Journal of Zhejiang University-SCIENCE C (Computers & Electronics)
ISSN 1869-1951 (Print); ISSN 1869-196X (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Road model prediction based unstructured road detection

Wen-hui ZUO[1], Tuo-zhong YAO[†‡2]

([1]*Department of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China*)

([2]*Ningbo Institute of Materials Technology & Engineering, Chinese Academy of Sciences, Ningbo 315201, China*)

[†]E-mail: thomasyao@zju.edu.cn

Received Apr. 14, 2013; Revision accepted July 15, 2013; Crosschecked Oct. 15, 2013

Abstract: Vision-based road detection is an important research topic in different areas of computer vision such as the autonomous navigation of mobile robots. In outdoor unstructured environments such as villages and deserts, the roads are usually not well-paved and have variant colors or texture distributions. Traditional region- or edge-based approaches, however, are effective only in specific environments, and most of them have weak adaptability to varying road types and appearances. In this paper we describe a novel top-down based hybrid algorithm which properly combines both region and edge cues from the images. The main difference between our proposed algorithm and previous ones is that, before road detection, an off-line scene classifier is efficiently learned by both low- and high-level image cues to predict the unstructured road model. This scene classification can be considered a decision process which guides the selection of the optimal solution from region- or edge-based approaches to detect the road. Moreover, a temporal smoothing mechanism is incorporated, which further makes both model prediction and region classification more stable. Experimental results demonstrate that compared with traditional region- and edge-based algorithms, our algorithm is more robust in detecting the road areas with diverse road types and varying appearances in unstructured conditions.

Key words: Road detection, Surface layout, Road model prediction, Temporal smoothing
doi:10.1631/jzus.C1300090 **Document code:** A **CLC number:** TP317.4; TP391

1 Introduction

As a development of robot techniques, mobile robots have been widely used in both civil and military domains. The objective of road detection is to accurately separate the road from its surroundings; this plays an important role in the autonomous driving and collision warning of robots. Detecting roads using a monocular vision system is a cheap option but also a difficult problem in outdoor environments as the detection algorithm must be able to deal with continuously changing backgrounds, different environments (urban, highways, and off-road), different road types (shape, color, and texture), and different imaging conditions (varying illumination, different viewpoints, and changing weather).

Research on structured road detection started a long time ago and many successful results have been obtained. Structured road detection can be considered a 'solved' problem (Bertozzi and Broggi, 1998; Lutzeler and Dick, 1998; Thorpe *et al*., 2003). However, separating the road from its surroundings in unstructured environments (Fig. 1) is a more challenging task because: (1) Road boundaries are ill-defined in cases where strong non-boundary longitudinal edges are present due to ditches, ruts, or the tire marks of other robots, (2) Illumination and appearance such as color and texture may change considerably from one area to another, and (3) Unlike human driving where the mobile robot is always on the road, in autonomous driving the assumption that the 'drivable' area is straight ahead does not always hold. To promote the development of road detection techniques, the Defense Advanced Research Projects Agency (DARPA) in the USA proposed the 'Battlefield 2000' Program, which included an Unmanned Ground Vehicle (UGV) Program in 1992 (Bellutta *et al*., 2000), a Perception for Off-road Robotics (PerceptOR) Program in 2000 (Kelly *et al*., 2006), and a

‡ Corresponding author

Grand Challenge Race in 2005 (Thrun *et al*., 2006). These programs focused on understanding complex unknown scenes, and finding the 'drivable' area in unstructured environments was one of the most important research fields.

Fig. 1 Diverse unstructured roads with varying appearances and road types

In this paper, we propose a new hybrid solution to detect the unstructured road. In our algorithm, high-level image semantics are first explored to determine the road type. Then the optimal road detection strategy from the candidates based on low-level image semantics can be selected automatically. Temporal cues are also used to make our algorithm more robust.

2 Related work

In the last decade, many different road detection approaches have been proposed and they can roughly be classified as bottom-up ones and top-down ones. Most of the bottom-up approaches are region-based and they usually track the road by grouping pixels or superpixels together with similar color features to form 'road' and 'non-road' areas, respectively. Researchers have used different kinds of classic parameterized learning (such as Gaussian mixture modeling (GMM) (Lookingbill *et al*., 2007)) or non-parameterized learning (such as fuzzy support vector machine (FSVM) (Zhou and Iagnemma, 2010) and Markov random field (MRF) (Guo *et al*., 2010)) algorithms to realize color clustering in different color spaces (e.g., RGB, HSV, or La^*b^*). However, the existing color-based approaches assume that there is a large difference in color between the road and its surroundings. If this assumption is unsatisfied in unstructured conditions, they will probably work poorly. Therefore, some researchers have tried to use other salient features to improve region-based road detection. Alon *et al*. (2006) and Zhou *et al*. (2010) obtained good results by analyzing a variety of texture characteristics in unstructured and off-road environments. Alvarez *et al*. (2008) applied photometric invariant images to road detection, which can provide some robustness to lighting conditions. Alvarez *et al*. (2010a) combined color, contextual, and temporal cues in a Bayesian framework for crowded scenarios. This approach works well in urban structured roads and implies that reasonable multi-feature fusion could improve overall performance in difficult road conditions.

The top-down approaches usually use edge cues or prior-probability models to reinforce the widely used approaches based upon the geometric constraint of the road. Most of the edge-based approaches assume that the road types are always straight ahead, so the two road boundaries can be defined by two parallel lines. Alon *et al*. (2006) extracted road boundaries based on the texture responses in the 'bird view' image during off-road navigation. Zhang *et al*. (2009) and Kong *et al*. (2009) also marked straight boundaries to define the drivable area by vanishing point estimation, which can effectively decrease the false negative rates caused by shadows or puddles on the road surface. However, none of them can fit curved roads or intersections well. To partly improve the weakness of these algorithms, He *et al*. (2004) proposed a novel approach that adaptively selects the optimal road boundaries through a number of different curvature models which fit curved roads more accurately. Another alternative technique was proposed by Wang *et al*. (2004; 2008). They divided one image into different strips horizontally and in each strip the corresponding vanishing point was estimated based on the image gradients. These vanishing points were then used to fit the parameters of the hyperbolic or B-spline models which can describe curved roads better. All of these approaches are based on the hard boundary constraint and usually work well in urban curved roads with well marked lanes or boundaries. However, they are inevitably less effective on unstructured roads with unclear road boundary description.

There are only a very few approaches that use a prior-probability model. They usually provide

several statistical probabilistic road masks to help a region-based algorithm detect road areas (Lombardi *et al*., 2005; Alvarez *et al*., 2009; Zhou and Iagnemma, 2010). Most of these models are manually pre-defined or learned, which can provide some flexibility to different road types and insensitivity to changes of appearance.

From what has been discussed above, it is clear that most of the existing region- or edge-based approaches work well only in specific environments. Region-based approaches lack the geometric model, while edge-based approaches focus only on boundary detection and ignore the region-based nature of the path. Thus, few of them have good adaptability to varying road types and appearances. If the region and edge cues could be integrated in a more efficient manner, the performance of existing road detection techniques could be improved significantly.

