Extending the Stixel World with Online Self-supervised Color Modeling for Road-versus-Obstacle Segmentation

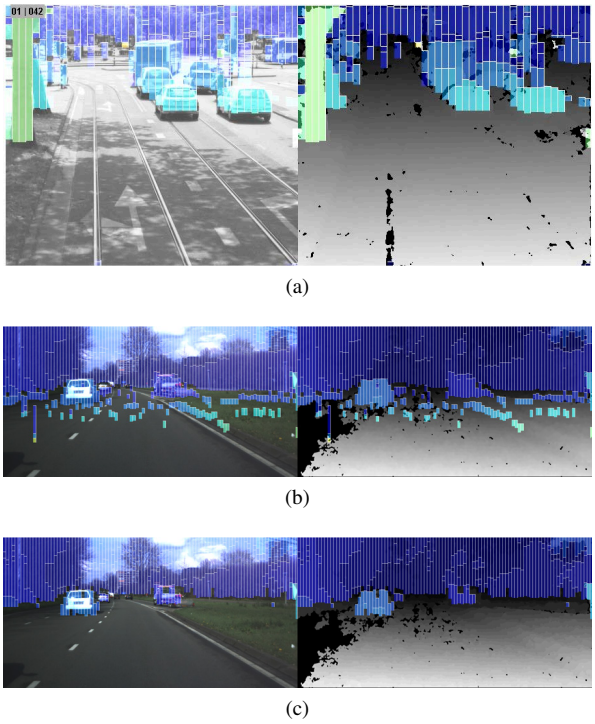
Willem P. Sanberg, Gijs Dubbelman and Peter H.N. de With

***Abstract*— This work concentrates on vision processing for ADAS and intelligent vehicle applications. We propose a color extension to the disparity-based Stixel World method, so that the road can be robustly distinguished from obstacles with respect to erroneous disparity measurements. Our extension learns color appearance models for road and obstacle classes in an online and self-supervised fashion. The algorithm is tightly integrated within the core of the optimization process of the original Stixel World, allowing for strong fusion of the disparity and color signals. We perform an extensive evaluation, including different self-supervised learning strategies and different color models. Our newly recorded, publicly available data set is intentionally focused on challenging traffic scenes with many low-texture regions, causing numerous disparity artifacts. In this evaluation, we increase the F-score of the drivable distance from 0.86 to 0.97, compared to a tuned version of the state-of-the-art baseline method. This clearly shows that our color extension increases the robustness of the Stixel World, by reducing the number of falsely detected obstacles while not deteriorating the detection of true obstacles.**

I. INTRODUCTION

In recent years, vehicles are made increasingly intelligent with so-called Advanced Driver Assistance Systems (ADAS). This development is expected to significantly reduce traffic accidents, traffic congestion and fuel consumption simultaneously. To ensure traffic safety, ADAS can e.g. indicate the location of potentially hazardous obstacles to the driver and the position of safely drivable road. On the longer term, ADAS and related technologies will allow the development of fully autonomous vehicles. In this work, we improve a state-of-the-art vision-based road-versus-obstacle detection system by exploiting a strong fusion of multiple image modalities.

To robustly facilitate situational awareness at a moving platform, several complementary sensor modalities should be employed. These modalities can include RADAR, LIDAR, ultrasound, and (thermal) imaging. The benefit of using vision-based systems is that they provide dense scene information in a cost-effective way. Image data is also a rich source of information, since it comprises of several submodalities. For stereo-based imaging, these informative features include *disparity, texture, color, shape*, and *optical flow*. All these submodalities can contribute to a robust situational analysis, such as e.g. the detection of partially occluded pedestrians who are about to cross the streets. Although LIDAR, RADAR or ultrasound provide valuable complementary information, in this work we solely focus on vision-based detection systems.



(a)

(b)

(c)

Fig. 1. Stixel segmentation results, superimposed on the left camera image (left) and the corresponding disparity image (right). Color indicates distance: blue is far away, green mid range and red is close by. Panel (a) illustrates good segmentation results, acquired with an HD resolution stereo camera with a 1.5-m baseline, panel (b) shows examples obtained with a a medium-quality stereo camera (with 0.12-m baseline and 1024×768 pixels resolution, lower dynamic range), resulting in false negatives in the road detection. In general, these false negatives are caused by inconsistencies in the disparity signal. Our contribution is illustrated in (c), where the same frame is segmented using our stixel algorithm. The disparity is fused with color, removing most of the inaccuracies.

Many vision-based ADAS systems are already used in practice. These are mainly monocular systems that detect pedestrians [1], lane markings [2], and traffic signs [3]. These systems are well suited for current ADAS solutions. Fully autonomous vehicles require more dense, accurate, and more reliable sources of scene information, in particular 3-D information. To obtain 3-D information, high-end laser-based systems (accompanied by RTK-GPS) are typically used [4], [5]. The current challenge in vision-based ADAS, is to be

Willem Sanberg, Gijs Dubbelman and Peter de With are with the Department of Electrical Engineering, Video Coding and Architectures Research Group, Eindhoven University of Technology, 5600 MB Eindhoven, The Netherlands w.p.sanberg@tue.nl

The research leading to these results has received funding from the European Unions Seventh Framework Programme managed by REA Research Executive Agency http://ec.europa.eu/research/rea(FP7/2007-2013) under grant agreement no *FP7-SME-2012*.

able to deliver comparable semantic 3-D scene information, at a more affordable price point than that of high-end laser-based systems.

