Neurocomputing 158 (2015) 174–183

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Adaptive road detection via context-aware label transfer

Qi Wang [a,*], Jianwu Fang [b,c], Yuan Yuan [b]

[a] Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, Shaanxi, PR China
[b] Center for OPTical IMagery Analysis and Learning (OPTIMAL), State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, Shaanxi, PR China
[c] University of Chinese Academy of Sciences, Beijing 100049, PR China

ARTICLE INFO

Article history:
Received 11 November 2014
Received in revised form
31 December 2014
Accepted 30 January 2015
Communicated by Qi Li
Available online 11 February 2015

Keywords:
Computer vision
Road detection
Depth map
Label transfer
Context-aware
MRF

ABSTRACT

The vision ability is fundamentally important for a mobile robot. Many aspects have been investigated during the past few years, but there still remain questions to be answered. This work mainly focuses on the task of road detection, which is considered as the first step for a robot to become moveable. The proposed method combines the depth clue with traditional RGB information and is divided into three steps: depth recovery and superpixel generation, weakly supervised SVM classification and context-aware label transfer. The main contributions made in this paper are (1) Design a novel superpixel based context-aware descriptor by utilizing depth map. (2) Conduct label transfer in an efficient nearest neighbor search and a temporal MRF model. (3) Update the learned model adaptively with the changing scene. Experimental results on a publicly available dataset justify the effectiveness of the proposed method.

1. Introduction

To autonomously navigate a robot in an outdoor environment, the vision system should be capable of perceiving the surrounding world. For example, the robot needs to know who he is interacting with [1], which way he could follow [2], and where he should stop to conduct his mission [3]. Among these abilities, road detection [4] is the primary one for a robot to become moveable. For this purpose, the road detection task needs to provide a clue of the drivable road in an input image or video so that the intelligent system can plan its path. In this paper, we put our focus on the road detection problem, which is fundamentally important not only for a robot, but also for an autonomous vehicle with an advanced driver-assistance system [5], such as object tracking [6] and anomaly detection [7].

Since the road images may differ with each other greatly, the detection task is actually not an easy task. Take Fig. 1 as an example. The roads have different pavements and are laid on different places, leading to various color, texture, and shape appearance. Along with these factors, the lighting condition is another influential one, inducing shadows on the road surface. These complexities in together make a reliable road detection difficult. Motivated by this fact, in this paper we propose a robust road detection method based on depth fusion and label transfer in a video sequence. The stable depth clue ensures the robustness of the proposed model and the ever-updating mechanism makes the transferred labels accurate (Fig. 2).

1.1. Overview of the proposed method

Though many works have been proposed in the past few years, most of them focus on the individual image. In fact, video sequences are the most frequently confronted situation instead of single images. Therefore, we lay our attention on the video sequence in this paper. The task is to infer the road area in each frame given a camera recorded street scene. The proposed method in this paper is named as context-aware label transfer (CALT), which is divided into the following three steps:

- *Preprocessing*: To facilitate the processing, the input video frames are firstly segmented by SLIC [8] to get superpixels. The subsequent label transfer is based on the obtained superpixels. Since we want to utilize the depth clue in the framework, the depth map of each frame is also reconstructed according to a consistent depth recovery technique [9].
- *Context-aware label transfer*: The preprocessed sequence is tackled frame by frame in this step. For the first frame, its ground truth labels (road or non-road) are manually marked. For the subsequent frames, their labels are sequentially

* Corresponding author.
E-mail addresses: crabwq@nwpu.edu.cn (Q. Wang), fangjianwu@opt.ac.cn (J. Fang), yuany@opt.ac.cn (Y. Yuan).

http://dx.doi.org/10.1016/j.neucom.2015.01.054
0925-2312/

transferred according to the previous one, taking the obtained result of current frame as the updated ground truth. This operation is iteratively conducted until all the frames are treated.

- *Result demonstration*: After the above procedure, each superpixel in a frame is allocated a label. To get a consistent labeling map without noisy labels, a Winner-Take-All (WTA) smoothing is applied to rule out the isolated inaccurate labels. Then a geometric triangle constraint is employed to restrict the road area. The final results are then demonstrated overlaid in the original sequence.

1.2. Contributions

Although there are existing road detection methods by employing depth map and label transfer, the proposed one in this paper is distinguished with them in the following aspects, which also makes the main contributions of this paper.

- Design a novel superpixel based context-aware descriptor by utilizing depth map. Most existing methods only consider the color information of the obtained image or video. Several works employ the depth map, but the way they incorporating it into the frameworks is simple and straightforward [10,11]. In this paper, we segment the image into superpixels and capture its characteristic by simultaneously concatenating the color and depth features. This combination is effective because it leverages the superiorities of the color's distinctiveness and depth's robustness. Based on this characteristic, a context-aware descriptor is developed to represent the superpixel's relationship with the adaptive circular neighborhood, which further paves the way for the optimization of label transfer.
- Conduct label transfer in an efficient nearest neighbor search. Label transfer can reduce the inference problem of training sophisticated parametric models for an unknown image to the problem of matching it to an existing set of annotated images [12]. But in this process, accurate registration is a challenging task [13]. For a more precise correspondence, we choose to transfer the labels between superpixels with the most similar contextual clues in adjacent images, instead of searching for the best match in a big training set. Since the two examined images are similar to each other, we utilize a new dense pixel correspondence method [14] to register the near superpixels respectively in the adjacent frames, which effectively exploits the video temporal relations between frames.
- Update the learned model adaptively. Traditional offline methods learn the model only once in the beginning. This strategy leads to an expensive training stage with large amount of images. More importantly, it means even if the actual scene changes much from the training ones, there is no adaptability. Based on this consideration, the proposed model updates the parameters of the classifier frame by frame, yet in an efficient manner. Novel properties of specific labels are dynamically updated, which ensures that the model can handle the changing scene.

Fig. 1. Road images with different colors, textures, shapes and lightings.

The rest of the paper is organized as follows. Section 2 reviews the related work. Section 3 introduces the recovery of depth map. Section 4 describes the main part of the proposed method – context-aware label transfer. Section 5 gives the experimental results to justify the effectiveness of the proposed method. Conclusions are finally made in Section 6.

2. Related work

The techniques for road detection can be categorized according to the types of road images, which are structured ones (e.g., a road in urban street) and unstructured ones (e.g., a road in rural area) [15].

For the structured road detection, the captured images have clear road markings and the designed algorithms are based on these extracted markings [16]. Among the earliest attempts, Bertozzi and Broggi [17] assume that the road markings are visible. Based on this assumption, the stereo image pair is first mapped to another 2D space to remove the perspective effect. Then the left image is used to recognize the road markings and both the two images are further employed to detect the free areas ahead the vehicle. Wang and Frémont [18] first use sky removal to enhance the axis-calibration stability. Then the stereo vision based extension is applied to extract the line function and reconstruct the ground plane. But the stereo images are vulnerable to weather conditions such as rain, snow, fog, and darkness. The radar sensor, on the contrary, being an active sensor and operating at millimeter wavelengths, can provide an alternate image of the scenario in front of the vehicle. For example, Ma et al. [19] propose a Bayesian model to interpret the radar and optical road images. In this procedure, they incorporate the lane and pavement prior to guide the boundary detection. Feng et al. [20] design a system equipped with a 2D laser radar. By measuring the distance from the radar to the road surface, a rectangle-searching algorithm is implemented to find the road rectangle containing the most road points. Besides the multi-sensor approaches, many other ones mainly focus on the feature representation, which can avoid the inconvenience of sensor setups and the undesired radioactive

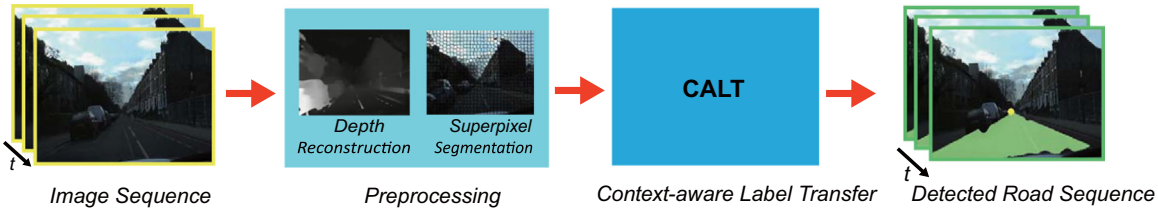

Fig. 2. General framework of the proposed method.

property. Han et al. [21] conduct road detection in the structured environment. They first extract line segments in polar coordinates using range data. Then the line segments are classified into road segments and obstacle segments by utilizing Markov chain propagation models and Bayesian update. Kühnl et al. [22,23] use SPatial RAY (SPRAY) feature and slow feature to learn the detection model that applies for both structured and unstructured roads. However, the main weakness of these structured road detection methods is that they highly rely on the prior structure. If the road markings are not clear in the image, the designed algorithms might not work well.

