Accepted Publication to the 2018 21st International Conference on Intelligent Transportation Systems (ITSC)
November 4-7, 2018, Maui, Hawaii, USA
DOI: 10.1109/ITSC.2018.8570004

Multi-Session Visual Roadway Mapping

Stefan Boschenriedter*, Phillip Hossbach*, Clemens Linnhoff*, Stefan Luthardt*, Siqian Wu*

***Abstract*— This paper proposes an algorithm for camera based roadway mapping in urban areas. With a convolutional neural network the roadway is detected in images taken by a camera mounted in the vehicle. The detected roadway masks from all images of one driving session are combined according to their corresponding GPS position to create a probabilistic grid map of the roadway. Finally, maps from several driving sessions are merged by a feature matching algorithm to compensate for errors in the roadway detection and localization inaccuracies. Hence, this approach utilizes solely low-cost sensors common in usual production vehicles and can generate highly detailed roadway maps from crowd-sourced data.**

I. Introduction

The development of advanced driver assistance systems and autonomous vehicles has drawn increasing attention in recent years. A central aspect of these systems is the precise perception of the surrounding environment including other traffic participants, infrastructure and the roadway. Besides the processing of sensor data, it is also necessary for these intelligent systems to employ highly detailed maps of the environment [1], especially maps specifying the roadway course. The creation of these maps with conventional surveying methods is very time consuming and costly [2].

A. Existing Roadway Mapping Approaches

One approach to generate roadway maps with highly detailed information is using aerial image data. Mattyus et al. [3] enhance open street map data with further information like the exact width and the centerline of the streets by utilizing aerial footage from multiple datasets. Pink et al. [4] employ aerial images to detect and map the exact position of lane markings on the roads. Although these algorithms yield useful additional map information, up-to-date aerial footage is hardly publicly available and important road parts might be occluded by surrounding trees or buildings.

Due to these issues, there are many ground-vehicle-based approaches to create roadway maps. Roh et al. [5] and Ishikawa et al. [6] present methods which employ a sophisticated sensor setup for the generation of digital maps. These systems are based on cameras and lidar sensors, which are mounted on the mapping vehicle, while the localization is done with a high precision GPS system, odometry and multiple inertial measurement units. Ishikawa et al. [6] make use of a multi-sensor fusion to determine the position of lane markings and traffic signs whereas Roh et al. [5] take advantage of a Simultaneous Localization and Mapping (SLAM) approach to create lane and 3D maps in sub-meter accuracy.

*Control Methods and Robotics, TU Darmstadt, Germany.
All authors contributed equally to this work.

(a) Google Maps satellite image with manually created roadway label. (aerial imagery © 2018, Google)

(b) Roadway grid map created with our approach.

Fig. 1. Comparison of a manually labeled roadway map and the result of our approach which uses solely camera images and GPS pose measurements. The map is a combination of data from 12 recording sessions.

Nevertheless, these methods rely on many highly specialized and expensive sensors. Therefore, further approaches try to employ a much reduced sensor setup. In [7] the road mapping is performed by only utilizing a stereo camera, an inertial navigation system (INS) and a consumer-grade GPS. The sensor data of the INS and GPS are combined to reduce their inaccuracies, while the stereo camera is employed to determine the 3D position of extracted landmarks.

When generating maps of complex road segments, it is necessary to cover all the relevant street parts with multiple recent measurement sessions. For the systems mentioned above this leads to a high demand for labor and a large effort to provide specialized measurement vehicles. Therefore, [8] and [9] present frameworks for precise road mapping with low-cost sensors, which are used in usual production vehicles. In order to improve the mapping results based on imprecise sensor data, they combine the measurements from multiple sessions of a certain road segment. However, Schreiber's graph-based SLAM approach [8] relies on lane markings in the middle of the roadway, because they are employed for the map representation. The approach of [9] works only on highways since they are using a Lane Marking Based Visual Odometry (LMBVO) to increase the accuracy, which is developed for highly standardized roads.

B. Our Roadway Mapping Approach

In this paper we propose an algorithm for an automatic, visual roadway mapping of urban areas with no restrictions to the properties of the street. The approach solely uses a stereo

camera and a GPS-based localization module, which are part of the standard equipment of commercially available cars today. Since there is no additional hardware needed nor any trained staff to use our system, the whole mapping procedure can be realized via crowd sourcing. All the measurement data from drives of regular customers can be recorded and combined to generate a large scale, high resolution and up-to-date map of the existing road system.

