Road Terrain Detection for Advanced Driver Assistance Systems

Dissertation

zur Erlangung des akademischen Grades
Doktor der Ingenieurwissenschaften (Dr.-Ing.)

vorgelegt an
der Technischen Fakultät der Universität Bielefeld

von

Tobias Kühnl

21.05.2013

Abstract

In recent years, automotive manufacturers have equipped their vehicles with innovative Advanced Driver Assistance Systems (ADAS) to ease driving and avoid dangerous situations, such as unintended lane departures or collisions with other road users, like vehicles and pedestrians. To this end, ADAS at the cutting edge are equipped with cameras to sense the vehicle surrounding. An important source of information for future ADAS is the road course, i.e., the future driving path of the ego-vehicle and other vehicles. Therefore, this thesis focuses on the camera-based analysis of road scenes and the detection of important types of road terrain, such as road area and ego-lane, which are necessary to draw inference about the actual road course and potential space for evasion maneuvers.

For this purpose, this thesis presents a generic concept for the visual and spatial analysis of the road environment. The core of the proposed method is a hierarchical feature extraction that combines local visual appearance with its spatial layout. In this sense, a novel vision-based approach for road terrain detection that goes beyond classical lane marking detection and image segmentation approaches is presented. Thus, the approach enhances the ability to cope with noise and appearance changes because the classification decision is not only based on local visual appearance but on a combination of visual and spatial aspects. This results in a higher robustness under various visual conditions due to different asphalt appearance, illumination changes, and shadows.

The approach's generic architecture internally represents certain visual properties, such as road area, road boundary, and lane marking information by means of a visuospatial representation. In contrast to many related approaches for road terrain detection, the proposed method does not employ an explicit road course model. Instead, the method learns classifying road terrain based on a combination of visual and spatial features by using machine learning. Especially the discrimination between ego-lane and other parts of the road area is very challenging, because a distinction based on local appearance is impossible. Extensive evaluations in urban scenarios show that the proposed system functions in spatially diverse road scenes and reliably detects ego-lane and road area even in challenging situations. Those situations may comprise bad-quality or missing lane markings, curbstones delimiting the road, and occlusion of lane delimiters, e.g., caused by parked cars.

Furthermore, the generic concept does not only have advantages in road terrain detection, but also in many other applications, benefitting from visual and spatial scene analysis. In order to prove this, the method is applied for pure vision-based ego-vehicle localization on the lane level. In this regard, a reliable classification allowing an inference about how many lanes exist adjacent to the ego-lane is presented on a large highway dataset.

In summary, this thesis presents a generic concept for visual and spatial analysis of the road environment and is therefore a substantial contribution to the development of future ADAS. Towards this end, the general approach is geared to problem-solving for complex situations that can not be handled by state-of-the-art methods, which has been shown for inner-city road terrain detection and ego-vehicle localization.

Contents

Acknowledgements

All the work in this thesis would definitely not be possible without the remarkable and intensive guidance of my PhD supervisor Jannik Fritsch. Furthermore, my gratitude goes to Professor Franz Kummert for his professional guidance and valuable support throughout the project. As an additional scientific advisor I would like to thank Mathias Franzius for helping me with his expertise. Thanks to the Honda Research Institute (HRI-EU) for funding the PhD project giving me this great opportunity and of course all my colleagues at HRI-EU. Thomas Weisswange, Marcus Kleinehagenbrock and Sebastian Herzog were great reviewers helping me a lot to improve my papers and thesis. My friends, my dear parents and especially my wife Christina gave me great support and encouragement over the years, thanks to all of them.

1. Introduction

Automobile manufacturers have equipped their vehicles with *Driver Assistance Systems*, like navigation systems and parking distance control, fore more than a decade. Even more advanced modules like *Adaptive Cruise Control* and *Lane Departure Warning* are nowadays affordable technologies even in middle-class cars. These modular technologies are only the beginning of new product generations in the field of *Advanced Driver Assistance Systems* (ADAS). In the future, ADAS will pave the way towards autonomous driving. In Figure 1.1, a vision of autonomous driving is illustrated, the daily way to work will not be lined with stress. Instead, drivers can spend their time preparing the next meeting or reading the newspaper.

Figure 1.1.: A group of vehicles in a platoon, succeeding vehicles are automatically following the one ahead. Image taken from the *Safe Road Trains for the Environment* project (see http://www.sartre-project.eu).

Improving traffic safety is the key aim of ADAS. In Germany, millions of motor vehicle accidents cause thousands of fatalities every year. Therefore, there is a high potential benefit in employing ADAS aiding the driver to prevent traffic accidents and therefore reducing fatalities. In contrast to passive safety systems, like seatbelts or airbags, active safety systems actively influence the steering, breaking and accelerating behavior of the vehicle. *Antilock Braking System* (ABS) is probably the most prominent and a default component in nowadays automobiles.

For the function of an ADAS the perception of the vehicle surrounding is essential. With the detection of the vehicle's lane and other vehicles, ADAS warn the driver in hazardous situations and are, beyond that, able to avoid or mitigate the consequences of a collision. Another type of ADAS are designed to ease the driving task and hence increase comfort in modern cars. An example for this type is steering assistance to support drivers in curves and for crosswinds. Beyond this, first autonomous functions are already close to the marked. For example, advanced emergency breaking, used in forward collision avoidance systems, works very effectively for crash avoidance by warning the driver, supporting his breaking, and even breaking autonomously in case he does not react. Because of technology advancements in the field of ADAS, systems will become more and more autonomously and therefore, full autonomous driving will consequently become reality some day in the future.

1.1. Motivation for Road Terrain Detection

The more tasks ADAS will accomplish, the more important becomes the holistic understanding of traffic scenes. Therefore, research on ADAS is shifting from the processing of single entities to a more complete understanding of the environment. Crash prevention systems typically focus on objects with which the vehicle could collide, e.g, other cars, pedestrians, and static obstacles. In contrast, this thesis concentrates on the detection of road terrain which refers to the driving space and its integral parts in the scene. Therefore, it is a specific aspect of scene understanding, but very important for ADAS where more in-depth knowledge about the driving space is required.

For perceiving the traffic environment, different sensors are typically used. Radars sensors are good in detecting objects, like other vehicles. Beyond this, laser scanners are additionally able to detect free space in the vehicle's surrounding by capturing a 3D view of the environment. However, for future ADAS it will be necessary to subdivide the free space into its semantic entities, i.e., the road area and its particular lanes. This requires capturing the visual appearance of road terrain and therefore can only be realized with a camera sensor.

The different categories, road terrain can be subdivided in, are illustrated in Figure 1.2. The definition of the distinct parts will be given in the following. Road terrain includes all ground areas that are relevant for traffic scene understanding. In contrast, non-road terrain refers to larger, elevated objects a vehicle encounters in road scenarios and that cannot be potentially driven over, like, e.g., other cars or buildings. The road area, one subcategory of road terrain, refers to the complete asphalt road region. Other terrain types, apart from road area, are sidewalk, traffic island, and other non-drivable flat areas (e.g., grass) which are referred to as off-limits terrain. The road area consists of the traffic lanes as a whole, named drivable road. On the contrary, non-drivable road describes road area regions that have a special purpose and are not meant to be driven on. A simple example for this is an emergency lane on the highway. Furthermore, drivable road is sub-divided into ego-lane and non-ego-lane.

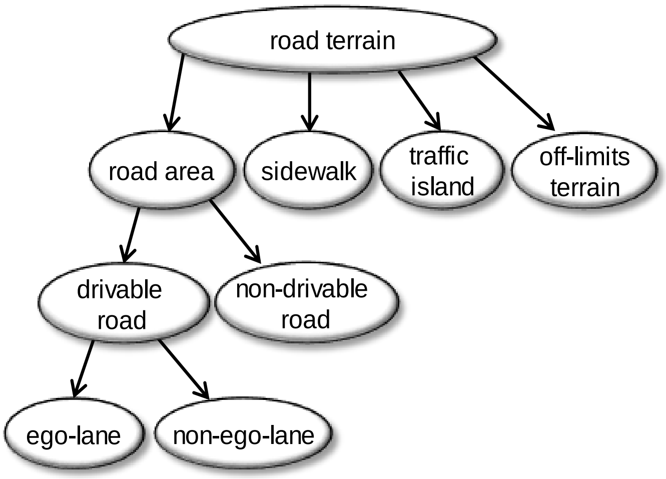

Figure 1.2.: Hierarchical order and dependence of road terrain categories.

Road terrain categories like road area and the ego-lane deliver important context information for all kinds of detection systems like, e.g., vehicles, or pedestrians, because they denote image regions where those road users potentially appear (see Guo and Mita, 2012). Furthermore, it can be used to predict where other traffic participants will probably move in the future (see Barth and Franke, 2008). Road terrain as context information can also be useful for judging the relevance of already detected objects, such as pedestrians. Imagine an assistance function which raises the drivers attention to a specific, most relevant pedestrian, as depicted in Figure 1.3. In case the system has to assess the potential risk or importance of mul-

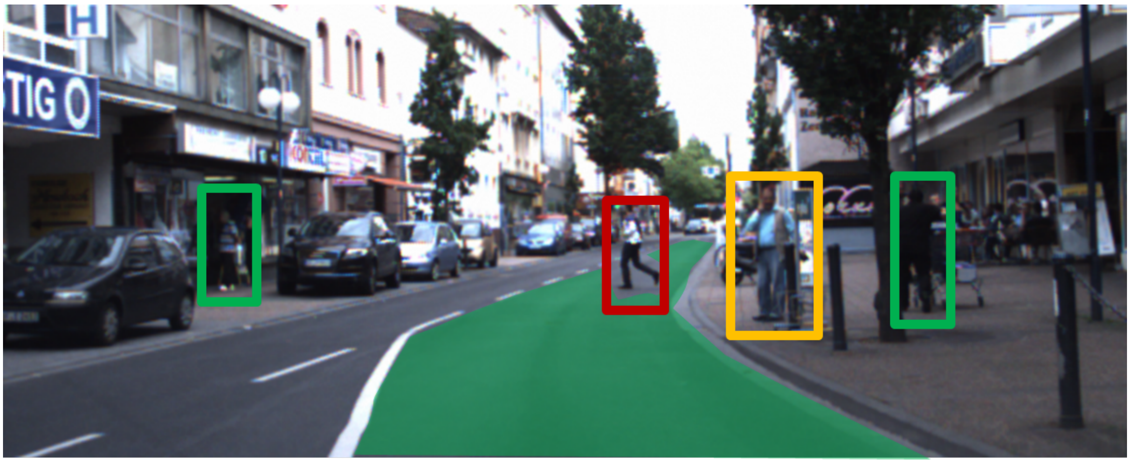

Figure 1.3.: Pedestrian detection problem: Which pedestrian is important? The colored rectangles reflect the importance of detected pedestrians. In this scenario, it is essential to considering the course of the ego-lane (green area).

tiple detections, it can be estimated based on the relative location of pedestrians to the ego-lane (area marked in green). In the shown example, the pedestrian crossing the road (red box) is more important than the pedestrian standing on the sidewalk (yellow box). The other detected pedestrians (green box) can be considered as unimportant.

Another useful information in a traffic scene which is closely related to road terrain is the knowledge about how many lanes exist and on which lane the ego-vehicle is currently driving. The identification of the lane the ego-vehicle is driving on is termed ego-vehicle localization in the context of this thesis. Combining a navigation system with a component for ego-vehicle localization would increase its usability by providing more adequate directions. An example for such a device is depicted in Figure 1.4. Example I shows the same scene from a different lane. In both situations the aim is to take the exit on the right lane in about 500m (behind the bridge). Only in example I-A a navigation message has to be provided, because the ego-lane in scene I-B will lead to the exit directly. This shows how a combination of navigation systems with ego-vehicle localization could avoid a lot of unnecessary messages. Furthermore, example II shows that the device could even give more helpful routing instructions, because it determines the required number of lane changes for reaching the target lane.

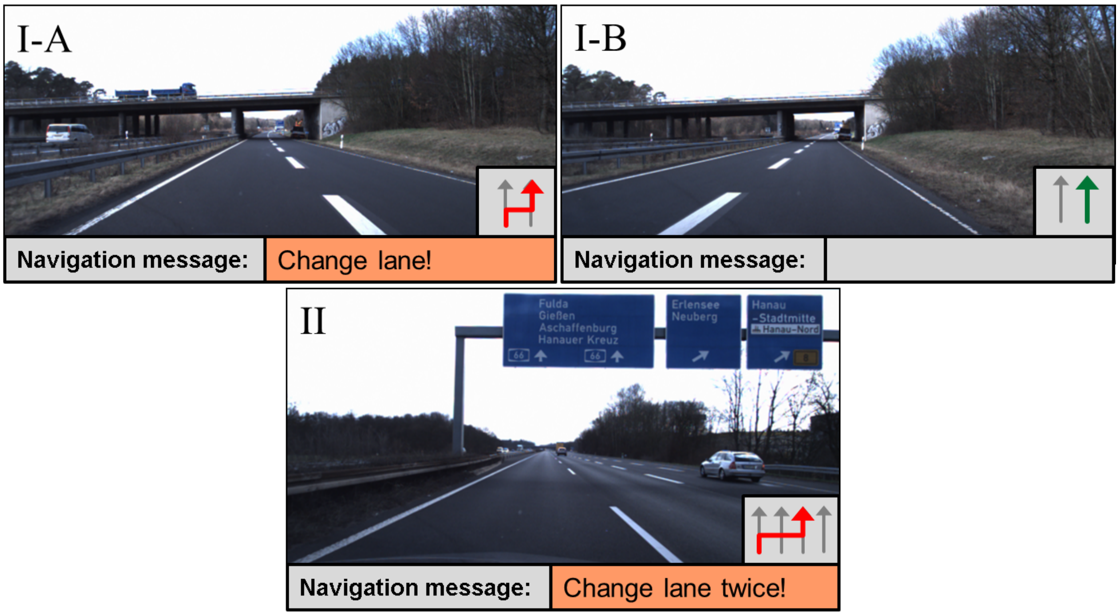

Figure 1.4.: Visualization of more adequate routing instructions by a navigation device combined with ego-vehicle localization. Scene I-A and I-B show the same situation from a different lane position. Only in example I-A, if a lane change is required to reach the target lane, a navigation message is provided. In example II, where reaching the target lane requires two lane changes, a more detailed navigation instruction is provided.

1.2. Restriction of State-of-the-Art

Driver assistance systems for lane keeping assistance or lane departure warning need specific knowledge about where the ego-vehicle will move. This is what is referred to as the ego-lane in this context. Nowadays commercial ADAS are mostly limited to specific scenarios, such as highway where certain conditions hold, e.g., low curvature of the lane and good quality of the lane markings.

However, robust recognition of the driving path on arbitrary roads will be needed for future ADAS operating in more complex traffic situations. On rural roads bad lane marking conditions and even unmarked road sections have to be considered. Furthermore, curves with higher steepness can occur, which implies that model conditions for ADAS have to be less restrictive. In urban scenarios (see Fig. 1.5), scene understanding becomes even more complex. Roads can have almost arbitrary shapes, and numerous lanes with different driving directions. Moreover, high traffic

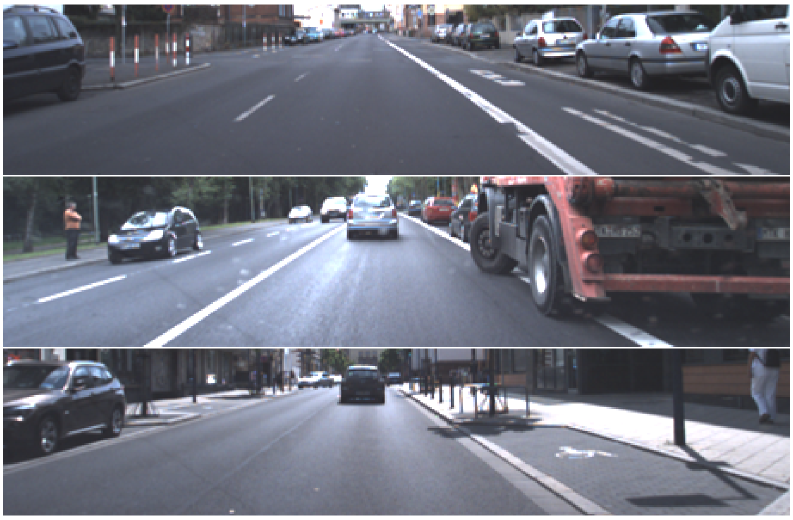

Figure 1.5.: Challenging scenarios in urban areas: Bad lane marking qualities, occlusion of road delimiters, flat curbstones and unmarked roads.

density and pedestrians in urban areas need to be handled adequately. Furthermore, city roads are often delimited by curbstones instead of lane markings which have to be detected as well. In addition to the above mentioned challenges on rural roads, the detection of explicit road boundaries (like curbstones and lane markings) is sometimes inappropriate because of, e.g., heavily worn out lane markings, or occlusion of the delimiters caused by parking cars on the side. In such situations, systems which are based on delimiter detection or use explicit lane models are not working. Even though, improvements in ADAS technologies and advances in research are apparently successful, all the above mentioned facts indicate that standard approaches reach their limit in some situations on rural road and definitely in urban areas. Therefore, it will be an important step for ADAS functioning on rural roads and in urban areas to not only to rely on delimiter-based detection but also to consider the actual road surface as source for road terrain detection.

A drawback of nowadays commercial ADAS is that they do only perceive information of the ego-lane. However, more advanced functions, such as an ADAS supervising lane change maneuvers, would need to identify if the neighboring lanes are blocked. In case the driver changes lane (intentionally or unintentionally) and the neighboring lane is blocked, a warning signal can be provided. In addition, the road area contributes information about the obstacle-free driving space which is relevant, e.g., for emerging maneuvers of the vehicle in case of a forthcoming accident. Such ADAS would require more general knowledge about the surrounding driving space, i.e., the road area and lanes beyond the ego-lane.

For nowadays navigation systems only an approximate vehicle position is available via *Global Positioning System* (GPS). This allows finding out on which road the vehicle is currently driving. However, GPS is not accurate enough to let us know which lane we currently are on. As mentioned above, in the context of this thesis this problem is named ego-vehicle localization. Although there is *Differential GPS* (DGPS), a technology enhancing the accuracy of GPS, it is very expensive and therefore not applicable to consumer cars.

1.3. Basic Concept of the Proposed Approach

The thesis at hand proposes a generic system that can be applied for road terrain detection and ego-vehicle localization. The proposed method has certain advantages compared to related approaches. Aim of this thesis is to detail these differences and point out novelties, limitations, and challenges of the proposed approach. In this section, an overview of the basic concept and the generic framework that comes along with it is given.

The proposed method gathers the necessary environment information by processing images from an in-car mounted camera. The framework comprises two main stages which reflect different levels of scene analysis. Firstly, the image is visually analyzed on a local scale. Following, the spatial layout of local visual properties is analyzed. Therefore, it can be seen as a complementation of traditional local visual features with spatial information using a two-stage process for visuospatial analysis of the road environment.

In the first stage, a visual base classification is proposed. Local visual appearance of an image region can be analyzed by extracting patches from the image. Figure 1.6 shows a road scene, with a close up view for several image patches. This leads to a pattern recognition problem, i.e., in order to infer whether an image region belongs to a certain type of road terrain, we have to find a way to distinguish the patterns and colors contained in these patches. For this purpose three visual appearance models are proposed that are considered as relevant for the given classification task. In particular the three models are: road appearance, road boundary appearance and lane marking appearance. After applying these models, image regions can be assigned to the corresponding model properties based on a comparison of whether a patch matches a given model. As these properties are only locally assigned, this can result in ambiguities. This means that only based on local visual appearance typical road terrain regions like the road surface, sidewalks, or other low-textured

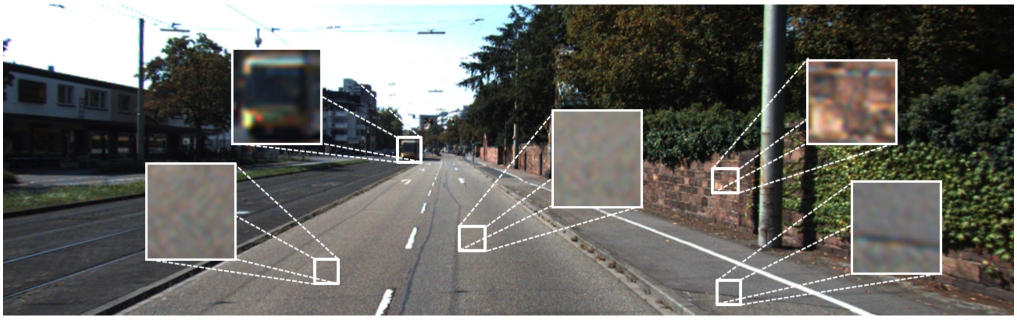

Figure 1.6.: Local visual appearance of image patches in a road scene. The side walk appearance is similar to the road region. The depicted non-road terrain patches are visually distinct.

gray surfaces can hardly be discriminated. This can be seen in Figure 1.6, where the visual appearance of sidewalk and road area is highly similar. As the road area has the same asphalt, there is no way of visually discriminating the lanes. Therefore, the detection of the ego-lane on a local scale is not feasible.

To overcome this limitation, a combination of local appearance with spatial features is proposed. Spatial features, in the context of this thesis, refer to information gathering strategies that are not restricted to a local surrounding. In contrast to local feature gathering (e.g. inside a window with a fixed size), spatial features can consequently combine feature information from multiple spatial locations and capture the extent of regions and shapes. Furthermore, based on a position's spatial relation to regions with certain properties, i.e., the above mentioned appearance properties, a useful visuospatial representation of the image for distinguishing road terrain is obtained. This visuospatial representation shows also to be beneficial for ego-vehicle localization, as it captures the extent of the overall road area and the distance to delimiting elements such as lane markings and curbstones.

The proposed SPatial RAY (SPRAY) features capture information in the shape of rays. An example can be seen in Figure 1.7. Starting from a defined base point (white dot), information is gathered along the ray in a specific direction. The course of a ray is denoted by the increasing color intensity (cyan to the left, magenta to the right). In this example, the feature values represent the distance from the base point to the road boundary. Based on this, the base point on the sidewalk can be distinguished from base points on the road area, because the features indicate the smaller width of the corresponding region. Furthermore, based on whether a location is closer to the left and respectively to the right boundary the road area can be subdivided into the left and right lane (assuming neighboring lanes have the

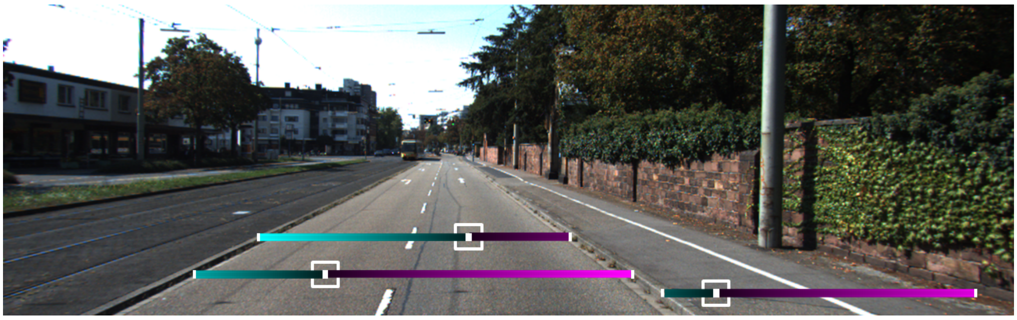

Figure 1.7.: Illustration of a spatial layout computation by rays to the left and right side, for three base points on different road terrain. SPRAY features are depicted by color intensity reflecting a distance to the base point (white dot). The distance until the road boundary, where the ray ends, is used for distinguishing ego-lane, non-ego-lane and sidewalk.

same width). Of course considering the location of lane markings would also lead to a clear distinction for the two road area positions. However, considering only lane markings would probably result in a wrong detection of the base point on the sidewalk because it is located within the area in between the closest lane marking to the left and right side[1]. This shows that it is beneficial to consider all delimiter types for a system operating in urban areas.

In the progress of this thesis, the benefits of combining local visual appearance and spatial layout will be discussed. The output of the system is a multi-usage visuospatial representation which is applied to road terrain classification and ego-vehicle localization. In particular, this thesis applies the generic framework to three different fields:

- Road area detection
- Ego-lane detection
- Ego-vehicle lane position assignment

The parameters of the system are automatically obtained using machine learning, which requires a training session. Therefore, the system can be adapted to function on various road conditions. In particular, conditions can change on the local visual appearance level, e.g., a different asphalt texture, or on the spatial level, e.g., a geometrically different road type.

[1] This assumption is typically applied for commercial ADAS relying on lane markings only.

1.4. Contribution to State-of-the-Art

The approach presented in this thesis aims at contributing solutions for state-of-the-art application problems. As the proposed system has a wide field of application, a differentiation from state-of-the-art can be done separately for road area detection, ego-lane detection and ego-vehicle lane position assignment.

Ego-lane detection suitable for arbitrary delimited roads: State-of-the-art ego-lane detection approaches have disadvantages in situations where the road geometry is complex. Usually, the ego-lane is represented by its delimiters and the spatial course is captured by a model. This model implies certain restrictions to the course, which are often problematic in complex situations. The proposed system does not only represent multiple delimiter types like lane markings and curbstones but also the road area. By learning the spatial layout of all these properties a more general ego-lane representation is obtained. This results in an ego-lane detector which enables reliable detection even in complex situations along with bad-quality or missing lane markings, flat curbstones, and occlusion of lane delimiters. This means the proposed system is suitable for all roads.

Robust visual road area detection: Typically, the road area can be detected by finding unelevated regions in the environment or analyzing visual image features like texture and color. The latter is less popular, because much more affected by noise and appearance changes due to changing scene and lighting conditions. For that reason, methods analyzing the height over ground in the image are robust for changing conditions. However, these methods have the disadvantage that actual road area can not be distinguished from other flat road terrain categories which can be inappropriate in some situations (e.g., for grass) but also dangerous (e.g., in case of a lake). Above mentioned disadvantages of both methods, perfectly lead to the advantages of the proposed system. Firstly, the sensitivity to noise and appearance changes will be reduced by the proposed method because the classification decision is not only based on local visual appearance but on a combination of visual and spatial aspects. Secondly, we include spatial aspects of man-made roads which enables to separate the road area from exterior non-road parts. Assuming the road area exhibits a structured constellation of visual properties, such as larger, gray, untextured asphalt regions, lane markings, and curbstones, this can be internally represented by the proposed system, and can be used to learn a decision where the actual road area is.

Pure visual ego-vehicle localization: State-of-the-art methods for ego-vehicle localization on the lane level mostly utilize expensive technologies such as DGPS. Furthermore, they require detailed digital map data which is only partially available. In contrast, ego-vehicle localization only based on vision is beneficial because camera sensors are cheap. However, pure vision-based ego vehicle positioning was, to the authors knowledge, never proposed before because it is very challenging to solve with the known methods. The proposed method comprises an internal visuospatial representation that appears to be very useful for this application because it captures the extent of the road area and the distance at which delimiters, like road boundaries and lane markings occur. Consequently, this thesis additionally presents a method for pure vision-based ego-vehicle localization.

1.5. Outline

The remainder of this thesis is structured as follows. Chapter 2 presents a thorough overview of the related work. Because road terrain detection is the application focus of this thesis, the proposed concept for visuospatial analysis of the road environment will be outline in Chapter 3 for this particular application. In the first main chapter of this thesis, road terrain detection using local visual features is discussed (see Chapter 4). This reflects the first stage of the road terrain detection system from Chapter 3. Subsequently, Chapter 5 details the generic concept of visuospatial classification. In this chapter, the approach from Chapter 4 is extended by incorporating the spatial layout of local visual appearance. This approach will be applied to road terrain detection as well but has a wider field of application. This will be additionally shown in Chapter 6 which discusses a different kind of application for visuospatial classification. To this end, the method is applied to visual ego-vehicle lane assignment using the same method as in Chapter 5. Chapter 7 gives a comparison to other state-of-the-art methods for road area detection. Furthermore, future steps for advancing the approach that are beyond the scope of this thesis are detailed. Finally, this thesis will be summarized and concluded in Chapter 8. At the end of each chapter, one can find a brief summary, containing the quintessence of each subject.

2. Related Work

"Image understanding is one of the Holy Grail problems in computer vision. Understanding a scene arguably requires parsing the image into its constituent objects" (Malisiewicz and Efros, 2009, pg. 1222).

There is a lot of ongoing research in the field of computer vision-based scene understanding. For a vision system, such a holistic task is very challenging. For non-computer scientists, this is sometimes hard to understand, because humans are excellent in scene understanding. Even if they perceive a scene only for a short period in time, they can easily get the gist of it (see Oliva and Torralba, 2001).

The particular field interest of this thesis is scene understanding of traffic scenes. On a rather macroscopic level, several approaches for traffic scene understanding have been proposed. For inner-city driving, Ess et al. (2009) presented a recognition of the traffic scene's coarse geometrical category (i.e., usual road, curve, crossing) and if there is a vehicle ahead, or a pedestrian by using holistic image segmentation (see Fig. 2.1).

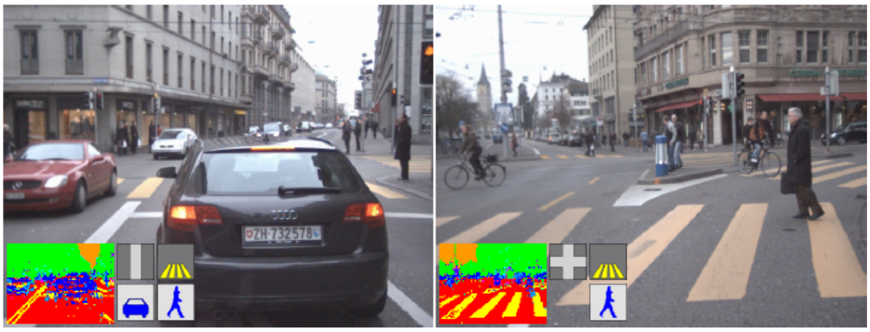

Figure 2.1.: Traffic scene understanding by classifying the scene category. Estimation is based on holistic segmentation of the image (bigger box on the lower left), inference of scene category is denoted by the darker and the existence of pedestrians and cars with the brighter small gray boxes. Pictures are extracted from Ess et al. (2009).

Road terrain detection is a particular problem of scene understanding for driving scenes. In contrast to holistic scene understanding, road terrain detection focuses on the ADAS-relevant driving space. Vision-based road terrain detection has been addressed in many papers in the last decade. The following Section 2.1 will give an overview of the main contributions. Section 2.2 presents approaches employing spatial features which are related to those proposed in this thesis. Finally, the related work chapter ends with a summary in Section 2.3.

2.1. Approaches for Road Terrain Detection

Strategies for detecting particular road terrain, such as the ego-lane or the complete road area, are split into three different groups (see Fig. 2.2). The most elaborated and frequently used approaches are delimiter-based, as presented in Section 2.1.1. Complementary to these, approaches using surface-based features for road terrain detection follow in Section 2.1.2. While typically those approaches are applied in a specific scenario using a fixed parameter set, the adaptation of internal parameters with respect to the current environmental conditions will be needed for ADAS that can handle multiple scenarios. Therefore, also adaptive approaches will be discussed in both Sections. Unlike most of the approaches that are rather bottom-up oriented, Section 2.1.3 details a special type of approaches using top-down scene context for road area detection.

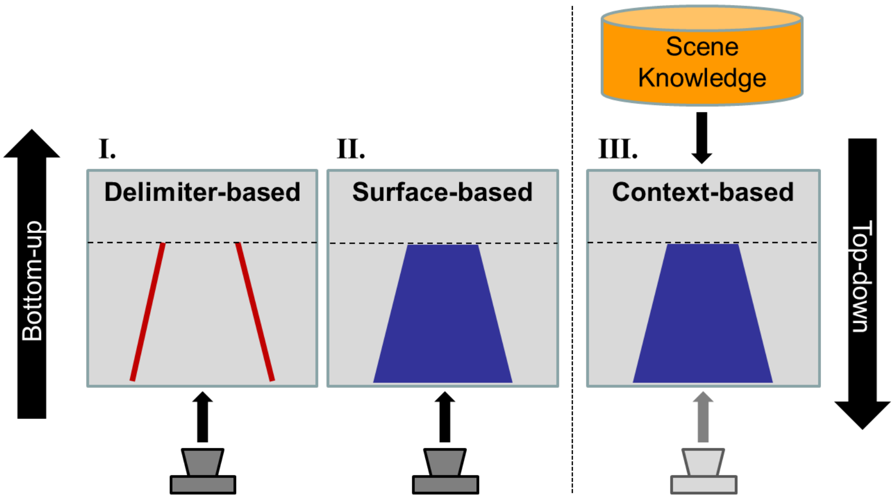

Figure 2.2.: Categorization of road terrain detection: Delimiter-based (I.) and surface-based (II.) methods are typically bottom-up approaches, another type of approaches are using top down scene context (III.).

2.1.1. Delimiter-based Road Terrain Detection

Most vision-based ego-lane detection methods follow the classical approach by Dickmanns and Zapp (1986) which is the basis, not only for delimiter-based driving space extraction but also for implementations in current commercial ADAS. After more than a decade of proposals, it is still an active field of research which shows that there are still unsolved issues.

The most obvious lane delimiter type in a driving scene are lane markings. By finding the closest lane marking to both sides, it can be assumed that the region in between is the ego-lane. Candidates for lane markings can be extracted by thresholding an image's intensity values or intensity gradients because usually white lane markings have a higher intensity value than surrounding road area.

A more advanced method, proposed by Gopalan et al. (2012), extracts lane marking candidates with a learning approach using visual inputs from a camera. A coarse overview of the approach is given in Figure 2.3. In detail, a hierarchical pixel-based feature descriptor is used to encode the spatial layout of the dark-light-dark transition of lane markings and the surrounding road area region. Then, a previously trained Boosting classifier is used to obtain lane marking candidates. Assuming the course of lane markings describe a second order polynomial curve, the generalized Hough transform (see Ballard, 1981) is used for grouping lane marking regions accordingly. This parametric description of the lane markings is then tracked over time with a *Particle Filter* (PF, see Gordon et al., 1993). This particular approach is suitable for tracking multiple lane markings, for each detected lane marking track one PF has to be used.

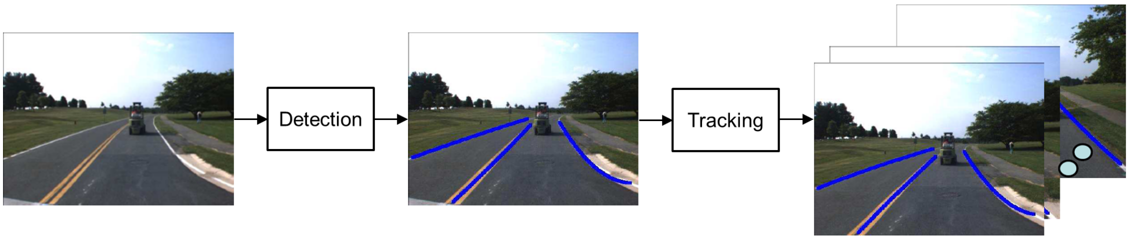

Figure 2.3.: Block diagram for lane delimiter extraction based on a lane marking detection and tracking. Pictures are extracted from Gopalan et al. (2012).

Basically all delimiter-based approaches assume that the course of delimiters follows a certain model. As these models have certain limitations, i.e., in terms of maximal curvature, this implies restrictions to the course of the road. A scenario where current lane markings based approaches fail are, e.g., road work zones on highways. This has mainly two reasons. Firstly, systems often use road models that restrict the road curvature to either left or right direction (e.g. clothoids) which sometimes does not hold in road work zones. Additionally, systems cannot cope with different types of lane markings, i.e., white and yellow lines, indicating validity. Therefore, Graf et al. (2012) proposes a probabilistic estimation of temporary lanes at road work zones. A road model which consist of two joined clothoids is used to capture the lane even in case of S-turns. The lane markings are extracted from grayscale images using an edge detection method. A *Joint Integrated Probabilistic Data Association* is used to find the most plausible associations of lane markings to lanes. This includes rules for German highway road works (e.g., width of lanes). Then, one PF per possible lane marking combination is used for tracking.

Again for the application in road work zones, another type of delimiters information can be extracted from elevated obstacles like barriers and guardrails. In order to estimate the road-course, Darms et al. (2010) fuse 3D-information from *Structure From Motion* using optical flow from a monocular camera and radar to obtain a 3D-boundary. Here the boundary is also modeled as a clothoid. Therefore, the approach reaches its limit in case of a more complex course of the delimiter. This would probably cause system failure due to mismatches between measurements and the clothoid model.

A lot of approaches for lane tracking assume that the number of lanes is known. Therefore, Deusch et al. (2012) presented their approach for multi-lane tracking (up to three lanes) on rural road without knowing the number of lanes. They propose an extension of the standard PF algorithm for tracking the ego-lane and optionally neighboring lanes. After lane marking candidates are extracted by a double gradient method an existence check is applied to infer whether neighboring lanes are existing which are then tracked with the proposed method.

For highways typically the curvature of the road is rather low. This makes it easier for systems operating only on highways because the range of possible parameters is also low. However, on rural roads more extreme road shapes can occur. Therefore, Shen and Ibrahim (2012) presented a model switching approach for road curvature estimation. Lane markings are extracted from a monocular image, for elevated road boundary objects stereo triangulation is used. The main contribution is a combination of lane models of first, second, and third order. A probabilistic switching of these models is realized using an so-called *Interacting Multiple Model*. This work implicitly shows that the application of higher complex models with more degrees of freedom is not necessarily leading to higher performance. We see, in simple scenarios complex models sometimes lead to unsatisfying results.

While early work only considered lane markings as delimiters, Danescu and Nedevschi (2011) also extract elevated curbstones as lane delimiter. The authors propose ego-lane detection based on a tracking of the course of ego-lane delimiter positions in structured urban scenarios. If a curbstone is sufficiently elevated, stereo vision can be used for detection. For lane marking candidates, a horizontal gradient method detects the dark-light-dark transition. Furthermore, a clothoid-shaped lane model is applied which models the course of the ego-lane. The authors show that using an enhanced road model for vertical and horizontal road course is beneficial. In the lane model the width of lane markings and the height offset for the vertical course are considered. The model parameters are tracked using a PF. In this context, the problem of changing pitch angles of the car while tracking is analyzed.

Especially in urban scenarios drawbacks of approaches applying clothoid lane models become clear. In complex situations model assumptions are likely to get violated (e.g., complex road shapes cause by parked cars) and therefore may lead to system failure.

A dedicated method for road area detection in the presence of elevated curbstones is proposed by Siegemund et al. (2011). The authors explicitly model the 3D-profile of the transitions between road and sidewalk to detect road and road adjacent regions. As features, the height over ground from a *Digital Elevation Map* is used. Assuming an elevated curbstone delimiting the road, a sigmoidal function is used for fitting the height profile of the road using a Conditional Random Field (CRF). The course of the curbstone is tracked over time using a *Kalman Filter*. This approach is a specific solution in situations where a curbstone is present and collateral to the driving direction.

Some authors propose detection of lane markings in the so-called *Birds Eye View* (BEV) of the image. Having a calibrated camera, the BEV can be obtained using *Inverse Perspective Mapping* (see Mallot et al., 1991). This is a nice representation because perspective changes of the road are compensated, and properties like parallelism of lane markings can easily be verified.

One of these authors, Liu et al. (2011) propose the detection of lanes within the BEV with a non-standard lane model for two-lane roads (see Fig. 2.4). They propose using a combined road model which is linear in near and parabolic in far range (see right image of Fig. 2.4). For extraction of lane marking candidates position, color and gradient information are employed. Furthermore, they propose a partitioned PF which has the property to recover from failure and maintain multiple hypothesis.

Lipski et al. (2008) propose a feature detection algorithm using local histograms

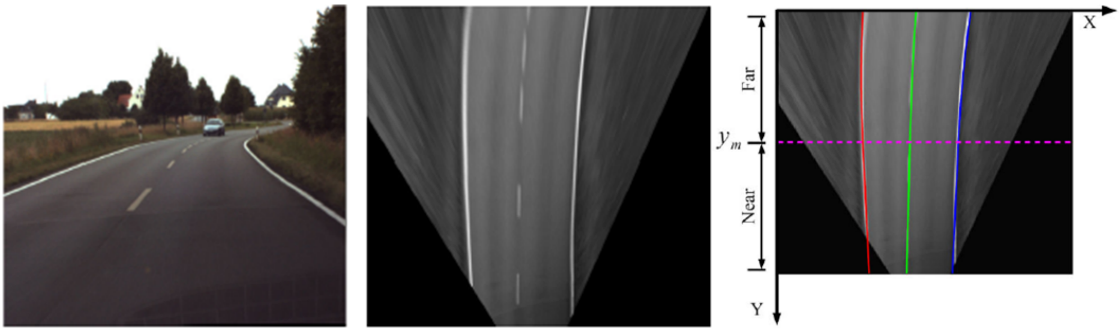

Figure 2.4.: Illustrated is the camera image (left) with the corresponding *Birds Eye View* (center) and the detected lanes (right). The lane model is separated in near and far range. Different lane shape models are applied for the two sectors. Pictures are extracted from Liu et al. (2011).

in the BEV of several input images to extract lane marking candidate positions. In order to fit these to a parametric representation, the authors propose a road model consisting of several connected lane segments which have a certain length and an orientation offset to their predecessor segment. Then, a fitting procedure is applied to map lane marking candidates to the model.

Another lane detector, also operating in the BEV, is proposed by Linarth and Angelopoulou (2011). Here, a *Histogram of Oriented Gradients* (HOG) is used as feature. A PF is used to temporally update the parameter of a second order lane model. Based on this, a feature expectation for lane markings is compared with the measured image features using the *Battacharyya distance* as a metric. The final lane delimiter is obtained by maximizing a weighting function including this metric and assuming a certain parallelism of neighboring markings.