Multi-view image processing, in particular stereo vision, has the potential to fulfill these requirements. In stereo vision, the disparity, which is analogous to depth, can be estimated densely and in real-time [6]. This gives a direct description of the geometry of the scene and it facilitates, for example, a separation of flat, drivable surfaces from erect obstacles [7], [8]. A state-of-the-art approach for this is called the Stixel World method [9]. This is a fully probabilistic framework to distinguish ground from obstacles in the disparity signal, and it can be implemented efficiently given several assumptions. This framework is generally more flexible and more robust than its predecessors, Figures 1a and 1b illustrate typical results. Section II discusses this method in more detail.

A common deficiency of all disparity-based methods is that their analysis relies on the single modality *disparity*, even though that modality generally suffers from errors such as noise, strong outliers and holes due to occlusions, or due to little texture information in large image regions. An example of these issues is depicted in Fig. 1b. Although these issues can be addressed to a certain extent by using high-quality cameras and more advanced disparity estimation, they can never be fully resolved, since traffic scenes will often contain image areas with, for instance, low illumination, shadows, sunny reflections or motion blur.

It is evident that the fusion of the estimated disparity with other (image) modalities is advantageous for obtaining more reliable information. Optical flow and texture analysis typically suffer from the same challenges as disparity estimation, as all require well-textured image regions. More orthogonal and complementary image modalities are therefore color, shape, and appearance. The color modality is dense by nature and is often used to automatically distinguish the road on which the vehicle is driving from its surroundings [10], [11], [12], which is also the focus of our work.

In recent work, advantages of the combination of using depth information with color information has been shown [13], [14], [15]. A particular interesting strategy is to use the dense disparity-based depth information, to on-line learn a color-based road-versus-obstacle model in a self-supervised manner [16], [15]. This model is used to classify image regions as either road or obstacle, which is then combined with the disparity-based analysis. This combination is typically performed with rather straightforward methods of fusion, and can be as simple as using the disparity analysis up to a distance from the vehicle and the color-based analysis after this distance [4].

A key property of the Stixel World method of [9] is that it allows for non-linear fusion of different dense (image) modalities in one probabilistic framework. Instead of analyzing each modality separately and then combining their results [13], [15] (i.e. *weak fusion*), the Stixel World method allows for an efficient analysis on the basis of all modalities simultaneously (i.e. *strong fusion*). To our knowledge, this property has received little attention in literature. In this work, we exploit this promising line of research, by combining disparity and color modalities for road-versus-obstacle image segmentation, where the color model is learned on-line and is executed in a self-supervised mode. Besides showing the advantages of strong fusion of these particular image modalities, our work is also informative for those who are interested in fusion of other modalities within the Stixel World framework.

The remainder of this paper is structured as follows. First, we will provide a short description of the disparity-based Stixel World in Section II, since it serves as a basis of our work. Our main contribution is put forward in Section III, where we describe how we fuse color information into an extended stixel framework. In Section IV we elaborate on our evaluation approach, including our publicly available data set, experiments and results. Lastly, conclusions are provided in Section V.

II. THE DISPARITY STIXEL WORLD

Let us now give a short overview of the Stixel World framework as it is presented in [9], which we use as a basis of our work. The main goal of stixel segmentation is to find the optimal labeling L^* of vertically stacked, piecewise planar ground or obstacle segments for the input disparity data $\mathbb{D}$. Finding L^* can be formulated as a MAP estimation problem as in (1), which can be solved efficiently using Dynamic Programming. Using Bayes' theorem and assuming (a) that columns are independent, (b) that disparity measurements $d_{u,v} \in \mathbb{D}$ at individual pixels (u,v) are independent and (c) that data within D_u is independent from the labeling in other columns, the posterior probability can be written as in (2). Here, u is the column index and w the image width. The probability $P(L_u)$ models a-priori world knowledge to constrain the labeling to avoid dispensable segments and physically unlikely situations. This world model offers a way to regularize the results for image-column optimality, whereas the methods of [7] and [8] potentially lead to sub-optimal results, since they analyze data mostly locally. The details concerning $P(L)$ are presented in [9]. Finally, in (3), the likelihood of the data given a certain labeling is provided, where n is the segment index, N_u the number of segments in L_u, and v_n^b and v_n^t the bottom and top row-index of segment s_n that has a label $l_n \in \{g,o\}$, representing the ground and obstacle classes, respectively. The previously mentioned equations are specified by

$$L^* = \arg\max_{L \in \mathbb{L}} P(L|\mathbb{D}), \tag{1}$$

$$P(L|\mathbb{D}) \sim \prod_{u=0}^{w-1} P(D_u|L_u) \cdot P(L_u), \tag{2}$$

$$P(D_u|L_u) \sim \prod_{n=1}^{N_u} \prod_{v=v_n^b}^{v_n^t} P(d_v|s_n,v). \tag{3}$$

The next step is accounting for invalid disparity measurements $d_v \notin [d_{min}, d_{max}]$. These will occur, for example, if the estimator cannot find a match in the stereo frames. To this

end, a probability of encountering a non-valid measurement is defined, p_{invalid}, as well as the probabilities that such a pixel will represent either ground or obstacle, $p(l_n|\text{invalid})$. With this, we can calculate the probability of invalid data for each class using Bayes' rule: $p^{l_n}_{\text{invalid}} = p(\text{invalid}|l_n) = p(l_n|\text{invalid}) \cdot p(\text{invalid}) / p(l_n)$. This then leads to

$$P(d_v|s_n,v) = \begin{cases} P_D(d_v|s_n,v) \cdot (1 - p^{l_n}_{\text{invalid}}) & \text{for valid } d_v \\ p^{l_n}_{\text{invalid}} & \text{otherwise.} \end{cases} \quad (4)$$