The dataset used for the evaluation within this paper was collected by an experimental vehicle in public road traffic. It consists of multiple recording sessions of the same route, driven in both directions, on different days with various lighting and weather conditions. A session is defined as a continuous drive of the whole course in one direction. The collected data comprises images from a stereo camera and its corresponding disparity maps. In our experiments a front view stereo camera is used, but in general there are no constraints to the type or position of the camera. A mono camera would also be sufficient with only minor limitations. Furthermore, for each image there is an associated *global pose*. This global pose is obtained from GPS measurements fused with the wheel odometry of the car.

The proposed algorithm performs the following steps. First, a convolutional neural network detects the roadway in the camera image. Subsequently, the generated roadway mask is transformed into bird's-eye view by an inverse perspective mapping. The roadway masks from all images of one single driving session are fused together to one probabilistic grid map of the roadway. Due to inaccuracies of the localization, the maps of several sessions are locally shifted compared to their actual positions. To compensate for these errors, we combine maps from various sessions by matching prominent roadway features and warping the maps locally. Additionally, we propose a method to combine maps which were created from driving sessions with different driving directions and therefore cover different spatial areas. Figure 1 depicts the resulting roadway map from our algorithm in comparison to a manually labeled ground truth.

C. Paper Overview

The remainder of this paper is organized as follows: Section II explains the necessary preprocessing steps to generate bird's-eye view roadway masks from the camera images. The main contribution of this work is illustrated in Section III, where the creation of a single map for one session and the following combination of multiple sessions is depicted. Afterwards, the proposed algorithm is evaluated in Section IV. Finally, Section V concludes the results and gives an overview of possible future work.

II. Preprocessing: Generating a Roadway Mask

In the preprocessing a bird's-eye view roadway mask is extracted from each image provided by the vehicle's camera. A convolutional neural network is used for the roadway detection. The result of this detection is improved further with a plausibility check based on depth information from the disparity map. Afterwards, the roadway mask is transformed into bird's-eye view, which simplifies the map creation. Figure 2 illustrates the whole preprocessing pipeline.

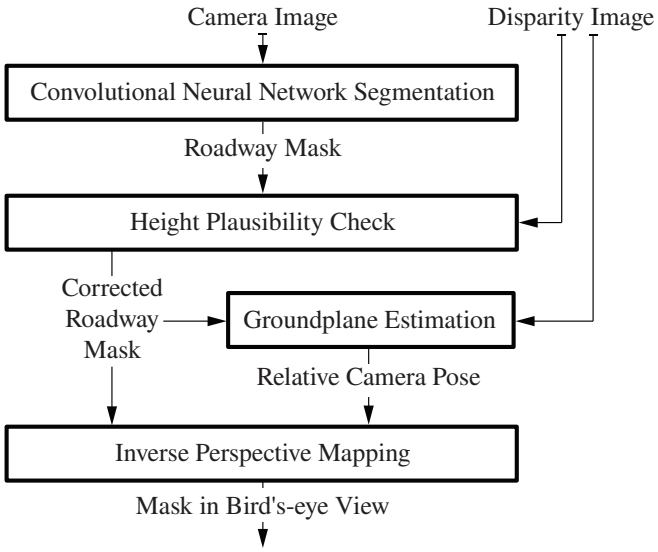


Fig. 2. Overview of the preprocessing procedure to generate a bird's-eye view roadway mask for a single image frame.

Fig. 3. Example of the roadway detection with a convolutional neural network. We used a pretrained segmentation network and retrained it with a small set of labeled images from our camera.

A. Roadway Detection

Today there are various approaches for roadway detection in images. However, in recent years it became apparent that deep convolutional neural networks can achieve an outstanding performance in image classification and segmentation [10]–[12]. Teichman et al. [13] developed an algorithm especially designed for roadway detection in RGB images, which won the first place in the "Kitti Road Detection Benchmark" in 2016 [14]. This neural network consists of an encoder and a decoder network. The encoder network is a multi purpose object detection network, which uses the pretrained VGG [15] model weights. The decoder part upsamples the output of the encoder to generate an output roadway mask with the size of the input image. We retrained this pretrained network with a set of 60 labeled images from our camera utilizing training data augmentation. With this retraining dataset we already obtained an accuracy of $97.3\,\%$ on our apart test dataset. In this case the small number of training samples is sufficient since the encoder is already pretrained and the decoder is initialized as bilinear upsampling. An example of the resulting roadway detection is illustrated in Fig. 3.

Despite the already satisfying performance of the roadway detection algorithm, the neural networks might still produce partially false roadway masks. Thus we implement a subsequent plausibility check. Since a stereo camera is used in our

setup, a disparity map is available for each input image pair [16]. The disparity map is employed to determine the 3D position of each image point [17]. This information is used to check if the height of every detected roadway point lies within a certain range around the ground level. If a point's position is too high, the point is presumably a false prediction and is removed from the roadway mask.

