2011 IEEE International Conference on Robotics and Automation
Shanghai International Conference Center
May 9-13, 2011, Shanghai, China

Adaptive Non-Planar Road Detection and Tracking in Challenging Environments Using Segmentation-based Markov Random Field

Chunzhao Guo, *Member, IEEE*, Seiichi Mita, *Member, IEEE*, and David McAllester, *Member, IEEE*

***Abstract*—Many roads made for land vehicles are not totally planar and present uphill and downhill slopes that follow the environment topography. Moreover, the road appearance is often affected by a number of factors in challenging conditions. In this paper, we present an adaptive non-planar road detection and tracking approach which overcomes these difficulties by a piecewise planar road model as well as a Markov Random Field (MRF)-based alternating optimization using belief propagation (BP) on segmented images and a hard conditional Expectation Maximization (EM) algorithm to achieve adaptability and optimality. The proposed framework incorporates image evidence, geometry information, and temporal support such that the graph we build and the well-defined energy minimization formulation can exploit the essence of the roads that is invariant in challenging environments. Experimental results in various real challenging traffic scenes show the effectiveness of the proposed approach.**

I. Introduction

ROAD detection is a key issue for autonomous vehicles because it provides information on the road geometry which is necessary for facilitating autonomous driving behaviors such as road following, obstacle avoidance, and off-road navigation. It has been extensively studied for several decades and dramatic development has been accomplished [1]-[7], which can be categorized into two main types of methods: vision-based methods and Light Detection And Ranging (LIDAR)-based methods. LIDAR-based systems can provide robust, accurate and stable detection in practice, which is crucial for the implementation of autonomous driving, due to the important merit of LIDAR – being insensitive to changing environments. However, LIDAR can only give the range information in a limited field. Such systems usually rely on GPS with other a priori information such as RNDF (Route Network Definition File) or GIS (Geographic Information System), although GPS has limitations on the spatial and temporal resolution and the map data may be outdated and inaccurate [8]. While such systems can perform extremely well in certain situations, vision can be utilized to perform well in a wide variety of situations since it can deliver a great amount of information, making it a powerful means for sensing the environment and identifying the objects. Therefore, vision is not only promising but also necessary for road detection and other applications related to autonomous vehicles, especially for those without any a priori knowledge. However, the vision system cannot work well in practice unless it is capable of getting accurate and robust results, especially in challenging environments where the road is not totally planar and presents uphill and downhill slopes because of the topography, and the road appearance is often affected by various factors.

In the field of vision-based road detection, some methods use a monocular camera to extract the road region by employing features with specific intensity, color and texture as visual cues on the road surface [1]–[3]. These methods are designed to be used mainly in structured environments such as urban streets and highways, since they depend on specific intensity patterns in the image, such as lane markings, pavement colors, etc. Others use a binocular camera (or cameras) for road detection by utilizing 3D structural information [4], [5], [9]. Stable and accurate reconstruction of a 3D structure by computing the disparity map is difficult due to the requirement of solving the correspondence problem for every pixel. Therefore, a number of assumptions are frequently made about the environment in order to facilitate the process. Some techniques assume that the road is planar [1], [4], which can cause false road detection or road detection failure since, in reality, many roads present uphill and downhill slopes. Others perform a dynamic estimate of the inclination of the road plane [9] by extracting and matching lane-markings in order to obtain the parameters of the road plane; however, this method is limited to the precision of the lane-marking extraction which is sometimes not reliable, e.g., poor quality of lane-markings or even no lane-markings on the road. Furthermore, while the hypothesis of flat road geometry is reasonable in the vicinity of the host vehicle, it may not be valid for the entire part of the road visible in the image. Road detection is therefore only reliable on the local planar area and the detection range is limited.

In order to cope with non-flat geometry roads as well as challenging environment conditions, we present in this paper an adaptive non-planar road detection and tracking approach which overcomes these difficulties using three key components which are also the main contributions of this paper: 1) a piecewise planar road model that simplifies the road geometry while maintaining effective road detection; 2) a segmentation-based Markov Random Field (MRF) formulation for road detection as well as a well-defined energy function that exploits the essence of the roads; 3) a hard conditional Expectation Maximization (EM) algorithm

Manuscript received September 14, 2010. This work was supported by the Research Center for Smart Vehicles of the Toyota Technological Institute, Japan. Partner institutions are the Toyota Technological Institute at Chicago, USA, and Toyota Central R&D Labs, Inc., Japan.

Chunzhao Guo and Seiichi Mita are with the Toyota Technological Institute, Nagoya, Aichi 468-8511 Japan. (e-mail: {guo, smita}@toyota-ti.ac.jp).

David McAllester is with the Toyota Technological Institute at Chicago, Chicago, IL 60637 USA. (e-mail: mcallester@tti-c.org).

that achieves adaptability as well as optimality in challenging environments without any a priori knowledge of the road. The flow diagram of the proposed approach is shown in Fig. 1.

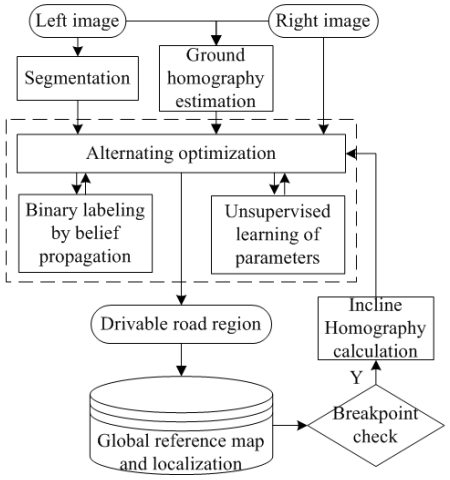


Figure 1. Flow diagram of the proposed approach

The proposed approach is developed in the scope of the stereovision-based navigation system integrated into our experimental intelligent vehicles shown in Fig. 2. The left one is a super-compact electric vehicle equipped with two computers, a stereo camera, four single beam laser scanners, a GPS, and other sensors. The right one is a hybrid vehicle equipped with six computers, a stereo camera, two four-beam laser scanners, a GPS, and other sensors. Our objective is to develop an approach for robust and accurate detection of a non-planar drivable road in challenging environments while maintaining a manageable computational load.