3 Framework of the proposed algorithm

The details of our algorithm are illustrated in Fig. 2. In this flowchart, a region-based method (Method 1) and an edge-based method (Method 2) are both applied. Method 1 utilizes on-line MRF to spatially cluster the pixels into 'road' and 'non-road' areas based on multiple salient features. It can be used for detecting curved roads and intersections for which edge-based methods usually do not work well (light-blue box). Method 2 integrates texture and geometric cues to obtain reliable road boundaries and is more appropriate for describing straight roads (deep-blue box). If the road type is known, we can choose the appropriate method to track the road. The key problem is to predict which type the current road belongs to.

The main feature of our approach is that, before road detection, an additional decision process is first run to predict the road type (pink box). However, road boundaries are usually not remarkable in an unstructured environment; when using only edge cues from images, it is difficult to decide the road type. One possible solution, proposed by Lombardi *et al*. (2005), fits a road model among three possible ones. The disadvantage of this approach is that a region-based road detection algorithm is required before the fitting process and the proposed road models are not diverse enough to fit all roads with different shapes. Inspired by Alvarez *et al*. (2009), we model and predict all the road types in the real world by the well-known scene classification technique (dotted box). This decision process can be very helpful because it provides optimal correspondences between the road type and the road detection approach, thereby increasing flexibility to diverse road types. Moreover, a temporal mechanism is added in both Method 1 and the road model decision process to improve the overall road detection performance (green box).

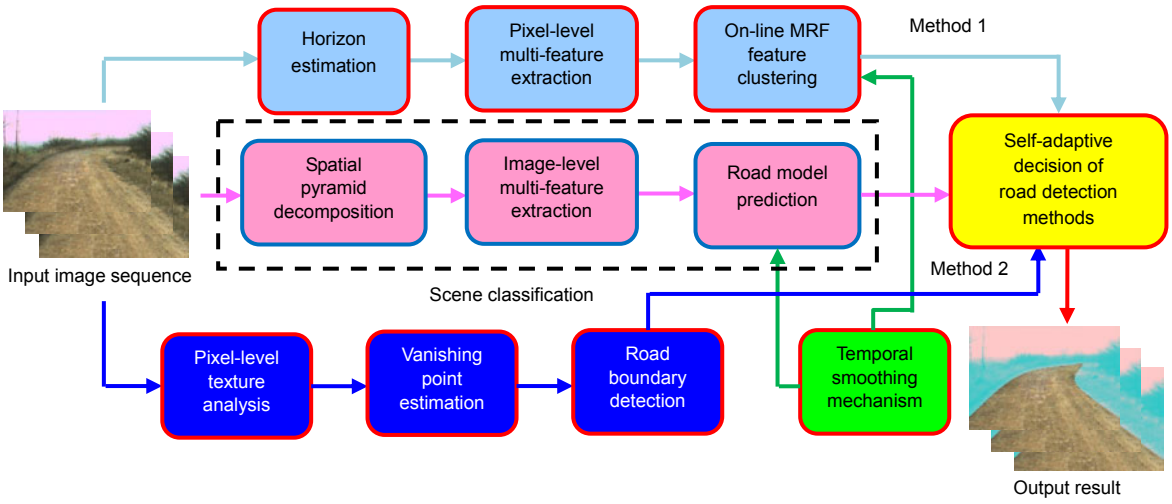

Fig. 2 Flowchart of the proposed road detection algorithm

For interpretation of the references to color in this figure legend, the reader is referred to the online version of this paper

4 Road model prediction

4.1 Road model

We decide the road type by a road model prediction which is similar to that proposed by Alvarez *et al*. (2009). However, unlike that approach, we do not use pre-learned road probability maps which are robust only in flat structured roads and ignore several unusual prior models designed for crowded urban scenes only. Instead, we build five general road models to roughly define all kinds of road types in the real world.

In Fig. 3, all the straight roads are denoted by Model 3, all the left/right curved roads are denoted by Model 1/Model 2, and all the left/right intersections are denoted by Model 4/Model 5. These road models will be predicted by a scene classification technique, which captures the complex statistics of natural images to train a classifier from examples of scene classes.

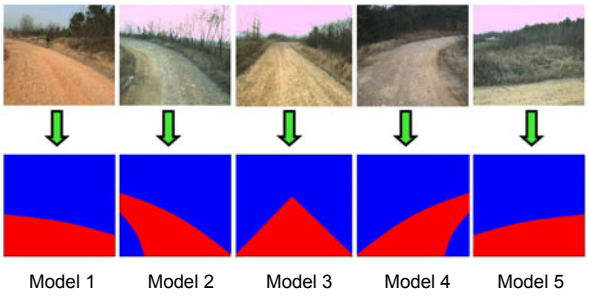

Fig. 3 General road models

4.2 Image representation

How to choose effective features for image representation is critical to the performance of road model prediction. In this study, both low- and high-level image cues are used to more effectively reflect the differences between the road and its surroundings.

4.2.1 Low-level cue

Similar to the scale-invariant feature transform (SIFT) descriptor (van de Sande *et al*., 2010), the histograms of oriented gradient (HOG) descriptor (Dalal and Triggs, 2005) is invariant to image scale and rotation and robust to changes in illumination, noise, and minor changes in viewpoint. Moreover, it is more computationally efficient than SIFT. In this study, we sample the R-HOG descriptor densely in the opponent color space, which is also invariant to illumination as a low-level cue. The sampling strategy is shown in Fig. 4.

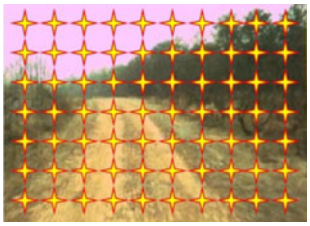

Fig. 4 Dense histograms of oriented gradient (HOG) sampling

Each sampled HOG descriptor is computed in dense grids at a single scale without dominant orientation alignment and used as part of a larger code vector that implicitly encodes spatial position relative to the detection window.

4.2.2 High-level cue

The high-level cue we use is a 3D surface layout proposed by Hoiem *et al*. (2007). It estimates the coarse geometric properties of a scene by learning appearance-based models of geometric classes: 'ground', 'vertical' (the regions perpendicular to the ground), and 'Sky' in cluttered natural scenes. Geometric classes can provide rough semantic annotation and 3D structure of an image region and can become an important hint for finding the drivable area. Alvarez *et al*. (2010a; 2012) were the first to use this high-level cue in road detection, which proved effective. In this study, we use a similar technique to obtain the road confidences of unstructured environments by estimating the surface layout of the scenes. The detailed procedure is outlined in Fig. 5.

