**Proceedings of the 11th International IEEE
Conference on Intelligent Transportation Systems
Beijing, China, October 12-15, 2008**

Novel Index for Objective Evaluation of Road Detection Algorithms

J.M. Álvarez and A. López

***Abstract*— Road detection is a relevant task within vision–based systems devoted to assist the driver. Although they have been improved during the last decade, these algorithms are usually validated using qualitative results. Nonetheless, quantitative evaluation is necessary either to enable the comparison between different algorithms or to achieve the optimal performance of a given one. In this paper we present a composite index to quantitatively assess the performance of road detection algorithms. The measure is based on a weighted combination of different evaluations which use a trade-off between precision and recall scores. Obtaining a single index score is a major benefit. It can be used to easily compare algorithms or to properly set their parameters. Moreover, innovatively our proposal includes a human perception criterion to improve its usefulness. Experiments on real–world data corroborate the usefulness of the proposed index.**

I. INTRODUCTION

Advanced driver assistance systems (ADAS) have arisen as a contribution to traffic safety. Within this field, on–board vision has been widely used since it has many advantages (higher resolution, low power consumption, low cost, easy aesthetic integration, non-intrusive nature) over other active sensors such as radar or lidar. Common vision–based systems use at least one on–board camera mounted facing forward the front windscreen of the car. Road detection (RD) is a very important task within these vision–based systems since it can be used either for assisting other driving systems [1] or for their own right [2]. The main idea of these algorithms is to classify each pixel of the incoming images as road or non–road, based on theory of machine learning using features such as color or texture (Fig. 1).

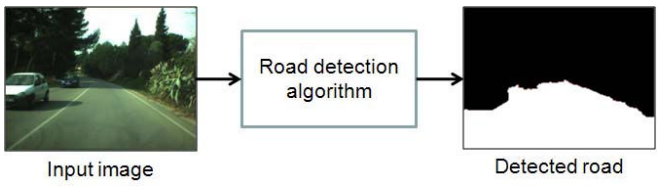

Fig. 1. Road detection algorithm: each pixel is classed as road (white) or background (black) depending on its similarity with a known road model.

Due to their relevance, important efforts have been done in order to improve RD algorithms. However, no agreement has been reached to properly evaluate their quality. Such evaluation is absolutely essential to characterize an individual algorithm and for comparing algorithms.

Currently, most authors validate their algorithms using visual results [3], [2], [4] and only a few of them use some quantitative criterion [5], [6]. In this sense, the most relevant assessment is based on the measure of quality $\hat{g}$ introduced in [7] and used in [8]. Nevertheless, none of these methods consider specific characteristics of the scenes imaged within the RD context: the perspective effect in the image, the ambiguity at the edge of the road and trade–off between hit and miss detection rates such as the trade–off between road detection and obstacle preservation (Fig. 2).

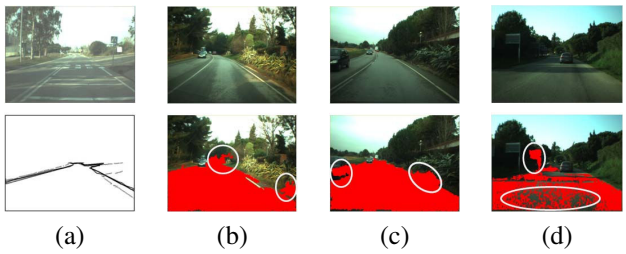

Fig. 2. The main lack of quality of current road detection methods is due to: a) boundary ambiguity, i.e., the uncertainty in the exact position of the edges of the road; b) different error perception when miss–classified pixels represent points closer to the camera c) final application requirements, e.g., miss–classifying objects is more relevant than miss–classifying the background; d) over–detection and under–detection do not have the same relevance.