Figure 2. Our experimental intelligent vehicles

II. PROPOSED APPROACH

A. Road and Camera Geometry and the Underlying Cues

As mentioned previously, the task of detecting the drivable road region is a difficult task for computer vision, as the road appearance is affected by a number of factors that are not easily measured and change over time, such as the road materials, lighting and weather conditions. Therefore, invariant cues are necessary to overcome these difficulties, especially in challenging traffic scenarios. Geometry information is a very important invariant cue, particularly for land vehicles. In order to deal with uphill and downhill slopes, we model the road as a succession of parts of planar planes, as shown in Fig. 3. In the piecewise planar road model, given the breakpoints $b^1 < b^2 < ... < b^k$, the longitudinal profile of the road is a continuous function that is linear on each interval $[b^i, b^{i+1}], i = 0,1,...,k$ (for convenience we let $b^0 = 0$ and $b^{k+1} = +\infty$). The proposed model is a natural and reasonable simplification of the road in traffic scenarios, which considers the characteristics of the land vehicle and the man-made roads. It creates a detection system with lower complexity, which is important for autonomous navigation.

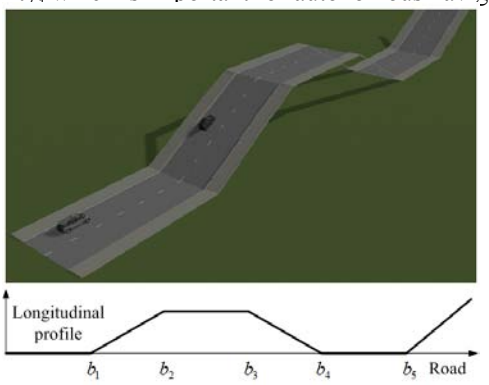


Figure 3. Piecewise planar road model

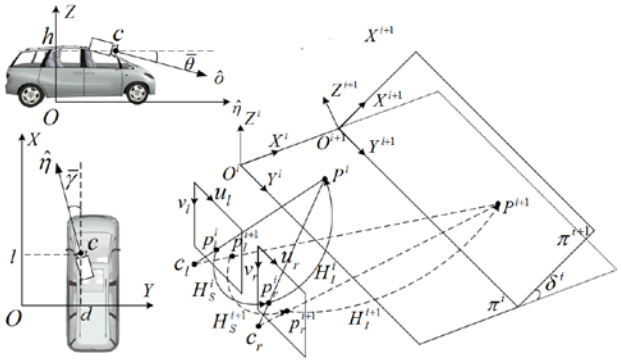


Figure 4. Road and camera geometry

The projective geometry of two cameras and the piecewise planar road model is shown in Fig. 4. The underlying geometry cues can be revealed by using the planar homography, which can relate points of images on a plane to corresponding image points in a second view. For any plane π^i in the road model, there is an inverse perspective mapping, $P^i = H_{I,l}^i p_l^i$, between the left image plane and π^i; and an inverse perspective mapping, $P^i = H_{I,r}^i p_r^i$, between the right image plane and π^i. The composition of the two inverse perspectivities is a homography induced by the plane π^i,

$$p_r^i = H_{I,r}^{i^{-1}} H_{I,l}^i p_l^i = H_S^i p_l^i \quad (1)$$

between the two image planes. As for the detection of piecewise planar roads when we employ $H_S^i, i = 0,1,...,k$ to find correspondences between the image pairs, only the road points that can comply with the homographies will have a good match while the other non-road points will not, according to (1). Therefore, all the regions that satisfy the piecewise planar road model should be contained in the matched area and classified as the drivable road region no matter how dramatically the road appearance changes in challenging conditions.

Textureless regions that connect to the road region, such as black cars, white walls, a blue sky, etc., may frequently cause errors in detection, since the intensities do not change in some parts of the textureless region although their position

does change. In order to solve this problem, all the textureless regions in the right image are extracted and their contours are recorded. For each textureless region, if parts of the transformed contour lie outside the detected road region, which indicates the shape of the textureless region has changed, the entire textureless region will be removed from the road region.

B. Homography Estimation

For a host vehicle on road segment π^i, which is always a ground plane relative to the vehicle, our strategy of detection with the piecewise planar road model is as follows:

Initialize the system [10], and iterate
- Detect the current drivable road region on π^i by the homography H_S^i induced by the ground plane.
- Register the detection result into the global reference map in the world coordinate and extract the skeleton of the drivable road to check if the road has junctions.
- Check all the far end positions of the detected road or road branches. If a position does not change along with the vehicle's movement, it means there is a potential breakpoint at this position.
- If a potential breakpoint comes into a certain range (e.g., the range of the local navigation map) from the host vehicle and blocks the entire drivable road in the heading direction, making the host vehicle unable to reach the following intended waypoints, this potential breakpoint will be regarded as the breakpoint b^i in the piecewise planar road model.
- Calculate the homography H_S^{i+1} induced by π^{i+1} based on the distance t^i between b^i and the host vehicle and the angle δ^i between π^i and π^{i+1}.
- Detect the drivable road region on π^i by H_S^i as well as the drivable road region on π^{i+1} by H_S^{i+1} at the same time and merge the two regions according to the road model.

Until the host vehicle reaches the goal point.

Figure 5. Detection strategy with the piecewise planar road model

The next task is to calculate the homography H_S^i induced by the ground plane and H_S^{i+1} according to the piecewise planar road model, if necessary. The inverse perspective mapping can be obtained by using the following camera parameters [11].