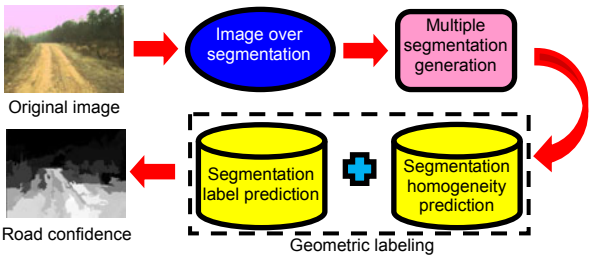

Fig. 5 Flowchart for estimation of the 3D surface layout

First, we use a graph-based algorithm (Felzenszwalb and Huttenlocher, 2004) in the over-

segmentation process. The advantage of this technique is that it can group large homogeneous regions of the image together while dividing heterogeneous regions into many smaller superpixels. We set the segmentation threshold k=150 and the minimum number of pixels in each superpixel min_size= 200. Then, pairs of same-label and different label superpixels (250 each) are sampled from the training set and a total of 26 salient features such as color, texture, shape, geometry, and position are extracted from each superpixel. We estimate the likelihood that two superpixels have the same label based on the absolute differences of their feature values to be $f_{ij}=\sum_k^M \left|f_{ik}-f_{jk}\right|$ and train a logistic regression version of Adaboost (logistic Adaboost) to predict the pairwise same-label likelihood P_{ij}. Logistic Adaboost weighted the output of a number of C4.5 decision trees. In this study, we set the number of decision trees n_t=20 and each decision tree has n_n=20 leaf nodes to prevent over-fitting during training.

Based on pairwise same-label likelihoods, multiple segmentation hypotheses are generated by a simple greedy algorithm that groups superpixels into larger continuous segments. The advantage of multiple segmentations is that they can acquire the spatial support necessary for complex cues while avoiding the risky commitment to a single segmentation. In this process, a diverse sampling of segmentations is produced by varying the number of segments n_s and using a random initialization. In our implementation, we generate five segmentations of the input image, with $n_s \in \{5, 10, 15, 20, 25\}$ (Fig. 6).

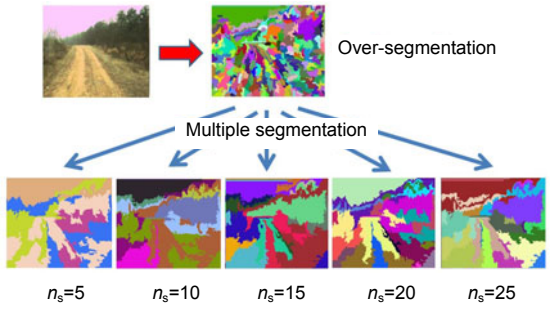

Fig. 6 Generation of multiple segmentation
Regions with different colors denote different segments

Finally, superpixel label confidence $P(y_i|I)$ can be determined by averaging the label likelihoods $P(\tilde{y}_j \mid s_j, I)$ of the segments that contain it, weighted by the homogeneity likelihoods $P(s_{ij}|I)$:

$$P(y_i = k \mid I) \propto \sum_{i \in s_j} P(s_{ij} \mid I) P(\tilde{y}_j = k \mid s_{ij}, I), \quad (1)$$

where y_i is the superpixel label, k={0, 1} is a possible label value ('1' denotes the 'road' class and '0' denotes the 'non-road' class), I is the image data, s_{ij} denotes the segment that contains the ith superpixel for the jth segmentation hypothesis, and $\tilde{y}_j$ is the segment label.

Fig. 7 shows several road confidence maps of different unstructured environments. Road confidences learned by geometric labeling are quantized to 0–255 grayscales, and brighter pixels mean they are more likely to belong to the 'road' class. Although the road confidence maps usually cannot reflect the real road distributions very accurately at the micro level, in most cases they can illustrate well the true shapes of unstructured roads at the macro level and are thus useful in helping improve the prediction accuracy of the road models.

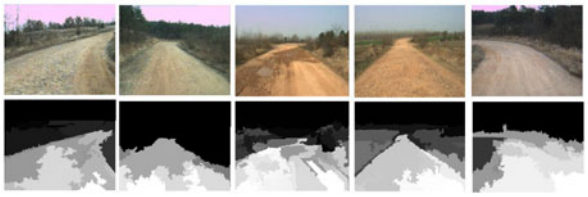

Fig. 7 Road confidence maps in different unstructured environments

4.3 Scene classification

To predict the road model, we use spatial pyramid matching (SPM) (Lazebnik *et al*., 2006) which can reflect the inherent spatial structure of road images, similar to the idea of Alvarez *et al*. (2013). The flowchart is illustrated in Fig. 8. In the SPM process, each image is firstly decomposed into three-level spatial pyramids with 1×1, 2×2, and 4×4 blocks, respectively. At each spatial level, HOG and surface layout (SL) are extracted to construct the corresponding visual histograms. Then, we align these histograms of all spatial levels in a weighted manner to obtain the pyramid histogram of visual words (PHOW):

$$H_j = (\alpha_0 \cdot H_{(L_0),j}) \oplus \left(\alpha_1 \cdot \sum_{i=1}^{4} H_{(i,L_1),j}\right) \oplus \left(\alpha_2 \cdot \sum_{i=1}^{16} H_{(i,L_2),j}\right). \quad (2)$$

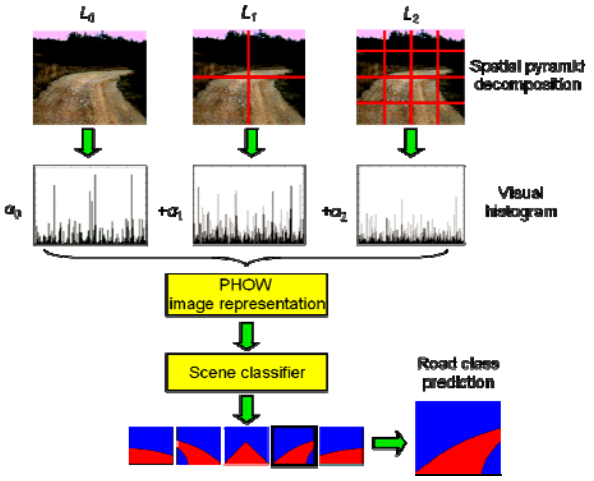

Fig. 8 Road model prediction by spatial pyramid matching (SPM)

In Eq. (2), $H_{(L_i),j}$ (i=1, 2, 3) is the visual histogram of block i at spatial level L_{i-1} by feature type j (j=1 refers to HOG and j=2 refers to SL) and symbol $\oplus$ is the alignment operation between visual histograms. For computing $H_{(L_k),1}$, all HOG descriptors in the training set are quantized to construct a codebook with the number of visual words being K=200 by K-means clustering. For computing $H_{(L_k),2}$, we divide each image into 5×5 overlapped patches and use a similar method to construct an alternative codebook with the number of visual words being K=10. Then all the visual histograms are obtained by occurrence statistics.

Different from the weight coefficient setting in Lazebnik *et al*. (2006), we allocate the smaller blocks with larger weight coefficients: α_0=0.2, α_1=0.4, α_2=0.4. Such a weight setting can better reflect shape dissimilarity between different road models. In this study, a traditional one-to-all multi-class support vector machine (SVM) classifier (Duan *et al*., 2003), in which specific classifiers are learned separately, is learned to decide the road models.

5 Road detection strategy

5.1 Straight road detection

The shape of straight roads can be simply illustrated by two intersection lines, and edge-based road detection approaches usually fit such types of roads well. Thus, we use an edge-based method proposed by Kong *et al*. (2009), which is quite robust in varying unstructured conditions if the road model is predicted as Model 3.