In this paper we propose a novel approach to validate RD results. The aim of this proposal is twofold: deciding which RD algorithm is the best based on the final application and finding the optimum parameters of an individual algorithm to process a complete sequence of images. We introduce the Road Detection Index: a composite measure which is a weighted combination of the evaluations of all interesting objects present in the scene (including the road). Each of these evaluations is based on common metrics such as precision and recall. The weighting strategy is used to meet the requirements of different final applications, e.g., car tracking applications prefers under–segmenting the road but completely preserve cars while road following algorithms will prefer miss–classifying car pixels while achieving a higher accuracy at the road region. Moreover, the procedure incorporates a human perception criterion to adapt the assessment to the real perception of the scene (cost–sensitive measures). Such criterion has been included earlier in the context of evaluating general segmentation algorithms [9], [10]. Finally, the proposed scheme also considers the ambiguity in determining the exact position of the edges

This work was supported by the Spanish Ministry of Education and Science under project TRA2007-62526/AUT and research programme Consolider Ingenio 2010: MIPRCV (CSD2007-00018)

J.M. Álvarez and A. López are with The Computer Vision Center and with the Dpt. of Computer Science, Universitat Autónoma de Barcelona, 08193 Cerdanyola (Barcelona), Spain jalvarez@cvc.uab.es

of the road (or vehicles) while generating the ground truth. The procedure weights down the errors within the boundaries more than errors elsewhere in the image.

The rest of this paper is organized as follows. First, in Sect. II, the proposed index is defined. In Sect. III the ground truth required and the procedure to set up the index are detailed. Experiments proving the usefulness of the index are shown in Sect. IV. Finally, in Sect. V, conclusions and future work are drawn.

II. Road Detection Index

First, in Sect. II-A, the index requirements are defined. Then, in Sect. II-B, the cost-sensitive measures are derived and the Road Detection Index is stated in Sect. II-C. This index is expended to tackle the boundary uncertainty in Sect. II-D.

A. Requirements

The first step is to highlight some desirable properties for the proposed index:

- The index must take into account positive and negative results. That is, over–detection and under–detection must be considered independently.
- The index must be capable to deal with boundary ambiguity.
- The values of the index must be bounded, so they can be easily analyzed and compared.
- The index must weight down the error committed to detect road pixels representing points far away from the camera.
- Although the main task is detecting the road surface, the index must penalize those algorithms which do not preserve other objects on the road.
- The scheme must be parameterizable to cope with different applications.
- Finally, the index must provide a continuous magnitude, so the adjustment of parameters of the algorithms can be carried out accurately.

Following these ideas, the computation of the index is summarized in Fig. 3. The process is divided in different parts corresponding with different categories of objects in the scene. Obviously, the main object is the road surface. Other objects may be vehicles or pedestrians. Each of these evaluations is done using the cost-sensitive precision recall measures defined in Sect. II-B. Finally, the Road Detection Index (RDI) is the overall combination of these evaluations.

B. Cost–Sensitive Precision-Recall Measures

Since our aim is knowing about the amount of over and under–segmented ratio in the result mask, we have focused the evaluation on two metrics: precision (P) and recall (R) [11]. These scores have been widely used in information retrieval systems. However, the interpretation of these values for the assessment of image segmentation is slightly different. In probabilistic terms, precision is the probability that the result is valid, and recall is the probability that the ground truth data was detected. Thus, the aim in

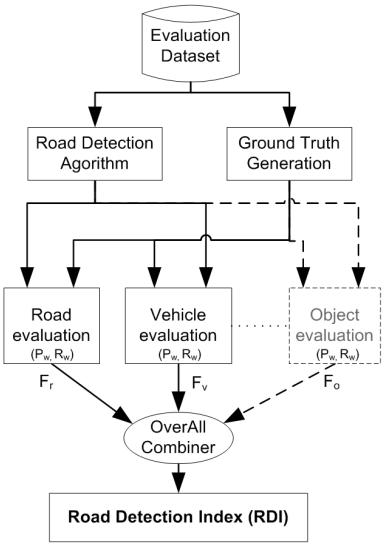

Fig. 3. Road Detection Index computation. A set of images is hand labelled and processed by the RD algorithm. Both results are passed to a set of blocks which evaluate the quality of the detection using the cost-sensitive precision recall measures defined in Sect. II-B. These measures are combined to obtain the final index.

image segmentation is to get both high precision and high recall scores. We consider using both scores since each one provides different information about the segmented image: a low recall value is typically the result of under–segmentation and indicates failure to capture salient image structures; a low precision value is typically the result of significant over–segmentation, or when a large number of boundary pixels have greater localization errors.

