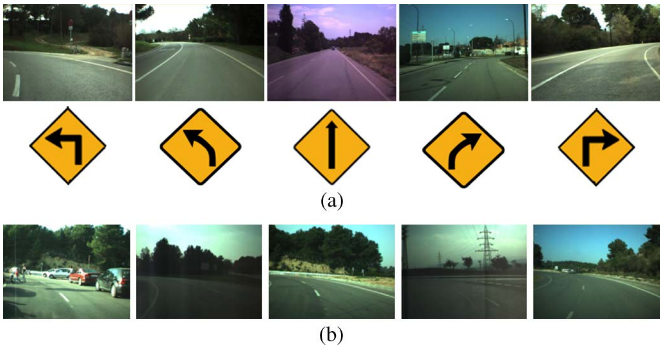
Road Geometry Classification by Adaptive Shape Models

José M. Álvarez, Theo Gevers, *Member, IEEE*, Ferran Diego, and Antonio M. López

***Abstract*—Vision-based road detection is important for different applications in transportation, such as autonomous driving, vehicle collision warning, and pedestrian crossing detection. Common approaches to road detection are based on low-level road appearance (e.g., color or texture) and neglect of the scene geometry and context. Hence, using only low-level features makes these algorithms highly depend on structured roads, road homogeneity, and lighting conditions. Therefore, the aim of this paper is to classify road geometries for road detection through the analysis of scene composition and temporal coherence. Road geometry classification is proposed by building corresponding models from training images containing prototypical road geometries. We propose adaptive shape models where spatial pyramids are steered by the inherent spatial structure of road images. To reduce the influence of lighting variations, invariant features are used. Large-scale experiments show that the proposed road geometry classifier yields a high recognition rate of 73.57% ± 13.1, clearly outperforming other state-of-the-art methods. Including road shape information improves road detection results over existing appearance-based methods. Finally, it is shown that invariant features and temporal information provide robustness against disturbing imaging conditions.**

***Index Terms*—GIST, holistic representation, illuminant invariant, image classification, road detection, scene classifier, scene recognition, spatial pyramids, support vector machine.**


(a)

(b)

Fig. 1. Road models are discrete prototypical road geometries. The main challenge in recognizing road shapes is dealing with (a) interclass versus (b) intraclass variability.

I. INTRODUCTION

VISION-BASED road detection aims at the detection of the free road surface ahead the ego-vehicle and is an important research topic in different areas of transportation systems such as autonomous driving [1], [2], vehicle collision warning [3], and pedestrian crossing detection [4]. Detecting the road in images taken from a mobile camera in uncontrolled cluttered environments is a challenging problem. The appearance of the road varies, depending on the time of the day, road type, illumination, and acquisition conditions. Many road detection algorithms have been proposed since the first autonomous vehicles were released. Common approaches to road detection use pixel-level features (such as color [2], [5]–[8] or texture [9], [10]) to characterize the appearance of the road and group pixels in two different groups, namely, drivable road and background. The performance of these algorithms is commonly improved considering temporal information based on heuristic rules [5] or temporal smoothing [11], [12]. For instance, in [11], Michalke *et al.* averaged past detection results to constraint the analysis of the current image. In [12], Alvarez *et al.* used time series analysis to predict the expected results instead of simple averages over past results. However, algorithms based on low-level features reach their limitations in situations where color or texture information is not reliable (e.g., severe lighting variations such as strong shadows and highlights). In these situations, the analysis of the context is more valuable for road detection algorithms.

Contextual information provides relevant information regarding the location of the road. For instance, in [13], Kong *et al.* exploited the perspective effect in an image to locate the vanishing point and provide a rough segmentation of the road. In [14], contextual information was exploited to recover the 3-D scene layout of a scene. Then, road detection was performed assuming that the road lies within horizontal regions. Another approach consists in estimating the composition of the road ahead the ego-vehicle to provide relevant information regarding the location of road regions. In fact, common road layouts can be divided in discretized prototypical versions, as shown in Fig. 1. An image of a road could be then assigned to one of

Manuscript received February 5, 2012; revised July 4, 2012; accepted September 10, 2012. Date of publication November 12, 2012; date of current version February 25, 2013. This work was supported by the Spanish Ministry of Education and Science through the Research Program Consolider Ingenio 2010: Multimodal Interaction in Pattern Recognition and Computer Vision, under Grant CSD2007-00018 and through the Ministry of Economy and Competitiveness Projects under Grant TRA2011-29454-C03-01, Grant TIN2011-25606, Grant TIN2011-29494-C03-02, and Grant TRA2007-62526/AUT. The Associate Editor for this paper was P. Grisleri.

J. M. Álvarez is with NICTA, Canberra, Australia, with the Computer Vision Center, 08193 Barcelona, Spain, and also with the Department of Computer Science, Universitat Autònoma de Barcelona, 08193 Barcelona, Spain (e-mail: jalvarez@cvc.uab.es).

T. Gevers is with the Faculty of Science, University of Amsterdam, 1098 Amsterdam, The Netherlands, and also with the Computer Vision Center, 08193 Barcelona, Spain (e-mail: th.gevers@uva.nl).

F. Diego is with the Heidelberg Collaboratory for Image Processing (HCI), University of Heidelberg, 69115 Heidelberg, Germany, and also with the Computer Vision Center, 08193 Barcelona, Spain (e-mail: ferran.diego@iwr.uni-heidelberg.de).

A. M. López is with the Computer Vision Center, 08193 Barcelona, Spain, and also with the Department of Computer Science, Universitat Autònoma de Barcelona, 08193 Barcelona, Spain (e-mail: antonio@cvc.uab.es).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TITS.2012.2221088

the discrete classes determining the possible locations of the road. We follow this paradigm and propose a road geometry classification system to improve road detection by exploiting road information provided by the analysis of the scene.

The aim of this paper is to classify road geometries for road detection by exploiting road information through the analysis of scene composition and temporal coherence. Road geometries are discretized versions of prototypical road compositions, including strong turns (left and right), soft turns (left and right), straight roads, and different traffic compositions (e.g., junctions or tunnels). Road geometry classification is based on learning corresponding models (hereafter, road shape models) from training images containing the different road geometries. These road shape models are defined as descriptions of images exhibiting aspects of typical road geometries (e.g., left turn, straight, right turn) to provide relevant semantic information regarding the position of the road in an image. Surprisingly, road shape models via scene classification have been largely ignored so far. A pioneer approach has been proposed in [15]. In this algorithm, images are globally described (i.e., using dense sampling) using scale-invariant feature transform (SIFT) in the opponent color space to reduce the influence of lighting variations. However, the algorithm discards all the information about the spatial layout of the features and exhibits limited performance to distinguish between different road models. Furthermore, the algorithm assumes that consecutive frames are independent without exploiting the possible correlation between adjacent images in a road image sequence.

Therefore, in this paper, we propose a novel road classification system for the purpose of road detection. The classifier is derived from the bag-of-words approach using spatial pyramids [16], that is, a global image description based on aggregating statistics of local features over fixed subregions. As spatial pyramids with fixed subregions fail to encapsulate road segments, we propose an adaptive shape model where spatial pyramids are steered by the inherent spatial structure of road images. In particular, we consider the horizon line to steer the spatial pyramid layout. This way, the spatial pyramids adapt their tessellation grids to the image layout. Hence, the proposed classification system uses adaptive shape models that are derived by the inherent structure of road images. Further, robustness to lighting variations is achieved by using (physics-based) illuminant–invariant feature space. Moreover, we incorporate temporal information into the road shape models to enforce coherence among road geometries in image sequences (i.e., a strong right turn cannot be immediately preceded by a strong left turn) using a probabilistic framework. In particular, coherence between images is learned by a hidden Markov model (HMM).

