

Text Detection and Recognition on Traffic Panels From Street-Level Imagery Using Visual Appearance

Álvaro González, Luis M. Bergasa, *Member, IEEE*, and J. Javier Yebes

Abstract—Traffic sign detection and recognition has been thoroughly studied for a long time. However, traffic panel detection and recognition still remains a challenge in computer vision due to its different types and the huge variability of the information depicted in them. This paper presents a method to detect traffic panels in street-level images and to recognize the information contained on them, as an application to intelligent transportation systems (ITS). The main purpose can be to make an automatic inventory of the traffic panels located in a road to support road maintenance and to assist drivers. Our proposal extracts local descriptors at some interest keypoints after applying blue and white color segmentation. Then, images are represented as a “bag of visual words” and classified using Naïve Bayes or support vector machines. This visual appearance categorization method is a new approach for traffic panel detection in the state of the art. Finally, our own text detection and recognition method is applied on those images where a traffic panel has been detected, in order to automatically read and save the information depicted in the panels. We propose a language model partly based on a dynamic dictionary for a limited geographical area using a reverse geocoding service. Experimental results on real images from Google Street View prove the efficiency of the proposed method and give way to using street-level images for different applications on ITS.

Index Terms—Bag of visual words (BOVW), computer vision, traffic panel detection, traffic panel inventory, traffic panel recognition.

I. INTRODUCTION

THIS paper presents a real application to intelligent transportation systems (ITS) of a method to detect and recognize text in images taken from natural scenarios proposed by the same authors in [1]. This text reading algorithm has proved to be robust in many kinds of real-world scenarios, including indoor and outdoor places with a wide variety of text appearance due to different writing styles, fonts, colors, sizes, textures and layouts, as well as the presence of geometrical distortions, partial occlusions, and different shooting angles that may cause deformed text. In this paper, this algorithm is applied, including some modifications and new functionalities, to read the information contained in traffic panels using the

images served by Google Street View. The aim of this work is, in the first place, to detect traffic panels and to recognize the information inside them, showing that the text detection and recognition method proposed in [1] can be generalized to other scenarios, which are completely different to those that have been tested, without needing to retrain the system. In the second place, we want to develop an application that enables the creation of up-to-date inventories of traffic panels of regions or countries that facilitate traffic signposting maintenance and driver assistance.

In this paper, we focus on traffic panels in the Spanish territory for two main reasons. First, unlike other countries, the coverage of Street View in Spain is near complete; thus, we can create a huge and diverse data set of images. Second, as far as we know, there is not any official database of all the traffic panels in Spain; thus, there are more possibilities that any government or institution responsible for managing the road network can be interested in having an up-to-date inventory of the traffic panels in Spain with the method here proposed. The reasons for which these organizations may be interested are various. Having a centralized database of all the traffic panels supposes a rapid and economic way of evaluating and analyzing the potential dangerous situations that may arise due to traffic panels that suffer from a bad visibility or show deteriorated or outdated information. Street-level panoramic image recording services, such as Street View, which have become very popular in the recent years and have reached a huge coverage of the road network, suppose a potential source to rapidly know the state of the vertical signposting of the road network, particularly when the street-level images are updated regularly. Computer vision techniques applied on this kind of images simplify and speed up the creation of traffic signposting inventories, minimizing the human interaction. In addition, these inventories can be useful not only for supporting maintenance but also for developing future driver assistance systems. In general, automatic text reading may be helpful to support drivers or autonomous vehicles to find a certain place by simply reading and interpreting street signs, road panels, variable-message signs, or any kind of text present in the scenario, when global positioning systems suffer from lack of coverage, particularly in high-density urban areas. Advanced driver assistance systems could also benefit from text recognition for automatic traffic signs and panels identification.

However, traffic panel detection still remains a challenging problem due to several reasons. First, there is a huge variability of traffic panels as each of them depicts different information, varying in size, color, and shape. Moreover, there are large viewpoint deviations due to the fact that the images are captured from a driving vehicle. There may be also occlusions due

Manuscript received February 5, 2013; revised May 7, 2013 and July 30, 2013; accepted August 2, 2013. Date of publication August 26, 2013; date of current version January 31, 2014. This work was supported in part by the Ministerio de Economía y Competitividad through the project Smart Driving Applications under Grant TEC2012-37104 and in part by the Comunidad de Madrid through the project Robocity2030 under Grant CAM-S-0505/DPI000176. The Associate Editor for this paper was M. Á. Sotelo Vázquez.

The authors are with the Department of Electronics, Polytechnic School, University of Alcalá, 28871 Alcalá de Henares, Spain (e-mail: alvaro.g.arroyo@depeca.uah.es; bergasa@depeca.uah.es; javier.yebes@depeca.uah.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.2013.2277662

to vegetation or other road users. In addition, weather and illumination conditions are a key problem in any kind of vision-based system. Apart from this, many elements in the roads or beside them can be easily confused with traffic panels, such as advertisement panels or truck bodies.

The rest of the paper is organized as follows. Previous attempts of developing systems to recognize the information depicted in traffic panels are explained in Section II. An overview of the data set created for this application is shown in Section III. The traffic panel detection method using color masks and bag of visual words (BOVW) is detailed in Section IV. A brief overview of the text detection and recognition method to extract the information contained in the traffic panels is presented in Section V. Experimental results and main conclusions are displayed in Sections VI and VII, respectively.

II. RELATED WORK

Traffic sign detection and recognition using computer vision techniques has been an active area of research over the past decade. A good survey about the main vision-based proposals of the state of the art for intelligent driver assistance systems can be found in [2], where a discussion about the future perspectives of this research line is there included. Additionally, the work in [3] presents a recent contribution about an intelligent road sign inventory based on image recognition, which is related to the application we propose in this paper, but for traffic signs instead of traffic panels and using images taken from a vehicle instead of images served by Google Street View.

Traffic panel detection and recognition has been out of the scope of researchers because, on the one hand, they are informative signs and then they have less priority than the regulatory or warning signs. On the other hand, there is a wide diversity of information contained in traffic panels, which is difficult to analyze. In conclusion, to date, there has not been much research on automatic detection and recognition of the information contained in road panels. From our knowledge, apart from a previous work of the authors in [4] where an automatic traffic signs and panels inspection system using active vision at night is presented, only two works have been developed in this matter.