Once selected the most appropriate metrics to be used, their definition has been modified to consider the perspective effect present in the image. That is, errors in pixels representing points closer to the camera position are more relevant than those errors in pixels representing further points. With this aim, a weighting strategy has been introduced to modify the contribution of each pixel to the final P and R scores. Thus, considering $S_1 = \{s_1, s_2, ..., s_N\}$ is the segmented result and given its associated ground–truth, $GT = \{gt_1, gt_2, ..., gt_N\}$, (where $gt_i = 1$ for "road" pixels and $gt_i = 0$ for "non–road" pixels), precision and recall scores can be defined as,

$$P = \frac{|S_1 \cap GT|}{|S_1|} = \frac{\sum_{i=1}^{i=N}(s_i \cap gt_i)}{\sum_{i=1}^{i=N} s_i}, \quad (1)$$

$$R = \frac{|S_1 \cap GT|}{|GT|} = \frac{\sum_{i=1}^{i=N}(s_i \cap gt_i)}{\sum_{i=1}^{i=N} gt_i}. \quad (2)$$

Given a weight map, $W = \{w_1, w_2, ..., w_N\}$, the value of the weighted precision (P_w) and weighted recall (R_w) scores can be defined as:

$$P_w = \frac{\sum_{i=1}^{i=N} w_i(s_i \cap gt_i)}{\sum_{i=1}^{i=N} w_i s_i}, \quad (3)$$

$$R_w = \frac{\sum_{i=1}^{i=N} w_i(s_i \cap gt_i)}{\sum_{i=1}^{i=N} w_i gt_i}, \quad (4)$$

where $w_i \in W$ denotes the weight on the i-*th* element of the data.

In addition, an advantage of having these two scores (P and R) is that they can be weighted differently to highlight their relevance. The *F–measure* (or *effectiveness*) is a single measure that trades–off precision versus recall values and is calculated using the weighted harmonic mean of precision and recall (further details on harmonic mean can be found in [11]):

$$F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} = \frac{(\beta^2+1)PR}{\beta^2 P + R}, \quad (5)$$

where $\beta^2 = \frac{1-\alpha}{\alpha}$, $\alpha \in [0,1]$ and thus $\beta^2 \in [0,\infty]$. F ranges from 0 to 1.

The default balanced *F–measure* equally weights precision and recall, which means $\alpha = \frac{1}{2}$ or $\beta = 1$ and $F = \frac{2PR}{P+R}$. However, using an even weighting is not the only choice. Values of $\beta < 1$ emphasize precision, while values of $\beta > 1$ emphasize recall. Two other commonly used *F–measures* are *F2–measure* and *F0.5–measure*. The former, weights recall twice as much as precision. The latter weights precision twice as much as recall.

The algorithm delivers a different F value for each block $(F_r, F_v..., F_o)$ as shown in Fig. 3.

C. Road Detection Index

Once defined the metrics used within each block (P_w and R_w) and the measure obtained as the output of these blocks (F_r, F_v...,F_i), the remaining is combining them into the final Road Detection Index (RDI). Having a single index is important in order to easily compare different algorithms or different results from the same algorithm.

Once more, a weighting strategy is used to emphasize the relevance of having an accurate road surface or to emphasize the relevance of preserving other objects in the scene. Thus, the final RDI is a weighted combination of $F_r, F_v..., F_i$:

$$RDI = \frac{\sum_{i=1}^{N} q_i}{\sum_{i=1}^{N} \frac{q_i}{F_i}}, \quad (6)$$

where N is the number of evaluation blocks. F_i are the measures delivered in each block, $i \in (road, vehicle..., object)$. $q_i \in [0...1]$ are the weights associated to each of these blocks.

RDI ranges between 0 and 1. The higher index, the higher quality of the algorithm.

D. Dealing with boundary ambiguity

Having the main index defined, the remaining is considering the boundary uncertainty. This error refers to the inherent ambiguity in the boundary perception when manually segmenting the images to generate the ground truth (Fig. 2a). That is, the exact localization of the boundaries (of the road or any object in the scene) may differ from one human segmenter to the others.