B. Inverse Perspective Mapping

After obtaining a validated roadway mask for each image, the mask is transformed from the camera's perspective into bird's-eye view. Therefore, an inverse perspective mapping from image coordinates to vehicle coordinates is employed, where the vehicle's coordinate system has its XY-plane on the roadway surface. This is done by assuming a pinhole camera model [17] and leaving out the height component Z_V since all road points are assumed to lie on the roadway plane, i. e. $Z_V = 0$. The resulting homography $\mathbf{H}$ yields a direct mapping from image points (x, y) to points (X_V, Y_V) on the roadway plane and is given by the equation

$$\lambda \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \underbrace{\begin{bmatrix} f & 0 & o_x \\ 0 & f & o_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & t_\mathrm{x} \\ r_{21} & r_{22} & t_\mathrm{y} \\ r_{31} & r_{32} & t_\mathrm{z} \end{bmatrix}}_{\mathbf{H} \in \mathbb{R}^{3\times 3}} \begin{bmatrix} X_V \\ Y_V \\ 1 \end{bmatrix}$$

$$\Leftrightarrow \begin{bmatrix} X_V \\ Y_V \\ 1 \end{bmatrix} = \lambda \, \mathbf{H}^{-1} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}. \quad (1)$$

This homography contains the intrinsic camera parameters f, o_x and o_y, as well as the extrinsic camera parameters r_{ij} and t_i. To improve the perspective mapping, we also incorporate camera movements in relation to the roadway plane. Accelerations of the vehicle and uneven roadway surfaces make the car pitch and roll which would distort the inverse perspective mapping. In order to compensate for these rotational changes, we again use the 3D points computed from the disparity map. We first select all 3D points which correspond to image points lying inside the roadway mask. From this roadway 3D point cloud we estimate the parameters of the roadway plane by using an iterative weighted linear least squares algorithm [18]. The relative pitch and roll angle of this plane are then used in the inverse perspective mapping. After these preprocessing steps the resulting bird's-eye view roadway masks are utilized in the actual roadway mapping procedure which is described in the following section.

III. Roadway Mapping

In this section, we first give an overview of the roadway mapping procedure followed by a detailed description, how a map is generated for a single driving session and how these maps are later combined into one final grid map.

A. Mapping Procedure Overview

Our proposed mapping procedure consists of three steps. In the first step, a map is generated for every session by accumulating the computed bird's-eye view roadway masks into a grid map. To determine the position and orientation of each mask within the map, we use the *global pose* which combines GPS measurements and wheel odometry. This single-session map generation is the first step in our procedure, as depicted in Fig. 4. Since the view of the roadway is sometimes partially obstructed by obstacles, like parked vehicles, and the global pose is imprecise, the map created by one session is insufficient for a precise and complete mapping of the roadway. Hence, several maps are combined to overcome these issues in the second step. Since the localization is locally varying, the single session maps cannot be combined by simply overlaying and averaging them. They first have to be locally warped to compensate for localization errors. Therefore, prominent feature points are detected and matched between the maps considering the variance of the global pose. Based on these feature point matches, the maps are warped and overlaid to form combined maps. The maps are combined separately for each direction, since parts of the roadway are only visible when driving in a certain direction. A direct combination of maps of different driving directions would therefore lead to unwanted results. Thus, in the last step of our mapping procedure, the two combined maps, one for each direction, are merged into one final roadway map with a special combination method.

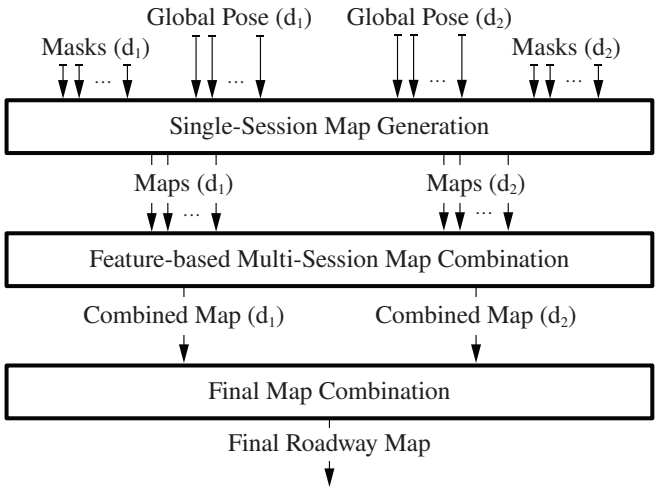


Fig. 4. Overview of the roadway mapping procedure consisting of single-session map generation, multi-session map combination and the final map combination of the driving directions d_1 and d_2.