In summary, the main contributions of this paper are the following: 1) road geometry classification based on scene geometry using 2) illuminant–invariant features to minimize the influence of lighting variations; 3) dynamic spatial pyramids to extract features from relevant (road) segments; and 4) temporal context using a probabilistic framework. Finally, we propose a 5) road detection algorithm based on road shape models and low-level image features.

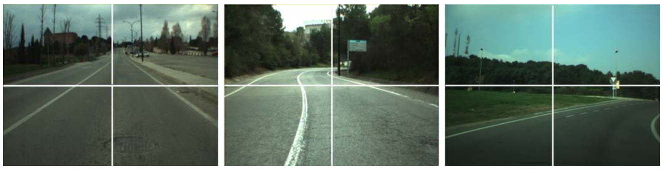


Fig. 2. Spatial pyramids capture the spatial layout of the scene by dividing the image into fixed regions. However, using fixed regions may fail to encapsulate relevant road geometry information.

The rest of this paper is organized as follows: Section III describes the algorithm to estimate road shape models. Then, the algorithm to detect roads based on adaptive road shape models is described in Section V. Experimental results and discussion are presented in Section VI. Finally, in Section VII, conclusions are drawn.

II. Related Work

Road scene classification aims to assign a semantic label according to the geometry of the road in the scene. Road scene classification is a challenging problem that is mainly due to intraclass and interclass variability. A common approach to scene classification consists in extracting features from training images and training a classifier for subsequent labeling of new images. A number of features have been proposed for scene classification [16]–[19]. For instance, in [19], Oliva and Torralba encapsulated the dominant spatial structure of the image using a GIST descriptor based on a Gabor filter bank. A different approach consists in using local descriptors based on the bag-of-words method using either dense sampling or salient point detection. These methods discard the information about the spatial layout of image features. Therefore, image pyramid representations are proposed [16], [18]. In [16], Lazebnik *et al.* proposed a global image description based on aggregating statistics of local features over fixed subregions. In particular, local features are estimated using SIFT [17]. Although a number of color descriptors have been proposed (HSV-SIFT, C-SIFT, OpponentSIFT, and HueSIFT, among others [20]), the performance of these algorithms is effected by changes in the illumination of a scene (particularly a road scene [8]). That is, descriptors based on transformed color spaces are, to a certain degree, still sensitive to lighting variations such as strong shadows and highlights.

As aforementioned, a way to incorporate spatial information is to use spatial pyramids such as in [18], which uses a spatial pyramid and the Histogram of Gradient Orientation (HoG) proposed in [21]. However, spatial pyramid schemes use fixed region subdivision (e.g., 2 × 2, 3 × 3, or 1 × 3). Fixed subregions may fail to encapsulate relevant road information, as shown in Fig. 2, where fixed regions do not align with road segments. Therefore, in the following section, we propose describing road images using adaptive road shape models. The main novelty of these models relies on steering spatial pyramids by the underlying structure of road data. For example, a horizon detector is used for initializing spatial pyramids.

III. Adaptive Road Shape Model: A Visual Descriptor for Road Scenes

Road shape models are descriptions of images exhibiting aspects of typical road geometries such as left turn, right turn,

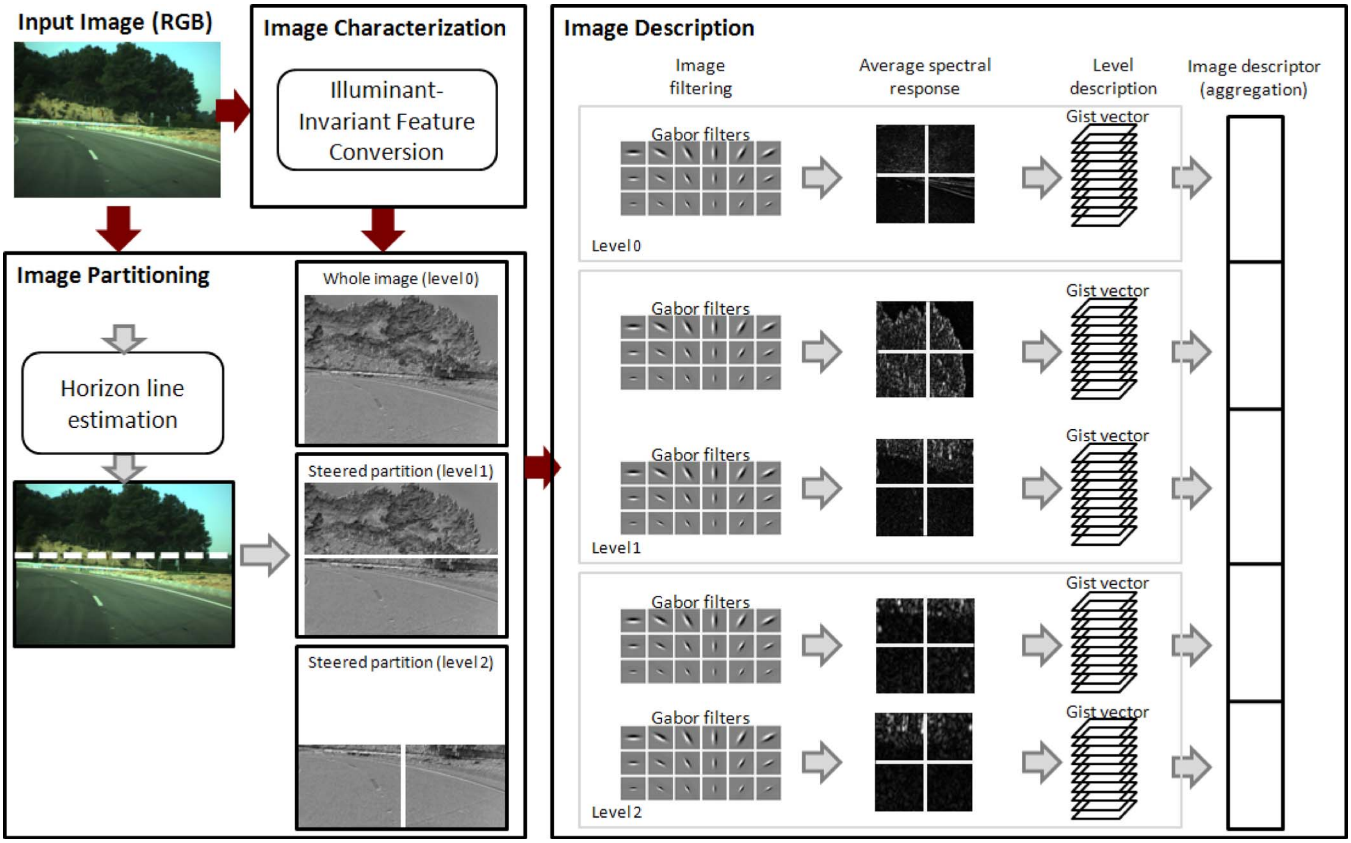


Fig. 3. Schematic of road image representation using adaptive road shape models (e.g., descriptions of images exhibiting aspects of typical road geometries such as left turn, straight, and right turn). The input image is first converted into illuminant–invariant feature space to minimize the effects of lighting variations. Then, the spatial partition is based on the horizon line to deal with the dynamic nature of road images and to encapsulate all the relevant information in the lower partitions.

straight road, or junctions. These models are learned from training images, and then, given a new image, the appropriate model is assigned by a scene classifier [15]. Learning and classifying road images is challenging due to the interclass versus intraclass variability (see Fig. 1). That is, there is a smaller variation between images in different classes than within a class itself. An important source of intraclass variability is shadows and lighting variations [see Fig. 1(a)], whereas the lower interclass variability arises from similar images exhibiting different road shapes that vary only local areas of the images [see Fig. 1(b)].