Here, $P_D(d_v|s_n,v)$ represents the probability of a single valid disparity measurement d_v at a certain row v, assuming that it would belong to a potential segment s_n. The distribution $P_D(d_v|s_n,v)$ is modeled as a mixture model that consists of a uniform distribution to handle outliers and a Gaussian distribution to model how well the measurement fits the potential segment:

$$P_D(d_v|s_n,v) = \frac{p_{out}}{d_{min} - d_{max}} + \frac{1 - p_{out}}{A_{norm}} e^{-\frac{1}{2}\left(\frac{d_v - f_n(v)}{\sigma^{l_n}(f_n,v)}\right)^2}. \quad (5)$$

In (5), p_{out} is the fixed probability of encountering an outlier. The normalization term A_{norm} and the modeled standard deviation σ^{l_n} are defined in [9]. The remaining term, $f_n(v)$, models the expected disparity within a segment for ground and object segments. For objects, $f^o_n(v) = \mu_n$ is adopted, assuming a fronto-parallel object surface at the mean disparity of the segment. For ground segments, $f^g_n(v) = \alpha \cdot (v_{horizon} - v)$ is used, assuming a linear ground plane surface with a slope α.

III. COLOR EXTENSION

As a key contribution, we incorporate a color signal $\mathbb{C}$ in the Stixel World model. To this end, the data term of (1)-(4) should now reflect both color and disparity information. Starting from $P(L|\mathbb{D},\mathbb{C})$, a derivation can be made analogous to the description in Section II. If we additionally assume that disparity and color modalities are independent, we can rewrite the data term of the likelihood as

$$P(D_u,C_u|L_u) \sim \prod_{n=1}^{N_u} \prod_{v=v^b_n}^{v^t_n} P(d_v|s_n,v) \cdot P(c_v|s_n,v), \quad (6)$$

where segment s_n has a label $l_n \in \{g,o\}$, as before. Note that the term $P(d_v|s_n,v)$ also incorporates the construction for invalid disparity measurements as in (4) but is left out here for compactness. Furthermore, we do not alter the definition of the world model $P(L)$ but focus on defining a suitable color model within $P(c_v|s_n,v)$. This term should capture the probability of a certain color measurement given a certain segment label. We let this be independent of the position v of the segment and merely consider the label of a segment, so that $P(c_v|s_n,v) = P(c_v|l_n)$. This is a reasonable simplification since $P(L)$ already constraints physically unlikely segmentations and we can assume that the color of the road surface is roughly constant within the image. When optimizing the Stixel World cost function $P(L|\mathbb{D})$, we use a weighing factor λ between the cost of the disparity term $P(d_v|s_n,v)$ and that of the color term $P(c_v|s_n,v)$. This is required to compensate for the oversimplification in the modeling of their theoretical joint probability density in two dis-joint terms, which use different probabilistic methods (i.e. Gaussian distributions and histograms).

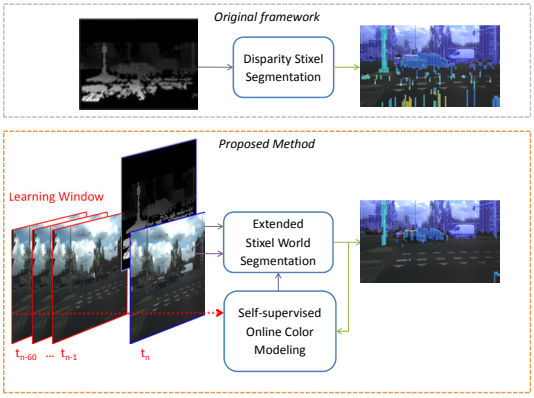


Fig. 2. The original stixel framework (top), relies on disparity images alone. In contrast, our proposed extension (bottom) exploits both disparity and color information.

Algorithm 1 Segmentation in the Extended Stixel World

Input: image I_{t_n}; disparity $\mathbb{D}$; learning window LW;

[Learn Color Models]
for each $t \in LW$ **do**
  $C_t \leftarrow \text{TransformRGB2Color}(I_t)$
  for $l \in \{ground, obstacle\}$ **do**
    $TM^l_t \leftarrow \text{GenerateTrainingMask}(L^*_t, TM^l_{prior})$
    $X^l_t \leftarrow \text{ExtractSamples}(C_t, TM^l_t)$
    $H^l_{t_0} \leftarrow \text{AddToHistogram}(H^l_{t_0}, X^l_t)$
  end for
end for
$P(C|l) \leftarrow \text{NormalizeHistogram}(H^l_{t_0})$

[Process Current Frame]
$\mathbb{C} \leftarrow \text{TransformRGB2Color}(I_{t_0})$
$L^* \leftarrow \text{StixelSegmentation}(\mathbb{D}, \mathbb{C}, P(C|l))$

Output: Optimal Labeling L^*

The main steps of our processing framework are conceptually captured in Fig. 2 and Algorithm 1. It comprises of two main steps: learning color models for road and obstacle image areas (in Algorithm 1: [Learn Color Models]) and segmenting the current frame using the Extended Stixel World method (In Algorithm 1: [Process Current Frame]). These will be addressed in detail in the next subsections.

A. Color Representation

A common approach in color analysis is employing color histograms as dense area descriptors. Color histograms can be defined with linear, non-linear or adaptive binning strategies. We apply the adaptive binning strategy *minimum variance quantization*, also known as *median cut quantization*, as described in [17]. This ensures that we optimally adapt the borders of our histogram bins to the color signal of a certain

traffic scene efficiently and accurately. We will compare this strategy to relying on unadapted, linearly spaced bins in our experiments.

In our analysis, we will mainly focus on the RGB color space since it yields good results in numerous color based experiments. We will compare this approach to a HS-based approach in several tests as well. The function that transforms each RGB image frame I_t to the desired color representation $\mathbb{C}_t$ is indicated as the 'TransformRGB2Color' in Algorithm 1.

