Multi-Sensor Self-Localization
Based on Maximally Stable Extremal Regions

Hendrik Deusch\textsuperscript{1}, Jürgen Wiest\textsuperscript{1}, Stephan Reuter\textsuperscript{1}, Dominik Nuss\textsuperscript{1}, Martin Fritzsche\textsuperscript{2} and Klaus Dietmayer\textsuperscript{1}

Abstract—This contribution presents a precise localization method for advanced driver assistance systems. A Maximally Stable Extremal Region (MSER) detector is used to extract bright areas, i.e. lane markings, from grayscale camera images. Furthermore, this algorithm is also used to extract features from a laser scanner grid map. These regions are automatically stored as landmarks in a geospacial data base during a map creation phase. A particle filter is then employed to perform the pose estimation. For the weight update of the filter the similarity between the set of online MSER detections and the set of mapped landmarks within the field of view is evaluated. Hereby, a two stage sensor fusion is carried out. First, in order to have a large field of view available, not only a forward facing camera but also a rearward facing camera is used and the detections from both sensors are fused. Secondly, the weight update also integrates the features detected from the laser grid map, which is created using measurements of three laser scanners. The performance of the proposed algorithm is evaluated on a 7 km long stretch of a rural road. The evaluation reveals that a relatively good position estimation and a very accurate orientation estimation (0.01 deg ± 0.22 deg) can be achieved using the presented localization method. In addition, an evaluation of the localization performance based only on each of the respective kinds of MSER features is provided in this contribution and compared to the combined approach.

I. INTRODUCTION

A precise and reliable localization of road vehicles is a prerequisite for future advanced driver assistance systems, especially in the context of highly automated driving. Tasks from planning a trajectory for overtaking to associating other road users to their respective lanes would highly benefit from the precise knowledge of the (local and global) pose of the vehicle. A localization accuracy for such tasks should be within the decimeter range. This cannot - and particularly not reliably - be achieved by using GPS alone. Therefore, a lot of research on vehicle localization has been going on in recent years. Especially algorithms focusing on camera localization have been published, caused by the fact that this sensor is already available in many new vehicles. Alonso et al. presented an algorithm that estimates the global position of the vehicle solely using visual odometry and a digital road map [1]. It uses visual odometry to estimate the trajectory which is then matched to the digital map. A different type of camera based localization was proposed by Badino et al. [2]. There, a large amount of visual features is collected and stored in a data base. Then a discrete Bayes filter is used to estimate the vehicle position based on detections of these features. The work of Schreiber et al. [10] presents an approach that is more similar to the one presented in this contribution. There, a map of lane markings and curbs is generated which is then matched with the online detections of a stereo camera system while making use of a Kalman filter to estimate vehicle pose. Nevertheless, also other sensors are very well suited for localization algorithms. For example, in [5] data from a LIDAR is used to build probabilistic maps for localization. This method allows for very precise localization within the decimeter range.

This contribution describes a particle filter based localization method that uses features from a camera image as well as from a 2D laser scanner grid map. The same feature detection algorithm, the maximally stable extremal regions detector, is used for both types of input. On the one hand it is used to detect bright regions on the dark ground in the camera images, i.e. usually lane markings on the tarmac. On the other hand, it is used to detect occupied regions in the laser scanner grid map, which can originate from arbitrary objects. Thus, this localization procedure is especially well suited for rural roads, due to the availability of lane markings and the mostly stationary surroundings of the roads.

The remainder of this paper is organized as follows: Section II explains the overall localization system along with the features that are used. Details on the filtering mechanism are given and also the fusion of the sensor data is outlined. In section III the performance of the presented method is evaluated. Finally, a conclusion is given in section IV.

II. LOCALIZATION ALGORITHM

This section is divided into four parts. The first part details the overall system design. Afterwards, the feature extraction algorithm is presented. Then it is shown how landmarks and online feature detections are used by a particle filter to estimate the vehicle’s pose. This section finally closes with a description of how the two types of features, i.e. those from the camera images and those from the laser scanner grid map (as used in [14]) can be easily combined in the estimation process.