Here, we propose a novel algorithm to describe road images using steered spatial pyramids (see Fig. 3). The algorithm consists of three main stages: image characterization, image partitioning, and image description.

A. *Image Characterization*

The first stage aims to minimize the influence of lighting variations and shadows by characterizing pixels using illuminant–invariant feature space. In particular, in this paper, we focus on the illuminant–invariant feature space introduced by Finlayson *et al.* [22] and successfully applied to road detection in [8]. The algorithm obtains an almost shadow-free image under the assumption of Planckian source of light (e.g., the sun), narrow-band imaging sensors, and Lambertian surfaces. The illuminant–invariant image $\mathfrak{I}(X)$ is a grayscale image obtained by projecting log-chromaticity values onto an *invariant-direction* θ as follows (see Fig. 4):

$$\mathfrak{I}(X) = r(X)\cos\theta + b(X)\sin\theta \tag{1}$$

where $X = \{x_1, \ldots, x_N\}$ is the set of pixels in the image, $r(X) = \log(R(X)/G(X))$ and $b(X) = \log(B(X)/G(X))$ being $R(X)$, $G(X)$, $B(X)$ the red, green, and blue color planes of the input image. The *invariant-direction* θ is device dependent, and it does not correlate with the lighting conditions. Hence, the calibration process for each camera need only to be done once using the calibration procedure described in [22].

B. *Image Partition*

The second stage aims to capture the spatial layout of the image by using dynamic image partitions. Common scene recognition approaches use spatial pyramids with fixed regions. However, using fixed regions may lead to image partitions that do not contain relevant information to describe the geometry of the road (see Fig. 2). Therefore, we propose an adaptive partitioning process to capture relevant road geometry information in each partition (see Fig. 3). Hence, we first divide the image in two parts based on the position of the horizon line. This line provides important information for inferring where the road is located in each image, and thus, the lower partition

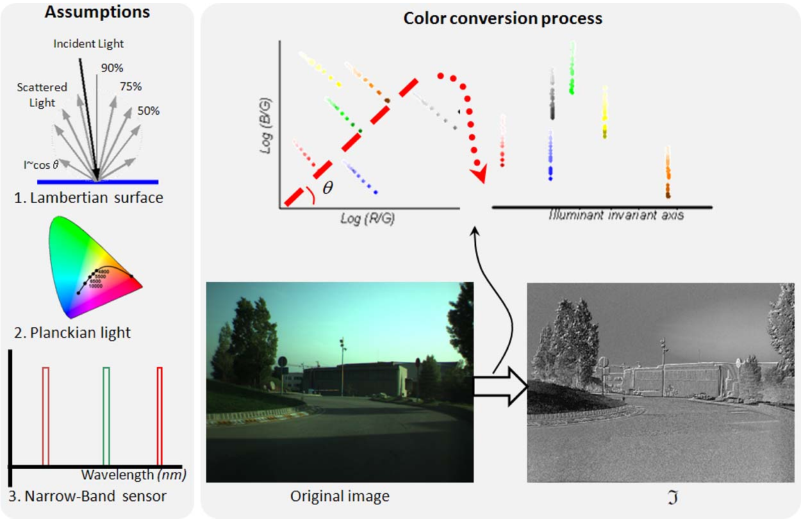


Fig. 4. Illuminant–invariant image $\mathfrak{I}$ is a grayscale image that is obtained from projecting the $\log(R/G)$, $\log(B/G)$ pixel values of the incoming data onto the direction orthogonal to the lighting change lines, i.e., invariant-direction.

encapsulates all the information on the road geometry. Then, an additional partition is performed to distinguish between left and right turns (see Fig. 3). That is, the image located below the horizon is partitioned into two characteristic road regions. In particular, for the first partition, we consider the horizon line estimation approach[1] proposed by Sivic *et al.* [23]–[25]. This method describes the image using GIST descriptors [19] and then estimates the horizon line by applying nonlinear mixtures of linear regressors to the description of the image. Example results of the horizon line detection algorithm are shown in Fig. 5. As shown, the position of the horizon line is correctly detected in different scenarios. Further, the lower partition encapsulates all the information about the geometry of the road. For the second partition, we divide the lower part of the image in half, as shown in Fig. 6.

C. Image Description

Finally, the third stage aims to obtain a compact and consistent representation of the image. This way, we consider global image descriptors applied to each partition. The strength of the approach relies on capturing information independent of specific objects present in each image block. Hence, we consider, for each partition, its holistic representation, as proposed in [19]. The idea of this holistic approach is characterizing images without explicitly detecting or recognizing objects in the scene. Following that approach, the holistic representation of each partition is obtained in three steps (see Fig. 3). First, the spectral information (edges) is extracted using a bank of Gabor filters. Second, each partition is described using a GIST vector [19] by dividing each region into a nonoverlapping 4 × 4 grid and averaging the spectral responses. Finally, the global description of the image is obtained by aggregating descriptors of each partition. As a result, we obtain an image descriptor with 1600 components: 320 × 5 (i.e., 1 for the whole image, 2 parts at level 1, and 2 parts at level 1, since the invariant image is a grayscale image).

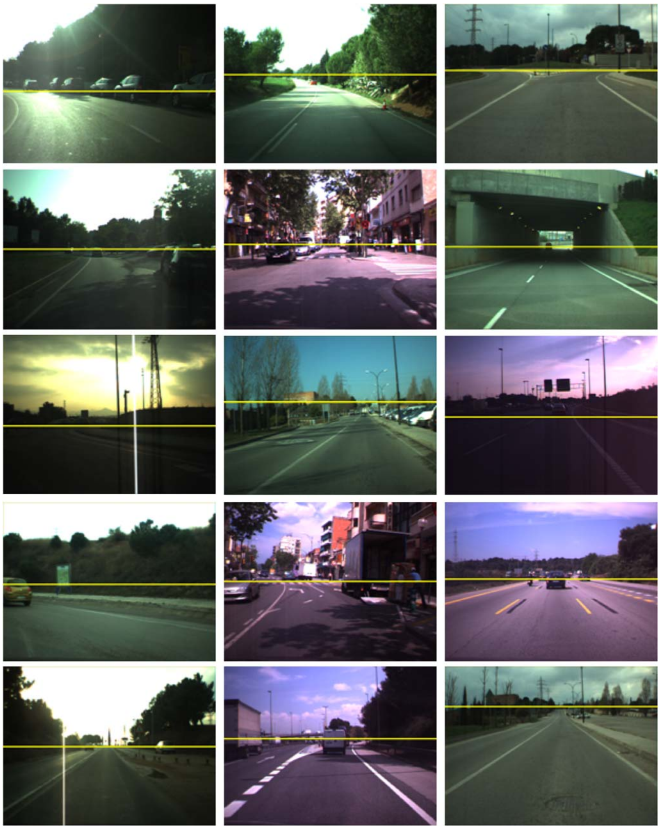


Fig. 5. Example horizon line estimation results. These results show the robustness of the algorithm to different situations such as lighting variations and road types.

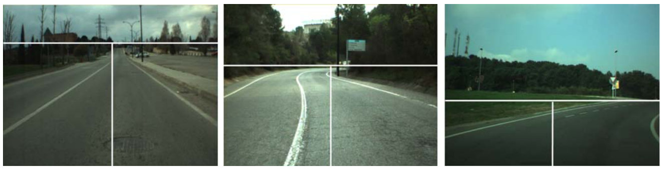


Fig. 6. Partition of the image is steered by the position of the horizon line to encapsulate all the road information, and then, the part below the horizon is divided in half to distinguish between left and right turns.