B. Self-supervised Online Learning of Color Models

Since we aim at a system that is highly adaptable to different traffic environments, we will learn our color models $P(c|l)$ online. This is an intuitive approach, since an offline learning strategy would require a single color model that is both general enough to be applicable to all potential road appearances and simultaneously discriminative enough to always separate that road from its surroundings. Our online learning approach is indicated in Algorithm 1 under the section 'Learn Color Models'.

In our framework, a learning window *LW* is defined, containing 1 or more frames that precede the current frame at $t = t_n$ with a maximum range of 60 frames back in time, denoted by t_{n-60}. These frames are transformed to the quantized color space. From this signal, training samples are selected that are believed to be representative of either the road or the obstacle class. These samples are then used to fill and normalize a color histogram for each class, providing the required $P(c_v|l_n)$.

To select ground and obstacle training samples from the preceding frames I_t within *LW*, we need to generate a *training mask* for each frame and each class $l \in \{g, o\}$, denoted by TM_t^l. To this end, we exploit the fact that each previous frame was already analyzed and segmented by our system at its corresponding time t. This process results in an estimate of obstacle and ground areas in each frame, to which we refer to as segmentation masks (*SegmMask*). Note that at the start of a sequence, the color model $P(c|l)$ is not yet learned and, hence, taken as a uniform distribution. As a consequence, the first frames are effectively segmented using only their disparity signal.

We explore two strategies to create training masks: we use the full *SegmMask* or we *Intersect* it with a prior mask. For road samples, this is illustrated in Fig. 3 with an example image (top left) and the corresponding estimation of the ground area, as provided by its (disparity-based) segmentation result (Fig. 3, middle left). We define the prior mask (TM^g_{prior}) as a fixed trapezoid in the bottom center of the image (Fig. 3, top right). By intersecting this prior mask with the segmentation result, we acquire a mask that contains the road area directly in front of the car, excluding detected obstacles (Fig. 3, middle right).

To generate a *training mask* to extract obstacle training samples, we employ a comparable strategy. The *SegmMask* is the inverted version of the one for ground (i.e., the black regions in Fig. 3 middle-left). For the *Intersect*, we apply a mask containing the area below the horizon. This makes the color modeling of obstacles more adapted to obstacles that are on the road, which are more relevant to detect than tree-leaves or rooftops.

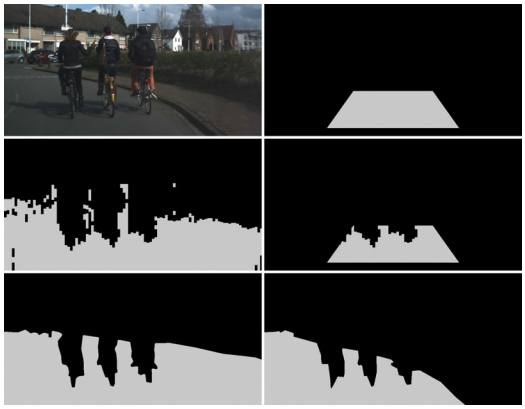


Fig. 3. Illustration of our *training mask* approach to acquire road samples. Top-left: input image; top right: TM^g_{prior}; middle left: ground segmentation mask; middle right: intersection of the two, TM^g_t. The bottom row illustrates the different ground-truth annotations, with left *drivable surface* and right *road*.

IV. EVALUATION

The goal of our evaluation is to show the benefit of our color extension and to assess its crucial design choices. These are: (a) the color space specifications, (b) the strategy of selecting training samples with TM^g_t and TM^o_t, and (c) the position and range of the learning window *LW*. The results give important insights on the most effective application of our methods and their efficiency in real-world situations. All our results will be compared to results that are acquired with the baseline approach of [9].

A. Test Setup

We have acquired an extensive data set in an urban environment, using a BumbleBee2 camera, mounted behind the windshield of a car, just below the rear view mirror. The camera has a baseline of 12 cm, a resolution of 1024x768 pixels and a frame rate of 20 FPS. This data set is made publicly available[1].

We have selected 74 representative frames from the set and manually annotated both road and drivable surface areas, as illustrated in the bottom row of Fig. 3. The frames contain a large variety of relevant traffic situations, such as small, crowded streets with cyclists, road repair sites, large crossings and high ways. It contains both asphalt and paved roads of several colors (black, gray, red) and both frames with low illumination due to heavily clouded skies or trees and frames with high illumination from clear sky with sunny reflections. Several example frames are provided in Fig. 4.

To obtain disparity measurements, we employ a multi-threaded version of the OpenCV's implementation of the Semi Global Block Matching algorithm of [18]. Due to

[1] http://www.willemsanberg.net/research

the many low-texture image regions in our data set, we have found that a matching window size 7×7 pixels and smoothing parameters $p_1 = 16 \cdot 7^2$ and $p_2 = 8 \cdot p_1$ provide the most acceptable results. We also employ a *winner margin* of 20, to force the algorithm to have a higher precision at the cost of recall. This is beneficiary for the baseline Stixel World method, since it can handle missing values better than erroneous ones. This can be seen as a simplification of the work presented in [19], in which disparity estimates are accompanied by a confidence measure to adaptively set a outlier probability. In our approach, this confidence is binary with a relatively strict threshold based on the *winner margin*.

As described, our camera has lower resolution and a smaller baseline than, for example, the camera used for the KITTI benchmark data set [20], resulting in lower quality disparity estimates. To compensate for this deficiency and to obtain more favorable results for the baseline method, we have made some improvements to the baseline framework. For instance, we learn the ground plane model $f_n^g(v)$ on-line instead of using a single fixed model. To this end, we exploit a v-disparity representation such as in [8] for several vertical slices of each frame, making our system more robust against ground plane deviations over time and non-horizontal ground areas.