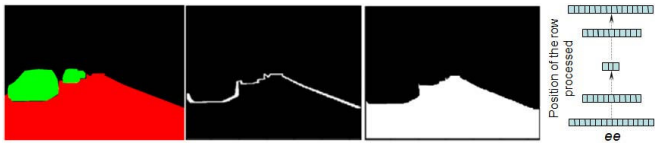

Fig. 4. Each ground truth is processed to differentiate between the inner part of the object (road) and its boundary. The boundaries are obtained with a dynamical structural element (ee) which size is modulated depending on position of the pixel being processed. Thus, pixels representing points closer to the camera use a bigger ee than those representing further away points. From left to right, complete ground truth, road boundary, road inner region and ee modulation.

To minimize this error the ground truth (I) is divided into two different masks: boundary mask (I_b) and inner mask (I_i). These masks are extracted considering not only the edge pixel but its surrounding region:

$$I_b = (I - (I \ominus ee)) \,|\, ((I \oplus ee) - I), \quad (7)$$

$$I_i = I - I_b, \quad (8)$$

$\ominus$ and $\oplus$ are the binary erosion and dilation respectively performed using the structural element ee. Its size varies dynamically depending on the position of the filtered pixel. Using this ee the method considers the perspective effect present in the scene. Its size is modulated depending on the boundary extracted: road or objects (Fig. 4).

Two different F values are worked out independently for the boundary and the inner part of the object. The output of each block $(F_r, F_v..., F_i)$ corresponds to the average of these two values. Nevertheless, this method does not produce major benefits if a single ground truth is considered. Hence, we have expanded the measure to compare the algorithm result against multiple ground truths using a *leave–one–out* procedure [12]. If we have K different ground truths available, we can obtain K different precision and recall scores considering pairwise evaluations (P_w^k and R_w^k respectively). A single recall score is obtained as the average of P_w^k ($k \in [1, 2..., K]$). A single precision is calculated using the number of pixels matching the mask of at least one ground truth. That is, it is obtained comparing the algorithm result against the logical *or* of all the ground truths available.

Finally, instead of arithmetical average of the boundary and the inner part scores we consider the number of ground truths available (K). Thus, the contribution of the boundary assessment to the final measure is proportional to K. That is, the bigger number of ground truths, the bigger contribution due to the boundary. This is quite logical since having a single ground truth for comparison does not reduce the localization error. This method slightly differs from the weighting strategy presented in [13] where pixels are directly weighted accordingly to how many observers have marked the given pixel as a boundary one.

III. Ground Truth and Index Set Up

Although road detection algorithms deliver binary masks (road and non-road pixels) the proposed evaluation needs more than one binary mask to cope with its requirements

(Fig. 4). Each ground truth provides information regarding the road region and those interesting objects (vehicles and/or pedestrians) in the scene. Furthermore, each image should be manually segmented more than once to assess properly the boundary region.

The perspective weights (W) have been adjusted as a quadratic function of the distance between the point represented by each pixel and the camera position. Thus, pixels closer to the center of the image give lower weights than further ones. Balanced weight have been used between over-segmenting and under-segmenting (β). Finally, two different adjustments for the weights of each block (q_i) have been used. The former (RDI from now on) is a balanced weighting between road precision and object preservation. The latter, (RDI$_{veh}$) emphasizes object preservation (twice relevance on the vehicles than on the road). These two configurations corresponds to two different common road detection applications: road following and vehicle detection. In addition, the simplest configuration has been considered (F). Such configuration sets perspective weights (W) to 1 and fully emphasizes road precision without considering objects. Thus, it incorporates neither the weighting nor the combination of indexes and equals to the *effectiveness*. Finally, three different users have generated ground truths for each image ($K = 3$).

IV. EXPERIMENTS

This section explores the benefits of using the proposed measure in front of the existing ones. We first compare different detection results of different images in terms of maximum performance and next the index is used to set the optimal parameters of an individual algorithm over a large number of images.