1) Viewpoint: the camera position is $c = (l, d, h) \in O^i$
2) Viewing direction: the optical axis $\hat{o}$ is determined by the following two angles, as shown in Fig. 4:
   $\bar{\gamma}$: the angle formed by the projection $\hat{\eta}$ of the optical axis $\hat{o}$ on the (X, Y) plane and the X axis;
   $\bar{\theta}$: the angle formed by the optical axis $\hat{o}$ and $\hat{\eta}$.
3) Aperture: the camera angular aperture is 2α.
4) Resolution: the camera resolution is $m \times n$.

After simple manipulations, the inverse perspective mapping from the image plane to π^i as a function of u and v is given by

$$x(u,v) = h \times ctg[(\bar{\theta} - \alpha) + u\frac{2\alpha}{m-1}] \times \cos[(\bar{\gamma} - \alpha) + u\frac{2\alpha}{n-1}] + l$$
$$y(u,v) = h \times ctg[(\bar{\theta} - \alpha) + u\frac{2\alpha}{m-1}] \times \sin[(\bar{\gamma} - \alpha) + u\frac{2\alpha}{n-1}] + d$$
$$z = 0 \qquad (2)$$

The camera parameters can be easily obtained by applying a simple external camera calibration with four reference points [12]. The homography H_S^i can be subsequently computed by (1).

Many stereovision-based road detection methods assume that the cameras are calibrated beforehand and the geometrical relations between the cameras and the road plane are known and fixed. However, these assumptions are not practical in real applications since the vehicle may tilt and the cameras may vibrate. Therefore, the geometrical relations, e.g. the homography induced by the road plane in the proposed approach, must be estimated and updated dynamically for each and every frame in order to achieve higher accuracy as well as robustness. In the present case, we estimate the homography matrix H_S^i by minimizing the compatibility cost $E(H_S^i)$ in (3) between the stereo image pair, using the Levenberg-Marquardt Algorithm (LMA) [13].

$$E(H) = \sum_{p \in R} \sum_k \lambda_k (\Phi_l^k(p) - \Phi_r^k(Hp))^2 \qquad (3)$$

where $\Phi_l(\cdot)$ and $\Phi_r(\cdot)$ are the feature vectors we constructed for the image pair, respectively, and λ_k is the weight of the k-th element. For color images in an RGB space, the vector is nine-dimensional, consisting of three color values plus a six-dimensional color gradient vector in the x and y direction. For grayscale images, it is a three-dimensional vector that consists of the intensity value plus a two-dimensional gradient vector. R is the computational region which is actually the planar road region of the image.

When estimating H_S^{i+1} induced by π^{i+1}, we simplify the situation by assuming the heading direction to be orthogonal to the Y^{i+1}-axis since the vehicle will enter the slope shortly. The relationship between π^i, π^{i+1} and the image plane is shown in Fig. 4, in which the transformation from O^i to O^{i+1} only needs to be translated by $(t^i + l)$ along the X^i-axis and rotated by δ^i about the Y^i-axis. The required camera parameters in O^{i+1} for calculating H_S^{i+1} can therefore be obtained by the following equation,

$$\begin{pmatrix} l^{i+1} \\ d^{i+1} \\ h^{i+1} \end{pmatrix} = \begin{pmatrix} \cos\delta^i & 0 & \sin\delta^i \\ 0 & 1 & 0 \\ -\sin\delta^i & 0 & \cos\delta^i \end{pmatrix} \begin{pmatrix} l^i \\ d^i \\ h^i \end{pmatrix} + \begin{pmatrix} -(t^i + l^i) \\ 0 \\ 0 \end{pmatrix}$$
$$\bar{\gamma}^{i+1} = \bar{\gamma}^i, \bar{\theta}^{i+1} = \bar{\theta}^i + \delta^i \qquad (4)$$

Thus, the inverse perspective mapping from the image plane to π^{i+1}, as a function of u, v, d^i and δ^i, can be obtained by substituting the derived parameters for (2).

t^i can easily be obtained from the global reference map; however, it is difficult to estimate δ^i if there is no a priori knowledge of the road. Some methods calculate δ^i by extracting and matching lane-markings on π^{i+1}; however, this method would not work well in roads with barely visible

lane-markings or even without lane-markings. In the proposed approach, we infer δ^i between $[-\pi/3, \pi/3]$ to calculate H_S^{i+1}, then compute the drivable roads for each inference. The best result that extends the drivable road over the breakpoint to include the next waypoints will be selected to update the navigation map.

C. Graph Construction for Road Detection

We use a MRF framework for finding the matched regions on π^i, and π^{i+1} if necessary, based on homographies induced by the planar road segments, since it models the spatial interactions present in the scene and uses the inference algorithm to find the most likely setting of each node in the MRF. In the conventional MRF model, each graph node corresponds to an image pixel and the inference algorithm finds the most likely setting for each pixel. However, human drivers do not see squared points (pixels) in the traffic scene but objects, and the decision whether the object is drivable or not is made over the region of each object instead of over every single pixel. Therefore, in order to mimic the perception behavior of human drivers and develop a more sophisticated system that utilizes some semantic knowledge, we first use a fixed oversegmentation for the reference image such that each image segment, which we call superpixel, corresponds to an object or object part of the traffic scene. We subsequently construct our MRF model in such a way that each superpixel is a graph node and adjacent superpixels are connected by edges. This segmentation-based MRF can capture the perceptual meaning of the scene in the image data. Therefore, we can allow for a sharp discontinuity of labeling between superpixels, while still maintain the smoothness assumption within each superpixel. As a consequence, our approach is less likely to suffer from the over-smoothing problem while the conventional interactions of the smoothness cost information across the whole image may lead to poor performance at the object boundaries. Furthermore, the size of the segmentation-based MRF is much smaller than the conventional one and the beliefs of the inference algorithm are propagated in a much more efficient manner, which allows fast labeling, since we also have to infer δ^i when a breakpoint appears.