B. Single-Session Map Generation

To generate a single-session grid map, each bird's-eye view roadway mask from the preprocessing is transformed into the map coordinate system using the corresponding global pose. The roadway map is a grid consisting of $20\,\mathrm{cm} \times 20\,\mathrm{cm}$ cells[1]. Each pixel of each roadway mask i is assigned to one cell $\mathbf{c}$ in this grid map according to its transformed 3D position. Before combining the individual

[1]We have chosen a $20\,\mathrm{cm}$-grid as a suitable trade-off between resolution and computational cost. However, higher map resolutions up to one centimeter are possible. The achievable map resolution depends on the image resolution, the used roadway detection distance and the positioning accuracy.

roadway pixels, they are weighted with the value

$$\xi_{\mathbf{c},i} = \frac{1}{\sqrt{X_{\mathbf{c},i}^2 + Y_{\mathbf{c},i}^2}}. \tag{2}$$

This weight contains the euclidean distance between the vehicle and the 3D point $[\, X_{\mathbf{c},i}\ Y_{\mathbf{c},i}\ 0\,]^{\mathrm{T}}$. We weight the roadway pixels with the inverse of this distance for two reasons. Firstly, the accuracy of the roadway detection decreases over distance, i. e. it becomes less reliable for more distant pixels. Secondly, the resolution of the 3D points is also declining with increasing distance to the camera. Using this weighting we compute for each map cell $\mathbf{c}$ the probability $p(m_{\mathbf{c}}|I)$ of being part of the roadway given all images I of one session. The combination formula is

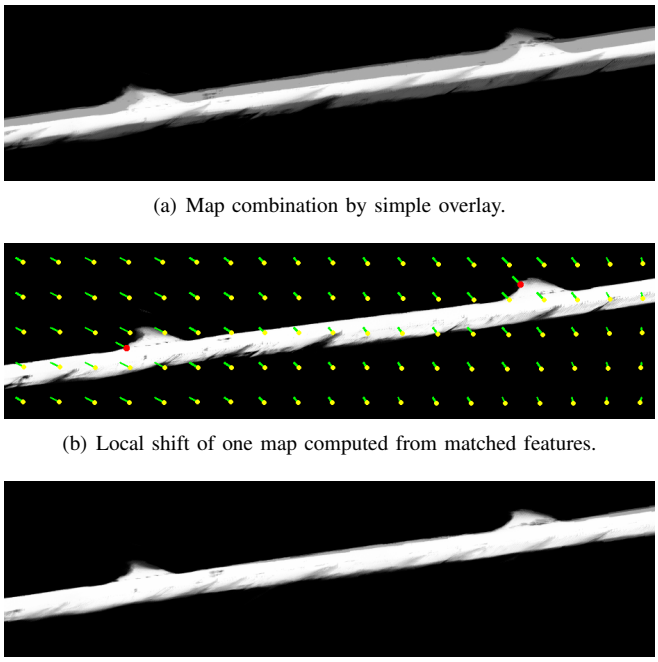
$$p(m_{\mathbf{c}}|I) = \frac{\sum_{i \in \tilde{I}} \xi_{\mathbf{c},i}}{\sum_{i \in \hat{I}} \xi_{\mathbf{c},i}}, \tag{3}$$

where the subset $\hat{I} \subseteq I$ consists of all images which contain a 3D point corresponding to cell $\mathbf{c}$. The images where the 3D point is also classified as part of the roadway, constitute the subset $\tilde{I} \subseteq \hat{I} \subseteq I$. Due to this combination of partly redundant roadway masks, errors in the roadway detection of individual images as well as the influence of moving objects on the roadway is reduced.

C. Feature-based Multi-Session Map Combination

In some sessions, there may be obstacles on the street, i. e. parked cars or construction sites, which are blocking the sight of parts of the roadway and thus prevent these roadway parts from being mapped. Additionally, the generated maps contain errors due to the localization inaccuracies of the global pose. In order to compensate for these deficiencies, maps from multiple sessions are combined by warping them on the basis of extracted feature points.

The Harris corner detector [19] is used to find prominent features in every roadway grid map. The first map is taken as reference and for every detected feature of the first map a small patch of $55\,\mathrm{px} \times 55\,\mathrm{px}$, centered on the feature point, is extracted. This corresponds to an area of approximately $11\,\mathrm{m} \times 11\,\mathrm{m}$ in metric size, which roughly covers a complete road intersection and this patch size has proven to yield the best results in our approach. The patch from the first map is compared via cross-correlation [20] to slightly larger patches of $65\,\mathrm{px} \times 65\,\mathrm{px}$ centered on the feature points in the other maps. Only pairs of features with a distance less than the assumed inaccuracy of the global pose are compared in this correlation analysis. By sliding the smaller patch from the first map over the larger patch in the other maps, a matrix of cross-correlation coefficients is created for each pair of features. If the maximal cross-correlation coefficient in this matrix is higher than a specified threshold, the features are considered to be a match. The position of the highest maximal coefficient is used to refine the shift estimation of the feature's position. For the warping we only consider features that have been successfully detected and matched in most of the roadway maps. Insignificant features which are contained in only a few maps are discarded. Therefore, each map k contains an individual number of features N_k.