Kong's method consists of two steps: vanishing point estimation (VPE) and road boundary selection (RBS). In the VPE process, vanishing point candidates are computed by estimating the confidences of the texture orientations and then a local adaptive voting strategy is integrated to select the optimal vanishing point. In Fig. 9b, the pixel with the highest voting score in the confidence map is the optimal vanishing point VP*. In the RBS process, a vanishing-point-constrained edge detection technique is used for road segmentation. We construct a set of 33 evenly distributed rays starting from VP*, excluding those rays whose angle relative to horizon is smaller than 10° or larger than 160° (Fig. 9c). Then, the orientation consistency ratio (OCR) and color differences in the regions between two adjacent rays (e.g., yellow region A1 and blue region A2 in Fig. 9d) are both used to decide the two optimal road boundaries for drivable area representation. OCR is the ratio between the number of discrete sampled points (green points in Fig. 9c) on each ray whose angle is between the point's orientation (blue arrows in Fig. 9c) and the ray's orientation is lower than a threshold and the total number of points on the ray. In Fig. 9e, two blue lines are extracted to describe the drivable area by the color pink.

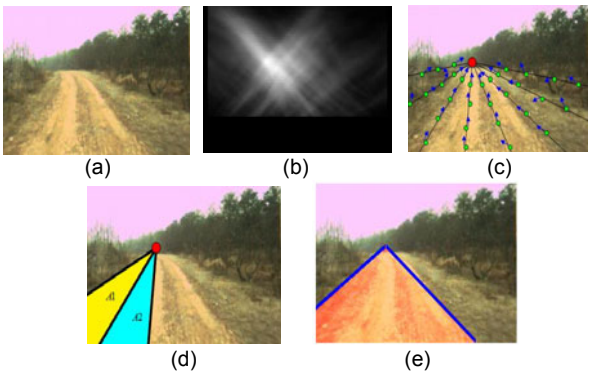

Fig. 9 Vanishing point estimation based road segmentation
(a) The original image; (b) Vanishing point candidates are estimated by the confidences of the texture orientations and the optimal vanishing point is determined by a local adaptive voting strategy; (c) A vanishing-point-constrained edge detection is used to estimate the optimal road boundaries; (d) The color differences of the regions between two adjacent rays are used to help road boundary decision; (e) The detected drivable area

5.2 Non-straight road detection

As discussed above, the edge-based approaches do not fit unstructured curved roads or intersections well. When the road models are predicted as Model 1, 2, 4, or 5, we use a region-based method to track the road instead.

Before road segmentation, a sky detection algorithm is used to avoid mistakenly judging sky regions as road regions because the sky usually has illumination similar to that of the road. In this algorithm, a green line is predefined to constrain the range search of the sky region in Fig. 10a. First, each image is decomposed into 4×4 patches and denoted by a feature vector consisting of color, texture, and position. Then, the K-means (K=2) clustering algorithm is used to separate the image into 'sky' (denoted by white pixels) and 'non-sky' (denoted by black pixels) regions roughly (Fig. 10b). The integral operation $L(y)$ is computed by

$$l^* = \arg\min_y L(y),\ \ L(y) = \int_0^W L(x,y)\mathrm{d}x. \tag{3}$$

If pixel (x, y) belongs to the 'sky' class, $L(x, y)$=1; otherwise, $L(x, y)$=0.

In Fig. 10c, the length of blue line segments along the horizontal direction means how many pixels are classified as 'sky' in the corresponding image rows, and the global minimum l^* is the estimated location of the horizon described by the red line in Fig. 10a. Finally, we define the region below the horizon as the region of interest (ROI) and the detected drivable area should be in ROI.

(a) (b) (c)

Fig. 10 Sky detection and horizon estimation

(a) A green line is predefined to constrain the range search of the sky region and the estimated horizontal line is described by the red line; (b) The 'sky' and 'non-sky' regions are both obtained by K-means clustering; (c) The length of blue line segments along the horizontal direction represents how many pixels are classified as 'sky' in the corresponding image rows

After sky removal, we use a self-supervised Markov random field (MRF) which models the spatial correlation between adjacent pixels for feature clustering. In the on-line MRF model, graph G=(V, E) is constructed, with node V consisting of pixel set X and undirected edge E defining the similarity between the nodes. The optimal annotation assignment Y^* of the pixels is estimated by minimizing the energy function $E(Y)$. This minimization can be solved by graph cut with α-expansion (Boykov *et al*., 2001):

$$\begin{cases} Y^* = \arg\min E(Y), \\ E(Y) = \sum_i E_{\text{local}}(y_i) + \sum_{i,j} E_{\text{smooth}}(y_i, y_j), \end{cases} \tag{4}$$

where the unary potential $E_{\text{local}}(y_i)$ defines the uncertainty of the pixel annotation and $\sum_i E_{\text{local}}(y_i)$ is denoted by a 13-dimensional feature vector combined with both low- and high-level features. In this study, we extract a 3D color histogram in the La*b* space as the low-level cue and a 10-dimensional visual histogram of the surface layout as the high-level cue. Pairwise potential $E_{\text{smooth}}(y_i, y_j)$ based on the Potts model is defined by Eq. (5) and it illustrates the discontinuity between adjacent pixels i and j. $f(x_i, x_j) = \sum_{k=1}^{13} \left\| f_{ik} - f_{jk} \right\|$ is the Euclidean distance of x_i and x_j in the feature space and $s(x_i, x_j)$ is the spatial distance of x_i and x_j in the image space.

$$\begin{cases} E_{\text{smooth}}(y_i, y_j) = \begin{cases} 0, & \text{if } y_i = y_j, \\ \beta \cdot \varphi(x_i, x_j), & \text{otherwise}, \end{cases} \\ \varphi(x_i, x_j) = \dfrac{1}{f(x_i, x_j)} \exp\left(\dfrac{-c(x_i, x_j)}{2\sigma^2} \right). \end{cases} \tag{5}$$

6 Temporal smoothing mechanism

During robot navigation, our proposed algorithm works on image sequences and handles input images one by one. Due to changes in the surrounding conditions and the prediction error of the scene classifier, large differences may occur in region-based feature clustering and road model prediction between the current frame and several previous neighbor frames. Thus, the temporal correlation constraint should be integrated to stabilize those results, further improving the overall performance.

In region-based feature clustering, there is an assumption that the drivable areas between adjacent frames vary progressively. In the on-line MRF model, the unary potential in Eq. (5) is denoted by a 13-dimensional feature vector from manually labeled 'road' and 'non-road' regions, respectively, for model learning in the initial frame. Starting from the second frame, the old feature vectors are gradually updated to reduce the artifacts occurring in the road segmentation process.

In road model prediction, the road models between adjacent frames should change continuously: Model 1⇔Model 2⇔Model 3⇔Model 4⇔Model 5. It means that the current state Model 1 can move only to the adjacent state Model 2 and is prohibited from jumping to the following remote states: Model 3, 4, or 5. In this study, we use the Kalman filter to smooth the model prediction results temporally. This process is effective and can further reduce the possibility of unexpected discontinuous state changes during road model prediction. Another advantage is that we define five road types (straight, curved, and intersected), not only two (straight and non-straight) is that by using more road models which change progressively one by one, the model prediction results can be improved by temporal filtering, while the results with fewer road models cannot.