2.1.2. Surface-based Road Terrain Detection

In contrast to many of the delimiter based approaches for ego-lane detection, presented in Section 2.1.1, road area detection is typically realized using properties of the road surface. Consequently, this Section is split in two parts: Firstly, approaches extracting physical properties from road terrain. Secondly, approaches capturing the visual appearance of road terrain.

Physical Features for Surface-based Road Terrain Detection

A well-known approach for surface-based detection of road terrain using physical features is based on correlation-based stereo vision. By simply assuming that the road area is flat, the height over ground can be used to extract unelevated road terrain (see Okutomi and Noguchi, 1998).

Following the same idea, "Stixels" (see Fig. 2.5) were proposed by Pfeiffer and Franke (2010) as efficient representation of traffic scenes, which can be obtained from stereo vision. Although, "Stixels" describe distance and height of an object in a certain image column (see Pfeiffer and Franke, 2011), it can also be used for free space detection. The base point where an obstacle touches the ground (see Fig. 2.5b) is obtained by separating each image column in its road-, obstacle-, and background-part (see Fig. 2.5a). Applied on the whole image this results in the vehicle's free space (see Fig. 2.5c) under the assumption that the road plane can be modeled with a B-spline.

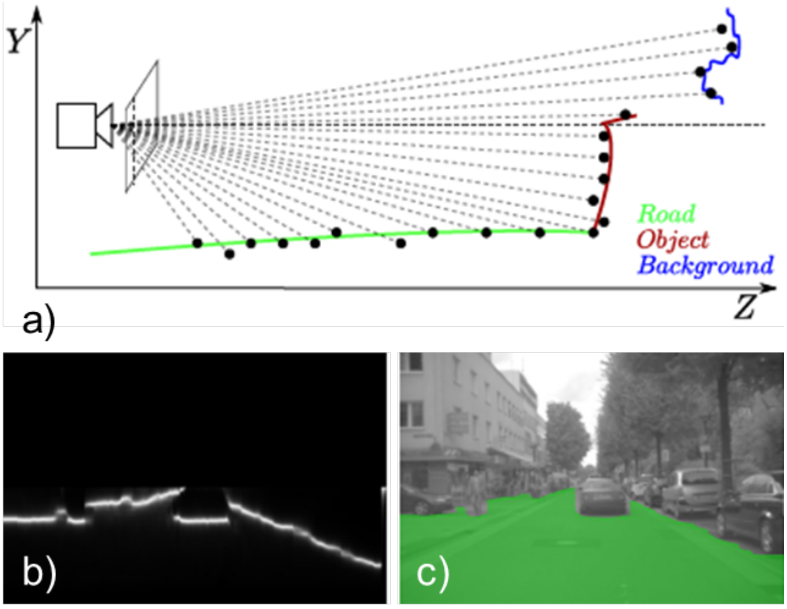

Figure 2.5.: "Stixels" can be used to extract unelevated areas in the field of view based on correlation-based stereo vision. Pictures are extracted from Pfeiffer and Franke (2010).

A closely related approach by Kang and Chung (2011) also proposes stereo-vision-based free space detection. Here a structural and traversability analysis is performed on a probabilistic volume polar grid map. This can be seen as an extension of the image column-based representation by Pfeiffer and Franke (2010).

While the latter two approaches modeled free space, objects and background as a whole, Schreier and Willert (2012) presented a dedicated free space extraction in a rural road scenario. Here, the flatness of road terrain is represented in an occupancy grid which is generated from an imaging radar sensor. In order to extract the actual road segment from the noisy data, well-known image processing methods like morphological operations and watershed segmentation are applied to the occupancy grid. The road boundary shape is represented by a dynamic B-spline which is tracked over time using an *Extended Kalman Filter*. This approach nicely shows how to gain a higher level road area representation from noisy data.

Adaptivity of a system means autonomously modifying internal system parameters to be better suited for changing environmental conditions. This is an important aspect for road terrain detection because ADAS as a product in consumer cars have to be flexible and work under many conditions (e.g., day and night). Guo et al. (2009) proposes adapting the relation of the cameras to the road plane combined with a recognition of image regions which correspond to unelevated road area. The authors propose an active adaptation of the camera-to-world transformation. Because of the cars' pitching and rolling behavior, the position and orientation of the camera relative to the fixed vehicle coordinate system is changing and therefore can be adapted to reduce transformation errors. The basic idea is to map corner points, extracted by a *Harris corner* detector (see Harris and Stephens, 1988) from stereo images into a homography using *Inverse Perspective Mapping*. Based on the distance of corresponding corner points in the homography, two issues can be considered. Firstly, optimizing a cost function, including all distances from corner point correspondences in the homography, leads to an updated camera-to-world transformation. Secondly, if a distance is low the point's pixel position is likely to be unelevated. Consequently, based on the assumption that a sufficient amount of corner points was detected and the road surface is planar, the road area segment can be extracted. It is an advantage that the approach does not require a calibrated stereo camera, only a sufficient initial transformation. The online adaption of the camera-to-world transformation makes this approach applicable, e.g., for slopy roads.

Using only physical properties of the road surface, the detection of ego-lane is not possible (see Section 1.3). Therefore, Loose and Franke (2010) proposed a 3D-lane recognition using a combination of stereo measurements, to model the 3D course of the road, with an optical flow based lane marking recognition from Gern et al. (2002). In this approach, the 3D lane course is modeled with a B-spline. The benefit of the proposed road model is visible in Fig. 2.6. A direct comparison of a planar clothoid Fig. 2.6b and the proposed 3D B-splines Fig. 2.6a shows the disadvantages (especially in far distances) of planar clothoid based models compared

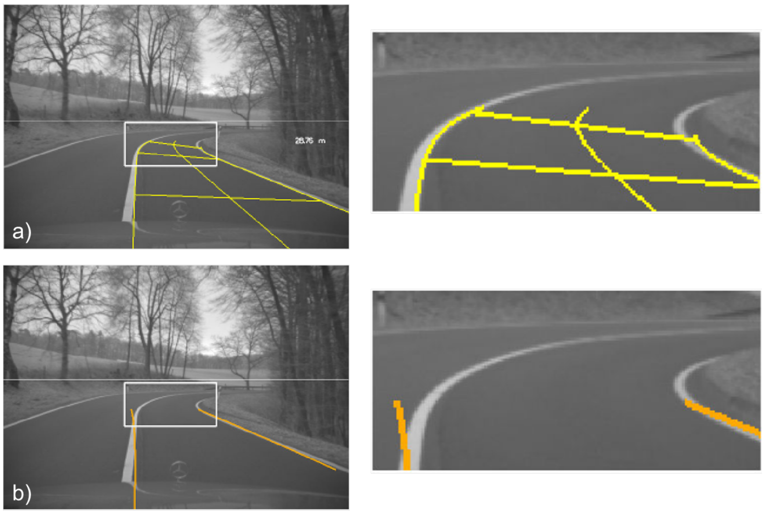

Figure 2.6.: Example showing the difference of the proposed 3D B-splines (part a) to the conventional planar model (part b) in an example scene. The zoomed image regions (right part) depicts the deviations of both models in far distances. Pictures are extracted from Loose and Franke (2010).

to the proposed model.

In case the usual lanes are not visible or blocked (e.g., due to an accident), convoy tracks, as presented by Weiherer et al. (2012), can be used as an alternative pathway for the ego-vehicle. In this work, occupied regions are represented in a 2D interval map, a map-based environment representation that is longitudinal discrete and lateral continuous. Consequently, the so-called convoy tracks (see Fig. 2.7) can be reconstructed from this representation. A convoy track describes locations that are or have been occupied by a group of succeeding vehicles. The convoy track can be seen as alternative option for maneuvering, in addition to the ego-lane.

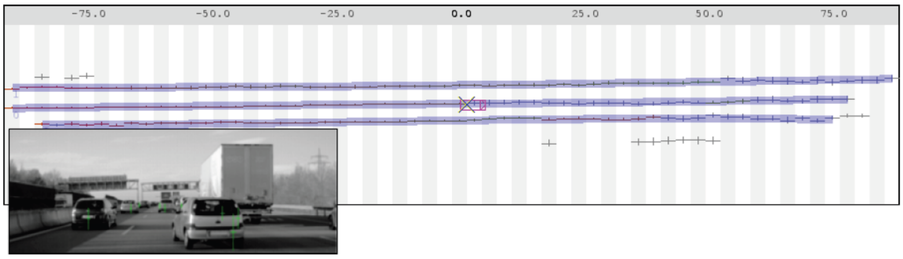

Figure 2.7.: Example showing three parallel convoy tracks on the highway. Picture is extracted from Weiherer et al. (2012).

Visual Features for Surface-based Road Terrain Detection

Delimiter-based approaches do not gain any knowledge about the free space status of the extracted ego-lane, i.e., if the driving path is actually not occupied by other road users or obstacles.

Therefore, Gumpp et al. (2011) estimate the non-occupied part of the lane based on visual surface features extracted from a monocular camera. First of all, an existing detection algorithm is used to extract the ego-lane (see left part of Fig. 2.8). The approach uses handcrafted appearance models for lane markings, gradients (sobel) and color which operate in metric space (see right part of Fig. 2.8). This model reflects a feature expectation, i.e., that the ego-lane is gray and untextured and has lane markings only in the border regions. By comparing the current measurement with the feature expectations, every single model outputs a confidence value which is then fused with the *Dempster Shafer* method. Based on the resulting confidences for the distinct metric coordinates an inference about the unoccupied lane length is applied.

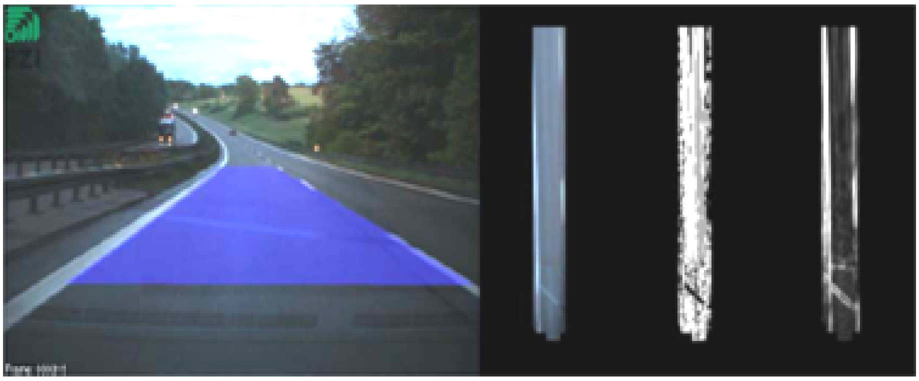

Figure 2.8.: Visual features extracted from the ego-lane surface (see blue region in left part) are analyzed in the *Birds Eye View* (right part). Picture is extracted from Gumpp et al. (2011).

For traffic scenarios in absence of lane markings, Franke et al. (2007) propose a driving space recognition fusing color, texture and edge cues. The system is applied on country roads which might not always have lane markings. Road hypotheses are generated from a maximum-a-posteriori function that optimizes parameters of a road model given an image sequence. The road model fuses delimiter-based as well as pixel-based road area information in a probabilistic fashion by having an internal representation for the road area and non-road regions. With that, the extracted road boundary is tracked using a PF.

Being adaptive to visual-changes in the scene is a highly desired property for any kind of road terrain detection. In this context, Michalke et al. (2009) presented an adaptive multi cue fusion for road area recognition. Training windows are used to adapt histogram-based internal road and non-road representations to the current visual appearance in the moment of capturing the image. Several cues, such as color and texture, are incorporated. The aim is to explore differences between road and non-road based on measurements from the training windows and start a segmentation process. This allows extracting the road area that is visually more similar to the area given by the road training window than the area in the non-road training window. Prerequisite for a good segmentation result is that the appearance in the road window area is representative for the whole road area. The problem of a directed segmentation is also that it might get stuck at a region with high gradient, such as a stop line or a transition between two different asphalt regions.

For capturing the visual appearance of road terrain from a camera-image, the contained pixel values have to be analyzed based on, e.g., their color. In order to get more information on an image position usually the local neighborhood of a single pixel, e.g., by using patches, is also examined (see, e.g., Sha et al., 2008). In contrast to isolated pixel-wise detection *Conditional Random Fields* (CRF) intend to improve the performance by modeling image-pixel neighborhood dependencies. This method can be used to identify multiple scene elements in the field of view, including the road area (see Sturgess et al., 2009). In contrast to prior CRF, Wojek and Schiele (2008) propose a dynamic CRF, which considers movement of the observer and objects in the scene which is obviously important for the automotive application. For this purpose multiple *Extended Kalman Filters* (one per object) are used.

CRFs are currently a very popular and powerful approach from the computer vision community. But the comparison of these rather holistic methods with dedicated road segmentation methods (e.g. Guo et al., 2009) shows that the concentration on the relevant road area results in a better classification performance.

Hoiem et al. (2007) proposes utilizing the so-called geometric context of the scene. It segments the image into superpixels and estimates a distribution over a set of discrete surface orientations for each superpixel. The method can be applied on a single image. Local appearance and context such as the location of the vanishing point is captured by using retinal position, color, texture and perspective features. Finally, the method generates a probability map for ground using a Boosting classifier. The aim of the approach is an applicability for arbitrary scenes (indoor and outdoor) by learning a generalizing model.

2.1.3. Inclusion of Top-down Scene Context

Complementary to the bottom-up detection approaches outlined before, top-down scene context as prior for the classification decision can be very helpful. Especially in situations where perception is tough, e.g, under adverse weather and lighting conditions, top-down knowledge like typical image-locations for the road area or knowledge from previous detections can help a lot.

Álvarez et al. (2013) proposes a switching of several road area models. The prototypical road models encode pixel locations of road regions (see Figure 2.9) for different geometrical road types. In particular the different models are: strong turns (left and right), soft turns (left and right), straight road, and additionally different traffic compositions (e.g., junctions or tunnels).

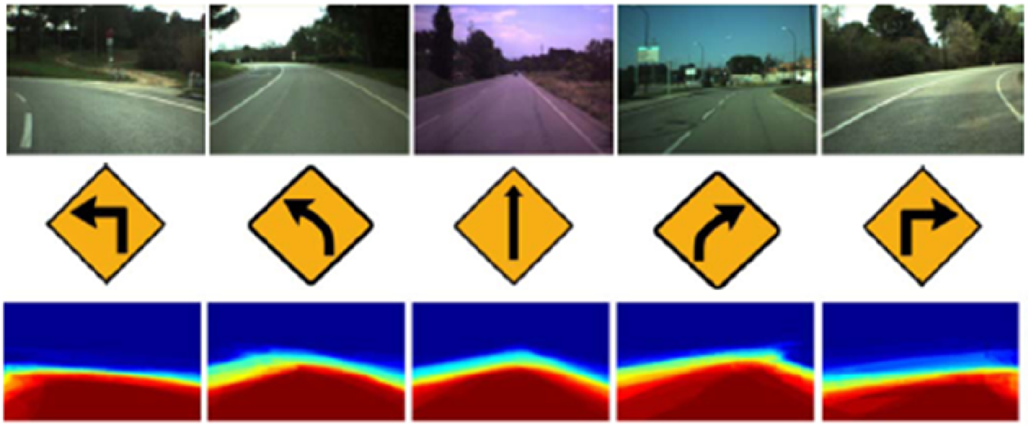

Figure 2.9.: Illustrated are different scene categories (top row): strong turns (left and right), soft turns (left and right), straight road. A probabilistic road model (bottom row) is generated for each scene category. Picture is extracted from Álvarez et al. (2013).

A particular road model corresponds to a specific scene category. By detecting the scene category, the corresponding model can be selected. Furthermore, the road location model is adjusted more specifically to the given scene by analyzing scene composition and temporal coherence. In order to be robust against illumination changes, the input camera image is transformed into an illumination normalized representation. Image statistics are aggregated by means of a spatial pyramid, containing subregions and different scales. In each subregion GIST is used to generate local features. By combining the local features of all levels a global image descriptor is obtained. The classification of the road model is carried out with a multi-class *Support Vector Machine* (SVM). Finally, temporal coherence of the road shape models is enforced by using a *Hidden Markov Model* (HMM).

The complete framework (see Fig. 2.10) does additionally contain a bottom-up road area detection. By combining top down and bottom-up detection this approach aims at increasing the robustness. For the bottom-up way, pixel based appearance information is used. Similar to Michalke et al. (2008) (approach already discussed in Section 2.1.2), an adaptive model based on a histogram extracted from a road region close to the ego-vehicle is used. Finally, both pathways are combined using a weighted harmonic mean.

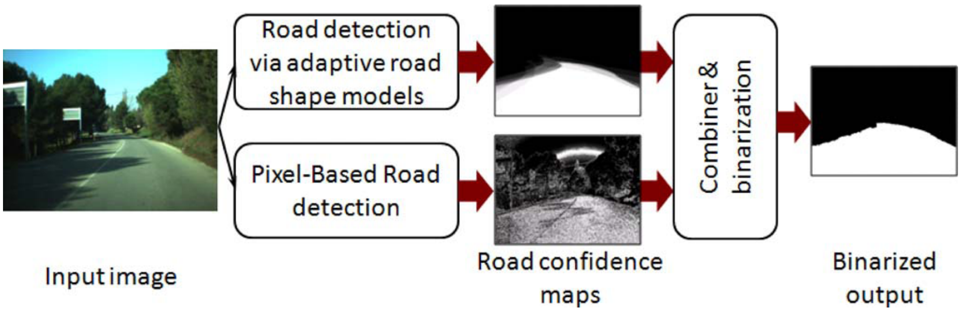
Road detection via adaptive road shape models
Pixel-Based Road detection
Combiner & binarization
Input image
Road confidence maps
Binarized output

Figure 2.10.: Blockdiagram showing the combination of top-down and bottom-up road detection. Picture is extracted from Álvarez et al. (2013).

Another source for top-down scene context can be obtained from the vanishing point. Kong et al. (2010) present a road detection from a single image. In this work, the road is modeled by two straight lines reaching the vanishing point. Based on texture orientations computed with a *Gabor filter bank*, several candidate positions for the vanishing point are extracted. Those are fused by applying an adaptive soft-voting scheme. Furthermore, a road boundary detection is used which is based on a most dominant edge detection. A set of edges oriented towards the vanishing point with fixed angles find the most suitable road boundary. The orientation with the highest score computed from orientation consistency and the color difference of left and right side of the edge is selected. All the previously mentioned approaches usually work only on more or less structured roads. In contrast, Kong et al. (2010) presented a very general method for driving space detection for arbitrary scenarios. Therefore, this method has advantages in environments where a high generalization is required, e.g., on off-road tracks. But in rather structured scenarios like asphalt roads more dedicated methods are anticipated to have better performance.

In a very related work, Wu et al. (2011) estimate the image position of the vanishing point based on an example database. For extracting features a *Gabor filter bank* (as in Kong et al., 2010) is used. From the database, the six nearest neighbors in feature space are extracted. Based on the labeled vanishing points in the database samples, a weighted position of the estimated vanishing point can be obtained.

Furthermore, map data (e.g., from a navigation system) can deliver useful scene knowledge. For an ADAS on rural roads it is very important to have geometrical knowledge of the road ahead. In order to reduce the error of visual road course estimation, Gackstatter et al. (2010) propose a combination of bottom-up and a top-down road course estimation. Firstly, the road course is estimated based on detected lane markings using a clothoid model. Secondly, information of the road ahead is extracted from map data. The authors show that the proposed combination can significantly reduce road course estimation errors.

2.2. Approaches employing Spatial Features

In contrast to approaches in previous sections where road terrain detection was the field of research, the following works have in common that they propose a special type of feature computation which is not locally restricted. Approaches presented in this section utilize spatial properties for classification and are therefore related to the spatial features applied in this thesis.

Kang et al. (2011) present a system for semantic segmentation and object recognition in road scenes using a hierarchical texture representation. In total eight classes are considered: road area, lane marking, sky, tree, car, pole, sidewalk and building. The set-up includes an expensive RGB and near-infrared multi band prism-based camera, in order to compute additional infrared based features. The multiband image is convolved with a filter bank in order to gain a representation of the image. Subsequently, the results are clustered using a standard k-means clustering approach in order to obtain a Texton representation (see Julesz, 1981). A spatial relation is then finally created by putting histograms of the Textons in a hierarchical representation using a bag-of-Textons approach (a modified version of the bag-of-words approach by Csurka et al., 2004; Lazebnik et al., 2006). This is depicted in Fig. 2.11. The grid windows S_1, S_2 and S_3 have different scales (S_3 is the largest) and are combined in the feature descriptor in a hierarchical manner. Lastly, the results are classified using a multi class Boosting approach, which classifies features on the basis of the localized frequency of Textons.

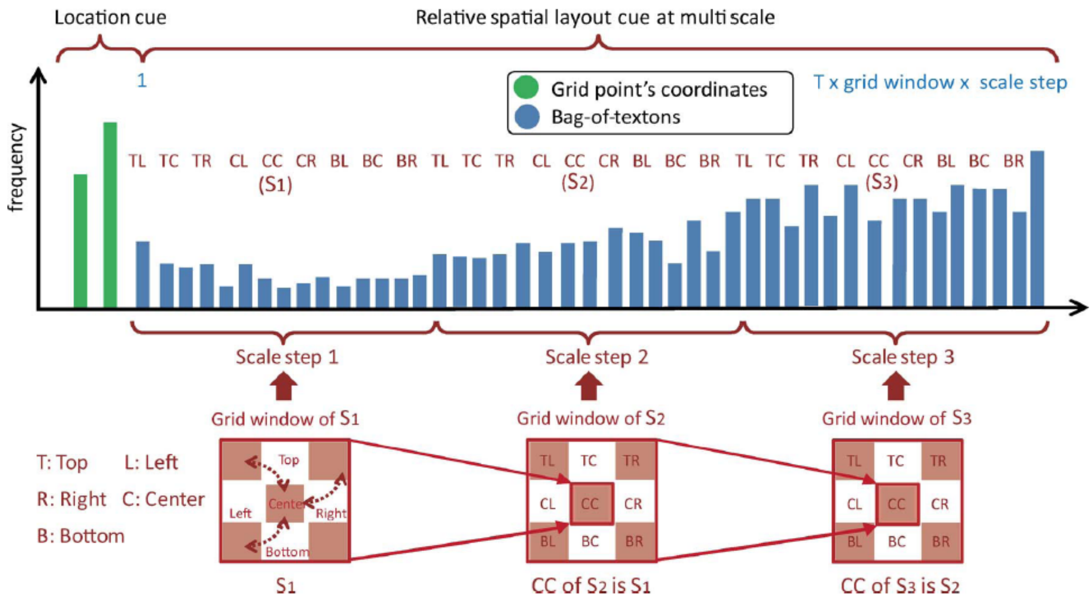

Figure 2.11.: Histogram of hierarchical bag-of-textons with multiscale grid windows. Picture is extracted from Kang et al. (2011).

As the approach by Kang et al. (2011) includes road area as one of the target classes it is comparable to the approaches discussed in Section 2.1.2. In contrast to approaches employing explicit road models (see, e.g., Loose and Franke, 2010), this approach comprises retinal position as a feature for each class. Therefore, the classifier implicit learns a retinal position model for the road area. However, using retinal position as a feature requires a very large and balanced dataset including all variations of locations for all the particular objects. Otherwise, this results in overfitting for object classes at the training locations. For instance, poles can only be detected when they are located in an image region that contained poles in the training dataset.

Álvarez et al. (2012) propose a fusion of two methods, a generic offline classifier and an adaptive online classifier. The first classifier is based on Convolutional Neuronal Networks CNN. The CNN is learned offline in an iterative manner to generate invariant feature descriptors. The resulting feature descriptors on different scales can be seen as a combination of visual and spatial information and is therefore similar to the approach followed in this thesis. Finally, an image can be classified into three classes (sky, vertical areas and horizontal areas) using a sliding window approach. As the approach is applied to road scenes, horizontal areas mostly correspond to the road area. The method needs ground truth labels for training. These are obtained using the generic method by Hoiem et al. (2007) already mentioned above. For the online classifier a method is proposed that maximizes the color uniformity of the road area on basis of a superpixel segmentation. In every timestep, the current appearance of the road area is captured from a static training window (lower-bottom part in the image). With respect to this reference for road area, a confidence is assigned to each superpixel region. Finally, both classification results (online and offline) are combined using a Naive Bayes approach. In Chapter 7, this approach will be compared with the road terrain detection system proposed in this thesis.

The approach by Shotton et al. (2008) uses a Texton representation for non-automotive image categorization and segmentation. Instead of using a filter bank for low-level image representation and k-means clustering for Texton generation (as, e.g., in Kang et al., 2008), the authors propose using a modified version of *Random Decision Forests* (RDF) that act directly on image pixels and implicitly represents semantic knowledge in the decision hierarchy. Compared to the above mentioned conventional Texton generation this can be computed more efficiently.

Point relational spatial features are useful for encoding an object's spatial alignment by comparing features at relative locations to base points located on the object. Shotton et al. (2011) present an approach using this type of spatial features in order to detect human body parts based on 2D-depth imagery. The used features are simple depth comparison features, which compare depth values at multiple locations with a relative displacement (i.e., angle and distance) to a base point. The extracted features are rather weak individually but perform very well in multiple combinations with a voting-based RDF classifier trained on the huge amount of synthetic data.

Ray-based approaches use RGB imagery in order to calculate characteristics alongside or at the end of rays cast through the 2D image plane. Smith et al. (2009) present an application in the area of automated cell microscopic imagery analysis for shape based classification. Following, in Lucchi et al. (2010) a fully integrated system is presented which is able to segment irregularly shaped cellular structures in EM images using these features. They introduce *Fast Ray Features* which essentially consider image characteristics at distant contour points. First, a gradient image is calculated from the original RGB image. Then, for any given point in the gradient image, four different features (see Fig. 2.12) are selected which test properties of the nearest edge location in certain direction. They differentiate between the difference, distance, orientation and norm of a distant contour to a base point, hence capture the spatial layout of a base points surrounding shape. These features show robustness against scaling and deformation, which is an improvement to the in this context usually employed locally restricted patch-based features (e.g., *Haar-like features* as used for face detection by Viola and Jones, 2001).

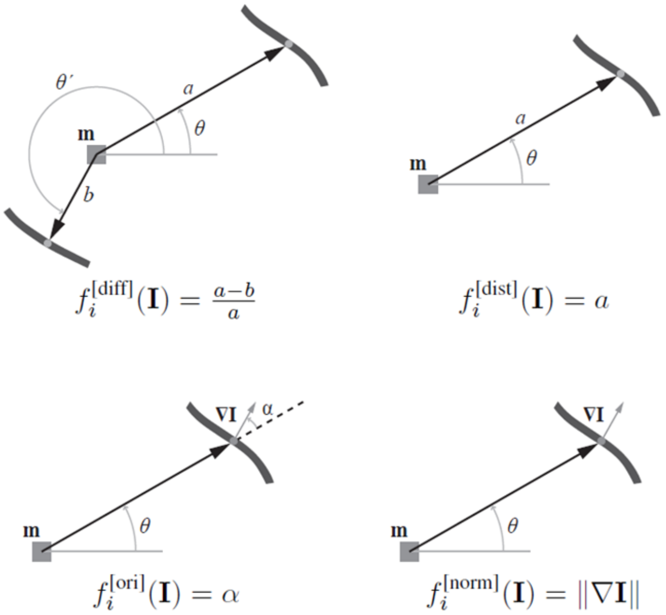
θ′
a
m
θ
b
$f_i^{[\text{diff}]}(\mathbf{I}) = \frac{a-b}{a}$
a
m
θ
$f_i^{[\text{dist}]}(\mathbf{I}) = a$
∇I
α
m
θ
$f_i^{[\text{ori}]}(\mathbf{I}) = \alpha$
∇I
m
θ
$f_i^{[\text{norm}]}(\mathbf{I}) = \|\nabla\mathbf{I}\|$

Figure 2.12.: Ray-like features are casted through the 2D image plane in order to classify differently-shaped cells. The four illustrations represent different type of features: difference, distance, orientation and norm. Picture is extracted from Smith et al. (2009).

In further research, Kumar et al. (2010) proposed *Radon-like features* which aggregate image statistics into compact feature descriptors. The approach also relies on casting two dimensional line segments across the image. But instead of extracting certain key features from the rays (as in Smith et al., 2009), *Radon-like features* are a generic spatial image descriptor, represented by a histogram computed with the Radon transform. The transform is applied extrinsically for every single image position with different angles and therefore orientations in the image plane. Consequently, the descriptor contains the integral value for all different angles for every image position which can be interpreted as the extrinsic spatial layout of the whole image. When observed from the outside, extracted features allow to distinguish enclosed areas very well.

2.3. Summary

Road terrain detection is an essential part of traffic scene understanding. This chapter gave an thorough overview of approaches detecting different kinds of road terrain.

One direction taken by a large number of researchers is the perception of road delimiting elements (e.g., curbstones and lane markings) for the detection of actual driving space. Features for these models are extracted from longitudinal road structures like lane markings or road boundary obstacles (e.g., curbstones and road barriers) by visual processing. This is mainly based on color and edge appearance, 3D information from stereo processing or *Structure From Motion*. From the extracted features, the road or lane shape can be tracked using different road shape models. However, especially for urban areas the applicability of these approaches is limited because road delimiters cannot be detected. Reasons are for instance missing or bad lane markings, parked cars occluding curbstones or very low curbstones. Furthermore, the typically applied models assume a clothoid-shaped road, and will fail in complex situations that can occur on rural roads and in urban areas.

Visual properties of the road area have also been used for estimating the road shape or for identifying the complete road area. Instead of delimiter-based road detection, characteristics of the road surface can be incorporated in the detection process. Besides the physical property of the road flatness extractable from stereo images, also visual properties like, e.g., the mainly gray and untextured asphalt region, can be used. Some approaches gain high generalization because they are adaptive to the current geometrical or visual conditions in the scene. Although, this is a desired feature for road area detection in unseen scenarios, adapting internal representations must be controlled carefully to avoid divergence.

Complementary to bottom-up detection approaches, also top-down scene context can be included as prior to further enhance robustness of bottom-up classification decisions. This context can be extracted directly from image data like, e.g., the vanishing point, the location of horizon line, and the scene category. Additionally, external information sources like map data or other sensory data can be incorporated. Top-down scene context is efficient in situations where the visual input conditions are challenging, e.g., under adverse weather conditions.

However, classifying the visual appearance on a local scale only, can lead to ambiguities. Spatial features refer to information gathering strategies that are not restricted to a local surrounding. In contrast to local feature gathering (e.g. inside a window with a fixed size), spatial features can consequently combine feature information from multiple spatial locations and capture the extent of regions and shapes. For tasks where object shapes or constellations play an important role, it was shown that using spatial features, such as rays casted from a base point through the 2D image plane, or point relational features, enhances detection quality compared to locally restricted methods.

3. Road Terrain Detection System

As discussed in the first two chapters of this thesis, road terrain detection is an important task for future ADAS. Beyond highways and other well structured roads, traffic scenes can be highly complex. The related work chapter pointed out the limitations of state-of-the-art methods in this context (see Chapter 2). Aim of this chapter is to give an overview of the proposed road terrain detection approach and detail differences to related work.

The thesis at hand details an approach that aims at improving appearance-based classification by incorporating the spatial layout of the road environment. Before we go into detail of the approach, a constructed example serves to recall different approaches for road terrain detection and their pros and cons: In the images depicted in Figure 3.1, four different concepts (I-IV) for road terrain detection are shown. Image I depicts an example result of a surface-based approach using the physical property of height over ground, e.g., the approach by Pfeiffer and Franke (2010) (cf. Section 2.1.2). The detection result is the complete unelevated road terrain, or the obstacle free space. However, for future ADAS road terrain categories

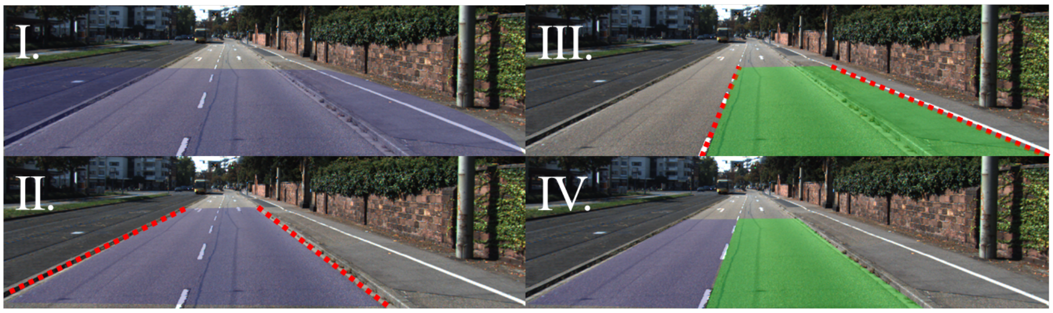

Figure 3.1.: Road terrain detection in urban areas for road area (blue) and ego-lane (green). (I) depicts an example result of a surface-based approach using the height over ground as feature. In (II), a curbstone-based road area detection approach is demonstrated. (III) represents the result of a lane marking based ego-lane extraction. In (IV), the desired result with a separation of the ego-lane from the road area is depicted.

reflecting knowledge about where the vehicle will go (i.e., ego-lane) or where it could go (i.e., road area) is more advantageous (see Chapter 1). For extracting the actual road area as in image II, we see that the delimiters, especially curbstones, have to be detected. However, in this scenario approaches like, e.g., Siegemund et al. (2011) might fail because the curbstones are hardly elevated. A lot of approaches for ego-lane detection solely rely on lane markings (cf. Section 2.1.1). In this case a false positive detection for the left part of the sidewalk, as depicted in image III, is expectable. Only few approaches mentioned in Section 2.1 would conceptually be applicable to detect the actual ego-lane in the given scenario (e.g., Danescu and Nedevschi, 2011; Franke et al., 2007). However, the curbstone detection from Danescu and Nedevschi (2011) would probably be inappropriate for the flat curbstone as well. The example in image IV illustrates the outcome of the proposed system trained on road area and on ego-lane. To the authors knowledge there is no other existing approach detecting road area and ego-lane with the same framework. Another advantage of the proposed system is that it can be trained to handle arbitrary delimiter types, e.g., lane markings, curbstones, and non-explicit delimiters like transitions from the road area to other vehicles, grass, and so forth.

Towards this goal, the proposed Road Terrain Detection System (RTDS) captures visual properties of the road surface, the boundary, and lane marking elements based on analyzing local visual appearance. From this basis, SPatial RAY (SPRAY) features that incorporate spatial properties of the road environment are calculated. Therefore, the proposed approach generates a visuospatial representation of driving scenes as the SPRAY features represent both local visual properties and their spatial layout. Subsequently, a decision can be taken by applying a classifier trained for a specific road terrain category. For the example in image IV, one would need two classifiers, one for road area, and one for ego-lane.

Considering the spatial layout of visual properties helps for any classification task where there is a clear correspondence between visual properties at different spatial locations. This differs from classical segmentation approaches which explicitly model the visual neighborhood in the image space, e.g., by using conditional random fields (see, e.g., Wojek and Schiele, 2008). The main novelties of the presented approach are:

- Proposal of spatial ray features for analyzing spatial properties in metric space and over arbitrary distances.

- Capability to handle different types of road terrain by separating the classification of appearance from the spatial classification task.

- Trainable approach that can be optimized for different kinds of road terrain independent from the road delimiter type.

The remaining part of this chapter is organized as follows. Section 3.1 will give an overview of the proposed RTDS. This chapter ends with a summary in Section 3.2.

3.1. System Overview

The vision-based RTDS is depicted in Figure 3.2.

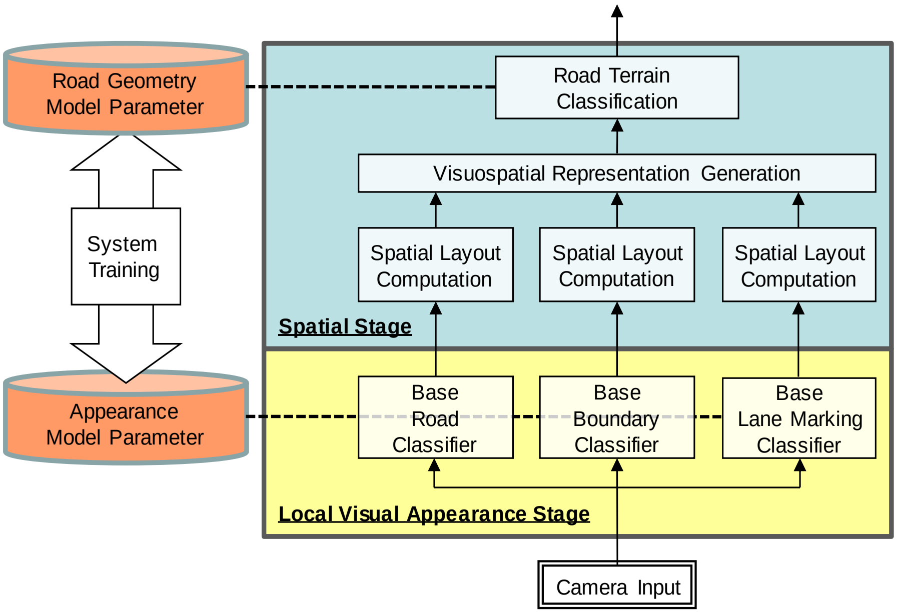
Road Geometry Model Parameter
Road Terrain Classification
Visuospatial Representation Generation
System Training
Spatial Layout Computation
Spatial Layout Computation
Spatial Layout Computation
Spatial Stage
Appearance Model Parameter
Base Road Classifier
Base Boundary Classifier
Base Lane Marking Classifier
Local Visual Appearance Stage
Camera Input

Figure 3.2.: Block diagram of the two system stages (from bottom to top): local visual appearance (yellow) and spatial stage (blue). Model parameters (orange boxes, left) are obtained during system training.

The two stages of the system reflect different levels of visual scene representation. On the lower level, local visual appearance in the image is represented by means of three base classifiers: base road, base boundary and base lane marking. On the higher level, based on the output of the first stage, the spatial arrangement of visual appearance properties is represented.

Images are handled instantaneously and independently, i.e., without considering temporal dependencies. In contrast to systems incorporating top-down scene context (see Section 2.1.3), the proposed system is purely bottom-up. The system parameters are obtained using machine learning. Both stages can be trained separately using ground truth annotations.

Subsequently, the two system stages of RTDS will be discussed. The representation of local visual appearance is detailed in Section 3.1.1, the generation of a visuospatial representation follows in Section 3.1.2.

3.1.1. Representation of Local Visual Appearance

The first stage of the block diagram depicted in Figure 3.2 is the so-called base classification. A particular base classifier can be seen as an appearance model for a certain visual property. Inside a base classifier the local visual appearance, such as color and texture, of an image region is extracted by using patches (sliding window approach). Subsequently, multiple appearance characteristics are compared to references. Based on the similarity of extracted characteristics and references, a confidence is assigned to every image position.

The first stage of the proposed road terrain detection system contains three base classifiers:

- Base road: An appearance model for road area.
- Base boundary: An appearance model for road boundary.
- Base lane marking: An appearance model for lane marking candidates.

These confidences then state whether the corresponding image region matches the given appearance model or not. The collectivity of base classifier outputs construct a local visual image representation in form of multiple confidence maps (see example in Fig. 3.3).

The base road and base boundary classifiers will be discussed in Chapter 4. Conceptually, the base classifier for road appearance is closely related to surface-based road terrain detection approaches detailed in Section 2.1.2. For the base lane marking classifier a different approach is applied. Similar to state-of-the-art methods (e.g., Gopalan et al., 2012; Veit et al., 2008), candidates for lane markings are extracted based on a dark-light-dark gradient (see Section 2.1.1).

Further processing is carried out in a metric image representation, the so-called *Birds Eye View* (BEV, see Mallot et al., 1991). The BEV is an image transformation which manipulates the location of image pixels as if the scene would have been

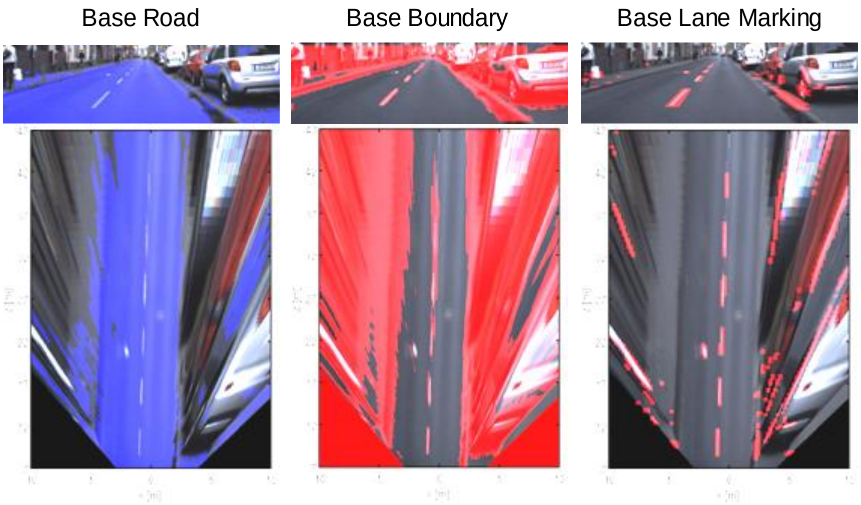

Figure 3.3.: Output of the base classification: Confidence maps for base road, base boundary and base lane marking. Results are shown in the perspective image (top) and corresponding *Birds Eye View* (bottom).

observed from above[1] (from the eyes of a bird). This has the advantage that perspective changes of flat surfaces on the ground, such as the road, are compensated. This can be seen in Figure 3.4, for any distance to the ego-vehicle the lateral extent of the road area and the ego-lane stays approximately constant. Furthermore, it can be seen as transformation of the perspective image into a metric representation because in the BEV a pixel directly correspond to a specific metric location (x, z) (see crosses in Fig. 3.4). However, this requires knowing the mounting position of the camera, i.e., the translation and rotation of the camera to the origin of the vehicle coordinate system. Throughout this thesis, the origin of the vehicle coordinate system is defined as the center of the rear axle projected onto the ground (see Fig. 3.4). More details on the BEV can be found in Appendix A.