A. Index Validation

Our first experiment consists in comparing the performance of different measures. The aim of this comparison is not to obtain a higher/lower score, but comparing the ability of the measure to correspond to the fidelity of the result accordingly with the index requirements (Sect. II-A). Five different measures have been compared: RDI, RDI$_{veh}$, F and two other measures currently used within the field: accuracy (ACC) and quality ($\hat{g}$). These two measures are derived using the four entries of a contingency table: *true positives* (TP) is the number of correctly labelled road pixels, *true negatives* (TN) is the number of non-road pixels detected, *false positives* (FP) is the number of non-road pixels classified as road pixels and *false negatives* (FN) is the number of road pixels erroneously marked as non-road. Accuracy ($ACC = \frac{TP+TN}{TP+FP+FN+TN}$) is the fraction of classifications that are correct. Quality ($\hat{g} = \frac{TP}{TP+FP+FN}$) takes into account the completeness of the extracted data as well as its correctness. The first comparison considers the case where no objects but the road are present. A set of results of a given image has been generated (Fig. 5) and assessed (Table- I). The most relevant thing is how the results are arranged. Top row shows the order obtained using accuracy, quality and F measures. Middle row shows the the order obtained using RDI and RDI$_{veh}$ (notice that since no interest objects are in the image these measures have the same value). The weighting process highlights those results where miss–classified pixels represent points further away from the camera. In addition it is shown in the last column how all the measures get closer to one (their maximum value) when the result is almost the ground truth.

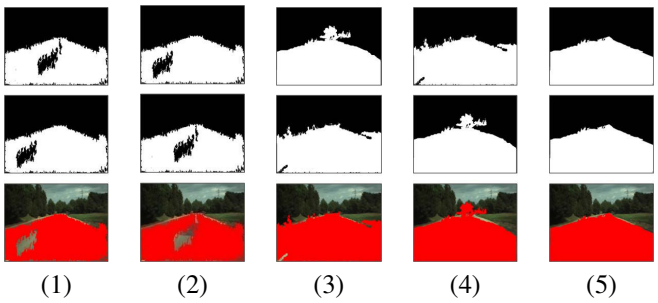

Fig. 5. Comparison of different measures performance applied to road detection results when no objects are present in the scene. Results in rows are arranged depending on its performance (the most to the left, the lower score). Upper row corresponds to the common order obtained using accuracy, quality and F (the three measures yield the same). Middle row shows the order obtained with RDI and RDI$_{veh}$ (the two measures the same). Bottom row shows the masks of second row overlapped on the corresponding original images.

TABLE I
QUALITY ASSESSMENT OF ROAD DETECTION RESULTS SHOWN IN FIG. 5

	Top row			Middle row	
	ACC	$\hat{g}$	F	RDI	RDI$_{veh.}$
1	0.9312	0.8682	0.9295	0.8468	0.8468
2	0.9344	0.8745	0.9330	0.9001	0.9001
3	0.9586	0.9236	0.9603	0.9385	0.9385
4	0.9642	0.9340	0.9659	0.9797	0.9797
5	0.9876	0.9761	0.9879	0.9922	0.9922

The second comparison refers to the advantages of using RDI when other objects are present in the scene. With this aim a few results for different images have been generated (Fig. 6) and assessed (Table- II). These results comprise different aspects such as non-preserving objects, miss–classifying road/background boundary pixels and different degrees of degradation.

This summary reveals that the selection of the best result (bold values) varies depending on the measure used. Once again, all these measures tend to the maximum value (one) when the result is really close to the ground truth (e.g., first image, fourth row). However, special attention should be paid on italic values. These rows reveals the differences in performance of the measures when cars are present in the scene: while RDI and RDI$_{veh.}$ vary depending on these errors, the rest of measures do not. Within these rows, RDI and RDI$_{veh}$ indexes are lower than the others due to the errors on the cars but not the other measures. As an example, in the first image, using accuracy, F or quality a

