Data-Driven Road Detection

Jose M. Alvarez, Mathieu Salzmann, Nick Barnes
NICTA; Canberra ACT 2601, Australia
{jose.alvarez, mathieu.salzmann, nick.barnes}@nicta.com.au

Abstract

In this paper, we tackle the problem of road detection from RGB images. In particular, we follow a data-driven approach to segmenting the road pixels in an image. To this end, we introduce two road detection methods: A top-down approach that builds an image-level road prior based on the traffic pattern observed in an input image, and a bottom-up technique that estimates the probability that an image superpixel belongs to the road surface in a nonparametric manner. Both our algorithms work on the principle of label transfer in the sense that the road prior is directly constructed from the ground-truth segmentations of training images. Our experimental evaluation on four different datasets shows that this approach outperforms existing top-down and bottom-up techniques, and is key to the robustness of road detection algorithms to the dataset bias.

1. Introduction

Reliable road detection is key to the success of many road scene understanding applications, such as autonomous driving, driver assistance (e.g., lane keeping [25]), and personal navigation [14]. Furthermore, it can serve as a preprocessing stage for higher level reasoning and understanding of road scenes [13]. Road detection is generally geometrically well-constrained since roads are designed to support vehicles with Ackermann steering. Despite recent progress, we are still far from having ideal solutions. Road detection remains a challenging computer vision problem due to the large variability of images acquired world-wide at different times of the day, with changing lighting, weather, and variable road conditions. Strong interest in the area has recently led to datasets and evaluation benchmarks (e.g., [12]).

Due to the lack of annotated data, early road detection methods have often focused on the online scenario [4, 28]. To this end, they typically make the assumption that pixels of the bottom part of the input image belong to the road surface, and that their appearance distribution can be used to classify the remaining pixels. While this simple approach can be effective, it will fail when the distribution of the bottom pixels is biased, e.g., due to the presence of strong shadows, or when those pixels do not belong to the road, but rather to another object such as another car.

Recently, a number of labeled datasets have become publicly available (e.g., CamVid [9] and KITTI [14]), giving rise to the possibility of designing supervised methods for road detection. Most existing supervised techniques follow a bottom-up approach and train pixel classifiers based on appearance and motion features [29]. Unfortunately, as noted in [9], the resulting classifiers typically suffer from a strong dataset bias, thus yielding poor performance when the test images are acquired in different conditions from the training data (e.g., dusk vs. day in CamVid). While it has been shown that such classifiers could be employed to learn better features from unsupervised data [3], the resulting technique suffers from the same drawback as bootstrapping approaches that can degenerate when the initial classifier performs poorly.

As an alternative, top-down road detection approaches have recently started to emerge. To the best of our knowledge, such approaches are currently limited to estimating a rough segmentation of the road by building a global [4], or three-class (i.e., straight, left curve, right curve) [2] road prior from training data. These priors, however, only yield a coarse estimate of road geometry, since they do not take into account the traffic conditions observed in the input image.

In this paper, we introduce two data-driven methods for road detection. Both methods fall into the label-transfer category, in the sense that they do not learn classifiers, but directly exploit the labels of training images to build road priors. In particular, we propose: (1) a top-down technique that generates an image-level road prior by retrieving a small number of training images with similar traffic pattern as the input image; and, (2) a nonparametric bottom-up approach to building a superpixel-level road prior based on appearance similarity between the input and training superpixels.

More specifically, our image-level method first makes
use of existing detectors to generate likelihood maps for typical street objects, such as cars and pedestrians, and attempts to reconstruct these maps as sparse linear combinations of similar maps computed on the training data. We then transfer the labels of the training images that have non-zero weights and merge them to form a pixel-level road likelihood. As evidenced by our results, this likelihood is more accurate than the ones used in the past [4]. For our bottom-up model, we build a likelihood for each image superpixel by retrieving its $K$ nearest neighbors amongst both road and non-road training superpixels. This nonparametric approach has the advantage over learned classifiers, such as [9, 29, 21], of not requiring re-training as new labeled data becomes available, and of being able to retrieve rare cases.

