The Role of Machine Vision for Intelligent Vehicles

Benjamin Ranft and Christoph Stiller, *Senior Member, IEEE*

***Abstract*—Humans assimilate information from the traffic environment mainly through visual perception. Obviously, the dominant information required to conduct a vehicle can be acquired with visual sensors. However, in contrast to most other sensor principles, video signals contain relevant information in a highly indirect manner and hence visual sensing requires sophisticated machine vision and image understanding techniques. This paper provides an overview on the state of research in the field of machine vision for intelligent vehicles. The functional spectrum addressed covers the range from advanced driver assistance systems to autonomous driving. The organization of the article adopts the typical order in image processing pipelines that successively condense the rich information and vast amount of data in video sequences. Data-intensive low-level "early vision" techniques first extract features that are later grouped and further processed to obtain information of direct relevance for vehicle guidance. Recognition and classification schemes allow to identify specific objects in a traffic scene. Recently, semantic labeling techniques using convolutional neural networks have achieved impressive results in this field. High-level decisions of intelligent vehicles are often influenced by map data. The emerging role of machine vision in the mapping and localization process is illustrated at the example of autonomous driving. Scene representation methods are discussed that organize the information from all sensors and data sources and thus build the interface between perception and planning. Recently, vision benchmarks have been tailored to various tasks in traffic scene perception that provide a metric for the rich diversity of machine vision methods. Finally, the paper addresses computing architectures suited to real-time implementation. Throughout the paper, numerous specific examples and real world experiments with prototype vehicles are presented.**

***Index Terms*—Advanced driver assistance systems, autonomous driving, computer vision, image processing, intelligent vehicles, machine vision.**

I. Introduction

AFTER three decades of intense research, vision based driver assistance systems have entered our passenger cars and trucks and are a significant pace maker for the progress towards fully automated and cooperative traffic [1]. Like no other sensor, vision sensors—due to the rich information included in images—potentially cover all relevant information that is required for driving. This includes but is by far not limited to lane geometry, drivable road segments, traffic signs, traffic lights, object position and velocity as well as object class. Recent approaches to scene understanding mark the way towards a potential future evolution.

Exploiting this potential of cameras, however, imposes a greater effort as compared to LIDAR, RADAR or ultrasonic sensors. The measurement data of the latter involve distance and/or velocity, i.e., information that resembles physical quantities used as reference for vehicle control. The brightness intensity pattern of a video camera requires computationally and algorithmically demanding processing procedures to extract information that can be used for vehicle control. Nevertheless, cameras have been deployed in commercially available driver assistance systems for many years. Monocular night vision systems, parking aids with a monocular rear or multiple surround view cameras, lane keeping support systems [2], traffic sign recognition [3], evasive pedestrian protection [4] or front collision warning or mitigation systems [5] are but few examples of camera-based driver assistance systems that have entered the automotive market. Even adaptive cruise control systems—traditionally a domain of RADAR or LIDAR sensors—have been implemented with a single camera [6]. Recently, longitudinal control systems have been supplemented with camera-based lateral control for autonomous lane-keeping on highways by multiple OEMs [7].

In the long term driver assistance systems will evolve to *cooperative automated driving* at a safety level significantly superior to that of a human driver, in cooperation with other traffic participants and in all traffic situations. Automated driving has already inspired researchers since the late 1980s. Remarkably, the vast majority of the impressive early results were based on visual perception. In 1994 two demonstrator vehicles drove in normal traffic on Autoroute 1 near Paris demonstrating lane keeping up to 130 km/h, convoy and lane change maneuvers. The latter still required a manual confirmation by a safety driver. About 50 transputers processed images from four cameras extracting lane geometry and the pose of other vehicles [8], [9]. In 1995 a passenger car travelled from Munich in Southern Germany, to Odense in Denmark about 95% in automated mode [9]–[11]. At about the same time another group demonstrated vision-based automated urban driving in the city of Karlsruhe at speeds of ca. 30 km/h [12]. The "No hands across America" tour led a vehicle from Washington DC to San Diego with 98% automated steering yet manual longitudinal control again based on machine vision [13].

Remarkably, the augmentation of map information has led from some 95% automation to full automation (cf., Section III). Yet supervision by a safety driver is still a necessity. While automated driving on structured roads, such as highways, is

Manuscript received February 11, 2016; accepted March 21, 2016. Date of publication April 15, 2016; date of current version July 18, 2016.

B. Ranft is with the FZI Research Center for Information Technology, Karlsruhe 76131, Germany (e-mail: ranft@fzi.de).

C. Stiller is with the Department of Measurement and Control, Karlsruhe Institute of Technology, Karlsruhe 76131, Germany (e-mail: stiller@kit.edu).