To segment the image, we use an efficient algorithm described in [14], which is based on a pairwise region comparison predicate. This predicate measures the dissimilarity between elements along the boundary of the two components relative to a measure of the dissimilarity among neighboring elements within each of the two components. This algorithm has an advantage over the color-based segmentation commonly used by other segmentation-based stereo methods. It captures perceptually important non-local image regions, which often reflect the global aspects of the image. This property strongly supports the assumption that the traffic scene structure can be approximated by a set of non-overlapping smooth surfaces in 3D space, each surface corresponding to an image segment/superpixel. In addition, this segmentation algorithm runs in time nearly linear in the size of the image and is fast in practice. Note that we over-segment the image in order to avoid arbitrary segmentation that may merge non-road regions with road regions, making either false positives or false negatives.

D. Binary Labeling by Belief Propagation

In the proposed approach, we present the road detection on each road segment in the piecewise planar road model as a binary labeling $\mathcal{L}: x \to \{0,1\}$ for each superpixel s in the segmented reference image, and the assigned label denoted by a random variable f is

$$f(s) = \begin{cases} 1 & s \in \text{Drivable road region} \\ 0 & \text{Otherwise} \end{cases} \tag{5}$$

Our goal is to find the correspondence that matches pixels of similar intensity as well as gradients based on the homography induced by the road plane while minimizing the number of discontinuities between the left and right images. We accomplish this by minimizing the following energy function, which describes the quality of labeling,

$$E = E_M + E_T + E_S \tag{6}$$

where E_M is a matching energy containing the costs of assigning the labels to the superpixels based on the image evidence. E_T is a tracking energy containing the costs of assigning the labels to the superpixels based on the temporal support. E_S is a smoothness term containing the cost of assigning labels to two neighboring superpixels, which enforces smoothness by penalizing the discontinuities. We employ the loopy belief propagation to find the labeling that minimizes the proposed energy function, which corresponds to the maximum a posteriori (MAP) estimation for MRF. Note that we infer the label for each superpixel s and any pixel p inside s is therefore assigned with the same label,

$$f(p \mid \forall p \in s) = f(s) \tag{7}$$

In the proposed system, we define E_M as

$$E_M = \sum_{s \in S} \sum_{p \in s} \left(\sum_k \lambda_k (\Phi_l^k(p) - \Phi_r^k(H_S p))^2 \cdot f(p) + \lambda_m (1 - f(p)) \right) \tag{8}$$

where S is the set of superpixels in the left image. $H_S p$ is the pixel in the right image that corresponds to the pixel p in the left image under H_S. $\Phi_l(p)$ and $\Phi_r(p)$ are the feature vectors and λ_k denotes the weights, as mentioned previously. The use of the gradient vector is aimed at improving the performance by encouraging the labeling discontinuities to be aligned with the intensity edges. The weighted sum of the squared difference of the two vectors is utilized to measure the compatibility cost, and λ_m is a thresholding factor designed to adjust the influence of the compatibility cost on the labeling.

We define E_T as

$$E_T = \sum_{s \in S} \sum_{p \in s} \min(\tau_t, \lambda_t I_l(p) \mid f^t(p) - f^{t-1}(\mathcal{T}(p)) \mid) \tag{9}$$

where λ_t is the weight of the tracking energy and τ_t is the truncation threshold. f^{t-1} and f^t are the labeling in the left image at time $t-1$ and t. $\mathcal{T}$ indicates the transformation according to the motion of the host vehicle from $t-1$ to t, and the transformed labeling map serves as the prediction of the new labeling.

We define E_S as

$$E_S = \sum_{p \in s_i, q \in s_j, (s_i, s_j) \in G} \min(\tau_s, \lambda_s \mid f(p) - f(q) \mid) \tag{10}$$

where G denotes the undirected edges in the graph. s_i, s_j are the neighboring superpixels in G. λ_s is the weight of the smoothness term and τ_s is the truncation threshold.

E. Unsupervised Learning of the Parameters

The parameters of our energy function were learned using a hard conditional EM algorithm applied to the stereo data to maximize the conditional likelihood. Some approaches, such as the one in [15], model the joint distribution of cues over the left and right images with various independence assumptions, which are not always correct assumptions. Our approach models the conditional probability distribution of the right image given the left image so that there is no danger of corrupting the model by modeling the distribution over both of the images poorly. The truncation thresholds τ_t and τ_s will not be estimated since they are insensitive to changing environments under our formulation of road detection. Therefore, we consider a general conditional probability model $P_\beta(I_r \mid I_l)$ over the input image pair I_l and I_r, and define it in terms of a parameter vector $\beta = \{\lambda_k, \lambda_m, \lambda_t, \lambda_s\}$ and the latent variable f,

$$P(I_r \mid I_l, \beta) = \sum_f P(I_r, f \mid I_l, \beta) \tag{11}$$

The hard conditional EM is an algorithm for locally optimizing the parameter vector β so as to maximize the probability of I_r given I_l.

$$\beta^* = \arg\max_\beta \ln P(I_r \mid I_l, \beta) \tag{12}$$

with the following two updates:

$$\text{Hard E step:} \quad f := \arg\max_f P(I_r, f \mid I_l, \beta) \tag{13}$$

$$\text{Hard M step:} \quad \beta := \arg\max_\beta \ln P(I_r, f \mid I_l, \beta) \tag{14}$$

In the proposed system, the hard E step is implemented using belief propagation in the segmentation-based MRF which computes f by minimizing the proposed energy function. The implementation of the hard M step relies on a factorization of the probability model into two conditional probability models.

$$P(I_r, f \mid I_l, \beta) = P(f \mid I_l) P(I_r \mid I_l, f, \beta) \tag{15}$$

where $P(f \mid I_l)$ is independent of β. Therefore, (14) can be written as the following update.

$$\beta := \arg\max_\beta \ln P(I_r \mid I_l, f, \beta) \tag{16}$$

F. Alternating Optimization

Here, we present an alternating algorithm to find the optimal binary labeling of the road image and learn the parameters from the stereo pair itself. Given f, we apply the hard conditional EM algorithm to compute the parameters in β using (16). Using the learned parameters, we then implement belief propagation to find the assignment of the labeling that minimizes (6). The alternating optimization procedure runs until convergence or a fixed number of iterations.