A. System Overview

The localization algorithm presented in this contribution is premised on a map of landmarks. As a first step, this map has to be created. For now, the map creation process inevitably requires a real-time kinematic (RTK) system with differential GPS (D-GPS) to determine the vehicle’s global
position with adequate accuracy. The procedure for the map functions is as follows (see Fig. 1):

While the vehicle is driven on the road to be mapped, the forward facing camera takes images with a frame rate of 15 Hz. In each frame, maximally stable regions are detected in region of interest placed on the ground plane in front of the vehicle. These regions are then transformed into global coordinates based on their position in the image and the D-GPS position of the vehicle making use of the flat world assumption. The thereby generated “landmark-candidates” are then tracked until they leave the field of view (FOV) of the camera, taking into account their estimated size and global position. All landmarks that have been tracked successfully for a certain amount of frames are stored in a geospatially enabled relational data base.

The landmarks from the laser scanner grid map are extracted at the same time in a similar fashion. A globally referenced occupancy grid map is generated taking into account the vehicle’s pose given by the RTK system and the laser scanner measurements. Then the MSER algorithm is used to extract the “landmark-candidates”, which are again tracked and then stored in the same data base if necessary. For more details about this process, see [14].

After a region has been mapped, the data base provides the landmarks used by the localization algorithm, see Fig. 2 and paragraph II-C. The localization algorithm itself starts off with the initialization of the particle filter based on a GPS position. During the filtering process, the vehicle’s speed and its yaw rate are used to predict the particles. Further, features are extracted from the images of a forward facing camera, from those of a rearward facing camera as well as from the occupancy grid. To assess the hypothesized vehicle poses (i.e. the particles), these features are then compared to their corresponding landmarks within the combined FOV of the cameras or the landmarks within the FOV of the laser scanners, respectively. The unity of the particles is then used to represent the final pose estimate.

B. Feature Extraction

Landmarks for any localization algorithms should have two major qualities: First, as the name already suggests, they should always stay at the same global position. Second, they should be reliably detectable. In the context of localization for road vehicles, lane markings satisfy both of these conditions. Although the position of lane markings may be modified due to road works, these changes are relatively rare events so that lane markings can be regarded as stationary landmarks. Nevertheless, lane markings are not always available and even if they are available, they are not always visible, e.g. due to fog or darkness. Thus, using a complementary sensor, e.g. a laser scanner, is definitely beneficial for a localization algorithm.

The detection of lane markings has been widely studied with the result that there exists a wide range of algorithms to accomplish this task. The Maximally Stable Extremal Regions algorithm has been chosen for this contribution, since it is fast and robust [3], [8]. MSERs were introduced by Matas et al. [6] in the context of wide-baseline stereo matching. Although the intended use of this algorithm was an wholly different one, it proves to be very useful in the context of lane marking detection [11], [15]. Moreover, the MSER algorithm is also very well suited to detect occupied spaces of any given size within an occupancy grid, [14]. The grid map can be interpreted as an graylevel image. Thus, MSERs can be detected on it. These regions therefore represent arbitrary, stationary objects which reflect the laser beam. Thus, the same exact algorithm can be used to detect entirely different features from entirely different sensor data.

As their name implies, MSER describe regions that possess an extremal property, i.e. the intensity $I$ of all pixels $p$ belonging to a region $R$ is higher than the intensity of all pixels $q$ that constitute the outer boundary $B$ of the region

$$\forall p \in R, \forall q \in B \rightarrow I(p) > I(q).$$  \hspace{1cm} (1)$$

This equation describes only maximum intensity regions, since only those are required to detect lane markings. Nevertheless, also minimum intensity regions can be detected by the MSER algorithm [6] by either inverting the image or by slightly modifying equation 1. This has to be done.
in order to extract the features from the occupancy grid, since the occupied areas have a lower intensity than the free space (see Fig. 4). The second property of the MSER ensures that they are “maximally stable”. This means, that the relative change of the area of a region is minimal, while the image is binarized with a sequence of increasing thresholds $g \in [\text{min}(I), \text{max}(I)]$. That is, for a stability range parameter $\Delta$ a “maximally stable” region $R_i$ is found when

$$\Psi(R_i) = \frac{|R_i + \Delta| - |R_i - \Delta|}{|R_i|}$$  \hspace{1cm} (2)

reaches a local minimum at $i$.