Finally, we contribute a thorough empirical comparison of our algorithms with existing approaches on four standard datasets and combinations of these datasets. Our evaluation clearly evidences the benefits of our label transfer approach over existing supervised methods. In particular, our framework proves robust to the dataset bias and provides an accurate solution to road detection.

2. Related Work

Over the years, the most popular approach to detecting road has been to perform bottom-up pixel, or superpixel, labeling. Since learning a general appearance model for non-road pixels is impractical, most methods focus on single-class classifiers and thus attempt to model the appearance distribution of road pixels only. To specifically handle the scene at hand, a number of works incorporate the assumption that road appears in the lower part of the image and is representative of the road appearance in the whole image. In this context, low-level cues such as color [6, 28, 19, 30], or a combination of color and texture [24] have been used. To account for the large intra-class variability of road pixels, these methods attempt to model data invariants by making use of different color spaces, such as HSV [28], normalized $rg$ [30], physics-based illuminant invariant color spaces [6, 19], and linear combinations of color planes [5, 3]. Unfortunately, the assumption underlying these methods has two major limitations: (1) In some scenarios the bottom part of the image does not belong to the road, e.g., it belongs to another vehicle; (2) When these pixels belong to the road surface, their appearance may not be representative of road pixels in the entire input image, e.g., strong shadows.

To overcome these drawbacks, several works make use of training images. For instance, in [7], the assumption that the bottom pixels are road pixels was also exploited, but in unsupervised training images instead of the input image, yielding a better model. More directly related to our work, the increasing availability of labeled images has lead to several supervised approaches. In particular, some papers follow a more traditional semantic labeling approach [29]. To this end, appearance and motion features are extracted to learn a pixel classifier, and smoothness is encouraged by modeling the image with a Conditional Markov Random Field (CRF). Other classifiers, such as boosting [33], have been employed to form a road model from training images. In [3], labeled data was used to learn a road model, which was then refined with unsupervised images to better reflect the scene at hand. Alternatively, several methods exploit the sequential nature of the data and use detection results in previous images as training data for the current frame [28, 30].

In contrast to the above-mentioned bottom-up techniques, top-down approaches build a road prior for the entire image. This class of methods can be thought of as exploiting contextual information to detect the road. For instance, a rough segmentation of the road can be obtained by locating the vanishing point [20, 27]. In [18], contextual information is exploited to recover 3D scene layout from which road segmentation is performed by assuming that the road lies within horizontal surfaces. In [17], the effectiveness of road priors based on restricted appearance was demonstrated by combining simple shape constraints with standard graph cut approaches to show significant improvement in pixel classification performance. Alternatively, a global road prior can be obtained by averaging over the ground-truth labels of all training images [4]. In [2], a more accurate prior was built by first automatically discriminating between different road geometries (i.e., left turn, straight, right turn).

Although road scenes are well structured and, in many cases, can thus be well-handled by standard classifier-based approaches, individual images are still subject to enormous appearance variations. Individual images can be heavily shadowed in unique ways, roads surfaces can be degraded, lanes can be brightly painted, and many different road shapes are possible. As a consequence, individual datasets will often be insufficient to characterize other sets. In learning-based methods this is known as the dataset bias, and was observed in [9]. A similar argument was presented in [16] regarding current appearance-based approaches to labeling road scenes. This presents a challenge for bottom-up and top-down methods. First, with strong shadowing or the appearance of a car close-by, the bottom of the image may not effectively characterize the appearance of the road. More broadly in classifier-based approaches, a model will be formed to characterize the distribution of training images. In such models, unusual images can be lost in the distribution (e.g., a small section of brightly painted bus lane). Similarly, existing top-down techniques may perform well in standard scenarios where no other vehicles or objects are present on the road. However, in urban scenes where the diversity is much larger, they tend to be highly inaccurate.
In these challenging situations, nonparametric approaches can help. Rather than trying to learn a classifier, nonparametric methods retrieve similar training images (or superpixels). Thus, even though there are few instances, these can be used to transfer labels to the current image. Another advantage of nonparametric approaches is that, as no training is used, new difficult images can be more easily added when difficult situations arise. In this paper, we therefore advocate the use of nonparametric approaches to road detection and demonstrate the benefits of the resulting algorithms on and across several datasets.