Moreover, we have tuned the label-based transition probabilities defined in $P(L)$ to boost the performance of the baseline method even further. Finally, we have added an artificial ground segment to the bottom of each stixel, denoted with *Seg0*. This segment represents the area below the camera view, which can safely be assumed to be road in this context, and it reduces false detections in the lower image regions due to noisy disparity estimates. To show the value of these additions, we report the performance of each of these three disparity baselines (original settings, tuned transition probabilities and with the artificial *Seg0* as ground).

The relevant Stixel World parameters, as described in Section II, are set as follows throughout all experiments: $p_{out} = 0.25$; $p_{\text{invalid}} = 0.25$; $p_g^{\text{invalid}} = 0.55$; $p_o^{\text{invalid}} = 0.45$; $p_{\text{g}} = p_{\text{o}} = 0.5$; $d_{min} = 1$; $d_{max} = 32$. Furthermore, we have adopted a stixel width of 10 columns and subsample the disparity and color signals vertically with a factor of 3 prior to segmentation. Note that we exploit the full image data to compute look-up tables and color models, which is comparable to the approach in [9]. The research version of the Extended Stixel World method is a MATLAB-based implementation. The added complexity of the color-processing is small compared to the disparity-analysis baseline complexity. Therefore, it is safe to assume that our proposed extension can executed as a real-time system, as can the original one [9].

B. Scoring Metrics

Inspired by the work of Fritsch *et al.* on performance metrics for road detection algorithms [21], we evaluate our road detection algorithm in a Bird's Eye View (BEV) representation of the scene. A BEV representation is corrected for geometric distortion to avoid that pixels near the car outweigh pixels further away in the segmentation score. In this representation, we have employed several metrics to assess the performance of our algorithm. First of all, we measure the recall and precision of the road area. We do this with the road annotation as a reference. However, pixels that are drivable but not belong to the road are ignored in the evaluation, such as curbs and grass. The purpose of this strategy is to assess in which areas improvements or deteriorations of the results appear. Since the goal of this work is to improve the road segmentation, improving the recall of the road region is the most important. If this results in a lower precision score, it is acceptable as long as that occurs mostly in drivable surfaces that are not road. In this evaluation, we consider the area up to 30 meters in front of the vehicle.

As the stixel representation approximates area contours with rectangular shapes, it is not possible to achieve a pixel-accurate segmentation. Consequently, achieving perfect recall and precision is also not realistic. To determine how close the performance of the tested methods come to the maximum attainable performance, we estimate realistic optimal recall and precision levels by eroding and dilating the ground truth segmentation masks by a square kernel with dimensions similar to the stixel width.

Next to this pixel-based metric that is vision inspired, we assess the added value of our method with a more practical application in mind, relevant to our context: measuring where a vehicle can drive. This can also be measured in the BEV representation of our road segmentation results. In that representation, we measure how far a vehicle can drive by calculating where the first object is that a vehicle of average width would drive into. Using the ground-truth annotations, we can define recall and precision scores, which take into consideration a 5 meter safety margin around obstacles. Recall indicates how much of the ground-truth drivable distance is detected correctly. The recall will be lower than unity when a false obstacle is detected in front of the first real obstacle. The precision represents how much of the detected drivable distance is correct. If the real obstacle is missed, the precision will be lower than unity. Consequently, for each frame either the recall or precision of the drivable distance is always equal to unity. This evaluation is performed over a range of up to 50 meters. For these metrics, we also calculate the resulting F-score, which is defined as the harmonic mean of recall and precision.

C. Experiments and Results

In the coming sections, the experiments assessing the critical design choices of our Extended Stixel World method are presented, together with their quantitative results. Fig. 4 shows qualitative results and illustrates the road segmentation scoring metric.

1) Color Model: We have evaluated several different color spaces and settings. We have varied the number of clusters k that is used to approximate the colors in the current learning window and the color-data weighing factor λ, for