The work proposed in [5] extracts candidates to be traffic panels, using a method that detects blue and white areas in the image using the hue and saturation components of the hue–saturation–intensity space. Then, candidates are classified according to their shapes, in order to extract the rectangular blobs. This is done through a method that correlates the radial signature of their fast Fourier transform with a pattern corresponding to an ideal rectangular shape. Then, panel reorientation is carried out using a homography that aligns the four vertices of each blob. Once the panels have been detected and reoriented, segmentation of the foreground objects from the background of the panel is done by analyzing the chrominance and luminance histograms. Connected components labeling and position clustering is finally done for the arrangement of the different characters on the panels. This algorithm is invariant to translations, rotations, scaling, and projective distortion, but it is severely affected by changing lighting conditions. In addition, there are many parameters and thresholds that are adjusted

ad hoc. Recognition is applied at character level, but no language model is applied to correct misspelled words. There is not any information on where and how the images are extracted. Moreover, the experimental results provided by the authors do not show any kind of performance evaluation; hence, it is impossible to know the robustness of their proposal and no comparisons are possible, as they use their own data set.

On the other hand, Wu *et al.* [6] proposes a method to detect text on traffic panels from video. First, regions of the same color are extracted using a k -means algorithm, and traffic panel candidates are detected by searching for flat regions perpendicular to the camera axis. The orientation of the candidate planes are estimated using three or more points in two successive frames; hence, this method needs an accurate tracking method to detect corresponding points in successive frames. Furthermore, a multiscale text detection algorithm is performed on each candidate traffic panel area. The text detection method integrates edge detection, adaptive searching, color analysis using Gaussian mixture models, and geometry alignment analysis. A minimum bounding rectangle is fitted to cover every detected text line. A feature-based tracking algorithm is then used to track all detected areas over the timeline as they are merged with other newly detected texts in the sequence. Finally, all detected text lines are extracted for recognition, but the authors do not comment how the recognition is carried out. In terms of text detection, this method provides good results under different lighting conditions, and it is not affected by rotations and projective distortions. It achieves an overall text detection rate of 89% in their own data set, which is not publicly available.

In this paper, we propose a novel approach to model traffic panels using visual appearance, specifically a BOVW technique from local descriptors extracted at interest keypoints, unlike the typical methods in the state of the art that use other features such as edges or geometrical characteristics. A previous color segmentation stage guides the keypoints searching in the image. The experimental results will show the effectiveness of the proposed method. The flowchart of the proposed application is shown in Fig. 1. First, the input images are downloaded from the Street View website using the air position indicator provided by Google. Then, a method based on color segmentation and the BOVW algorithm is applied on each frame to detect a traffic panel. In case a panel is detected, the text detection and recognition algorithm developed in [1], including some modifications and new functionalities, is applied, and a geolocalization method is carried out to estimate the geographic coordinates of the panel.

III. IMAGE CAPTURE AND DATA SET CREATION

The images used in this work have been obtained from the Street View service developed by Google. It provides high-resolution views from various positions along many streets and roads in the world. These images are taken at discrete geographical locations defined by a pair (LAT , LON) (latitude and longitude in decimal degrees, respectively). Each position is around 10 or 20 m from each other.

A total of 16 277 images has been extracted, and two independent subsets of images have been created, i.e., one for

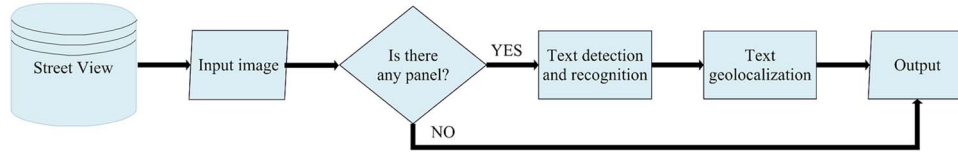


Fig. 1. Flowchart of the proposed application.



Fig. 2. Roads from which the images have been obtained: train set (red) and test set (blue).

training the system composed of 5514 images (1047 positive samples of 509 different panels and 4467 negative samples) and the other for testing the system composed of 10763 images. All the images have been obtained from street-level images of the Spanish road network, specifically from the roads shown in Fig. 2 (the train set from the roads shown in red and the test set from the roads shown in blue). We have chosen a wide variety of different situations (landscapes, weather conditions, and times of the day) for training and all the panels of two different road sections for testing. Then, some complex scenarios may occur, as it will be explained in Section VI-B. In addition, the panels may have different degrees of deterioration.

In the data set, there are two kinds of traffic panels, i.e., those with blue background and those with white background. They can be located above the road and on the right margin of the road. Table I shows the number of panels of each type in both train and test sets. Since there can be several samples of each panel taken at different distances, we also show the number of images.

IV. TRAFFIC PANEL DETECTION USING VISUAL APPEARANCE

Our proposal is to apply our text detection and recognition algorithm only on those images in which there is a traffic panel, in order to increase the efficiency of the system. For this purpose, a traffic panel detection method has been developed.

TABLE I
NUMBER OF PANELS AND IMAGES IN THE DATA SET

| | | | Train | | Test | |
|-----------|---------|-------|--------|--------|--------|--------|
| | | | Panels | Images | Panels | Images |
| Positives | Lateral | Blue | 314 | 613 | 84 | 480 |
| | | White | 35 | 68 | 24 | 87 |
| | Upper | Blue | 79 | 167 | 45 | 164 |
| | | White | 81 | 199 | 32 | 123 |
| Negatives | | | - | 4467 | - | 9909 |
| Total | | | 509 | 5514 | 185 | 10763 |

It is based on color segmentation and a BOVW approach [7]. We have chosen this technique since it has become one of the most popular in terms of classifying images. In this paper, we want to prove that BOVW is suitable to model traffic panels, despite the challenge that supposes their huge variability, and we want to show that geometrical characteristics are not strictly needed in the detection process. The diagram of blocks of our proposal is shown in Fig. 3. For both training and test images, the BOVW technique is applied only over certain areas of the image given by blue and white color masks. We will explain later in this section the reason why we are applying these color masks. However, first, we are going to briefly explain the BOVW technique.

The BOVW method stems from text analysis, wherein a document is represented by word frequencies without regard to their order. These frequencies are then used to perform document classification. The BOVW approach to image representation follows the same idea. The visual equivalent of words are local image features. Therefore, the BOVW technique models an image as a sparse vector of occurrence counts of the vocabulary of local image features. In other words, it translates a very large set of high-dimensional local descriptors into a single sparse vector of fixed dimensionality (a histogram) across all images.