1) Computing λ_k, λ_m given f, H_S : As mentioned previously, the points in the road region have correspondences between the stereo image pair based on homography while other points do not. Therefore, considering the matching energy (8), the conditional probability model $P(I_r \mid I_l, f, H_S, \lambda_k)$ for the hard M step is defined as

$$P(\Phi_r^k(p') = x \mid \Phi_l^k, f, H_S, \lambda_k) = \begin{cases} \varsigma \exp(-(\Phi_l^k(p) - x)^2 / (2\sigma_k^2)) & p' = H_S p \;\&\; f(p) = 1 \\ 1/N & \textit{Otherwise} \end{cases} \tag{17}$$

where σ_k^2 is the variance for the Gaussian distribution and ς is a normalization factor. N is the number of possible intensity/gradient values for the feature vector. Substituting (17) for (16) results in an equation with a closed form solution, i.e. $\lambda_k = 1/(2\sigma_k^2)$, which could be used to update λ_k. σ_k^2 can be obtained from the input stereo image pair as follows,

$$\sigma_k^2 = \frac{1}{N_R} \sum_{p \in R} (\Phi_l^k(p) - \Phi_r^k(H_S p))^2 \tag{18}$$

where N_R is the size of road region R.

As described previously, λ_m is a thresholding factor to discriminate good matches from bad matches. If we take the compatibility cost term as a 2D grayscale plot of the compatibility cost per pixel, the good matches should be dark (background) and the bad matches should be bright (foreground). Inspired by Otsu's thresholding method [16], we determine λ_m by maximizing the inter-class variance σ_b^2 defined as

$$\sigma_b^2 = \omega_1'(c)\omega_2'(c)(\mu_1(c) - \mu_2(c))^2 \tag{19}$$

where the weights ω_i' are the probabilities of the two classes separated by a threshold c, and μ_i are class means. In our hard M step, the inter-class variance is calculated using (19) given f, and the result is written as $\sigma_{b,f}^2$. Subsequently, λ_m iterates through all the possible threshold values to calculate the inter-class variance σ_{b,λ_m}^2 and is set to the value that minimizes the absolute difference between $\sigma_{b,f}^2$ and σ_{b,λ_m}^2.

2) Computing λ_t given f^t, f^{t-1} : For the use of tracking, the conditional probability model is actually $P(I_r \mid I_l, f^t, f^{t-1}, \beta)$ which can be factorized as

$$P(I_r, \mid I_l, f^t, f^{t-1}, \beta) = P(f^t \mid I_l, f^{t-1}, \lambda_t) P(I_r, \mid I_l, f^t, \beta_{\overline{\lambda_t}}) \tag{20}$$

where $\beta_{\overline{\lambda_t}}$ indicates the parameters, apart for λ_t. $P(I_r, \mid I_l, f^t, \beta_{\overline{\lambda_t}})$ is irrelevant to the tracking energy. Considering the tracking energy (9), we define $\Delta f_t(p) = \mid f^t(p) - f^{t-1}(\mathcal{T}(p)) \mid$ and then the conditional probability for the tracking step is defined as

$$P(\Delta f_t(p) = x \mid I_l, f^{t-1}, \lambda_t) = \begin{cases} \eta \exp(-\mu x) & f^t(p) = 1 \\ 1/2 & \text{Otherwise} \end{cases} \quad (21)$$

where μ is the rate parameter for the exponential distribution and η is a normalization factor. Substituting (21) for (16) results in an equation with a closed form solution to update λ_t, which can be obtained from the adjacent left image pair as follows,

$$\lambda_t = 1/(\frac{1}{N_R}\sum_{p \in R}(I_l(p) \mid f^t(p) - f^{t-1}(\mathcal{T}(p)) \mid)) \quad (22)$$

3) Computing λ_s given f : λ_s can be calculated in the same way as λ_t if we define $\Delta f_s(p,q) = \mid f(p) - f(q) \mid$. However, in our formulation of road detection, we force the detected drivable road to be a connected region in front of the vehicle. Therefore, in the experiments, we set λ_s by using the squared difference between the mean intensity of the two superpixels s_i and s_j. The idea is that the smoothness violation between adjacent superpixels should be penalized more for superpixels that look similar, and less for superpixels that look different.

4) Computing f given β : As mentioned previously, we implement an efficient belief propagation approach for inference to find the assignment of the labeling that minimizes (6).

III. Experimental Results

In the experiments, the proposed approach has been implemented to test a wide variety of typical but challenging scenarios with uphill and downhill slopes as well as bending road surfaces without code optimization. The experimental specification is shown in Table I. We implement belief propagation on segmented images to find the optimal labeling, and implement the hard conditional EM algorithm to learn the parameters. In all the experiments, we alternate between BP and EM twice and the average computational time is approximately under 300 ms/frame for the images with a resolution of 320×240. The computational time may be longer for roads with many slopes; however, in these conditions the vehicle should be driving slowly as human drivers do. The system therefore has enough time to find the inclined road regions. The computational time will be further reduced by utilizing special image processing hardware, such as GPU and FPGA, as well as optimizing the codes of the proposed method. Hardware acceleration with GPU is under development and the rate of more than 10Hz can be expected.