(a) Map combination by simple overlay.

(b) Local shift of one map computed from matched features.

(c) Combined map by feature matching and local warping.

Fig. 5. Comparison of simple map overlay and the map combination by feature matching and warping. Subfigure (a) shows the combination by simple overlay. In subfigure (b) the local shifts (green lines) for some exemplary cells (yellow dots) are shown. These shifts are computed from the shifts of the feature points (red dots) by interpolation. Subfigure (c) depicts the resulting combined map using the warping displayed in (b).

To average out the localization error, the computed feature points are shifted to the center of all corresponding features. The cluster center ϵ_n of the features $\mathbf{f}_{n,k}$ from all K_n maps containing the corresponding feature is calculated by

$$\epsilon_n = \frac{1}{K_n} \sum_{k=1}^{K_n} \mathbf{f}_{n,k}. \tag{4}$$

Consequently, the shift $\mathbf{d}_{n,k}$ for a feature point in map k is $\mathbf{d}_{n,k} = \epsilon_n - \mathbf{f}_{n,k}$. These feature shifts are used to determine the shift for all map cells to align the individual maps with the combined map. For map k the shift $\mathbf{d}_{\mathbf{c},k}$ of cell $\mathbf{c}$ is computed from all feature shifts $\mathbf{d}_{n,k}$ by

$$\mathbf{d}_{\mathbf{c},k} = \frac{1}{\sum_{n=1}^{N_k} \zeta_{\mathbf{c},n}} \sum_{n=1}^{N_k} \zeta_{\mathbf{c},n}\, \mathbf{d}_{n,k}, \tag{5}$$

$$\text{with weights} \quad \zeta_{\mathbf{c},n} = e^{-\frac{||\mathbf{c}-\epsilon_n||^2}{2\sigma^2}}. \tag{6}$$

The Gaussian weighting function $\zeta_{\mathbf{c},n}$ ensures that only the shifts of close feature points have an impact on the calculation of (5). The σ is chosen such that it covers an area, in which the localization errors presumably correlate, e. g. $30\,\mathrm{m}$. With the computed shifts for each map cell the roadway probability $p(\overline{m}_{\mathbf{c}})$ in the combined map yields

$$p(\overline{m}_{\mathbf{c}}) = \frac{1}{M} \sum_{k=1}^{M} p(m^k_{(\mathbf{c}+\mathbf{d}_{\mathbf{c},k})}). \tag{7}$$

To illustrate this procedure, Fig. 5 shows a map section with a straight road and two junctions. In Fig. 5(a) three maps are simply overlaid by averaging the cell values without warping. It is clearly recognizable that the roadway grids are not coextensive, due to the localization inaccuracies. Hence, each map is warped as depicted in Fig. 5(b). At the exemplary yellow dots, the local shifts are visualized by green lines. These cell shifts are based on the shifts of the red marked feature points. Figure 5(c) depicts the resulting combined map, generated by feature matching and local warping. The roadways from the individual maps, which are merged together by this procedure, are now coextensive on their average position. The localization error of the single-session maps is thereby averaged out.

D. Final Map Combination

The resulting combined roadway maps from the previous step are still separated by driving direction. In the last step these two maps are joined together. Since some areas of the roadway can only be seen when driving in one certain direction, this combination should not be done by simple averaging. Important details would be lost if such a naive combination would be used. Therefore, we developed a combination formula where a high cell confidence in either of both maps results in a high confidence in the final map, even though the confidence in the other map might be very low. On the other hand, if the confidences are low in the maps of both directions, the confidence in the final map should be even lower. Following these considerations, we compute the confidence for the cells in the final roadway map by

$$p(\hat{m}_{\mathbf{c}}) = \kappa_{max}^2 + (1 - \kappa_{max}) \cdot \kappa_{min}, \tag{8}$$

where κ_{max} and κ_{min} are

$$\kappa_{max} = \max\big(p(\overline{m}_{\mathbf{c}}^{d_1}), p(\overline{m}_{\mathbf{c}}^{d_2})\big) \text{ and}$$
$$\kappa_{min} = \min\big(p(\overline{m}_{\mathbf{c}}^{d_1}), p(\overline{m}_{\mathbf{c}}^{d_2})\big).$$

While the first term of (8) is dominant if there is a high roadway confidence in one map, the second term becomes dominant if there is only a low or moderate confidence for both directions. The resulting final roadway map, created by the complete procedure described above is depicted in Fig. 1(b) and is evaluated in detail in the next section.