[1] code available at http://labelme.csail.mit.edu/LabelMeToolbox/index.html

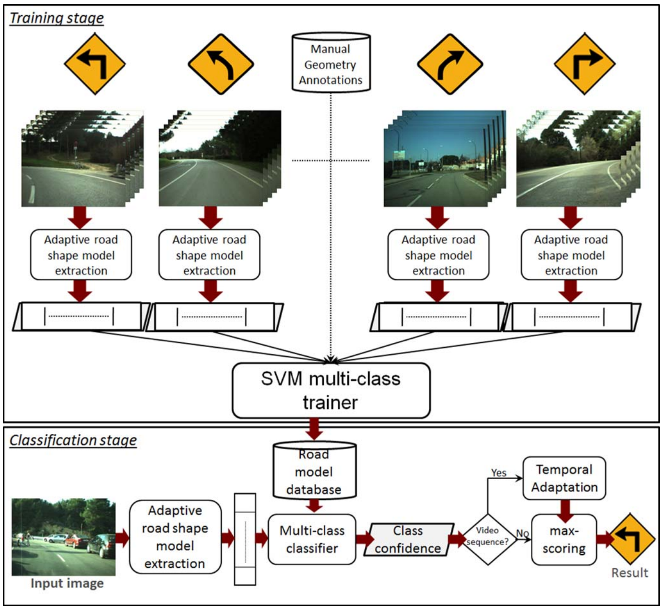

Training stage
Manual Geometry Annotations
Adaptive road shape model extraction
Adaptive road shape model extraction
Adaptive road shape model extraction
Adaptive road shape model extraction
SVM multi-class trainer
Classification stage
Road model database
Input image
Adaptive road shape model extraction
Multi-class classifier
Class confidence
Video sequence?
Yes
No
Temporal Adaptation
max-scoring
Result

Fig. 7. Proposed algorithm for road geometry classification based on adaptive road shape models. Road models are learned in a training stage and then inferred in the classification stage.

IV. Road Geometry Classification via Adaptive Road Shape Models

Here, we propose an algorithm for road geometry classification based on adaptive road shape models. The outline of the algorithm is devised in Fig. 7. During training, adaptive road models are extracted from different images, and a multiclass support vector machine (SVM) classifier is trained using these descriptions and their manual geometry annotations. Then, in the testing process, the adaptive road model description is computed for each incoming image and used as input to the multiclass SVM classifier. The output of the classifier is a confidence (in the range [0, 1]) for each possible geometry in the database (e.g., left turn, right turn, straight road, and so on). Finally, a semantic geometry label (e.g., straight road, left turn, right turn) is assigned by selecting the highest scoring class.

For video sequences, we use a probabilistic framework to exploit the temporal coherence between road models computed in consecutive frames. In particular, in this paper, we consider an HMM to model this temporal coherence. The hidden states of the model are the true class of the incoming image, and the observed evidence is the confidence provided by the multiclass classifier. Then, edges between hidden states are the temporal dependence values between classes (see Fig. 8). Then, the goal of the algorithm is

$$\arg\max_{C} P(C|M) = \arg\max_{C} P(M|C)P(C) \tag{2}$$

where $C = \{C_1 \ldots C_N\}$ is the classification vector of a sequence of N images. This vector contains the true class label. Further, $M = \{M_1 \ldots M_N\}$ is the evidence vector (real valued) for these images [26]. To perform the system online, that is, without considering the complete sequence of images, we use a fixed-lag smoothing approach [27]. Furthermore, transition probabilities are learned by manually counting class transitions in the training set.

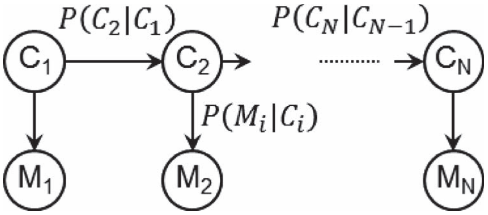


Fig. 8. HMM is used to exploit the inherent correlation between geometry classes in consecutive frames. Class transition probabilities are learned by counting transitions in the training set.

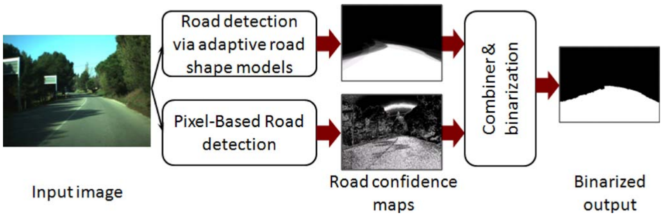

Road detection via adaptive road shape models
Pixel-Based Road detection
Combiner & binarization
Input image
Road confidence maps
Binarized output

Fig. 9. Proposed road detection algorithm based on adaptive road shape models and low-level road information.

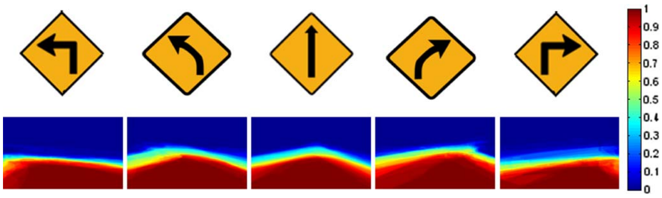


Fig. 10. Road detection via geometry classification is performed by associating a road confidence map to each semantic label.

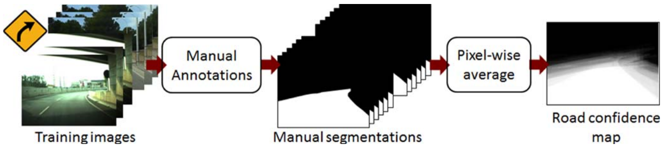


Fig. 11. Road confidence map providing a rough detection of the road is associated to each road model. These confidence maps are learned by combining manual segmentations of images in the training set of each model.

V. Road Detection via Adaptive Road Shape Models

A direct application of road geometry classification using adaptive road shape models is detecting the road surface ahead of a moving vehicle. Inferring the road geometry provides strong prior information regarding the location of the road in an image. Therefore, here, we propose a road detection algorithm based on adaptive road shape models (see Fig. 9). As shown, the algorithm combines the road information provided by road geometry classification with the pixel-level information provided by low-level road detection algorithms.

The first part of the approach consists in extending the road geometry classifier algorithm to obtain road confidence maps. The main idea behind this extension is associating a road confidence map to each semantic label (e.g., left turn, right turn, straight road, strong left turn) available (see Fig. 10). These confidence maps provide a rough description of the road ahead the moving vehicle and are learned from training images as follows (see Fig. 11): First, road areas in each training image are manually segmented. Then, binary annotations for each