First, features at some keypoints are extracted in the train images and converted into feature descriptors, which are high-dimensional vectors. Good descriptors should be able to handle intensity, rotation, scale, and affine transformations. In this paper, we compare different descriptors of the state of the art, as it will be explained later in this section. Then, the sampled features are clustered in order to quantize the space into a discrete number of visual words using k -means clustering. The visual words are the cluster centers and can be considered as a representative of several similar local regions. The image can be represented by the histogram of the visual words, which counts how many times each of the visual words occurs in the image. To account for the difference in the number of interest points between images, the BOVW histogram is normalized to have unit L1 norm. The classes or categories of the input train

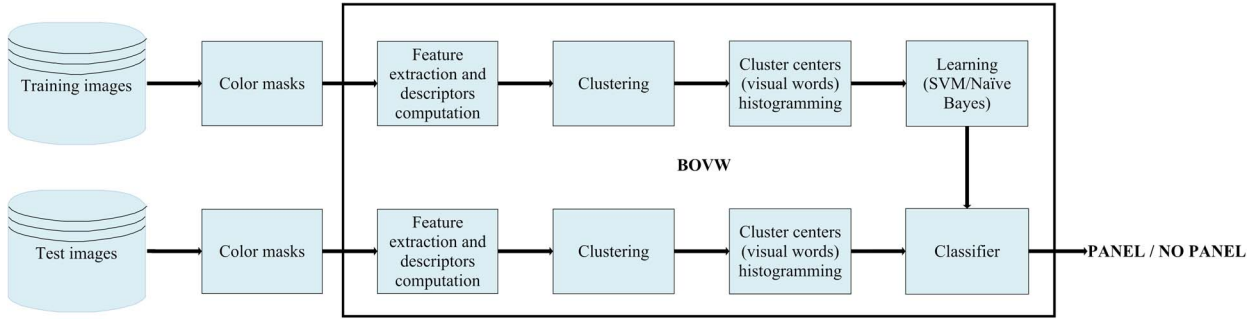


Fig. 3. Traffic panel detection.

images are learned by a classifier. In this paper, we compare two classifiers: support vector machines (SVM) [8] and Naive Bayes [9].

An SVM performs classification by constructing an N -dimensional hyperplane that optimally separates the data into two categories. The goal of SVM modeling is to find the optimal hyperplane that separates clusters of data in such a way that cases with one category of the target variable are on one side of the plane and cases with the other category are on the other side of the plane. The data points near the hyperplane are the support vectors, and the distance between the support vectors is called the margin. The optimum separation is achieved by the hyperplane that maximizes the margin, since, in general, the larger the margin is, the lower the generalization error of the classifier is. In our approach, the input data points are the histograms representing every train image.

On the other hand, the basic assumption of the Naïve Bayes model is that each category has its own distribution over the visual vocabulary and that the distributions of each category are observably different. Suppose N is the number of visual words. Let be each image represented by $m = [m_1, m_2, \dots, m_N]$, where m_i is an N -dimensional vector whose i th component measures the occurrence frequency of the i th visual word (cluster center) in the image. Let c represent the category of the image. Given a collection of training examples, the Naive Bayes classifier learns the different distributions for different categories. The classification decision is made by

$$c_{\text{MAP}} = \arg \max_c p(c|m) \quad (1)$$

which finds the class c that maximizes the posterior probability $p(c|m)$.

Applying Bayes' rule, (1) can be expressed as

$$c_{\text{MAP}} = \arg \max_c \frac{p(m|c)p(c)}{p(m)}. \quad (2)$$

$p(m)$ can be dropped out, and assuming that the distributions on each category are independent, (2) reduces to

$$c_{\text{MAP}} = \arg \max_c p(c) \prod_{n=1}^N p(m_n|c). \quad (3)$$

$p(c)$ is the prior probability of class c .

Given a test image, the nearest visual word is identified for each of its features, using the Euclidean distance between

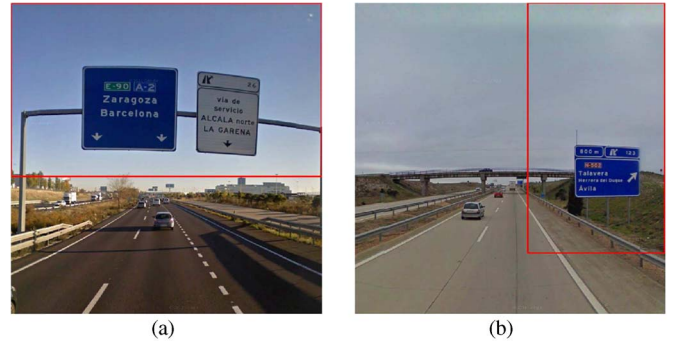


Fig. 4. Regions of interest on the images. (a) Upper region of interest. (b) Lateral region of interest.

the cluster centers (visual words) and the input descriptors. A BOVW histogram is computed to represent the whole image, and the classification decision is made by the classifier previously trained, which is either SVM or Naïve Bayes.

After this general explanation of the BOVW technique, we are going to focus in its first step, i.e., the feature extraction process. Since the traffic panels are located above the road or on the right side, two independent regions of interest are applied on the images. These regions are shown in Fig. 4. Feature extraction, training, and testing are done separately on each region of interest. The features are extracted at some keypoints, which are obtained using the Harris–Laplace salient point detector [10]. It uses a Harris corner detector and subsequently the Laplace operator for scale selection. Because BOVW does not account for spatial information, a prior segmentation stage is mandatory in order to guide the searching of keypoints over the potential areas to be panels in the image. In this way, we maximize the panels modeling again other areas of the image. As traffic panels can be characterized in a global way by their background color, the local features are extracted only on those regions of the images, which are blue or white, through two color masks. Our color masks are hypotheses to be confirmed through the extracted keypoints and bag of features approach. Then, using a prior color segmentation mask and a BOVW technique in cascade is an alternative approach for traffic panel detection without using edges or other geometrical features, as it has been done up to now in the literature. This technique can be generalized to detect any object characterized by a uniform color background in the image.

An efficient method to segment blue and white pixels in the images has been developed. We propose to segment the blue

regions in the image as a combination of three independent methods using a logical AND operation as in

$$\text{Blue Mask} = g_1(x, y) \text{ AND } g_2(x, y) \text{ AND } g_3(x, y) \quad (4)$$

where the first two methods have been proposed by other authors but the third method is a proposal that we are making in this paper, as well as the combination of the three methods.

$g_1(x, y)$ is computed using

$$g_1(x, y) = \begin{cases} 255 & \text{if } R(x, y) \leq T_r \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

as it is proposed in [11]. $R(x, y)$ is the red channel of the image, and $T_r = 90$ is the optimum value according to the source article. This method has been proved to be really useful to discard the blue regions corresponding to the sky, while keeping the blue regions corresponding to the panels, which are typically darker than the sky. On the other hand, this method has the disadvantage that it is not able to reject dark regions in the image (black, gray, and dark colors). This is solved using the next two methods.