TABLE I. Experimental Specifications

CPU	Xeon(R) X5550 @2.67 GHz
Memory (RAM)	12.0GB
Operating system	Window Xp Professional Sp2
Programming language	C++ with OpenCV library
Experimental data	Grayscale images (maximum visible range of the road > 100m)
Iterations	BP×2+EM×2+LMA×10
Resolution	320×240
Computational time	<300 ms/frame

We prove the effectiveness of the proposed approach to overcome the difficulties mentioned previously by experiments. Our first experiment is to demonstrate our segmentation-based MRF. Here, we use the publicized source code from Pedro Felzenszwalb [17] for segmentation. The segmentation result and the corresponding segmentation-based MRF as well as the detection results are shown in Fig. 6. Each segment with a random color corresponds to a graph node and the edges between neighboring nodes are indicated by blue lines. (a)–(b) are the input image pair. The left image contains strong backlight while the right one does not. (c) is the segmentation result with the suggested parameters: $\sigma = 0.5$, $k = 500$ and the minimum component size = 20 [17]. (d) is the detection result, which is poor due to the arbitrary segmentation where the bush and part of the road are segmented as one component. (e) is the segmentation result with the parameters: $\sigma = 0.5$, $k = 100$ and the minimum component size = 30. Although the segmentation is somewhat above that required, the segments can still capture the perceptual meaning of the scene in the image data. As shown in (f), the following belief propagation in the segmentation-based MRF merges the segments into road/non-road regions by minimizing the proposed energy function.

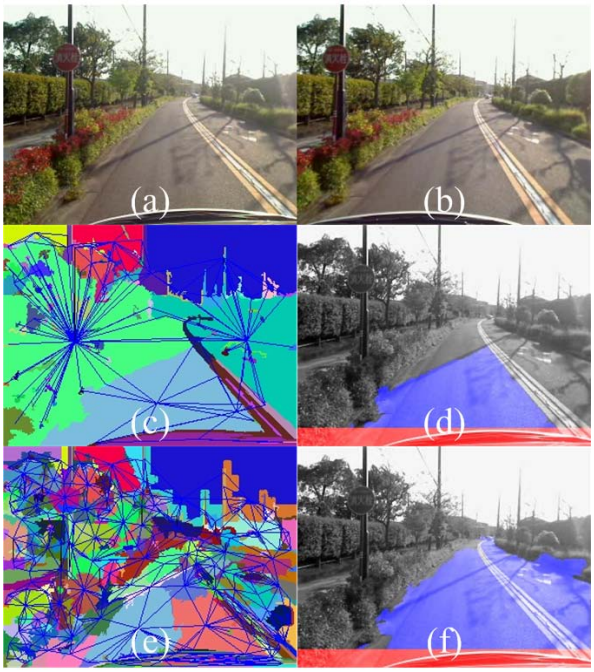


Figure 6. Examples of segmentation-based MRF and the detection results

Our second experiment reveals the significance of dynamic ground homography estimation. Fig. 7 shows an example of the comparison of the results without the dynamic estimation of the ground homography [(a) and (c)], and with it [(b) and (d)]. In (a), the road points around the pedestrian crossing are not contained in the detected road region because the two images do not coincide well, as shown in (c). When dynamic estimation was applied, the homography is optimized. Subsequently, the match of the road points becomes very good, as shown in (d), and the road points around the road markings are contained in the detected road region as shown in (b).

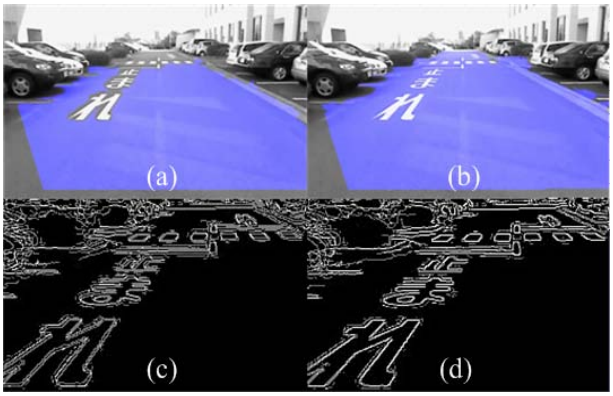


Figure 7. Comparison of results with/without dynamic ground homography estimation

Our third experiment proves the effectiveness of the piecewise planar road model in coping with slopes. Fig. 8 gives an example of the host vehicle's movement from π^i to π^{i+1}. In the first 3 images, the breakpoint appeared and is detected. The homography H_S^{i+1} was subsequently estimated so that the drivable road region on π^{i+1} was detected and merged into the final result (frame 1735–1755). In frame 1760, the vehicle entered into π^{i+1}, which became the ground plane for the vehicle's driving.

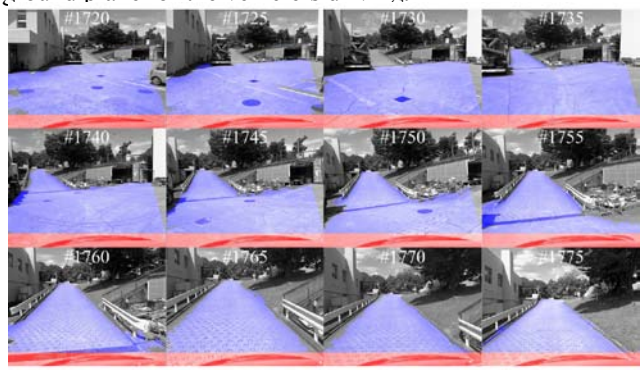


Figure 8. Examples of the road detection with slopes

Our forth experiment evaluates the optimality and adaptability of the proposed system. First, we fix all the other parameters with the learned values but vary each of the $\lambda_1, \lambda_2, \lambda_3, \lambda_m$ values singly in a certain range and implement BP to find the binary labeling by minimizing the proposed energy function. We plot the error rate as a function of each evaluated parameter in Fig. 9. The red bar indicates our learned values, whose corresponding error rates are quite close to the minimum error rates of the graphs.