3.1.2. Inclusion of the Spatial Layout of Visual Appearance

Aim of the second stage of the RTDS is to incorporate the spatial layout of the given appearance properties and therefore encode both: visual and spatial characteristics.

Thus, the presented approach intends to overcome limitations of state-of-the-art delimiter-based road terrain classification. The base cues (see above) represent delimiter information, i.e., lane markings and curbstones (base boundary), like

[1] Note that the underling assumption is that each pixel corresponds to a ground location.

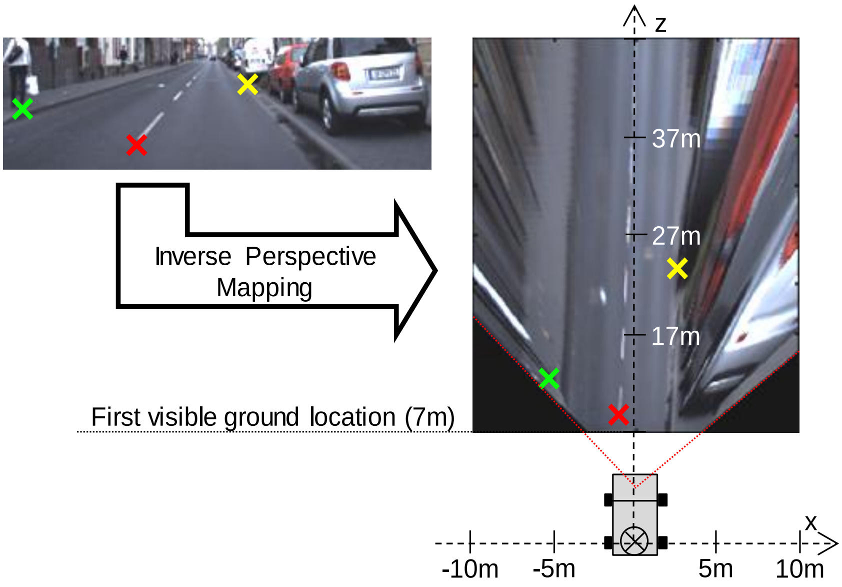
Inverse Perspective Mapping
First visible ground location (7m)
z
37m
27m
17m
x
-10m
-5m
5m
10m

Figure 3.4.: Transformation from perspective image into metric BEV using Inverse Perspective Mapping. Image pixels (see red, green, and yellow, crosses) correspond to a metric location in the coordinate system (x, z).

most of the approaches addressed in Chapter 2. Beyond this, another base cue captures the characteristics of the road area. If road delimiters are occluded or have bad quality, visual features from the actual road surface are expected to be more relevant as information from delimiters. The combination of visual and spatial characteristics of road terrain (including delimiter information) is anticipated to be more robust than relying solely on local visual features.

Each of the base classifiers delivers a metric confidence map as depicted in Figure 3.3. In order to capture spatial constellations, so called *SPatial RAY* (SPRAY) features are computed on every metric confidence map. SPRAY features encode the existence of a given visual property in a certain angular orientation with a certain distance relative to a specific metric location named base point. In contrast to the approaches presented in Section 2.2, SPRAY features are more suited to encode spatial relations in confidence maps. By combining SPRAY features from the different appearance cues corresponding to the same base point, a visuospatial representation is obtained.

The visuospatial representation has a wide range of application. This will be shown by training a road terrain classifier for road area and ego-lane as mentioned earlier in this chapter (see image IV of Fig. 3.1). In Figure 3.5 the result of the RTDS is depicted, more details will follow in Chapter 5. As discussed in the introduction, especially ego-lane detection is a challenging task for a visual appearance system, because the category ”ego-lane” is visually not discriminable from other regions on the road area. As pointed out in the related work, that is the reason why many approaches for ego-lane detection are purely delimiter based.

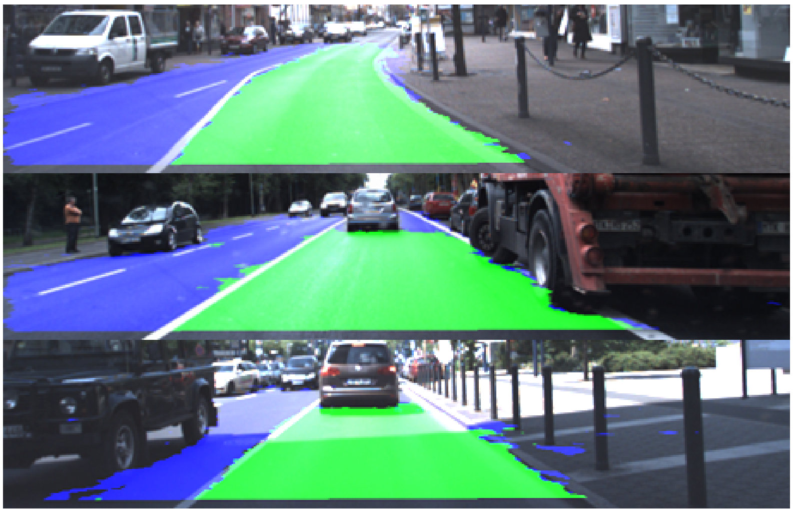

Figure 3.5.: Exemplary inner-city scenes of the road terrain detection with marked ego-lane (green) and road area (blue). In the visualization, the blue road area is occluded by the green ego-lane.

In most of the related approaches (see Section 2.1) explicit road models in combination with filtering approaches, e.g., *Particle Filter* or *Kalman Filter*, are used which generally improve the mean performance in structured scenarios. However, in more complex environments the applied models lead to restrictions in applicability. In the proposed RTDS, no explicit road model is applied. By using machine learning, the system can learn complex relations of visual and spatial features in order to classify road terrain. This enables the application of the approach to multiple scenarios and road terrain categories.

The system can be seen as a generic framework that can be applied to various tasks. As mentioned above, combining the visuospatial representation with a road terrain classifier can be used to identify the road terrain category for every image position. This will be shown for road area and ego-lane in Chapter 5. Furthermore, this representation can be used to infer spatial scene characteristics. An example for this is detailed in Chapter 6. Here, an application for visual ego-vehicle localization is discussed. Based on the SPRAY features computed from a fixed base point, the number of lanes left and respectively right beside the ego-lane can be estimated.

3.2. Summary

This chapter introduced the Road Terrain Detection System (RTDS) and highlighted differences of the approach compared to related work (previously presented in Chapter 2). The RTDS is a two stage process. Firstly, local visual appearance of the road area and its delimiters is captured by means of base classifiers. In Chapter 4 a more in-depth discussion of appearance-based road and boundary classification will follow. The second stage of RTDS combines the visual information with spatial features in order to construct a visuospatial representation. The proposed SPRAY features capture geometric characteristics of road environments. Combining visual appearance and spatial layout helps to compensate errors, e.g., in a situation where the visual properties do not indicate the presence of road, but incorporating the spatial layout making the detection possible. Through applying machine learning techniques for training the classifiers, the approach can be tuned for extracting different road terrain types from road scenes. For the most relevant road terrain types, i.e., road area and ego-lane, dedicated RTDS will be discussed in Chapter 5.

Beyond that, the generic system can be used for tasks different from road terrain detection. This will be shown in Chapter 6 where the visuospatial representation from the proposed system is applied for ego-vehicle localization.

4. Road Terrain Detection using Local Visual Features

This chapter deals with road terrain detection based only on visual characteristics of road terrain. In correlation to Chapter 3, in this chapter an elementary part on the road terrain detection system, the base classification is discussed (see Section 3.1.1). The base classifiers are the basis of the whole approach because they operate on the lower level, i.e., directly on the camera images. In particular, the approach learns a model for Local Visual Appearance (LVA) with respect to a certain type of road terrain which is then applied for detection (see Kuehnl et al., 2011). Because the method comprises features extracted from the road area, it is closely related to approaches mentioned in Section 2.1.2.

The remaining part of this chapter is structured as follows: The next Section 4.1 gives a system overview of the LVA-system. Patch-based features for representing LVA are detailed in Section 4.2. Accordingly, Section 4.3 explains how LVA-features are classified using Boosting. Classifiers will be trained for two purposes: for road area and for road boundary detection. Thus, Section 4.4 details how dedicated training data can be generated. Subsequently, conducted experiments on local visual appearance based classification follow in Section 4.5. This chapter ends with a summary in Section 4.6.

4.1. System for Local Visual Appearance Classification

The LVA-system processes camera images in order to extract local visual appearance information. Here, local refers to a local neighborhood which indicated that the approach extracts visual appearance features based on a combination of an image region's pixel information.

As input the LVA-system uses RGB images from an in-car mounted monocular camera (see Fig. 4.1). More than the half of the image, as we can see in Figure 4.1, is irrelevant for road terrain detection. Therefore, it is reasonable that a region of interest is selected before applying the method. In the example, a vertical cropping to remove the ego-vehicle hood and the sky decreases the total amount of input data that has to be processed by the system.

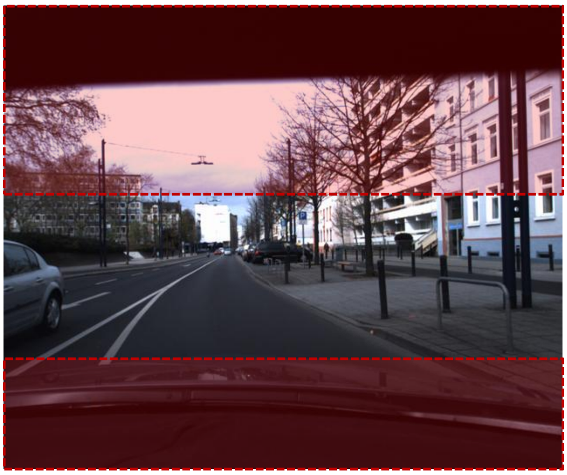

Figure 4.1.: Input image of the LVA-system captured by an in-car mounted camera. The colored part denotes regions that are cropped before processing.

The processing of the LVA-system can be split into three major parts as depicted in Figure 4.2. First of all, a patch extraction is applied on the cropped input image, where subsequently, LVA-features are computed on each patch. In particular, a regular grid with a constant step size s_{grid} is used to define the position where patches with size a_P are extracted. Let vector $c_i = [u_i, v_i]$ denote a particular grid location with linear index i. Then, c_i is the center of the i-th patch that is extracted. Subsequently, three different LVA-features, color $y_c(u_i, v_i)$, Walsh Hadamard $y_{\text{WH}}(u_i, v_i)$, and slow features $y_{\text{SFA}}(u_i, v_i)$ are computed. By merging these vectors we obtain a LVA-feature vector $x_{\text{LVA}}(u_i, v_i)$ for each center c_i. In the final step of the process (see Section 4.3), all LVA-feature vectors x_{LVA} are classified and the result is mapped back onto the image plane. Because patches were not extracted for every image position (the image was sampled using s_{grid}) bi-linear interpolation is applied to obtain an output with the same size as the input.

The whole LVA-system basically reflects an appearance model, i.e., the classifier has the ability to compare appearance characteristics (reflected by the LVA-features) with learned references. In order to learn these references for a certain category, e.g., road area or road boundary, supervised learning is applied. Based on the similarity of the extracted features with the learned reference a confidence

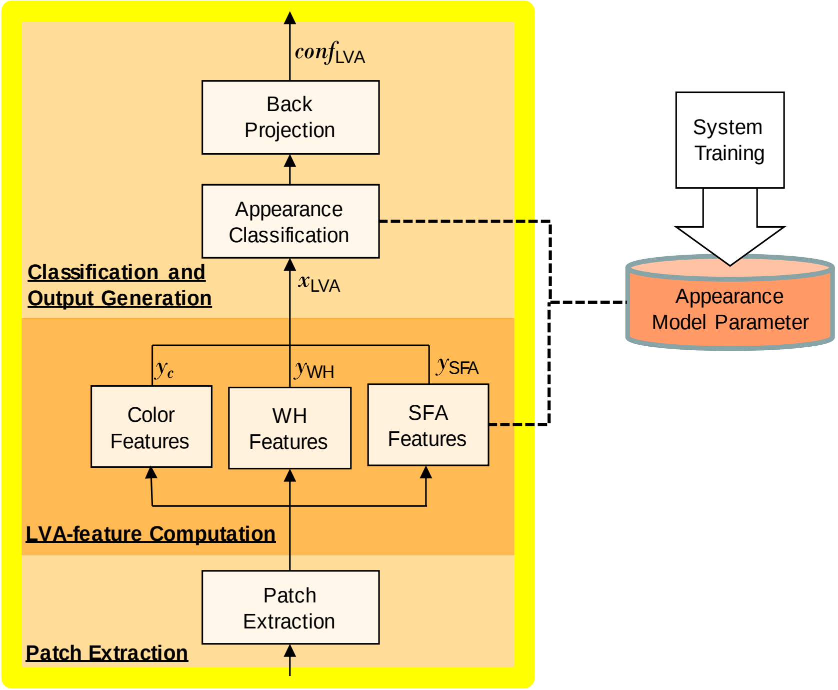

Figure 4.2.: Block diagram depicting the main steps of the LVA-system: Patch extraction, LVA-feature computation, and classification/output generation.

is computed (see Section 4.3). Because the procedure is applied on many image locations, the system output is a confidence map $conf_{\mathrm{LVA}}(u, v)$, indicating for each pixel position (u, v) whether it is likely to belong to a given road terrain category or not.

Two system components require explicit training (see right part of Fig. 4.2). The module computing the SFA-features (see Section 4.2.3) and the GentleBoost classifier (see Section 4.3) have to be trained offline once. Afterwards, the system parameters are static and input images are processed with the learned parameters.

The LVA-system is a pure bottom-up approach for road terrain classification. This means that no prior information is given, e.g., by explicit (see, e.g., Loose and Franke, 2010) or implicit (see Kang et al., 2011) road models. Furthermore, no temporal dependencies such as knowledge from previous detections (see, e.g., Danescu and Nedevschi, 2011) are considered.

4.2. Extraction of Local Visual Appearance using Patch-based Features

For extracting appearance characteristics from an image pixel's local neighborhood, a patch extraction method and subsequent LVA-feature computation is used. A patch, defined as a rectangular window which is centered around a certain pixel position (u_i, v_i), contains the information of the three color channels (RGB).

As mentioned above, a regular grid to define positions where patches are extracted and LVA-feature are computed is used in order to save computational resources. It is expected that using patch-based features is less sensitive to noise and has a higher performance compared to directly computing on the RGB-color values (as, e.g., Michalke et al., 2008). In this work, appearance is captured by color and texture features. The specific details about the distinct features are discussed in the following sections.

4.2.1. Color Features

The color of an image pixel is given by its red, green, and blue (RGB) component. In this thesis a patch-based feature representation is proposed. There are multiple ways of representing the color of a patch, i.e., an image region. For example, color segmentation methods which aim at finding regions with homogeneous color, typically apply advanced methods, such as, e.g., Gaussian Mixture Models (see Permuter et al., 2006). In this work a rather simple approach based on the mean and variance of the patch (or patch sub-regions) is used to obtain a color feature vector y_c.

First of all, a feature vector y'_c is computed (see Equation 4.1). The mean color μ_c of the pixels $\rho(i)$ of an image patch's R, G, or B channel is given by Equation 4.2, where the number of pixels in a patch is denoted by N. Based on this, the color-variance for each channel σ_c can be computed with Equation 4.3.

$$y'_c = [\mu_c^{\mathrm{RGB}}, \sigma_c^{\mathrm{RGB}}] \tag{4.1}$$

$$\mu_c = \frac{1}{N} \sum_{n=1}^{N} \rho(i) \tag{4.2}$$

$$\sigma_c = \frac{1}{N-1} \sum_{i=1}^{N} (\rho(i) - \mu_c)^2 \tag{4.3}$$

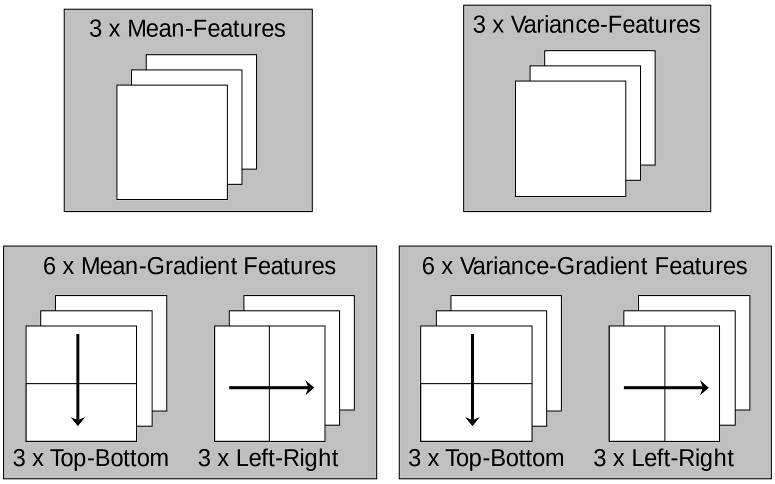

Figure 4.3.: Illustration of color features. Upper boxes: Mean and variance are computed on each channel of the whole patch. Lower boxes: Color gradient features are computed using the mean and variance offset in patch subregions

For every patch this results in a six dimensional feature vector $y_c \triangleq y'_c$ including mean and variance as depicted in the upper boxes in Figure 4.3.

The mean and variance of patches are rather primitive features for encoding the appearance of an image. Especially for transitions from one scene element to another, e.g., at a curbstone, also the color difference is important (see, e.g., Kong et al., 2010).

Therefore, additionally gradient-based color features are used. The color gradient is composed of left-right and top-down gradient by splitting each patch horizontally and vertically in half (see lower boxes in Fig. 4.3). The gradient is computed by a subtraction of mean and variance of right and left, and respective bottom and top patch channel content (see Eq. 4.5-4.6). The superscript $^{\mathrm{RGB}}$ illustrates that this is applied on each color channel (Ψ is a place holder for μ and σ). The corresponding mean or variance for each of the channels can be computed with Equation 4.2 and 4.3.

$$y_{c,\mathrm{grad}} = [\mu^{\mathrm{RGB}}_{c,\mathrm{hor}}\,,\ \mu^{\mathrm{RGB}}_{c,\mathrm{ver}}\,,\ \sigma^{\mathrm{RGB}}_{c,\mathrm{hor}}\,,\ \sigma^{\mathrm{RGB}}_{c,\mathrm{ver}}] \tag{4.4}$$

$$\Psi^{\mathrm{RGB}}_{c,\mathrm{hor}} = \Psi^{\mathrm{RGB}}_{c,\mathrm{right}} - \Psi^{\mathrm{RGB}}_{c,\mathrm{left}} \tag{4.5}$$

$$\Psi^{\mathrm{RGB}}_{c,\mathrm{ver}} = \Psi^{\mathrm{RGB}}_{c,\mathrm{bottom}} - \Psi^{\mathrm{RGB}}_{c,\mathrm{top}} \tag{4.6}$$

By merging all mentioned color features, we obtain a vector $y_c \triangleq y''_c = (y'_c, y_{c,\mathrm{grad}})$ with a dimension of 18.

4.2.2. Walsh Hadamard Texture Features

The Walsh Hadamard transform can be used to encode the texture of grayscale image patches as a feature (see, e.g., Alon et al., 2006). For the application as texture representation the transformation describes a conversion of a spatially sampled signal in form of a 2D-patch into its spectral components (frequency-sampled). The transform is orthogonal, symmetric, involutional, and linear. Basically, it describes a decomposition of an arbitrary grayscale input patch into a weighted superposition of 2D-Walsh matrices (see Fig. 4.4). The output of the transform is a feature vector y_{WH} containing the weights of the 2D-Walsh matrices. Because the transform describes a decomposition, summing up the y_{WH}-weighted 2D-Walsh matrices will reconstruct the input patch. The 2D-Walsh matrices are ordered by frequency which is also called sequency. In Figure 4.4, depicting the first 16 2D-Walsh matrices, we see that the order is increasing from left to right and respectively from top to bottom. The order is reflected by the number of zero-crossings, i.e., sign changes, in the matrix.

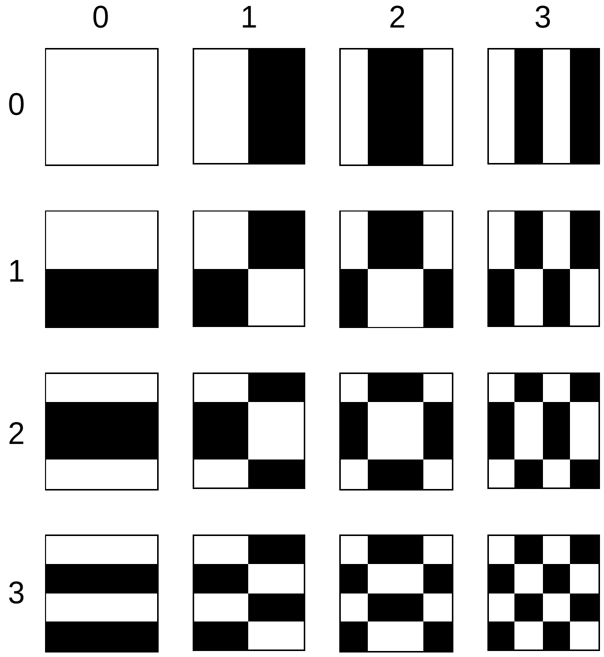

Figure 4.4.: 2D-Walsh matrices are the basis functions of the Walsh Hadamard transform. Each image patch can be decomposed into a weighted set of 2D-Walsh matrices. For the depicted 4 × 4 px Walsh matrices the white color corresponds to one and black to zeros.

For fast computation of the weights for each 2D-Walsh matrix (i.e., the elements of y_{WH}) the Fast Walsh Hadamard algorithm (see Hel-Or and Hel-Or, 2005) can be used. The input of the Fast Walsh Hadamard transform is restricted to real numbered matrices with a dimension that is a power of two. Assume having a patch with a size of $a_P = d \times d$ px ($d = 2^k$) the 2D Fast Walsh-Hadamard transformation can be split into a 1D column-wise and subsequent 1D row-wise transformation. Both 1D transformations can be efficiently computed with the same recursive algorithm. Initially, the column-wise transformation is applied on the input patch for each column independently resulting in an output with the same size as the input. Then, the row-wise transformation is applied on the output of the column-wise transformation for each row independently resulting in the final 2D weight matrix containing y_{WH}. The algorithm for computing the 1D-Walsh Hadamard transformation, e.g., for a row or column of a patch, employs a binary tree for computing the projections of a 1D vector onto all Walsh functions. Imagine a tree with $k+1$ layers where each branch of the tree connects one nodes at level i with another node at level $i+1$. Every node has two branches, one is assigned with $\alpha_i = 1$ and the other with $\alpha_i = -1$. Note that the root of the tree is on level $i = 0$ and the leafs on $i = k$. For each branch of the tree $(0 > i > k)$ a signal s is computed: The signal on level i is obtained by combining entries of the signal from the predecessor node (level $i-1$). On the one hand side, the combination depends on α_i which represents either an addition or a subtraction of two entries, and on the other hand side on the index offset $\mathit{\Delta}$. Let $\mathit{\Delta} = 2^i$ be the index offset of combination entries, then Equation 4.7 states that each entry j of signal s on level i is equal to entry j combined with the entry $j + \mathit{\Delta}$ of the predecessor signal.

$$s^{(i)}(j) \triangleq s^{(i-1)}(j) + \alpha_i \cdot s^{(i-1)}(j + \mathit{\Delta}) \tag{4.7}$$

"Thus, the 2 signals at level 1 are obtained by adding or subtracting consecutive entries in the signal of level 0 which is the original signal. The 4 signals at level 2 are obtained by adding/subtracting entries at distance 2 in the signals at level 1, and so on"(Hel-Or and Hel-Or, 2005, pg. 9).

At the leaves of the tree, i.e., after applying the k filters steps accordingly, we obtain the weights for the d Walsh functions. After applying the row-wise transformations for all d rows on the column-wise result of all d columns, the weights for the corresponding $d \times d$ Walsh matrices, i.e., the features y_{WH}, are obtained. For additional details on the Fast Walsh Hadamard transform, e.g., the derivation of the Walsh functions and the ordering of α_i so that the output signal is ordered according to the sequency of the Walsh functions, see Hel-Or and Hel-Or (2005).

4.2.3. Learning Road Appearance Descriptors using Slow Feature Analysis

Another type of visual feature analyzed in this thesis are slow features which encode the appearance and texture of image regions. The Slow Feature Analysis (SFA) is an unsupervised learning method which has been shown to be very efficient in obtaining class specific appearance descriptors for the task of patch-based classification (see Franzius et al., 2008). This Section is subdivided in two parts. Firstly, an introduction to SFA of temporal signals is given. Subsequently, it is explained how this concept is applied to learn specialized road terrain appearance descriptors.

Slow Feature Analysis of Temporal Signals

SFA is a learning technique which enables to find useful and invariant representations by using unsupervised learning (see Wiskott and Sejnowski, 2002). During training the algorithm performs an optimization in order to obtain a static transformation from a highly varying multidimensional temporal input signal to a slowly-varying output signal. This concept is illustrated in Figure 4.5. For vision-based tasks the rapidly changing sensory inputs, namely the pixel values, encode the behaviorally relevant visual information like class membership only indirectly. In our patch-based classification system the temporal signal corresponds to the change of pixel values x_i. The temporal change is generated by spatially shifting a patch over image areas belonging to one class and sampling the function value for each pixel.

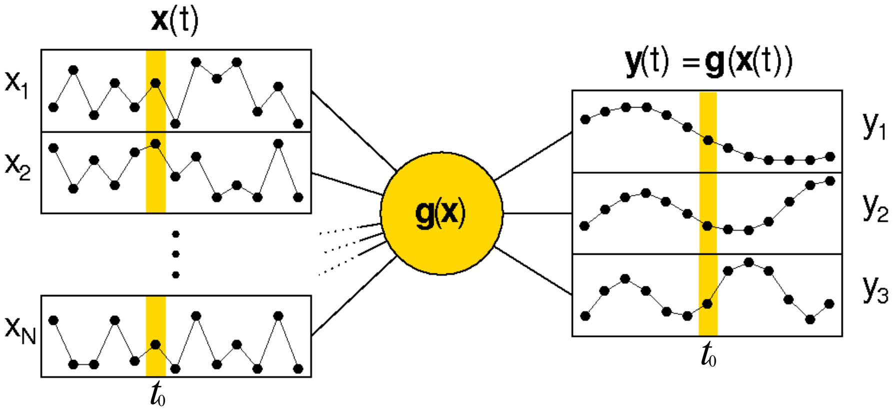

Figure 4.5.: Schematics of the optimization problem solved by slow feature analysis. Timestep t_0 marked in yellow, illustrating the instantaneous transformation.

Consider a two class problem. In order to easily separate input signals from the particular classes in feature space, a transformation creating output signals with low variance from arbitrary input signals belonging to one class is desired. This can be achieved with SFA because it creates a class specific representation for our type of input signals.

Additionally it can be used for order reduction, because in general a specified number of slow features that are able to distinguish inputs from different classes can be found (see Franzius et al., 2008). Mathematically spoken, we search the quantity of functions $g_j(x)$ that is generating the slowest varying output functions $y_j(t)$ from a multidimensional input signal $x(t)$ (see Eq. 4.8).

$$y_j(t) = g_j\left(x(t)\right) \tag{4.8}$$

Given Equation 4.8, one can formulate an optimization problem: Finding the transfer function $g_j(x)$ that minimizes the temporal variance of the output signals $\Delta(y_j)$ (see Eq. 4.9).

$$\Delta(y_j) = \left\langle \dot{y}_j^2 \right\rangle_t \tag{4.9}$$

We require uncorrelated output signals, having an equal variance and zero mean, which leads to the constraints in Equation 4.10-4.12. Accordingly, Equation 4.10 forces the output signals to be decorrelated, and Equation 4.11-4.12 exclude trivial solutions.

$$\forall i < j : \left\langle y_i \cdot y_j \right\rangle_t = 0 \tag{4.10}$$

$$\left\langle y_j^2 \right\rangle_t = 1 \tag{4.11}$$

$$\left\langle y_j \right\rangle_t = 0 \tag{4.12}$$

In Equation 4.9-4.12 $\langle f \rangle_t := \int_{t_0}^{t_1} \frac{1}{t_1 - t_0} f(t)\, \mathrm{d}t$ means averaging the function f over time and with the temporal derivative of f being $\dot{f}$. The detailed solution for the optimization problem can be found in Wiskott and Sejnowski (2002). For additional information on SFA please also refer to Franzius (2008).

Unsupervised Training of Road Terrain Appearance Descriptors using Slow Feature Analysis

Purpose of utilizing slow features in this thesis is to obtain an appearance descriptor for road terrain that captures additional local visual appearance information that is not contained in the other feature types. To this end, the SFA procedure explained above is applied to distinguish road and non-road areas.

As mentioned before, patches can be extracted in order to serialize the patch-pixel values into signals needed for SFA training. This spatial image sampling is realized by using predefined constant paths, which define how the point of patch-extraction moves over the image plane. There are two paths for each class, one horizontal p_{hor} and one vertical path p_{ver}, as illustrated in Figure 4.6. The advantage of using two paths is that the higher variability of the spatial input signal increases the likeliness of finding a useful transformation.

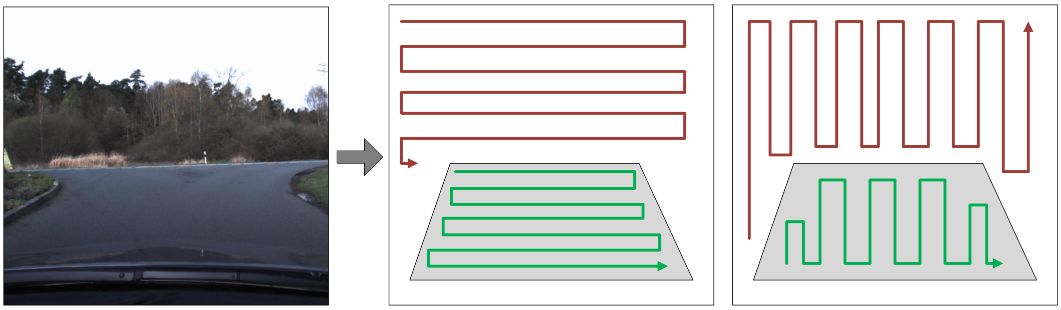

Figure 4.6.: Path for spatial patch sequence extraction for SFA training: on the left the horizontal and on the right the vertical path is illustrated. The paths are partitioned into road (green) and non-road (red) sections.

Given a patch $P_i = f(c_i, a_P)$, defined by its center $c_i = [c_{i,u}\,, c_{i,v}]$ and size $a_P = [a_{P,u}\,, a_{P,v}]$, one can sequentially extract patches by shifting the center c_i along a path $p_{\text{hor|ver}}$. Using a constant step size s_{grid} for this path, this results in a spatial signal $x_{\text{SFA}}(k_t)$. Note that the spatial index k_t corresponds to t from Equation 4.8. A signal $x_{\text{SFA}}(k_t)$ is a $d_k \times d_x$ matrix, where d_k describes the number of samples and $d_x = a_{P,u} \cdot a_{P,v} \cdot 3$ the input dimension of an image patch. In order to minimize the temporal variance for each class, temporal signals corresponding to road $x_{\text{SFA,R}}(k_t)$ and non-road $x_{\text{SFA,NR}}(k_t)$ are extracted, as it is illustrated in Figure 4.6. With ground truth information, given by a binary matrix (road = true), the assignment for every patch along the path can be found by thresholding the number of patch-pixels belonging to the road class. Every patch containing more than 50% of true pixels in the ground truth is interpreted as belonging to the road. Applying this for every training image we are able to train a model (linear SFA), defined by the transfer function $g(x(k_t))$, by presenting the system a certain number of signals $x_{\text{SFA,R}}(k_t)$ and $x_{\text{SFA,NR}}(k_t)$, using the SFA-TK Toolbox (see Berkes, 2003).

With the trained transformation, it is now possible to extract a slowly varying output signal $y_{\text{SFA}}(k_t)$ for every input image patch P_i. The signal $y_{\text{SFA}}(k_t)$ has the dimension n_{slow} ($n_{\text{slow}} \leq d_x$) which is the number of slow features.

In principle it should be sufficient to use the first slowest feature ($n_{\mathrm{slow}} = 1$) to separate the slowly varying road from the rapidly varying non-road (cf. Franzius et al., 2008), but due to noise and additional influences like changes in illumination and appearance (e.g., road markings, different surface colors, noise), the classification results improve for multiple slow features. Typically a limited number of slow feature n_{slow} with $n_{\mathrm{slow}} << d_x$ can be found which implies a huge reduction of the feature space, compared to the input dimension d_x.

4.3. Appearance Classification using Boosting

In order to classify the patch-based features into a specific road terrain category an offline-learning approach is applied. In contrast to online learning or adaptive approaches (see, e.g., Michalke et al., 2009) offline learned systems are time invariant. This has the advantage that robustness for certain conditions can be ensured in an offline learned system while parameters in an adaptive approach may diverge.

For training a classifier based on the proposed features boosting is applied, because it has been shown to be very successful in feature selection and classification (cf. Sha et al., 2008). In this work, the GentleBoost classification method is used. It is a modification of the AdaBoost algorithm proposed by Freund and Schapire (1997). GentleBoost is known to have similar performance as AdaBoost but is less sensitive to noise (cf. Friedman et al., 2000, p.354). The algorithm generates a weighted set of weak tree classifiers that build a strong classifier in combination (see Fig. 4.7).

This has the advantage that, in contrast to a strong classifier, weak classifiers are easy to construct. The method works iteratively and in every iteration the method adds another weak learner to an ensemble (a set of trees) in order to improve the classification step by step. The procedure uses a weighting in the input samples, starting with a uniform distribution for all samples, the method increases weights of those which were wrongly learned and decreases weights of those which were correctly learned. This allows focusing on the hard examples in the training data.

The following Section 4.3.1 concentrates on how training of GentleBoost is applied while Section 4.3.2 discusses the classification with previously trained parameters.

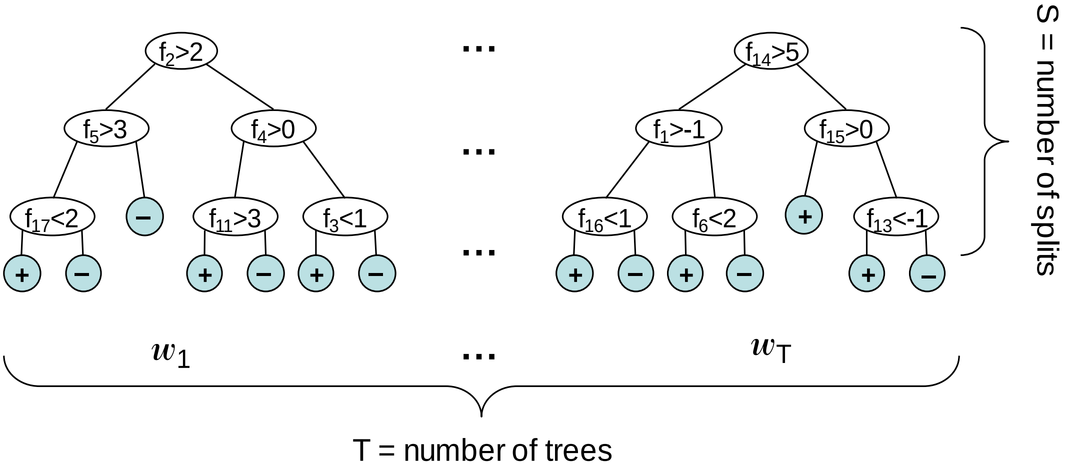

Figure 4.7.: Structure of an ensemble of weak tree classifiers as applied in the GentleBoost classifier. Each of the T binary decision trees has a number of S decisions and a weight w_i. Each decision is based on thresholding a feature f_j. The prediction function of each tree h_t produces a binary output {-1,1} denoted by +/- in the blue leafs.

4.3.1. Training of GentleBoost

GentleBoost, as described by Friedman et al. (2000), is a supervised machine learning method. Given are training samples $X = (x_1, ..., x_m)$ with known labels $L = (l_1, ..., l_m)$. Each training sample x_i is a vector consisting of multiple features f_j. We assume to have binary class labels $Y = \{-1, 1\}$ which refer to negative and positive training samples. For the training samples a weighting set $D_t = [d_t(1), ..., d_t(m)]$ is used. The method iteratively modifies the weights (each corresponding to a specific sample) in order to minimize the error of a regression function $\mathcal{F}_t$. The number of iterations is $t = [1, .., T]$. The weight of a training sample i on round t is denoted $d_t(i)$. In the first iteration the set is initialized as equal distribution $d_1(i) = 1/m$. The following regression algorithm describes how D_t is adapted according to the GentleBoost approach (step 1-4) in combination with decision trees as weak learners (in step 1):

1. Train a new weak learner, i.e., find a new prediction function $h_t(x_i)$ according to the current weighting of the samples D_t. Output of a weak learner is a prediction $\{-1, 1\}$ for a given feature vector x_i. A weak learner is a decision tree with a number of S tree splits. In every node of the weak learner a decision is based on thresholding of a specific feature value (see Fig. 4.7). Starting from the root of the tree, the methods selects for every node in the

tree a feature f_j optimizing an error function. Let $h^*_t(f_j, x_i)$ be the prediction function of the corresponding sub tree of a node in the weak learner which has a binary output and depends on the f_j-th feature of x_i. For every node the feature f_j that minimizes the weighed mean square error over all samples can be obtained with Equation 4.13.

$$f_j = \min_k E(f_k) \tag{4.13}$$
$$\text{with } E(f_k) = \sum_{i=1}^{m} d_t(i) \cdot (h^*_t(k, x_i) - l_i)^2$$

2. Assign a weight w_t for the weak learner with Equation 4.14 which considers the weighted training error represented by s_+ and s_-.

$$w_t = \frac{s_+ - s_-}{s_+ + s_-} \tag{4.14}$$
$$\text{with } s_{+|-} = \sum_{k=1}^{m} d_t(k) \cdot h_t(x_k) \;\; (\forall k | l_k = \{1| - 1\})$$

3. Update the regression function $\mathcal{F}_t$ with Equation 4.16 by computing the output of the current weak learner $\Delta^{\mathcal{F}}_t(x_i)$ (see Eq. 4.17). Where P_w is a weighted prediction of a weak classifier given a training sample x_i.

$$\mathcal{F}_t(x_i) = \sum_{v=1}^{t} \Delta^{\mathcal{F}}_v(x_i) \tag{4.15}$$
$$\mathcal{F}_t(x_i) = \mathcal{F}_{t-1}(x_i) + \Delta^{\mathcal{F}}_t(x_i) \tag{4.16}$$
$$\Delta^{\mathcal{F}}_t(x_i) = P_w(l_i = 1|x_i) - P_w(l_i = -1|x_i) = w_t \cdot h_t(x_i) \tag{4.17}$$

4. Update weight distribution D_{t+1} for the next iteration with respect to the current output using Equation 4.18. Weights of those samples which were wrongly learned are increased, weights of those which were correctly learned are decreased. Z_t is a normalization constant.

$$d_{t+1}(i) = \frac{d_t(i) \cdot \exp(-l_i \Delta^{\mathcal{F}}_t(x_i))}{Z_t} \tag{4.18}$$

After the T iterations are over, i.e., the ensemble is completed, Equation 4.19 can be used to obtain the weighted output of the tree ensemble $y(x_i)$.

$$y(x_i) \triangleq \mathcal{F}_T(x_i) \tag{4.19}$$

The value of $y(x_i)$ represents a confidence in the prediction, the sign $|y(x_i)|$ gives the binary classification.

4.3.2. Classifier Processing and Back Projection

In the processing phase of the LVA-system, patches are extracted from the image using a regular grid with a constant offset of s_{grid}. Similar to the training, a feature vector $x_{\text{LVA}}(i)$ is computed for every patch. This vector contains color, Walsh Hadamard and slow features. Using Equation 4.19, a confidence $y_{\text{LVA}}(x_i)$ for every patch is obtained. Based on patch center position c_i we map the confidences $y_{LVA}(x_i)$ onto the perspective image plane and obtain the image-based confidence map $conf_{\text{LVA}}(u, v)$ by applying a 2D bi-linear interpolation.

4.4. Dedicated Generation of Training Data

Basically, the LVA-system can be trained on all different kinds of visual categories by means of positive and negative training samples.

Training data refers to a set of feature vectors $X = (x_1, ..., x_m)$ where each x_i contains several appearance features. A combination of all features described in Section 4.2 is given by the vector x_i in Equation 4.20. Additionally it shows that each sample corresponds to a patch location in the perspective image (u_i, v_i).

$$x_i \longleftarrow [y_c(u_i, v_i)\,,\, y_{\text{WH}}(u_i, v_i)\,,\, y_{\text{SFA}}(u_i, v_i)] \tag{4.20}$$

Aim of this section is to detail how labels $L = (l_1, ..., l_m)$ for road area and road boundary can be extracted in order to use them for training a classifier as described in Section 4.3. The source for extracting labels are manually annotated ground truth polygons as depicted in Figure 4.8.