The regions found by the MSER algorithm can be used as landmarks by first transforming from image coordinates into vehicle coordinates and subsequently transforming to world coordinates using a kind of “ground truth” vehicle position (i.e. UTM or WGS 84). The features from the grid map are already defined in world coordinates during the mapping phase. Then, one can either store the whole region or one only stores the centroid and the size of the regions (also calculated in world coordinates). To simplify the storage process and to save disk space, the latter is done in this contribution. Further, storing the exact region does not provide any benefit in many cases, especially for the lane markings. They can be either distinguished by their size (e.g. arrows and line segments) or they can not be distinguished at all (e.g. two line segments of the same length).

It is also possible to define a minimal area and maximal area of a MSER easily. This is particularly useful if one wants to ignore very large lane markings (e.g. a continuous lane marking) which do not provide any longitudinal position information. In addition, very small regions can be neglected, since lane markings usually have a well defined minimal size. This property is of equal importance for the occupancy grid features, in order to be able to only consider objects of certain sizes.

To provide more MSER detections in any situation, the algorithm is applied to images of a forward facing camera and a rearward facing camera at each time step. The cameras are time triggered such that the images are taken at exactly the same point in time.

In Figure 3, two examples of the MSER detections are shown. It can be seen, that the algorithm is able to find the exact contours of most lane markings within the region of interest on the ground plane. Only lane markings that either intersect any edge of the ROI or that do not match the size constraints are neglected. Figure 4 shows an example of the MSER detections in an occupancy grid.

C. Localization Particle Filter

The presented localization filter is closely related to the one presented in [14]. As described in [9], [12], a set of $N$ particles $P_k = [p_1^k, ..., p_N^k]$ and their associated weights $W_k = [w_1^k, ..., w_N^k]$ are used to represent the state to be estimated.

1) Particle State and Initialization: Since the global pose of the vehicle is to be estimated, the state comprises three components: a UTM easting coordinate $E$, a UTM northing coordinate $N$ and an orientation angle towards the east axis of the UTM cell $\psi$. The UTM cell number and letter are neglected here. Therefore, the state vector of particle $i$ at time step $k$ is given by

$$p_i^k = [E \ N \ \psi]^T.$$  \hspace{1cm} (3)
The particles are initialized using the position provided by a standard GPS receiver and the orientation given by a compass. A random pose within an uncertainty ellipsoid around these values is assigned to each particle.

2) Prediction: To account for the movement of the vehicle within a time span $\Delta t$, the hypothesized vehicle poses have to be predicted, before an update of the particle weights can be performed. To do this, standard on-board vehicle sensors are used to determine the speed $v_k$ and yaw rate $\psi_k$ of the vehicle. This control input is then – for each particle individually – superimposed with noise to also respect the measurement uncertainties of these sensors. Then, the well known dead reckoning formulation is used to determine the predicted poses:

$$\Delta \psi_k = \psi_k \cdot \Delta t$$

(4)

$$E_{k+1} = E_k + \frac{v_k}{\dot{\psi}_k} \cdot \left[ \sin (\psi_k + \Delta \psi_k) - \sin (\psi_k) \right]$$

(5)

$$N_{k+1} = N_k + \frac{v_k}{\dot{\psi}_k} \cdot \left[ \cos (\psi_k) - \cos (\psi_k + \Delta \psi_k) \right]$$

(6)

$$\psi_{k+1} = \psi_k + \Delta \psi_k.$$  

(7)

Since a rather simple motion model is used, which amongst other things neglects slipping, process noise has to be added as a final step.