7 Experiments

In the experiments, road image sequences were captured from a CCD camera equipped on top of the Pioneer 4 platform designed by ActiveMedia Co. Ltd. to evaluate our algorithm. Our robot ran at 3.5 km/h and the resolution of each captured image was 320×240 pixels. The algorithm was run on a Linux platform with Pentium IV Dual-Core 2.8 GHz CPU and 4 GB RAM.

7.1 Datasets and the evaluation criterion

We collected several image sequences in unstructured roads with diverse road types in different lightening and weather conditions (Fig. 1) to build the datasets for training and testing. In our training sets, there were 68 949 images in total, while the testing sets were sampled in a certain interval from collected image sequences which had 2026 images. The percentages of straight roads, curved roads (Models 2 and 4), and intersections (Models 1 and 5) in the complete training sets were 70.5%, 22.6%, and 6.9%, respectively.

In this study, we described five different experiments and used average segmentation precision (ASP) as the quality measure for evaluating the performance of our proposed algorithm. ASP is defined as follows:

$$\mathrm{ASP}=\frac{\mathrm{r\_tp}}{\mathrm{r\_tp}+\mathrm{r\_fp}+\mathrm{r\_fn}}\times 100\%, \qquad (6)$$

where r_tp denotes the number of pixels which are correctly classified as the 'road' class, r_fp denotes the number of pixels which belong to the 'non-road' class and are misclassified as the 'road' class, and r_fn denotes the number of pixels which belong to the 'road' class and are misclassified as the 'non-road' class.

7.2 Experiment 1: evaluation of multi-cue combination in road model prediction

In Experiment 1, we first design a simple algorithm to evaluate the effect of the 3D surface layout (SL) in road detection. This SL-based road detection algorithm can be illustrated as follows: a road label is assigned to superpixel x if the road confidence road_conf(x)>0.5; otherwise, a background label is assigned to x. The confusion matrix of the SL-based method is outlined in Fig. 11.

Fig. 11 The confusion matrix of surface layout based road detection with an average accuracy of 87.67%

In Fig. 11, we can see that the segmentation accuracies of all the three main geometric classes 'sky', 'vertical', and 'road' are higher than 0.8 and all

the false alarms are lower than 0.2. These statistics demonstrate that we have a higher probability of successfully estimating the road shapes by integrating the 3D surface layout into scene classifier learning. The more accurate the scene contextual information, the higher the ASP in road detection and the higher the prediction accuracy of the road model.

Second, we learn the road model classifiers by different feature combinations. In Fig. 12, the first column outlines representative scenes whose road models are easily mistakenly predicted. The second column shows the road types predicted by using HOG only (RMP−HOG). The third column gives the road confidence maps of the corresponding scenes, and the road types predicted by using HOG and surface layout simultaneously (RMP−(HOG+SL)) are given in the fourth column. The fifth column gives the ground truth road models.

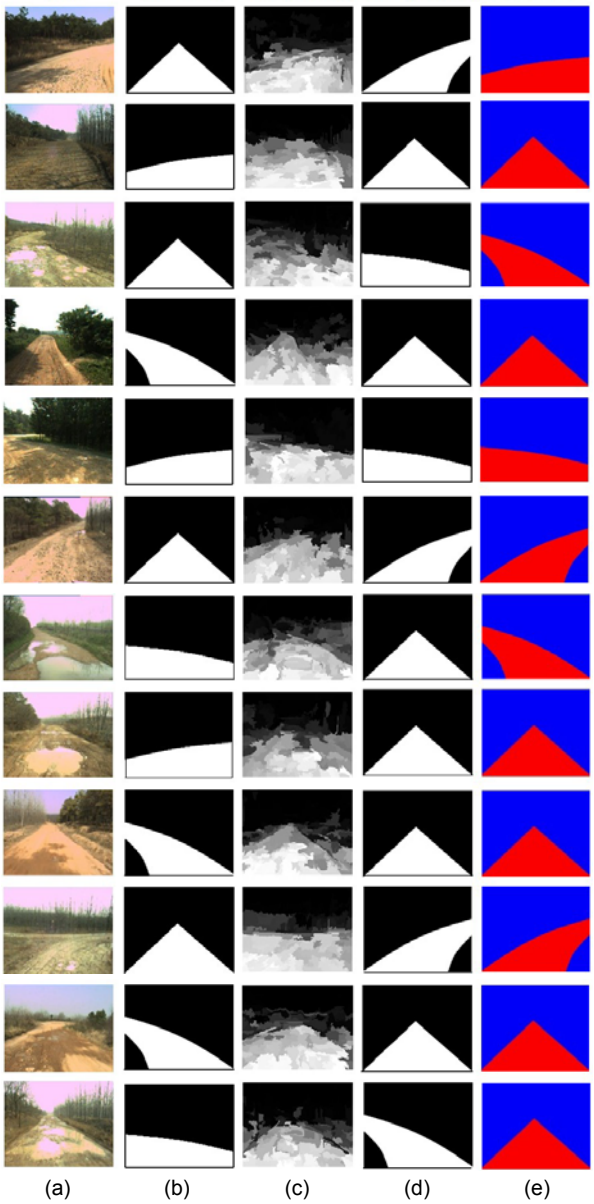

Fig. 12 Comparison between RMP−HOG and RMP−(HOG+SL): (a) Scene; (b) RMP−HOG; (c) confidence map; (d) RMP−(HOG+SL); (e) ground truth

In Fig. 12, the scene classifier outputs the uncorrected road types due to strong shadow, fuzzy road boundaries, or big puddles, with greater probability by RMP−HOG. It implies that using only low-level cues usually cannot accurately illustrate the shapes of diverse unstructured roads in complex conditions. By contrast, after combining low- and high-level cues in scene classifier learning, most of the model prediction results are corrected or get closer to the ground truth results because SL features which can provide a rough description of the road shapes substantially help improve the road model prediction.

The confusion matrices in Fig. 13a and 13b illustrate the road model prediction in two different conditions: without and with high-level cues (temporal smoothing is not used).

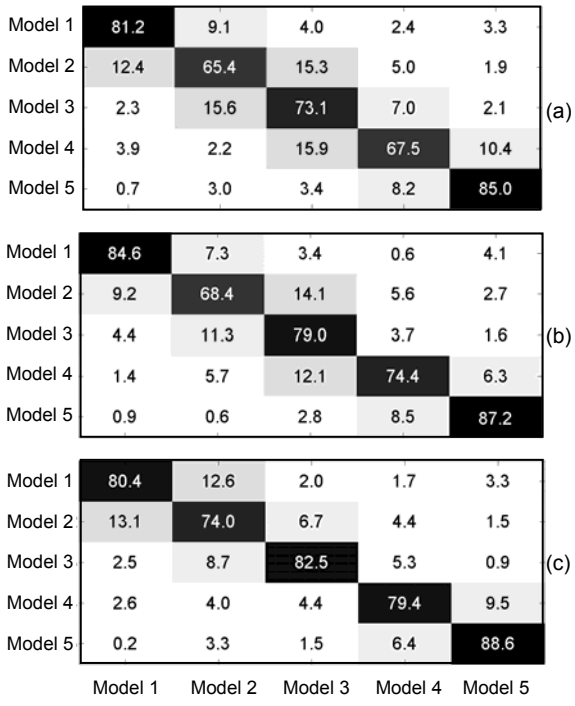

Fig. 13 Confusion matrices of different road detection strategies: (a) RMP−HOG (average accuracy is 74.4%); (b) RMP−(HOG+SL) (average accuracy is 78.7%); (c) RMP−(HOG+SL)+KF (average accuracy is 80.9%)

Through the comparison between RMP−HOG and RMP−(HOG+SL), it can be easily found that the multi-cue combination strategy significantly improves the prediction accuracy of all the five models and the average accuracy of RMP−(HOG+SL) is higher than that of RMP−HOG by 4.3%.