On the other hand, $g_2(x, y)$ is computed using

$$g_2(x, y) = \begin{cases} 255 & \text{if } H(x, y) \geq T_1 \text{ and } H(x, y) \leq T_2 \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

as it is proposed in [12]. $H(x, y)$ is the hue component of the image, and $T_1 = 200^\circ$ and $T_2 = 280^\circ$ are the optimum values of the thresholds according to the authors. Unlike the previous method, this method is not able to distinguish between the blue regions in the sky and the blue regions in the panels, and it is not able to discard white regions in the image, but it is very useful to reject colors whose tonality is completely different to blue, such as green, red, or orange.

Finally, our proposal, apart from (4), consists of computing $g_3(x, y)$ using

$$g_3(x, y) = \text{Otsu}(|R(x, y) - B(x, y)|) \quad (7)$$

which applies Otsu's segmentation method [13] on the image obtained by subtracting the blue color component $B(x, y)$ from the red color component $R(x, y)$. Otsu's method reduces the input image to a binary image, assuming that the input image contains two classes of pixels or a bimodal histogram. It computes the optimum threshold that separates both classes so that their intraclass variance is minimal. Unlike the first method, this method is not able to discard the blue regions that correspond to the sky, but it improves the performance of the first method by rejecting dark regions in the image and it improves the performance of the second method by rejecting white regions in the image.

On the other hand, the method to segment white regions is based on the maximally stable extremal regions method [14], which is a region detector that allows to detect bright-on-dark regions in the image.

A comparison of different gray- and color-based descriptors has been carried out. Specifically, the following descriptors have been used: Scale Invariant Feature Transform (SIFT) [15], Colored Scale Invariant Feature Transform (C-SIFT) [16], Hue Scale Invariant Feature Transform (Hue-SIFT) [17], RGB Scale

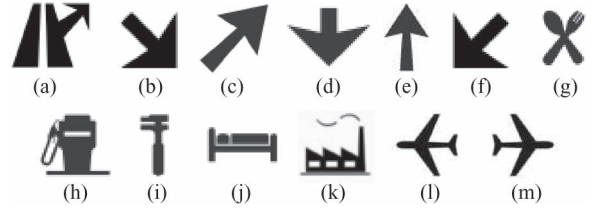


Fig. 5. Considered symbols.

Invariant Feature Transform (RGB-SIFT) [18], Hue Histogram [17], and Transformed Color Histogram (TCH) [18]. They have been computed using the ColorDescriptor library.¹ Results will be shown in Section VI.

V. TEXT DETECTION AND RECOGNITION IN TRAFFIC PANELS

Once the previous method finds that there is a traffic panel in an image, our text location and recognition method explained in [1] is applied on the image. However, some modifications and new functionalities have been proposed in order to increase the efficiency and reduce the number of false positives. Instead of applying the text location method in the whole image, it is done only on those areas of the image given by the blue and white color masks.

Then, character and word recognition is applied. Our character recognizer described in [1] was developed to recognize letters from "A" to "Z" and from "a" to "z" and digits from "0" to "9." However, traffic panels contain not only words and numbers but also symbols such as direction arrows and petrol station indications. Therefore, the system has been modified to recognize also this kind of symbols. Some of the most common symbols that appear in traffic panels have been chosen, and several samples for each symbol have been added to the train set. The chosen symbols are shown in Fig. 5.

The character recognizer may fail when panels are far, as text is small and difficult to segmentate and recognize. However, it is not necessary to recognize all the characters perfectly. They are just an estimation, because a word recognizer is applied later. The word recognizer is based on a unigram probabilistic language model that constrains the output of the character recognizer to a set of meaningful words weighted to their prior probabilities. The model used in [1] to recognize single words in natural images was based on the British National Corpus (BNC), which is a compendium of all the words of the modern English language. However, in this case, we are not recognizing English words but text that appears in Spanish traffic panels. Therefore, instead of using the BNC, we use a dictionary of words that includes all the words that the system is able to recognize, i.e., name of cities, places, and other common words that typically appear in traffic panels, such as "cambio de sentido" (U-turn), "via de servicio" (service road), or "centro comercial" (shopping center). However, we do not have any information available on the frequency of each word; hence, it is not possible to compute the prior probabilities of the words.

¹<http://www.colordescriptors.com>

Therefore, we are assuming equal prior probability for all the words.

In order to increase the effectiveness of the recognition algorithm, we make use of a Web Map Service to reduce the size of the dictionary to a limited geographical area, i.e., to the nearest places. In this paper, we use a reverse geocoding service provided by the project Cartociudad,² which is an information system based on an official database of the Spanish road network, including any kind of routes, highways, urban thoroughfares, and streets. It is supported by different public state Spanish institutions, and it is updated every short time. The reverse geocoding service allows to get certain geographic data, such as street addresses, names of roads, postal codes, milestones, municipalities, provinces, and autonomous communities, from geographic coordinates (latitude and longitude). In addition, this pair (LAT , LON) is known for every image, as it was shown in Section III.

Therefore, instead of using a unique dictionary of words for the whole country, we have created a dictionary for every province in Spain, i.e., each dictionary contains the names of all the municipalities in the province, and another dictionary that has a set of typical words that appear in traffic panels and do not depend on the geographical position, like “centro comercial,” “via de servicio,” or “cambio de sentido.” The names of the capital cities of each province have been also added to the second dictionary, in order to deal with those situations in which a capital city of a province is referenced in a panel that is not located in the same province. The sizes of the dictionaries depend on the province itself, but each one is composed of several hundreds of words.

Given an input image with its associated pair of latitude and longitude coordinates, we make a request to the Cartociudad server using these coordinates. Then, we use the name of the province given by the reverse geocoding service in order to choose the corresponding dictionary. As well as this, the dictionary that contains the common words is also used. Therefore, the language model used to recognize single words is partly based on a fixed dictionary and partly based on a dynamic dictionary that depends on the province where the image was taken.

VI. EXPERIMENTAL RESULTS

A. Traffic Panel Detection

As it was stated in Section IV, a comparison of different descriptors has been carried out in order to check which are the most suitable for the proposed application of traffic panel detection using visual appearance. Specifically, the following descriptors have been compared:

- SIFT [15];
- C-SIFT [16];
- Hue Histogram [17];
- Hue-SIFT [17];
- RGB-SIFT [18];
- TCH [18].