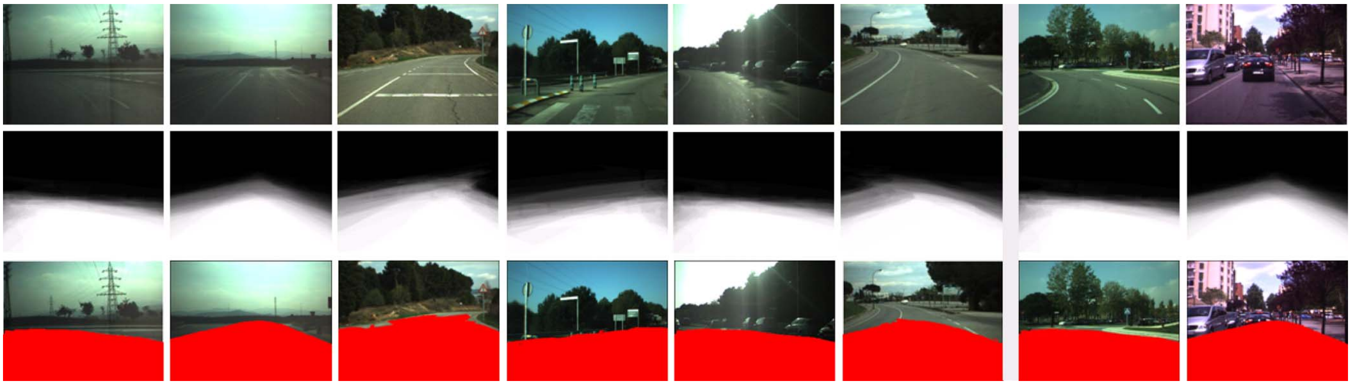


Fig. 12. Example road detection results by road geometry classification. (a) Original images. (b) Road confidence maps assigned by the algorithm considering the max-scoring class. (c) Binarized result overlapping the original image. These binary masks are obtained using 0.5 as threshold. That is, a road label is assigned to a pixel if that pixel is a road pixel in at least half of the training images for that class.

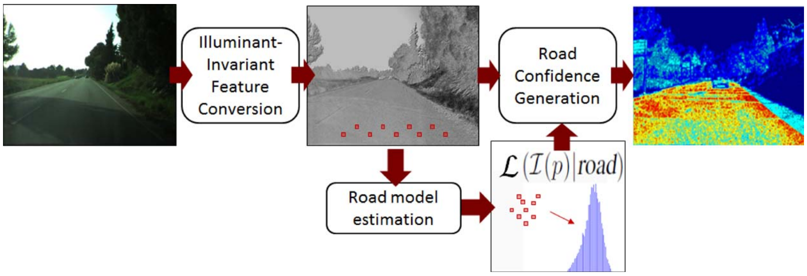

Illuminant-Invariant Feature Conversion
Road Confidence Generation
Road model estimation
$\mathcal{L}(\mathcal{I}(p)|road)$

Fig. 13. Pipeline used to obtain a road confidence map based on pixel-level features. This pipeline is based on the illuminant–invariant algorithm proposed in [8].

semantic class are pixelwise averaged to obtain the confidence map associated to each class. Finally, given a new image, we perform road geometry classification and assign the confidence map associated to the semantic label exhibiting the highest score.

Example road detection results by road geometry classification are shown in Fig. 12. Each image is classified, and a road confidence map is assigned according to the max-scoring class. As shown, the algorithm provides a rough detection of the road in different lighting conditions and road types. However, this rough detection does not provide the required pixel-level accuracy. Therefore, the proposed algorithm combines this result with a pixel-based approach.

The second block consists in obtaining a road confidence map based on low-level (e.g., pixel-level) features. We consider the illuminant–invariant approach [8], as shown in Fig. 13. Following this approach, images are first converted into the illuminant–invariant feature space [see (1)]. Then, a training set is used to build a model that provides insight for predicting the label (road or background) of each image pixel. The training set consists of the surrounding areas of several pixels placed at the bottom part of the image, and the model is the normalized histogram of these pixels. Hence, the histogram is used as a likelihood function indicating the support of each bin (possible pixel values) depicting road surface. Finally, a confidence map is computed by mapping each pixel value to this road likelihood.

Finally, the last step consists in combining the confidence map associated with the road shape model and the confidence map obtained using a pixel-based road detection algorithm (see Fig. 9). The former provides a general view of the road geometry, whereas the latter provides the accuracy required. These confidence maps are continuous valued and can be interpreted as the pixel potential for being a road pixel. Then, these outputs are combined using a weighted harmonic mean as follows:

$$RP(x) = \sum_i w_i \left(\frac{1}{2} \sum_i \frac{w_i}{s_i(x)} \right)^{-1} \qquad (3)$$

where RP is the final road confidence, s_i is the road confidence assigned by the ith classifier ($i \in \{$geometry model, appearance$\}$), and w_i is the weight (relevance) assigned to each of these classifiers. $RP(x)$ ranges from 0 to 1. The higher the value, the more likely the pixel x belongs to the road. Finally, the road mask is obtained by thresholding $RP(x)$ using a fixed threshold λ.

VI. Experiments

Experiments to validate the proposed method are conducted on a large-scale data set of road image sequences acquired using an onboard camera with the Sony ICX084 sensor. This is a charge-coupled device chip of 640 × 480 pixels and 8 bits per pixel that makes use of a Bayer pattern for collecting color information. Standard Bayer pattern decoding (bilinear interpolation) is used to obtain RGB color images. The camera is equipped with a microlens of 6-mm focal length. The frame acquisition rate is 15 ft/s. The data set consists of thousands of images taken on different days, at different times of the day, and in different scenarios. Thus, images exhibit different backgrounds, different lighting conditions and shadows, and the presence of other vehicles due to different traffic situations.

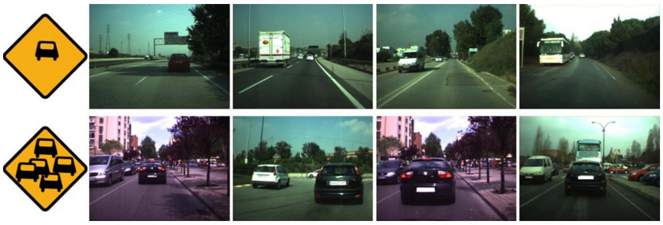
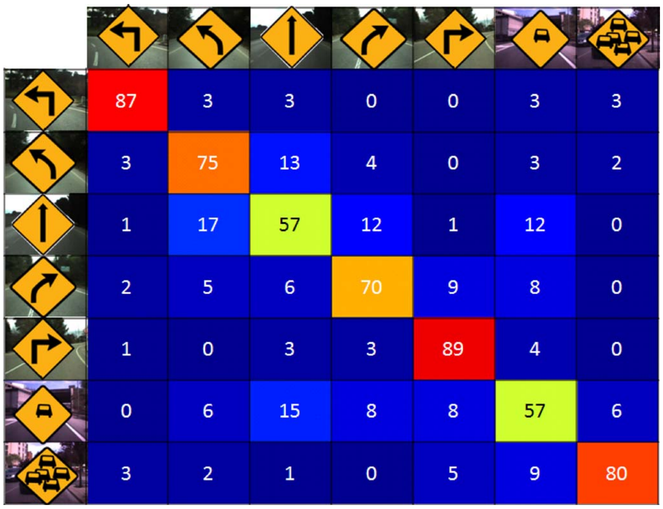


Fig. 14. Road models (see Fig. 10) are extended with two road models containing different traffic situations.



87	3	3	0	0	3	3
3	75	13	4	0	3	2
1	17	57	12	1	12	0
2	5	6	70	9	8	0
1	0	3	3	89	4	0
0	6	15	8	8	57	6
3	2	1	0	5	9	80

Fig. 15. Confusion matrix for S1 (training/testing subset) including the illuminant–invariant feature space in the pipeline. The average recognition rate is 73.57% ± 13.1. Each road model has the same occurrence.