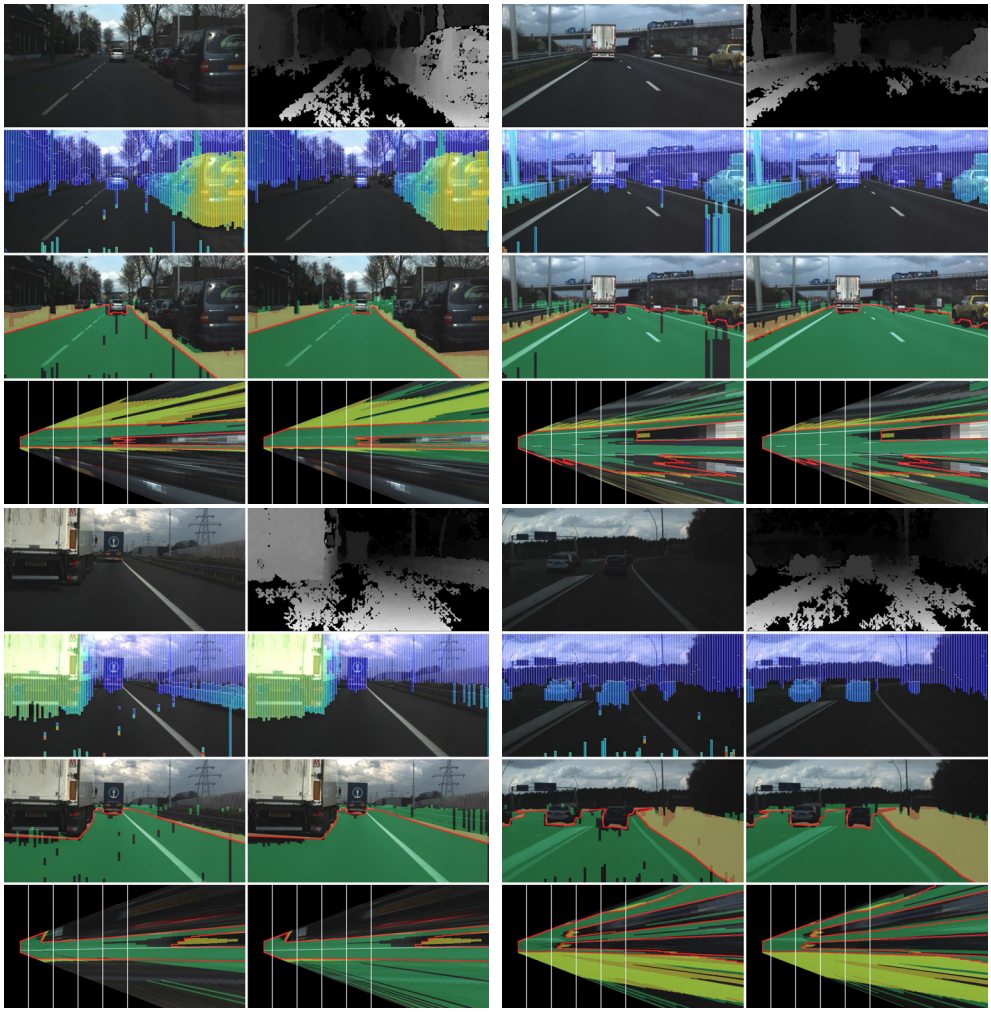


Fig. 4. Qualitative results of our proposed method. For each frame, we show eight images. The top row contains the rectified and cropped left camera image and the corresponding disparity image. Below that, we present the baseline result in the left column and our results in the right column. We show three representations: first the stixel results, where color depicts obstacle distance (red (close) to blue (far)); then an overlay of the ground mask (in green) with the road ground truth outlined in red, and ignored pixels that are drivable but not road in orange. Finally, we show the BEV of the result with lines at ranges of 10, 20, 30, 40 and 50 meters. All frames show that our method improves the recall of road regions, even with relatively low light conditions.

both indexed RGB and linear-binned HS representations. For linear binning, we allow k bins per dimension, resulting in k^2 bins for HS.

The most relevant results of these experiments are provided in Fig. 5 and Table I. Based on the results illustrated in Fig. 5, the HS color representation (bright green crosses) performs better at increasing the recall, and RGB (cyan crosses) tends to improve the precision of the road segmentation. The experiments with high a λ and a low k generally led to deterioration of the (tuned) baseline results. This makes sense, since the color representation cannot contain much discriminating information, and yet the confidence is boosted, leading to false results. We have obtained the best results with an indexed RGB color model, using $k = 64$ and $\lambda = 4$ ($F = 0.968$). The closest HS-based score ($F = 0.967$) was obtained using $k = 16$ and $\lambda = 2$, where we ignored experiments with worse precision than the baseline, since road-precision is key for safe ADAS. Although these pixel-based scores are very similar, the RGB run outperforms the HS-run 0.919 to 0.861 in the recall of the drivable distance (Table I). This means that more false obstacles are detected with the HS experiment, even though the HS model uses 16×16 bins and the RGB only 64. In the left graph of Fig. 6, two runs with $\lambda = 1$ are shown. These perform similar or worse than the (tuned) baseline method, illustrating the importance of correct normalization when fusing different signal modalities.

2) Training Mask: To evaluate our choice of *training mask* (TM_t^l) for the selection of training samples, we compare the two approaches described in Section III-B. This involves using the segmentation mask alone, or intersecting it with a fixed, a-priori defined trapezoid in front of the vehicle (for ground samples) or the area below the horizon (for obstacle samples). In the middle graph of Fig. 6, the results of these strategies are shown, using the best settings for RGB and HS, as found in Section IV-C.1. In this metric,

the training mask has little influence, since all graphs nearly overlap. However, in the ROC plot (Fig. 5), the influence is remarkable, visible from the dotted green (HS) and blue (RGB) lines. The central points are the experiments with one segmentation mask and one intersection. The scores with high recall are obtained using the segmentation masks for both ground and obstacles, and the scores with high precision are obtained using the intersection masks for both classes. So, the system becomes more conservative (higher precision) when the color models are focused on relevant areas. We have selected the use of the segmentation mask for ground samples and the intersect method for obstacle samples, since it results in the highest F-score and is a good, rational compromise.

3) Learning Window: The most relevant settings of the learning window are its range and position with respect to the new frame. First, we vary the length of the learning window by extending it further into the past with a maximum of 60 frames before the current frame (equivalent to 3 seconds ago). With this sub-experiment, we will validate if the added complexity of taking more frames into account translates into more robustness. Next, we analyze whether or not it is possible to leave a gap between the frames in the learning window and the frame currently under analysis. This is an important sub-experiment, since if there is more time available to analyze the frames of the learning window, either the constraints on execution time can be loosened, or more complex algorithms can be employed. As a third sub-experiment on the learning window parameters, we limit the range to a single frame and vary its position. Effectively, this combines the extreme cases of the first experiment (varying the length) with the idea of the second (leaving a gap). In the most extreme case of this third experiment, a color model is learned on a single frame, 60 frames back in time. We found that the effects of the LW settings are similar for both our metrics. In Fig. 5, the results are marked magenta. They mostly overlap, even though we tested several extreme cases. In the right graph of Fig. 6, the runs with the 3-seconds-old frame are marked explicitly. They perform slightly worse, but still outperform the baseline with more than 25%. This signifies that the system is very flexible in the selection of LW frames: a single frame, somewhere from the last 3 seconds, can provide enough information to learn a reliable color model for road area.