For the unstructured ones, no explicit road marking can be incorporated. In this case, different kinds of low-level image clues and multi-model data are generally employed [24,25]. For example, He et al. [26] propose to detect roads in campus and urban scenes. They first estimate the boundary in the intensity image using basic mathematical morphology operators – erosion and dilation. Then the color distribution of the road area is modeled as a multivariate Gaussian to extract it from the color image. Unsatisfyingly, the detected boundary is always disconnected. Franke et al. [27] achieve the country road detection by fusing color, texture and edge features. These features are modeled by a maximum-a posteriori estimation which is solved by particle filter. Apart from the color, texture, boundary/edge clue, the vanishing point and superpixel are alternative valuable information that can be used for inferring the road region. For instance, Rasmussen [28] deals with the problem of road following in ill-structured roads, in which the dominant texture orientations are computed with multi-scale Gabor wavelet filters. Then the obtained orientations vote for a consensus vanishing point. The subsequent tracking of vanishing point among frames is updated by a particle filter. Kong et al. [29] detect the straight part of road area with a set of weighted Gabor filters. Instead of using the hard voting strategy, a locally adaptive soft-voting is employed to determine the vanishing point. Later, they extent this work in the segmentation of road area by incorporating more features [30] and designing a novel generalized Laplacian of Gaussian (gLoG) filter [31]. As for the superpixel feature, Song et al. [32] detect roads using weighted aggregation based segmentation. Firstly, a road identifier is trained with supervised learning algorithm. Secondly, road regions are detected by combining a posteriori probability and the visual information using segmentation. There are still others trying to involve the multi-modality clues. Guo et al. [33] tackle the road detection problem in the stereo vision setup. They get the estimated camera parameters first, in order to infer the underlying road geometry. Subsequent detection is then formulated in the framework of MRF optimization, enforcing image evidence, geometry information and temporal constraint. Zhou et al. [34] use a monocular clue from the dark channel prior to get the estimated depth. Then six features are integrated to segment the input image into regional groups. By this means, the road region can be extracted. Different from the above works towards the road related clues, other literatures focus on the roads themselves. A few cases are listed as follows. Wang et al. [35] employ a hyperbola road model to cope with roads with varying curvatures. The geometrical and statistical reasoning in vanishing point leads to an accurately estimated parameter set. In the end, they integrate the hyperbola road model with a condensation particle filter to track the road in real time. Álvarez and López [36] focus on the situation that the road surface has different lighting conditions. They utilize a shadow-invariant feature, together with a likelyhood-based classifier to fulfill this task. Álvarez et al. [37] use a convolutional neural network based algorithm to learn features from noisy labels to recover the 3D scene layout of a road image. Its main contribution lies in utilizing the machine generated labels to learn the road detection model. This type of methods is comparatively more robust than the structured ones. However, in situations of particular lighting conditions, existing method is still far from perfect.

Apart from the road detection methods, the semantic scene segmentation is closely related to the proposed method. In the following, we will review the different techniques from this aspect. Hoitem et al. [38] assign the region labels through exploiting a combination of features, such as location, shape, color, texture, and perspective context. By learning boosted decision trees for each classifier, the class of each superpixel is determined. Kang et al. [39] integrate a color and near-infrared images to conduct scene segmentation, which utilizes hierarchical bag-of-textons features. But these two models are both learned offline and cannot handle the new confronted environment. Liu et al. [12] propose a label transfer framework called nonparametric scene parsing. They first construct a large database containing fully annotated images. Then, the system establishes dense correspondence between the input image and the nearest neighbors in the database using the dense SIFT flow algorithm. Finally, the existing annotations and segmentations are warped to the input image. However, this method needs large annotated database to supervise the image segmentation and extremely long time to train its model. Li et al. [40] propose a hierarchical generative model to segment the image into different levels of expression. Its success depends on the generative model that jointly expresses the visual and textual clues, and a fully automatic learning framework that is able to learn robust scene models from noisy web data. But the hierarchical level determination and the error accumulation between different image levels are the main problem for image segmentation. Zheng et al. [41] use the 3D cloud points of the examined scene to infer the semantic objects. They first form a 3D volumetric primitive by filling the missing voxels. Then a compact graph of the primitives is established for a more efficient representation. With this graph, a physical reasoning process is applied to merge the related nodes to get a stable segmentation. Similar to [12], the physical reasoning process relies on large external image dataset. Farabet et al. [42] use a convolutional network to learn the semantic features of different scales. By owing the complex interaction of various features to the deep learning process and combining a prerequisite segmentation, an input image can be well labeled with the trained model. Although this type of methods is widely studied in the vision community, most of them focus on the static images and the extreme large image datasets are needed to train their models. As for the video sequences, the considerations are absolutely different.

3. Depth recovery

Traditional depth map is obtained from stereo cameras or radar lasers. Nevertheless, the former needs to configure two cameras, which is a tedious work, and the latter has limited perception range. In this work, consistent depth maps are reconstructed from a video sequence [9] captured by a moving camera. Suppose the video sequence to be processed is $\mathcal{I}=\{I_t \mid t=1,\ldots,n\}$, where $I_t(x)$ represents the color of pixel x at frame t. Our objective is to calculate the corresponding depth map $\mathcal{D}=\{D_t \mid t=1,\ldots,n\}$. There are four main steps for this procedure.

(1) *Recovery of camera parameters*: The camera parameters are denoted as $\mathcal{C}=\{K_t,R_t,T_t\}$, where K_t is the intrinsic matrix, R_t is the rotation matrix, and T_t is the translation vector. They are recovered by the shape from motion (SFM) technique [43], which can handle long sequences with varying focal lengths. The estimated parameters are used for subsequent depth refinement.

(2) *Depth initialization*: In this step, an initial depth map is obtained for each frame independently. By minimizing an energy function containing a data term and a smoothness term with loop belief propagation [44], each pixel is assigned a depth label. The

energy function is defined as

$$E(D_t \mid \mathcal{I}) = \sum_x \left[1 - u(x) L_{init}(x, D_t(x)) + \sum_{y \in N(x)} \lambda(x,y) \cdot \rho(D_t(x), D_t(y)) \right], \tag{1}$$

where $u(\cdot)$ is the adaptive normalization factor, $\lambda(\cdot,\cdot)$ is the adaptive smoothness weight, $N(x)$ is the neighborhood of x, and $\rho(\cdot,\cdot)$ is the smoothness cost. L_{init} is the disparity likelyhood defined as

$$L_{init}(x, D_t(x)) = \frac{\sigma_c}{\sigma_c + \| I_t(x) - I'_t(x') \|}, \tag{2}$$

where σ_c controls the shape of the differentiable robust function and x' is the corresponding pixel (given a specified disparity) of x within frame I'.

(3) *Bundle optimization*: The depth map obtained from the above step is a rough estimation. Here each frame is associated with others to refine the result. For a pixel x in frame t, its corresponding pixel x' in frame t' is computed by the epipolar geometry as

$$x'^h \sim K_{t'} R_{t'}^T R_t K_t^{-1} x^h + D_t(x) K_{t'} R_{t'}^T (T_t - T_{t'}), \tag{3}$$

where $D_t(x)$ is the estimated disparity and h denotes the homogeneous coordinate system. According to this relationship, an energy function is derived and minimized to refine the initial depth map.

