Towards Multi-Cue Urban Curb Recognition

Markus Enzweiler, Pierre Greiner, Carsten Knöppel, and Uwe Franke

Environment Perception Group
Daimler AG Group Research & Advanced Engineering
71059 Sindelfingen, Germany

E-Mail: firstname.lastname@daimler.com

Abstract—This paper presents a multi-cue approach to curb recognition in urban traffic. We propose a novel texture-based curb classifier using local receptive field (LRF) features in conjunction with a multi-layer neural network. This classification module operates on both intensity images and on three-dimensional height profile data derived from stereo vision.

We integrate the proposed multi-cue curb classifier as an additional measurement module into a state-of-the-art Kalman filter-based urban lane recognition system.

Our experiments involve a challenging real-world dataset captured in urban traffic with manually labeled ground-truth. We quantify the benefit of the proposed multi-cue curb classifier in terms of the improvement in curb localization accuracy of the integrated system. Our results indicate a 25% reduction of the average curb localization error at real-time processing speeds.

I. INTRODUCTION

Over the last years, visual cues derived from in-vehicle camera sensors have become the primary sensing device in many advanced driver assistance systems (ADAS). In those systems, visual data provides the main foundation for scene interpretation and understanding. A key requirement for such systems is the sound knowledge of possible driving space. On highways and rural roads there are usually explicit high-contrast lane markings available that are straightforward to detect using machine vision algorithms. Hence, many successful lane detection algorithms have been proposed and are available on the market today, see [5] for a review.

In urban environments however, things look quite different. There are often no explicit lane markings available that designate the boundary of the road. Here, curbs are of special interest since they define the distinct boundary between drivable space on the road and the sidewalk. From a machine vision point-of-view, curb detection is a much harder problem than lane marking detection, given that the contrast between curb, road and sidewalk may not be very pronounced and can significantly vary due to different height profiles along the curb, e.g. see Fig. 2.

II. RELATED WORK

In this overview, we do not cover methods for lane marking detection, see [5]. Instead, our focus is on image- and stereo-based curb detection, where several approaches have been proposed.

Se and Brady extracted curbs from clusters of parallel straight lines in the image that are recovered using the Hough transform of edge pixels [15]. This method has been further extended by Turchetto and Manduchi to exploit 3D information from stereo vision. They proposed to weight each edge pixel within the Hough accumulator by a function of the brightness gradient and the 3D elevation gradient [17]. Additionally, the estimated 3D surface curvature has been incorporated into the weighting function [8].

In [9], the 3D points are arranged in a horizontal height grid, denoted as Digital Elevation Map (DEM). A height value is determined for each grid cell. Curb candidates are then detected from height discontinuities of neighboring cells.
using a combination of Canny edge detection and Hough transform on the DEM.

A common drawback of the previously mentioned approaches is their restriction to the detection of straight-line curbs. Furthermore, curb detection via the elevation gradient is very sensitive to artifacts caused by measurement noise and outliers. The observed height discontinuity gets increasingly blurred with larger distances to the camera. In order to detect low and distant curbs, small edge detection thresholds are required which generally cause many false curb candidates.

With reliable egomotion information at hand, Oniga et al. have developed a way to remove those false candidates by temporal integration [10]. They further extend their curb detection scheme to curved curbs which are modeled using chains of straight-line curb segments. Their approach is capable of detecting curbs up to a distance of 10 m but is restricted to curbs with a constant height.

To overcome this drawback, Siegemund et al. present a graph-based approach to curb reconstruction which is able to reliably detect and reconstruct curved curbs at varying heights up to a distance of 20 m [16]. Their method is based on 3D point observations which are arranged in a DEM scheme. The main assumption is that - even at large distances - the average measured height levels of the curb adjacent surfaces, i.e. road and sidewalk, are different. This constraint is exploited within a parametrized environment model, where the curb is defined as horizontal separation of its adjacent surfaces. Each DEM cell is then assigned to one part of the environment using a Conditional Random Field (CRF) model. As such, the problem of curb classification is cast as a labeling problem of DEM cells given 3D point measurements and the environment model. The best labeling is then obtained by inference on the CRF.