7.3 Experiment 2: effect of temporal smoothing in road model prediction

In Experiment 2, we compared the performance of road model prediction with and without temporal smoothing mechanism. Fig. 14a outlines a representative example: a group of 30 successive frames are sampled from the dataset which describes the progressive changes of road models as follows: Model 3⇒Model 2⇒Model 1⇒Model 2⇒Model 3. We map tags 1, 2, 3, 4, and 5 in the *Y*-axis to Models 1, 2, 3, 4, and 5, respectively.

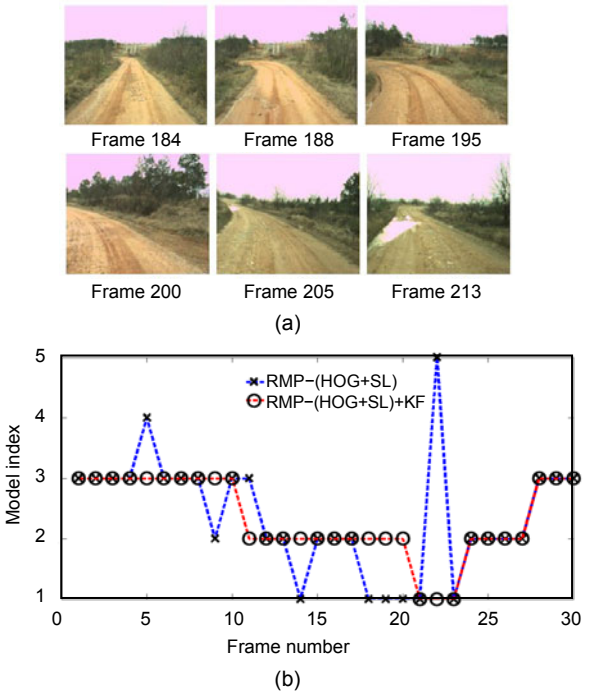

Fig. 14 Temporal smoothing for road model prediction
(a) A group of 30 successive frames are sampled for test; (b) The comparison between the two algorithms

In Fig. 14b, two curves with different colors illustrate the road model prediction by the following two approaches: road model prediction without temporal mechanism (RMP−(HOG+SL)) and road model prediction with Kalman filter (RMP−(HOG+SL)+KF). Discontinuous state changes between adjacent frames by RMP−(HOG+SL) occur because the inter-class similarities between road scenes generate the prediction error of the scene classifier. After using the temporal smoothing mechanism, the prediction results by RMP−(HOG+SL)+KF get closer to the ground truth.

The confusion matrix in Fig. 13c illustrates the road model prediction with KF. It is clear that the prediction accuracy of Models 2, 3, and 4 is improved significantly, while false alarms are decreased. It demonstrates that the temporal smoothing mechanism can effectively decrease the prediction error of the scene classifier. Although some road classes are incorrectly predicted as adjacent ones and the prediction precision of Model 1 decreases because of over-smoothing of KF, it occurs only in a few non-straight road scenes, which does not significantly impact the average prediction accuracy.

7.4 Experiment 3: evaluation of the edge-based approach

In Experiment 3, the edge-based approach is evaluated in the testing datasets where the road types of the images are predicted as Model 1. In Fig. 15, the predicted horizontal lines are denoted by the yellow lines in the first row, the second row shows the confidence maps of the vanishing point candidates, the pink regions in the third row are the edge-based road segmentation results, and the red regions in the fourth row are manually labeled ground truth road regions.

In Fig. 15a, in most cases Kong's algorithm gives good results, even though there are large areas of puddles which make the road boundaries fuzzy. The road segmentation results in Fig. 15b also demonstrate that Kong's algorithm is quite robust in difficult conditions where ponding and shadows frequently appear on the road. These prove that this method works well for straight roads.

7.5 Experiment 4: evaluation of the region-based approach

In Experiment 4, we decompose the region-based algorithm into the following three sub-algorithms for evaluating the effects of different smoothing mechanisms in road detection: (1) pixel-level clustering only, without any smoothing mechanism (PC); (2) pixel-level clustering, by using spatial smoothing (PC+SS); (3) pixel-level clustering, by using both spatial smoothing and temporal smoothing (PC+SS+TS). Fig. 16 shows several

representative curved road scenes and the corresponding road detection results (indicated by red regions) by these three sub-algorithms.

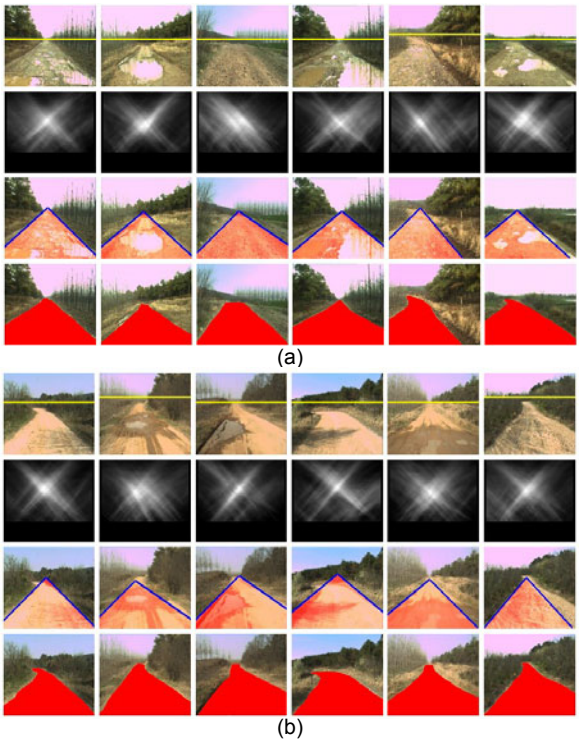

Fig. 15 Horizontal line prediction, vanishing point estimation and road detection results in different straight road datasets by the edge-based approach
(a) The road scenes where there are large areas of puddles which make the road boundaries fuzzy; (b) The road scenes where ponding and shadows frequently appear on the road

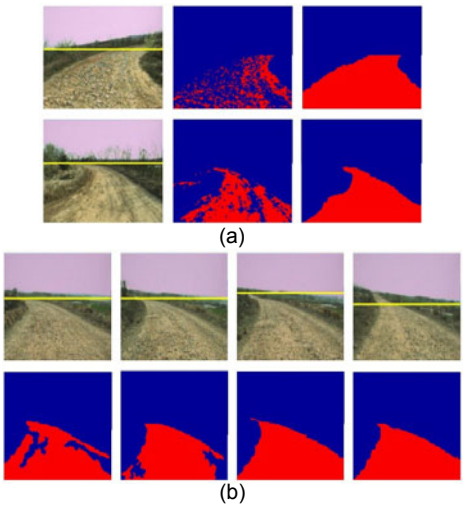

Fig. 16 Region-based road detection results obtained using different smoothing mechanisms
(a) The comparison between pixel-level clustering (PC) only and pixel-level clustering by spatial smoothing (PC+SS); (b) The temporal change of the road detection result by PC+SS+TS

In Fig. 16a, some road areas in the first column which are quite similar to the surroundings are misclassified as surroundings by PC in the second column because stones and weeds on the road often make the road boundaries unclear. By comparison, the results of PC+SS in the third column seem more accurate because they efficiently integrate the spatial relationship between the pixels with different labels. In Fig. 16b, the first row lists several successive frames which illustrate a left turning process of the robot. The second row in Fig. 16b provides the corresponding results by PC+SS+TS and demonstrates that the road detection results are progressively refined by temporal smoothing.