(4) *Space-time fusion*: Though bundle optimization can improve the accuracy of depth maps greatly, there are still reconstruction noises. To get a better reconstruction map, a space-time fusion algorithm is employed to reduce the disparity noises. It is actually an optimization operation. The main idea is that spatial continuity, temporal coherence, and sparse feature correspondence are simultaneously considered to constrain the depth map. By defining an energy function and optimizing it, the previously obtained results can be enhanced with fewer errors. More details can be found in [9].

4. Context-aware label transfer

In this part, the details of the proposed context-aware label transfer are introduced. Suppose the frame I_t have been processed, which means each superpixel s_i^t has been assigned a label $\ell_i^t \in \{0,1\}$ and the label map $L_t = \{\ell_i^t \mid i = 1, \ldots, N\}$ is treated as the ground truth of I_t. Here N is the number of superpixels in I_t and 1 and 0 respectively represent the road and non-road area. Our task is to estimate L_{t+1} of I_{t+1} according to the previous result. There are mainly three steps for this procedure as illustrated in Fig. 3. Firstly, a rough label map is predicted according to a weakly supervised SVM classifier. Then the previously obtained label map L_t is transferred to current result to refine the estimation, by employing the contextual information. In the end, a smoothing operation is applied to get a more consistent result and the detected road region is demonstrated on the original sequence.

(1) *Weakly supervised SVM prediction*: To decide the label of a superpixel is actually a classification problem. After the former frame I_t had been processed, the obtained result was employed to train a kernel SVM classifier [45]. Therefore, when I_{t+1} is being examined, the previously trained SVM classifier is applied to determine the current labelings of superpixels. The classifier is later updated with the newly obtained L_{t+1} when the cycle at time $t+1$ is finished. This updating strategy makes the model adaptable to the changing scenery.

As for the training of SVM classifier, there are several questions to be answered. The first one is how to select the samples. If all the pixels are involved in the training set, the efficiency cannot be ensured. Therefore, we select the collection of each superpixel's central pixel (red dots in the RGB image of Fig. 4) as the training data, which is representative enough and fast to be realtime. The second one is how to select the label and feature. For each central pixel, its label is set as the average of the labels within the superpixel, weighted by a Gaussian filter located at the center pixel. The feature vector is concatenated by two parts, one is from the color image and the other is from the depth map. For the color image, the HOG+LBP+Color is used, which has been proved to be effective in [46]. For the depth map, since there is no obvious gradient, only the LBP+Intensity is used.

There is still one important factor that should be pointed out. Except for the first frame, not all the sample features are recalculated for the updating procedure of SVM classifier when the label transfer at time $t+1$ is finished. We only put those samples that are not correctly classified into the previous training set to replace the same number of samples that are correctly classified with high confidence. As for the correctness, we employ the previous frame result as the ground truth, which means if the classification result at frame $t+1$ differs with the result at frame t, the related pixels are treated as the incorrect classified ones. Though this strategy is not perfect, it is proved to be effective in our experiment because the consecutive frames are supposed to have no much difference. This implies that many unchanged samples do not need to conduct label assignment and feature extraction again, which makes the update effective. Because of the partially updated samples and the sparsely sampled points, instead of the whole, the classifier is called weakly supervised.

(2) *Context-aware label transfer*: The obtained labeling result by the above SVM classifier is a rough estimation. To get a more precise prediction, the contextual clue is employed to optimize the label map. For this purpose, we first introduce the contextual descriptor of each superpixel as illustrated in Fig. 4.

Suppose the label of the examined superpixel is ℓ_i^{t+1} (either 0 or 1), its feature vector is sp_i^{t+1} and its central pixel is $C_i^{t+1} = (x_i^{t+1}, y_i^{t+1})$. Then we draw a series of equally sampled points (green points) around C_i^{t+1}, forming a circle with a radius r of the size of the superpixel's maximum edge. The sampling interval is denoted as $\Delta\theta$ and there are in total $M = 2\pi/\Delta\theta$ sample points. Suppose $\ell_{i,k}^{t+1}$ is the label of the $k^{th}(k \in \{1:M\})$ sample point. The contextual descriptor for the examined superpixel is then represented as $T_i^{t+1} = [\ell_{i,1}^{t+1}, \ell_{i,2}^{t+1}, \ldots, \ell_{i,M}^{t+1}]$. The defined descriptor is simple yet efficient and robust to noise, as described in the following. For a visual illustration, we draw a circular area around C_i^{t+1}, with different colors indicating different label percentages around the examined superpixel. This is illustrated in the bottom-right of Fig. 4.[1]

With the above contextual descriptor, we will check if the estimated label for each superpixel in the SVM prediction step is appropriate. This is achieved by a temporal MRF optimization procedure. For this purpose, we first find the corresponding counterpart at time t of the examined superpixel at time $t+1$ by registering the two adjacent frames with a recent dense pixel correspondence method [14]. After that, the obtained label map is further refined by minimizing an energy function. It is defined as

$$E(\mathcal{L}) = \sum_{i=1}^{N} D(\ell_i^{t+1}) + \sum_{i=1}^{N} w_{i,j} V(\ell_i^{t+1}, \ell_j^t), \tag{4}$$

[1] In order to demonstrate the contextual difference of different superpixels, we draw the largest percentage of road/no-road labels surrounding the examined superpixel with yellow/orange colors. And the drawn color mask is superimposed in the original RGB image. Because of the dark environment of the video scene, the RGB image shows almost a black color. Meanwhile, there are many superpixels lying on the boundary of different label regions. Therefore, the superimposed color mask may not take up the entire circle region centered with these superpixels. Hence, it occurs that the contextual descriptor has three colors (road, no-road and the color of the original RGB).

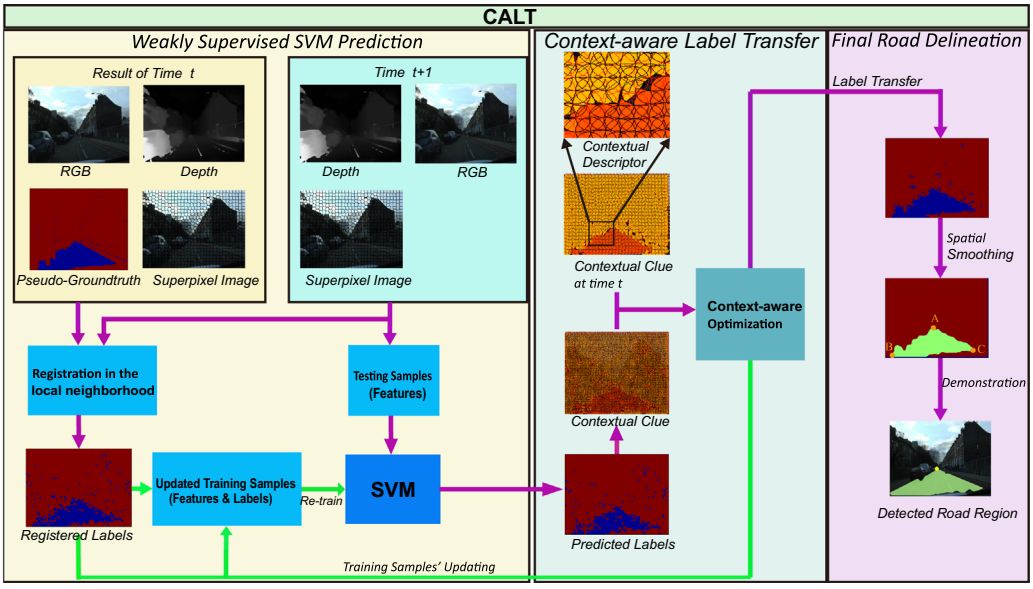

Fig. 3. Illustration of the proposed context-aware label transfer (CALT).