TABLE II
DETECTION FOR BLUE LATERAL PANELS

| Descriptor | Panel detection rate | Sensitivity | Specificity | f |
|---------------|----------------------|-------------|-------------|--------|
| SIFT | 64.28% | 0.2500 | 0.9192 | 0.5846 |
| Hue Histogram | 90.47% | 0.6625 | 0.8782 | 0.7704 |
| TCH | 95.23% | 0.6042 | 0.9253 | 0.7674 |

TABLE III
DETECTION FOR BLUE UPPER PANELS

| Descriptor | Panel detection rate | Sensitivity | Specificity | f |
|---------------|----------------------|-------------|-------------|--------|
| SIFT | 84.43% | 0.5366 | 0.9789 | 0.7577 |
| Hue Histogram | 97.77% | 0.9512 | 0.8438 | 0.8975 |
| TCH | 97.77% | 0.8963 | 0.9536 | 0.9300 |

TABLE IV
DETECTION FOR WHITE LATERAL PANELS

| Descriptor | Panel detection rate | Sensitivity | Specificity | f |
|---------------|----------------------|-------------|-------------|--------|
| SIFT | 41.66% | 0.1724 | 0.9264 | 0.5494 |
| Hue Histogram | 54.66% | 0.3563 | 0.6107 | 0.4835 |
| TCH | 91.66% | 0.6552 | 0.5079 | 0.5815 |

TABLE V
DETECTION FOR WHITE UPPER PANELS

| Descriptor | Panel detection rate | Sensitivity | Specificity | f |
|---------------|----------------------|-------------|-------------|--------|
| SIFT | 72.65% | 0.3740 | 0.9542 | 0.6641 |
| Hue Histogram | 91.40% | 0.8293 | 0.6827 | 0.7560 |
| TCH | 94.53% | 0.7480 | 0.8998 | 0.8238 |

Only with SIFT, Hue Histogram, and TCH, it has been possible to successfully cluster the descriptors and train the classifier, because convergence was not reached using the other descriptors.

Tables II–V show the results for each defined class: blue-background lateral panels, blue-background panels above the road, white-background lateral panels, and white-background panels above the road, respectively. Table VI shows the results for all the panels on the right of the road regardless of their color, whereas Table VII shows the results for the panels above the road regardless of the color. The panel detection rate is evaluated in two different ways. On the one hand, we use the sensitivity and the specificity per frame. These figures give the performance of the system in terms of the recall and the true negative rate, respectively. On the other hand, we give a per-panel detection rate using a multiframe validation process. This figure is normally used by researchers when a tracking process is involved, because the final goal is to provide a percentage of correctly detected panels. Images provided by Google Street View are taken every 10–20 m. Then, a panel can be seen from

²<http://www.cartociudad.es>

TABLE VI
DETECTION INCLUDING ALL THE LATERAL PANELS

| Descriptor | Panel detection rate | Sensitivity | Specificity | f |
|---------------|----------------------|-------------|-------------|--------|
| SIFT | 60.80% | 0.3304 | 0.8511 | 0.5907 |
| Hue Histogram | 84.97% | 0.7625 | 0.5385 | 0.6505 |
| TCH | 94.68% | 0.7464 | 0.4772 | 0.6118 |

TABLE VII
DETECTION INCLUDING ALL THE UPPER PANELS

| Descriptor | Panel detection rate | Sensitivity | Specificity | f |
|---------------|----------------------|-------------|-------------|--------|
| SIFT | 79.38% | 0.5708 | 0.9394 | 0.7551 |
| Hue Histogram | 95.04% | 0.9292 | 0.5975 | 0.7634 |
| TCH | 96.38% | 0.8821 | 0.8817 | 0.8819 |

one to six consecutive frames, depending on the panel size and the occlusions. Additionally, a BOVW classifier does not give geometrical information about the panels in order to implement a traditional tracking process. In our case, we validate a panel when it is detected in at least two consecutive frames. We apply this multiframe detection strategy as a simple way of using tracking. Thus

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (8)$$

$$\text{Specificity} = \frac{TN}{TN + FP}. \quad (9)$$

Sensitivity and specificity are defined as in (8) and (9), respectively. TP stands for the number of true positives, FN stands for the number of false negatives, TN is the number of true negatives, and FP is the number of false positives. The sensitivity relates to the system's ability to identify positive samples, whereas the specificity relates to the system's ability to identify negative samples, i.e., if a panel is present or not in a frame and if it has been detected or not, regardless of if the same panel appears in previous or subsequent frames. Finally, the f -measure is defined as the mean of specificity and sensitivity.

It can be seen that the best results are obtained for the color descriptors, with TCH being the best one. The reason is their invariance to scale and shift of intensity, which make them more robust to illumination changes, shadow, and viewpoints in the scene than other descriptors based only in luminance. Additionally, TCH transforms RGB into normal distributions of the color channels for the image patches pointed by the detected keypoints. Therefore, they become discriminative features to distinguish panel colors from background scene colors (image patches not containing traffic panels) in presence of light changes and arbitrary offsets on intensity values. The per-panel detection rate is above 91% for the four situations under study, and the value of the f -measure is the highest in all cases, except for the blue panels located on the side of the road, although it is very close to the highest value which is obtained with the Hue Histogram descriptor. However, the highest value of the



Fig. 6. Correct panel detection under illumination changes.

specificity is achieved in most cases for the SIFT descriptor. It means that the number of false positives for this descriptor is very low. Nevertheless, the sensitivity is much lower for SIFT, with respect to the other descriptors. Thus, the number of false negatives is very high, with respect to the number of true positives. In other words, the classifier trained with the SIFT descriptor categorizes most of the images as if there is not any panel present in the image. Hence, the detection rate for SIFT is much lower than for the other descriptors.

All the previous experiments have been carried out using a Naïve Bayes classifier. However, another classifier based on SVM has been tested. The best combination of parameters of this classifier for this specific application is a linear kernel and a cost parameter $C = 100$. We have found that, in general, the number of false positives using SVM is much lower than the number of false positives using Naïve Bayes and, therefore, the specificity is higher. However, the number of false negatives (when the algorithm does not detect a panel but there is one in reality) is higher and, consequently, the sensitivity is lower than if a Naïve Bayes classifier is used. The panel detection rate is also lower and, in addition, we have seen that an SVM requires a much higher computational time than a Naïve Bayes classifier. Therefore, in this application, it is preferred to use Naïve Bayes.

B. Traffic Panel Detection in Challenging Scenarios

Here, we include a discussion on the challenging scenarios found during the experiments and several sample images of the system behavior in these cases. No additional preprocessing stage was carried out on the Google Street View images, except the already described methodology in this paper. Two image sets were created, as explained in Section III.

Google images are publicly available, but they introduce some constraints: they are taken at daytime, usually during summer to avoid bad weather conditions (rain, fog, snow, etc.), they are rectified and stitched to generate a panoramic view from several cameras, and they are also filtered to blur vehicle plates and human faces, among others. One might think that these restrictions could help, but traffic panel detection is not straightforward because the images still include challenging scenarios for computer vision processing. Hereafter, several examples are discussed for a system configuration with a TCH descriptor and a Naïve Bayes classifier, which are the ones that yielded the best detection performance.