3) Weight Update: The weight, which is calculated during the weight update step, describes how well the set of observations $Z_k = \{z^1, \ldots, z^m\}$, i.e. the online MSER detections transformed into vehicle coordinates, matches the stored landmark map $M_k$. To allow for a fair comparison of those sets, the map has to be restricted to the FOV of the cameras (or laser scanners respectively). Thus, the set $M_k^{\text{ROI}} = \{m^1, \ldots, m^{\text{m}}\}$ only contains landmarks that are within that FOV. Then, for each particle, the landmarks of $M_k^{\text{ROI}}$ are transformed into the hypothesized vehicle coordinate frame originating at the pose described by the particle state. Thereafter, a likelihood for all combinations of transformed landmarks $m^l$ and detections $z^j$ is determined

$$g(z^j|m^l) = \frac{p_D}{(1-p_D) \cdot \kappa(z^j)} \cdot \exp\{-0.5 \cdot \Lambda\},$$

(8)

with

$$\Lambda = \frac{d_{\text{lat}}^2}{\sigma_{\text{lat}}^2} + \frac{d_{\text{lon}}^2}{\sigma_{\text{lon}}^2} + \frac{a_j^2}{\sigma_a^2}. $$

(9)

Here, $d_{\text{lat}}$ and $d_{\text{lon}}$ describe the lateral and longitudinal distance between the landmark $m^l$ and the detection $z^j$, while $a_j$ describes the respective difference in the area of the MSERs. The state independent detection probability of any landmark is denoted as $p_D$, and $\kappa(z^j)$ describes the intensity of a Poisson distributed clutter process, cf. [13].

In contrast to the likelihood function presented in [14], equation 8 explicitly takes into account different measurement uncertainties of the longitudinal position and the lateral position. This is of course due to the different measurement principles of the employed sensors. One certainly has to assume a larger uncertainty in the longitudinal position than in the lateral position, when a camera is used. In contrast to this, the longitudinal uncertainty and the lateral uncertainty should be identical when a laser scanner is utilized.

For determining a particle’s final weight, the best association hypothesis $\theta : \{1, \ldots, m_k\} \rightarrow \{0, 1, \ldots, z_k\}$, has to be found. The measurement denoted by “0” represents missed detections. Missed detections within this context occur primarily when other traffic participants either occlude a lane marking or when they restrict the laser scanners’ FOV. The likelihood for a missed detection is given by

$$g(z^0|m^l) = 1 - p_D. $$

(10)

All measurements, except for the missed detection measurement “0”, may at most be associated to one landmark. The Munkres [7] algorithm is used to find the optimal association hypothesis $\theta$ respecting this constraint. Based on this hypothesis, the weight of particle $p_k$ is

$$w^i_k = \prod_{l=1}^{m} g(z^0(l)|m^l). $$

(11)

These importance weights are of vital importance within the resampling step. In this contribution, a low variance sampler is used to perform the resampling [12].

4) Resulting State Estimate: The final state estimate is calculated by averaging over the states of all particles. The results presented in section III suggest that this simplistic calculation of the estimated state is sufficient in this use case.

D. Fusion

In [14] a localization algorithm based solely on MSER detections in laser scanner grid maps is presented. This contribution proposes the fusion of those features with features detected in camera images. The main advantage lies in the increased availability of features, i.e. lane markings may be present in situations where no proper features can be found in the grid map and vice versa.

Both kinds of features can be easily used altogether in one localization algorithm, as long as the sensors are synchronized. If this is the case, one can calculate two separate weights $w_{\text{Laser}}$ and $w_{\text{Camera}}$ each according to equations 8 and 11 with their respective sets of measurements and landmarks. The final weight can then be obtained by

$$w_k^i = w^i_{\text{Laser}} \cdot w^i_{\text{Camera}}. $$

(12)

This final weight then reflects how good a hypothesized pose matches the lane marking landmarks as well as the laser scanner grid map landmarks.