3. Traffic-Aware Image-Level Road Prior

In this section, we introduce a label transfer approach to building an image-level road prior that accounts for the traffic pattern in the input image. Our algorithm is inspired by the fact that roads are also distinguished by the objects that occupy them, such as vehicles and pedestrians. Intuitively, similar objects in similar places in the images would imply similar road shapes. Therefore, we build a prior by exploiting existing object detectors that have proved relatively robust to the dataset bias [32]. In particular, we employ the Deformable Part Models (DPMs) of [11] trained on the Pascal VOC2009 images. While ObjectBank [22] also makes use of similar detectors, our approach significantly differs from it in the sense that we do not use object detector responses as image features, but directly employ them to build a road prior.

More specifically, let \( x_* \in \mathbb{R}^{N_p} \) be the vector concatenating the response at each pixel of one specific detector, e.g., car, evaluated in a sliding-window manner over a given input image. Furthermore, let \( X \in \mathbb{R}^{N_p \times N} \) be the matrix of similar detector responses for \( N \) training images. Our goal is to find a small number of training images whose detector responses can be combined to approximate \( x_* \). This can be expressed as

\[
\min_w \frac{1}{2} \| Xw - x_* \|_2^2 + \lambda \| w \|_1, \tag{1}
\]

where \( w \in \mathbb{R}^N \) is the vector of combination weights. To encourage selecting a small number of training images, we make use of an \( L_1 \)-norm sparsity-inducing regularizer on \( w \), whose influence is regulated by the scalar \( \lambda \). Problem (1) is a convex optimization problem, which we solve using the efficient publicly available SLEP [23] software.

Given the weights \( w \) obtained by solving (1), we can then directly build a road prior by leveraging the availability of ground-truth pixelwise road labels for the training images. To this end, let \( Y \in \mathbb{R}^{N_p \times N} \) be the matrix of binary labels indicating the presence or absence of road at each pixel of the \( N \) training images. The probability of each pixel in the input image belonging to the road surface can be estimated as

\[
p(y_* = 1|w) \propto Yw, \tag{2}
\]

where \( y_* \in \mathbb{R}^{N_p} \) is the vector of binary labels for the test image that we seek to estimate.

To increase the robustness of our approach to different traffic conditions, we combine the predictions obtained from the detector responses for multiple objects. To this end, we follow a simple product rule, which lets us write

\[
p(y_* = 1) = \prod_{i=1}^{N_d} p(y_* = 1|w_i), \tag{3}
\]

where \( N_d \) is the number of different detectors, and \( p(y_* = 1|w_i) \) is the probability estimated as in Eq. 2 for the \( i^{th} \) detector with parameters \( w_i \) obtained by solving a problem of the form given in (1).

In practice, to reduce the computational cost of solving (1), we do not work at pixel-level, but employ an oversegmentation of the images (i.e., \( N_p \) is not the number of pixels, but the number of segments). A value in \( x_* \), or \( X \) is then computed as the mean detector response value in one specific segment. Furthermore, to prevent having to recompute this oversegmentation for each test image, we make use of a fixed oversegmentation. This allows us to precompute \( X \) and keep it fixed for all test images. As will be shown in our experiments, this approximation entails virtually no loss of accuracy.

4. Nonparametric Superpixel Road Prior

In this section, we introduce our nonparametric approach to building a superpixel-level road prior. To this end, we first compute an oversegmentation of the input image and, for each superpixel, follow an approximate nearest-neighbor strategy to retrieve similar superpixels from the training data and transfer their label to form a prior. While a related label transfer approach was proposed in [31], their SuperParsing algorithm works by first retrieving a small set of similar images based on image-level features. This typically is ill-suited for road detection due to the presence of local appearance variations (e.g., strong shadows). The set of images retrieved based on global similarity will not contain the full diversity of superpixels of all road images in the dataset, and may be inadequate to represent the road in the input image. Here, instead, we directly work at superpixel level for retrieval.