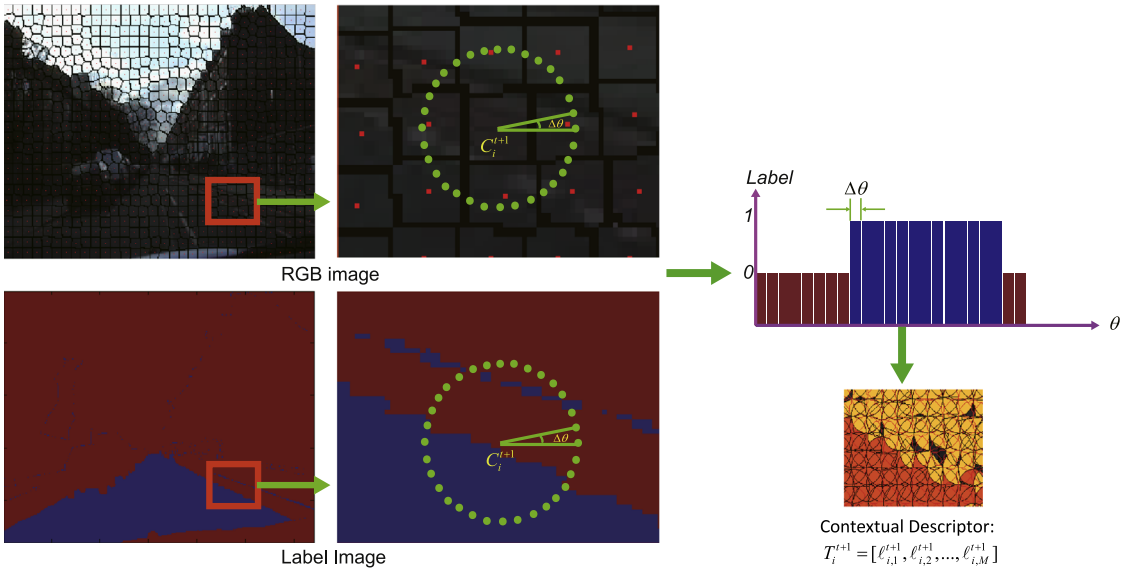

Fig. 4. Illustration of contextual descriptor. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

where j is the superpixel index at time t which is matched with the ith superpixel at time $(t+1)$, and $w_{i,j}$ is temporal varying weight, which balances the contextual consistency between the examined superpixel at time $(t+1)$ and its corresponding counterpart at time t. The larger $w_{i,j}$ is, the poorer the contextual consistency is.

The first term is the data term, reflecting the likelihood probability. It measures the consistency with the previously learned road/nonroad priors, assuming each class label with a Gaussian Mixture Model (GMM). In this paper, it is defined as

$$\sum_{i=1}^{N} D(\ell_i^{t+1}) = \sum_{i=1}^{N} \log\left(\frac{1}{p(\ell_i^{t+1} \mid sp_i^{t+1}, \mathbf{\Phi_t})}\right), \tag{5}$$

where $\mathbf{\Phi_t}$ is the parameter set of GMM estimated by the EM algorithm [47].

As for the smoothness term in this paper, it considers the temporal contextual consistency between the two adjacent frames. Since the two adjacent frames are nearly similar, our assumption is that those superpixels with its contextual descriptor different from the previous ones at the same locations might get its label transferred from the previous one. Whether this transfer is allowed or not depends on the value of the objective energy function. If the transfer can get a lower energy value, then the label change is encouraged; otherwise, it is rejected. To be specific, different from the traditional pair-wise label constraint [48,49] in spatial neighborhood, the smoothness term balances the temporal contextual consistency and is defined as

$$\sum_{i=1}^{N} w_{i,j} V(\ell_i^{t+1}, \ell_j^{t}) = \sum_{i=1}^{N} w_{i,j} \mid \ell_i^{t+1} - \ell_j^{t} \mid, \tag{6}$$

where $w_{i,j} = \| T_i^{t+1} \odot T_j^{t} \|_1 / M$, and $\odot$ represents the "bitxor" operation with an output of a vector whose elements are 0 or 1. When T_i^{t+1} and T_j^{t} are prone to be the same, which means the assigned label at frame $t+1$ is more consistent with previous label, $w_{i,j}$ tends to be smaller towards 0. This implies that the smoothness term encourages a smooth labeling between frames. Otherwise, if T_i^{t+1} and T_j^{t} are

prone to be the different, $w_{i,j}$ tends to be larger towards 1, which indicates that the smoothness term discourages the inconsistent labeling between frames.

The above description about the smoothness term implies if the labels of two corresponding superpixels are the same, the examined superpixel maintains its label unchanged. Otherwise, we need to consider the contextual consistency between the examined superpixel and its associated one matched by the dense correspondence method. Different from the appearance consistency constraints in video foreground segmentation [50] and target detection [51,52], the contextual consistency in this work concentrates on label domain. Because of the only two road/non-road regions, the temporal label constraint is more efficient and effective than the appearance based methods.

Since the road detection in this paper is a binary-labeling procedure, it is obvious that the smoothness term in the energy function is a metric in ℓ_i^{t+1}, ℓ_j^t. The optimization procedure can be conducted by the α-expansion algorithm in [44]. Note that the proposed method utilizes pairwise energy function in a temporal manner, i.e., the between-frame constraint. Although the temporal and spatial constraint can be jointly modeled in the pairwise energy function, we focus on the temporally contextual consistency of the MRF model in this work. At the same time, the spatial constraint is carried out by a post-processing Winner-Take-ALL (WTA) strategy in the following.

(3) *Final road delineation*: After the above steps, the label map is basically fixed, except for a small amount of noisy superpixels. To get a smoother map, an 8-neighborhood Winner-Take-All filter is applied, which encourages the pixel's label should comply with its surrounding labels and the obtained results should be smooth. We also note that after the perspective transformation, the road area always becomes a triangle, as shown in Fig. 5 by points A, B and C. Thus a prior geometric restriction is further considered. To be specific, we assume that the pixels lie outside this triangular region (the green areas in Fig. 5) do not belong to the road area. Though this constraint is a little coarser for the curved road, our focus is mainly on the large road area ahead of the vehicle. Therefore, this refinement can cover most of the road regions. Besides, our primary contribution in this work is in the previous part. After the above steps, the road detection results are generally satisfying. Thus we do not lay more emphasis on the step.

With this constraint, the obtained results can rule out the outliers and keep neat. In the end, the detected road region is laid on the original sequence as a display.

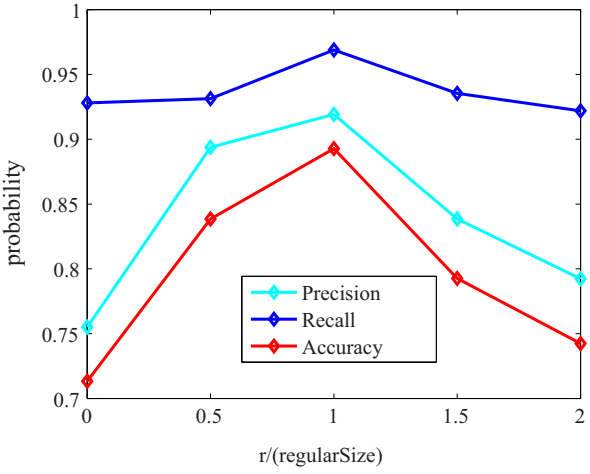

Fig. 5. Illustration of the final road delineation. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

5. Experiments

5.1. Dataset

In this work, we selected three video clips to test the proposed method. Among them, one video clip is from Cambridge-driving Labeled Video Database (CamVid) [53] ("Sequence1"), and the others ("sequence2 and Sequence3") are downloaded from web site database. CamVid Database is a popular collection of *urban* videos with object class semantic labels, together with metadata. All the videos are captured by a camera mounted inside a car and the frame size is 960×720. Instead of using bounding boxes or approximate boundaries, the database provides pixel-precision ground truth, allowing for more accurate learning and inference purpose. We select 250 frames from *seq01TP* because this clip is shot at dusk. Objects in the scene can still be identified, but the road and its surrounding street scene have similar brightness. This makes the recognition very difficult, in which case we can just see the robustness of the proposed method. Unfortunately, the ground truth labeling is provided every 30 frames, which we think is too few to get a precise evaluation. Therefore, in this work we re-label the sequence every 5 frames to produce more ground truths. In addition to the urban sequence [53], we also prove the efficiency of the proposed method in the sequences downloaded from the web site. These sequences are all captured in the *highway* circumstance, by which we can see the performance of our method when driving fast. The frame size of these highway videos is 500×280. In order to validate the performance of our method, the ground truths are provided by ourselves for every 5 frames.