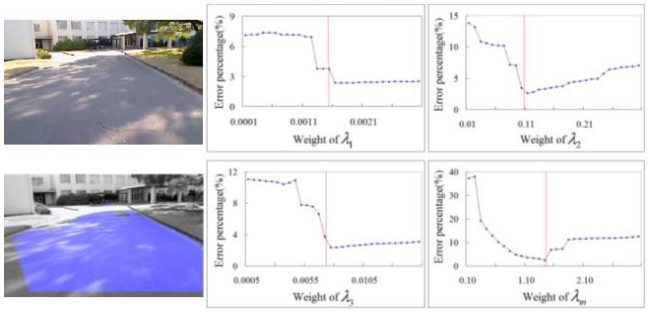


Figure 9. Optimality evaluation of the unsupervised learning

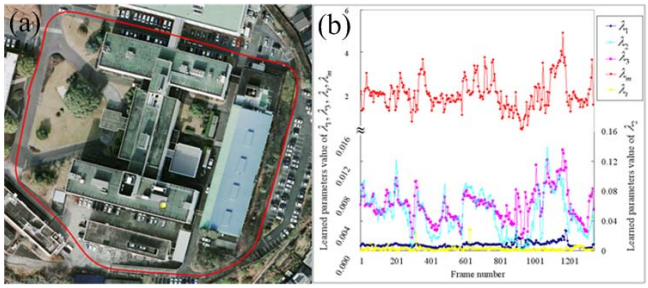
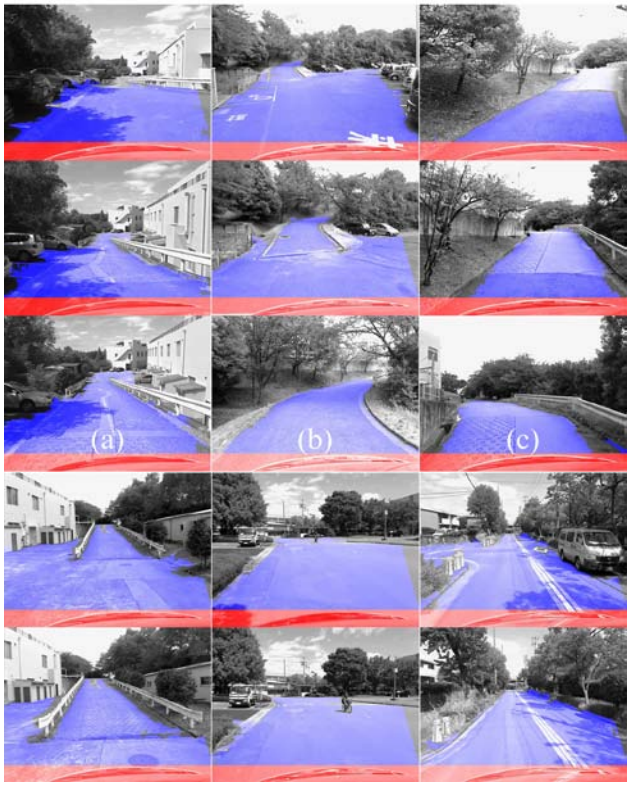


Figure 10. Parameter adaptation in long range transition senarios

We also implement the proposed system in a long range loop inside our campus, as shown in Fig. 10(a), which includes complex road scenarios with changing environments. We plot all the estimated parameters every 5 frames in Fig. 10(b), which shows the long term adaptation of the parameters learned from the input image pair itself. This experiment proves that the automatic parameter tuning procedure of the proposed system improves both the accuracy and robustness of the road detection.

Next, we give some example results of road detection in various typical but challenging scenarios, especially for the non-planar roads with affected appearances. Example results are shown in Fig. 11, in which, the detected drivable road region is indicated in blue. The red area was not detected due to the occlusion of the car body. The difficulties include downhill slopes (a),(g), uphill slopes (b),(c),(d),(i), bending road surface without obvious breakpoints (h), dark shadows (a),(f),(g),(i), and extreme inhomogeneous surface. As we can see from the figure, the proposed system overcame these difficulties and the detected drivable road regions are in good agreement with the real boundaries. (e),(f) also show that our system can detect pedestrians, bicycles and cars effectively since they are not true of the piecewise planar model.



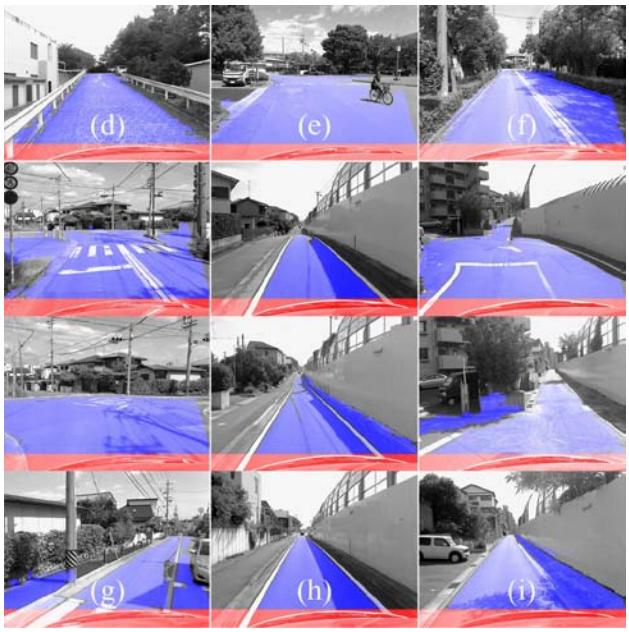


Figure 11. Example results in challenging traffic scenarios, in which, there are 3 frames per sequence in a column.