Nevertheless, if the sensors are not synchronized, as it is the case with the sensors used in this contribution (the laser scanners run at 12.5 Hz, while the cameras run at 15 Hz), one has to resort to other methods. First of all, the particles have to be predicted between the different measurement updates, since they originate from a different point in time. Furthermore, another problem arises with the sensors not being synchronized [4]: Out of sequence measurements may
occur due to the different data preprocessing methods, which require a different amount of time that may or may not be constant. This means, it is possible that the features extracted from e.g., the cameras are ready just after the features have been extracted from the occupancy grid, although the cameras recorded their pictures well before the laser scan was triggered. If this happens, a retrodiction has to be done in order to be able to integrate the delayed features or they have to be discarded. It is also possible to avoid such a retrodiction step, if data buffering is applied. If the data is buffered, the feature detections can be sorted according to the recording time stamps of their respective sensors. Subsequently, the particles can be updated with the sorted buffered measurements while predicting in between the updates. Although this approach leads to an inherent delay in the system, it is used in the presented localization algorithm. Also, a time out can be integrated to avoid that the system waits arbitrarily long for data of a malfunctioning sensor.

### III. Evaluation

#### A. Experiment Description

The experiment has been conducted on a German rural road of approximately 7 km length. The test vehicle was a Mercedes-Benz E-Class. It has two additional cameras installed. The forward facing camera is a Baumer TXG14f with a resolution of 1392 × 1040 px and a horizontal FOV of about 40 deg. The rearward facing camera is a Baumer TXG06 with a resolution of 776 × 582 px and a horizontal FOV of about 56 deg. Both cameras are synchronized and time triggered, so that they take exactly 15 frames per second. The vehicle also disposes of three IBEO LUX laser scanners. They are mounted in the front bumper in a way, that they provide a total horizontal opening of 210 deg. The laser scanners’ horizontal angular resolution is 0.125 deg and their radial resolution is 0.04 m. They are synchronized and time triggered with a frequency of 12.5 Hz. Further, the standard sensors of the vehicle measure the speed of the vehicle and its yaw rate with a rate of 50 Hz.

The vehicle is equipped with a computer (Intel Core i7 3.2 GHz, 12 GB of RAM, running Ubuntu 12.04 64 Bit) which runs the localization algorithm in real time using 2500 particles. The map of landmarks was created by again driving the same route and localizing the vehicle with the help of a RTK system and differential GPS. This way, the location of the vehicle is known with an accuracy of approximately 2 cm, so that the detected features have been transformed to a relatively precise global position. Thus the accuracy of the resulting map of landmarks should also be roughly within the centimeter range. The map for this 7 km long road contains 930 lane marking landmarks and 1644 landmarks from laser scanner grid maps. Storing this as raw data, i.e. without spatial index and additional information, would lead to a memory consumption of circa 18 kB and 31 kB respectively (2 doubles for the position and one float for the area of the MSERs). That implies that less than 3 kB of data per kilometer is sufficient for a map of lane marking landmarks, or respectively less than 5 kB for the grid map features.

#### B. Results

All results were averaged over 50 Monte Carlo runs of the particle filter. Only one recording of a test drive was used for the evaluation. The map was generated based on a recording from a different date (approximately 3 months earlier). The ground truth vehicle pose to which the localization estimation was compared, was given by the RTK system (as during mapping).