IV. Evaluation and Results

In the following, the results of our approach are discussed and evaluated utilizing labeled aerial images as ground truth. First, the quality of the generated roadway map is surveyed by projection into the original camera images. Subsequently, we demonstrate the advantage of our method over a naive averaging approach. Finally, we compare our final result to different intermediate results and to standard roadway map data by inspecting precision, recall and F_1-score.

A. Roadway Map Projection into the Camera Image

For a first assessment of the quality of our final roadway grid map, we projected the roadway map into the original camera images. The final roadway map is generated from a set of driving sessions and thus can not directly be projected into the images of one specific session due to the global pose errors of that specific session. This issue is addressed by using the calculated shift data to re-warp the combined map to fit the specific session containing the image. In Fig. 6 we show the warped combined roadway map visualized in the image as a green lattice. In the image, the roadway map nicely covers the actual road and is neither shifted nor skewed. Currently unrecognizable parts of the roadway, like junctions and far away road sections, are already indicated by the projected roadway map. In addition, due to the combination of several driving sessions, even areas which are currently obscured by other vehicles are labeled as roadway.

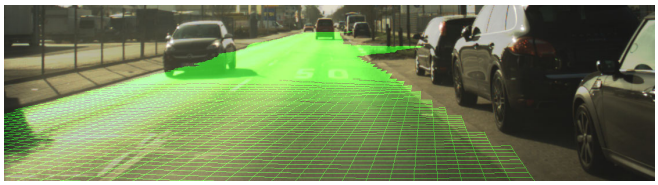


Fig. 6. Projection of our roadway grid map into the camera image. The map nicely covers the visible roadway and furthermore indicates roadway in currently occluded or distant areas.

B. Used Reference Data

For a quantitative evaluation of our results we manually labeled the roadway of our course in Google Maps aerial imagery with a pixel size of $9\,\mathrm{cm} \times 9\,\mathrm{cm}$. Figure 7(a) shows an example section of the created ground truth map. The ground truth is downscaled to fit the resolution of our roadway map. To compare our roadway map, which contains continuous confidence values $p(\hat{m}_{\mathbf{c}})$, to this binary ground truth, we apply a threshold. In the following evaluation we will also discuss the influence of the threshold value.

C. Comparison to Naive Averaging

First, we compare our proposed local map warping method, shown in Fig. 7(d), to a naive averaged overlay approach (Fig. 7(b)). For both approaches their specific optimal threshold value was determined, i. e. the threshold yielding the highest F_1-score compared to the ground truth. Clearly, the thresholded result of the proposed method, shown in Fig. 7(e), outperforms the simple averaging approach in Fig. 7(c), where entire road parts are lost.

We further investigated the effect of the threshold value: When varying that value for the naive averaging approach, it is possible to achieve a similarly high F_1-score for the complete course as with our approach. However, the dependency on the threshold is quite different. By simply averaging the single-session maps, the confidence in the middle of the actual roadway is very high. It then gradually bottoms out to the roadsides. This confidence topology provides a sweet spot, where an optimal threshold can be found to fit the ground truth quite well, resulting in a high F_1-score. However, the threshold range in which the F_1-score is good is very small and important details are lost as shown in Fig. 7. This behavior is not very robust and furthermore, it is hard to find the optimal threshold when no ground truth

(a) Aerial image with manually created label in green. (imagery © 2018, Google)

(b) Combination of 12 driving session by naive averaged overlay.

(c) Resulting roadway mask when applying an optimal threshold of 0.36 to (b).

(d) Combination of 12 driving session with our presented approach.

(e) Resulting roadway mask when applying an optimal threshold of 0.53 to (d).

Fig. 7. Exemplary comparison of roadway map results for a naive averaging approach and our method.

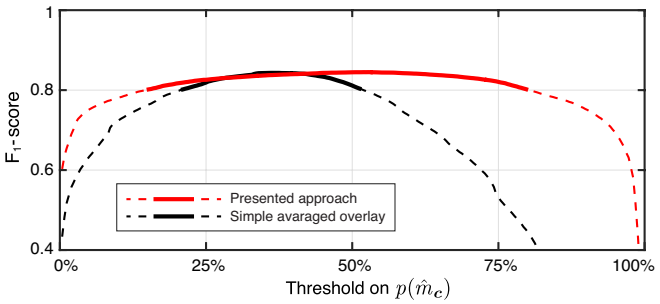


Fig. 8. Comparison of the F_1-scores depending on different thresholds for a naive averaged overlay approach and our proposed combination method. The presented method has a much wider range of thresholds which yield a high F_1-score. The solid lines highlight F_1-scores of 0.8 and higher.