TABLE II. QUANTITATIVE EVALUATIONS OF ROAD DETECTION USING THE PROPOSED SYSTEM AND SAD MATCHING METHOD

	Method using SAD matching			The proposed system		
	FPR (%)	FNR (%)	Error rate (%)	FPR (%)	FNR (%)	Error rate (%)
(a)	2.33	5.07	4.38	0.24	1.20	0.86
(b)	2.10	2.13	2.12	0.31	0.33	0.32
(c)	2.54	5.34	4.33	0.71	1.09	0.83
(d)	9.07	7.49	8.90	0.21	1.51	1.13
(e)	1.21	2.55	3.05	0.01	0.74	0.21
(f)	2.83	9.07	6.38	0.34	2.01	1.36
(g)	2.80	5.13	4.80	0.81	2.43	1.46
(h)	5.54	13.34	12.33	0.51	8.09	7.33
(i)	11.13	9.49	10.60	2.48	2.81	2.58

Furthermore, Table II gives the quantitative evaluation results of the image sequences as well as a comparison with the road detection method using SAD (Sum of Absolute Difference) matching [10]. It utilizes three ratios, the false positive ratio: FPR = N_{FP} /N_P×100%, the false negative ratio: FNR = N_{FN} /N_N×100%, and the error rate: Error rate = (N_{FP} +N_{FN})/ (N_P +N_N)×100%. N_{FP}, N_{FN}, N_P and N_N are the number of false positives, false negatives, true positives and true negatives, respectively. The ground truth are obtained by a human operator. The quantitative evaluation shows the high accuracy of the proposed system. From the experimental results we can see that, the proposed system can provide the road information accurately as well as robustly and give a safe path for the host vehicle to facilitate autonomous driving behaviors such as road following, obstacle avoidance, and off-road navigation.

IV. CONCLUSION

In this paper we presented an adaptive drivable road detection and tracking approach designed for autonomous driving in challenging environments, such as non-planar roads with affected appearance. Our first contribution is the piecewise planar road model and the estimation of the homographies induced by the ground and inclined planes. Our second contribution is the formulation of road detection in a segmentation-based MRF and a well-defined energy function that exploits the essence of the roads. The third one is the unsupervised learning of the parameters in the segmentation-based MRF from the stereo pair itself using a hard conditional EM algorithm. The proposed framework incorporates image evidence and geometry information as well as temporal support such that higher accuracy and robustness could be expected.

REFERENCES

[1] M. Bertozzi, A. Broggi, and A. Fascioli, "Vision-based intelligent vehicles: State of the art and perspectives," *Robot. Autonom. Syst.*, vol. 32, pp. 1–16, 2000.

[2] Y. Sha, G. Zhang, and Y. Yang, "A road detection algorithm by boosting using feature combination," in *Proc. IEEE Intelligent Vehicles Symp.*, Istanbul, Turkey, June 2007, pp. 364–368.

[3] Y. He, H. Wang, and B. Zhang, "Color-based road detection in urban traffic scenes," *IEEE Trans. on Intelligent Transport System*, Vol. 5, No. 4, 2004, pp. 309–318.

[4] M. Okutomi, K, Nakano, J. Maruyama, and T. Hara, "Robust estimation of planar regions for visual navigation using sequential stereo images," in *Proc.of the 2002 IEEE International Conference on Robotics & Automation*, Washington, DC, May 2002, pp. 3321–3327.

[5] N. Hautiere, R. Labayrade, M. Perrolaz, and D. Aubert, "Road scene analysis by stereovision: a robust and quasi-dense approach," in *Proc. of the Ninth International Conference on Control, Automation, Robotics and Vision*, Singapore, Dec. 2006, pp. 1–6.

[6] A.V. Reyher, A. Joos, H. Winner, "A lidar-based approach for near range lane detection," in *Proc. Conf. on Intelligent Vehicles Symposium*, 2005, pp. 147–152.

[7] D. Munoz, N. Vandapel, M. Hebert, "Onboard contextual classification of 3-D point clouds with learned high-order Markov random fields," in *Proc. Conf. on Robotics and Automation*, 2009, pp. 2009–2016.

[8] Z. Kim, "Robust lane detection and tracking in challenging scenarios," *IEEE Trans. Intell. Transp. Syst.*, vol. 9, no. 1, pp. 16–26, Mar. 2008.

[9] D. Koller, T. Luong, J. Malik, "Binocular stereopsis and lane marker flow for vehicle navigation: lateral and longitudinal control," *Technical Report UCB/CSD 94-804*, 1994.

[10] C. Guo, and S. Mita, "Drivable Road Region Detection Using Homography Estimation and Efficient Belief Propagation with Coordinate Descent Optimization," in *Proc. Conf. on Intelligent Vehicles Symposium*, 2009, pp. 317–323.

[11] M. Bertozzi, A. Broggi, "GOLD: a parallel rea-time stereo vision system for generic obstacle and lane detection," *IEEE Trans. on Image Processing*, Vol. 7, No. 1, 1998, pp. 62–81.

[12] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[13] P. Gill and W. Murray, "Algorithms for the Solution of the Nonlinear Least-squares Problem,", SIAM Journal on Numerical Analysis, 1978, pp. 977–992.

[14] P. Felzenszwalb, D. Huttenlocher, "Efficient graph-based image segmentation," *IJCV*, vol. 59, no. 2, 2004.

[15] L. Zhang, S. Seitz. "Parameter estimation for MRF stereo," in *Proceedings of CVPR,* 2005, pp. 288–295.

[16] N. Otsu, "A threshold selection method from gray-level histograms,", *IEEE Transactions on Systems, Man and Cybernetics*, 1979, pp. 62–66.

[17] http://people.cs.uchicago.edu/pff/bp/