More specifically, let \( f_* \in \mathbb{R}^F \) be the vector of image features for one superpixel in the input image, obtained by extracting features at the centroid of the superpixel. Similarly, let \( \{ f_i \}_{i \in \mathcal{R}} \) and \( \{ f_i \}_{i \in \mathcal{R}} \) be the sets of image feature vectors for the training road \( (i \in \mathcal{R}) \) and non-road superpixels, respectively. We make use of an approximate approach to retrieve the \( K \) nearest neighbors of \( f_* \) among the
Figure 1. Example of the images used in our experiments. We show images from: First row, KITTI [14]; Second row, CamVid [9]; Third row, BCN-1 [6]; Last row, BCN-sunny and BCN-rain [6].

road and non-road superpixels independently. Given these nearest neighbors, we estimate the probability of an input superpixel belonging to road as

$$p(y_\star = 1) = 1 - \frac{\sum_{k \in N_R(f_\star)} ||f_\star - f_k||}{\sum_{k \in N_R(f_\star)} ||f_\star - f_k|| + \sum_{k \in \bar{N}_R(f_\star)} ||f_\star - f_k||}$$

(4)

where $N_R(f_\star)$ and $\bar{N}_R(f_\star)$ denote the sets of $K$ approximate road and non-road nearest neighbors, respectively.

In practice, we make use of the ANN library [26] to compute the approximate nearest neighbors. In particular, we employ their kd-tree algorithm with Euclidean distance. This involves setting one approximation parameter $\epsilon$. As will be shown in Section 5, the final accuracy of our approach is quite robust to the value of this parameter.

5. Experimental Evaluation

In this section, we provide a thorough evaluation of our algorithms and compare them against several baselines. We particularly use multiple datasets and explore robustness to the dataset bias.

5.1. Setup

Datasets: We conducted experiments on a set of 3119 images from 4 different road datasets: CamVid [9], KITTI [14], BCN-1 [6] and Barcelona-2 [6]. Our data comprises images acquired with different cameras, in different real driving scenarios, and at different times of the day. The different datasets are summarized in Table 1, and representative images are shown in Fig. 1. Note that, in our experiments, we will show results obtained by training on one dataset and testing on the others, as well as training on combinations of datasets and testing on the remaining data.

Parameters: Our two algorithms rely on a number of parameters that were set as defined below and kept fixed in all our experiments. Note that the sensitivity of our results to these parameters will be studied in Section 5.3.

For our image-level prior, we used 5 different object detectors: bicycle, bus, car, person, motorbike. These detectors are DPMs trained on Pascal VOC2009 images with the code of [10] (release 4). For each detector, to obtain the weights $w$ from (1), we ran 100 iterations of the SLEP optimization library [23].

For our superpixel-level prior, we employed the SLIC oversegmentation of [1] to obtain roughly 1800 superpixels per image. We computed a grayscale SIFT descriptor at the centroid of each superpixel and reduced the dimensionality of these descriptors to 35 using a random projection [8]. We then retrieved $K = 25$ road and non-road nearest neighbors for each superpixel. In the ANN library, we empirically set $\epsilon$ to 3, which gives a good trade-off between speed and accuracy.

5.2. Empirical Evaluation

We compare our results against the baselines on all four datasets. In the following discussion, we report the accuracy of each method as the area under the ROC curve computed for the entire test set (i.e., not image-by-image).

Baselines: We make use of two baselines: a top-down (image-level prior) and a bottom-up (pixel-level) prior. The top-down baseline consists of building a global road prior by averaging all training road segmentations [4]. The bottom-up baseline performs pixel labeling based on one-versus-all boosted decision tree classifiers trained on multiple image features, including the responses of 17 filters, RGB colors, dense HOG features, and intensity averages over image rows and columns. The output of the decision trees are then calibrated using a multi-class logistic regression classifier. We employed the publicly available DARWIN library [15], which yields very accurate results for many pixel labeling tasks.