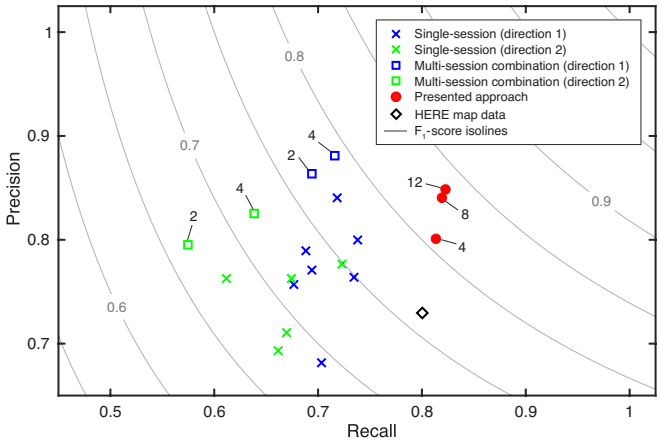


Fig. 9. Precision, recall and F_1-score for different stages of our approach and for standard roadway map data (HERE). The attached numbers indicate the quantity of the combined driving sessions.

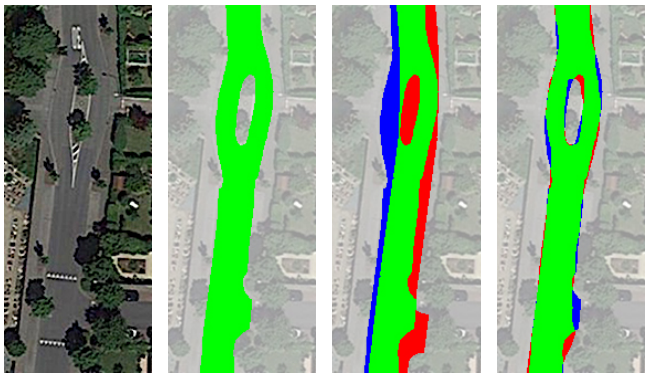
is available. Figure 8 illustrates this issue. Furthermore, the optimal threshold value for the averaging approach is below 50 %, which is a rather illogical choice, since it includes also roadway cells which were classified as road in less than half of the images. Using our proposed approach, a high F_1-score can be reached by a much wider and more meaningful range of thresholds. This is due to the fact, that our created confidence map has much sharper edges and more distinct details than a simply averaged map.

D. Roadway Map Comparison

To assess the benefits of the presented multi-session map combination in more detail, various roadway maps from different stages of our approach and standard roadmap data from HERE are compared to our ground truth. In order to achieve comparability, the same threshold is applied to all of the maps. We only consider cells with confidences over 0.66 as roadway which is a quite conservative but reasonable choice for the threshold. Figure 9 shows the resulting precision, recall and F_1-score values. The final results of our approach are visualized as red circles for different numbers of combined sessions. To illustrate how the combination of multiple sessions improves the mapping results, evaluation points from intermediate combination stages are also visualized. The crosses represent maps from one single session, either in direction d_1 or d_2. In these cases, the average F_1-score is about 0.73. In addition, measures from multi-session maps of only one driving direction are displayed as squares. For those maps precision and recall both increase with the number of considered maps. Combining both driving directions to one final roadway map leads to an F_1-score of 0.84 for 12 sessions. However, the precision decreases slightly since the maps of d_1 and d_2 still suffer from minor localization inaccuracies, which leads to a slightly broader roadway after their combination.

When comparing our final map to roadway maps that are the common standard today, clear advantages are apparent. The F_1-score of the standard roadmap evaluated with our ground truth is illustrated as a black diamond in Fig. 9. Its F_1-score of 0.76 it is much lower than the F_1-score of our map and it additionally has quite low precision. The advantage of our proposed method in capturing roadway details is also visible in Fig. 10, where a section of the roadway containing a traffic island is depicted. The real course of the road is shown in the aerial image in Fig. 10(a) and the manually labeled roadway in Fig. 10(b). In the color coded evaluation of the standard roadway map in Fig. 10(c), the lack of precision can be seen by the red false-positive areas and the blue false-negative areas. It is clear that the traffic island is not represented in this standard map. In addition, the position of the roadway is slightly shifted. On the other hand, the map created by our approach (Fig. 10(d)) contains the traffic island with great detail and only small misclassifications (red and blue areas) are visible.

In conclusion, we could identify two main benefits of our approach. Firstly, the combination of various maps from multiple driving sessions increases the map quality significantly



(a) Aerial image from Google Maps. (b) Manually created roadway label. (c) Evaluation of a standard map (HERE). (d) Evaluation of our crowd-sourced map.