V. CONCLUSIONS

We have presented a color extension to the disparity-based Stixel World algorithm, to more robustly segment road versus obstacles in traffic scenes by on-line learning color models in a self-supervised way. This extension particularly improves the robustness of the segmentation against erroneous disparity estimates, which inevitably occur during challenging low-texture imaging situations, regardless of the quality of the stereo camera being used. Fusing the disparity information with other (image) modalities, as is done in this work, is

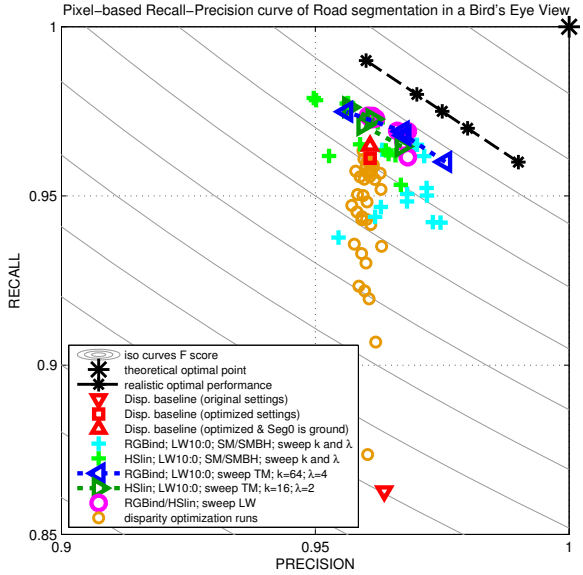


Fig. 5. Resulting ROC plot of the described experiments. Note that the range of the axis is only $[0.9,1]$ and $[0.85,1]$. This figure is best viewed in color. Explanation of the legend: LW[start:end]: Learning Window with range $[t_0 - t_{start}, t_0 - t_{end}]$; TM: training mask, which can be the segmentation mask or the intersection of that with a prior mask. The scores are based on the *road* ground-truth annotation and ignore pixels that are drivable but not part of the road region, such as grass and pavement.

therefore crucial for safe and reliable road-versus-obstacle segmentation.

Critical design choices for our color extension have been evaluated on a data set that was specifically focused on challenging imaging situations. The correct selection of color model settings, such as the number of bins and the normalization factor for data fusion, is shown to be crucial to increase the segmentation performance. The use of an on-line optimized indexed color representations allows for highly descriptive and efficient color models for road and obstacle classes. Moreover, it was shown with our experiments that even a single frame, captured seconds earlier, provides our system with sufficiently reliable color information. This offers opportunities for time-efficient or more complex color modeling, if required.

The combination of these aspects result in an increased pixel-based F-score on road segmentation from 0.96 to 0.97, compared to a heavily optimized baseline method. Without optimization, the baseline method scored 0.91. In detecting drivable distance, our method increases the F-score from 0.86 to 0.97. These results clearly show that our Extended Stixel World method, based on strong fusion of disparity and color modalities, is an accurate and robust method for road versus obstacle segmentation.

REFERENCES

[1] M. Enzweiler and D. M. Gavrila, "Monocular pedestrian detection: survey and experiments." *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 31, no. 12, pp. 2179–95, Dec. 2009.

[2] Z. Kim, "Robust lane detection and tracking in challenging scenarios," in *IEEE Trans.on Intelligent Transportation Systems (TITS)*, vol. 9, 2008, pp. 16–26.

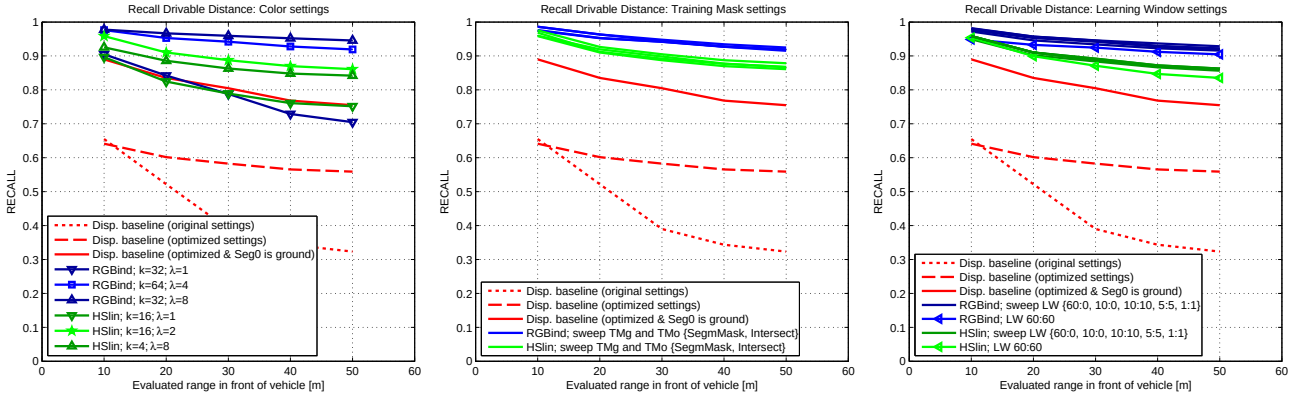


Fig. 6. Recall plot of the drivable distance for several runs, evaluated over increasing ranges. Average precision scores are not shown, since they are 0.997 ± 0.004 over all frames, runs and ranges (meaning that our methods very rarely miss a true obstacle within 50 meters). Explanation of the legend: LW[start:end]: Learning Window with range $[t_0 - t_{start}, t_0 - t_{end}]$; TM: training mask, which can be the segmentation mask or the intersection of that with a prior mask; k: the number of bins (for HS: k bins per dimension).