A label value $l_i = l(u_i, v_i)$ can be assigned for any pixel position. Basically, there are three possible assignment options: Either it is positive ($l_i = 1$), negative ($l_i = -1$), or "don't care" ($l_i = \emptyset$). The latter is important because some of the patches are hard to assign to the either positive or negative part or are for any other reason not desired to be part of the training data. The "don't care" samples will be neglected, i.e., not part of the training data.

The following Section 4.4.1 and Section 4.4.2 detail the generation of training data for road area and road boundary respectively.

4.4.1. Generation of Training Data for Road Area

The most significant road terrain category is probably the road area. The typical appearance of road area, i.e., the gray untextured asphalt including markings, potholes, etc, is termed *road-like area*. A distinction of *road-like area* and road area is

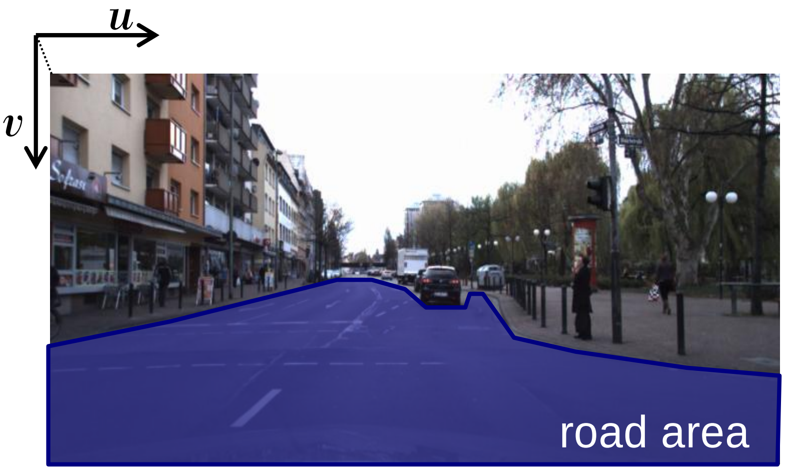

Figure 4.8.: Ground truth for road area: image regions are labeled using polygons. In the perspective image positive samples for road area, i.e., the region within the polygon, are colored blue.

important because it emphasizes that road area does not have a unique appearance that differs from other road terrain categories. For example, sidewalks and gray cars usually have similar appearance to the road area. Therefore, it can not be the goal of the appearance classification to find a perfect separation of road area and non-road area. Aim is to learn a generalizing model that gives a confidence for belonging to the class road area given the local visual appearance of a particular image region.

Let B_{RA} be a binary ground truth matrix with the size of the image (see Fig. 4.8), where each value $B_{\mathrm{RA}}(u_i, v_i) \in \{0, 1\}$ contains an annotation for road area. Based on the following rules a label $l_{\mathrm{RA},i}$ can be assigned for every sample at position $[u_i, v_i]$:

Rule A A positive sample ($l_{\mathrm{RA},i} = 1$) conforms $B_{\mathrm{RA}}(u_i, v_i) = 1$ and the patch at $[u_i, v_i]$ contains an amount of positive labels greater than a threshold $Th_{\mathrm{RA},+}$.

Rule B A negative sample ($l_{\mathrm{RA},i} = -1$) conforms $B_{\mathrm{RA}}(u_i, v_i) = 0$ and the patch at $[u_i, v_i]$ contains an amount of negative labels greater than a threshold $Th_{\mathrm{RA},-}$.

Rule C A "don't care" sample ($l_{\mathrm{RA},i} = \emptyset$) neither conforms to Rule A nor Rule B.

4.4.2. Generation of Training Data for Road Boundary

This section explains how training data for a visual appearance model for the road boundary is generated. The road boundary here refers to the boundary of road area, i.e., the transition from road area to other road terrain beyond or elevated obstacles such as pedestrians or cars.

If we think of a reasonable LVA-system that detects the road boundary, this system should have the following properties: Firstly, it should provide high confidences for locations where the appearance indicates the existence of a road boundary, e.g., the transition from road area to sidewalks. Secondly, it should provide low confidences on the actual driving space, i.e., the road area.

In the experiments in Appendix D it has been verified that training a dedicated road boundary classifier on the road boundary region results in a better suited LVA-system employing more discriminative features compared to the inverse approach from Section 4.4.1, i.e., a non-road area classifier.

Let B_{RB} be a binary ground truth map where each element $B_{\mathrm{RB}}(u_i, v_i) \in \{0, 1\}$ is zero except of a line of ones delineating the course of the road boundary[1] (see green line in Fig. 4.9). Furthermore, a binary ground truth map B_{RA} for road area as depicted in Figure 4.8 is required. Then, a label $l_{\mathrm{RB},i}$ can be assigned for each sample at location $[u_i, v_i]$ based on the following rules:

Rule A A positive sample ($l_{\mathrm{RB},i} = 1$) conforms $B_{\mathrm{RB}}(u_i, v_i) = 1$.

Rule B A negative sample ($l_{\mathrm{RB},i} = -1$) conforms $B_{\mathrm{RB}}(u_i, v_i) = 0$, and $B_{\mathrm{RA}}(u_i, v_i) = 1$, and the patch at $[u_i, v_i]$ contains an amount of positive labels greater (in B_{RA}) than a threshold $Th_{\mathrm{RB},-}$.

Rule C A "don't care" sample $l_{\mathrm{RB},i} = \emptyset$ neither conforms to rule A nor rule B.

It was discovered that additionally excluding lane markings from the negative samples is beneficial. In the system that will be detailed in Chapter 5 this allows focusing the classifier on curbstones, instead of also covering lane markings which are separately represented by another system part. For this purpose, a dark-light-dark transition detection is applied to extract a binary map $B_{\mathrm{LM}}(u_i, v_i) \in \{0, 1\}$ indicating the locations lane markings (more details on this method will follow in Section 5.2). In order to incoorporate $B_{\mathrm{LM}}(u_i, v_i)$ into the generation of training data an additional rule B* is used:

Rule B* A negative sample, conforming rule B, must also conform $B_{\mathrm{LM}}(u_i, v_i) = 0$.

[1] Note that the road boundary line can be easily extracted from the road area annotation.

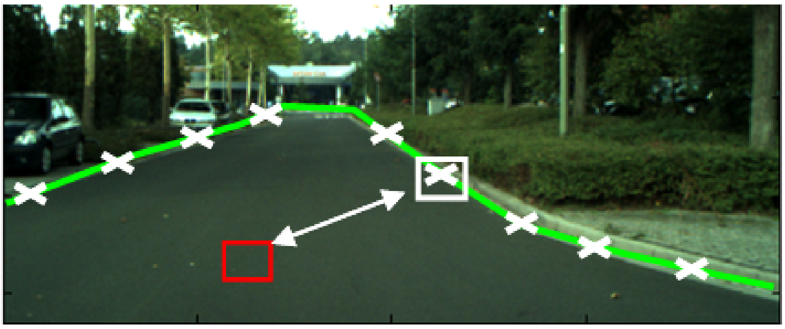

Figure 4.9.: Illustration how training samples are extracted. Green line showing the border between road and non-road. White crosses are exemplary extraction points for image patches for the positive training set. The red rectangle illustrates the negative training set consists of samples extracted from the road area.

Those samples which conform rule B but not Rule B* will be handled as "don't care" sample.

For generating a binary map of lane marking candidates standard techniques such as a dark-light-dark transitions detection can be used (e.g., proposed by Gopalan et al., 2012). This method will be further discussed in Section 5.2.

4.5. Experiments and Results of Local Visual Appearance based Classification

In this section several experiments on the LVA-system are conducted. There are two main objectives. The first is finding out which parameterization of the LVA-system achieves the highest performance. As there are dozens of free parameters for the system only the most relevant ones are discussed here. The experiments are based on road area classification because it is assumed that this is the most relevant road terrain category. Chapter 5 will evaluate the differences of the LVA-system trained on road area with a road terrain detection system combining visual and spatial features. It can be anticipated that good parameters found based on road area experiments are reasonable for the road boundary classification as well, because there is a high degree of overlap of the training material for road area and road boundary. The second main objective of this section is comparing the performance of the LVA-system with a baseline (see Appendix C) in order to assess its capabilities.

The used evaluation datasets contain images with an original resolution of 1024x1280 px. Before an RGB-image is processed a vertical cropping is applied which reduces the height to 271 px (see Fig. 4.1). Two types of datasets are used for the experiments. Firstly, a benchmark dataset which is rather small but containing a high variety of different appearances. And secondly, a sequential inner-city dataset which is extracted from three video streams recorded on different days containing different weather and lighting conditions. The temporal offset of image frames in each part of the dataset is 8 seconds in order to have a sufficient decorrelation in the training material. Additional information on the used datasets can be found in Appendix B.

All the experiments detailed in this section have several parameters of the LVA-system in common. For patch extraction, a patch size a_P of 21 × 21 px and a step size of $s_{\text{grid}} = 10$ px is used as this has been found as a reasonable setup in initial experiments. For classifying the resulting features, the GentleBoost algorithm is set up to generate 100 weak learners with 4 tree splits. For the evaluation of the resulting value-continues confidence maps with the ground data, a threshold of zero was applied. This was done to keep the evaluation simple although a different threshold might result in slightly better results. However, in other experiments it was found that this difference is very low for the GentleBoost classifier applied to the given task.

In the following section, the evaluation measures are detailed (see Section 4.5.1). The actual experiments are subdivided into four parts. In Section 4.5.2 local and global image normalization is compared to non-normalized RGB images. Subsequently, the influence of feature order and the combination of features are analyzed in Section 4.5.3. Accordingly, in Section 4.5.4, experiments for the road area detection on a larger inner-city dataset is performed in order to validate findings from previous experiments and to assess the generalization to different weather conditions. Finally, Section 4.5.5 gives results for the road boundary detection.

4.5.1. Perspective and Metric Evaluation Measures

For evaluation criteria from Álvarez and López (2008) are used. Based on the number of true positives TP, false positive FP, true negatives TN, and false negatives FN the corresponding rates, i.e., true positive rate TPR, false positive rate FPR

and so forth, can be obtained with Equation 4.21-4.24.

$$FNR = \frac{\sum_{i=1}^{n} FN}{\sum_{i=1}^{n} P} \cdot 100\% \tag{4.21}$$

$$TPR = 100\% - FNR \tag{4.22}$$

$$FPR = \frac{\sum_{i=1}^{n} FP}{\sum_{i=1}^{n} N} \cdot 100\% \tag{4.23}$$

$$TNR = 100\% - FPR \tag{4.24}$$

The quality Q (see Eq. 4.25) is a measure combining correct and wrong predictions of a classifier (see Michalke et al., 2009). In the thesis at hand, this measure is applied in the perspective image space (denoted by superscript ^{P}) and for obtaining the metric space quality Q^{BEV}. In Equation 4.26, the error ratio Σ, i.e., the ratio of TP to all errors, is given.

$$Q = \frac{\sum_{i=1}^{n} TP}{\sum_{i=1}^{n} (TP + FN + FP)} \cdot 100\% \tag{4.25}$$

$$\Sigma = \frac{TP}{FP + FN} = \frac{Q}{100\% - Q} \tag{4.26}$$

Equation 4.26 shows for example that a quality $Q < 50\%$ implies more errors $(FN + FP)$ as true detections (TP), i.e., $\Sigma = 1$.

For obtaining the metric quality Q^{BEV} the *Birds Eye View* (BEV) which was already introduced in Section 3.1 is used (see also Appendix A). Thus, each classification result and ground truth can be transfered into a metric representation and subsequently evaluated. For all the following experiments the BEV is defined for a range of -10m to 10m in x direction (lateral) and 7m to 47m in z direction (longitudinal). The origin of the coordinate system is under the center of the vehicle's rear axle on the road. This means the metric representation starts roughly 3m in front of the vehicle bumper. With a resolution of 5cm, the metric representation contains 400×800 data points.

4.5.2. Influence of Local and Global Image Normalization

In order to increase the robustness against illumination changes in the images a series of experiments is conducted. Here the default RGB color-space is compared to local and global image normalization applied on the RGB images.

Alon et al. (2006) reported that using image whitening, i.e., transforming gray scale images to have zero mean and unit variance, improves classification rates for road terrain detection compared to non-whitened gray scale images. Because in the conducted experiments color images are used, a modified version of the global image normalization method from Alon et al. (2006) is proposed. Instead of applying the method on every channel independently, we apply it on the three channels as a whole as defined by Equation 4.27-4.29. Here, the mean color value $\mu_{\text{glo,RGB}}$ and the variance $\sigma_{\text{glo,RGB}}$ are computed over all channels. This has the advantage that there is no shift in the relative difference of the color channels because the same linear scaling is applied.

$$R_{\text{glo,norm}}(u,v) = \frac{R(u,v) - \mu_{\text{glo,RGB}}}{\sigma_{\text{glo,RGB}}} \tag{4.27}$$

$$G_{\text{glo,norm}}(u,v) = \frac{G(u,v) - \mu_{\text{glo,RGB}}}{\sigma_{\text{glo,RGB}}} \tag{4.28}$$

$$B_{\text{glo,norm}}(u,v) = \frac{B(u,v) - \mu_{\text{glo,RGB}}}{\sigma_{\text{glo,RGB}}} \tag{4.29}$$

Another method which is known to achieve illumination invariance is a local image normalization (see Grgic et al., 2009), where each color value is normalized with the sum over all color values at the same position (u, v) as given in Equation 4.30-4.32.

$$R_{\text{loc,norm}}(u,v) = \frac{R(u,v)}{R(u,v) + G(u,v) + B(u,v)} \tag{4.30}$$

$$G_{\text{loc,norm}}(u,v) = \frac{G(u,v)}{R(u,v) + G(u,v) + B(u,v)} \tag{4.31}$$

$$B_{\text{loc,norm}}(u,v) = \frac{B(u,v)}{R(u,v) + G(u,v) + B(u,v)} \tag{4.32}$$

In order to infer the influence of global and local image normalization on the LVA-system nine experiments are conducted (see Table 4.1): Three for each feature type (color, Walsh Hadamard (WH), and slow features (SFA)) each without, with local and with global normalization. Accordingly, nine LVA-system for road area detection were trained on 90 road scene images from the benchmark dataset (see Appendix B.1). It is assumed that findings for the classification of road area also apply for road boundary. In order to evaluate the performance, the quality measure on the test dataset Q^P_{test} (see Section 4.5.1) was computed using a 30-fold cross validation. This means for each fold there are 87 training frames and 3 test frames. The averaged quality over all folds is listed in Table 4.1. This experiment setup allows analyzing the gain of image normalization for each of the proposed features separately.

For color features we see an improvement for local and global image normalization. For the WH-texture features a loss for local and a small gain in performance

Table 4.1.: Effects of normalization. The experiments are conducted on the benchmark dataset using 30-fold cross validation. Evaluation is carried out in the perspective image domain.

Feature	applied method	Q^P_{test} [%]
Color	without normalization	70.82
Color	local normalization	73.32
Color	global normalization	75.95
WH	without normalization	68.95
WH	local normalization	67.90
WH	global normalization	70.08
SFA	without normalization	68.56
SFA	local normalization	69.62
SFA	global normalization	73.22

for global normalization is visible. Therefore, the statement by Alon et al. (2006) also holds for this experiment. Applying local image normalization slightly improves the results using slow features. Surprisingly, the performance also increased significantly for the SFA features when using global image normalization, although the SFA itself includes a whitening of the input signal. The difference of the proposed whitening is that each frame is whitened independently while the SFA whitening uses a global mean and variance for all the data which is obtained during training. We see that global normalization leads to the best performance and is therefore a good trade-off between low information loss and illumination invariance.

4.5.3. Influence of Feature Order and Different Feature Combinations

For evaluating the LVA-system parameterization efficiently, a benchmark dataset with a total of 100 images containing a wide variety of different roads and weather conditions is used (see Appendix B.1). Having a small dataset allows conducting experiments quickly. Using a wide variety of different scenes in the training data increase the likelihood that the found parameters generalize well to unseen scenarios.

The LVA-system is trained for road area as it is assumed that the best found parameterization can be applied for road boundary detection as well. Section 4.5.2

verified that global normalization improves the quality for all features. Therefore, all evaluation experiments are combined with a preprocessing step applying the normalization.

Aim of the experiments in this section is finding a good individual feature setting and a good combination of features for the LVA-system. Therefore, the performance is evaluated for a varying number of features.

In the evaluation, the quality measure from Equation 4.25 is applied, where both, perspective image space quality Q^P and metric space quality Q^{BEV} (see Section 3.1.1) are included.

For comparison, the evaluation in the following subsections contains a baseline quality generated by summing up all binary ground truth maps and normalizing the result (see also Appendix C). This baseline is therefore essentially a scene prior similar to the one used in Álvarez et al. (2010). All results are obtained using a 20-fold cross validation and are discussed in the following paragraph. The training quality $Q^{P|BEV}_{\text{train}}$ measures how well the resulting classifier can handle the *known* training data, this value is averaged over all 20 trained classifiers. The test quality $Q^{P|BEV}_{\text{test}}$ measures how well *unknown* data can be classified and is also averaged over all runs. For the slow features it was found that training the linear SFA-model on the actual training part of the dataset is inconvenient because it might result in a slight overfitting. Therefore, for the following experiments the linear SFA-model is not trained on the benchmark dataset but a subpart of the inner-city dataset (see Appendix B.2).

Individual Features

The first experiment is conducted to find a suitable feature order for the features explained in Section 4.2. This is performed for each feature individually. The evaluation results for the slow features, Walsh Hadamard features, and the color features are listed in Table 4.2.

By comparing the results of the individual feature to the baseline it becomes visible why metric evaluation is relevant: In the perspective evaluation, up to a certain number of features the baseline outperforms the individual features. However, this is not the case for the evaluation in metric space. But why is this the case? In the perspective image the density of image pixels per square meter is very high for short distances (close to the observer) and low for far distances. In the near sector the baseline performs already very well and obtains a high Q^P. Although the perspective evaluation gives a nice measure how well a classifier performs in average, this leads to the problem that pixels corresponding to distant regions be-

Table 4.2.: Performance of the LVA-system with individual features on the benchmark dataset

	Q^{P}_{train} [%]	Q^{P}_{test} [%]	Q^{BEV}_{train} [%]	Q^{BEV}_{test} [%]
Baseline	-	73.0	-	40.7
# SFA-features				
3	59.7	53.6	52.5	50.1
10	76.2	71.6	62.4	59.4
20	78.0	73.0	63.5	59.6
30	78.8	73.3	64.2	59.9
40	79.2	73.9	64.1	59.9
# WH-features				
16	73.5	71.2	57.0	54.5
36	76.0	73.5	59.7	56.5
64	76.8	74.2	60.5	57.2
100	77.7	75.0	61.1	57.4
# RGB-features				
6	79.0	74.9	62.8	60.0
18	81.7	77.4	65.4	62.1

come statistically unimportant because the amount of analyzed samples in the close range is much higher than those in the far range. This is not the case in the metric domain where there is an equal number of samples for each metric distance. Apparently, feature-based road terrain detection is superior to the baseline especially in far distances because the difference in quality between LVA-system and baseline using Q^{BEV} is much higher than Q^{P}.

Furthermore, the results show that an increase of the feature order, i.e., the number of features, results in an increase of the overall classification quality. The gain saturates, i.e., for a certain amount of features the quality only increases insignificantly. In the following the findings of the distinct features are shortly discussed.

SFA Features: Evaluations have been performed by varying the number of features n_{slow} presented to the GentleBoost classifier. The perspective quality Q^{P}_{test} continues to increase for higher number of features but the metric quality Q^{BEV}_{test} reaches a maximum for 30 slow features.

Walsh Hadamard Features: The evaluation shows that perspective and metric quality increases with growing feature order but for higher feature orders the benefit becomes smaller.

RGB Color Features: While already the simple 6 RGB features show a good performance, a further increase is obtained by adding the 12 color gradient features. Note that both color feature sets achieve better performance than any SFA and Walsh Hadamard feature set.

Feature Combination

A further performance increase on the benchmark set can be achieved by combining the features. Table 4.3 shows the evaluation results for different subsets of feature types as well as for the full combination.

Table 4.3.: Feature combinations on the benchmark dataset.

Feature type	# features	Q^{P}_{train} [%]	Q^{P}_{test} [%]	Q^{BEV}_{train} [%]	Q^{BEV}_{test} [%]
Baseline	2	-	73.0	-	40.7
RGB & SFA	18+30	84.4	78.3	68.0	63.1
WH & SFA	64+30	83.3	77.5	67.3	62.0
RGB & WH	18+64	84.5	79.6	68.8	64.5
RGB & WH & SFA	18+64+30	86.1	80.2	70.3	64.9
RGB & WH & SFA	18+64+20	86.1	80.1	70.2	65.0
RGB & WH & SFA	18+100+20	86.3	80.2	70.2	64.6

The results show that each different feature type contributes to a performance increase. The maximum increase in metric Quality is 3 percentage points (pp) compared to the best single feature. Furthermore, it is visible that color is the most important feature for the classification, as feature combinations with color outperform feature combinations without. Picking the classifier with individual feature parameterization that performed best in isolation (see Table 4.2) is, however, not always a good choice as in the combination of multiple features it can lead to overfitting of the joint classifier to the training data. For instance, although in the individual evaluation 30 SFA features performed best, in combination with color and Walsh Hadamard features the use of 20 SFA features is sufficient and results even in a slightly better performance. A similar effect is observable for Walsh Hadamard features, where an order of 64 are sufficient.

The LVA-system with any of the feature combinations in Table 4.2 outperforms the baseline in both, the perspective and metric evaluation. Consequently, in the following evaluation of the overall approach the best performing configuration is selected using 20 slow features, a set of 64 Walsh Hadamard texture features and 18 RGB color features. This results in a feature vector x_{LVA} with 102 elements.

4.5.4. Evaluation and Results of Road Area Detection

Previous experiments are conducted on a rather small benchmark dataset. This means, that frames from different road scenes with different visual appearance characteristics were merged. This is useful for finding suitable feature setups that can be applied to multiple environmental conditions. Contrary, in the following experiment for road area detection a large dataset with sequential frames from one recording is used. Accordingly, aim of this section is to verify if the findings from Section 4.5.3 apply for larger sequential datasets as well. Therefore, validation experiments (see Table 4.4) are conducted on three sequential inner-city datasets (see Appendix B.2) recorded under different weather conditions: overcast (IC1), sunny (IC2), and mixed weather (IC3). Consecutive frames have a temporal offset of 8 seconds. Each dataset was recorded on an inner-city loop in Offenbach and consists of three rounds, this allows measuring the train and test quality using a 3-fold cross validation (1 round $\triangleq$ 1 fold). Furthermore, a merged dataset (all), consisting of all frames of the three datasets, is used. For the experiments the best feature setup from Section 4.5.3 was used (18 color, 64 Walsh Hadamard, and 20 slow features). The results, given in Table 4.4, are averaged over the three folds.

Table 4.4.: Experiments conducted on the inner-city dataset with different weather/lighting conditions.

	Perspective Evaluation			Metric Evaluation		
Dataset	Q^{P}_{BL} [%]	Q^{P}_{train} [%]	Q^{P}_{test} [%]	Q^{BEV}_{BL} [%]	Q^{BEV}_{train} [%]	Q^{BEV}_{test} [%]
Cloudy	78.1	91.7	85.5	50.2	74.1	67.2
Sunny	70.3	85.2	77.5	36.9	65.4	56.8
Mixed	75.0	87.6	80.9	47.6	70.3	62.8
All	73.9	83.3	79.8	43.0	65.6	62.1

In a performance ranking for the LVA-system applied on different weather conditions the cloudy dataset is on top because the highest values for Q^{P}_{test} and Q^{BEV}_{test} are obtained. Followed by the mixed weather dataset, which contains rainy, overcast and sunny parts, with a decrease of approximately 4.5pp in Q^{P}_{test} and Q^{BEV}_{test} compared to the overcast dataset. Finally, the sunny dataset obtains the worst performance with 3.5pp less in Q^{P}_{test} and 6pp less in Q^{BEV}_{test} compared to the mixed dataset.

Obviously, sunny is the hardest dataset to perform LVA classification, because of changes in illumination due to direct sunlight, light and shadow transitions on the road area caused by street canyons or trees. Surprisingly, if we consider the relative gain in Q^{BEV}_{test} compared to the baseline performance, which is a static model for the road area and therefore independent from appearance, the highest gain is obtained by the sunny dataset. The reason for this is that especially in the sunny dataset the static assumption for road area is violated frequently due to a higher traffic density which leads to a worse baseline performance. Furthermore, the evaluation shows a good generalization of the system when trained on all datasets. Here, Q^{BEV}_{test} is slightly minor than the mean of the three dedicated systems which indicates that the classification performance is comparable.

Finally, we see that the experiments conducted on Section 4.5.3 leaded to a good configuration for the sequential dataset as well. For all conducted experiments on the inner-city datasets the LVA-system clearly outperforms the baseline.

Figure 4.10 depicts 12 exemplary images from the 'all' dataset (testing frames only) overlaid by binary detection results. The raw images without detection can be found in Appendix B.2. The qualitative results show that the general local visual appearance of the road area, i.e., the mainly gray asphalt and markings on the road area, can be captured by the LVA-system. Therefore, the LVA-system must be very general to handle all depicted conditions with one configuration. Some examples (see, e.g., IX and XII in Figure 4.10) show that gray low textured areas on side walks sometimes cause false positive detections. Furthermore, there can be false positives on house walls or on cars due to a similar appearance to the road area. We see on a local appearance level there are a lot of ambiguous scene parts. The ambiguity problem (see Ho and Basu, 2002) can not be fully solved by taking different types of local visual features (a solution will be discussed in Chapter 5). Also bigger markings on the road area such as road signs (see VII in Figure 4.10) are not covered by the internal appearance model of the LVA-system. In example IX the transition from a dark shadow area to a very light area almost has no negative effect on the detection. However, in example X where the shadow is collateral to

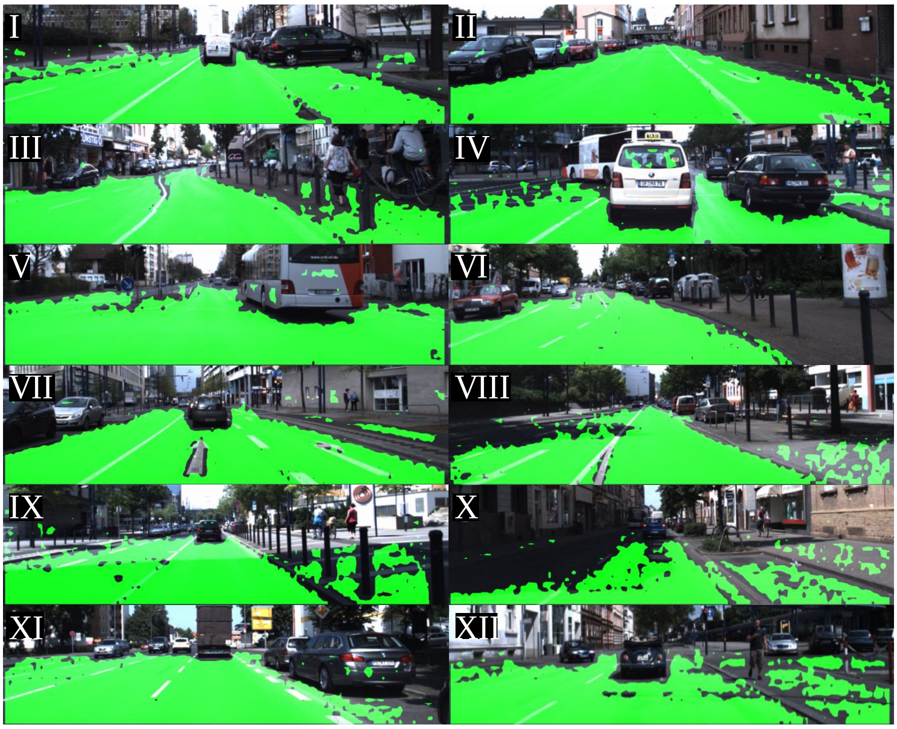

Figure 4.10.: Example scenes with depicted road area detection result marked in green. The raw images can be found in Appendix B.2.

the driving direction, the whole left lane is falsely not detected as part of the road area.

4.5.5. Results of Road Boundary Detection

Analogously to the road area, a LVA-system for road boundary can be trained (see Section 4.4.2). Figure 4.11 illustrates some exemplary detection results for road boundary. The results show that road delimiting elements, such as curbstones, are well covered (see, e.g., I, VI and VIII in Figure 4.11). Furthermore, the system detects the transition to obstacles on the road area such as cars (see, e.g., IV and VII in Figure 4.11) and traffic islands (see V in Figure 4.11). Lane markings and road signs are detected as road boundary (see, e.g., VII in Figure 4.11). Beyond that, the road boundary detector, wrongly marks shadow edges as positive which can be seen in example X and XI.

Further analysis of the LVA-system for road boundary integrated in a bigger system, and its value for road terrain detection will be discussed in Chapter 5.

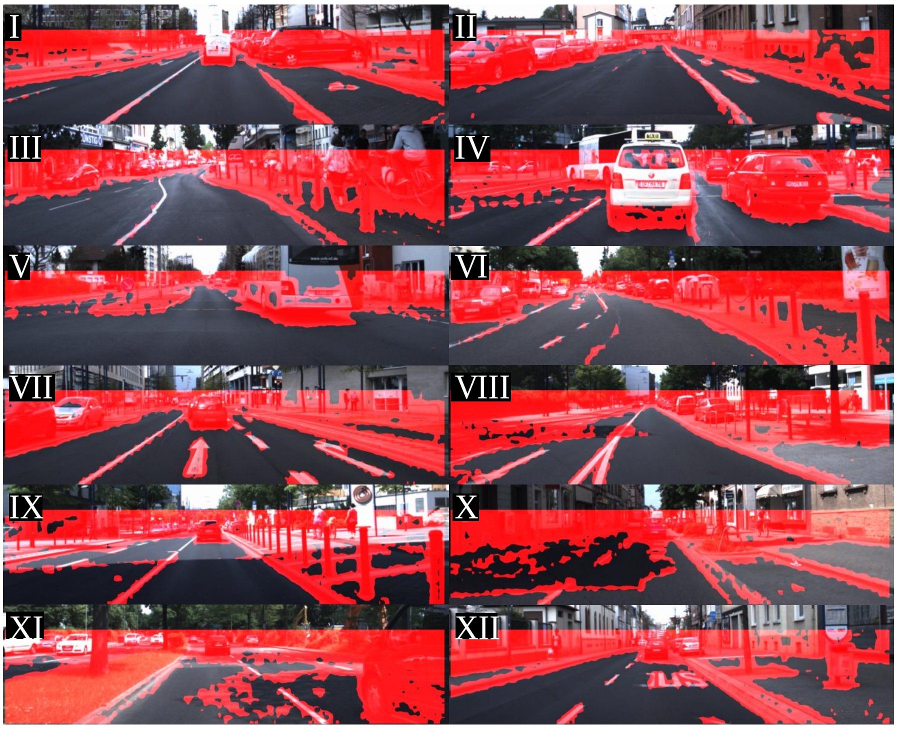

Figure 4.11.: Example scenes with depicted road boundary detection result marked in red. The raw images can be found in Appendix B.2.

4.6. Summary

This chapter analyzed road terrain detection based on Local Visual Appearance (LVA). In this context, a LVA-system is presented that learns the typical appearance of road terrain categories. This system captures LVA by means of RGB-color, Walsh Hadamard texture, and slow features extracted from image patches. A GentleBoost classifier is applied to generate a confidence map indicating for each pixel position whether it is likely to belong to the trained road terrain category or not. The generation of databases with dedicated training material for road area and road boundary is obtained by applying a set of rules which are based on ground truth annotations.

In the conducted experiments, it was found that the proposed global image normalization significantly improves the classification performance compared to non-normalized images and local image normalization. Another series of experiments was applied to find a suitable parameterization of the LVA-system on a small benchmark dataset.

The system trained for the purpose of road area detection was evaluated on a large inner-city dataset with three different weather conditions. The approach clearly outperforms the baseline, which is an optimal static model of the road area. This becomes visible especially for datasets where the static assumption of the road is violated frequently, i.e., for heavy traffic. The results show a comparable performance of general training (applied on a merged dataset with all weather conditions) compared to dedicated training. However, local visual appearance-based classification suffers from the ambiguity problem (see Ho and Basu, 2002). Meaning, without spatial knowledge, e.g., an implicit model based on retinal position (see, e.g., Sha et al., 2008; Kang et al., 2011), or an explicit road shape model (see, e.g., Danescu and Nedevschi, 2011), false positive detections on areas, such as sidewalks, cars or house walls having a similar road-like appearance, have to be expected.

The LVA-system for road boundary detection shows that the appearance of delimiting elements can be distinguished from general road area. The proposed training strategy neglecting non-road area regions in the training data outperforms binary road versus non-road training. Qualitative results showed that the proposed method for road boundary detection enables to correctly classify boundary elements such as curbstones, the transition to cars or other obstacles on the road area.

The following chapter introduces spatial features which capture the spatial layout of local visual appearance in order to overcome drawbacks of pure local appearance based classification (e.g., the above mentioned ambiguity problem).

5. Incorporating the Spatial Layout of Local Visual Appearance

The following chapter presents a generic concept for the visual and spatial (visuospatial) analysis of the road environment which has been developed in the course of this thesis. An in-depth overview and motivation of the approach applied to road terrain detection was already discussed in Chapter 3.

Chapter 4 found that using local visual features for road terrain classification may result in ambiguities. This means that on a local scale the road area is hard to discriminate from, e.g., a gray wall from a house, or low textured regions on a metallic gray car. The generic approach detailed in this chapter enables to overcome these ambiguities by joining visual and spatial aspects. This will be shown for the example of road area detection which can be directly compared to the results in Chapter 4. Beyond that the proposed approach can be applied as trainable detector for arbitrary road terrain types. The utilization of visuospatial classification enables to discriminate road terrain regions with identical visual but different spatial characteristics (see Kuehnl et al., 2012). One example for this is the detection of ego-lane, as all lanes on the road usually have the same appearance.

The remaining chapter is organized as follows: A structural overview of the generic concept for visuospatial classification is given in Section 5.1. Afterwards, Section 5.2 details the base classification of the proposed approach which is making use of the local visual appearance classification detailed in Chapter 4. Subsequently, in Section 5.3 the core of the approach, the spatial layout computation is discussed. The application of the approach for road terrain classification follows in Section 5.4. Subsequently, experiments applying the proposed system to road area detection are discussed in Section 5.5. Then, Section 5.6 presents experiments and results for ego-lane detection. Finally, this chapter will be summarized in Section 5.7.

5.1. Basic Concept for Visuospatial Classification

This section will give an overview of the basic concept of visuospatial classification. The consecutive steps for visuospatial classification are illustrated in Figure 5.1.

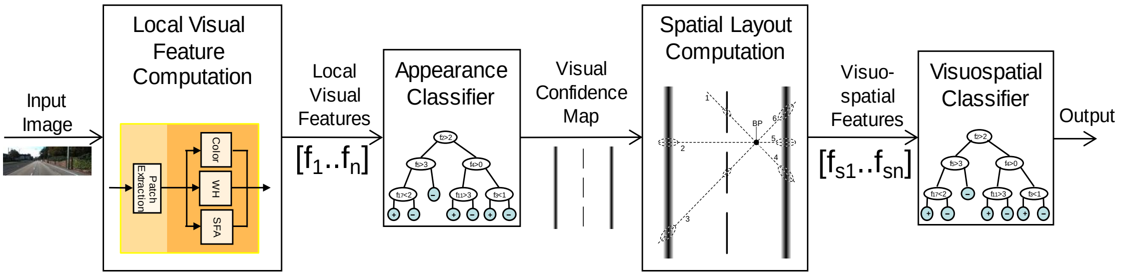

Figure 5.1.: The basic concept for visuospatial classification comprises 4 steps: Local visual feature computation, appearance classification, spatial layout computation, and visuospatial classification. The visuospatial classifier can be applied for road terrain detection or lane index classification.

Input of the approach is an image. Firstly, local visual appearance features $[f_1..f_n]$ are computed at multiple locations in the image. Subsequently, these features are classified to generate confidences with respect to a certain class. Both steps can be realized by the LVA-system presented in Chapter 4. For instance, the LVA-system can be trained on the road boundary as described in Section 4.3. The output is than a confidence map representing this particular visual property (as the illustrated visual confidence map in Figure 5.1). The spatial layout computation operates on this confidence map to compute visuospatial features $[f_{s1}..f_{sn}]$. These features capture spatial characteristics of the confidence maps property with respect to a defined computing position which is called a base point (BP) of the visuospatial classification (see Fig. 5.2).

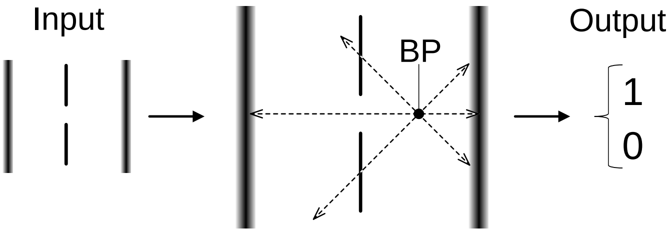

Figure 5.2.: Visuospatial classification is applied with respect to a defined base point (BP) at the input (i.e., a confidence map). The output is a classification decision.

Further details on the spatial layout computation and the proposed SPatial RAY (SPRAY) features will follow in Section 5.3. Finally, the visuospatial feature vector is processed by the visuospatial classifier. For instance, assuming a two lane road and a visuospatial classifier trained on ego-lane, the result of the example in Figure 5.2 should be 1 (true) because the base point is located on the ego-lane.

The basic concept for visuospatial classification can be incorporated in a bottom-up system architecture as depicted in Figure 5.3. This visuospatial classification system consists of three parts (see left part of Fig. 5.3): base classification, spatial feature generation, and visuospatial classification.

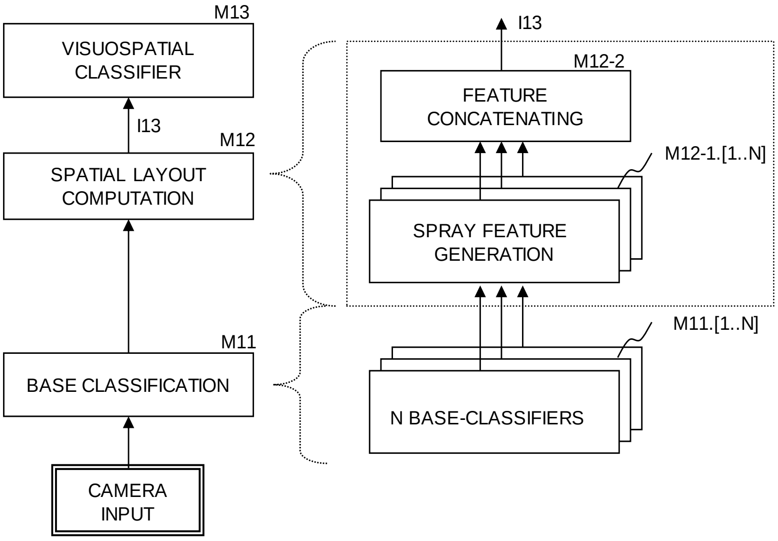

Figure 5.3.: System architecture for visuospatial classification showing the main processing steps (left) and a more fine grained structure (right).

The camera input is fed into each of the N base classifiers (M11) as can be seen in the right part of Figure 5.3. Each base classifier (M11.i) obtains a confidence map for a specific visual property. An example for these properties is the road boundary. More details on the base classification will follow in Section 5.2. In order to increase the spatial homogeneity, it is reasonable to represent the confidence maps in the metric domain because perspective changes are compensated here. On each metric confidence map, a spatial layout computation (M12-1) that captures spatial aspects of this confidence map's property is applied. Therefore, the resulting features capture both, visual appearance and the spatial layout. The spatial feature generation operates on value-continuous confidence maps which retains more information from the original image compared to taking explicit decisions based on visual appearance in the base classification, e.g., by binarization. All the individual features computed from the N different base classifier cues are concatenated (M12-2) to obtain a SPRAY feature vector (I13) for a base point. Finally, the SPRAY feature vector (I13) is used for performing the visuospatial classification (M13) with respect to a desired application.

For instance, for the purpose of road terrain classification, the approach can draw inference whether the base point is located, e.g., on or off the road area, or the egolane. The road terrain detection system that will be discussed later in this chapter uses multiple base points, distributed all over the metric space, for detecting a particular road terrain category. Beyond that, classification of visuospatial features is useful for identifying the lane index with respect to a single representative base point. This will be discussed in Chapter 6.

5.2. Base Classification for Capturing Local Visual Appearance

The block diagram in Figure 5.4 shows for proposed system for visuospatial classification with a concrete setup employing three base classifiers in module (M11.1-3) capturing local visual appearance as discussed in Chapter 4.

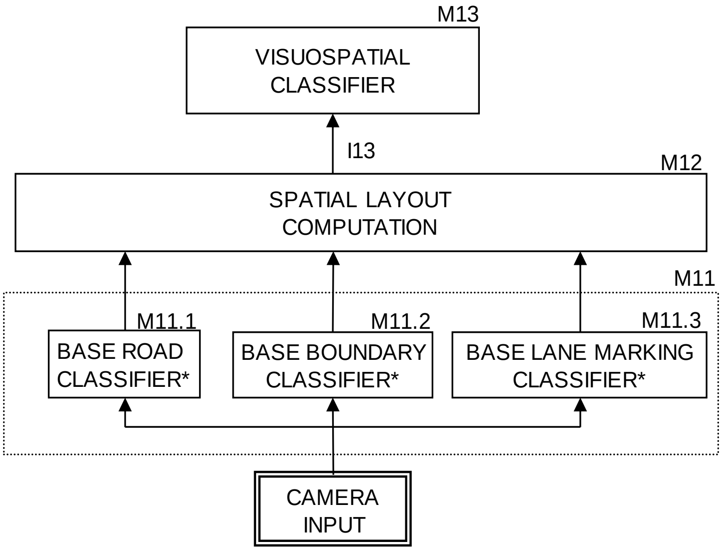

Figure 5.4.: System block diagram showing the setup for base classification in module M11. Note that all base classifiers (*) include preprocessing and inverse perspective mapping to provide metric confidence maps.