TABLE I
SUMMARY OF ROAD SCENE CLASSIFICATION METHODS USED IN THE FIRST EXPERIMENT

	Features	Spatial Layout	Descriptors	Descriptor Size
Our proposal using $\mathfrak{I}$	$\mathfrak{I}$	3–levels/steered	Gist	1600
Our proposal (without $\mathfrak{I}$)	gray–level	3–levels/ steered	Gist	1600
$\mathfrak{I}$–GIST (Single level)	$\mathfrak{I}$	none	Gist	320
GIST [19] (Single level)	gray–level	none	Gist	320
Dense OpponentSIFT [15]	opponent color	none	SIFT	75264
SIFT (Single level)	gray–level	none	SIFT	bag–of–words
SIFT [16] (Spatial pyramid)	gray–level	3–levels/ fixed regions	SIFT	bag–of–words
PHoG [18]	gray–level	3–levels/ fixed regions	HoG	680

TABLE II
AVERAGE CLASSIFICATION RATE OF DIFFERENT ALGORITHMS AND DIFFERENT INSTANTIATIONS OF OUR APPROACH. OMITTED NUMBERS REFER TO ALGORITHMS WITH VERY LOW PERFORMANCE

	Training/testing S1	Extended test S2
Our proposal (using $\mathfrak{I}$)	73.57% ± 13.1	44,3%
Our proposal without $\mathfrak{I}$	74.57% ± 14.0	31,3%
$\mathfrak{I}$–GIST (Single level)	8.85% ± 15.2	−−
GIST [19] (Single level)	7.57% ± 13.0	−−
Dense OpponentSIFT [15]	14.00% ± 20.9	9.8%
SIFT (Single level)	9.53% ± 18.4	−−
SIFT [16] (Spatial pyramid)	20.30% ± 16.2	14.8%
PHoG [18]	14.29% ± 38.8	10.4%

TABLE III
CONTINGENCY TABLE USED TO COMPUTE MCNEMAR'S TWO-SIDED SIGNIFICANCE TEST (GIVEN SIGNIFICANCE OF 0.025) IN THE S1 SUBSET OF IMAGES. THE TEST IS USED TO COMPARE THE EFFECT OF INCLUDING THE ILLUMINANT–INVARIANT FEATURE SPACE. THE RESULT REVEALS THAT THE DIFFERENCE IN PERFORMANCE SHOWN IN THE FIRST TWO ROWS IN TABLE II ARE NOT SIGNIFICANT. SUCCEEDED REFERS TO IMAGES CORRECTLY CLASSIFIED, AND FAILED REFERS TO CLASSIFICATION ERRORS

		Using $\mathfrak{I}$	
		Failed	Succeeded
Without $\mathfrak{I}$	Failed	56 (8.0%)	122 (17.4%)
	Succeeded	129 (18.4%)	393 (56.2%)

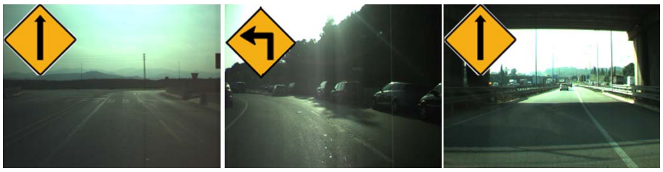


Fig. 16. Example of images correctly classified using the proposed approach. These images are miss-classified if the illuminant–invariant feature space is not considered.

A. Road Geometry Classification in Still Images

The first experiment consists in evaluating the road classification system. To this end, a first subset of 2000 images is annotated into one of the seven different prototypes. Five of these shapes refer to typical road geometries: strong left turns, left turns, straight roads, right turns, and strong right turns, as shown in Fig. 1(a). The other two types refer to common traffic situations (see Fig. 14). The subset of images is subdivided again into two parts: S1, where 700 images are used as training/testing set, and S2, where the rest of the images are considered for extended testing. Multiclass classification is performed with an SVM. In particular, we use the LibSVM implementation [28]. The position of the horizon line is estimated using the approach by Sivic *et al.* [23].

The performance of the algorithm to recognize road geometries is evaluated using S1 (700 images). Evaluations are repeated ten times by randomly selecting the same number of images from each class providing the same relative occurrence (i.e., prior probability) for each model in the subset. We record the average per-class recognition rate for each run, and the final result is reported as their mean and standard deviation. Classification rates given by the total number of correctly classified images divided by the total number of images are shown in Fig. 15. These results indicate that a number of classes are robustly detected. This contingency table suggests that miss-classifications are mainly located in straight and soft left and right turns due to strong similarity between these types of images. Further, the algorithm exhibits lower performance in classifying traffic images due to the higher intraclass variability in these two classes.

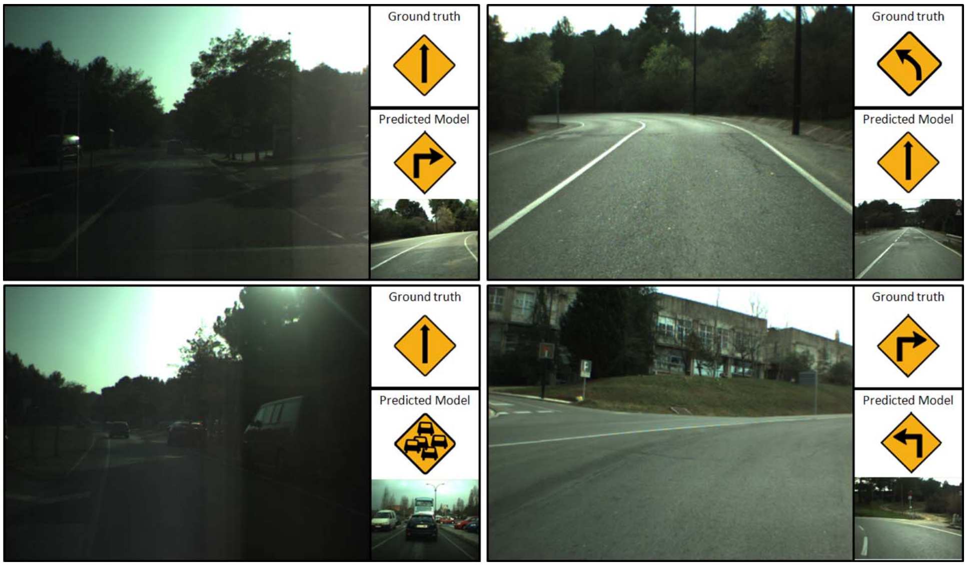


Fig. 17. Example of miss-classified images. The upper right part of each image shows the ground truth. The predicted label is shown at the bottom right part of the image. These particular examples show unclear road classes.

B. Comparison to State of the Art

Now, the performance of the proposed algorithm is compared with other state-of-the-art algorithms. The summary of these methods is listed in Table I. The first scene classification algorithm adapted to roads is from [15]. This approach extracts SIFT features using dense sampling in the opponent color space. The second approach uses SIFT-based spatial pyramids [16]. The third and fourth approaches are based on GIST descriptors using a single level (i.e., the entire image). The third approach uses a gray-level image, and the fourth approach uses illuminant–invariant feature space. The fifth algorithm is based on the global Pyramid of Histogram of Gradient Orientation (PHoG) descriptor, as described in [18]. Finally, two different instances of the proposed pipeline are evaluated. The first instance uses the steered pyramid on a grayscale image (i.e., without considering the illuminant–invariant feature space). The second instance uses the steered pyramid on the illuminant–invariant feature space. Pairwise comparisons between algorithms are conducted using McNemar's test [29]. For completeness, additional comparisons between algorithms are performed. The comparison consists in training these classifiers using all 700 images in the S1 subset. Then, images in the S2 subset are used as a testing set. These images contain more lighting variability than the images in S1. Classification rates of each algorithm are shown in Table II. The results reveal poor performance for state-of-the-art methods. In the case of dense sampling, the poor performance is mainly due to variation of spatial information. That is, these algorithms require road edges located in the same grid position.