7.6 Experiment 5: comparison between the hybrid approach and single approaches

In Experiment 5, the following three different road detection algorithms are compared in all testing datasets: (1) edge-based vanishing point estimation algorithm (VPE); (2) region-based pixel-level clustering algorithm (PLC); (3) hybrid algorithm guided by the road model classifier (RMC). Fig. 17 shows the results.

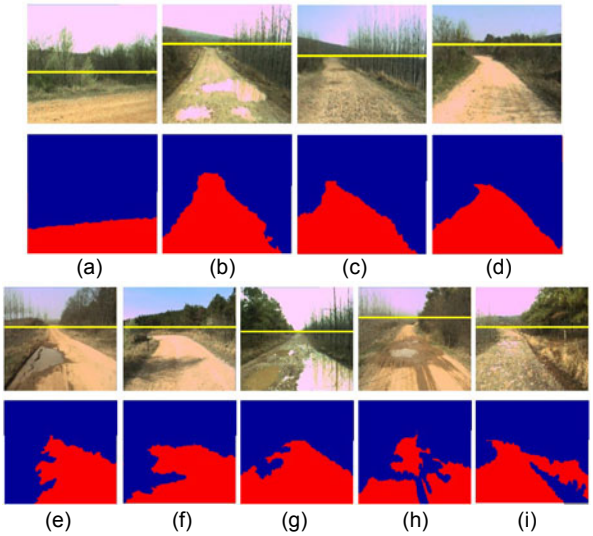

Fig. 17 Road detection results in different unstructured scenes by PLC
The first row of (a)–(i) shows different unstructured road scenes and the second row (a)–(i) gives the corresponding results by PLC

When a large area of puddle (Fig. 17a) or weak shadow (Fig. 17b) appears on roads, PLC can usually obtain good results. However, in many cases, its

performance is far from robust; e.g., in Fig. 17f, the strong shadow increases the detection error severely; In Fig. 17h, a big puddle makes the drivable area heterogeneous and separates the road into three discrete small regions. In Fig. 17i, the appearance of the grass area near the road is quite similar to that of the road itself, and most of the grass is misclassified as 'road'. Comparison of Figs. 15 and 17 shows that the VPE, which is insensitive to varying conditions, can fit the straight roads much better than PLC.

Fig. 18 outlines some challenging video sequences in the testing dataset in different weather (sunny and rainy), illumination (light and dark), and road type (shape, color, and texture) conditions. The road model prediction results are tagged in the second rows. The road detection results by VPE and PLC are shown in the third and fourth rows, respectively, and the results by RMC are in red rectangles. The fifth and sixth rows are ground truth road models and road detection results, respectively. Comparison of Figs. 18a–18c shows that, by using road model prediction through learning both low- and high-level cues, the optimal strategy can be chosen with a high probability, making our algorithm more robust.

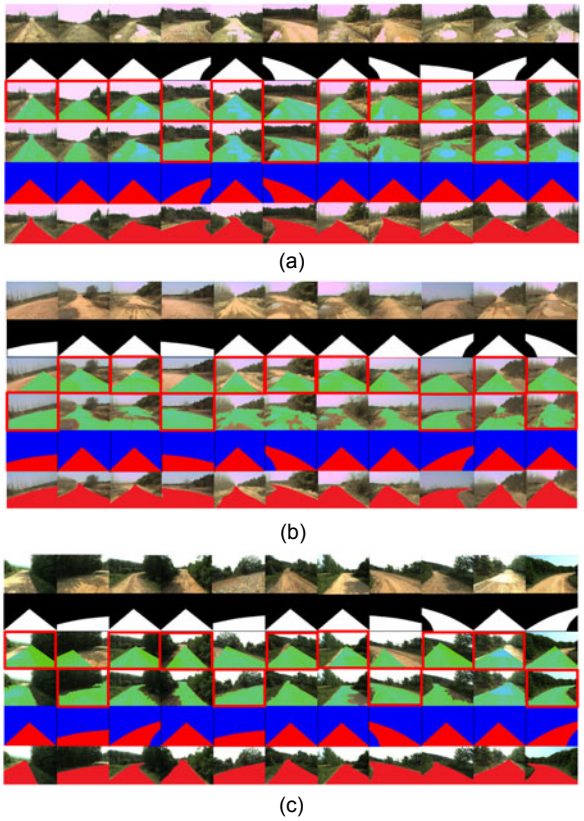

Fig. 18 Road detection results obtained using the proposed algorithm: (a) sequence 1; (b) sequence 2; (c) sequence 3

Fig. 19 shows the ASP statistics of VPE, PLC, and RMC in straight, non-straight, and hybrid roads respectively in the testing dataset. VPE performs better in straight road scenes than PLC; the ASP of VPE is higher than that of PLC by 14.63%.

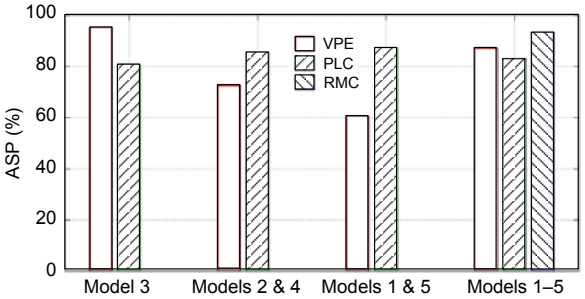

Fig. 19 Average segmentation precision (ASP) statistics of three algorithms VPE, PLC, and RMC

8 Conclusions

In this paper, we propose a novel top-down based approach, which adaptively selects the proper region- or edge-based approach in each frame to detect diverse unstructured roads. In our algorithm, a scene classifier is pre-learned to tell the robot the current road type, and thus indicate the optimal road detection strategy. A temporal smoothing mechanism is integrated to further stabilize the overall performance. Experiments show that our proposed algorithm has more flexibility and insensitivity to diverse road types and varying appearances than traditional region- or edge-based approaches.