5.2. Comparative methods

The success of the proposed method depends on the cooperation of several components. They are the SVM classifier and its updating strategy, the context-aware based MRF optimization and depth fusion. For evaluating the effects of different components, we select "Sequence1" to compare because of the most difficulty of this clip. Therefore, experiments with different component combinations are conducted and then evaluated according to the ground truth (GT). They are denoted as SD (SVM+Depth), SMD (SVM+MRF+Depth), SUM (SVM+Update+MRF), and SUMD (SVM+Update+MRF+Depth).

Besides, we also select three popular works representing the state-of-the-art to be compared with the proposed method on all the video sequences selected in this work. The competitors are a vanishing point based method (VP) [30], an illuminate invariance based method (ILL) [36], and a causal graph based video segmentation method (CA) [54]. The reasons for choosing the three methods are explained as follows. First, VP does not need any training samples. It can prove the superiority of using road sample training in our context-aware label transfer process. Second, ILL is designed to test the illuminate invariant ability. Its involvement can demonstrate the robustness of our method. At last, CA utilizes optical flow to correspond the labeled semantic regions, whose label is predicted by the state-of-the-art scene parsing method [42]. It can be used to test the effectiveness of our label transfer operation.

5.3. Evaluation metrics

For the evaluation of different methods, two sets of metrics are employed. The first one contains precision, recall and F-measure [55–57]. Precision reflects the rightness of the detected results, and recall indicates the ability to retrieve the desired information. F-measure is a balance between them. They are defined as

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}},$$

$$F\text{-measure} = \frac{\text{Precision} \times \text{Recall}}{(1-\alpha) \times \text{Precision} + \alpha \times \text{Recall}}, \quad (7)$$

where α is set as 0.5. The second one is accuracy [58]. It reflects the overlapping ratio of the detected results with the ground truth, which is formulated as

$$\text{Accuracy} = \frac{\text{area}(D \cap G)}{\text{area}(D \cup G)}, \quad (8)$$

where D is the detected results and G is the ground truth benchmark.

5.4. Parameter setup

There are several parameters to be set. The first one is for SLIC segmentation [8]. There are two parameters, *regularSize* and *regularizer*. *regularSize* controls the size of the superpixel and is set to 10. Smaller value of this parameter will lead to more training samples of SVM classifier and larger value will result in less accurate segmentation of semantic objects. *regularizer* is empirically set to 1 to ensure the superpixel has a smooth boundary.

The second one is the SVM classifier. *Gaussian* is selected as the kernel function and C is set to 10^5. The third one is the radius r for context-aware descriptor. We set the size from small to large and then calculate the performance under the three metrics. Fig. 6 shows the results on 10 training frames (each frame has about 1000 samples). From the curves, we can see clearly that when r is set equal to *regularSize*, the model performs best. Therefore, r is set to 10.

The last set of parameters is the σ_c, $\Delta\theta$, and M. In our experiments, they are all empirically determined as $\sigma_c = 10$ (according to [9]), $\Delta\theta = \pi/72$, and $M = 2\pi/\Delta\theta = 72$.

5.5. Results

In this part, the experimental results are demonstrated and analyzed. First of all, we present some typical example results in Fig. 7 to compare the effects of different components in the proposed method. From the figure, we can see that SD generates results with many unpleasant isolated superpixels. This is mainly due to the absence of optimization. With the MRF optimization added, SMD gets better results. This is true for SUM even though the

Fig. 6. The performance under different r values.

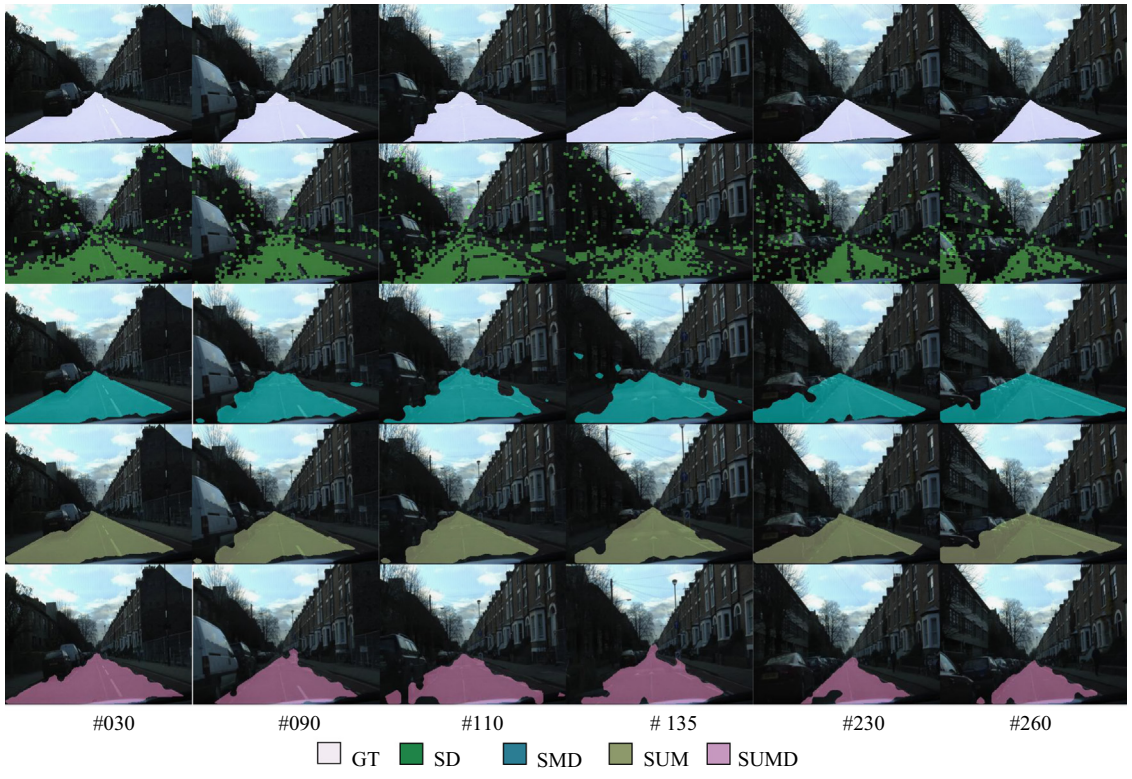

Fig. 7. Typical road detection examples of different component integrations.

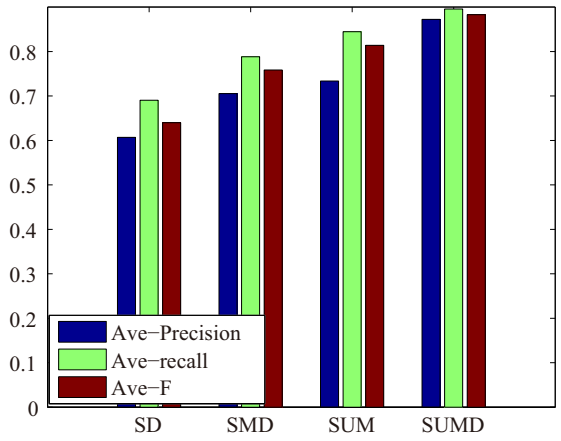

Fig. 8. Quantitative comparison of the averaged precision, recall and F-measure across all the frames. Different components of the proposed method are compared in this experiment.