For the S1 subset, the illuminant–invariant feature space provides similar performance as using a gray-level image. In fact, the differences in performance are not significant according to McNemar's test (see Table III). However, there is a significant difference when the algorithm is applied to the extended testing set (S2). The drop-off in performance is mainly due to more challenging situations (shadows and lighting variations). In this case, using the illuminant–invariant feature space clearly outperforms the gray-level instantiation of the algorithm. The invariant feature space reduces the influence of shadows and lighting variations, thus improving the robustness against real-world driving situations. Example results of correctly classified images using the illuminant–invariant but miss-classified images for gray-level are shown in Fig. 16. Finally, examples of miss-classified images are shown in Fig. 17. These images reveal situations with which it is difficult to associate a specific class.

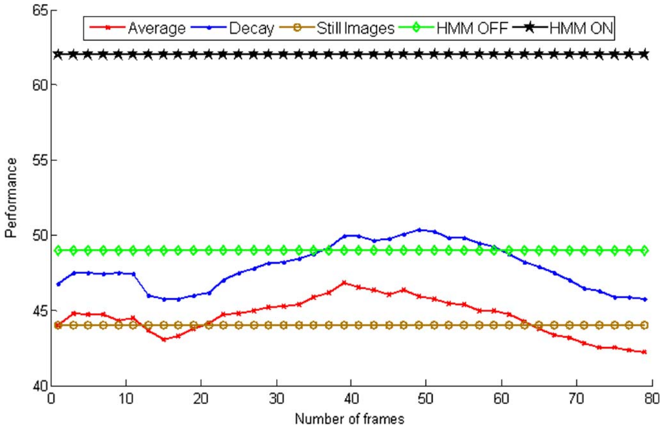


Fig. 18. Average classification rate when temporal information is considered. Average consists in averaging the output of the classifier over a period of time. Decay consists in using a decay factor to provide more relevance to recent observations than distant observations. HMM off is the proposed framework using a uniform transition matrix. HMM on is the proposed method.

C. Temporal Information

Here, the evaluation consists in quantifying the effect of including temporal information. To this end, the proposed approach is compared with two different methods based on integrating the obtained probabilities for each class over periods of time. The former uses uniform weighting to consider each previous frame. The latter uses a decay factor that weights the

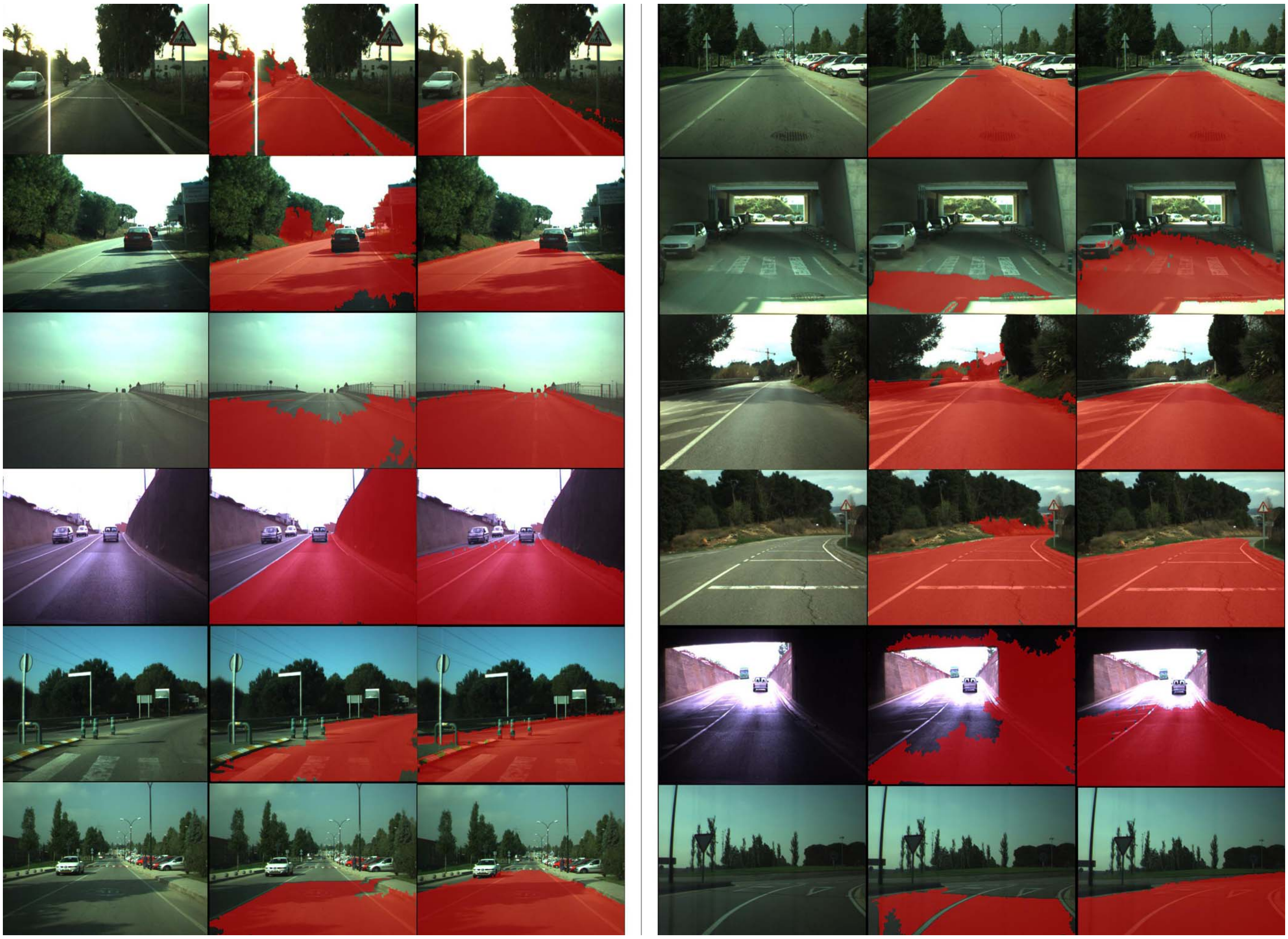


Fig. 19. Example road detection results. From left to right in each column: original image; result using the pixel-based classifier; final result combining road shape models and appearance-based information.

contribution of each past frame. Hence, more recent observations receive higher weights than older observations. Finally, two different instances of our method are considered: using a uniform transition matrix (e.g., without HMM) and learning the transition matrix of the HMM from training sequences. The parameters of the HMM (i.e., the transition matrix) are learned from labeled data by counting the number of transitions from one road geometry to another. Fig. 18 shows the effects of including temporal information into the classifier. As shown, using the proposed probabilistic framework, the performance increases up to 62%.

D. Road Detection Using Adaptive Road Shape Models

The goal of the last experiment is to evaluate the proposed road detection algorithm based on temporal road models (see Section V). The weights of the combination are fixed to 0.6 for the road shape and 0.4 for the pixel-based information. Example results are shown in Fig. 19. As shown, including road shape information improves road detection results provided by appearance-based algorithms. Further, an evaluation is performed on 500 randomly selected images. Ground truth for each image is manually generated. As a result, combining shape and appearance increases the *effectiveness* by 15% over appearance alone. From these results, it can be derived that road shape models provide relevant information for road detection.

VII. Conclusion

In this paper, a road geometry classification system has been proposed. The system uses adaptive shape models in which spatial pyramids are controlled by the inherent spatial structure of road images. To further reduce the influence of lighting variations, invariant features and temporal information have been used.

Large-scale experiments show that the proposed road geometry classifier yields the highest recognition rate of 73.57% ± 13.1, clearly outperforming other state-of-the-art methods. Including road shape information improves road detection results over existing appearance-based methods. Invariant features and temporal information provide robustness against disturbing imaging conditions.