In contrast to previous approaches, we propose to address curb recognition as an image-based classification problem exploiting powerful multi-cue texture-based classifiers which have been widely applied to many object detection problems with great success. A detailed review of such classification models is out-of-scope for this paper. We refer to [3], [7] for in-depth surveys.

An overview of our proposed urban curb recognition system is shown in Fig. 1. Based on an initial curb location expectation, we use the responses of a multi-cue curb classifier to get a refined curb estimate. Details are given in the following section.

III. SYSTEM ARCHITECTURE

A. Data Acquisition

Our main sensor is a vehicle-mounted stereo vision camera system that provides intensity image pairs (grayscale) as well as dense disparity images using Semi-Global Matching (SGM) stereo [4], [6]. We further filter out large elevated objects such as walls, cars, pedestrians and other traffic participants and infrastructure elements using Stixel freespace information [11], see Fig. 3.

Fig. 3. Data acquisition pipeline. (a) Stereo image pair. (b) Corresponding SGM disparity image. (c) Stixel World [11] computed from (b). (d) Stixel freespace, where large elevated objects are filtered out.

B. Curb Classification

For curb classification, we use sub-regions of an image at a given size, e.g. 120 x 15 pixels at a fixed distance (15 m) as regions-of-interest (ROIs), see Sec. III-C. The aim of curb classification is to determine the posterior probabilities $P(\omega_c|x_k)$ and $P(\omega_{n}|x_k) = 1 - P(\omega_c|x_k)$ with respect to object classes $\omega_c$ (curb) and $\omega_n$ (no curb) for every ROI cut-out sample $x_k$. Hence, we address curb classification as a two-class problem.

A visual inspection of an exemplary cut-out containing a curb, see Fig. 4, reveals that curbs usually have a distinct contrast in the grayscale intensity domain along with a discriminative height profile. Both domains are highly uncoupled which motivates our multi-cue approach. In particular, we train individual classifiers for each domain and follow a high-level fusion strategy which allows to tune features specifically to each modality and base the final decision on a combined vote of the individual classifiers.

For classification, 12 bit grayscale images are used in the intensity classifier without any further photometric processing, see Fig. 4a. We only correct for perspective shearing effects based on the rough expectation of curb orientation at the image location where the classifier is applied. This expectation can arise from an initial estimate or prior knowledge, see Sec. III-C. In cases where no prior information is available, we assume parallel lines with a common vanishing point. As a result, curbs are predominantly vertically oriented in our cut-outs, see Fig. 4. Remaining variance in the data samples is then learned during classifier training.

To incorporate 3D height information, we utilize the SGM disparity map to compute an elevation image where each pixel denotes the elevation of a 3D measurement above the dynamically estimated ground-plane, see Fig. 4b. This eleva-
tion image is normalized to have zero mean and constitutes the input domain of the elevation classifier. Thus, a cut-out sample \( x_k \) consists of both the intensity representation \( x_k^e \) and the elevation \( x_k^c \): \( x_k = \{ x_k^e, x_k^c \} \).

Let \( P^e(\omega_c | x_k^e) \) denote the posterior probability estimate of the intensity classifier, and \( P^c(\omega_c | x_k^c) \) the corresponding estimate of the elevation classifier for the same sample \( x_k \).

We employ a support vector machine (SVM) [18] as a fusion classifier to discriminate between object classes \( \omega_c \) (curb) and \( \omega_n \) (no curb) in the space of posterior probabilities of the individual classifiers, c.f. [13].