Fig. 10. Comparison between a standard roadway map and the results of our approach. The colors represent true-positives (green), false-positives (red) and false-negatives (blue). Manually labeled aerial images are used as ground truth. (aerial imagery © 2018, Google)

Fig. 11. By comparing our roadway map to the current roadway detection, additional useful information can be extracted, e. g. parking spots on the roadside.

in comparison to single-session maps. Secondly, by locally warping and combining different maps, a considerable quality improvement in comparison to a naive map combination and to a common standard map is achieved.

V. Conclusion

This paper proposes an algorithm for the creation of highly detailed roadway maps of urban areas based on low-cost sensors. Therefore, roadway masks are computed from camera images by a convolutional neural network and are subsequently stitched together using the global pose of the vehicle. The main contribution of this paper is the fusion of multiple detections in a probabilistic roadway grip map and the combination of multiple driving sessions by a feature-based map warping to correct localization inaccuracies.

While our roadway maps can be directly used in intelligent vehicles, they can also be employed to generate even more information. After obtaining a map as proposed, the comparison of the map and the current roadway detection can be used to identify temporarily occupied roadway space. This allows an easy detection of moving objects, parking spots or construction sites. For example, Fig. 11 illustrates how parking spots are identified. Furthermore, the proposed mapping pipeline could also map additional information extracted from the images like lane markings or the texture of the road surface. For these attributes our procedure with perspective mapping and map combination can be utilized in exactly the same way as for the roadway probability.

In the future, it should be investigated, how the combination of even more driving sessions would improve the results and how to merge maps from different overlapping routes. The proposed mapping procedure easily scales to larger mapping areas since all procedure steps influence only a limited local area. Moreover, a specially designed feature detector and matcher could be employed that utilizes more roadway-specific knowledge to improve and accelerate the combination of multiple sessions.

Acknowledgment

We kindly thank Continental for their great cooperation within Proreta 4, a joint research project of TU Darmstadt and Continental to investigate future concepts for intelligent and learning driver assistance systems.

References

[1] N. Mattern, R. Schubert, and G. Wanielik, "High-accurate vehicle localization using digital maps and coherency images," in *IEEE Intelligent Vehicles Symp.*, 2010.

[2] P. Vansteenwegen, W. Souffriau, and K. Sörensen, "Solving the mobile mapping van problem: A hybrid metaheuristic for capacitated arc routing with soft time windows," *Computers & Operations Research*, 2010.

[3] G. Máttyus, S. Wang, S. Fidler, and R. Urtasun, "Enhancing road maps by parsing aerial images around the world," in *IEEE Int. Conf. on Computer Vision*, 2015.

[4] O. Pink and C. Stiller, "Automated map generation from aerial images for precise vehicle localization," in *IEEE Int. Conf. on Intelligent Transportation Systems*, 2010.

[5] H. Roh, J. Jeong, Y. Cho, and A. Kim, "Accurate mobile urban mapping via digital map-based SLAM," in *Sensors (Basel)*, 2016.

[6] K. Ishikawa, J. ichi Takiguchi, Y. Amano, and T. Hashizume, "A mobile mapping system for road data capture based on 3D road model," in *Computer Aided Control System Design*, 2006.

[7] G. He, "Design of a mobile mapping system for GIS data collection," in *Int. Archives of Photogrammetry and Remote Sensing, Vienna*, 1996.

[8] M. Schreiber, A. Hellmund, and C. Stiller, "Multi-drive feature association for automated map generation using low-cost sensor data," in *IEEE Intelligent Vehicles Symp.*, 2015.

[9] M. Naumann and A. Hellmund, "Multi-drive road map generation on standardized high-velocity roads using low-cost sensor data," in *IEEE Int. Conf. on Intelligent Transportation Systems*, 2016.

[10] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2017.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, 2017.

[12] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[13] M. Teichmann, M. Weber, J. M. Zöllner, R. Cipolla, and R. Urtasun, "MultiNet: Real-time joint semantic reasoning for autonomous driving," *CoRR*, vol. abs/1612.07695, 2016, arXiv.

[14] "The KITTI vision benchmark suite," http://www.cvlibs.net/datasets/kitti/eval_road.php, accessed: Jan 12, 2018.

[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Int. Conf. on Learning Representations*, 2015.

[16] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. Computer Vision*, 2002.

[17] Y. Ma, S. Soatto, J. Kosecka, and S. S. Sastry, *An Invitation to 3-D Vision: From Images to Geometric Models*. Springer, New York, 2003.

[18] L. Wasserman, *All of Nonparametric Statistics*, ser. Springer Texts in Statistics. Springer, New York, 2006.

[19] C. Harris and M. Stephens, "A combined corner and edge detector," in *Proc. of the 4th Alvey Vision Conf.*, 1988.

[20] F. Zhao, Q. Huang, and W. Gao, "Image matching by normalized cross-correlation," in *IEEE Int. Conf. on Acoustics Speed and Signal Processing*, 2006.