TABLE I

OVERVIEW OF QUANTITATIVE RESULTS FOR DIFFERENT SETTINGS AND OUR TWO SCORING METRICS

						Road Segmentation			Drivable distance ($\leq$ 50m)		
						F-score	Recall	Precision	F-score	Recall	Precision
Disparity baseline (original settings)						0.910	0.863	0.964	0.489	0.323	**1.000**
Disparity baseline (optimized settings)						0.961	0.961	0.961	0.717	0.559	**1.000**
Disparity baseline (optimized & Seg0 is ground)						0.963	0.965	0.961	0.858	0.755	0.993
Color space	LW	TMg	TMo	k	λ						
indexed RGB	10:0	SegmMask	Intersect	64	4	**0.968**	0.969	0.967	0.956	0.919	0.995
indexed RGB	10:0	SegmMask	Intersect	32	8	0.963	0.964	0.963	**0.968**	**0.946**	0.992
linear HS	10:0	SegmMask	Intersect	16x16	2	0.967	0.973	0.961	0.923	0.861	0.995
indexed RGB	10:0	SegmMask	SegmMask	64	4	0.965	**0.975**	0.956	0.958	0.925	0.995
indexed RGB	10:0	Intersect	Intersect	64	4	**0.968**	0.960	**0.976**	0.956	0.915	**1.000**
indexed RGB	60:60	SegmMask	Intersect	64	4	0.965	0.961	0.968	0.950	0.905	0.999
linear HS	60:60	SegmMask	Intersect	16x16	2	0.967	0.973	0.961	0.909	0.835	0.997

[3] A. De la Escalera, J. M. Armingol, and M. Mata, "Traffic sign recognition and analysis for intelligent vehicles," *Image and Vision Computing*, vol. 21, pp. 247–258, 2003.

[4] S. Thrun and M. Montemerlo, "Stanley: The robot that won the DARPA Grand Challenge," *J. of Field Robotics*, vol. 23, pp. 661–692, 2006.

[5] C. Urmson, C. Baker, J. Dolan, P. Rybski, B. Salesky, W. Whittaker, D. Ferguson, and M. Darms, "Autonomous driving in urban environments: Boss and the Urban Challenge," *AI magazine*, vol. 30, pp. 17–28, 2008.

[6] W. Van Der Mark and D. M. Gavrila, "Real-time dense stereo for intelligent vehicles," *IEEE Trans. on Intelligent Transportation Systems (TITS)*, vol. 7, no. 1, pp. 38–50, 2006.

[7] G. Dubbelman, W. van der Mark, J. C. van den Heuvel, and F. C. A. Groen, "Obstacle detection during day and night conditions using stereo vision," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*. IEEE, Oct. 2007, pp. 109–116.

[8] R. Labayrade, D. Aubert, and J.-P. Tarel, "Real time obstacle detection in stereovision on non flat road geometry through "v-disparity" representation," *IEEE Intelligent Vehicle Symp. (IV)*, vol. 2, pp. 646–651, 2002.

[9] D. Pfeiffer, "The Stixel World," Ph.D. dissertation, Humboldt-Universitat Berlin, 2011.

[10] J. M. Alvarez, T. Gevers, Y. LeCun, and A. M. Lopez, "Road Scene Segmentation from a Single Image," in *Eur. Conf. on Computer Vision (ECCV)*, 2012, pp. 376–389.

[11] J. M. Alvarez, M. Salzmann, and N. Barnes, "Learning appearance models for road detection," *IEEE Intelligent Vehicles Symp. (IV)*, no. Iv, pp. 423–429, June 2013.

[12] A. Neto and A. Victorino, "Real-time estimation of drivable image area based on monocular vision," in *IEEE Intelligent Vehicles Symp. (IV)*. Gold Coast, Australia: IEEE, 2013, pp. 63–68.

[13] C. G. Keller, M. Enzweiler, M. Rohrbach, D. F. Llorca, C. Schnörr, and D. M. Gavrila, "The benefits of dense stereo for pedestrian detection," *IEEE Trans. on Intelligent Transportation Systems (TITS)*, vol. 12, no. 4, pp. 1096–1106, 2011.

[14] T. Scharwächter, M. Enzweiler, U. Franke, and S. Roth, "Efficient Multi-Cue Scene Segmentation," in *LNCS Volume 8142, German Conf. on Pattern Recognition*. Springer-Verlag Berlin Heidelberg, 2013, pp. 435–445.

[15] M. L. L. Rompen, W. P. Sanberg, G. Dubbelman, and P. H. N. de With, "Online Self-supervised Learning for Road Detection," in *WIC/IEEE Symp. on Information Theory and Signal Processing in the Benelux (SITB)*. IEEE, May 2014, pp. 148–155.

[16] M. Bajracharya, A. Howard, L. H. Matthies, B. Tang, and M. Turmon, "Autonomous off-road navigation with end-to-end learning for the LAGR program," *J. of Field Robotics*, vol. 26, pp. 3–25, 2009.

[17] P. Heckbert, "Color image quantization for frame buffer display," *Computer Graphics*, vol. 16, no. 3, pp. 297–307, 1982.

[18] H. Hirschmüller, "Stereo processing by semiglobal matching and mutual information," *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 30, no. 2, pp. 328–41, Feb. 2008.

[19] D. Pfeiffer, S. Gehrig, and N. Schneider, "Exploiting the Power of Stereo Confidences," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Portland, USA, 2013.

[20] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2012, pp. 3354–3361.

[21] J. Fritsch, K. Tobias, and A. Geiger, "A New Performance Measure and Evaluation Benchmark for Road Detection Algorithms," in *IEEE Conf. on Intelligent Transportation Systems (ITSC)*. IEEE, 2013.