Let \( p_k^c = (P^e(\omega_c | x_k^e), P^c(\omega_c | x_k^c)) \) denote the vector of individual posteriors for sample \( x_k \) with respect to object class \( \omega_c \). The corresponding SVM hyperplane is defined by:

\[
f_j(p_k) = \sum_i y_i \alpha_i \cdot K(p_i, p_k) + b
\]

(1)

Here, \( p_i^c \) denotes the set of support vectors with labels \( y_i \) and Lagrange multipliers \( \alpha_i \). \( K(\cdot, \cdot) \) represents the SVM kernel function. We use a non-linear RBF kernel in our experiments. The SVM decision value \( f_j(p_k^c) \) (distance to the hyperplane) is converted to a posterior probability, using a sigmoidal mapping with parameters \( A \) and \( B \) learned from the training set by Maximum-Likelihood [12]:

\[
P(\omega_c | x_k) \approx \frac{1}{1 + \exp(A \cdot f_j(p_k^c) + B)}
\]

(2)

Note, that this SVM-based high-level classifier fusion gave better results than simpler fusion schemes such as the product of individual posteriors in our preliminary experiments.

Regarding features and classifier architectures we opted for neural networks with local receptive field features (NN/LRF) for both the intensity and elevation domains [3, 13, 19]. Given that there are only slowly varying contrasts in both the intensity and elevation images, see Fig. 4, gradient-based approaches such as HOG [2] have limited applicability. Adaptive features such as the aforementioned local receptive fields are easily able to model such data characteristics. In contrast to multi-layer perceptrons (MLP), where the hidden layer is fully connected to the input layer, NN/LRFs introduce the concept of \( N_b \) branches \( B_i \), where every neuron in each branch only receives input from a limited local region of the input layer, its receptive field, see Fig. 5. Since synaptical weights are shared among neurons in the same branch, every branch acts as a spatial feature detector on the whole input pattern and the amount of parameters to be determined during training is reduced, alleviating susceptibility to overfitting. In our experiments, we use a NN/LRF consisting of \( N_b = 24 \) branches with \( 5 \times 5 \) pixel receptive fields shifted at a step size of two pixels over the \( 120 \times 15 \) pixel intensity and elevation images.

C. Integrated Urban Curb Recognition System

The proposed multi-cue curb classifier is easily integrated into a multitude of existing urban lane recognition systems. The curb classifier outputs can be regarded as uncertain 3D point measurements of curb locations where uncertainty is given by the curb posterior probability \( P(\omega_c | x_k) \).

One example is a commonly used Kalman filter based lane recognition system originally build for rural roads. It uses a vehicle centered clothoidal road model and utilizes gradient-based measurements specifically tuned towards lane markings, e.g. [5].

In our experiments, as described in the next section, we utilize a more sophisticated system that exhibits a significantly higher flexibility with regard to possible road course. As such, it is more applicable to the urban scenario. This system reformulates the curb recognition problem as a self-localization problem with a given digital map that explicitly contains and distinguishes between lane markings and curbs. Coarse ego-position is available through GPS and inertial sensors. The most likely curb location is obtained by the best estimate of the vehicle’s position relative to that digital map in world coordinates. Again, gradient-based measurements and a Kalman filter estimator are used. Details are given in [14].

Both approaches provide position and orientation of expected markings and curbs, respectively. Given that curbs have significantly lower image gradients than lane markings, associated gradient-based measurements on curbs tend to be very noisy which results in erroneous curb location estimates. At this point, the responses of the curb classifier can be incorporated as additional uncertain 3D curb measurements into
Fig. 6. Example output of the integrated urban curb recognition system. (a) Input image and corresponding SGM disparity image. (b) Stixel World [11] and freespace (green line), curb classifier search area (red horizontal line), the curb classifier ROI with the highest posterior probability \( P(\omega_c|x_k) \) (green box), recovered curb location (yellow line).