References

[1] A. Lookingbill, J. Rogers, D. Lieb, J. Curry, and S. Thrun, "Reverse optical flow for self-supervised adaptive autonomous robot navigation," *Int. J. Comput. Vis.*, vol. 74, no. 3, pp. 287–302, Sep. 2007.

[2] C. Thorpe, M. Hebert, T. Kanade, and S. Shafer, "Vision and navigation for the Carnegie-Mellon Navlab," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 10, no. 3, pp. 362–373, May 1988.

[3] Z. Sun, G. Bebis, and R. Miller, "On-road vehicle detection: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 5, pp. 694–711, May 2006.

[4] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 743–761, Apr. 2012.

[5] M. Sotelo, F. Rodriguez, and L. Magdalena, "VIRTUOUS: Vision-based road transportation for unmanned operation on urban-like scenarios," *IEEE Trans. Intell. Transp. Syst.*, vol. 5, no. 2, pp. 69–83, Jun. 2004.

[6] Y. He, H. Wang, and B. Zhang, "Color-based road detection in urban traffic scenes," *IEEE Trans. Intell. Transp. Syst.*, vol. 5, no. 24, pp. 309–318, Dec. 2004.

[7] J. M. Alvarez, T. Gevers, and A. M. Lopez, "Learning photometric invariance from diversified color model ensembles," in *Proc. IEEE CVPR*, Miami, FL, 2009, pp. 565–572.

[8] J. M. Alvarez and A. Lopez, "Road detection based on illuminant invariance," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 1, pp. 184–193, Mar. 2011.

[9] P. Lombardi, M. Zanin, and S. Messelodi, "Switching models for vision-based on-board road detection," in *Proc. IEEE Int. ITSC*, Vienna, Austria, 2005, pp. 67–72.

[10] C. Rasmussen, "Grouping dominant orientations for ill-structured road following," in *Proc. IEEE CVPR*, Washington, DC, 2004, pp. I-470–I-477.

[11] T. Michalke, R. Kastner, M. Herbert, J. Fritsch, and C. Goerick, "Adaptive multi-cue fusion for robust detection of unmarked inner-city streets," in *Proc. IEEE IV Symp.*, Jun. 2009, pp. 1–8.

[12] J. M. Alvarez, T. Gevers, and A. M. Lopez, "3D scene priors for road detection," in *Proc. IEEE CVPR*, Jun. 2010, pp. 57–64.

[13] H. Kong, J. Audibert, and J. Ponce, "Vanishing point detection for road detection," in *Proc. IEEE CVPR*, Miami, FL, 2009, pp. 96–103.

[14] D. Hoiem, A. A. Efros, and M. Hebert, "Recovering surface layout from an image," *Int. J. Comput. Vis.*, vol. 75, no. 1, pp. 151–172, Oct. 2007.

[15] J. M. Alvarez, T. Gevers, and A. M. Lopez, "Vision based road detection using road models," in *Proc. IEEE ICIP*, Cairo, Egypt, 2009, pp. 2073–2076.

[16] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE CVPR*, Jun. 2006, pp. 2169–2178.

[17] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[18] A. Bosch, A. Zisserman, and X. Munoz, "Representing shape with a spatial pyramid kernel," in *Proc. ACM CIVR*, New York, 2007, pp. 401–408.

[19] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, May/Jun. 2001.

[20] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek, "Evaluating color descriptors for object and scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1582–1596, Sep. 2010.

[21] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE CVPR*, San Diego, CA, 2005, vol. 1, pp. 886–893.

[22] G. Finlayson, S. Hordley, C. Lu, and M. Drew, "On the removal of shadows from images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 1, pp. 59–68, Jan. 2006.

[23] J. Sivic, B. Kaneva, A. Torralba, S. Avidan, and W. T. Freeman, "Creating and exploring a large photorealistic virtual space," in *Proc. 1st IEEE CVPRW*, Anchorage, AK, Jun. 2008, pp. 1–8.

[24] D. Hoiem, "Seeing the world behind the image: Spatial layout for 3D scene understanding," Ph.D. dissertation, Robotics Inst., Carnegie Mellon Univ., Pittsburgh, PA, Aug. 2007.

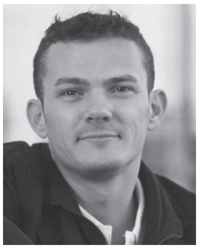
[25] A. Torralba and P. Sinha, "Statistical context priming for object detection," in *Proc. IEEE ICCV*, Vancouver, BC, Canada, 2001, vol. 1, pp. 763–770.

[26] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[27] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*. Upper Saddle River, NJ: Pearson, 2003.

[28] C. C. Chang and C. J. Lin (2011, Apr.). LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* [Online]. vol. 2, no. 3 , pp. 1–27. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[29] E. Alpaydin, *Introduction to Machine Learning (Adaptive Computation and Machine Learning)*. Cambridge, MA: MIT Press, 2004.



José M. Álvarez received the M.Sc. degree in computer science from La Salle School of Engineering, Barcelona, Spain, in 2005 and the Ph.D. degree from the Universitat Autònoma de Barcelona in 2010.

He is currently a researcher with NICTA, Canberra, Australia. He was a Postdoctoral Researcher with the Advanced Driver Assistance Systems Group, Computer Vision Center, Barcelona. In 2011, he was a Visiting Researcher with the Computational and Biological Learning Group, New York University, New York, NY. His main research interests include road detection, color, photometric invariance, machine learning, and fusion of classifiers.

Theo Gevers (M'12) received the Ph.D. degree in Computer Science from the University of Amsterdam, The Netherlands, in 1996 for a thesis on color image segmentation and retrieval.

He is an Associate Professor of computer science and the Teaching Director of the M.Sc. degree program in Artificial Intelligence with the University of Amsterdam, Amsterdam, The Netherlands. His main research interests include the fundamentals of content-based image retrieval, color image processing, and computer vision, specifically in the theoretical foundation of geometric and photometric invariants.

Mr. Gevers currently holds a VICI-award (for excellent researchers) from the Dutch Organisation for Scientific Research. He is the Chair of various conferences and is an Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING. He is a Program Committee Member of a number of conferences and an Invited Speaker at major conferences. He is a Lecturer of postdoctoral courses given at various major conferences (Computer Vision and Pattern Recognition, International Conference on Pattern Recognition, International Society for Optics and Photonics, and Computer Graphics, Imaging and Visualization).

Ferran Diego received the higher engineering degree in telecommunications from the Polytechnic University of Catalonia (UPC), Barcelona, Spain, in 2005 and the M.Sc. degrees in language and speech from UPC and the University of Edinburgh, Edinburgh, U.K., in 2005 and in computer vision and artificial intelligence from the Computer Vision Center, Barcelona, and the Universitat Autònoma de Barcelona in 2007. He is currently working toward the Ph.D. degree with the Universitat Autònoma de Barcelona.

His research interests include video alignment, optical flow, sensor fusion, machine learning, and image retrieval.

Antonio M. López received the B.Sc. degree in computer science from the Universitat Politècnica de Catalunya, Barcelona, Spain, in 1992 and the M.Sc. degree in image processing and artificial intelligence in 1994 and the Ph.D. degree in 2000 from the Universitat Autònoma de Barcelona (UAB).

Since 1992, he has been lecturing with the Department of Computer Science, UAB, where he is currently an Associate Professor. In 1996, he participated in the foundation of the Computer Vision Center, UAB, where he has held different institutional responsibilities, presently with the Advanced Driver Assistance Systems research group. He has since been responsible for public and private projects. He is a coauthor of more than 100 papers, all in the field of computer vision.