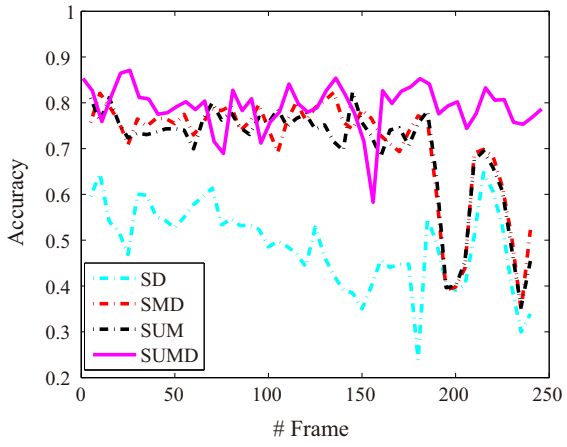

Fig. 9. Quantitative comparison of accuracy for each frame. Different components of the proposed method are compared in this experiment.

depth information is removed from the model. But SUMD is the most superior method in all comparisons.

Then we conduct intensive quantitative analysis to give a more objective evaluation. Fig. 8 presents the averaged precision, recall and F-measure across all the frames. It is manifest that the SVM classifier plus the depth (SD) performs worst. If we add the MRF optimization component without updating mechanism (SMD), the performance improves a bit. But if we replace the depth clue with MRF optimization together with the updating mechanism (SUM), the model demonstrates a lot more improvement. However, the best performance is achieved by combining all these factors in a unified model (SUMD). The reason is that compared with SMD, the updating procedure can infer the correctness of predicted road region, and compared with SUM, the depth map can provide a clearer feature description for road and background. Fig. 9 illustrates a similar result for the metric of accuracy, except that SMD and SUM are hard to say which one is better. But in all cases, SUMD, even with some small defect regions, is the most outstanding one, which means utilizing those components together is the appropriate choice. Note that the significant drop in the accuracy curve of SUMD is caused by the triangle region constraint of road delineation in Fig. 5. Because the upcoming car in the roadside from 150th frame to 170th frame in the CamVid sequence, the constrained road region by our method abandons some pixels belonging to the actual road region in ground truth.

We also compare our results with three typical methods representing the state-of-the-art on all the video sequences selected in this work. Fig. 10 shows the typical examples. It is obvious that our method outperforms the competitors because we have an updating scheme that can adapt to the changing scene. To be specific, the VP method without any training data, approximates the road region through a computed vanishing point. It has poor adaptation to the street roads because of the cars near the sidewalk. As for ILL, it aims to tackle the illuminative variance in the road surface (such shadows). However, the video sequence in this paper has low environmental illumination, which causes the features in the road and the background difficult to be distinguished. In terms of CA, it utilizes an efficient graph segmentation to cluster the semantic regions. Then the region labels are given by a state-of-the-art scene parsing method. However, CA only corresponds the regions in the adjacent frames by an optical flow

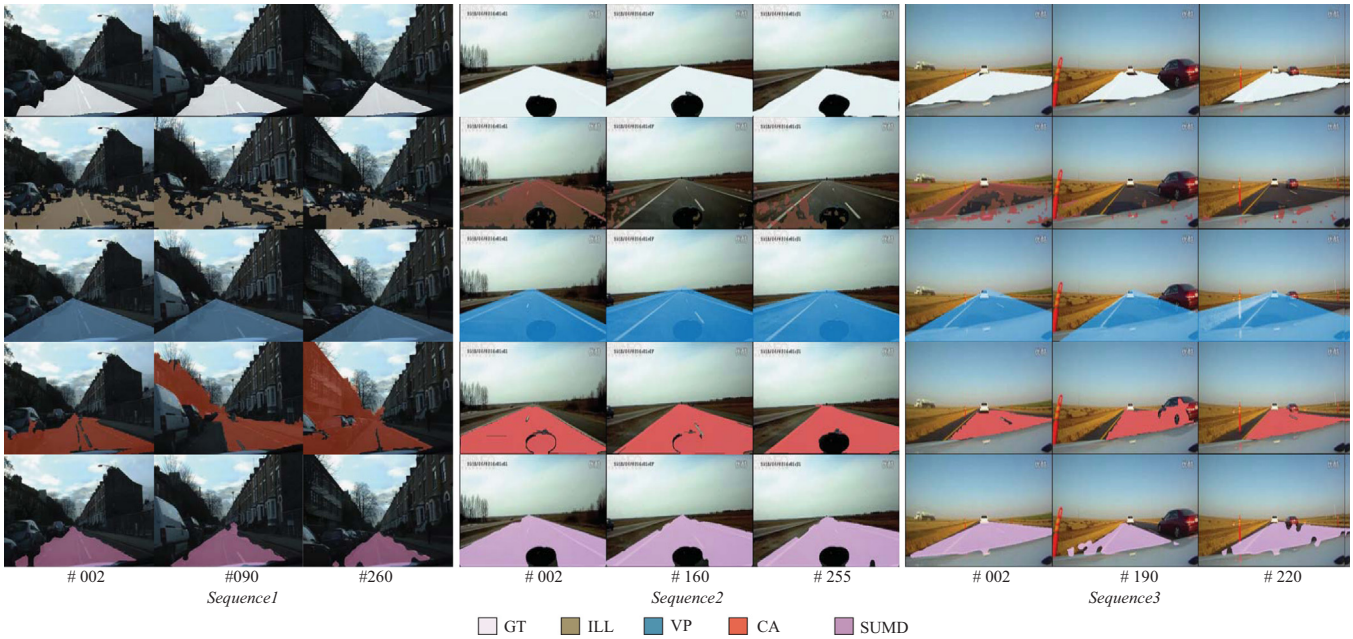

Fig. 10. Typical road detection examples for different methods.

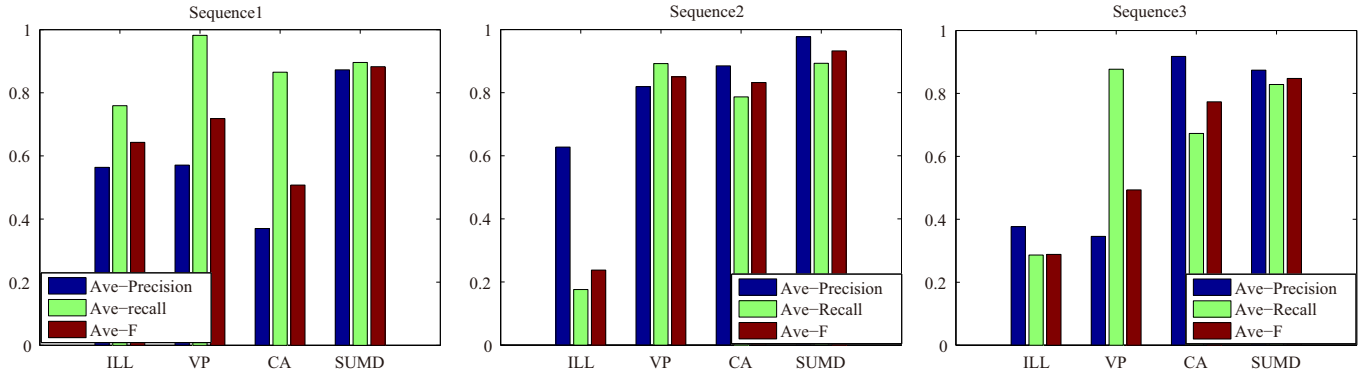

Fig. 11. Quantitative comparison of the averaged precision, recall and F-measure across all the frames. Different road detection methods are compared in this experiment.