We use the initial curb expectation estimates as a starting point to generate regions-of-interest (ROIs), \( x_k \), for the proposed curb classifier. The curb classifier is scanned in a sliding window fashion at a step size of 5 pixels, e.g. [3], through a search area that is centered at the expected curb position and covers a horizontal range of roughly \( \pm 40 \) cm in world coordinates. Any ROIs that violate the Stixel freespace constraint, see Sec. III-A, are discarded at this point. The 3D center point of each valid ROI along with its estimated curb posterior \( P(\omega_c|x_k) \) is then used as measurement in the Kalman filter. Fig. 6 shows an example of the output of the integrated urban curb localization system.

IV. EXPERIMENTS

A. Experimental Setup

The presented approach has been tested in real-world experiments regarding curb recognition performance. To train our curb classifiers, we use 5914 curb and 6118 non-curb cut-outs (including mirrored samples) in both intensity and evaluation domains, as shown in Fig. 4. Non-curb samples are extracted from arbitrary road and sidewalk surfaces. Our test sequence consists of 4800 images captured from a vehicle-mounted stereo camera in urban traffic, see Fig. 2, with hand-labeled ground-truth curb locations.

B. Classification Performance

In a first experiment, we evaluate the performance of the curb classifiers in isolation, that is not in the context of curb localization but in an ordinary two-class classification setting. To do so, we split our cut-out training data (see above) into two non-overlapping sets and use two-thirds for training and one-third to test the classifiers. Results are given in terms of ROC performance. Fig. 7 shows that the intensity-only curb classifier outperforms the elevation-only curb classifier at higher detection rate levels and vice versa at lower false positive rates. The proposed SVM-based fusion scheme significantly outperforms each individual classifier variant.

C. Localization Accuracy

We now turn to an evaluation of curb localization performance of four different system variants at a fixed distance of 15 m from the camera. Our baseline system is the map-based curb localization system without any texture-based classification, as presented in Sec. III-C.

The second system variant under consideration is the so called integrated system which additionally incorporates multi-cue curb classifier measurements.

A third variant, referred to as curb classifier only, does not use any initial map/gradient-based curb estimate or recursive Kalman filtering at all. Instead, the multi-cue curb classifier is applied on a frame-by-frame basis in a sliding window fashion to the full image scanline (left to right) that corresponds to a distance of 15 m from the camera. The resulting multitude of classifier responses and their confidence form a probability density which is estimated using a non-parametric Gaussian kernel density estimator. The most probable curb locations are obtained as the modes of this density which are computed using Mean-Shift [1].

A fourth system variant further applies a threshold on the estimated curb posterior probability \( P(\omega_c|x_k) \) for each classification window before estimating the density and most likely curb location. This focuses the curb localization on
high-confidence classifier responses at the expense of a reduced system availability; there is no curb location estimate for image frames without any high-confidence classifier outputs above the posterior threshold. The threshold is chosen from the ROC evaluation in Sec. IV-B at a single-window detection rate of 90%.

For each system variant, we determine the (horizontal) localization error in pixels between the hand-labeled ground-truth curb location and the estimated position in the image. Results are given in terms of localization error histograms and the mean absolute localization error, see Fig. 8 and Table I.

Our results show that the integration of the proposed multi-cue curb classifier into the baseline system reduces the mean absolute localization error by roughly 25% (Table I) to 9.15 pixels, which corresponds to approximately 11 cm in world coordinates, given our camera setup. In addition, the distribution of the localization error shows a more prominent peak at lower error values, c.f. Fig. 8a vs. Fig. 8b. The system variants utilizing the curb classifier in isolation are surprisingly competitive as well, at the drawback of significantly higher computational costs, see Sec. IV-E. The system which additionally incorporates a posterior threshold on the curb classifier responses reaches the lowest mean error, that is however paid for with a reduced system availability. This may disqualify this system variant for those real-world applications that require maximum system availability, e.g. autonomous driving. We take these results as strong indication for the large discriminative power of our proposed multi-cue texture-based curb classifiers.

D. Dependency on Curb-to-Camera Distance

In the previous experiments, we applied the proposed curb classifier to regions-of-interest at a constant distance of 15 m to the camera. Many applications would certainly benefit from curb recognition at a larger distance to gain additional confidence and reaction time. Hence, we evaluate the dependency of the mean localization error from curb-to-camera distance.