The results of the solely lane marking based localization are depicted in first three rows of Table I. It is obvious that the algorithm is able to reliably estimate the vehicle pose with high accuracy. Especially the orientation error is really low with almost zero mean and a standard deviation below 0.22 degrees. Also the lateral position error is very low. The performance of the estimation of the longitudinal position is inferior and even exhibits an offset. The offset is most likely due to suboptimal calibration of the cameras. The higher standard deviation of the longitudinal position error is founded on the measurement principle of a mono camera. Further, also the pitch angle of the vehicle is not taken into account, which leads to relatively high errors whenever the vehicle accelerates or decelerates. Nevertheless, the overall performance of this localization algorithm is good enough for almost any advanced driver assistance system (e.g., a semi-autonomous overtaking assistant). This holds especially true since the global lateral position and orientation are more important than the longitudinal position for most systems. Usually only the longitudinal position relative to other traffic participants, which can be measured directly by e.g. a radar, is of interest. The performance of the algorithm based exclusively on laser scanner grid maps has also been evaluated: See the three central rows Table of I. It is obvious that the laser scanner based approach outperforms the camera based approach significantly in the estimation of the longitudinal position. On the other hand the lane markings are a very strong cue for the lateral position and the orientation. This fact leads to a slightly better performance in these domains. Using a camera has two significant advantages over using a laser scanner: availability and cost. Another advantage of this method is, that it only uses bright regions detected on the

<table>
<thead>
<tr>
<th>Combination</th>
<th>mean</th>
<th>standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Longitudinal error in m</td>
<td>-0.0298</td>
<td>0.1338</td>
</tr>
<tr>
<td>Lateral error in m</td>
<td>0.0124</td>
<td>0.2117</td>
</tr>
<tr>
<td>Orientation error deg</td>
<td>0.0038</td>
<td>0.2527</td>
</tr>
<tr>
<td>Laser Scanner Grid Map Only</td>
<td>mean</td>
<td>standard deviation</td>
</tr>
<tr>
<td>Longitudinal error in m</td>
<td>-0.0719</td>
<td>0.1641</td>
</tr>
<tr>
<td>Lateral error in m</td>
<td>-0.1272</td>
<td>0.1578</td>
</tr>
<tr>
<td>Orientation error deg</td>
<td>0.0017</td>
<td>0.2285</td>
</tr>
<tr>
<td>Camera Only</td>
<td>mean</td>
<td>standard deviation</td>
</tr>
<tr>
<td>Longitudinal error in m</td>
<td>-0.2143</td>
<td>0.3613</td>
</tr>
<tr>
<td>Lateral error in m</td>
<td>-0.0670</td>
<td>0.1134</td>
</tr>
<tr>
<td>Orientation error deg</td>
<td>0.1578</td>
<td>0.2904</td>
</tr>
</tbody>
</table>

#### TABLE I: Localization performance of the proposed approach.
ground plane, i.e. almost solely actual lane markings and thus strong landmarks. The other approach uses arbitrary regions of a occupancy grid, which may not always be as stationary as they seem (e.g., pedestrians or vegetation that changes shape over the seasons). Nevertheless combining both types of sensors makes the system more reliable and adaptable. The combination of the lane marking features and those from the laser scanner grid map are shown in the bottom rows of Table I. Figure 5 shows the results of an exemplary run of the filtering algorithm. When both sensors are used in combination, the overall performance increases. The combined approach offers the best estimation of the lateral position. Further, the longitudinal position is estimated equally well as with the grid map features, while the orientation estimation is almost as good as with the lane marking features.

IV. CONCLUSION

This paper presented a particle filter based localization algorithm. Maximally Stable Extremal Region features detected in grayscale camera images and in laser scanner generated grid maps are used for the importance weighting of the particles. It is described how those kind of features can be used in combination. The approach has been evaluated and the results presented in section III indicate that this localization approach is able to estimate the pose of a road vehicle robustly and very precisely on a German rural road. Also, using features extracted from camera images alongside those from the laser scanner data does increase the performance. Of course one has to take into account that computational cost also increases. But this drawback becomes more or less insignificant when also regarding the higher reliability of such a localization system. Since the two types of features complement each other, using them both significantly reduces the cases in which the system fails. For example, the purely laser scanner based localization may fail in the presence of crash barriers (because it cannot find any features) and the camera based localization will fail if there are no lane markings, the overall system will only fail if both of these conditions occur at the same time.

The major drawback of the proposed localization schemes is that a very precise map is required. This map has to be generated using a vehicle that is equipped with very expensive reference system. To overcome this shortcoming, SLAM approaches will be investigated.

REFERENCES