Each base classifier generates a metric representation of confidence values, wherein each location corresponds to a certain location in the BEV which is internally obtained using inverse perspective mapping (see Appendix A). An entry of this confidence map contains confidence information about whether a corresponding cell in metric space has a certain property, i.e., whether its visual appearance indicates, e.g., the presence of road area.

The proposed system architecture uses three base classifiers (M11.1-3) which work on preprocessed camera images and result in three confidence maps in a metric representation. These are base road classifier (M11.1), base boundary classifier (M11.2) and base lane marking classifier (M11.3). The two base classifiers for road (M11.1) and boundary (M11.2) are based on the LVA-system presented in Chapter 4. This approach enables to learn the typical visual appearance of a given class based on texture and appearance features extracted from image patches. The only difference in the two LVA-systems is the training that is used to specialize each classifier on the specific task (see Section 4.4).

The base road classifier (M11.1) is specialized to generate high confidences on the *road-like area* and low confidences on non-road terrain (including, e.g., cars and pedestrians). The base boundary classifier (M11.2) is specialized on detecting boundaries between the road area and adjacent regions, like sidewalks, traffic islands, off-limits terrain, or non-road terrain. This base boundary classifier generates low confidences on *road-like area* and high confidences at locations that correspond to boundaries. Both base classifiers, base road and base boundary, make use of global image normalization for preprocessing, as this has been found to be advantageous in the experiment in Section 4.5.2. After computing the base classification result as detailed in Chapter 4 the confidences are mapped into the BEV to obtain a metric confidence map. Subsequently, these confidences are normalized to a range $\hat{y}_{LVA} \in [-1..1]$. The normalization employs a constant for the maximal confidence $\mathrm{conf}_{\max}$ which is empirically estimated.

For the base lane marking classifier (M11.3) a dark-light-dark transition detection is applied. The method is similar to standard techniques (see, e.g., Gopalan et al., 2012; Veit et al., 2008) and tuned to obtain only few false negative while having a lot of false positive detections. Multiple filter kernels are applied on the luminance channel mapped into the BEV. Subsequently, thresholding is applied on each filter result to select image regions corresponding to lane markings with a certain width and exhibiting the typical dark-light-dark illumination transition. A confidence is computed by summing up the binary results and normalizing. After applying the method for every pixel in the image, a confidence map, typically showing high confidences at locations corresponding to lane markings and low confidences on low textured road terrain (e.g., *road-like area* or sidewalk), is obtained. The output of

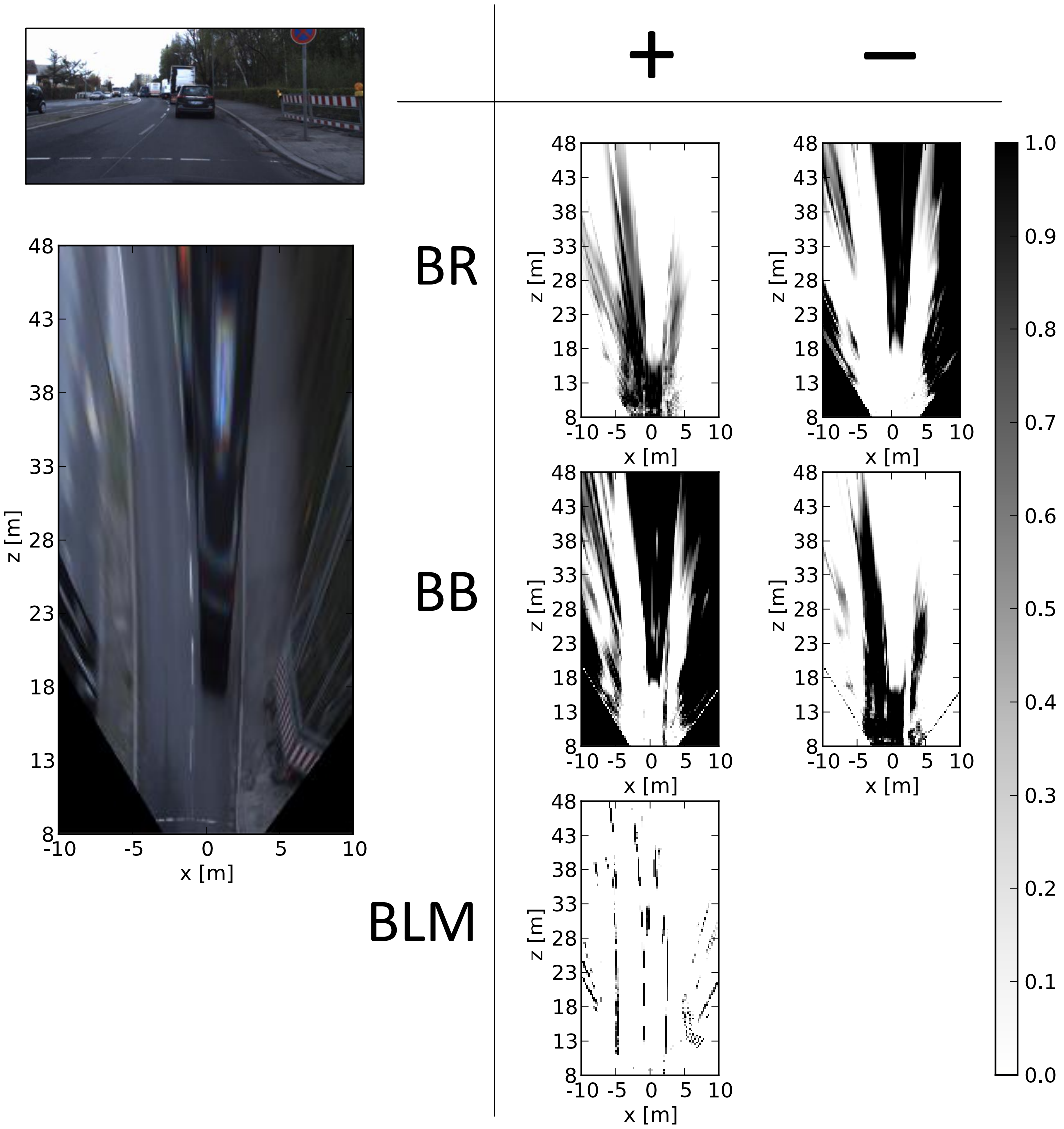

Figure 5.5.: Result of the base classification showing an illustration of the base road (BR), base boundary (BB) and base lane marking (BLM). Additionally, the positive (+) and negative (-) part of the confidence maps are depicted separately.

the base lane marking classifier is a metric confidence map with values $\hat{y}_{LM} \in [0..1]$.

Figure 5.5 shows exemplary results of the base classification in the metric BEV. The combination of all base classifier outputs, i.e., the three confidence maps, build a value-continuous visual representation of the image. These confidence maps are used as inputs for the spatial layout computation which will be detailed in the following.

5.3. Spatial Layout Computation

To overcome limitations of sole local appearance based decisions, a combination of local visual appearance based classification with a spatial layout computation is proposed. The approach followed here, is capturing the spatial layout by means of SPatial RAY (SPRAY) features that have been developed during this thesis. Note that spatial features refer to information gathering strategies that are not restricted to a local surrounding. In contrast to local feature gathering (e.g., inside a window with a fixed size), spatial features can consequently combine feature information from multiple spatial locations and capture the extent of regions and shapes. As the SPRAY features are computed using the appearance information from the base classification, it can be seen as a complementation of traditional local visual features with spatial information.

Approaches employing other spatial features are detailed in the related work chapter (see Section 2.2). For example, a bulk of features extracted at different spatial locations relative to a base point is a state-of-the-art approach for body part recognition (see Shotton et al., 2011). Furthermore, it has been shown by Smith et al. (2009) that spatial features in the shape of rays can be very beneficial for shape based classification. This ray approach captures the intrinsic shape characteristics with respect to binary cell boundaries. In contrast to this, the proposed SPRAY features encode spatial characteristics even in larger distances, e.g., beyond road delimiting elements such as curbstones and lane markings. Furthermore, SPRAY features operate directly on the continuous confidence values of the confidence maps from the base classification. Therefore, no explicit decisions are made based on local visual appearance. This procedure retains more information from the original image which is expected to be superior to an approach using explicit decisions, e.g., a binarization after the base classification.

In Section 5.3.1 follows an overview of the method for spatial layout computation. Section 5.3.2 details the basic concept of the proposed SPRAY feature computation. Subsequently, Section 5.3.3 gives some algorithmic details on the approach. How a visuospatial representation is generated from the extracted features is discussed in Section 5.3.4. Section 5.3.5 details the discrimination of ego-lane and other parts of the road area.

5.3.1. Method for Spatial Layout Computation

The proposed spatial feature computation is a process with consecutive steps which is applied on a single, non-negative, metric confidence map from a base classifier. This procedure is therefore applied as many times as the number of input confidence maps. The spatial feature generation (M12) process is illustrated in Figure 5.6. The left part of Figure 5.6 shows the general processing steps for feature generation.

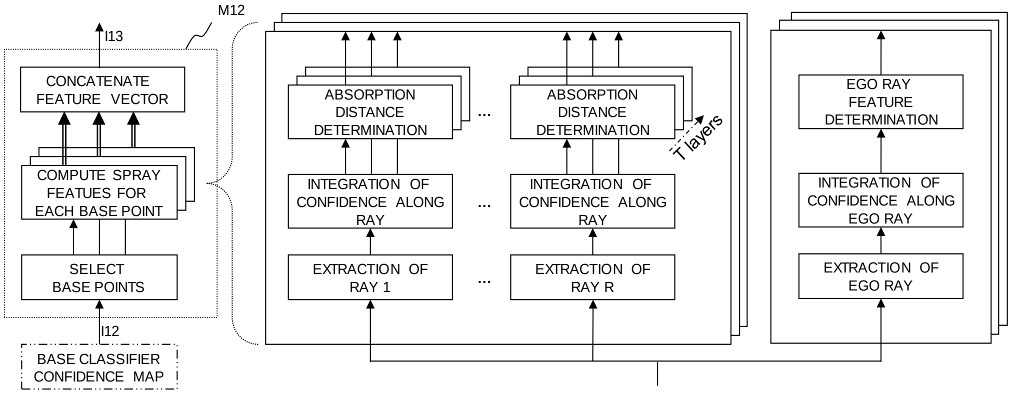
M12
I13
CONCATENATE FEATURE VECTOR
COMPUTE SPRAY FEATUES FOR EACH BASE POINT
SELECT BASE POINTS
I12
BASE CLASSIFIER CONFIDENCE MAP
ABSORPTION DISTANCE DETERMINATION
...
ABSORPTION DISTANCE DETERMINATION
T layers
INTEGRATION OF CONFIDENCE ALONG RAY
...
INTEGRATION OF CONFIDENCE ALONG RAY
EXTRACTION OF RAY 1
...
EXTRACTION OF RAY R
EGO RAY FEATURE DETERMINATION
INTEGRATION OF CONFIDENCE ALONG EGO RAY
EXTRACTION OF EGO RAY

Figure 5.6.: System block diagram showing the general processing steps of spatial feature generation (left) and a fine grained illustration of the SPRAY feature computation which is applied for each base point (right).

Input of the spatial layout computation is a confidence map from a base classifier (I12) with values in the range [0..1]. The computation is applied for defined positions in the metric representation, i.e., the previously introduced base points (BP).

For every base point several SPRAY features are generated to capture spatial aspects of the base classifiers property (see right part of Fig. 5.6). The features are ray-like. Consequently, each ray is emitted with a certain angle. Based on the captured information along the ray, several SPRAY feature values are extracted. Beside the ordinary SPRAY features, a special type of SPRAY feature named ego-SPRAY is computed. Finally, all kinds of SPRAY features belonging to one base point are merged. More details on the features follow in the next section.

Note that the base road and base boundary comes with positive and negative confidences, i.e., not the above mentioned range of [0..1] (see Fig. 5.5). Therefore, optionally one can use the positive part of the confidences only or split the confidence map into positive and negative part and compute the spatial layout for both. The question if this has a benefit will be evaluated in the experiments in Section 5.5

and Section 5.6. This is not an option for the lane marking cue as there are no negative confidences.

5.3.2. Spatial Ray (SPRAY) Feature Generation

In the literature one can find several methods for extracting spatial characteristics in an image (see Section 2.2). This thesis proposes a ray casting method which can be seen as an extension of the ray approach for cell microscopic imagery by Smith et al. (2009). Similar to this approach, rays are casted from a base point in a certain angular direction which allows extracting features that reflect distances from the base point to locations where a certain property is existing. However, outdoor vision system operate in more challenging conditions causing a lot of noise for the perception. In the context of this thesis, this affects the base classifier outputs, i.e., the confidence maps to contain more errors. For this reason, a ray casting method which spatially integrates confidences along the rays is proposed. This allows the direct usage of the noisy confidence maps as input for the feature generation without applying binarization (which is a prerequisite for the approach by Smith et al. (2009).

As mentioned above, the computation of SPRAY features is carried out with respect to a base point. An example for a distribution of base points, defined by a grid, is shown in Figure 5.7 (circles in left illustration). The spatial layout with respect to the confidence map is captured at each individual base point by means of rays (see right part of Fig. 5.7).

The example in Figure 5.7 illustrates a confidence map of one base classifier in the metric BEV which is used for explanation purpose. It reflects a simplified confidence map (example for I12) for a two-lane road with lane markings in the center and curbstones on the left and right side (dark color indicates high confidences). The simulated base classifier generates high confidences on curbstones (depicted by the bigger bars) and lane markings and low confidences on road terrain and is therefore comparable to the base boundary classifier (see Section 4.5.5).

In the following subsections the two SPRAY features types, ordinary and ego-SPRAY features, will be detailed.

Ordinary SPRAY Features

The spatial layout with respect to a base point in a non-negative confidence map is captured by radial vectors, which are called rays. A ray-vector R_α includes all confidence values along a line, with a certain angular orientation α, starting from a

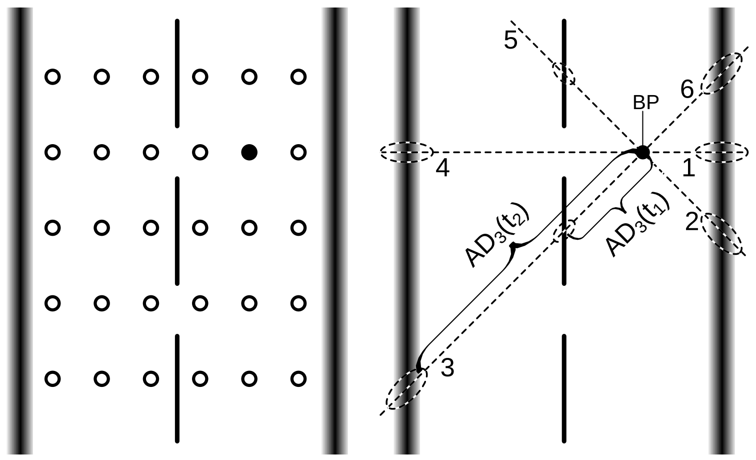

Figure 5.7.: Distribution of base points over the metric space (left) and the spatial feature generation procedure illustrated for one base point (right).

specific base point (x_{BP}, z_{BP}) and ending at the border of the metric representation. The example in Figure 5.7 (right) shows six rays (1-6) enumerated clockwise, all starting from the same base point.

To convert this information into a defined number of feature values $f_{\mathrm{SPRAY},\alpha}$, the integral of the confidence values $A_\alpha(\rho)$ along the ray R_α is computed (see Eq. 5.1). This integral can be interpreted as absorption of confidences along the ray. Therefore, the confidence integral value, i.e., the absorption value, encodes a likeliness of how likely it is that the ray runs over locations with a given property. In particular, for the example of positive boundary confidences, $f_{\mathrm{SPRAY},\alpha}$ represents the locations where a certain amount of boundary confidences were absorbed. These locations are defined by the ray angle α and the distance from (x_{BP}, z_{BP}), i.e., the absorption distance AD_α. Then, a specific SPRAY feature $f_{\mathrm{SPRAY},\alpha}(t_i)$ is equal to the absorption distance $AD_\alpha(t_i)$, i.e., the locations where the integral value A_α reaches a certain threshold t_i (see Eq. 5.2).

$$A_\alpha(\rho) = \int_0^\rho R_\alpha(\gamma)\,\mathrm{d}\gamma \tag{5.1}$$

$$f_{\mathrm{SPRAY},\alpha}(t_i) = AD_\alpha(t_i) = \operatorname*{argmin}_\rho \left(\rho \,|\, A_\alpha(\rho) > t_i\right) \tag{5.2}$$

By defining a certain number N_T of absorption thresholds and a number N_α of different ray orientations, a SPRAY feature vector consisting of $N_T \times N_\alpha$ elements is generated for one specific base point.

For the third ray (3) in Figure 5.7, the absorption distances $AD_3(t_1)$ and $AD_3(t_2)$ for two thresholds are depicted. Additionally, the graph in Figure 5.8 shows a principal sketch of the integral over the third ray given the confidence map mentioned above (see Figure 5.7). In the example, the thresholds are optimally chosen so that, threshold t_1 and t_2 lead to absorption distances that correspond to the distances from the base point to a lane marking $AD_3(t_1)$ and the left curbstone $AD_3(t_2)$. Apparently, the selection of 'good' thresholds is crucial for the method.

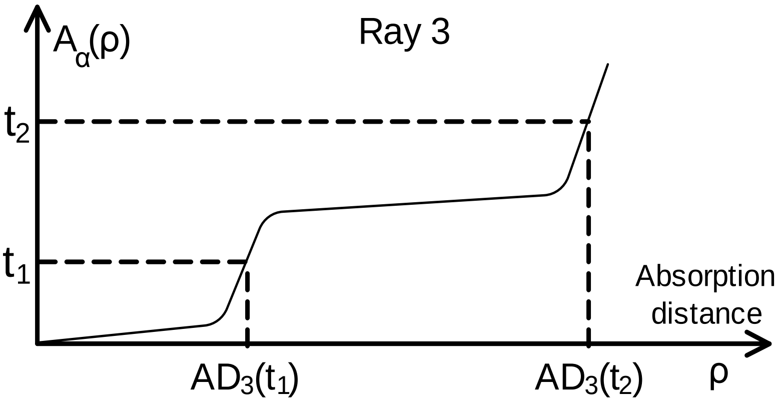

Figure 5.8.: Integral over the confidences (absorption) for a specific ray (here ray 3 in the scenario from Fig. 5.7). Two spatial features $AD_3(t_1)$ and $AD_3(t_2)$ are obtained, which reflect in this case the distance to the lane marking and the left road boundary.

Smaller absorption thresholds result in shorter distances than bigger thresholds. Therefore, the features are monotonic increasing. The absorption distance offset from one threshold to another is a good indicator, e.g., for a transition of lane marking to road area. Consider computing SPRAY features on the road boundary confidence map using two thresholds where the second threshold is only slightly bigger than the first. If one coincidently finds a big difference in the feature values it can be argued that the distance of the first feature value is possibly the end of a boundary, e.g., a lane marking.

Ego-SPRAY Features

Ego-SPRAY features capture a properties of the ray in between the base point and the location of the ego-vehicle, i.e., the lower center position in the BEV.

The idea is to use the absorption value of the integral produced from a ray, send from the base point (x_{BP}, z_{BP}) to the ego-position $(x_{\text{ego}}, z_{\text{ego}})$ in the metric representation, as a feature. This is for example an indicator on which lane a base point is located. For instance, with increasing absorption value, e.g., for lane marking, also the likeliness for not being located on the ego-lane is increasing. The feature value f_{ego} can be obtained with Equation 5.5 after obtaining the angle α_{ego} (see Eq. 5.3) and the distance d_{ego} (see Eq. 5.4) from BP to the ego-position.

$$\alpha_{\mathrm{ego}} = \arctan\left(\frac{z_{BP} - z_{\mathrm{ego}}}{x_{BP} - x_{\mathrm{ego}}}\right) \tag{5.3}$$

$$d_{\mathrm{ego}} = \sqrt{\left(z_{BP} - z_{\mathrm{ego}}\right)^2 + \left(x_{BP} - x_{\mathrm{ego}}\right)^2} \tag{5.4}$$

$$f_{\mathrm{ego}} = A_{\mathrm{ego}} = A_{\alpha}(\rho = d_{\mathrm{ego}}) \text{ with } \alpha = \alpha_{\mathrm{ego}} \tag{5.5}$$

In contrast to the standard ray features the orientation α_{ego} is changing for different base points. This is for instance beneficial for encoding ego-lane specific spatial properties.

For example, consider a confidence map for boundary as depicted in Figure 5.7. For most of the base points on the opposing lane the integral value from the base point to the ego-position is higher than for base points on the ego-lane. Therefore, this feature is expected to be beneficial for ego-lane classification. In the experiments (see Section 5.6), this will be evaluated thoroughly.

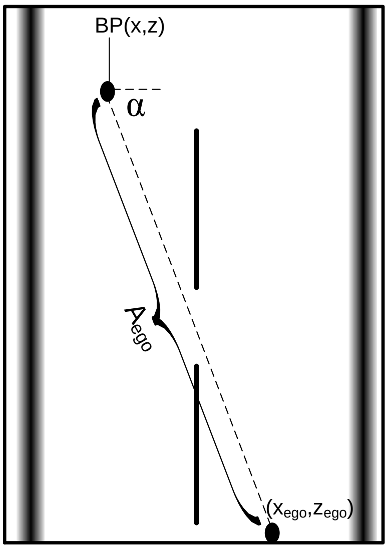

Figure 5.9.: Illustration for an ego-SPRAY feature f_{ego}. The feature is equal to the integral A_{ego} from BP to $(x_{\mathrm{ego}}, z_{\mathrm{ego}})$.

5.3.3. SPRAY Algorithm

Basis of the SPRAY algorithm is the computation of several features for one base point. Although, for one base point different angles and thresholds are considered, the calculation procedure is the same for each angle.

A regular grid, as illustrated in Fig. 5.7, is convenient to distribute base points over the metric space. The base points then define the locations where SPRAY features are computed. Furthermore a certain amount of angles is defined. For each base point angle combination, a ray vector $R_\alpha(\rho)$ is defined as a vector containing all confidences starting at the base point $BP = (x_{BP}, z_{BP})^T$ and ending at the boundary of the metric space. In order to get the corresponding content from the metric confidence map, the ray position $(x_R, z_R)^T$ for increasing ray length ρ can be calculated with Equation 5.6.

$$\begin{pmatrix} x_R \\ z_R \end{pmatrix} = \begin{pmatrix} x_{BP} \\ z_{BP} \end{pmatrix} + \rho \cdot \begin{pmatrix} \cos(\alpha) \\ \sin(\alpha) \end{pmatrix} \tag{5.6}$$

Because ray positions are static for every base point, the mapping of the discretized metric positions to rays can be precomputed once for all ρ.

The flow diagram in Figure 5.10 shows the processing steps of the algorithm for calculating the absorption distance given one ray and a certain threshold t_i. As the calculation of each base point and angle combination is independent from each other the features can be computed efficiently by applying parallelization.

One issue of the algorithm is the handling the situation when the ray integral reaches the end of the defined metric space while not all features are computed, i.e., not all absorption thresholds are exceeded yet. It was found that this boundary behavior of the algorithm affects the spatial smoothness of the resulting feature values and therefore the visuospatial classification. Spatial smoothness means that one can expect similar feature values for neighboring base points. The proposed solution for this problem is a continuation of the integration when reaching the boundary of the metric space. This approach can be interpreted as a prediction of the confidence values in the unobservable part of metric space.

Let ρ^* be the distance from the base point to the boundary of the defined metric space, then is the predicted value of ray vector R^*_α (see Equation 5.7) a constant given by Equation 5.8 reflecting the average of absorbed confidences along the ray A_α (computed with Equation 5.1).

$$R^*_\alpha(\rho) = R^* \qquad (\forall \rho > \rho^*) \tag{5.7}$$

$$R^* = \frac{1}{\rho^*} \cdot A_\alpha(\rho^*) \tag{5.8}$$

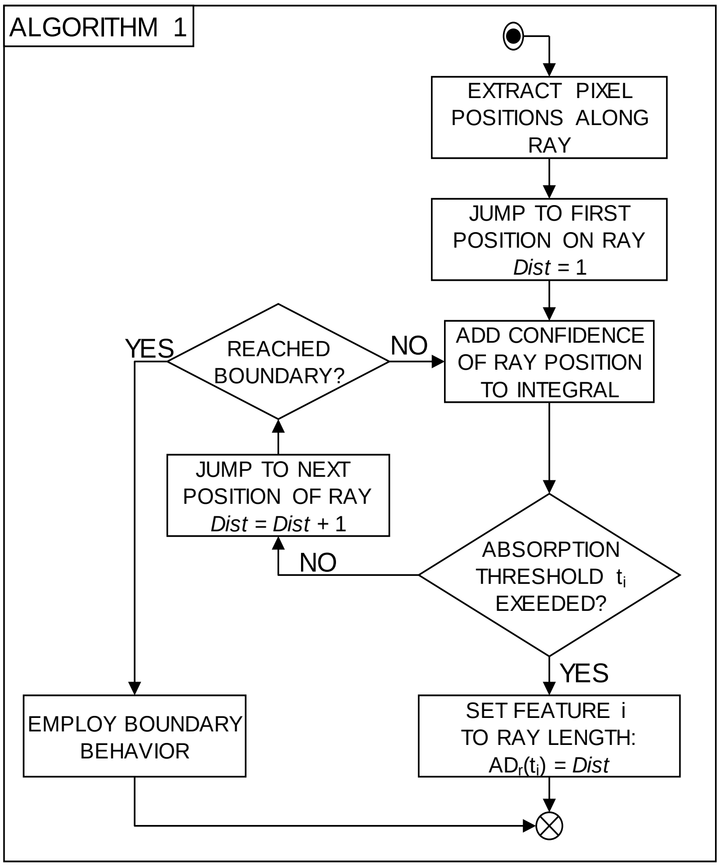
ALGORITHM 1
EXTRACT PIXEL POSITIONS ALONG RAY
JUMP TO FIRST POSITION ON RAY *Dist* = 1
ADD CONFIDENCE OF RAY POSITION TO INTEGRAL
REACHED BOUNDARY?
YES
NO
JUMP TO NEXT POSITION OF RAY *Dist* = *Dist* + 1
ABSORPTION THRESHOLD t_i EXEEDED?
NO
YES
EMPLOY BOUNDARY BEHAVIOR
SET FEATURE i TO RAY LENGTH: $AD_r(t_i)$ = *Dist*

Figure 5.10.: Flow chart illustrating the extraction algorithm for SPRAY feature generation given one ray and one threshold. *Dist* refers to ρ from Equation 5.6.

This procedure achieves a higher spatial smoothness and a wider distribution of absorption distances AD_α compared to setting the not yet computed features to a constant value (e.g., ρ^*).

5.3.4. Combining SPRAY Features to a Visuospatial Representation

After computing the distinct SPRAY features, i.e., ordinary and ego-SPRAY features for the three base cues, a visuospatial representation is generated. Basically all features belonging to one base point are concatenated to a single SPRAY feature vector. Remind that the SPRAY features where computed on a value-continuous local visual representation, i.e., the confidence maps from the three base cues.

Therefore, SPRAY features implicitly encode both, visual appearance and its spatial layout for a specific base point. The compilation of SPRAY features for the distinct base points form a visuospatial representation which is used as input for the road terrain classification detailed in the next section.

Figure 5.11 depicts an exemplary scene and the resulting SPRAY features with two ray angles send in opposite directions (180° and 0° from left to right). Both SPRAY features have a low threshold. Therefore, the features reflect a distance (i.e., absorption distance) where the first boundary is hit. In both examples one can clearly see the cap between the lane markings in the near range. The absorption distances above 4m (color: orange, red) are obtained for rays casted thought the gap and therefore reflecting the distance from the base point to the left or right curbstone.

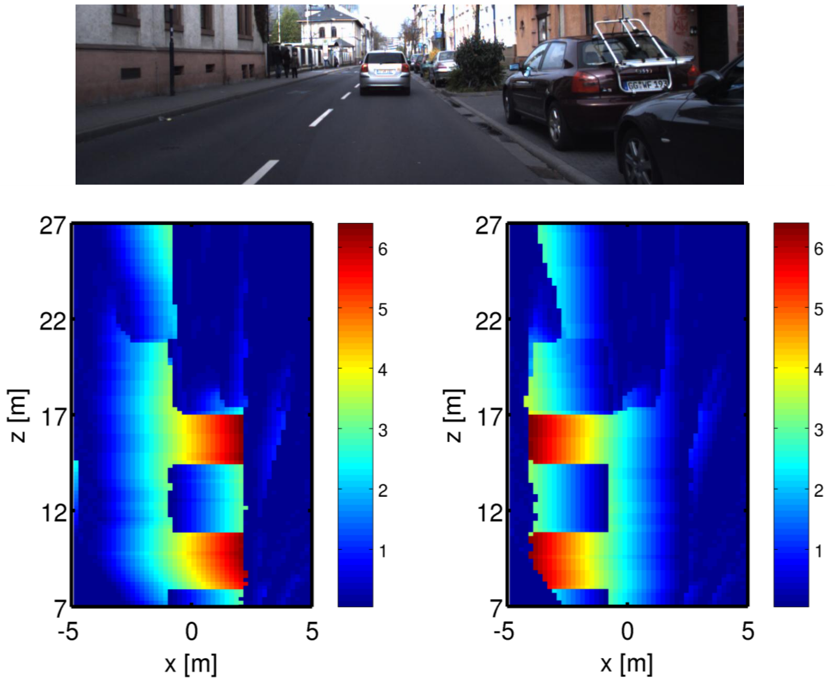

Figure 5.11.: Two exemplary SPRAY features for a road scene with lane markings. The left plot is a SPRAY feature casted to the left (with $\alpha = 180°$), the right illustration is casted in the opposite direction ($\alpha = 0°$). Both are computed on the positive part of the boundary cue.

In the following section it will be shown that SPRAY feature are a very useful representation for discriminating the ego-lane from the opposing lane.

5.3.5. Discrimination of Ego-Lane from other Lanes

As discussed in the related work chapter ego-lane detection is typically performed by combining detection of lane delimiters, i.e., lane markings or curbstones, with tracking using a particular lane model (see Section 2.1).

It has been shown that a vast number of spatial feature combinations enable to discriminate different body parts (see Shotton et al., 2011) or differently shaped areas in images Smith et al. (2009). Aim of this section is to show that the properties of the proposed SPRAY features enable to distinguish visually similar but spatially different road terrain, i.e., for a two lane road, the ego-lane and the opposing lane. For explanation purpose serves the example depicted in Figure 5.12.

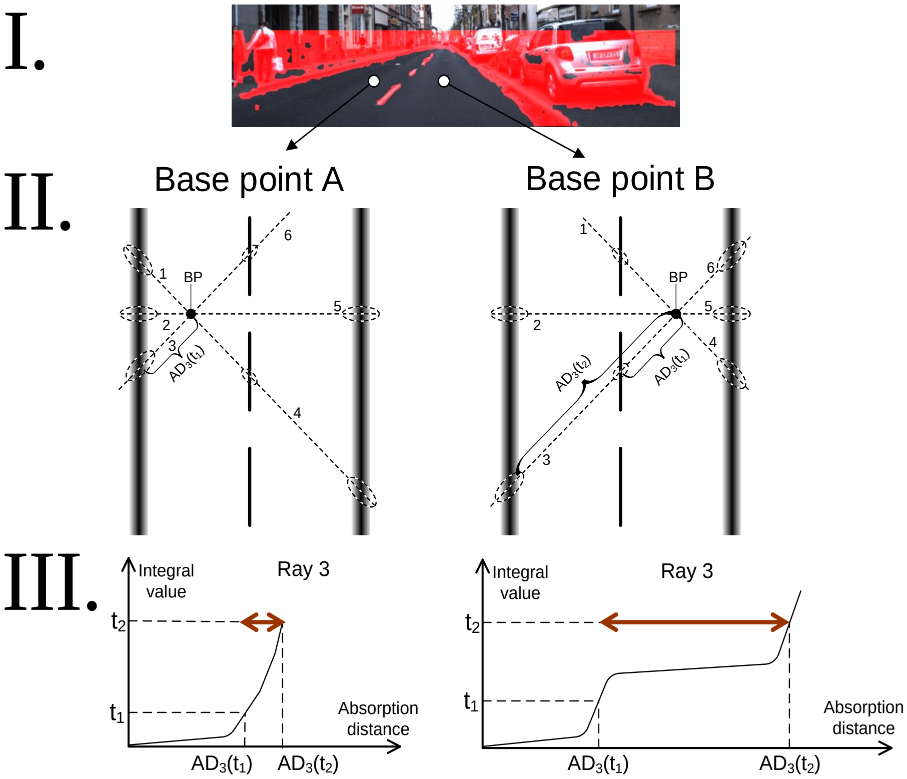

Figure 5.12.: Example for discrimination of ego-lane from the opposing lane on a two lane road. Part I shows an example image with detection result for base boundary (red) and two base points on the road area. Base point A is on the opposing lane, B is on the ego-lane. Additionally, the spatial rays in the BEV (II) and the ray integral for the 3rd ray (III) corresponding to A and B are illustrated.

Consider a base point A on the opposing lane and a base point B on the ego-lane. In order to assign one of the base points to the correct road terrain category one has to find a discriminative characteristic, i.e., a SPRAY feature having different values on the ego-lane than on the opposing lane. Figure 5.12 (II) shows the course of the rays where SPRAY features are extracted. Having a closer look at ray number 3 in Figure 5.12 (III) one can see for base point A that one thresholds is reached after the ray crosses the left curbstone of the road area on the opposing lane. There is a low distance between feature $AD_3(t_1)$ and $AD_3(t_2)$ because directly after the first also a second threshold is reached. This is not the case for base point B on the ego-lane, as here the first threshold is reached at the lane marking, and the second at the left curbstone of the road area. Therefore, there is a higher offset between $AD_3(t_1)$ and $AD_3(t_2)$ which is potentially a good feature for distinguishing opposing lane and ego-lane in this particular example. With the GentleBoost classification method applying threshold-based decision trees as weak classifiers this feature difference can not be represented directly. However, multiple combination of $AD_3(t_1)$ and $AD_3(t_2)$ with other SPRAY features would result in an even better combination, i.e., suitable for more examples than the given one.

The decision for a road terrain category relies on several feature combinations by means of decision trees that have been learned in a training session. To get an idea of the complexity of the resulting classifier, assume that a decision tree has $S = 4$ tree splits. Therefore, there are $2^S = 16$ possible leafs of the decision tree that can be reached. The leafs than represent a combination of 4 features. Using, e.g., an ensemble of $T = 100$ trees, this results in an amount of 1600 nodes while 100 are non-zero given one feature vector. A weighted sum of these 100 values are used as final classification decision (cf. Section 4.3).

5.4. Road Terrain Classification

This section explains how a Road Terrain Detection System (RTDS) is obtained using the proposed visuospatial classification. To this end, a road terrain classifier is trained in the metric domain based on the proposed SPRAY features. The GentleBoost method using decision trees as weak learners is used. GentleBoost was already used for base road and base boundary classification in the perspective image domain. Therefore, all basic details on the method can be found in Section 4.3, the transfer of the method for road terrain detection is detailed in the following.

Section 4.3 detailed how the method can be trained for a specific road terrain category based on local visual features and binary ground truth maps in the perspective image domain. In contrast to this, for the visuospatial classification metric ground truth information is required. In this chapter, the proposed system is applied to road area and ego-lane classification. Therefore, BEV-based ground truth maps are generated for these two road terrain categories.

Basically, for generating training material for road area in the metric domain, one can follow most of the instructions detailed in Section 4.4.1. However, as the road terrain classifier operates in metric space, a binary matrix B_{RA}^{BEV} is obtained by applying Inverse Perspective Mapping (IPM) on B_{RA} (see Appendix A). Then, following the rules as in Section 4.4.1, and replacing B_{RA} with B_{RA}^{BEV}, will result in the desired label $l_{\mathrm{RA}}^{BEV} \in \{1, -1, \emptyset\}$ for a SPRAY feature vector computed at any base point location in the metric representation.

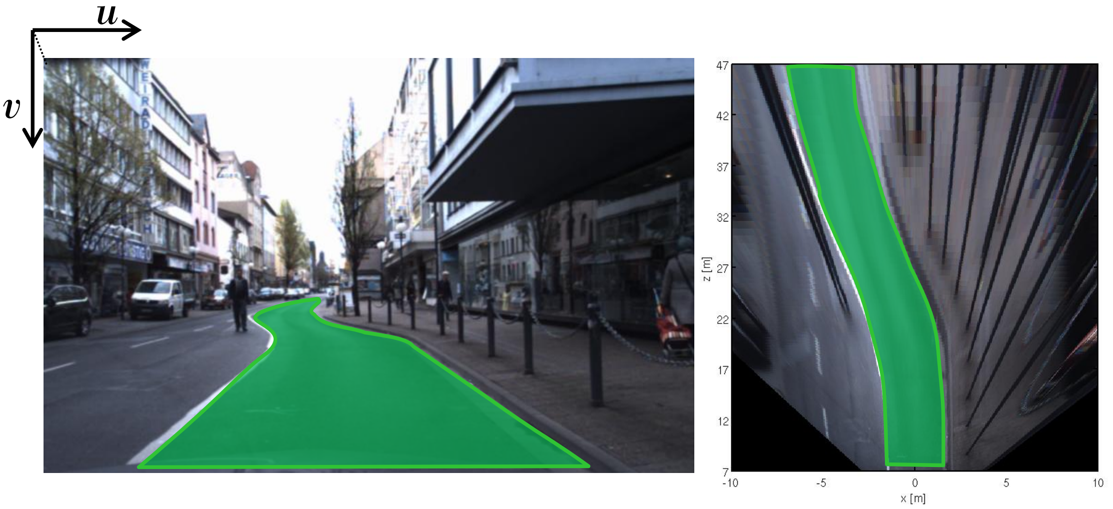

Figure 5.13.: Ground truth for ego-lane: image regions are labeled using polygons. In the perspective image positive samples for ego-lane, i.e., the region within the polygon, are colored green. On the right, the metric BEV is depicted

For training a classifier that learns the distinction between ego-lane and non-ego-lane it is assumed that there is a binary matrix B_{EGO}^{BEV} for ego-lane with elements $B_{\mathrm{EGO}}^{BEV}(x_i, z_i) \in \{0, 1\}$. The source for extracting labels are manual ground truth annotations as depicted in Fig. 5.13. Then, proceeding as for road area, using the Rules A,B, and C of Section 4.4.1, will result in the training labels $l_{\mathrm{EGO}}^{BEV} \in \{1, -1, \emptyset\}$.

After training for a specific road terrain category is finished, the classifier generates a confidence value for a given SPRAY feature vector, indicating whether the corresponding base point is likely belonging to the trained category or not. The confidences are normalized to a range of $[-1..1]$. Analogously to Section 4.3.2, based on the confidences results for the distinct base points, 2D-linear interpolation is applied to create a confidence map with the same size as the BEV.

An exemplary classification result for ego-lane is given in Figure 5.14.

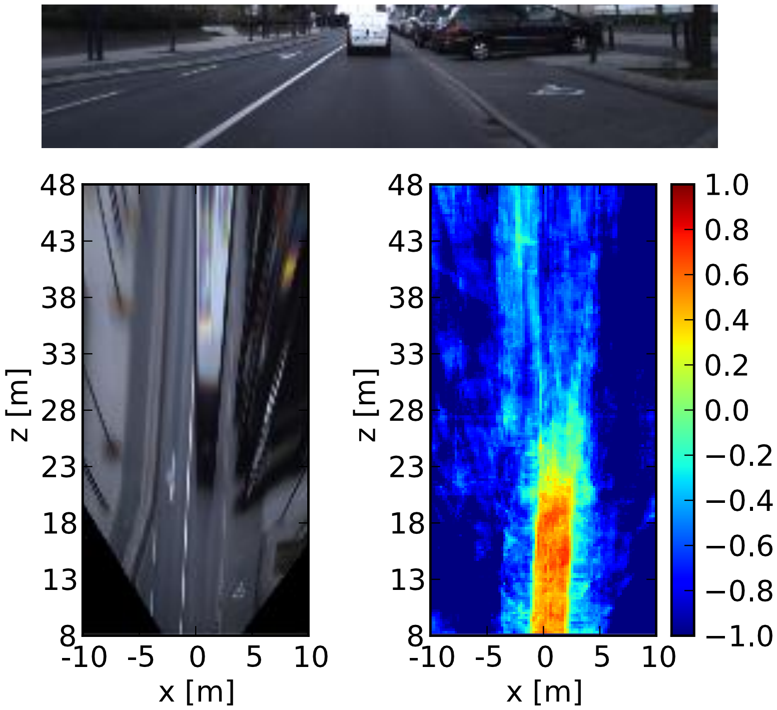

Figure 5.14.: Result of the road terrain classification showing the input image (top), the BEV (bottom left) and the classification result for ego-lane as confidence map (bottom right). Here, the ego-lane is unoccupied up to a distance of 23m.