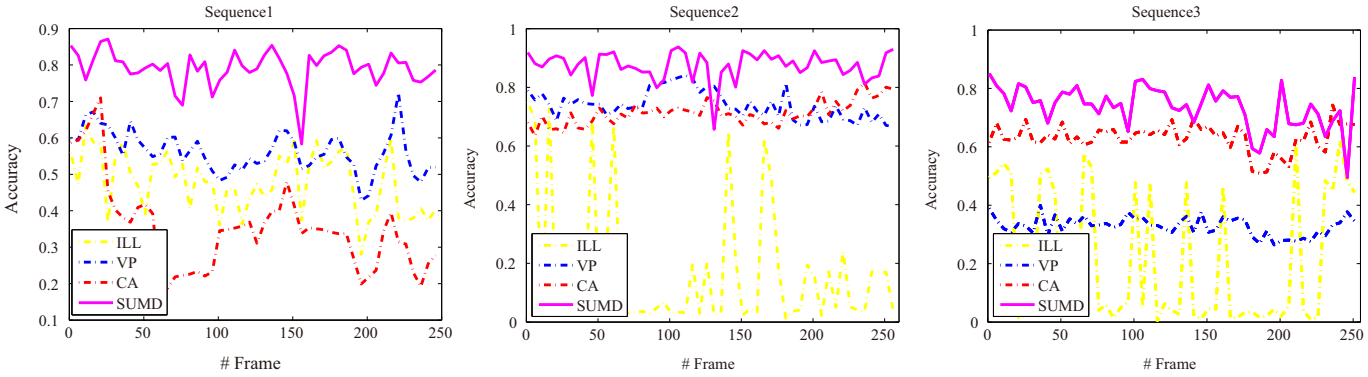

Fig. 12. Quantitative comparison of accuracy for each frame. Different road detection methods are compared in this experiment.

computation, and the correctness of the clustered regions is not ensured.

Fig. 11 shows the averaged precision, recall and F-measure for the four methods. We can see that on the precision and F-measure, our method is obviously better than the other three ones. Though the VP method has a high recall value, its corresponding precision is very low, which means it has an inferior performance. The same conclusion can be found in Fig. 12, where for all the frames, it is manifest that the proposed method has a higher accuracy. All these evidences prove that our method is more effective than the competitors.

5.6. *Efficiency analysis*

The proposed method and two competitive ones (VP and ILL) are run on MatLab 2010, while AC is run on C++. The employed computer is consistent, with Intel(R) Core(TM), i3-2130 @3.40 GHz and 4 GB RAM. For the proposed method, we get the statistics of each component. They are feature extraction (0.8 s), SVM prediction (1.02 s), label transfer (0.45 s) and postprocessing (6.06 s). For the vanishing point based method (VP), it takes 50 s to process one single frame, which is the slowest one. For the CA and ILL methods, they respectively take 50 ms and 1.25 s. These two methods are more efficient than the proposed one, but their performances are not satisfying. From this point, our method can be thought as a compromise, having the best performance and a moderate computational time.

6. Conclusion

In this work, we propose a road detection method based on context-aware label transfer. Based on the combination of depth clue with traditional RGB colors, the method utilizes label transfer to predict the road and nonroad area. The whole procedure can be divided into three steps. Firstly, the input images are segmented into superpixels and the depth maps are extracted. Then the road areas are estimated with a weakly supervised SVM classifier, and an MRF optimization process is followed to refine the obtained results, during which the context-aware information plays an important role expressing the neighborhood correlation. In the end, the optimized results are smoothed by Winner-Take-All and restricted by a geometric triangle. Experiments on a publicly available dataset demonstrate the effectiveness of the proposed method.

In the future, we plan to extend our work to the detection of multiple class objects. The ultimate goal is to understand the street scene automatically.

Acknowledgment

This work is supported by the State Key Program of National Natural Science of China (Grant no. 61232010), the National Natural Science Foundation of China (Grant nos. 61172143, 61379094 and 61105012) and the Fundamental Research Funds for the Central Universities (Grant no. 3102014JC02020G07).

References

[1] T. Lee, S.K. Park, M. Park, An effective method for detecting facial features and face in human-robot interaction, Inf. Sci. 176 (21) (2006) 3166–3189.
[2] J. He, H. Gu, Z. Wang, Multi-instance multi-label learning based on gaussian process with application to visual mobile robot navigation, Inf. Sci. 190 (2012) 162–177.
[3] S. Park, S. Kim, M. Park, S.K. Park, Vision-based global localization for mobile robots with hybrid maps of objects and spatial layouts, Inf. Sci. 179 (24) (2009) 4174–4198.
[4] J. Fritsch, T. Kuhnl, A. Geiger, A new performance measure and evaluation benchmark for road detection algorithms, in: 2013 16th International IEEE Conference on Intelligent Transportation Systems – (ITSC), 2013, pp. 1693–1700.
[5] S. Li, H. Yu, J. Zhang, K. Yang, R. Bin, Video-based traffic data collection system for multiple vehicle types, IET Intell. Transp. Syst. 8 (2) (2014) 164–174.

