin, Heidelberg, 2010), ECCV'10, Springer-Verlag, pp. 352–365.

[27] TORRALBA, A. Contextual priming for object detection. *Int. J. Comput. Vision* 53, 2 (July 2003), 169–191.

[28] TORRALBA, A., AND EFROS, A. A. Unbiased look at dataset bias. In *CVPR'11* (June 2011).

[29] YAO, J., FIDLER, S., AND URTASUN, R. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (June 2012), pp. 702 –709.

[30] ZHANG, H., XIAO, J., AND QUAN, L. Supervised label transfer for semantic segmentation of street scenes. In *Proceedings of the 11th European conference on Computer Vision*: Part V (Berlin, Heidelberg, 2010), ECCV'10, Springer-Verlag, pp. 561–574.

[31] ZHU, S. C., WU, Y., AND MUMFORD, D. Filters, random fields and maximum entropy (FRAME) – towards a unified theory for texture modeling. *International Journal of Computer Vision* 27, 2 (1998), 1–20.