Our algorithm, however, has high computation complexity, which makes it unsuitable for high-speed robot navigation. In future work, we intend to do some optimization based on the following ideas: First, the search space of candidate vanishing points in the edge-based method will be more properly constrained (Moghadam *et al*., 2012) to speed up this time-consuming process. Second, the process of superpixel merging and labeling in estimating the 3D scene layout could be computed in parallel. Furthermore, geographic information from aerial remote sensing images (Alvarez *et al*., 2010b) could be integrated. It can be pre-captured by unmanned aerial vehicle (UAV) or obtained by Google Map to provide

prior information of road shapes, which will be helpful in road model prediction.

References

Alon, Y., Ferencz, A., Shashua, A., 2006. Off-Road Path Following Using Region Classification and Geometric Projection Constraints. IEEE Conf. on Computer Vision and Pattern Recognition, p.689-696.

Alvarez, J.M., Lopez, A.M., Baldrich, R., 2008. Illuminant Invariant Model-Based Road Segmentation. Proc. IEEE Intelligent Vehicles Symp., p.1175-1180.

Alvarez, J.M., Gevers, T., Lopez, A.M., 2009. Vision-Based Road Detection Using Road Model. 16th IEEE Int. Conf. on Image Processing, p.2073-2076.

Alvarez, J.M., Gevers, T., Lopez, A.M., 2010a. 3D Scene Priors for Road Detection. IEEE Conf. on Computer Vision and Patten Recognition, p.57-64.

Alvarez, J.M., Lumbreras, F., Gevers, T., Lopez, A.M., 2010b. Geographic Information for Vision-Based Road Detection. IEEE Intelligent Vehicles Symp., p.621-626.

Alvarez, J.M., Gevers, T., LeCun, Y., Lopez, A.M., 2012. Road Scene Segmentation from a Single Image. European Conf. on Computer Vision, p.376-389.

Alvarez, J.M., Gevers, T., Diego, F., Lopez, A.M., 2013. Road geometry classification by adaptive shape models. *IEEE Trans. Intell. Transp. Syst.*, **14**(1):459-468. [doi:10.1109/TITS.2012.2221088]

Bellutta, P., Manduchi, R., Matthies, L., Owens, K., Rankin, A., 2000. Terrain Perception for DEMO III. Proc. IEEE Intelligent Vehicles Symp., p.326-331.

Bertozzi, M., Broggi, A., 1998. GOLD: a parallel real-time stereo vision system for generic obstacle and lane detection. *IEEE Trans. Image Process.*, **7**(1):62-81. [doi:10.1109/83.650851]

Boykov, Y., Veksler, O., Zabih, R., 2001. Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Anal. Mach. Intell.*, **23**(11):1222-1239. [doi:10.1109/34.969114]

Dalal, N., Triggs, B., 2005. Histogram of Oriented Gradient for Human Detection. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, p.886-893.

Duan, K.B., Keerthi, S.S., Chu, W., Shevade, A.K., Poo, A.N., 2003. Multi-category Classification by Soft-Max Combination of Binary Classifiers. 4th Int. Workshop on Multiple Classifier Systems, p.125-134. [doi:10.1007/3-540-44938-8_13]

Felzenszwalb, P., Huttenlocher, D., 2004. Efficient graph-based image segmentation. *Int. J. Comput. Vis.*, **59**(2):167-181. [doi:10.1023/B:VISI.0000022288.19776.77]

Guo, C.Z., Mita, S., McAllester, D., 2010. MRF-Based Road Detection with Unsupervised Learning for Autonomous Driving in Changing Environments. IEEE Intelligent Vehicles Symp., p.361-368.

He, Y.H., Wang, H., Zhang, B., 2004. Color based road detection in urban traffic scenes. *IEEE Trans. Intell. Transp. Syst.*, **5**(4):309-318. [doi:10.1109/TITS.2004.838221]

Hoiem, D., Efros, A.A., Hebert, M., 2007. Recovering surface layout from an image. *Int. J. Comput. Vis.*, **75**(1):151-172. [doi:10.1007/s11263-006-0031-y]

Kelly, A., Stentz, A., Amidi, O., Bode, M., Bradley, D., Diaz-Calderon, A., Happold, M., Herman, H., Mandelbaum, R., Pilarski, T., *et al*., 2006. Toward reliable off road autonomous vehicles operating in challenging environments. *Int. J. Robot. Res.*, **25**(5-6):449-483. [doi:10.1177/0278364906065543]

Kong, H., Audibert, J.Y., Ponce, J., 2009. Vanishing Point Detection for Road Detection. IEEE Conf. on Computer Vision and Pattern Recognition, p.96-103.

Lazebnik, S., Schmid, C., Ponce, J., 2006. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, p.2169-2178.

Lombardi, P., Zanin, M., Messelodi, S., 2005. Switching Models for Vision-Based On-Board Road Detection. IEEE Intelligent Vehicles Symp., p.67-72.

Lookingbill, A., Rogers, J., Lieb, D., Curry, J., Thrun, S., 2007. Reverse optical flow for self-supervised adaptive autonomous robot navigation. *Int. J. Comput. Vis.*, **74**(3):287- 302. [doi:10.1007/s11263-006-0024-x]

Lutzeler, M., Dick, E.D., 1998. Road Recognition with Marveye. Proc. IEEE Intelligent-Vehicles Symp., p.341-346.

Moghadam, P., Starzyk, J.A., Wijesoma, W.S., 2012. Fast vanishing point detection in unstructured environments. *IEEE Trans. Image Process.*, **21**(1):425-430. [doi:10.1109/TIP.2011.2162422]

Thorpe, C., Carlson, J., Duggins, D., 2003. Safe Robot Driving in Cluttered Environments. 11th Int. Symp. of Robotics Research, p.271-280.

Thrun, S., Montemerlo, M., Dahlkamp, H., Stavens, D., Aron, A., Diebel, J., Fong, P., Gale, J., Halpenny, M., Hoffmann, G., *et al*., 2006. Stanley: the robot that won the DARPA grand challenge. *J. Field Robot.*, **23**(9):661-692. [doi:10.1002/rob.20147]

van de Sande, K.E.A., Gevers, T., Snoek, C.G.M., 2010. Evaluating color descriptors for object and scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, **32**(9):1582-1596. [doi:10.1109/TPAMI.2009.154]

Wang, Y., Teoh, E.K., Shen, D.G., 2004. Lane detection and tracking using B-snake. *Image Vis. Comput.*, **22**(4):269-280. [doi:10.1016/j.imavis.2003.10.003]

Wang, Y., Bai, L., Fairhurst, M., 2008. Robust road modeling and tracking using condensation. *IEEE Trans. Intell. Transp. Syst.*, **9**(4):570-579. [doi:10.1109/TITS.2008.2006733]

Zhang, G., Zheng, N., Cui, C., Yan, Y.Z., Yuan, Z.J., 2009. An Efficient Road Detection Method in Noisy Urban Environment. IEEE Intelligent Vehicles Symp., p.556-561.

Zhou, S.Y., Iagnemma, K., 2010. Self-Supervised Learning Method for Unstructured Road Detection Using Fuzzy Support Vector Machines. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, p.1183-1189.

Zhou, S.Y., Gong, J.W., Xiong, G.G., Chen, H.Y., Iagnemma, K., 2010. Road Detection Using Support Vector Machine Based on Online Learning and Evaluation. IEEE Intelligent Vehicles Symp., p.256-261.