5.5. Experiments on Road Area Detection

In this section experiments are conducted to assess the performance of the RTDS applied to road area classification. This allows a direct comparison to the results of the LVA-system in Section 4.5.4. For the proposed system RGB images with a resolution of 1024×1280 pixels are used. The metric representation is defined for a range of -10m to 10m in x direction (lateral) and 8m to 48m in z direction (see Fig. 5.14). With a resolution of 5cm, a representation with a block size of 800 × 400 is obtained. Because the GentleBoost road terrain classifier has the ability to select the best out of a large variety of features it is proposed to use a feature setup, which is a trade-off between brute force (take all features one can get) and training duration. A regular grid with 7px offset was used for distributing base points over the metric space. Here the spatial feature generation is set up to have 8 ray orientations $\phi = [-20, 0, 20, 90, 160, 180, 200, 270]$ ($\phi = 0$ is rightwards, counted clockwise positive). As mentioned above selecting good absorption thresholds is significant for the performance of the system. In the example of Figure 5.8, thresholds are selected in an optimal way in order to encode the distance to relevant scene elements. In reality, this selection is of course not trivial. Choosing a high amount of absorption thresholds would be probably the most high-performance parameters but it would also lead to a very high feature order which is computationally complex to handle. Based on parameter optimization experiments, it was found that for a fixed number of absorption thresholds many different threshold combinations lead basically to a similar performance. The absorption thresholds are set to $th = [1.5, 5, 15, 35, 60]$, as this was found to be a reasonable compromise of computational effort and detection quality. Note that a finer graduation would lead to a slightly better detection quality but also more computational effort. In the parameter optimization the same configuration was used for each base classifier. However, it would be more appropriate to alter thresholds and angles for each confidence map.

In order to compare the results of the road terrain detection system to results from the base classifier in Section 4.5 the experiments are conducted on the same innercity datasets (see also Appendix B). Furthermore, a metric baseline is computed for comparison (see Appendix C). This baseline was previously used in Section 4.5. Consequently, the detection results of the RTDS applied to road area classification can be compared with this baseline. In contrast to Chapter 4, all evaluations are carried out in the metric domain only.

For the base road and base boundary classifier the best setup found in Section 4.5 is selected: 18 color, 64 Walsh-Hadamard texture and 20 slow features. The GentleBoost road terrain classifier uses the same setup (100 iterations and 4 tree-splits for the weak learners) as in Section 4.5 so that the results can be compared adequately.

The conducted experiments in the remaining sections will be about an assessment of the single cue performance in Section 5.5.1 (separately for road, boundary or lane marking), followed by an evaluation of the performance when combining the different cues in Section 5.5.2. Section 5.5.3 shows qualitative results of road area detection.

5.5.1. Assessing Single Cue Road Area Classifier Performance

In this experiment the classification of road area is analyzed for the single cues separately. As in Section 4.5, the three inner-city datasets with different weather conditions are taken. Each of the datasets has three folds, one fold is used for testing, and two for training. Note, that one training fold is used for the base classifier and the other for the road terrain classifier. For each of the folds from particular dataset a RTDS is trained with a given configuration. Then, the performance of the resulting classification system is tested on the corresponding test fold. It is important that the road terrain classifiers is trained on unseen data, i.e., images that where not used for training of either the road terrain classifier or the base classifiers. Because the appearance-based detection for unseen images results in noisy base confidence maps, the system has a chance to learn feature combinations that are significant in the presence of noise. This results in a detection performance that has a better generalization, i.e., can better cope with unseen situations.

In Table 5.1 the metric testing quality for the road, boundary and lane marking cue are listed. The subscripts for Q indicate different options. Q_{BL} is the baseline quality, $Q_{default}$ uses only the positive part of the confidence maps, $Q_{+/-}$ is combining positive and negative parts of the confidences (as detailed in Section 5.3.1). Q_{ego} is the result for the ego-SPRAY features (see Section 5.3.2). Multiple indexes stand for a combination of the distinct feature types. The testing qualities are listed for the three inner-city datasets with different weather conditions: 'cloudy', 'sunny', and 'mixed'. For analyzing the generalization, the dataset 'all' contains all conditions from the dedicated datasets.

The results for the road and boundary cue show a substantial increase in test quality when using the ego-SPRAY features together with the combination of positive and negative confidences (rightmost column in Table 5.1). For the lane marking cue a combination of negative and positive confidences is not available, here the best result is obtained for the combination of ordinary and ego-SPRAY features. For the three cues, the results show a good generalization performance, as the testing quality for general training using the 'all' dataset is close to the mean of testing quality for dedicate training. In the following subsections the results for the single cues are discussed.

I. Results for the road cue

The results show that the proposed approach clearly outperform the baseline Q_{BL}^{BEV} for all system configurations using only the road cue. In order to measure the per-

Table 5.1.: Metric quality on test dataset using the single cues for road area classification with different configurations. The metric quality on the training dataset is given in parenthesis.

dataset	Q_{BL}^{BEV} [%]	$Q_{default}^{BEV}$ [%]	$Q_{+/-}^{BEV}$ [%]	Q_{ego}^{BEV} [%]	$Q_{+/-,ego}^{BEV}$ [%]
I. Road cue					
Cloudy	50.2	71.7 (79.1)	74.0 (87.9)	72.3 (80.6)	**75.1** (89.6)
Sunny	36.9	59.2 (65.0)	63.4 (78.7)	60.2 (67.5)	**65.5** (81.4)
Mixed	47.6	66.0 (73.0)	69.9 (84.4)	67.4 (75.9)	**70.7** (86.4)
All	43.0	67.2 (70.7)	69.2 (78.3)	68.0 (71.9)	**70.2** (80.0)
II. Boundary cue					
Cloudy	50.2	69.2 (74.1)	72.8 (85.0)	72.8 (79.1)	**76.2** (88.3)
Sunny	36.9	55.0 (60.9)	60.9 (74.6)	59.5 (66.7)	**64.5** (78.7)
Mixed	47.6	63.2 (72.2)	68.8 (79.9)	68.4 (73.3)	**72.5** (84.3)
All	43.0	66.6 (69.8)	68.8 (76.4)	69.8 (73.4)	**71.6** (79.4)
III. Lane marking cue					
Cloudy	**50.2**	43.6 (53.3)	N/A	44.6 (55.8)	N/A
Sunny	**36.9**	28.8 (36.3)	N/A	30.6 (39.5)	N/A
Mixed	**47.6**	37.3 (46.9)	N/A	38.6 (49.7)	N/A
All	**43.0**	37.9 (42.8)	N/A	40.3 (45.9)	N/A

formance gain of visuospatial classification compared to pure visual classification, a direct comparison of the road cue using SPRAY features and the base classification results from Table 4.4 (see Section 4.5.4) is carried out. By applying the proposed visuospatial classification on the road cue we gain 7.9pp-8.8pp on the distinct inner-city datasets. The highest gain is obtained on the sunny dataset. While on this dataset the performance is the worst in Section 4.5.4, there is apparently a bigger potential for improvements when applying visuospatial classification. On the general dataset including all weather conditions ('all') a gain of 8.1pp is obtained.

II. Result for the boundary cue

Compared to the road cue, Table 5.1 shows slightly better qualities on the 'cloudy' and 'mixed' dataset but slightly worse quality on the 'sunny' dataset for the boundary cue. For the general dataset an increase of 1.4pp compared to the road cue is obtained. This corresponds to an increase of 9.5pp compared to the base classification quality of road area.

III. Result for the lane marking cue

For the lane marking cue the results show worse results compared to the other cues and the base classifier trained on road area. This is obviously caused by the fact that a large part of the dataset consists of road scenes which are unmarked, i.e., the road area is delimited by curbstones on the very left and very right of the road area without explicit road markings. Therefore, using only the lane marking cue for road area classification is apparently not suitable.

5.5.2. Influence of Road Area Classifier Cue Combinations

The following experiment analyzes the combination of the three different cues and the influence on the overall system performance. The results, listed in Table 5.2, show the different combinations for road (R), boundary (B), and lane marking (LM) cues. For the setup, the best result found in Section 5.5.1 is chosen, i.e., using ego-SPRAY features and the combination of positive and negative confidences for the road and boundary cue and the combination of ordinary and ego-SPRAY features for the lane marking cue.

Table 5.2.: Combining road (R), boundary (B) and lane marking (LM) cues for road area classification. The metric quality on the training dataset (if available) is given in parenthesis.

dataset	Q_{BL}^{BEV} [%]	Q_{R+B}^{BEV} [%]	Q_{R+LM}^{BEV} [%]	Q_{B+LM}^{BEV} [%]	Q_{R+B+LM}^{BEV} [%]
Cloudy	50.2	79.0 (93.2)	77.5 (91.7)	77.8 (90.8)	**79.3** (94.3)
Sunny	36.9	69.8 (84.7)	67.6 (82.1)	68.0 (80.4)	**70.3** (86.6)
Mixed	47.6	75.2 (91.0)	73.3 (88.9)	74.5 (87.0)	**76.3** (92.2)
All	43.0	75.1 (82.7)	73.1 (79.4)	74.6 (80.1)	**75.9** (84.3)

The results show that a combination of road and boundary cue for road area classification is beneficial because the testing quality improves by 2.7pp-4.3pp for the dedicated datasets and 3.5pp for the general dataset. The best result is obtained by the combination of all cues, although the increase by adding the lane marking cue is rather low. On the general training dataset, the system outperforms the baseline by approximately 33pp and the local visual classification by approximately 14pp (cf. Section 5.5).

5.5.3. Qualitative Results

In addition to the evaluations carried out before, this section illustrates the detection performance by means of qualitative results in order to discuss further advantages and weaknesses of the proposed approach which can not be derived from evaluation metrics. Exemplary qualitative results for road area detection in challenging scenes with various asphalt types, curbstones delimiting the road area and shadows causing strong appearance changes are depicted in Figure 5.15. The detection result are generated using the most general dataset ('all') as it contains all of the analyzed appearance conditions. The original images (without detection results) can be found in Appendix B.2.

The depicted examples in Figure 5.15 correspond to the examples for base road classification in Figure 4.10 (see Section 4.5.4). Therefore, the results, which are illustrated by overlaying the image with the binarized RTDS output, can be directly compared. The results show that the increase in quality compared to the base road classifier as discussed in Section 5.5.2 is visible in the qualitative detection results as well. The depicted results contain less false positive detections, e.g., on sidewalks and cars, compared to the base road classifier (see, e.g., XII in Figure 5.15). Furthermore, the results are less noisy, i.e., small erroneous regions on the road area are compensated, and the overall shape better reflects the actual course of the road area.

Another important aspect is that the road area delimiters may be occluded by a parking car (see XI in Figure 5.15). An approach based on delimiter detection relying solely on the explicit detection of lane markings and curbstones (as, e.g., proposed by Danescu and Nedevschi, 2011) may fail in such situations. The depicted examples show that vehicles generally do not cause false positive detections. Therefore, the proposed method apparently handles occlusions far better than delimiter based approaches.

The presence of strong shadows is known to be very challenging for an appearance

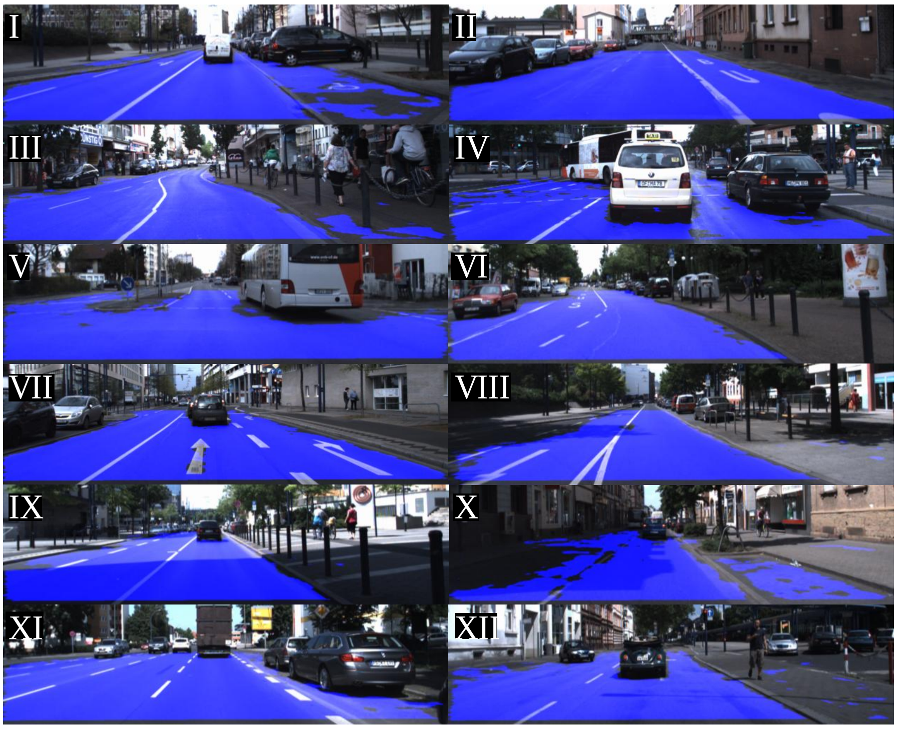

Figure 5.15.: Example scenes with depicted road area detection result marked in blue. The images are identical with the ones used in Figure 4.10. The raw images can also be found in Appendix B.2.

based road area detection (see Michalke et al., 2009). As can be seen, e.g., by comparing example VIII and IX in Figure 5.15 and Figure 4.10 (see Section 4.5.4), the proposed system clearly improves the detection of road area in the presence of shadows compared to pure visual based detection. However, example X shows a situation where the bigger part of the opposing lane is not correctly detected. This is apparently caused by the worse detection of the base road and boundary classifiers.

5.6. Experiments on Ego-Lane Detection

Beside road area detection visuospatial classification allow detecting categories that are visually similar but spatially different such as the ego-lane. In the following, the detection performance of an RTDS applied to ego-lane classification is evaluated. As discussed earlier, a direct comparison of visuospatial classification and local visual appearance based classification is not possible for ego-lane classification. Therefore, only a baseline for ego-lane is used for comparison in the following. This baseline can be computed in the same way as for road area using ground truth annotations. More details on that can be found in Appendix C.

Training data for the conducted experiments is generated as described in Section 5.4. The parameters for the system are kept constant and are chosen in the same way as in Section 5.5.

In the following, an assessment of the performance for the distinct cues will be detailed in Section 5.6.1. The experiments in Section 5.6.2 evaluate the combination of the cues. Subsequently, Section 5.6.3 discusses qualitative results for ego-lane detection.

5.6.1. Assessing Single Cue Ego-Lane Classifier Performance

The results for the different cues, i.e., road, boundary, and lane marking cue are given in Table 5.3. As in Section 5.5, the subscripts for Q indicate different options. Q_{BL} is the baseline performance, $Q_{default}$ uses only the positive part of the confidence maps, $Q_{+/-}$ is combining positive and negative parts of the confidences. Q_{ego} is the result for ego-SPRAY features. Multiple indexes stand for a combination of feature types.

The results in Table 5.3 show that the system configurations influence the performance in the same way as for road area detection (cf. Section 5.5), i.e., using positive and negative confidence maps together with ego-SPRAY features lead to the best performance. Also the results of the single cues outperform the baseline. A good generalization is achieved which can be seen by observing the results on the general dataset ('all') and comparing it with the dedicate training performance. In the following subsection, results for the three single cues are further detailed.

I. Results for the road cue

For the road cue the results show a clear improvement compared to the baseline with an increase of metric test quality of 6.9pp-12.8pp on the dedicated datasets and 10.2pp on the general dataset. Furthermore, having a closer look at the particular configurations, one can see that the influence of the ego-SRPAY features on the road cue are comparably lower than the inclusion of positive and negative confidences.

Table 5.3.: Metric quality on test dataset using the single cues for ego-lane classification with different configurations. The metric quality on the training dataset (if available) is given in parenthesis.

dataset	Q^{BEV}_{BL} [%]	$Q^{BEV}_{default}$ [%]	$Q^{BEV}_{+/-}$ [%]	Q^{BEV}_{ego} [%]	$Q^{BEV}_{+/-,ego}$ [%]
I. Road cue					
Cloudy	42.7	40.8 (59.5)	49.5 (86.0)	40.0 (58.9)	**49.6** (87.9)
Sunny	37.5	39.0 (57.0)	49.1 (88.5)	40.4 (60.2)	**50.3** (90.5)
Mixed	41.5	38.7 (50.2)	47.8 (84.8)	38.4 (52.3)	**49.7** (87.5)
All	39.1	44.1 (67.0)	47.9 (85.0)	44.6 (69.9)	**49.3** (88.4)
II. Boundary cue					
Cloudy	42.7	45.8 (56.3)	55.6 (88.7)	47.4 (65.4)	**57.9** (91.3)
Sunny	37.5	40.0 (53.8)	51.6 (87.8)	45.9 (64.5)	**54.3** (91.6)
Mixed	41.5	40.1 (54.0)	49.1 (84.4)	43.3 (61.0)	**52.0** (88.4)
All	39.1	45.0 (67.0)	50.0 (88.5)	49.5 (74.0)	**54.0** (91.2)
III. Lane marking cue					
Cloudy	42.7	46.5 (60.1)	N/A	**48.9** (65.2)	N/A
Sunny	37.5	39.2 (53.7)	N/A	**40.7** (60.1)	N/A
Mixed	41.5	43.4 (54.9)	N/A	**45.7** (61.0)	N/A
All	39.1	44.1 (65.8)	N/A	**46.2** (70.9)	N/A

II. Result for the boundary cue

Compared to the road cue, using only the boundary cue, a significant increase for the three dedicated datasets especially for the cloudy dataset with 8.9pp is visible. For the merged dataset an increase of 4.7pp compared to the road cue is obtained. Interestingly, the gain of applying the configuration with positive and negative confidences and ego-SPRAY feature (subscript: $+/-, ego$) compared to the default configuration is much higher than for all previous experiments (including the experiments in Section 5.5.1).

III. Result for the lane marking cue

The result using only the lane marking cue is worse than both the other cues. However, on all datasets it outperforms the baseline quality (which was not the case in the experiments for road area detection in Section 5.5.1).

5.6.2. Influence of Ego-Lane Classifier Cue Combinations

Aim of the experiments in this section is to show that the performance increases when combining multiple cues for the RTDS applied to ego-lane classification. The results applied to the dataset for the three possible combinations of road (R), boundary (B) and lane marking (LM) cue are given in Table 5.4.

Table 5.4.: Combining road (R), boundary (B) and lane marking (LM) cues for ego-lane classification. The metric quality on the training dataset (if available) is given in parenthesis.

dataset	Q^{BEV}_{BL} [%]	Q^{BEV}_{R+B} [%]	Q^{BEV}_{R+LM} [%]	Q^{BEV}_{B+LM} [%]	Q^{BEV}_{R+B+LM} [%]
Cloudy	42.7	59.0 (97.9)	67.0 (97.6)	67.5 (97.7)	**67.5** (99.4)
Sunny	37.5	55.0 (97.9)	60.2 (97.3)	60.3 (96.9)	**61.3** (99.1)
Mixed	41.5	53.2 (97.1)	61.2 (96.5)	**61.4** (96.2)	61.2 (98.8)
All	39.1	56.0 (95.3)	60.7 (95.6)	61.2 (95.2)	**62.1** (97.3)

The results show a significant increase in metric quality for any combination compared to the baseline. With respect to the best single cue results, an increase of 0.7pp - 2.1pp for the dedicated and 2pp for the general training set is obtained for the combination of base road and base boundary cue. From this we see that the base road cue plays a secondary role for the visuospatial classification of ego-lane. This can also be seen for the "mixed" dataset where the combination of boundary and lane marking cue performs slightly better than the the combination of all the three cues. However, the general performance of using all cues is better than all other combinations. Compared to the best single cue it shows an increase of 8.1pp on all test datasets. Compared to the baseline an increase of 23pp is obtained.

5.6.3. Qualitative Results

This section illustrates the detection performance with concrete examples and qualitative results in order to discuss further advantages and weaknesses of the proposed ego-lane detection. Exemplary qualitative results for ego-lane detection in scenes with marked and unmarked lanes are depicted in Figure 5.16. In Chapter 2 it was found that ego-lane detection approaches that handle such situation are rare. The detection result are generated using the most general dataset ('all') as it contains all of the analyzed appearance conditions.

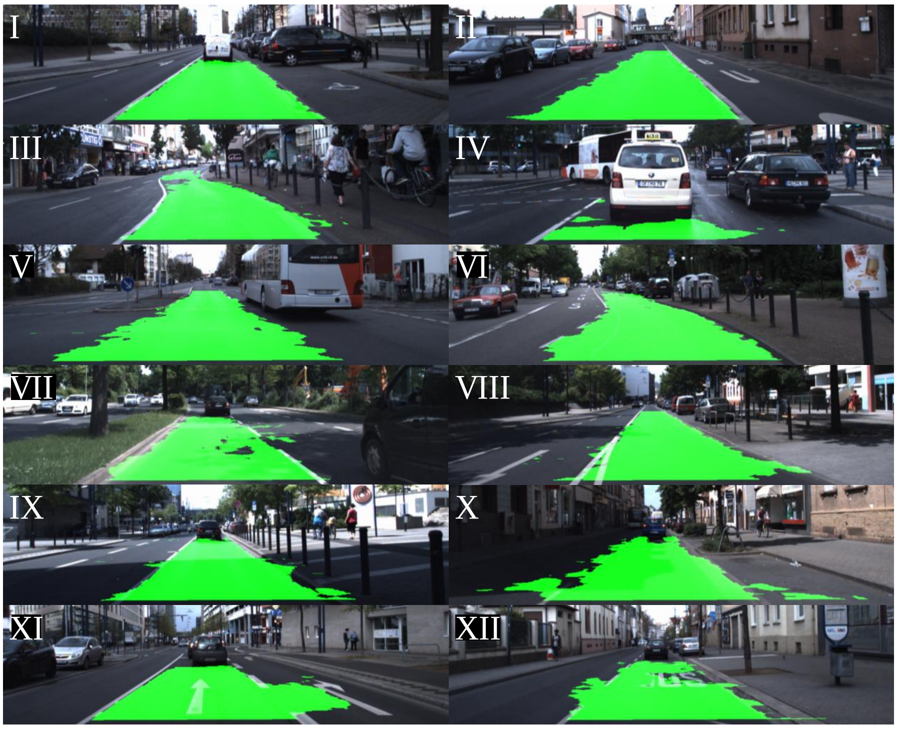

Figure 5.16.: Example scenes with depicted ego-lane detection result marked in green. The raw images can also be found in Appendix B.2.

The depicted scenes show that the proposed method trained on the ego-lane can handle various delimiter types such as lane markings, and flat curbstones on both sides (see VII and VIII in Fig. 5.16). Transition from light to shadow are very challenging due to the changing appearance of the road area (see VII and IX in Fig. 5.16). The proposed method enables a reliable detection of the ego-lane in the presence of shadows on the road area without employing a specialized shadow compensation. This is for instance proposed by Álvarez et al. (2013), where an illumination invariant image transformation is applied.

In some situations large marked areas on the ego-lane such as road signs, which are not explicitly represented in the system, may cause false negative detection because apparently those regions can not be distinguished from lane delimiters such as lane markings (see XII in Fig. 5.16). Example XI shows another typical error behavior of the system which is a flowing out effect of the detection from the ego-lane in the gap between lane markings.

5.7. Summary

This chapter introduced the concept of visuospatial classification that has been developed in the course of this thesis. The core of the approach is the incorporation of the spatial layout of local visual appearance. Complemented by a classifier, a generic approach for visual and spatial analysis of the driving environment was derived which can be applied to multiple classification tasks.

A spatial layout computation is applied on metric confidence maps of multiple local visual appearance properties such as road, boundary, or lane marking. The utilization of metric representations is a good basis for spatial layout computation because compared to perspective representations, they have the advantage that perspective changes, e.g., of the road geometry, are compensated. Beyond that it was found, that local visual appearance in an image allows only limited inference about a certain class due to the visual ambiguity of different road terrain categories (see Chapter 4). The incorporation of the spatial layout complements the visual information and improves the detection of road terrain categories.

In order to compute the spatial layout of local visual properties SPatial RAY (SPRAY) features were proposed. SPRAY features capture the spatial layout by means of a directed ray-based integration process with respect to base points located at fixed positions in metric space, gathering confidence information of local visual properties. With this method, information of road environments can be transferred into a visuospatial feature representation. Complemented by a classifier it can be used as Road Terrain Detection System (RTDS).

The proposed RTDS is a combination of the base classification presented in Chapter 4 with the spatial layout computation. By applying machine learning concepts, a road terrain classifier is trained for a specific road terrain category using ground truth annotations. Possible application for RTDS is the utilization for ADAS (as motivated in the beginning of this thesis) or the construction of higher-level spatial representations of driving scene elements as proposed by Kastner et al. (2010). One fundamental aim of RTDS is to reliably extract the course of the road area or ego-lane even in challenging inner-city scenarios including various asphalt appearances and multiple road/lane delimiters, and occlusion of delimiters.

A series of conducted experiments for road area classification shows that incorporating the spatial layout helps to compensate errors. For instance, in situations where the visual properties of an image region does not indicate the presence of road area, the combination of visual features with its spatial layout (encoded by the proposed SPRAY features) makes the detection possible. A comparison to the results in Chapter 4 showed that the combination of visual and spatial features clearly outperforms local visual appearance based classification. The approach reduces false detections due to the presence of shadows or other appearance changes, such as dirt, gully covers, or broken asphalt, on the road area.

Approaches for ego-lane detection, as detailed in the related work chapter, typically apply an explicit model and temporal tracking of the parameters (see Section 2.1.1). In contrast to these, the experiments show that the proposed approach enables to implicitly learn a model that captures the spatial characteristics of ego-lane. In contrast to the road area, the ego-lane is visually more challenging to detect because the local visual appearance of all lanes is typically the same. The conducted experiments show that incorporating the spatial layout makes a clear distinction available. This results in a generic and pure visual bottom-up ego-lane detector that handles challenging urban scenarios including marked and unmarked lanes and operates on single images without employing temporal integration.

The following chapter will discuss visual ego-vehicle lane assignment, a further application of the proposed visuospatial classification.

6. Visual Ego-Vehicle Lane Assignment using Spatial Ray Features

One highly relevant aspect of the environment is information about how many lanes there are and on which lane the ego-vehicle is currently driving (see Kuehnl et al., 2013).

Current vision-based ADAS perform lane departure warning and lane keeping assistance by only using information about the ego-lane. However, information on the relative location of the ego-lane on a multi-lane road, i.e., the ego-lane index, is not available from these systems. Identifying the ego-lane index is not straightforward. Without information from a high-precision GPS signal, as used, e.g., in the DARPA Challenges, many approaches rely on combining image processing with detailed digital maps (see e.g. Mattern et al., 2010; Popescu et al., 2012).

In this thesis, an approach was developed which assigns the ego-vehicle to a lane without having any prior knowledge, e.g., on the total number of lanes. The approach is purely vision based and estimates the number of lanes adjacent to the ego-lane in both directions. This is realized by combining the system presented in Chapter 5 with an ego-lane index classification.

An easily realizable benefit of the proposed method is to provide the ego-vehicle's lane position to a navigation system which allows for more adequate routing instructions as motivated in the introduction of this thesis. More importantly, for future ADAS operating on highways, information about other lanes provides context to predict behavior of other drivers. Thus, visuospatial classification can contribute to a low-cost solution for the mentioned applications because it solves the problem by relying solely on monocular image processing.

The remainder of this chapter is organized as follows. Section 6.1 gives an overview about other approaches for ego-vehicle localization. Details on the proposed approach and the system architecture follow in Section 6.2. Subsequently, the details of the lane index classification are discussed in Section 6.3. Then, experiments evaluating the detection performance on highways are presented in Section 6.4. Finally, Section 6.5 provides a summary of this chapter.

6.1. Related Work for Ego-vehicle Localization

For safety applications such as lane keeping or lane departure warning GPS sensors are not accurate enough. Therefore, a lot of publications tackle the problem of improving longitudinal ego-vehicle localization on a digital map which can be achieved by fusing visual sensors with GPS (see e.g. Chausse et al., 2005; Szczot et al., 2010). In contrast to those approaches, lateral ego-vehicle localization, i.e., assigning the ego-vehicle to a specific lane, requires even more accuracy.

For future ADAS operating on highways, information about other lanes provides context to predict behavior of other drivers. For example, knowing the ego-lane index and encountering an entrance lane allows to predict cut-in maneuvers by entering road users (see Bonnin et al., 2012).

In previous work, a number of sensing techniques have been employed to achieve lateral ego-vehicle positioning. A classical approach that was also extensively used in the DARPA challenges (see Buehler et al., 2009) is the localization on a highly detailed digital map using costly high-precision GPS receivers. There is also promising research on affordable and accurate GPS techniques by Knoop et al. (2012). However, for safety reasons additional environment sensing, e.g., by camera, will be needed for future ADAS. As alternative sensor technologies, inter-vehicle communication has been used for lane positioning (cf. Dao et al., 2007) as well as a high-precision Inertial Measurement Unit (IMU) in combination with a terrain map (cf. Dean and Brennan, 2009). Furthermore, Kloeden et al. (2011) applied radio-frequency-based landmarks for localization on inner-city intersections.

If a vehicle is equipped with a radar system that can detect several preceding vehicles on the same lane (e.g., using a tunneling radar), vehicles can be grouped into a convoy track, indicating the track driven by the vehicles (see Weiherer et al., 2012). Depending on the opening angle of the radar, this approach implicitly provides information about the number of lanes (the different convoy tracks) and the own position. Applying this approach for localization would require a certain traffic density, i.e., there is at least one vehicle on each lane.

Avoiding the need for extra and costly sensor hardware, Mattern et al. (2010) apply an extended digital map that contains the precise locations of the road and lane delimiters (see Fig. 6.1). These landmarks are used in a generative approach where a hypothetical sensor image is generated from the landmarks and subsequently matched to the actual camera data to obtain precise positioning. Popescu et al. (2012) also use GPS and an extended digital map containing road infrastructure information (e.g., number and width of lanes, lane delimiter types). A Bayesian network is used for positioning on the lane level where the network parameters are adapted to the current map information. A visual detection of the delimiter types of the ego-lane serves as input data together with information about the dynamic environment (position and movement of other vehicles).

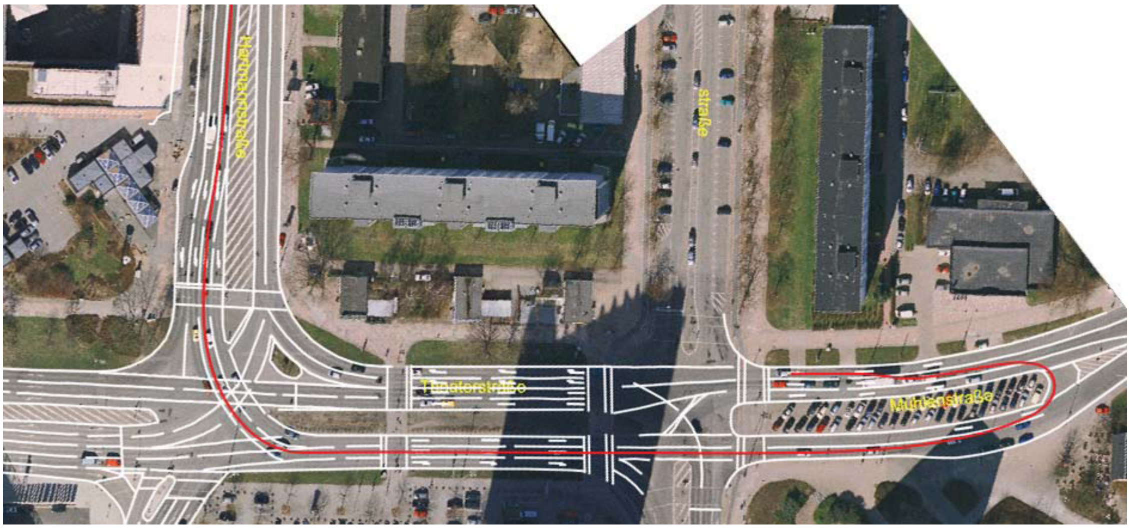

Figure 6.1.: Localization of the ego-vehicle in an extended digital map. The extended data contains lane delimiter and road signs. The red line shows the mapping of the ego-vehicle path into the map. The picture is extracted from Mattern et al. (2010).

Using a navigation map and a fixed road width, Konrad et al. (2012) convert this map data into virtual road border information. Road borders are extracted in the form of lane markings, color differences, or obstacles and matched to the virtual road borders to perform the positioning. The approach relies on successful border detection and seems to be limited to two-lane roads unless an extended digital map is used.

Hold et al. (2010) presented a vision-based method for detecting if the ego-vehicle is on an exit lane. The approach is based on extracting the spatial frequency of the closest lane marking to the left side. While a low spatial frequency indicates an usual highway lane marking, a high spatial frequency indicates being on an exit lane.

6.2. System Description

The proposed approach for lane assignment is a pure vision-based approach which aims at capturing the free-space in the road scene in order to detect the lane index in both lateral directions. Therefore, it can be seen as a complementary contribution to the related work.

Instead of estimating the total number of lanes, the left and right ego-lane index can be separately identified. A combined system output is illustrated in Figure 6.2, the boxes in the top right corner show the assigned ego-lane indexes and implicitly the total number of lanes. The black box denotes the ego-lane. In this example the left ego-lane index is two, meaning that there are two more lanes to the left. Because of the emergency lane, the right ego-lane index is one.

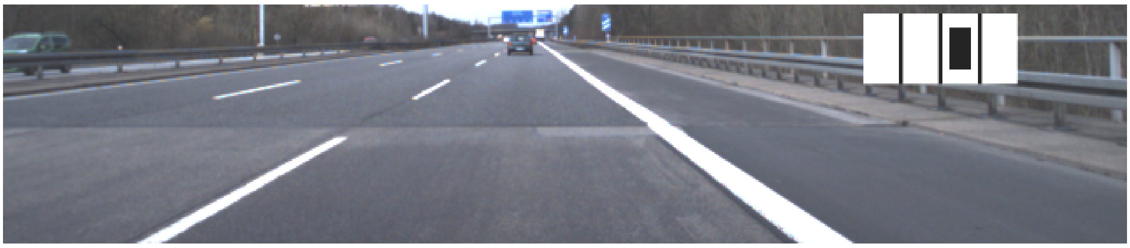

Figure 6.2.: Demo image showing a typical scene on a German highway. Here the left ego-lane index is 2, and the right ego-lane index is 1.

The overall system for visual ego-vehicle lane assignment using spatial ray features, depicted in Figure 6.3, consists of three stages: Firstly, base classifiers capturing local visual properties (which was discussed in Chapter 4). Secondly, a SPRAY feature generation capturing spatial properties of the road scene as discussed in Chapter 5. And thirdly, an ego-lane index classification.

Essentially, the general system architecture does not differ from the visuospatial classification system which was presented in Chapter 5. Instead of applying the system for road terrain classification (as in Section 5.4), an ego-lane index classification is proposed. In the next section, the lane index classification will be explained detailedly.

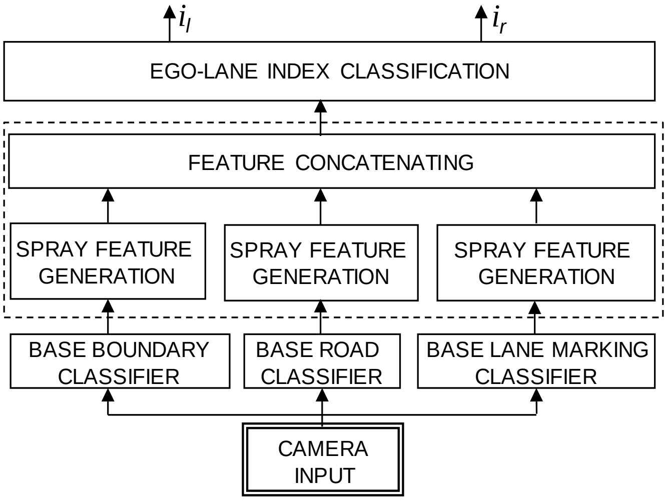

Figure 6.3.: Block diagram showing the main three stages of the proposed system: Base classification, SPRAY feature generation and ego-lane index classification. Output is the left and right ego-lane index (i_l, i_r).

6.3. Lane Index Classification

The system makes use of the visuospatial representation (see Section 5.3.4) as discussed in Chapter 5. However, there are some differences that will be detailed in the following.

The problem of finding the ego-lane index and the number of lanes can be reformulated into an ego-lane index classification to the left and respectively to the right. The ego-lane index $i_{l,r} = [0..n]$ is given by a discrete number, where zero indicates the ego-lane is the left- or rightmost lane, one indicates second lane and so on. Figure 6.4 depicts an example for a three lane road ($n = 2$) showing all combinations of left and right ego-lane indexes.

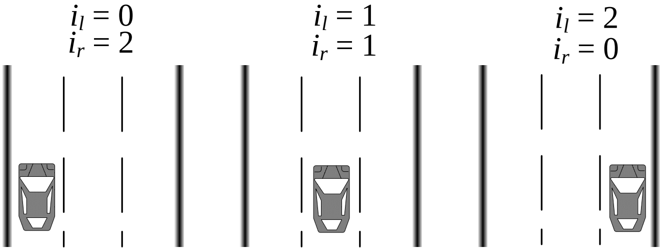

Figure 6.4.: Three possible combinations of left (i_l) and right (i_r) ego-lane indexes on the three lane road.

The utilization of the spatial layout of the road, boundary and lane markings (as depicted in Fig. 6.3) aims at capturing the lateral extent of the road area and the amount of lanes in a certain direction. Therefore, the presented lane index classification approach could also be used for the application to inferring the number of unoccupied lanes. However, for the given task, in case a neighboring lane is occupied by a vehicle the approach may underestimate the lane index to the corresponding direction. Thus, combining the approach with a method for detecting vehicles to have an estimate which lanes are occupied, or digital map data to know the total amount of lanes, would allow better inference.

For the conducted experiments a maximum ego-lane index of three (both sides) was considered. In order to capture the extent of a four-lane highway even if the ego-vehicle is on the left- or right-most lane our system uses a metric representation for computing SPRAY features with a lateral range from $-16m$ to $16m$. In longitudinal direction a range from $5m$ to $30m$ (measured from the front bumper of the car) was analyzed. With a discretization of $0.1m$ per pixel we obtain a BEV-resolution of $250 \times 320px$.

For the base classification the same setup as detailed in Section 5.2 is used. The base classifiers use pre-learned parameters, obtained from the combined inner-city dataset (see Appendix B.2). Because the datasets comprise versatile visual appearance conditions caused by the different weather/lighting (overcast, sunny and mixed weather) the parameters are suitable for general practice.

In contrast to the road terrain detection system from Chapter 5, a single base point for SPRAY feature generation, set to a distance of $11m$ centered in front of the vehicle is used. Because in the near range the lateral extent of the road is tough to capture and in the far range the SPRAY features are more noisy, this distance was found to be a good compromise. Furthermore, for driving on highway or motorways one can assume that this metric location is always unoccupied.

Obviously, for ego-lane index classification laterally oriented SPRAY features of the road scene are more relevant than those measured along the driving direction. Thus, SPRAY features are configured with 6 rays, 3 to the right $[-30°, 0°, 30°]$ and 3 to the left $[150°, 180°, 210°]$. The $\pm 30°$ rays are useful in order to have a higher likelihood to measure the full lateral extent of the road even if cars are driving on neighboring lanes. Beyond that, the same configuration as in Chapter 5 was chosen. Consequently, five thresholds are used.

Therefore, the SPRAY features encode the scene with 30 distance values per confidence map. We obtain a feature vector length 150 plus 5 ego-rays from the 5 confidence maps (2× road, 2× boundary, and 1× lane marking).

A GentleBoost classifier is used to automatically select good features and obtain class separation. The basic problem is a multiclass classification. Consequently, the number of classes is equal to the number of ego-lane indexes. Note, that the GentleBoost method detailed in Section 4.3 is originally designed for binary classification problems. The strategy for applying it to the multi-class problem to obtain a confidence for each class is described by Friedman et al. (2000). Finally a decision on the ego-lane index can be taken based on the class with the maximum confidence. In Figure 6.5 two examples for the ego-lane index classification are depicted. The upper image shows the correct estimation on a five-lane highway. Note, that the system is not able to distinguish an emergency lane from an usual traffic lane. The lower image shows a challenging situation because of hardly visible lane markings (between leftmost and second leftmost lane) and cast shadow caused by the fence. The proposed system can handle these situations, while a pure lane marking based approach would probably fail.

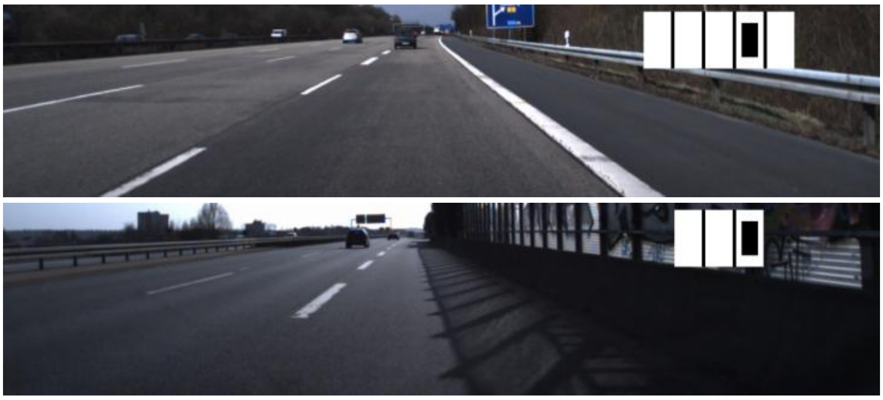

Figure 6.5.: Exemplary classification results where estimated ego-lane indexes are illustrated with white boxes. The black box denotes the ego-lane.