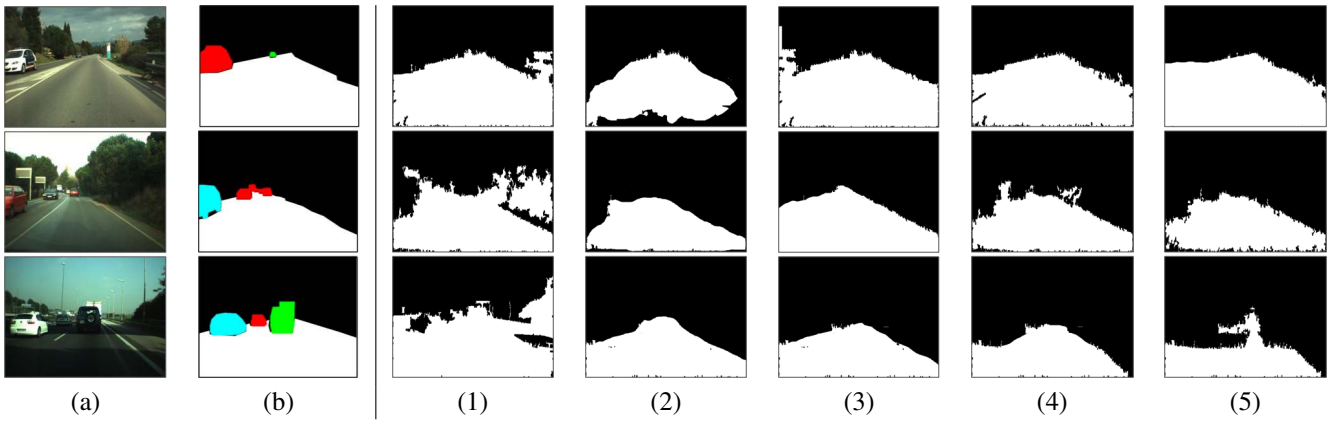

Fig. 6. Road detection results used to validate RDI. (a) original image. (b) Ground truth. (1) - (5) Different assessed results which represent either the variations of an individual algorithm or the results using different RD algorithms.

TABLE II
QUALITY ASSESSMENT OF ROAD DETECTION RESULTS SHOWN IN FIG. 6

		RDI	$RDI_{veh.}$	F	ACC	$\hat{g}$
	1	0.7914	0.7133	0.9581	0.9574	0.9195
	2	0.7562	0.7053	0.8417	0.8633	0.7267
Img. 1	*3*	*0.5021*	*0.3903*	*0.9618*	*0.9614*	*0.9265*
	4	**0.9890**	**0.9956**	**0.9749**	**0.9755**	**0.9510**
	5	0.8441	0.7822	0.9735	0.9734	0.9484
	1	0.8695	0.8840	0.8226	0.8493	0.6987
	2	0.9645	**0.9855**	0.9123	0.9421	0.8387
Img. 3	*3*	*0.7283*	*0.6369*	***0.9531***	*0.9648*	***0.9103***
	4	0.9752	0.9843	0.9520	**0.9654**	0.9084
	5	**0.9792**	0.9915	0.9448	0.9623	0.8954
	1	*0.4553*	*0.3581*	*0.8207*	*0.8380*	*0.6960*
	2	0.8253	0.7954	**0.8745**	0.9082	**0.7771**
Img. 4	3	**0.8909**	0.8964	0.8738	**0.9102**	0.7759
	4	0.8730	**0.8966**	0.8217	0.8793	0.6974
	5	0.7859	0.7974	0.7410	0.8333	0.5886

high performance is achieved while the closest car is almost undetected. RDI weights down this lack of precision and gives a low score. This effect is emphasized if more relevance is given to the object evaluation block (RDI_{veh}). Thereby, the proposed index is more suitable to highlight and differentiate between results.

B. Algorithm Optimization

The second experiment consists in tuning a given algorithm. A simply probability classifier on intensity values has been used to decide whether a pixel belongs or not to the road class (Fig. 7). The road model at each frame is built under the assumption that the bottom part of the image belongs to the road [3]. Thus, the parameter to be optimally fixed is λ, the threshold probability of being road (it runs in [0,1]). The lower threshold the more permissive classifier.

To properly select λ, the algorithm has been run using different threshold values and its results have been analyzed.