Images are not taken at early morning nor sunset, but light changes can be very challenging even at daytime. Fig. 6 depicts some examples of correct detections when some of the following artifacts are present: lateral sunshine, low contrast, saturated images, and shadows or glints on the panel. The panels on these



Fig. 7. Correct panel detection under occlusions and clipping.



Fig. 8. Panel detection under cluttered background. The two leftmost images are correct detections and the rightmost images are false positives.

images are correctly detected due to the robustness of the TCH descriptor against illumination changes.

Other challenging scenarios may occur because of partial occlusions or clipped panels. Fig. 7 shows some examples of correctly detected panels, but partially occluded due to other traffic signs and vegetation or clipped due to image patch boundaries. Our proposal is quite robust for these cases because it does not consider geometrical properties of the panels, except visual appearance. For hard occlusions or clipping (over 50%), our system fails.

In roads near urban areas, the background of the traffic panels is usually cluttered, which affects the detection process. Fig. 8 depicts two correct detections and two false positives related to these challenging scenarios. There are some objects with uniform white or blue backgrounds in the image of a size similar to the panels and with letters inside them, such that our system will consider them as traffic panels generating false positives. These are the worst scenarios for our system. The performance will depend on the background object color. In the leftmost image, we can see a red banner that does not prevent the successful detection of the traffic panel, but in the rightmost image, a blue factory panel is incorrectly classified as a traffic panel.

Google image rectification and filtering processes, unlike one might expect, sometimes introduce new challenging situations similar to the those displayed in Fig. 9. Those images present blurring and mismatches because of the image stitching. All of the them are correctly detected as panels by our system. In these cases, the performance of our system is reduced in the text recognition stage, because, as can be easily observed, the images are also difficult for a human reader without prior knowledge.

So far, we have shown the strength of our system in several challenging situations. In relation to false positives, they are mainly due to cluttered images, as shown in Fig. 8. Additionally, they can also appear as a bridge, a vehicle, or a bush because of other objects in the road that are incorrectly



Fig. 9. Correct panel detection under rectification and stitching artifacts due to Google treatment.



Fig. 10. Additional examples of false positives.

TABLE VIII
TEXT DETECTION AND RECOGNITION INCLUDING ALL
DETECTED LATERAL PANELS

| Data | Distance | Detection rate | Recognition rate |
|---------|----------|----------------|------------------|
| Words | Short | 85.48% | 79.25% |
| | Medium | 64.37% | 46.43% |
| | Long | 20.22% | 0% |
| Numbers | Short | 62.07% | 48.61% |
| | Medium | 26.44% | 10.87% |
| | Long | 5.81% | 0% |
| Symbols | Short | 43.33% | 71.79% |
| | Medium | 36.36% | 60.71% |
| | Long | 18.29% | 60.00% |

classified as traffic panels. Fig. 10 presents a few more false positive cases.

C. Text Detection and Recognition

A total of 145 km of two different highways of the Spanish road network have been analyzed (see Fig. 2). In this stretch, there are 185 traffic panels of which 77 are above the road and 108 are located at the right side of the road. Typically, there are several samples for every panel, because each panel usually appears in consecutive frames at different distances. The detection and recognition have been carried out for every frame independently. Therefore, the text detection and recognition rates shown in Tables VIII–X have been computed in a single frame. We show the results as a function of the distance from the vehicle to the panel, in order to show how the distance to the panels affects to the algorithm performance. We have defined three ranges: short distance, when the panel is less than 40 m far; medium distance, when it is in the range 40–70 m; and long distance, when the panel is further than 70 m, approximately. In addition, the detection and recognition rates have been computed separately for words, numbers, and symbols. In general, the nearer the panel is, the better the performance is.

TABLE IX
TEXT DETECTION AND RECOGNITION INCLUDING ALL
DETECTED UPPER PANELS

| Data | Distance | Detection rate | Recognition rate |
|---------|----------|----------------|------------------|
| Words | Short | 92.92% | 79.19% |
| | Medium | 85.33% | 70.06% |
| | Long | 46.32% | 26.98% |
| Numbers | Short | 83.44% | 61.42% |
| | Medium | 46.62% | 48.39% |
| | Long | 18.63% | 10.53% |
| Symbols | Short | 46.27% | 90.32% |
| | Medium | 66.09% | 97.37% |
| | Long | 31.25% | 85.00% |

TABLE X
TEXT DETECTION AND RECOGNITION INCLUDING ALL
THE DETECTED PANELS

| Data | Distance | Detection rate | Recognition rate |
|---------|----------|----------------|------------------|
| Words | Short | 90.18% | 79.21% |
| | Medium | 78.60% | 63.85% |
| | Long | 36.00% | 20.99% |
| Numbers | Short | 74.46% | 56.93% |
| | Medium | 38.64% | 38.24% |
| | Long | 13.09% | 8.51% |
| Symbols | Short | 45.09% | 83.17% |
| | Medium | 54.17% | 87.50% |
| | Long | 23.97% | 74.29% |

The best performance is achieved for words, with the detection and recognition rates above 90% and close to 80%, respectively, when the panel is at short distance, as shown in Table X. The performance hardly decreases when the panel is at medium distance (up to 70 m far). However, the detection rate for numbers and symbols is lower compared with words because our text detection method, as explained in [1], focuses on detecting text lines that has at least three elements; thus, numbers and symbols that may appear isolated in the panel could not be detected. A lower threshold than 3 could be used for this specific application, although the number of false positives may increase; however, what we wanted to show in this paper is that the proposed text detection and recognition algorithm trained with a concrete data set, which is composed of real-world images that are completely different to the Street View images used in this paper, can be generalized to any other situation similar to the one shown in this work, achieving a reliable performance.

From the tables, it can also be seen that the recognition rate for symbols remains above 70% even when the panel is at long distance. This is due to the fact that, in general, symbols in the traffic panels are typically bigger than letters and numbers; thus, it is easier to be recognized even when the panel is further than 70 m.