To do so, the curb classifier-only system is utilized in this experiment to disregard any possible adverse effects of the initial gradient- or map-based curb estimate on location accuracy. In particular, we adopt the same ROI-generation method as given in Sec. III-C but use hand-labeled ground-truth as initial curb estimate - hence the slightly better results at 15 m distance, c.f. Fig. 9 and Table I. This scheme

<table>
<thead>
<tr>
<th>System</th>
<th>Mean Loc. Error</th>
<th>Availability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline without curb class.</td>
<td>11.71 px</td>
<td>100 %</td>
</tr>
<tr>
<td>Integrated system with curb class.</td>
<td>9.15 px</td>
<td>100 %</td>
</tr>
<tr>
<td>Curb class. only (mean-shift)</td>
<td>13.05 px</td>
<td>100 %</td>
</tr>
<tr>
<td>Curb class. only (mean-shift + threshold)</td>
<td>6.56 px</td>
<td>91 %</td>
</tr>
</tbody>
</table>

**TABLE I**

Mean Curb Localization Error for Different System Variants in Pixels.

---

Fig. 8. Localization error distribution for different system variants. (a) Baseline without curb classification. (b) Integrated system with curb classification. (c) Curb classifier only (mean-shift based localization). (d) Curb classifier only (posterior threshold and mean-shift based localization).
is applied at various distances between 13 m and 25 m. Note, that beyond the upper bound of 25 m the stereo reconstruction quality of (low) curbs decreases significantly in our camera setup.

Fig. 9 shows that the mean localization error increases almost monotonously with larger curb-to-camera distance. On absolute terms, the localization error at the largest distance of 25 m is only doubled compared to the closest distance of 13 m. This indicates, that the curb classifier is able to learn a mostly distance-independent model and can cope with different stereo reconstruction accuracies.

E. Computational Costs

Table II examines the computational complexity of the approaches under consideration in terms of computation time per frame on off-the-shelf PC hardware (Intel Core i7-3770K CPU at 3.5 GHz). We disregard the computation of SGM stereo at this point, given that it can be computed on real-time FPGA hardware [4].

The integration of the proposed multi-cue curb classifier into the map-based self-localization baseline system significantly increases the computation time per frame by approximately 15 ms. However, with a total computation time of 17.55 ms per frame, the integrated system is clearly real-time capable. The classifier-only system variants, where the curb classifier is scanned horizontally through the full image (at a fixed distance to the camera) in a sliding window fashion consume significantly more processing time at 105 ms per frame which disqualifies those approaches for real-time use, at this time.

V. Conclusion

In this paper, we proposed a novel approach towards curb recognition in urban environments that uses powerful multi-cue discriminative curb classifiers as foundation. Our curb classifiers operate on intensity and elevation images from in-vehicle stereo vision. In extensive experiments, the benefit of the proposed method has been quantified in terms of curb localization accuracy and computational costs: a 25% reduction of the average curb localization error at real-time processing speeds has been demonstrated.

Of all considered approaches, the system variant that integrates curb classification with map-driven self-localization (integrated system with curb classification, see Sec. IV-C) represents the best compromise regarding localization performance, system availability and computational efficiency.

REFERENCES


<table>
<thead>
<tr>
<th>System</th>
<th>Time per Frame</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline without curb class.</td>
<td>2.46 ms</td>
</tr>
<tr>
<td>Integrated system with curb class.</td>
<td>17.55 ms</td>
</tr>
<tr>
<td>Curb class. only (mean-shift)</td>
<td>105.29 ms</td>
</tr>
<tr>
<td>Curb class. only (mean-shift + threshold)</td>
<td>105.29 ms</td>
</tr>
</tbody>
</table>

TABLE II

COMPUTATIONAL COSTS OF DIFFERENT SYSTEM VARIANTS.