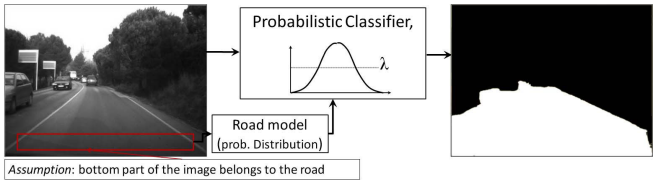

Fig. 7. Simple road detection algorithm. The pixels (intensity values) of the incoming image are classified as road (white) depending on their probability to belonging or not to the road class. The road model is built using the lower part of the image.

The λ value which produces the highest score is the optimal λ for the given image. Since different measures have been used, different optimal thresholds have been obtained (Fig. 8). As shown in Fig. 8 the threshold obtained using accuracy, quality of F are more permissive than RDI. Nevertheless, is one of those thresholds is used the car present in the scene is not preserved (top right result). RDI imposes a higher threshold value (more restrictive) which preserves the car in the scene (bottom right example). In addition, RDI clearly differentiate those results which do not hold the requirements (bottom left example).

However, our aim is not optimizing a single image but obtaining the optimal performance for a complete sequence of images that can be thought as a training set for the algorithm under development. Therefore, instead of looking for the maximum score on a given image, the maximum of the averaged values for each threshold for all the training images has been used (Fig. 9). Reported results (Fig. 10) show the ability of preserving objects when the RDI index is used (red segmentation) in front of the lack of precision when other measures are used (yellow segmentation).

V. CONCLUSIONS

In this paper, we have defined a novel index to quantitatively assess the performance of road detection algorithms. The proposed index is focused on specific requirements of the final application and specific properties of the scenes imaged by these algorithms rather than relying on a general

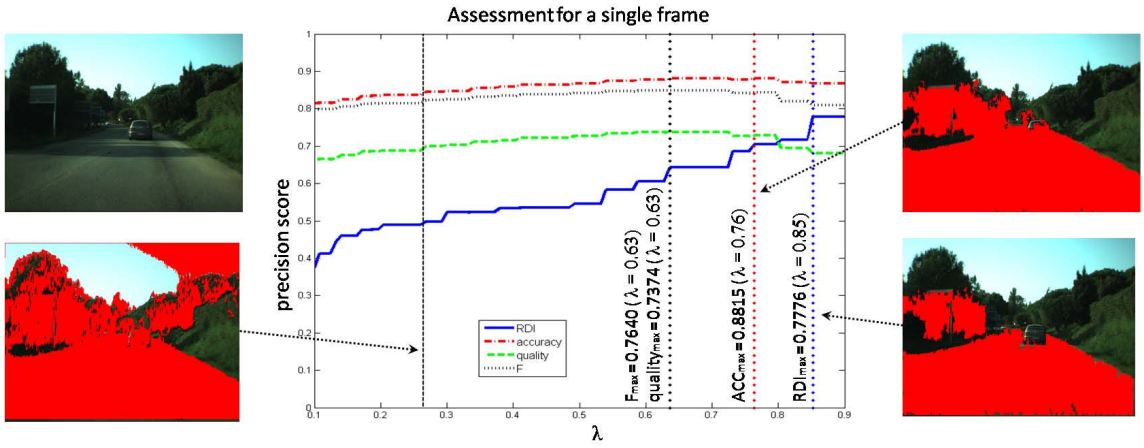

Fig. 8. Different scores obtained by varying the parameters of an individual algorithm on a single image. The optimal values of λ (highest precision scores) vary depending on the measure used. Results on the right show the differences between setting λ using quality ($\hat{g}$) (top) and RDI (bottom) for a given image (top left). RDI score weights down those results which do not preserve objects in the scene (bottom left).

criterion to evaluate them. The index combines information from different evaluations so it is capable to deal with major problems such as boundary ambiguity and application dependency. In addition, the index provides a perceptual evaluation which corresponds with the different error perception depending on the position of the pixel in the image.