Another conclusion that can be extracted from the tables is that, in general, the performance of the algorithm at medium distance is worse for panels located beside the road than for panels above the road, whereas it happens the opposite at very short distance. The reason for this is that there are some

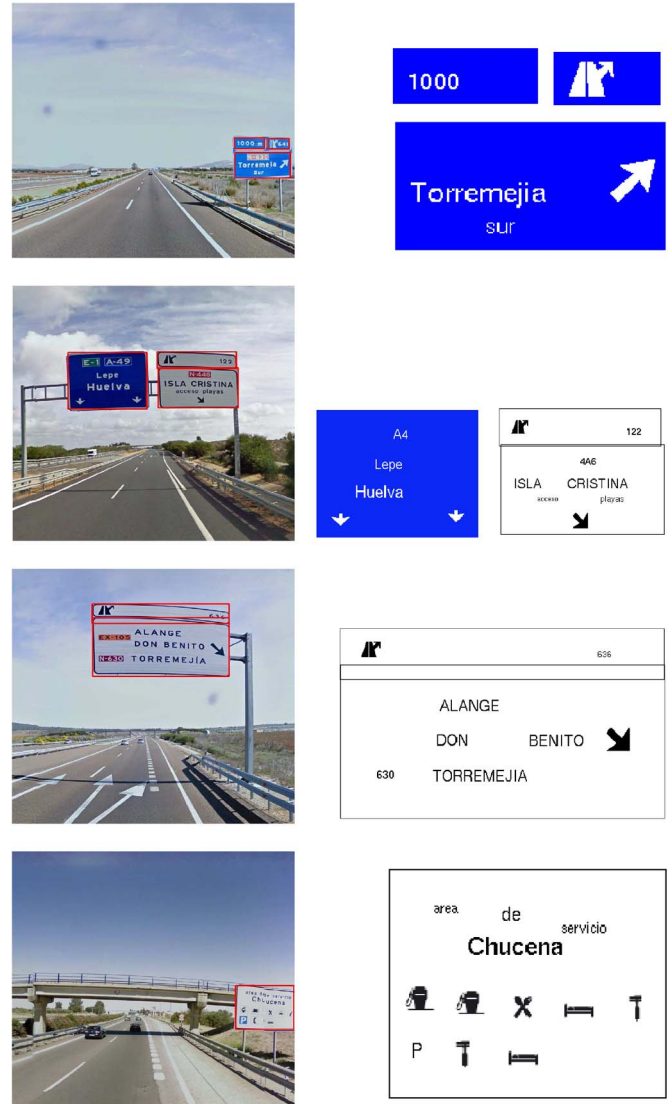


Fig. 11. Image results. Panel detections are on the left column and text recognition on the right column.

deformations at the margins of the image, both left and right and top and bottom, which affect the detection and recognition.

Some examples of the detection and recognition are shown in Fig. 11.

D. Text Detection and Recognition in Challenging Scenarios

Challenging scenarios do not only affect to the traffic panel detection performance but also influence in the posterior text recognition stage. More complex scenarios derivate in worse text detection and recognition ratios. Nevertheless, the text recognition works well in some challenging images but fails in others with strong artifacts, as shown in Fig. 12.

After the analysis of the test set, we estimate that around the 15% of the images could be considered as challenging ones, according to the discussed cases in Section VI-B and D.

VII. CONCLUSION AND FUTURE WORKS

In this paper, we have presented a real application of the text detection and recognition algorithm proposed in [1] by the

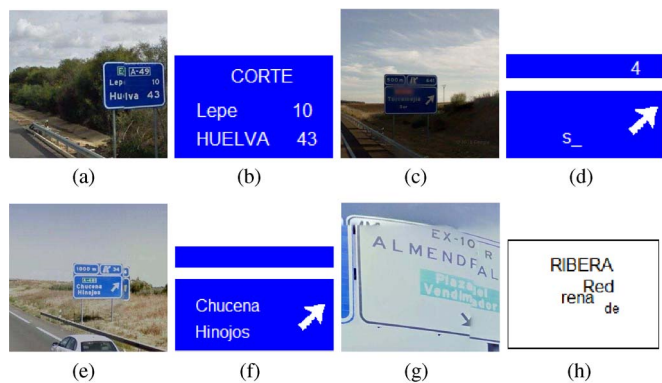


Fig. 12. Examples of text recognition in challenging images. (a), (c), (e), and (g) are the detected traffic panels, and (b), (d), (f), and (h) are the recognized texts. As can be seen, (c) and (g) are some very complex scenarios where our system is not able to correctly recognize the text and symbols.

same authors, including some adaptations and new functionalities. It consists in reading the information depicted in traffic panels using panoramic images downloaded from the Google Street View service. The main use of this application is to automatically create up-to-date inventories of traffic panels of whole regions or countries. This information is very useful for supporting road maintenance and for developing future driver assistance systems.

One of the main contributions of this paper is the modeling of traffic panels using a BOVW technique from local descriptors extracted at interest keypoints, instead of using other features such as edges or geometrical characteristics as it has been done up to now in the literature. This is not an easy task due to the immense variability of the information included in traffic panels. Nevertheless, the experimental results show the effectiveness of the proposed method. An efficient segmentation method based on color masks has been implemented to guide the keypoints searching in the image. A new method to detect blue areas in the image has been proposed. Different gray- and color-based descriptors have been compared, and the TCH descriptor has proved to be more suitable for this application. In addition, the dimensionality of this descriptor is small (only 37 and 45 elements, respectively); thus, the training time is lower than using other descriptors of higher dimensions. Other additional contributions are: a symbol recognizer for traffic panels, a method to reduce the size of the dictionary to a limited geographical area using a reverse geocoding service, the generalization of our previous text detection and recognition method to traffic panels without retraining, and the idea of using Google Street View images as inputs of our system for traffic panel inventory.

Although the experimental results obtained with our proposal are quite good, there are some improvements to carry out as future works in order to reach an application able to be commercialized. The BOVW technique automatically detects traffic panels on image patches, but it does not account for 2-D spatial information. We have constrained BOVW using color masks to restrict the possible panel locations in the image and guide keypoint extraction. However, as future work, we intend to use spatial extensions to BOVW, such as sliding window approaches, branch and bound [19], or structured

SVM [20]. Moreover, the reliability of part-based models [21] and semantic segmentation [22] has been demonstrated in several data sets and challenges, such as PASCAL VOC [23], but it has not been proven that they perform better than a constrained BOVW for traffic panel detection. Further research of these techniques for this application and its comparison with our current proposal is a good plan for the near future. In addition, false panel detections need to be reduced, particularly for lateral panels. Lens distortion removal and improved visual appearance description can help to obtain higher specificity values.

The text location and recognition method described in [1] was applied only on those images where a panel was found, in order to reduce the number of false positives and increase the efficiency of the proposed algorithm. A unigram language model conducted the word recognition. In addition, our proposed model was partly based on a fixed dictionary that contained common words that can be found everywhere and partly based on a dynamic dictionary that depends on the province where the traffic panel is located. The model assumed equal prior probability for all the words. As future work, we intend to compute the prior probabilities of all the words in the dictionary to allow a more precise and reliable recognizer. In the same way, the use of a unigram language model does not take into account the likelihood of two or more words appearing together. Using language models of a higher order would allow to recognize more precisely the names of places composed of several words.