<table>
<thead>
<tr>
<th>Code</th>
<th>Resolution</th>
<th>Train</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>UK</td>
<td>CamVid-Day</td>
<td>960 x 720</td>
<td>406</td>
</tr>
<tr>
<td></td>
<td>CamVid-Dusk</td>
<td>960 x 720</td>
<td>62</td>
</tr>
<tr>
<td>GB</td>
<td>KITTI</td>
<td>1241 x 376</td>
<td>323</td>
</tr>
<tr>
<td></td>
<td>BCN-1</td>
<td>640 x 480</td>
<td>500</td>
</tr>
<tr>
<td></td>
<td>BCN-2 Sunny</td>
<td>640 x 480</td>
<td>−</td>
</tr>
<tr>
<td></td>
<td>BCN-2 Rain</td>
<td>640 x 480</td>
<td>−</td>
</tr>
</tbody>
</table>

Table 1. Summary of the datasets used in our experiments.
Overall, we see that the image-level road prior consistently yields accurate road segmentation and outperforms the baseline in the cases where the training data contains a good diversity of vehicles and pedestrians, as well as different road shapes and sizes. In cases where there is a lack of diversity in the training set, such as in BCN-1, as expected, the image-level road prior performs similarly to the baseline. For our bottom-up method, we see that when training on CamVID, and on KITTI, we outperform or perform similarly to the baseline in all cases, which shows robustness to dataset bias when there is good appearance diversity in the training set. However, when our bottom-up method trains on BCN, the performance of the baseline is often superior. If you consider the appearance of the BCN images in Fig. 1, they are clearly different from the others. That is this dataset does not contain sufficient diversity for a label transfer method to draw upon. Overall, this demonstrates that the label transfer approach can make excellent use of diversity, but only when it is present in the dataset. Fortunately, in a nonparametric approach, it is easy to add extra images.

We now focus on row K of Table 2 to provide a detailed analysis. In this row, the training images correspond to the entire CamVID dataset, which, as can be observed in the

---

**Table 3. Definition of datasets A to Q.**

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Num. Imgs.</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>CamVID-Day test</td>
<td>171</td>
</tr>
<tr>
<td>B</td>
<td>CamVID-Day train</td>
<td>406</td>
</tr>
<tr>
<td>C</td>
<td>CamVID-Dusk train</td>
<td>62</td>
</tr>
<tr>
<td>D</td>
<td>CamVID-Dusk test</td>
<td>62</td>
</tr>
<tr>
<td>E</td>
<td>subset of BCN-1</td>
<td>300</td>
</tr>
<tr>
<td>F</td>
<td>subset of BCN-1</td>
<td>500</td>
</tr>
<tr>
<td>G</td>
<td>BCN-2 Rain</td>
<td>251</td>
</tr>
<tr>
<td>H</td>
<td>BCN-2 Sunny</td>
<td>754</td>
</tr>
<tr>
<td>I</td>
<td>KITTI train</td>
<td>468</td>
</tr>
<tr>
<td>J</td>
<td>CamVis train</td>
<td>701</td>
</tr>
<tr>
<td>K</td>
<td>CamVis</td>
<td>800</td>
</tr>
<tr>
<td>L</td>
<td>BCN-1</td>
<td>1005</td>
</tr>
<tr>
<td>M</td>
<td>BCN-2</td>
<td>1005</td>
</tr>
<tr>
<td>N</td>
<td>BCN-1 + BCN-2</td>
<td>1805</td>
</tr>
<tr>
<td>O</td>
<td>CamVis + BCN-1 + BCN-2</td>
<td>2506</td>
</tr>
<tr>
<td>P</td>
<td>BCN-1 + BCN-2 + KITTI</td>
<td>2128</td>
</tr>
<tr>
<td>Q</td>
<td>CamVis + KITTI</td>
<td>1024</td>
</tr>
</tbody>
</table>