[6] J. Fang, Q. Wang, Y. Yuan, Part-based online tracking with geometry constraint and attention selection, IEEE Trans. Circuits Syst. Video Technol. 24 (5) (2014) 854–864.
[7] Y. Yuan, J. Fang, Q. Wang, Online anomaly detection in crowd scenes via structure analysis, IEEE Trans. Cybern. 45 (3) (2015) 562–575.
[8] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, S. Süsstrunk, Slic superpixels compared to state-of-the-art superpixel methods, IEEE Trans. Pattern Anal. Mach. Intell. 34 (11) (2012) 2274–2282.
[9] G. Zhang, J. Jia, T.T. Wong, H. Bao, Consistent depth maps recovery from a video sequence, IEEE Trans. Pattern Anal. Mach. Intell. 31 (6) (2009) 974–988.
[10] C. Oh, B. Kim, K. Sohn, Automatic illumination invariant road detection with stereo vision, in: IEEE Conference on Industrial Electronics and Applications, 2012, pp. 889–893.
[11] P. Lombardi, M. Zanin, S. Messelodi, Unified stereovision for ground, road, and obstacle detection, in: IEEE Intelligent Vehicles Symposium, 2005, pp. 783–788.
[12] C. Liu, J. Yuen, A. Torralba, Nonparametric scene parsing via label transfer, IEEE Trans. Pattern Anal. Mach. Intell. 33 (12) (2011) 2368–2382.
[13] C. Zhang, L. Wang, R. Yang, Semantic segmentation of urban scenes using dense depth maps, in: European Conference on Computer Vision, 2010, pp. 708–721.
[14] J. Kim, C. Liu, F. Sha, K. Grauman, Deformable spatial pyramid matching for fast dense correspondences, in: IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 2307–2314.
[15] C. Ünsalan, K.L. Boyer, Review on building and road detection, in: Multispectral Satellite Image Understanding, Advances in Computer Vision and Pattern Recognition, Springer London, 2011, pp. 139–144.
[16] G. Zhang, N. Zheng, C. Cui, G. Yang, Z. Yuan, An efficient road detection method in noisy urban environment, in: IEEE Intelligent Vehicles Symposium, 2009, pp. 556–561.
[17] M. Bertozzi, A. Broggi, Gold: a parallel real-time stereo vision system for generic obstacle and lane detection, IEEE Trans. Image Process. 7 (1) (1998) 62–81.
[18] B. Wang, V. Frémont, Fast road detection from color images, in: IEEE Intelligent Vehicles Symposium (IV), 2013, pp. 1209–1214.
[19] B. Ma, S. Lakshmanan III, A.O. Hero, Simultaneous detection of lane and pavement boundaries using model-based multisensor fusion, IEEE Trans. Intell. Transp. Syst. 1 (3) (2000) 135–147.
[20] M. Feng, P. Jia, X. Wang, H. Liu, J. Cao, Structural road detection for intelligent vehicle based on a 2d laser radar, in: International Conference on Intelligent Human-Machine Systems and Cybernetics, 2012, pp. 293–296.
[21] J. Han, D. Kim, M. Lee, M. Sunwoo, Enhanced road boundary and obstacle detection using a downward-looking lidar sensor, IEEE Trans. Veh. Technol. 61 (3) (2012) 971–985.
[22] T. Kuhnl, F. Kummert, J. Fritsch, Spatial ray features for real-time ego-lane extraction, in: Proc. IEEE Conference on Intelligent Transportation Systems, 2012, pp. 288–293.
[23] T. Kuhnl, F. Kummert, J. Fritsch, Monocular road segmentation using slow feature analysis, in: Proceedings of IEEE Conference Intelligent Vehicles Symposium, 2011, pp. 800–806.
[24] Y. Matsushita, J. Miura, On-line road boundary modeling with multiple sensory features, flexible road model, and particle filter, Robot. Auton. Syst. 59 (5) (2011) 274–284.
[25] D. Obradovic, Z. Konjovic, E. Pap, I.J. Rudas, Linear fuzzy space based road lane model and detection, Knowl. Based Syst. 38 (2013) 37–47.
[26] Y. He, H. Wang, B. Zhang, Color-based road detection in urban traffic scenes, IEEE Trans. Intell. Transp. Syst. 5 (4) (2004) 309–318.
[27] U. Franke, H. Loose, C. Knöppel, Lane recognition on country roads, in: Proceedings of IEEE Conference on Intelligent Vehicles Symposium, 2007, pp. 288–293.
[28] C. Rasmussen, Grouping dominant orientations for ill-structured road following, in: IEEE Conference on Computer Vision and Pattern Recognition, 2004, pp. 470–477.
[29] H. Kong, J.Y. Audibert, J. Ponce, Vanishing point detection for road detection, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 96–103.
[30] H. Kong, J.Y. Audibert, J. Ponce, General road detection from a single image, IEEE Trans. Image Process. 19 (8) (2010) 2211–2220.
[31] H. Kong, S.E. Sarma, F. Tang, Generalizing Laplacian of gaussian filters for vanishing-point detection, IEEE Trans. Intell. Transp. Syst. 14 (1) (2013) 408–418.
[32] T.T. Son, S. Mita, A. Takeuchi, Road detection using segmentation by weighted aggregation based on visual information and a posteriori probability of road regions, in: Proceedings of IEEE Conference on Systems, Man, and Cybernetics, 2008, pp. 3018–3025.
[33] C. Guo, S. Mita, D.A. McAllester, Robust road detection and tracking in challenging scenarios based on Markov random fields with unsupervised learning, IEEE Trans. Intell. Transp. Syst. 13 (3) (2012) 1338–1354.
[34] W. Zhou, L. Lin, B. Lou, X. Wei, Monocular depth cue fusion for image segmentation and grouping in outdoor navigation, in: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2010, pp. 3201–3206.
[35] Y. Wang, L. Bai, M.C. Fairhurst, Robust road modeling and tracking using condensation, IEEE Trans. Intell. Transp. Syst. 9 (4) (2008) 570–579.
[36] J.M. Álvarez, A.M. López, Road detection based on illuminant invariance, IEEE Trans. Intell. Transp. Syst. 12 (1) (2011) 184–193.
[37] J.M. Alvarez, T. Gevers, Y. LeCun, A.M. Lopez, Road scene segmentation from a single image, in: Proceedings of European Conference on Computer Vision, 2012, pp. 376–389.
[38] D. Hoiem, A.A. Efros, M. Hebert, Recovering surface layout from an image, Int. J. Comput. Vis. 75 (1) (2007) 151–172.
[39] Y. Kang, K. Yamaguchi, T. Naito, Y. Ninomiya, Multiband image segmentation and object recognition for understanding road scenes, IEEE Trans. Intell. Transp. Syst. 12 (4) (2011) 1423–1433.
[40] L.J. Li, R. Socher, L. Fei-Fei, Towards total scene understanding: classification, annotation and segmentation in an automatic framework, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 2036–2043.
[41] B. Zheng, Y. Zhao, J.C. Yu, K. Ikeuchi, S.C. Zhu, Beyond point clouds: scene understanding by reasoning geometry and physics, in: IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3127–3134.
[42] C. Farabet, C. Couprie, L. Najman, Y. LeCun, Learning hierarchical features for scene labeling, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2013) 1915–1929.
[43] G. Zhang, X. Qin, W. Hua, T.T. Wong, P.A. Heng, H. Bao, Robust metric reconstruction from challenging video sequences, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8.
[44] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, et al., A comparative study of energy minimization methods for Markov random fields with smoothness-based priors, IEEE Trans. Pattern Anal. Mach. Intell. 30 (6) (2008) 1068–1080.
[45] A.R. Canu, Estimation de la concentration en ozone par svm, in: Proc. 1re soumission Actes de Automatique et Environnement, 2001.
[46] X. Wang, T.X. Han, S. Yan, An hog-lbp human detector with partial occlusion handling, in: Proceedings of IEEE International Conference on Computer Vision, 2009, pp. 32–39.
[47] S. Calinon, F. Guenter, A. Billard, On learning, representing, and generalizing a task in a humanoid robot, IEEE Trans. Syst. Man Cybern. Part B: Cybern. 37 (2) (2007) 286–298.
[48] B. Fulkerson, A. Vedaldi, S. Soatto, Class segmentation and object localization with superpixel neighborhoods, in: Proceedings of IEEE Conference on Computer Vision, 2009, pp. 670–677.
[49] J. Xiao, L. Quan, Multiple view semantic segmentation for street view images, in: Proceedings of IEEE Conference on Computer Vision, 2009, pp. 686–693.
[50] J.W. Waggoner, Y. Zhou, J. Simmons, M.D. Graef, S. Wang, 3d materials image segmentation by 2d propagation: a graph-cut approach considering homomorphism, IEEE Trans. Image Process. 22 (12) (2013) 5282–5293.
[51] L. Zhang, L. Zhang, D. Tao, X. Huang, Sparse transfer manifold embedding for hyperspectral target detection, IEEE Trans. Geosci. Remote Sens. 52 (2) (2014) 1030–1043.
[52] Q. Zhang, L. Zhang, Y. Yang, Y. Tian, L. Weng, Local patch discriminative metric learning for hyperspectral image feature extraction, IEEE Geosci. Remote Sens. Lett. 11 (3) (2014) 612–616.
[53] G.J. Brostow, J. Fauqueur, R. Cipolla, Semantic object classes in video: a high-definition ground truth database, Pattern Recognit. Lett. 30 (2) (2009) 88–97.
[54] C. Couprie, C. Farabet, Y. LeCun, Causal graph-based video segmentation, CoRR, 2013, abs/1301.1671.
[55] Q. Wang, Y. Yuan, P. Yan, X. Li, Saliency detection by multiple-instance learning, IEEE Trans. Cybern. 43 (2) (2013) 660–672.
[56] Q. Wang, Y. Yuan, P. Yan, Visual saliency by selective contrast, IEEE Trans. Circuits Syst. Video Technol. 23 (7) (2013) 1150–1155.
[57] Q. Wang, G. Zhu, Y. Yuan, Multi-spectral dataset and its application in saliency detection, Comput. Vis. Image Underst. 117 (12) (2013) 1748–1754.
[58] M. Everingham, L. van Gool, C.K.I. Williams, J. Winn, A. Zisserman, The pascal visual object classes (voc) challenge, Int. J. Comput. Vis. 88 (2) (2010) 303–338.

Qi Wang received the B.E. degree in automation and Ph.D. degree in pattern recognition and intelligent system from the University of Science and Technology of China, Hefei, China, in 2005 and 2010 respectively. He is currently an associate professor with the Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision and pattern recognition.

Jianwu Fang received the B.E. degree in automation and M.E. degree in Traffic Information Engineering and Control from the Chang'an University, Xi'an, China, in 2009 and 2012 respectively. He is currently a candidate Ph.D. with the Center for Optical Imagery Analysis and Learning, State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, China. His research interests include computer vision and pattern recognition.

Yuan Yuan is currently a Full Professor with the Chinese Academy of Sciences, Beijing, China. She has authored or coauthored over 150 papers, including about 100 in reputable journals such as IEEE Transactions and Pattern Recognition, as well as conference papers in CVPR, BMVC, ICIP, and ICASSP. Her current research interests include visual information processing and image/video content analysis.