6.4. Evaluation

Two highway dataset recorded with 20 fps are used to evaluate the performance of lane index classification. The first dataset ”Highway 1”, capturing approximately 30 km of driving is evaluated using four-fold cross validation. Here, for every fold $\frac{3}{4}$ of the dataset was used for training the GentleBoost classifier (5 tree-splits and a maximum of 100 iterations). In order to show the generalization capabilities of the approach a second dataset ”Highway 2” (11 km) is used for testing. Consequently, no additional training is applied. The learned classifiers from ”Highway 1” are simply reused. Note that "Highway 2" contains more parts with dense traffic than "Highway 1" but both exhibit mostly low traffic density.

For both datasets the left and right ego-lane index was manually annotated for every frame. In situations of unclear class membership, e.g., lane changes or other unclear situations, the frame was not used for evaluation and training. The variety of different situations in the dataset go from ramps ($i_{l,r} = 0$) up to four-lane highways ($i_{l,r} = [0..3]$). Note that traffic, entrance, and emergency lanes are handled equally by the proposed system.

The datasets were recorded at daytime with mostly sunny weather conditions. Especially cast shadows (see lower image in Figure 6.5) can cause noisy behavior of the base classifiers. Furthermore, various asphalt types require to have very general base classifiers.

In the following sections, conducted experiments for the two highway datasets for left and right ego-lane index classification are separately evaluated. In all experiments SPRAY features were only computed for a single base point as there was no significant improvement when using multiple base points. This stands in contrast to what was necessary for road terrain detection in previous chapters.

6.4.1. Left and Right Ego-Lane Index Classification (Highway 1)

The evaluation of the ego-lane index classification on "Highway 1" was carried out on 19884 frames for left and 18522 frames for right using 4-fold cross validation. The distribution of the ego-lane index classes is given in Table 6.1.

Table 6.1.: Class distribution of ego-lane indexes.

	Highway 1		Highway 2	
class	left	right	left	right
0	9803	7358	2290	3661
1	6922	7776	3111	3813
2	2756	3287	920	400
3	403	101	34	0
$\sum$	19884	18522	6355	7874

The confusion matrices for the two classifiers are depicted in Table 6.2. The proposed method obtains similar results for both sides with recognition rates of $R_{l,1} = 97.64\%$ (left) and $R_{r,1} = 97.62\%$ (right) for all classes. Having a closer look

Table 6.2.: Confusion matrix for ego-lane index classification (Highway 1). Recognition rates are given in parenthesis.

Left				
	0	1	2	3
0	9639 (98.3 %)	137 (1.4 %)	21 (0.2 %)	6 (0.1 %)
1	58 (0.8 %)	6786 (98.0 %)	75 (1.1 %)	3 (0.0 %)
2	44 (1.6 %)	78 (2.8 %)	2631 (95.5 %)	3 (0.1 %)
3	0 (0.0 %)	19 (4.7 %)	25 (6.2 %)	359 (89.1 %)
Right				
	0	1	2	3
0	7302 (99.2 %)	34 (0.5 %)	22 (0.3 %)	0 (0.0 %)
1	123 (1.6 %)	7567 (97.3 %)	86 (1.1 %)	0 (0.0 %)
2	107 (3.3 %)	69 (2.1 %)	3111 (94.6 %)	0 (0.0 %)
3	0 (0.0 %)	0 (0.0 %)	0 (0.0 %)	101 (100 %)

to the confusion matrix one finds the minimum recognition rate per class for three lanes to the left with 89.1% and for two lanes to the right with 94.6%. Table 6.1 shows that the higher lane indexes are underrepresented in the dataset. Therefore, the amount of frames correspond only to a couple of different scenes which results in a strong fitting of the classifiers to the presented training samples.

6.4.2. Generalization Experiments (Highway 2)

In order to analyze the generalization capabilities of the classification system a second dataset, recorded on a different day with a different route, was used. Instead of using one single of the trained classifier from the previous step, a majority decision of all four classifiers is computed. The ego-lane index result is obtained by summing up the per-class confidence, and select the class with the maximum argument. As shown in the confusion matrices in Table 6.3 a high generalization performance for the first two ego-lane indexes (class 0 and class 1) is obtained. The overall recognition rates on the ”Highway 2” dataset are $R_{l,2} = 95.04\%$ to the left and $R_{r,2} = 93.80\%$ to the right. One cause for the worse performance for ego-lane index class 2 and 3 is the higher amount of traffic on ”Highway 2”. Additionally we see in Table 6.1 that the ego-lane index class 2 and 3 are underrepresented, so that

Table 6.3.: Confusion matrix for ego-lane index classification (Highway 2). Recognition rates are given in parenthesis.

Left				
	0	1	2	3
0	2272 (99.2 %)	17 (0.7 %)	1 (0.0 %)	0 (0.0 %)
1	101 (3.2 %)	3008 (96.7 %)	2 (0.1 %)	0 (0.0 %)
2	12 (1.3 %)	147 (16.0 %)	760 (82.6 %)	1 (0.1 %)
3	0 (0.0 %)	0 (0.0 %)	34 (100 %)	0 (0.0 %)
Right				
	0	1	2	3
0	3591 (98.1 %)	29 (0.8 %)	41 (1.1 %)	-
1	215 (5.6 %)	3500 (91.8 %)	98 (2.6 %)	-
2	21 (5.2 %)	84 (21.0 %)	295 (73.8 %)	-
3	-	-	-	-

the classifiers do not generalize that well for these classes. One would need to use a more balanced dataset in terms of class occurrence in the training for obtaining better generalization.

6.4.3. Discussion

Exemplary classification results are depicted in Figure 6.6. The examples I-XVI show correct classification. Even though, very high recognition rates are obtained, there are several reasons for false detections: Although, the SPRAY-feature ray casting strategy increases the likeliness of capturing the full extent of the road, in some cases vehicles occluding a significant part of the road lead to false detections. This can be seen in example XVIII of Figure 6.6 where the ego-lane index is underestimated. This could be compensated by incorporating vehicle detections, e.g., using radar or camera, into the system as proposed by Popescu et al. (2012).

A situation where the ego-lane index is overestimated is shown in Figure 6.6 XVII, this effect is caused by the road-like appearance of the barrier on the right. This problem might be solvable with an approach capturing the height over ground in the scene (e.g., by stereo).

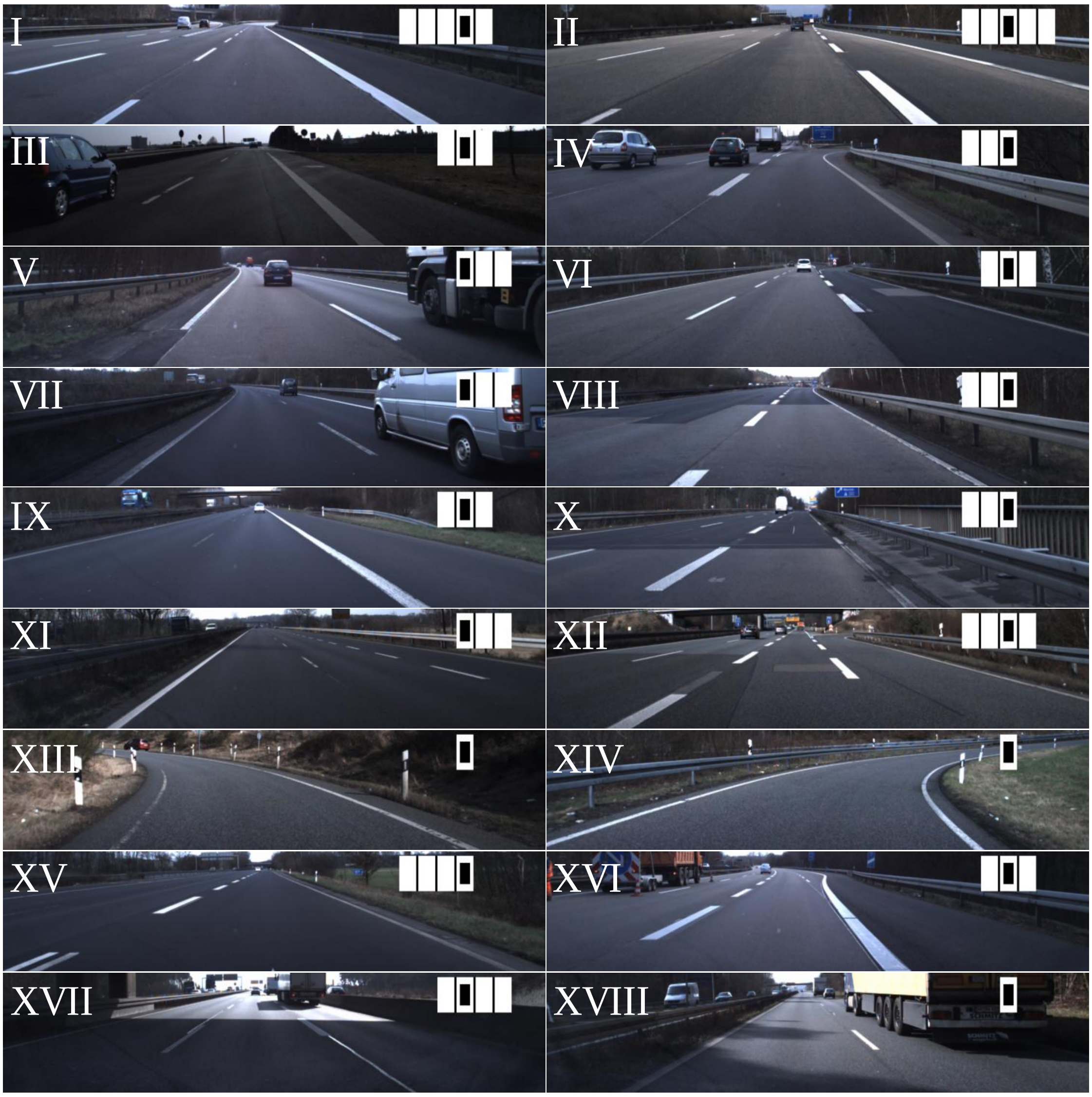

Figure 6.6.: Exemplaric classification results in multiple scenes.

Experiments on temporal smoothing of the results were not successful. Applying a median of the last 5 detections as a simple temporal integration did not significantly improve the results. This is because errors occur in larger temporal blocks, e.g., during an overtaking maneuver. A noisy behavior of the lane index classifier is not observed. However, a more sophisticated temporal integration may result in a performance increase.

6.5. Summary

This chapter demonstrated that SPRAY features provide an effective representation of spatial environments like highway roads. Using a purely vision-based approach operating on individual images, the resulting ego-lane index detection provides recognition rates of 93%-97%. Towards the application in real ADAS, a combination of the presented approach with other information sources will be needed. Most importantly, having vehicle detections from stereo, radar, or object recognition (see, e.g., Kastner et al., 2011) will allow to reduce the miss-detections occurring during take-over maneuvers and to ensure correct operation also in dense highway traffic. In addition, incorporating temporal filtering and lane change notifications from lane tracking, turning lights, a.s.o., will ensure robustness during continuous driving. Even though there is still some work to do in order to incorporate this approach into an ADAS (e.g., a prediction of cut-in maneuvers as proposed by Bonnin et al., 2012) the simplicity of the feature calculation and lane index classification are promising aspects of the proposed approach, supporting its implementation in low-cost hardware of a future ADAS.

7. State-of-the-Art Comparison and Future Embedding of the Concept

One application of the approach developed in this thesis is road terrain detection in challenging inner-city scenarios. In order to evaluate the influence of particular components and parameters of the Road Terrain Detection System (RTDS), Chapter 4 and 5 used the classification quality. Beyond that, a baseline comparison served as a measure to rank the overall performance of RTDS with respect to a naive approach. In order to assess the performance more generally, a comparison to state-of-the-art approaches is necessary. To this end, Section 7.1 compares the proposed RTDS applied to road area detection (see Chapter 5) with two state-of-the-art approaches.

It was found that the proposed system performs well even on datasets with challenging weather and lighting conditions. However, for applying the approach in real ADAS an even higher performance and robustness is required. This issue was already discussed in the related work (see Chapter 2), for ADAS operating in varying outdoor conditions generalization is a basic necessity. To this end, Section 7.2 outlines an adaptive concept to extend RTDS which is a combination of an arbitrary number of offline-learned classifiers on two stages. An important role plays the model switching based on scene context. This allows selecting different classifiers with respect to current weather and lighting conditions and to current spatial conditions. However, the implementation of this important concept is beyond the scope of this thesis and remains future work. After the two main sections in this chapter follows a summary in Section 7.3.

7.1. Road Area Detection Performance Comparison with State-of-the-Art

This section provides a comparison of the method presented in Chapter 5 with two state-of-the-art methods which have been discussed in the related work (see Chapter 2). The methods are the ones by Hoiem et al. (2007) and by Álvarez et al. (2012).

Geometric Context (GC, i.e., the method by Hoiem et al., 2007) is a very generic approach that can be applied on arbitrary scenes using a very general parameter set. The method basically detects all horizontal surfaces on the ground which is not always corresponding to the road area. Therefore, a direct comparison using the road area ground truth as basis for the evaluation is apparently unfair, but it serves as a good baseline. The executable for generating the *Geometric Context* is available on their website[1].

The method by Álvarez et al. (2012) is a very recent approach. This method combines a generic offline-trained Convolutional Neural Network (CNN) and an adaptive online classifier. Note that the CNN is, in contrast to the procedure detailed in the related work (see Section 2.2), trained using the provided training ground truth for road area. The online classifier adapts an internal representation based on the current appearance of the road area captured from a training window and is therefore an adaptive classification approach. The results for the method by Álvarez et al. (2012) were provided in the context of a conference publication (see Fritsch et al., 2013).

For applying the performance comparison, a benchmark dataset is used which is part of the KITTI dataset (see Geiger et al., 2012). The KITTI-ROAD dataset is specialized for road area detection algorithms. It comprises images recorded in urban scenarios consisting of 600 frames with challenging sunny conditions. The images have a resolution of 375x1242 px. In order to guarantee decorrelation of the images a spatial offset of the frames of at least 20m (extracted based on GPS position) is used. The benchmark comes with 300 training frames. For those also ground truth annotations of the road area are available. Testing is then applied on another 300 frames. This dataset is publicly available, therefore other authors may also directly compare their approaches with the RTDS results.

For comparison of the three methods several evaluation scores are used. These scores are computed on the test set firstly in the perspective image and secondly in the metric BEV. The perspective evaluation is carried out on the image region that corresponds to the metric extent of the BEV in order to restrict the evaluation to the important parts. Therefore, less important regions (e.g., the sky, far-off scene part) are neglected. For each method a threshold is chosen that maximizes the F-measure (F_{max}) on the test dataset. The F-measure (see, e.g., Álvarez and López, 2008) is derived from the precision and recall values (see Equation 7.3). The precision (Prec) is given in Equation 7.1. The recall (Rec) is equal to the true

[1] http://www.cs.uiuc.edu/homes/dhoiem/projects/software.html

positive rate (see Equation 7.2).

$$\text{Prec} = \frac{TP}{TP + FP} \tag{7.1}$$

$$\text{Rec} = \frac{TP}{TP + FN} \tag{7.2}$$

$$\text{F-measure} = (1 + \beta^2)\frac{\text{Prec} \cdot \text{Rec}}{\beta^2 \cdot \text{Prec} + \text{Rec}} \tag{7.3}$$

The F-measure contains a parameter β which can be used for balancing the influence of precision and recall for a particular purpose. But as there is no concrete application, the harmonic mean is used which can be realized by setting $\beta = 1$ ("F1-measure"). Beside the quality Q and the false positive rate FPR which where introduced in Section 4.5.1, another measure that can be found in Table 7.1 is the Average Precision (AP) (see Equation 7.4).

$$\text{AP} = \frac{1}{11} \sum_{r \in 0,0.1,..,1} \max_{\tilde{r}:\tilde{r}>r} \text{Prec}(\tilde{r}) \tag{7.4}$$

In contrast to the other measures, AP provides insights into the performance over the full recall range (see Everingham et al., 2010). In Table 7.1 all mentioned evaluation scores are listed for the tree methods.

Table 7.1.: Comparison of the proposed RTDS to state-of-the-art approaches on the KITTI-ROAD dataset. The results are obtained using a threshold optimizing the F-measure on the test set.

Perspective Evaluation						
Method	AP	F_{max}	Prec.	Rec.	FPR	Q
Hoiem et al. (2007)	65.4	67.8	54.4	89.9	57.3	51.3
Álvarez et al. (2012)	82.1	80.4	75.3	86.3	21.5	67.3
RTDS	91.8	87.1	89.9	84.6	7.2	77.2

Metric Evaluation						
Method	AP	F_{max}	Prec.	Rec.	FPR	Q
Hoiem et al. (2007)	60.5	60.7	50.4	76.3	36.2	43.6
Álvarez et al. (2012)	68.3	67.7	64.7	70.8	18.6	51.1
RTDS	88.3	83.5	85.2	81.8	6.8	71.6

Beyond that, a graphical impression of the overall performance can be gained using precision-recall curves as depicted in Figure 7.1.

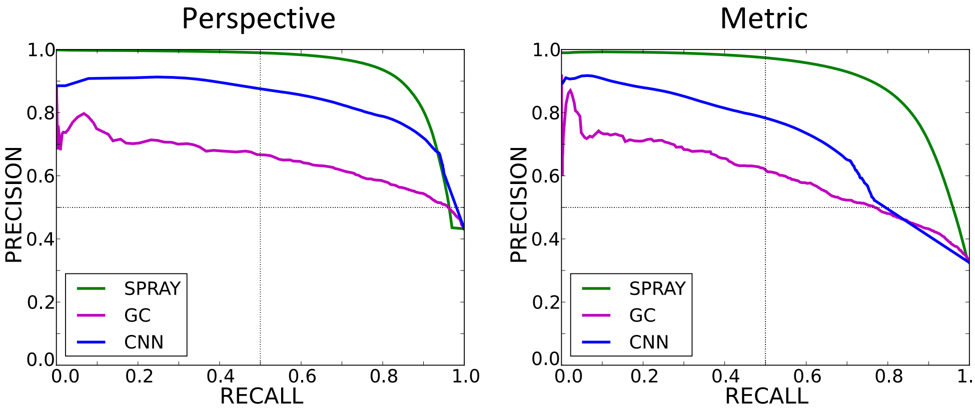

Figure 7.1.: Precision-recall curves of the three approaches in perspective (left) and metric domain (right). SPRAY denotes the proposed method.

All the evaluation measures in Table 7.1 and the precision-recall curve in Figure 7.1 show that the proposed approach for road area classification generally outperforms both compared methods. However, as mentioned above the method by Hoiem et al. (2007) detects all kinds of horizontal ground surfaces and can therefore be only seen as a baseline because the method is not intended for road area detection. Comparing RTDS to the approach by Álvarez et al. (2012) shows that the proposed method ensures a higher detection precision. The quality in the perspective evaluation of RTDS is 10pp above the method by Álvarez et al. (2012). Even higher is the gain in the metric domain where approximately 20pp are obtained. Therefore, the quality in metric evaluation exhibits a bigger offset as in the perspective evaluation. This can be interpreted through the better detection of RTDS in farther distances compared to the method by Álvarez et al. (2012).

For a qualitative comparison of all three methods, Figure 7.2 depicts detection results on some exemplary images from the KITTI-ROAD dataset (test set). The *Geometric Context* basically detects all ground regions independent from the semantics. The approach has some trouble with shadows on the ground (see example VI in Figure 7.2) and with objects close to the observer (see, e.g., the vehicle in example VII). However, the method uses parameters which were generated on a different dataset. It can be assumed that the authors could tune the parameters (e.g., by using the 300 training frames) in order to perform better on the test set.

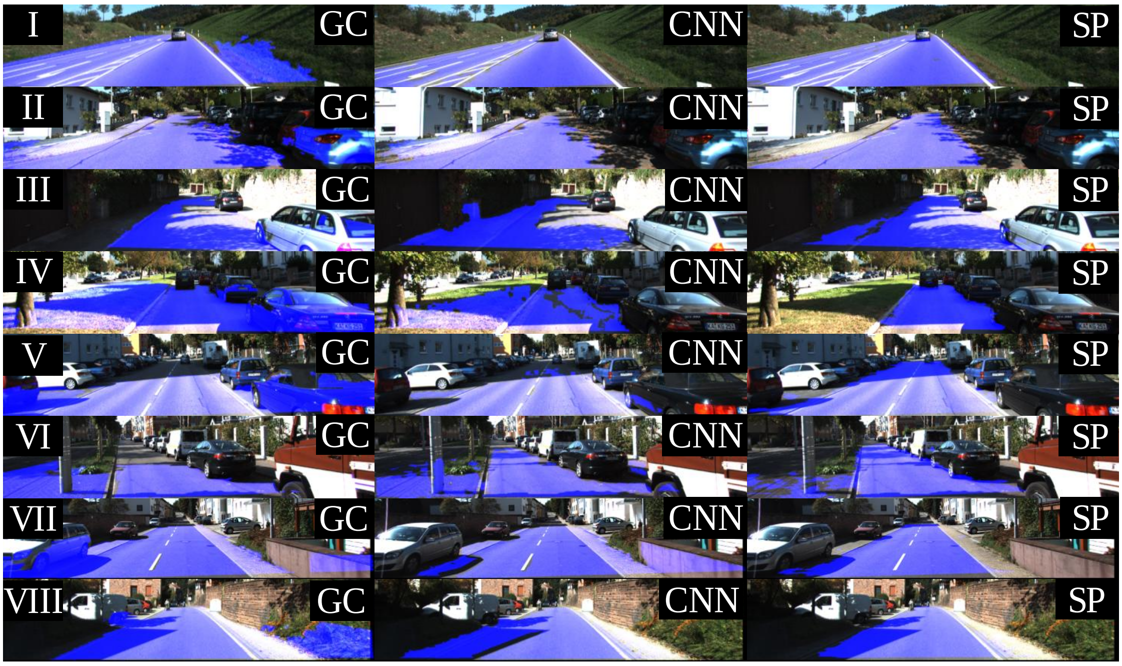

Figure 7.2.: Qualitative comparison of the three approaches. Depicted are eight example images with detection results for road area of the *Geometric Context* by Hoiem et al. (2007) (GC), the approach by Álvarez et al. (2012) (CNN), and RTDS using the proposed SPRAY features (SP).

The method by Álvarez et al. (2012) also has problems with shadows on the road area. Furthermore, the separation between the road area and other regions apart from the road area such as sidewalks is not always given (see example VII in Figure 7.2). In the depicted examples it becomes visible that the method developed in this thesis handles unmarked scenarios and shadows on the road area much better than the compared methods.

7.2. Enhanced Robustness Through Adaptation

There are several proposals for adaptation in the field of road terrain detection which automatically adapt to different visual conditions (see, e.g., Álvarez et al., 2012; Michalke et al., 2009; Franke et al., 2007). However, current approaches seem to be worse in performance compared to offline-learned classifier in situations where the appearance is manifold. This can be explained in the following way. Typically, the adaptive methods try to capture the current road (and non-road) appearance by extracting local visual appearance information from a certain region in the image (see windows in the images I-A and II-A in Figure 7.3). This allows to infer whether any position in the image is likely to be part of the road area by checking if it is similar to the extracted local visual appearance in the window. However, the extracted local visual appearance must be a good representative for all road area regions in the scene. This can be very complex for a scene with light and shadows or various kinds of road and non-road appearances.

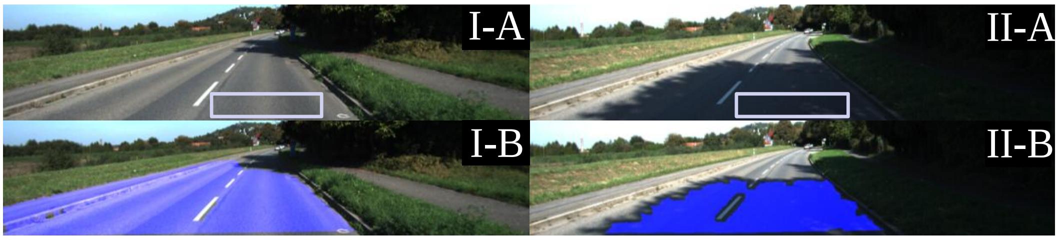

Figure 7.3.: Illustration of the problem when adapting to visual conditions by using an window approach. There is a window for measuring appearance characteristics of the current scene (see boxes in images with tag A). The images with tag B show the resulting road area classification (blue).

An example for this is the changing over space of the local visual appearance, e.g, due to a shadow on the road area or simply due to a changing asphalt texture and color. Figure 7.3 illustrates this conceptual problem with two exemplary images. In the first image, the detection gets stuck at the shadow (see I-B) because the appearance differs from the road area in the window (see I-A). In the second image, the algorithm adapts to the appearance of the road area in the shadow (see II-A). However, the detection gets stuck at the transition from the shadow to the bright road area (see II-B). This additionally shows that it is apparently hard to find a good extraction region for road area (and non-road area) that are representative for the whole scene.

The benchmark presented in the previous Section 7.1 reflects how well the detection system performs on an unseen dataset with a few hundred samples. In this case, the system is specialized on this specific test run by training it on a similar dataset, i.e., containing more or less the same appearance conditions. However, implemented in a real ADAS a detection system needs to handle much more different conditions than contained in a single benchmark. Although, it was shown in Chapter 5 that a generalization to multiple appearance conditions is possible, at a certain point increasing the number of different training scenes becomes inconvenient because the increase in generality comes together with a loss in specificity of the resulting classifier.

Therefore, finding a good trade-off of generality and specificity enabling a system to operate on a high amount of different appearance conditions and to handle changing conditions with high precision is a key aspects for bringing the proposed concept to an automotive vision system.

As we saw in the comparison in the last section, adaptive approaches that come with high generality lack in specificity to ensure high precision in challenging detection situations. However, a higher specificity in certain conditions is the advantage of the proposed offline-learned method. Therefore, a strategy that combines multiple offline-learned classifier models, each being dedicated to a certain condition (specificity), in conjunction with a classifier switching, which ensures generality, is reasonable. This kind of structure variability is contrary to the approach by Álvarez et al. (2012) where the adaptivity is generated within the feature-space of the adaptive online classifier.

The switching of multiple models with the proposed system architecture, being separated in a visual and a spatial stage, allows utilizing the superposition principle as illustrated in Fig. 7.4. This results in appearance and geometrical road models that can be switched independently for adapting the proposed method to the current scene. This is useful because on the one hand conditions change on the local visual appearance level, e.g., different types of asphalt, or on the other hand on the spatial level, e.g., geometrically different road types. This separate switching of the models for appearance and road geometry generally reduces the complexity of adaptation to multiple environmental conditions. This is therefore an advantage of the proposed system architecture because it comprises the corresponding stages which can be adapted separately. Different appearance models for adapting to visual conditions in the base classification (lower stage) could be, e.g., models for different asphalt colors and lighting conditions. Examples for different geometrical models for road terrain classification in the higher stage are, e.g., highway, rural road, or urban area.

From the basis of an arbitrary number of appearance and road geometry models, this results in the problem of identifying the most suitable model for a given scene. To this end, the proposed approach allows the utilization of scene context for switching the models for both stages. In this scope, scene context refers to higher level scene information that describes visual or spatial characteristics of the current scene. As scene context for selecting a particular appearance model simple features

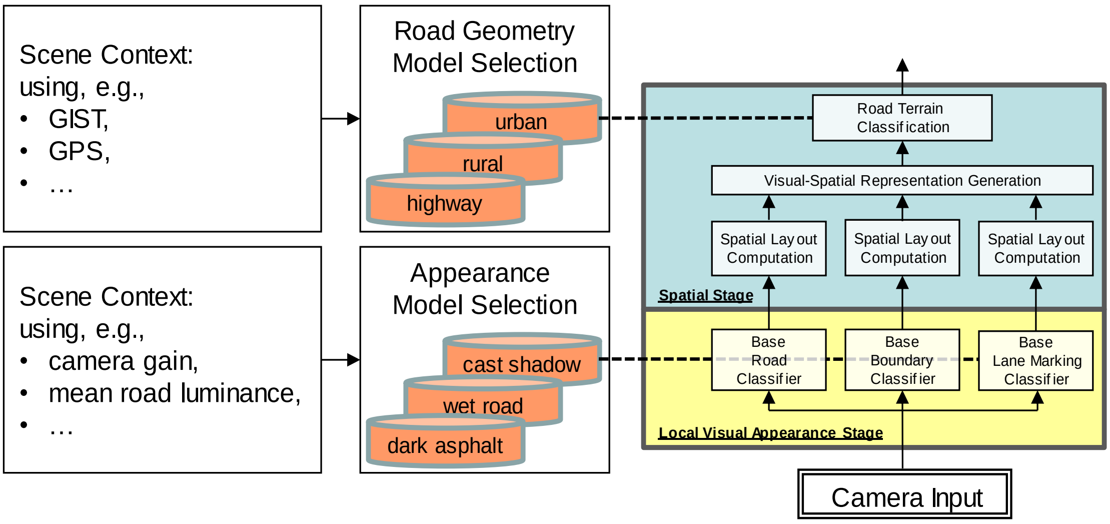
Scene Context:
using, e.g.,
• GIST,
• GPS,
• …
Road Geometry Model Selection
urban
rural
highway
Scene Context:
using, e.g.,
• camera gain,
• mean road luminance,
• …
Appearance Model Selection
cast shadow
wet road
dark asphalt
Road Terrain Classification
Visual-Spatial Representation Generation
Spatial Layout Computation
Spatial Layout Computation
Spatial Layout Computation
Spatial Stage
Base Road Classifier
Base Boundary Classifier
Base Lane Marking Classifier
Local Visual Appearance Stage
Camera Input

Figure 7.4.: Model switching concept for adapting to visual and geometrical changes in the scene. The right part corresponds to the road terrain detection system detailed in previous chapters (see Chapter 3).

capturing the current lighting conditions such as the camera gain or the mean road luminance can be used (see Harville, 2002). Scene context for the spatial stage needs to capture the coarse road category, i.e., whether the ego-vehicle is currently driving on a highway, rural road, or urban road. Kastner et al. (2009) showed that this information can be identified from an image sequence using a GIST-like scene descriptor. The lateral location on the road, i.e., the lane index and the number of lanes can be obtained with the system presented in Chapter 6. This is therefore also useful context information for selecting a particular road geometry model. Furthermore, map data could be used to get knowledge about the current road curvature or, e.g., if the ego-vehicle is currently driving on a highway or on an urban road.

7.3. Summary

This chapter compared the approach proposed in this thesis to state-of-the-art approaches for a specific road terrain detection task. The comparison of three road area detection approaches on a benchmark dataset showed that the method presented in Chapter 5 outperforms all of the other state-of-the-art methods. The proposed approach shows far better detection of the road area under challenging lighting conditions, such as shadows, as the compared methods. Furthermore, the proposed method shows a higher detection precision for road area in farther distance than the compared ones.

In general, adaptive methods are better suited for changing conditions than the proposed method because employing adaptive concepts aims at high generalization of detection systems while an offline-learned classifier (as comprised in the proposed method) is limited to the operation in scenes that are comparable to the training conditions. This generalization is a key issue for bringing any approach to application in an ADAS because in outdoor environments a high variety of different visual and spatial conditions occur.

To this end, the proposed system can be extended with a model switching to ensure high detection performance by adapting to multiple environmental conditions. A model switching of trained classifiers which was discussed in this chapter is a good trade-off between generalization to a lot of varying conditions and specialization for different conditions. The central aspect in correlation to the proposed two-step approach is that the adaptation of the models can be separately applied on the local visual appearance stage and on the spatial stage using scene context. The realization and implementation of the proposed adaptive approach remains future work.

8. Summary and Conclusion

The thesis at hand presented a novel and general approach for visuospatial analysis of road environments from a single image. In the course of this thesis, it was shown that this approach is generic and useful for many aspects in the field of road terrain detection. By implementing the concept in the form of a hierarchical system the advantages and disadvantages compared to other vision-based approaches were analyzed.

The fundamental motivation for this thesis was that road terrain detection is an important source of information for future Advanced Driver Assistance Systems (ADAS). Road terrain reflects in-depth knowledge about the driving space and therefore knowledge about future positions of the ego-vehicle and locations where other road users might appear. The drawback of state-of-the-art approaches implemented in current ADAS is that they are restricted to well-structured environments. The boundary conditions in the employed models restrict the applicability of those approaches for instance to roads courses with low curvature. This leads to the fact that current ADAS are only functioning on highways and some rural roads and definitely reach their limit in urban scenarios. Especially because of unmarked roads and complex road boundary shapes in inner-cities the above mentioned boundary conditions of state-of-the-art approaches do not hold. This poses a big challenge for researchers and is the main reason why some ADAS, such as lane keeping assistance, are currently not applicable on urban roads. However, there is a high potential gain in bringing ADAS to inner-cities because many fatalities happen in urban traffic. Because of a lack of approaches enabling robust detection of road terrain under various visually and spatially complex scenarios, this thesis proposed a different approach to the typically employed delimiter-based detection of nowadays ADAS.

The concept for visuospatial road environment analysis discussed in this thesis represents a major advance towards resolving the above mentioned challenges. In an initial study, systems based on local visual appearance for road terrain detection, so-called base classifiers were analyzed. It was found that detection based on local visual appearance of road terrain categories is ambiguous and therefore error-prone. This limitation motivated to extend the approach in a hierarchical manner. To this end, this thesis presented a bottom-up approach including multiple base classifiers on a lower stage combined with a spatial layout computation of local visual appearance on a higher stage. The basic idea of the spatial layout computation is to spatially capture the confidence distribution from the base classifiers with respect to base points located at fixed positions in metric space. This computation is performed in metric space because perspective changes are compensated, which results in a higher homogeneity of the spatial layout of road scenes. It was shown that the spatial layout of local visual appearance can be represented by means of visuospatial features. In this thesis, this was realized by introducing SPatial RAY (SPRAY) features that capture the spatial layout of local visual appearance by means of a directed ray-based integration process. SPRAY features are directly computed on a base classifier output, which is a value-continuous confidence map in the metric space. By combining several base classifiers with spatial layout computation, this results in a metric representation of the image capturing visual and spatial characteristics of road environments. It was exemplarily shown that base classifiers for road, boundary and lane marking combined with the proposed SPRAY features allow far better discrimination of road terrain categories compared to local visual appearance based processing alone.

Related approaches typically apply explicit models to represent the course of the driving path. This thesis shows that based on the SPRAY features an implicit model can be learned that allows detection of road terrain categories and the classification of road geometrical properties. The approach was evaluated on multiple datasets and showed to handle complex situations, multiple visually and spatially different conditions, and arbitrary delimited roads.

As presented in the course of this thesis, the generic nature makes the approach suitable for a lot of different functions. Through applying machine learning techniques for training a classifier the approach can be automatically tuned to a particular task. In this context, this thesis showed that the proposed system contributes solutions for three challenging applications:

- Robust visual road area detection.
- Ego-lane detection suitable for arbitrary delimited roads.
- Pure visual ego-vehicle localization on the lane level.

The road area is an important type of road terrain because it describes the ground area which is meant for the task of driving, i.e., the composition of all lanes. The detection of road area is challenging because it is visually diverse, especially because of different asphalt colors and textures as well as markings on the road area. Additionally, in urban scenarios the road area is spatially diverse because of multiple lanes with almost arbitrary shapes and parking cars on the side causing occlusion of the road delimiters. Because of all these reasons the visual detection of road area has to cope with a lot of noise. This thesis showed that the proposed method enables learning visual and spatial characteristics of road area by offline-training a classifier. To this end, the GentleBoost classification method using decision trees as weak learners was applied to classify road area based on the proposed SPRAY features. By experimentally comparing the proposed road area detection with base classifiers on real-world video data two effects could be observed. Firstly, smaller regions of false detections are removed, which can be seen as noise compensation effect of the SPRAY-based classification. The outcome was a smoother road area detection compared to the base classifiers. Secondly, and more importantly, it was shown that erroneous bigger regions apart from, but with a similar appearance to the road area can be compensated by incorporating the spatial layout. This results in a far better separation of the road area to road adjacent, road-like regions, such as sidewalks, with the proposed method compared to the base classifiers.

Additional experiments demonstrated a direct comparison of the proposed road area detection with other approaches. On a publicly available and challenging dataset the proposed method proofed its performance compared to a very generic approach by Hoiem et al. (2007) and an adaptive approach by Álvarez et al. (2012). In metric and perspective evaluations using ground truth for road area the proposed method outperformed both methods. The results highlight that the proposed offline learning approach using SPRAY features significantly improves the detection in real world scenarios and is apparently better suited for visually complex situation, e.g., due to shadows on the road area. In quantitative and qualitative evaluation the proposed method handled these challenging conditions far better as the compared methods. Because of all the above, the proposed approach has a high potential for future ADAS. Even though more experiments have to be carried out in the future, the general approach highlighted its advantages compared to other methods and could be useful for systems requiring more holistic knowledge of the driving scene such as ADAS executing emerging maneuvers when a collision is unavoidable.

Beyond that, for future ADAS it will be necessary to subdivide the free space into its semantic entities which are the lanes on the road area. From the perspective of nowadays ADAS, the most important part of the road area is the ego-lane, i.e., the lane the ego-vehicle is currently driving on. Especially ADAS such as lane departure warning and lane keeping assistance systems rely on a robust extraction of the ego-lane. In contrast to the road area, the ego-lane is visually more challenging to detect because the local visual appearance of all lanes is typically the same. In current

ADAS, ego-lane detection is done by combining extraction of the left and right lane delimiters (such as lane markings and curbstones) with temporal tracking using an explicit lane model. As this results in the above mentioned restrictions, this thesis proposed a more generic approach. To this end, a combination of multiple cues reflecting visual and spatial characteristics of lane delimiters and the road area was presented. Using the GentleBoost algorithm, a classifier was learned that detects ego-lane based on spatial relations of local visual properties. This results in an ego-lane detector handling scenarios with challenging visually and spatially diverse conditions. Tests with challenging real-world video data showed that the proposed detection approach can cope with arbitrary roads leading to robust extraction of the driving path even without temporal integration utilizing only a single camera. The approach is also applicable in situations without any or with occluded lane markings, varying asphalt appearances and shadows on the road and is therefore beneficial for future ADAS operating in urban scenarios. A combined classification result for road area and ego-lane is shown in Figure 8.1.

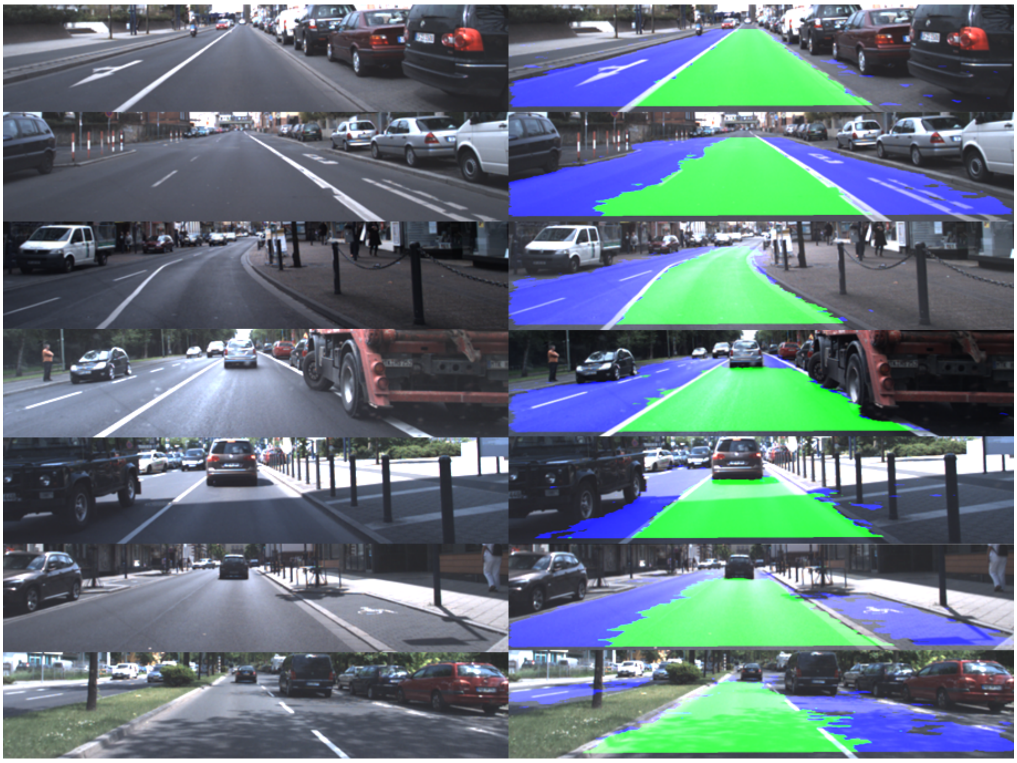

Figure 8.1.: Example scenes showing a combined visualization of road area and ego-lane classification.

In order to show that visuospatial processing provides an efficient representation for spatial road environments which goes beyond road terrain detection, this thesis demonstrated the classification of a holistic scene property. In this context, a system for lane index classification was implemented. This purely vision-based approach enables ego-vehicle localization on the lane level and operates on individual images. This function is for instance useful for navigation systems. Combining a navigation system with a component for ego-vehicle localization would increase its usability by providing more adequate directions. In conducted experiments on highway roads with low traffic density, the ego-lane index detection provided high recognition rates. This demonstrated that the approach captures the extent of the road area and detects the number of lanes in a specific direction. In dense traffic a big amount of the road area is occluded. Therefore, the extent of the road is very challenging to detect with a pure vision-based approach. Consequently, for bringing the approach to an application in real ADAS, a combination with other information sources will be needed. Most importantly, having vehicle detections will allow to reduce the miss-detections occurring during take-over maneuvers and to ensure correct operation also during very dense highway traffic. In addition, incorporating temporal filtering and lane change notifications from lane tracking, turning lights, etc., will ensure robustness during continuous driving.