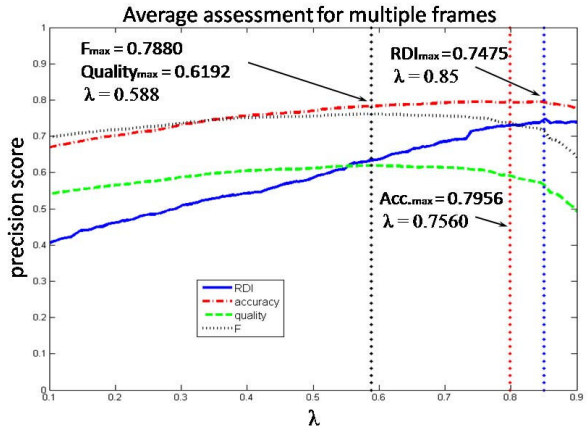

Fig. 9. Assessment scores for all possible thresholds are averaged for all the training images. Different measures yield different optimal thresholds.

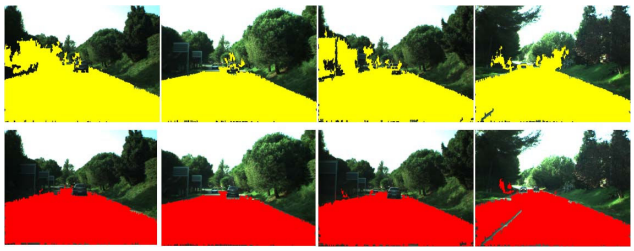

Fig. 10. Different results obtained when the threshold is selected using $\hat{g}$ (upper row) and RDI (lower row).

Experiments have proved the benefits of using this index instead of other existing measures. It is suitable for deciding not only the best road algorithm from an application point of view, but it is also suitable for characterizing and tuning an individual algorithm.

In the future, we aim to use this index to characterize and compare current road detection algorithms.

REFERENCES

[1] F. Dornaika and A. D. Sappa, "Real time on board stereo camera pose through image registration," in *Intelligent Vehicles Symposium*, 2008.

[2] M. Sotelo, F. Rodriguez, L. Magdalena, L. Bergasa, and L. Boquete, "A color vision-based lane tracking system for autonomous driving in unmarked roads," *Autonomous Robots*, vol. 16, no. 1, 2004.

[3] J. M. Álvarez, A. M. López, and R. Baldrich, "Illuminant-invariant model-based road segmentation," in *Intelligent Vehicles Symposium*, 2008.

[4] C. Rotaru, T. Graf, and J. Zhang, "Extracting road features from color images using a cognitive approach," in *IEEE Intelligent Vehicles Symposium*, 2004.

[5] T. Hong, A. Takeuchi, M. Foedisch, and M. Shneier, "Performance evaluation of road detection and following algorithms," in *Proceedings of SPIE Optics East 2004*, vol. 25 - 28, October 2004.

[6] P. Conrad and M. Foedisch, "Performance evaluation of color based road detection using neural nets and support vector machines," *AIPR*, vol. 00, p. 157, 2003.

[7] C. Heipke, H. Mayer, C. Wiedemann, and O. Jamet, "Evaluation of automatic road extraction," in *ISPRS Conference, Vol. 32, 3-2W3.*, 1997, pp. 47–56.

[8] P. Lombardi, M. Zanin, and S. Messelodi, "Switching models for vision-based on–board road detection," in *International IEEE Conference on Intelligent Transportation Systems*, 2005.

[9] E. D. Gelasca, T. Ebrahimi, M. C. Q. Farias, M. Carli, and S. K. Mitra, "Towards perceptually driven segmentation evaluation metrics," in *CVPRW '04: Proceedings of the 2004 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'04) Volume 4*. Washington, DC, USA: IEEE Computer Society, 2004, p. 52.

[10] P. Villegas and X. Marichal, "Perceptually-weighted evaluation criteria for segmentation masks in video sequences," vol. 13, no. 8, pp. 1092–1103, August 2004.

[11] C. D. Manning, P. Raghavan, and H. Schiütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008, ch. 16.3.

[12] D. R. Martin, "An empirical approach to grouping and segmentation," Ph.D. dissertation, EECS Department, University of California, Berkeley, Aug 2003.

[13] F. J. Estrada, "Advances in computational image segmentation and perceptual grouping," Ph.D. dissertation, Department of Computer Science, University of Toronto, June 2005.