We show the results in Table 2, where the rows and columns (A to Q) correspond to different training and test sets, respectively, as described in Table 3. Note that we only show test sets up to I, since the remaining ones are combinations of those shown.
second row Fig. 1, depicts a broad diversity of road images in many aspects of common variation including different illumination conditions, density of traffic, and size of roads. In general, it can be seen that the global road prior of our top-down baseline performs reasonably, but is too general and does not yield accurate segmentation. In contrast, the bottom-up baseline does yield accurate results when also tested on CamVid (columns A, B, C, D), as well as reasonably good results when tested on the KITTI dataset (column I). This shows that this baseline has some degree of robustness to the dataset bias. However, when the test images become too different, as is the case with the BCN datasets (columns E, F, G, H), the bottom-up baseline accuracy drops dramatically. In contrast, our superpixel-level label transfer approach consistently yields good accuracy. This demonstrates that accurate labels can be retrieved from rarely occurring superpixels, the case where statistical classifiers are ill-suited. Similarly to our superpixel-level prior, our image-level road prior yields accurate detections, since the traffic patterns observed in the training images cover a large range of scenarios and are thus well-suited for label transfer. Importantly, despite the fact that the other datasets are not as diverse as CamVid in terms of vehicles and pedestrians, our image-level results remain accurate in all training/testing combinations.

For the sake of completeness of comparison, we also computed the results of two online methods that assume that the bottom part of the image contains road pixels. In particular, one method relies on computing SIFT features at the centroids of the superpixels of the image, and the other on histograms of LBP descriptors in the superpixels [3]. In both cases, the probability of road is estimated based on the distance to the nearest superpixel in the bottom part of the image. When evaluated on the CamVid Day test set, these techniques yield AUCs of 0.87 and 0.86, respectively, which is typically lower than our methods when the diversity of the training images is sufficient.

5.3. Parameter Sensitivity

We now study the robustness of our algorithms to their parameters. To this end, we employed a leave-one-out procedure on the entire CamVid dataset.

Image-level Prior: We first study how the accuracy of our top-down approach is affected by the number of iterations used to solve (1). In Fig. 2, we plot the resulting ROC curve corresponding to different number of iterations, which shows that accuracy is quite robust to the number. In Fig. 3, we show the evolution of the weights \( w \) as a function of the iteration number. A clear sparsity pattern can be observed around 100 iterations, which also yields a good trade-off between speed (1.26 sec) and accuracy (AUC 0.9746).

In Fig. 4, we illustrate the effect of using a fixed oversegmentation of the images for our image-level road prior. The black line corresponds to using the same fixed (pre-defined) superpixels for all test images, whereas the red one corresponds to computing the true superpixels for each test image. As can be observed from the ROC curves, this approximation yields virtually no loss of accuracy compared to the baseline. However, our approximate superpixels lead to a much faster runtime (1.26 sec vs. 705.9 sec for the exact superpixels). Note that, for the exact case, each image in the training set needs to be processed at test time.

Superpixel-level Prior: For our bottom-up approach, we study the influence of the number of approximate nearest neighbors, the approximation parameter \( \epsilon \) and the dimensionality of the random projections of the SIFT features.
While we ran a full set of experiments by letting all the parameters vary simultaneously, here, we report the accuracies obtained by varying a single parameter while keeping the other ones fixed to the values used in our experiments. These results are representative of the full set. Figs. 5b, 5c, and 5d depict the AUC as a function of $K$, $\epsilon$ and the number of random projections, respectively. Note that the results are quite robust to different values of these parameters. The parameters chosen for our experiments represent a good trade-off between speed and accuracy.

6. Conclusion

In this paper, we have introduced two data-driven approaches to road detection: a top-down method that builds an image-level road prior by exploiting the traffic pattern in the scene, and a bottom-up technique that exploits a non-parametric approach to predict the presence of road in each image superpixel. We have evaluated our methods and compared them against several baselines on four datasets containing images acquired in different conditions. Our experiments have shown that our algorithms are less prone to dataset bias than the baselines, and that our image-level road prior yields accurate road detection across all these datasets. This confirms our intuition that label transfer is robust to the domain shift occurring between different datasets and is therefore a promising direction for road detection. While the speed of our algorithms is already competitive with state-of-the-art learned classifier, in the future, we will focus on further improving these runtimes.

References

Figure 5. a) Influence analysis of the approximation parameter $\epsilon$ fixing $K = 25$ and the random projection to 35 dimensions. b) Influence analysis of the number of neighbors parameter $K$ fixing $\epsilon = 3$ and the random projection to 35 dimensions. c) Influence analysis of the number of random projections fixing $\epsilon = 3$ and $K = 25$. As shown, the performance drops off when projecting to fewer dimensions (e.g., 5 dims.).