Other limitations of the system are the distance where acceptable text recognition can be carried out and daytime working only. In case the proposed method was aimed at driver assistance, the recognition would need improvements at far distances to the panel, and it should work at nighttime. Higher resolution images and tracking techniques could enhance the performance of our system, and a new nighttime functionality based on the previous work of the authors in [24] should be added.

Finally, the recognition of the information depicted in the traffic panels was done frame by frame. Typically, a panel appeared in several consecutive frames. As future work, we intend to do a multiframe integration of the recognized information at each single frame. In addition, the use of the *a priori* knowledge that we know about the design of traffic panels would improve the recognition rates, because certain objects, particularly symbols and numbers, are located only at certain parts of the panels.

ACKNOWLEDGMENT

The authors would like to thank Google for the images, which have been used only for research purposes.

REFERENCES

- [1] A. González and L. M. Bergasa, "A text reading algorithm for natural images," *Image Vis. Comput.*, vol. 31, no. 3, pp. 255–274, Mar. 2013.
- [2] A. Mogelmose, M. Trivedi, and T. Moeslund, "Vision-based traffic sign detection and analysis for intelligent driver assistance systems: Perspectives and survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 4, pp. 1484–1497, Dec. 2012.

- [3] Z. Hu, "Intelligent road sign inventory (IRSI) with image recognition and attribute computation from video log," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 28, no. 2, pp. 130–145, Feb. 2013.
- [4] A. González, M. Garrido, D. Llorca, M. Gavilán, J. Fernandez, P. Alcantarilla, I. Parra, F. Herranz, L. M. Bergasa, M. Sotelo, and P. Revenga, "Automatic traffic signs and panels inspection system using computer vision," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 2, pp. 485–499, Jun. 2011.
- [5] A. V. Reina, R. J. L. Sastre, S. L. Arroyo, and P. G. Jiménez, "Adaptive traffic road sign panels text extraction," in *Proc. 5th WSEAS ISPRA*, 2006, pp. 295–300.
- [6] W. Wu, X. Chen, and J. Yang, "Detection of text on road signs from video," *IEEE Trans. Intell. Transp. Syst.*, vol. 6, no. 4, pp. 378–390, Dec. 2005.
- [7] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Proc. Workshop Stat. Learn. Comput. Vis. ECCV*, 2004, pp. 1–22.
- [8] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995.
- [9] D. D. Lewis, "Naive (Bayes) at forty: The independence assumption in information retrieval," in *Proc. ECML*, 1998, pp. 4–15.
- [10] K. Mikolajczyk and C. Schmid, "Indexing based on scale invariant interest points," in *Proc. ICCV*, 2001, pp. 525–531.
- [11] N. Kulkarni, "Color thresholding method for image segmentation of natural images," *Int. J. Image, Graph. Signal Process.*, vol. 4, no. 1, pp. 28–34, Feb. 2012.
- [12] H. Gómez-Moreno, S. Maldonado-Bascón, P. Gil-Jiménez, and S. Lafuente-Arroyo, "Goal evaluation of segmentation algorithms for traffic sign recognition," *IEEE Trans. Intell. Transp. Syst.*, vol. 11, no. 4, pp. 917–930, Dec. 2010.
- [13] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man Cybern.*, vol. SMC-9, no. 1, pp. 62–66, Jan. 1979.
- [14] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide baseline stereo from maximally stable extremal regions," in *Proc. BMVC*, 2002, pp. 384–393.
- [15] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. ICCV*, 1999, pp. 1150–1157.
- [16] A. E. Abdel-Hakim and A. A. Farag, "CSIFT: A SIFT descriptor with color invariant characteristics," in *Proc. CVPR*, 2006, pp. 1978–1983.
- [17] J. van de Weijer, T. Gevers, and A. D. Bagdanov, "Boosting color saliency in image feature detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 1, pp. 150–156, Jan. 2006.
- [18] 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.
- [19] C. Lampert, M. Blaschko, and T. Hoffman, "Efficient subwindow search: A branch and bound framework for object localization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 12, pp. 2129–2142, Dec. 2009.
- [20] T. Joachims, "Learning to align sequences: A maximum-margin approach (technical report)," in *New Algorithms for Macromolecular Simulation*. New York, NY, USA: Springer-Verlag, 2003, pp. 57–68.
- [21] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.
- [22] E. Borenstein and S. Ullman, "Combined top-down/bottom-up segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 12, pp. 2109–2125, Dec. 2008.
- [23] M. Everingham, L. V. Gool, C. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.
- [24] A. González, L. M. Bergasa, J. Yebes, and M. Sotelo, "Automatic information recognition of traffic panels using SIFT descriptors and HMMS," in *Proc. ITSC*, 2010, pp. 1289–1294.



portation systems.



commercializes computer vision systems for road infrastructure inspection. He is the author of more than 150 refereed papers in journals and international conferences and the corresponding author of six national patents and one patent from the Patent Cooperation Treaty. His research interests include real-time computer vision and its applications, particularly in the field of intelligent vehicles, vehicle safety, and assistant robotics.

Mr. Bergasa is an Associate Editor of the *Journal of Physical Agents*, a member of the editorial board of the *International Journal of Vehicular Technology*, and a Habitual Reviewer in 12 journals indexed by the Journal Citation Reports. He has served on Program/Organizing Committees in more than 15 conferences. During 2004–2012, he has been awarded nine Spanish prizes related to robotics and automotive fields.



Álvaro González received the M.S. and Ph.D. degrees from the University of Alcalá, Alcalá de Henares, Spain, in 2008 and 2013, respectively, both in telecommunications engineering.

Since 2008, he has been with the Robesafe Research Group, University of Alcalá. He is a Co-founder of Vision Safety Technologies Ltd., which is a spin-off company established to commercialize computer vision systems for road infrastructure inspection. His research interests include real-time computer vision and intelligent traffic and trans-

Luis M. Bergasa (M'06) received the M.Sc. degree in electrical engineering from the Technical University of Madrid, Madrid, Spain, in 1995 and the Ph.D. degree in electrical engineering from the University of Alcalá (UAH), Alcalá de Henares, Spain, in 1999.

He is currently a Full Professor with the Department of Electronics, UAH, where he was the Head during 2004–2010. Since 2010, he has been the Coordinator of the RobeSafe Research Group. Since 2009, he has been the Research Director of Vision Safety Technologies Ltd., a spin-off company that

J. Javier Yebes received the M.Sc. degree in electrical engineering in 2009 from the University of Alcalá (UAH), Alcalá de Henares, Spain, where he is currently working toward the Ph.D. degree at the Department of Electronics.

He has collaborated in several intelligent transportation systems (ITS) projects as a member of the RobeSafe Research Group, UAH. His research interests include computer vision, service robotics, ITS, 3-D scene understanding, semantic learning, and visual inference.