One benefit of the proposed approach for visuospatial road environment analysis is that it can learn models with high specificity allowing to handle complex scenes. Beyond that, this thesis demonstrated that the approach has good generalization capabilities when applied for road terrain detection under multiple visual and spatial conditions. However, for application in ADAS even more generality will be needed to allow a stand-alone system to work robustly under multiple environmental conditions and in various scenarios. This is for instance necessary when road terrain detection is performed for two types of scenes that are visually extremely dissimilar, e.g., day and night. To this end, a structure-variable combination of an arbitrary number of offline-learned classifiers was discussed which is expected to further enhance the generalization capability and the robustness of the current approach. This extended approach relies on a switching of the corresponding classification models based on scene context. Scene context can be extracted directly from the image or using additional systems indicating the current visual appearance or the geometrical road properties in the current scene. Consequently, the switching can be separately done for the local visual appearance stage of the system, i.e., the low level base classifiers, or the higher level spatial stage. This adaptive system extension will be an important step for applying road terrain detection to a very high number of different conditions with a single system.

In summary, this thesis presented a novel and generic approach for analyzing road scenes by combining visual and spatial information. The pursued approach showed its capabilities for three exemplary applications but in the future other fields of application will be explored as well. In several conducted experiments road terrain detection was demonstrated in challenging real-world scenarios with varying visual and spatial conditions. For all of the above reasons, it is concluded that the presented approach has a high potential for bringing existing ADAS to more complex environments and therefore contributes a significant component for future inner-city safety applications.

List of Publications by the Author

J. Fritsch, T. Kuehnl, and A. Geiger. A new performance measure and evaluation benchmark for road detection algorithms. In *Proc. IEEE Intelligent Transportation Systems*, 2013. submitted.

R. Kastner, T. Kuehnl, J. Fritsch, and C. Goerick. Detection and motion estimation of moving objects based on 3d-warping. In *Proc. IEEE Intelligent Vehicles Symp.*, pages 48–53, 2011.

T. Kuehnl, F. Kummert, and J. Fritsch. Monocular road segmentation using slow feature analysis. In *Proc. IEEE Intelligent Vehicles Symp.*, pages 800–806, 2011. doi: 10.1109/IVS.2011.5940416.

T. Kuehnl, F. Kummert, and J. Fritsch. Spatial ray features for real-time ego-lane extraction. In *Proc. IEEE Intelligent Transportation Systems*, pages 288–293, 2012.

T. Kuehnl, F. Kummert, and J. Fritsch. Image-based lane level positioning using spatial ray features. In *Proc. IEEE Intelligent Vehicles Symp.*, 2013. accepted.

References

Y. Alon, A. Ferencz, and A. Shashua. Off-road path following using region classification and geometric projection constraints. In *Conf. on Computer Vision and Pattern Recognition*, pages 689–696, 2006.

J. Álvarez, T. Gevers, and A. López. 3D scene priors for road detection. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 57–64, 2010. doi: 10.1109/CVPR.2010.5540228.

J. Álvarez, T. Gevers, Y. LeCun, and A. López. Road scene segmentation from a single image. In *Computer Vision–ECCV 2012*, pages 376–389. Springer, 2012. URL http://link.springer.com/chapter/10.1007/978-3-642-33786-4_28.

J. Álvarez, T. Gevers, F. Diego, and A. López. Road geometry classification by adaptive shape models. 2013. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6342913.

J. M. Álvarez and A. López. Novel index for objective evaluation of road detection algorithms. In *Proc. IEEE Intelligent Transportation Systems*, pages 815–820, 2008. doi: 10.1109/ITSC.2008.4732651.

D. Ballard. Generalizing the hough transform to detect arbitrary shapes. *Pattern recognition*, 13(2):111–122, 1981. URL http://www.sciencedirect.com/science/article/pii/0031320381900091.

A. Barth and U. Franke. Where will the oncoming vehicle be the next second? In *Proc. IEEE Intelligent Vehicles Symp.*, pages 1068–1073. IEEE, 2008. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4621210.

P. Berkes. SFA-TK: Slow feature analysis toolkit for matlab (v.1.0.1), 2003. URL http://itb.biologie.hu-berlin.de/berkes/software/sfa-tk/sfa-tk.shtml.

S. Bonnin, F. Kummert, and J. Schmudderich. A generic concept of a system for predicting driving behaviors. In *Proc. IEEE Intelligent Transportation Systems*, pages 1803–1808, 2012.

M. Buehler, K. Iagnemma, and S. Singh, editors. *The DARPA Urban Challenge: Autonomous Vehicles in City Traffic, George Air Force Base, Victorville, California, USA*, volume 56 of *Springer Tracts in Advanced Robotics*, 2009. Springer. ISBN 978-3-642-03990-4.

F. Chausse, J. Laneurit, and R. Chapuis. Vehicle localization on a digital map using particles filtering. In *Proc. IEEE Intelligent Vehicles Symp.*, pages 243–248. IEEE, 2005. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1505109.

G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, volume 1, page 22, 2004.

R. Danescu and S. Nedevschi. New results in stereovision based lane tracking. In *Proc. IEEE Intelligent Vehicles Symp.*, pages 230–235, 2011. doi: 10.1109/IVS.2011.5940434.

T.-S. Dao, K. Y. K. Leung, C. M. Clark, and J. P. Huissoon. Markov-based lane positioning using intervehicle communication. *IEEE Transactions on Intelligent Transportation Systems*, 8(4): 641–650, 2007.

M. Darms, M. Komar, and S. Lueke. Map based road boundary estimation. In *Proc. IEEE Intelligent Vehicles Symp.*, pages 609–614, 2010. doi: 10.1109/IVS.2010.5548011.

A. J. Dean and S. N. Brennan. Terrain-based road vehicle localization on multi-lane highways. In *Proc. ACC '09. American Control Conf*, pages 707–712, 2009.

H. Deusch, J. Wiest, S. Reuter, M. Szczot, M. Konrad, and K. Dietmayer. A random finite set approach to multiple lane detection. In *Proc. IEEE Intelligent Transportation Systems*, pages 270–275, 2012. doi: 10.1109/ITSC.2012.6338772. URL http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6338772.

E. Dickmanns and A. Zapp. A curvature-based scheme for improving road vehicle guidance by computer vision. In *Proceedings of the SPIE Conference on Mobile Robots*, volume 727, pages 161–198, 1986. URL http://spie.org/x648.html?product_id=937795.

A. Ess, T. Mueller, H. Grabner, and L. J. V. Gool. Segmentation-based urban traffic scene understanding. In *Proc. British Machine Vision Conference*, 2009.

M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. URL http://link.springer.com/article/10.1007/s11263-009-0275-4.

U. Franke, H. Loose, and C. Knoeppel. Lane recognition on country roads. In *Proc. IEEE Intelligent Vehicles Symp.*, pages 99–104, 2007.

M. Franzius. *Slowness and sparseness for Unsupervised Learning of Spatial and Object Codes from Naturalistic Data*. PhD thesis, Humboldt-Universität zu Berlin, 2008.

M. Franzius, N. Wilbert, and L. Wiskott. Invariant object recognition with slow feature analysis. In *Proc. Conf. on Artificial Neural Networks*, volume 5163 of *Lecture Notes in Computer Science*, pages 961–970. Springer, 2008. ISBN 978-3-540-87535-2.

Y. Freund and R. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55:119–139, 1997.

J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting. *The Annals of Statistics*, 28:337–407, 2000.

C. Gackstatter, P. Heinemann, S. Thomas, B. Rosenhahn, and G. Klinker. Fusion of clothoid segments for a more accurate and updated prediction of the road geometry. In *Proc. IEEE Intelligent Transportation Systems*, pages 1691–1696, 2010. doi: 10.1109/ITSC.2010.5625270.

A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Computer Vision and Pattern Recognition (CVPR)*, Providence, USA, June 2012.

A. Gern, R. Moebus, and U. Franke. Vision-based lane recognition under adverse weather conditions using optical flow. In *Proc. IEEE Intelligent Vehicles Symp.*, volume 2, pages 652–657. IEEE, 2002. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1188025.

R. Gopalan, T. Hong, M. Shneier, and R. Chellappa. A learning approach towards detection and tracking of lane markings. *IEEE Transactions on Intelligent Transportation Systems*, 13(3): 1088–1098, 2012. doi: 10.1109/TITS.2012.2184756.

N. Gordon, D. Salmond, and A. Smith. Novel approach to nonlinear/non-gaussian bayesian state estimation. In *Radar and Signal Processing, IEE Proceedings F*, volume 140, pages 107–113. IET, 1993. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=210672.

R. Graf, A. Wimmer, and K. C. J. Dietmayer. Probabilistic estimation of temporary lanes at road work zones. In *Proc. IEEE Intelligent Transportation Systems*, pages 716–721, 2012. doi: 10.1109/ITSC.2012.6338764. URL http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6338764.

M. Grgic, K. Delac, and M. Ghanbari. *Recent advances in multimedia signal processing and communications*, volume 231 of *Studies in Computational Intelligence*. Springer, 2009. ISBN: 978-3-642-02899-1.

T. Gumpp, D. Nienhuser, and J. M. Zollner. Lane confidence fusion for visual occupancy estimation. In *Proc. IEEE Intelligent Vehicles Symp.*, pages 1043–1048, 2011. doi: 10.1109/IVS.2011.5940497.

C. Guo and S. Mita. Semantic-based road environment recognition in mixed traffic for intelligent vehicles and advanced driver assistance systems. In *Proc. IEEE Intelligent Transportation Systems*, pages 444–450, 2012. doi: 10.1109/ITSC.2012.6338871. URL http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6338871.

C. Guo, S. Mita, and D. McAllester. Drivable road region detection using homography estimation and efficient belief propagation with coordinate descent optimization. In *Proc. IEEE Intelligent Vehicles Symp.*, pages 317–323, 2009. doi: 10.1109/IVS.2009.5164297.

C. Harris and M. Stephens. A combined corner and edge detector. In *Alvey vision conference*, volume 15, page 50. Manchester, UK, 1988.

M. Harville. A framework for high-level feedback to adaptive, per-pixel, mixture-of-gaussian background models. *Computer Vision – ECCV 2002*, 2352:543–560, 2002. URL http://www.springerlink.com/index/F36MR35R3NR7A6LC.pdf.

Y. Hel-Or and H. Hel-Or. Real-time pattern matching using projection kernels. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(9):1430–1445, 2005. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1471708.

T. K. Ho and M. Basu. Complexity measures of supervised classification problems. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(3):289–300, 2002. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=990132.

D. Hoiem, A. A. Efros, and M. Hebert. Recovering surface layout from an image. *International Journal of Computer Vision*, 75(1):151–172, 2007. URL http://link.springer.com/article/10.1007/s11263-006-0031-y.

S. Hold, S. Gormer, A. Kummert, M. Meuter, and S. Muller-Schneiders. Ela-an exit lane assistant for adaptive cruise control and navigation systems. In *Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on*, pages 629–634. IEEE, 2010. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5625216.

B. Julesz. Textons, the elements of texture perception, and their interactions. *Nature*, 290:91–97, March 1981. URL http://psycnet.apa.org/psycinfo/1982-00297-001.

J. Kang and M. J. Chung. Stereo-vision based free space and obstacle detection with structural and traversability analysis using probabilistic volume polar grid map. In *Proc. IEEE Conf. Robotics, Automation and Mechatronics (RAM)*, pages 245–251, 2011. doi: 10.1109/RAMECH.2011.6070490. URL http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6070490.

Y. Kang, K. Kidono, T. Naito, and Y. Ninomiya. Multiband image segmentation and object recognition using texture filter banks. In *Proc. Int. Conf. on Pattern Recognition*, pages 1–4, 2008.

Y. Kang, K. Yamaguchi, T. Naito, and Y. Ninomiya. Multiband image segmentation and object recognition for understanding road scenes. *IEEE Transactions on Intelligent Transportation Systems*, 12(4):1423–1433, 2011. doi: 10.1109/TITS.2011.2160539.

R. Kastner, F. Schneider, T. Michalke, J. Fritsch, and C. Goerick. Image-based classification of driving scenes by hierarchical principal component classification (hpcc). In *Proc. IEEE Intelligent Vehicles Symp.*, pages 341–346, 2009. doi: 10.1109/IVS.2009.5164301. URL http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5164301.

R. Kastner, T. Michalke, J. Fritsch, and C. Goerick. Towards a task dependent representation generation for scene analysis. In *Proc. IEEE Intelligent Vehicles Symp.*, pages 731–737, 2010.

H. Kloeden, D. Schwarz, E. Biebl, and R. Rasshofer. Vehicle localization using cooperative RF-based landmarks. In *Proc. IEEE Intelligent Vehicles Symp.*, pages 387–392, 2011.

V. Knoop, P. Buist, C. Tiberius, and B. van Arem. Automated lane identification using precise point positioning: An affordable and accurate gps technique. In *Proc. IEEE Intelligent Transportation Systems*, pages 939–944, 2012.

H. Kong, J. Audibert, and J. Ponce. General road detection from a single image. *Image Processing, IEEE Transactions on*, 19(8):2211–2220, 2010. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5443611.

M. Konrad, D. Nuss, and K. Dietmayer. Localization in digital maps for road course estimation using grid maps. In *Proc. IEEE Intelligent Vehicles Symp.*, pages 87–92, 2012.

R. Kumar, A. Reina, and H. Pfister. Radon-like features and their application to connectomics. In *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 186–193, 2010. doi: 10.1109/CVPRW.2010.5543594. URL http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5543594.

S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, volume 2, pages 2169–2178, 2006. doi: 10.1109/CVPR.2006.68. URL http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1641019.

A. Linarth and E. Angelopoulou. On feature templates for particle filter based lane detection. In *Proc. IEEE Intelligent Transportation Systems*, pages 1721–1726, 2011.

C. Lipski, B. Scholz, K. Berger, C. Linz, T. Stich, and M. Magnor. A fast and robust approach to lane marking detection and lane tracking. In *IEEE Southwest Symposium on Image Analysis and Interpretation*, pages 57–60. IEEE, 2008. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4512284.

G. Liu, F. Worgotter, and I. Markelic. Lane shape estimation using a partitioned particle filter for autonomous driving. In *Proc. IEEE Robotics and Automation (ICRA)*, pages 1627–1633. IEEE, 2011. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5979753.

H. Loose and U. Franke. B-spline-based road model for 3d lane recognition. In *Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on*, pages 91–98. IEEE, 2010. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5624968.

A. Lucchi, K. Smith, R. Achanta, V. Lepetit, and P. Fua. A fully automated approach to segmentation of irregularly shaped cellular structures in em images. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2010*, 13:463–471, 2010. URL http://www.springerlink.com/index/8289K5750P63KV00.pdf.

T. Malisiewicz and A. A. Efros. Beyond categories: The visual memex model for reasoning about object relationships. In *NIPS'09*, pages 1222–1230, 2009.

H. A. Mallot, H. H. Bulthoff, J. Little, and S. Bohrer. Inverse perspective mapping simplifies optical flow computation and obstacle detection. *Biological Cybernetics*, 64:177–185, 1991.

N. Mattern, R. Schubert, and G. Wanielik. High-accurate vehicle localization using digital maps and coherency images. In *Proc. IEEE Intelligent Vehicles Symp.*, pages 462–469, 2010.

T. Michalke, R. Kastner, J. Fritsch, and C. Goerick. A generic temporal integration approach for enhancing feature-based road-detection systems. In *Proc. IEEE Intelligent Transportation Systems*, pages 657–663, 2008. doi: 10.1109/ITSC.2008.4732574.

T. Michalke, R. Kastner, M. Herbert, J. Fritsch, and C. Goerick. Adaptive multi-cue fusion for robust detection of unmarked inner-city streets. In *Proc. IEEE Intelligent Vehicles Symp.*, pages 1–8, 2009. doi: 10.1109/IVS.2009.5164243.

M. Okutomi and S. Noguchi. Extraction of road region using stereo images. In *Proc. Fourteenth Int Pattern Recognition Conf*, volume 1, pages 853–856, 1998.

A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

H. Permuter, J. Francos, and I. Jermyn. A study of gaussian mixture models of color and texture features for image classification and segmentation. *Pattern Recognition*, 39(4):695–706, 2006. URL http://www.sciencedirect.com/science/article/pii/S0031320305004334.

D. Pfeiffer and U. Franke. Efficient representation of traffic scenes by means of dynamic stixels. In *Proc. IEEE Intelligent Vehicles Symp.*, pages 217–224, 2010.

D. Pfeiffer and U. Franke. Modeling dynamic 3d environments by means of the stixel world. 3 (3):24–36, 2011.

V. Popescu, R. Danescu, and S. Nedevschi. On-road position estimation by probabilistic integration of visual cues. In *Proc. IEEE Intelligent Vehicles Symp.*, pages 583–589, 2012.

S. J. D. Prince. *Computer vision: models, learning, and inference*, volume 2. Cambridge University Press, 2012.

M. Schreier and V. Willert. Robust free space detection in occupancy grid maps by methods of image analysis and dynamic b-spline contour tracking. In *Proc. IEEE Intelligent Transportation Systems*, pages 514–521, 2012. doi: 10.1109/ITSC.2012.6338636. URL http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6338636.

Y. Sha, X. Yu, and G. Zhang. A feature selection algorithm based on boosting for road detection. In *Proc. Conf. Fuzzy Systems and Knowledge Discovery*, volume 2, pages 257–261, 2008. doi: 10.1109/FSKD.2008.550.

T. Shen and F. Ibrahim. Interacting multiple model road curvature estimation. In *Proc. IEEE Intelligent Transportation Systems*, pages 710–715, 2012. doi: 10.1109/ITSC.2012.6338884. URL http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6338884.

J. Shotton, M. Johnson, and R. Cipolla. Semantic texton forests for image categorization and segmentation. In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition*, pages 1–8, 2008. doi: 10.1109/CVPR.2008.4587503. URL http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4587503.

J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition*, pages 1297–1304, 2011. doi: 10.1109/CVPR.2011.5995316.

J. Siegemund, U. Franke, and W. Forstner. A temporal filter approach for detection and reconstruction of curbs and road surfaces based on conditional random fields. In *Proc. IEEE Intelligent Vehicles Symp.*, pages 637–642, 2011. doi: 10.1109/IVS.2011.5940447.

K. Smith, A. Carleton, and V. Lepetit. Fast ray features for learning irregular shapes. In *Proc. IEEE Int Computer Vision Conf*, pages 397–404, 2009. doi: 10.1109/ICCV.2009.5459210.

P. Sturgess, K. Alahari, L. Ladicky, and P. H. S. Torr. Combining appearance and structure from motion features for road scene understanding. In *Proc. British Machine Vision Conference*, 2009.

M. Szczot, M. Serfling, O. Lohlein, F. Schule, M. Konrad, and K. Dietmayer. Global positioning using a digital map and an imaging radar sensor. In *Proc. IEEE Intelligent Vehicles Symp.*, pages 406–411, 2010. doi: 10.1109/IVS.2010.5548043.

T. Veit, J.-P. Tarel, P. Nicolle, and P. Charbonnier. Evaluation of road marking feature extraction. In *Proc. IEEE Intelligent Transportation Systems*, pages 174–181, 2008. doi: 10.1109/ITSC.2008.4732564.

P. Viola and M. Jones. Robust real-time face detection. In *Proc. Int. Conf. on Computer Vision*, volume 2, 2001. doi: 10.1109/ICCV.2001.937709. URL http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=937709.

T. Weiherer, E. Bouzouraa, and U. Hofmann. A generic map based environment representation for driver assistance systems applied to detect convoy tracks. In *Proc. IEEE Intelligent Transportation Systems*, pages 691–696, 2012.

L. Wiskott and T. Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural Computation*, 14(4):715–770, 2002.

C. Wojek and B. Schiele. A dynamic CRF model for joint labeling of object and scene classes. In *European Conference on Computer Vision*, volume 5305, pages 733–747, 2008. doi: DOI: 10.1007/978-3-540-88693-8.

Q. Wu, W. Zhang, and B. Kumar. Example-based clear path detection assisted by vanishing point estimation. In *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, pages 1615–1620. IEEE, 2011. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5979617.

A. Inverse Perspective Mapping

Inverse perspective Mapping proposed by Mallot et al. (1991) refers to a transformation of the perspective image into a top view of the visual scene in metric coordinates, the so-called *Birds Eye View* (BEV). The basic assumption is that every position is a ground location, i.e., its height over the ground plane $y_{\text{road}} = 0$.

Let the homogeneous matrix P (3x4) map a point in metric camera coordinates $(x_{\text{cam}}, y_{\text{cam}}, z_{\text{cam}})^T$ in unnormalized image coordinates $(u', v', w')^T$ (see Eq. A.1). This projection matrix P can be obtained using the pinhole camera model (see Prince, 2012).

$$\begin{bmatrix} u' \\ v' \\ w' \end{bmatrix} = P \cdot \begin{bmatrix} x_{\text{cam}} \\ y_{\text{cam}} \\ z_{\text{cam}} \\ 1 \end{bmatrix} \tag{A.1}$$

The image pixel coordinates on the image plane $(u, v)^T$ can be obtained by normalization with $u = u'/w'$ and $v = v'/w'$. A 3D location $(x_{\text{road}}, y_{\text{road}}, z_{\text{road}})^T$ in the vehicle coordinate system can be transformed into 3D camera coordinates with Equation A.2. The transformation matrix $T_{r,\text{road2cam}}$ (4x4) contains elements for rotation r_{ij} as well as for translation t_i.

$$\begin{bmatrix} x'_{\text{cam}} \\ y'_{\text{cam}} \\ z'_{\text{cam}} \\ w'_{\text{cam}} \end{bmatrix} = T_{r,\text{road2cam}} \cdot \begin{bmatrix} x_{\text{road}} \\ y_{\text{road}} \\ z_{\text{road}} \\ 1 \end{bmatrix} = \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_1 \\ r_{21} & r_{22} & r_{23} & t_2 \\ r_{31} & r_{32} & r_{33} & t_3 \\ 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} x_{\text{road}} \\ y_{\text{road}} \\ z_{\text{road}} \\ 1 \end{bmatrix} \tag{A.2}$$

Furthermore, assume a rectification matrix $R_{0,\text{rect}}$ (4x4) (which is the identity matrix for rectified images). Combining the equations from above, we obtain the transformation from metric vehicle coordinates $(x_{\text{road}}, y_{\text{road}}, z_{\text{road}})$ into image coordinates $(u, v)^T$ with Equation A.3.

$$T_r = P \cdot R_{0,\text{rect}} \cdot T_{r,\text{road2cam}} \tag{A.3}$$

Consequently, T_r is a 3x4 transformation matrix for obtaining the unnormalized image coordinates $(u', v', w')^T$ for a metric location $(x_{\text{road}}, y_{\text{road}}, z_{\text{road}}, 1)^T$ in vehicle coordinates. For the basic assumption of the BEV that $y_{\text{road}} = 0$, the second column of T_r can be neglected which results in a 3x3 matrix $T_{r,33}$.

$$\begin{bmatrix} u' \\ v' \\ w' \end{bmatrix} = T_{r,33} \begin{bmatrix} x_{\text{road}} \\ z_{\text{road}} \\ 1 \end{bmatrix} \quad \text{for } y_{\text{road}} = 0 \tag{A.4}$$

Let the values $\{x_{\text{min}}, x_{\text{max}}\}$ and $\{z_{\text{min}}, z_{\text{max}}\}$ denote the lateral and respectively longitudinal extent of the BEV. With a constant metric discretization of s_{grid} in the BEV, all metric positions are defined. By applying Equation A.4 on the metric BEV coordinates all corresponding image coordinates can be computed. Therefore, the BEV mapping can be computed conveniently by means of a look up table. Consequently, any image or image-based representation can be transformed into a metric representation.

B. Inner-city Evaluation Datasets

In the main part of this thesis two types of datasets are used for performance evaluation. Both include several images from an in-car mounted camera. The used datasets are:

- A benchmark dataset
- An inner-city dataset

Firstly, a rather small benchmark dataset containing a high variety of different scenes. And secondly, a sequential inner-city dataset which better reflects the typical conditions a vehicle encounters during continous driving, i.e., different scenes for a specific weather condition. In the following both datasets will be discussed. In section Appendix B.1 follows the benchmark dataset. Subsequently, information about the inner-city dataset will be given in Appendix B.2.

B.1. Benchmark Dataset

This dataset consists of 100 images and is therefore a rather small dataset. Having a small dataset allows conducting experiments quickly. The 100 frames are randomly taken from different datasets to capture a high variety of different road appearances. Consequently the benchmark dataset comprises images from sunny, rainy and overcast weather and various road conditions, e.g., different asphalt types. Using a wide variety of different scenes for training a classifier is convenient because it increases the likelihood that the found parameters generalize well to unseen scenarios. The images have an original resolution of 1024x1280 px. In (see Figure B.1) 12 examplary images from the benchmark are depicted.

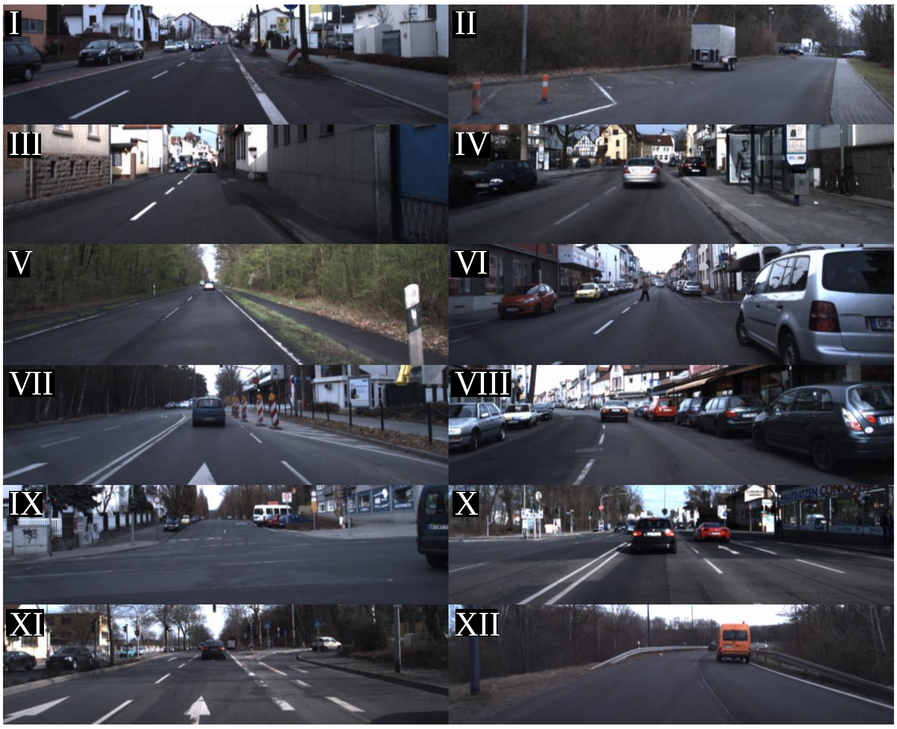

Figure B.1.: Example images for the BENCH dataset.

B.2. Sequential Inner-city Dataset

The inner-city dataset is split into three sub-datasets: overcast (IC1), sunny (IC2) and mixed (IC3). In each dataset the recording vehicle follows a round track in Offenbach (Germany) and contains three driven rounds. The round track is illustrated in Figure B.2.

Each dataset was recorded on a different day with different weather and lighting conditions. The temporal offset of image frames in each part of the dataset is 8 seconds in order to have a sufficient decorrelation in the dataset for training a classifier. Note that similar frames, e.g., caused by the car standing at traffic lights were removed. There is a high variety of different lighting conditions in the dataset. In the following details about the sub datasets will be shortly discussed.

The inner-city dataset IC1 was recorded on the round track in Offenbach under mostly overcast conditions. Twelve exemplary images of the IC1 dataset from the three rounds are depicted in Figure B.3. This part of the inner-city dataset exhibits the most homogeneous lighting conditions.

The inner-city dataset IC2 contains three rounds with sunny weather conditions

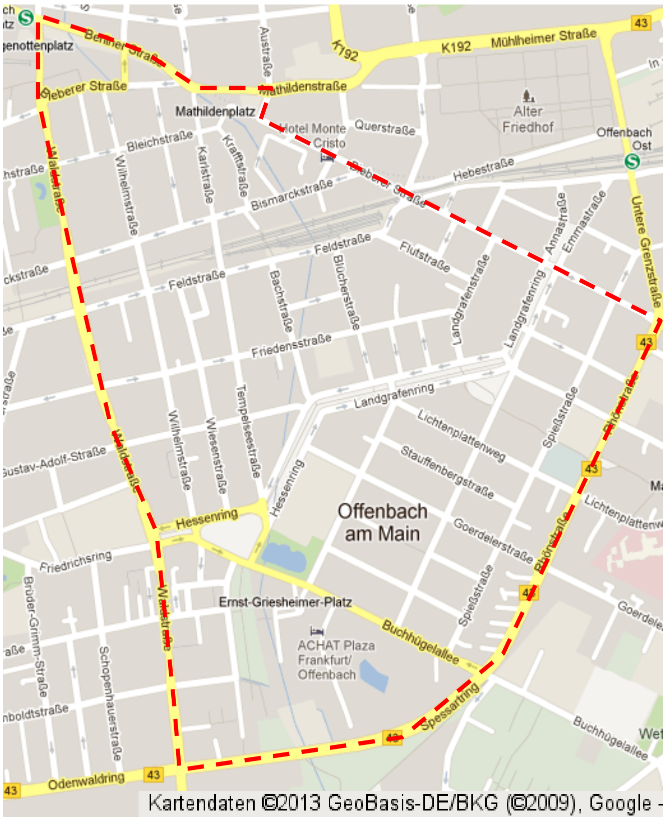

Figure B.2.: The location where the inner-city dataset is recorded is a round track in Offenbach (red line). The picture was extracted from Google Maps.

(see Fig. B.4) from the Offenbach round track. On the one hand side there are rather dark illumination conditions in street canyons but also alternating dark and bright section due to light and shadow. Additionally, this dataset was recorded at a time with high traffic density. Therefore, this is a very challenging part of the dataset.

The inner-city dataset IC3 contains three rounds with changing weather conditions which is very challenging for vision systems (see Fig. B.5). In some passages of the steam there is wet asphalt. There are also image contained with sunny and overcast conditions.

In the evaluation (see Chapter 4 and Chapter 5) there is also a merged dataset containing all images of the three datasets ('all').

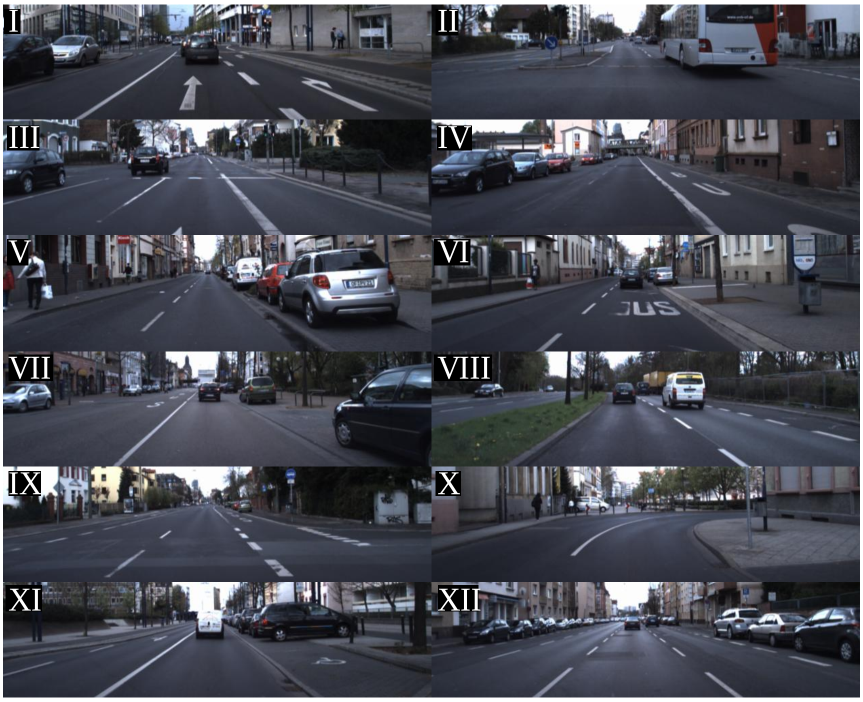

Figure B.3.: Example images for IC1 dataset.

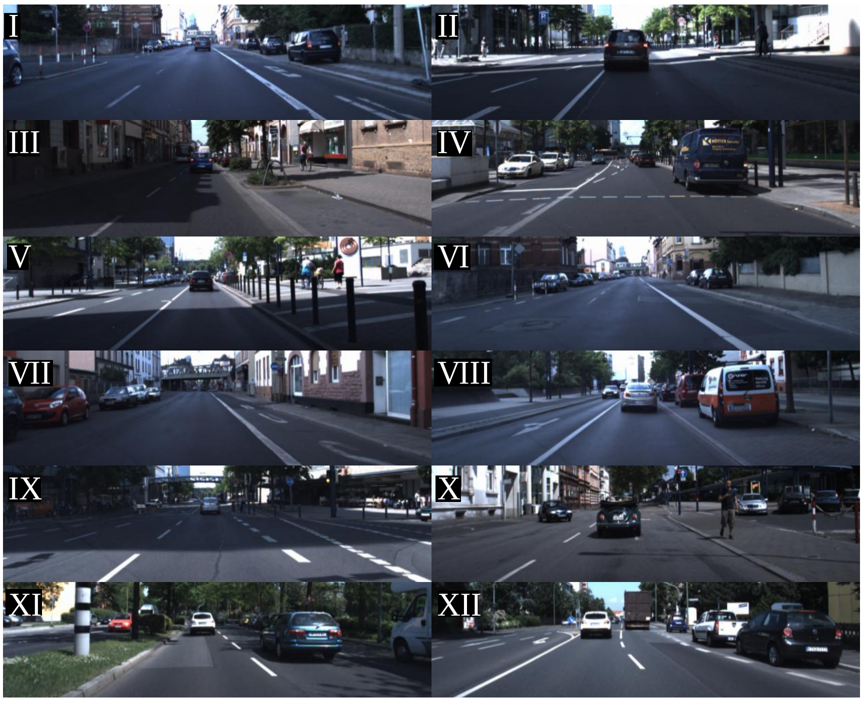

Figure B.4.: Example images for the IC2 dataset.

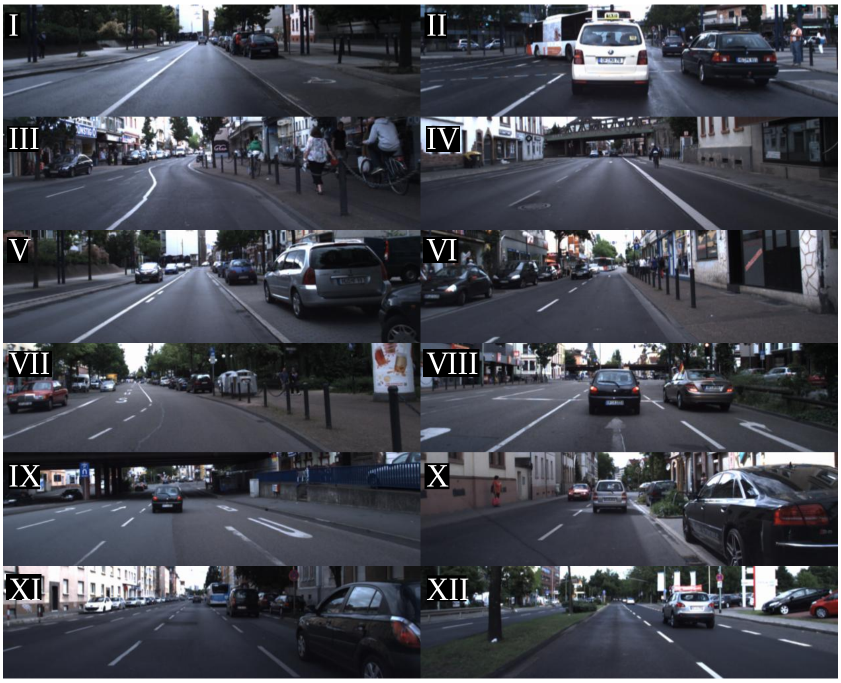

Figure B.5.: Example images for the IC3 dataset.

C. Baseline Performance Evaluation

In order to provide a lower bound for the performance any road detection algorithm should achieve, a static baseline model for road area and ego-lane is computed. Using ground truth for a specific road terrain category (e.g, road area), a baseline model can be obtained by averaging over all binary ground truth maps from a particular dataset. This average reflects basically the likelihood of every image pixel for belonging to the particular class. Assuming a dataset with a defined training and testing set, this model can be generated on the training set. Choosing a threshold ($th = 0.5$ is reasonable), an evaluation can than be performed on the testing set.

As illustrated in Figure C.1, the baseline can not only be created in the perspective image but also in the metric representation by applying Inverse Perspective Mapping (see Appendix A). This results in confidence maps indicating for each perspective/BEV location the confidence for being road area or ego-lane. These baselines can be viewed as scene priors similar to the one used as input to the method by Álvarez et al. (2010).

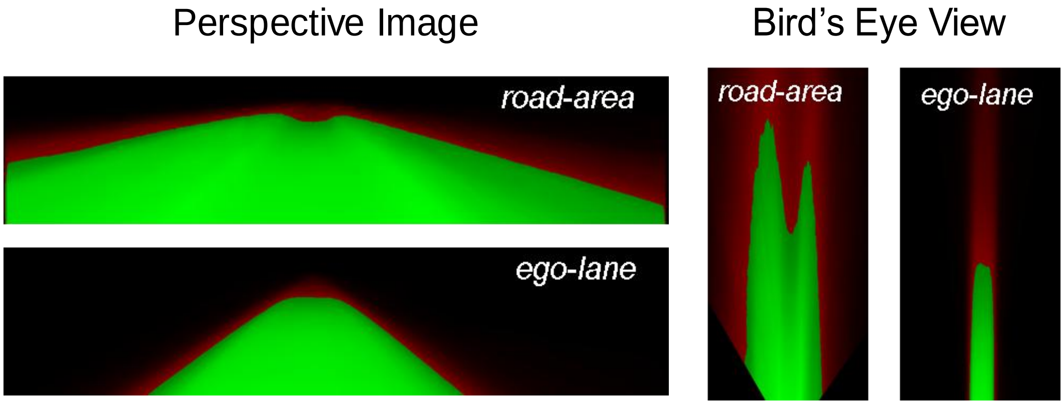

Figure C.1.: Example baseline models for road area and ego-lane in the perspective image (left) and the metric BEV (right). Green colors indicate a confidence above 0.5. Red colors indicate rather unlikely locations of road area / ego-lane.

The baseline is therefore static and independent from the actual input image. However in many situations a large part of the road area can be correctly classified with it. This is because road area typically appears in the same image regions, i.e., in the lower center of an image with a high likelihood.

D. Evaluation of Road Boundary Detection

For evaluating the LVA-system (see Chapter 4) trained on road boundary, experiments are conducted using RGB images with a resolution of 800×600 pixels. Those images are extracted from video streams which are manually annotated with 1Hz (recorded with 20 Hz) and a total stream length of about 4.5 minutes (267 annotated frames). The resulting dataset is split into training and testing part by using N-fold cross validation with a blocking of 20 seconds resulting in 12 blocks.

The evaluation criteria detailed in Section 4.5.1 are used. The false-negative rate, given in Eq. 4.23 is used to evaluate the classification performance on the boundary line (FNR_{Bnd}). Another focus of the system, especially because it is meant to be used as a support for road area detection (see Chapter 5), is the system capabilities not to cause false-positives on the actual driving space. This is nicely captured by the false-positive rate (FPR_{Drv}) given by Equation 4.23 evaluated only on the road area.

The evaluation compares three different training strategies: Firstly, the GentleBoost classifier of the LVA-system is trained as described in Section 4.4.2 (see "Boundary" in Table D.1). Secondly, two specialized classifiers for distinct image regions are obtained using the same approach but splitting the training data into samples corresponding to left and respectively right part of the image (see "Bound-

Table D.1.: Evaluation of the LVA-System trained on road boundary

Training strategy	FNR_{Bnd}	FPR_{Drv}
Boundary	13.46%	10%
Boundary (left/right split)	12.54%	10%
Baseline	21.45%	10%
Boundary	10%	12.20%
Boundary (left/right split)	10%	11.78%
Baseline	10%	17.32%

ary (left/right split)" in Table D.1). The classification results are combined binary, i.e., the left classifier is responsible for the left part and the right classifier vice versa. It is assumed that during training more discriminating features for the road boundary (e.g., edge orientation represented by texture features) can be selected and therefore reduce FNR_{Bnd}. And thirdly, a baseline is generated by using a LVA-system as a non-road area classifier (see "Baseline" in Table D.1). Instead of training a new LVA-system and using non-road area as positive samples and road area as negative samples, the strategy explained in Section 4.4.1 is taken and the confidence result is inverted.

By using different thresholds on the output confidence map of the classifier (see Section 4.3.2) different working points, with different relation of FNR_{Bnd} and FPR_{Drv}, can be employed. In Table D.1 two different working points, one with a $FNR_{Bnd} = 10\%$, and another one with a $FPR_{Drv} = 10\%$ are listed. Table D.1 shows that the proposed training strategy clearly outperforms the baseline. The results show that the combination of the two dedicated classifiers only lead to slightly better performance than non-dedicated classification.

In contrast to the baseline, the proposed training strategy allows focusing the classifier on important boundaries, such as curbstones or the transitions from the road area to non-road. Disregarding lane markings in the training dataset (cf. Section 4.4.2) leads to a reduction of false negatives at road delimiters such as curbstones because features extracted from lane markings on the road area are quite similar to those extracted on the road boundary.