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIV.2016.2551553

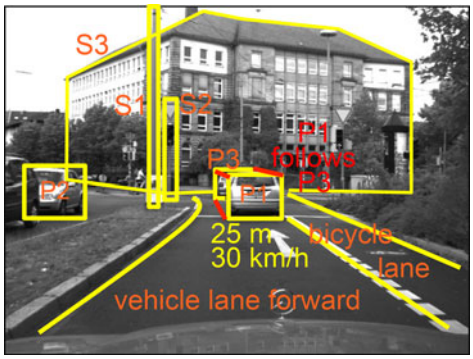

Fig. 1. Metric (yellow), symbolic (orange), and conceptual (red) knowledge for cognitive automobiles.

expected to enter the market within a few years, automated driving on rural roads and in urban environments is still an open research issue.

Machine vision is expected to advance the knowledge acquisition and representation for vehicles as a basis for automated decisions. Naturally, while this article focusses on machine vision, safety-critical functions will not be based on a single sensor principle, but will fuse redundant sensor and map information. As illustrated in Fig. 1, driving involves knowledge representation of various nature. *Metric knowledge*, such as the geometry of static and dynamic objects is required to keep the vehicle on the lane at a safe distance to others. This includes multi-lane geometry, position and orientation of the own vehicle and the position or velocity of other traffic participants. *Symbolic knowledge*, e.g., a classification of lanes as either "vehicle lane forward," "vehicle lane rearward," "bicycle lane," "walkway," allows to conform with basic rules. Finally, *conceptual knowledge*, e.g., specifying a relationship between other traffic participants allows to anticipate the expected evolution of the scene and to drive foresightedly [14].

The remainder of this paper gives an overview on machine vision tasks and methods in the order of typical processing pipelines, in which raw image data is "condensed" to usable information through multiple stages: Section II presents various common "early vision" methods to obtain intermediate results for the following and other functions. As an important provider of static environment information, localization w. r. t. detailed maps is described in Section III. Supplementarily, Section IV highlights the recognition and classification of (dynamic) objects in images. Section V bridges the gap to vehicle control by showing different abstractions of environment perception results for obtaining the aforementioned usable information. Two further sections go beyond this processing pipeline: The benchmarks presented in Section VI are a tool for the scientific community to quantitatively compare and validate computer vision methods. Finally, Section VII gives an overview of hard- and software useful for implementing computer vision systems in mainly prototype vehicles.

II. Early Vision

The term "early vision" is used in both machine vision and neurobiology with certain similarities [15]. We will focus on the former context, where the term however lacks a sharp definition: Early vision methods generally occur first in image processing pipelines and yield only intermediate results rather than symbolic or conceptual information [16]. They are often characterized as similarly data-intensive as their input images, and as general-purpose w. r. t. different applications beyond intelligent vehicles.

For the following sections, we will follow a categorization similar to [17] into single-, dual- and multi-image methods, but only after covering image undistortion as an important prerequisite: By correcting lens distortions one creates a virtual image which complies with a defined camera model, such as pinhole for common or equiangular for fisheye lenses. This greatly simplifies relating two-dimensional (2-D) image and three-dimensional (3-D) world coordinates in downstream processing steps. The estimation of lens distortions is called intrinsic calibration, and can be conducted both offline with specific targets like checkerboards [18] as well as online during the operation of a machine vision system [19]. Some methods additionally require extrinsic calibration of cameras w. r. t. another frame of reference—either another camera [18] as, e.g., in stereo vision, a different type of sensor [20] or the vehicle itself.

A. Single Frame

Since we do not consider "early vision" to include the detection and classification of objects by their appearance (see Section IV), at this point relatively few information can be obtained from single images: Common approaches detect, e.g., edges based on brightness and/or color gradients [21], [22]: Intersections of straight lines can be used to estimate vanishing points and thereby the camera's orientation within a mostly rectangular "Manhattan world" [23], while circular or triangular edges are a basic indicator of traffic signs [24]. Such known-in-advance patterns could also be applied as convolution kernels to a whole image in order to detect their presence and location in a very simple way. Other, e.g., Gaussian kernels as well as non-linear transforms [25], e.g., reduce image noise while ideally preserving detail.

B. Dual Frame

1) Fundamentals: Observation of a scene either from different viewpoints or from a moving camera at different time yields images bearing significantly more information than single ones. The typical corresponding "early vision" tasks are stereo vision and optical flow estimation, respectively. Both methods fundamentally aim at finding precise and accurate matches between images, i.e., pairs of corresponding pixel coordinates. An important characteristic of any method is the density of results: Dense methods provide a match for ideally every pixel, while sparse methods usually only yield one match per tens or thousands of pixels.

A stage only required by sparse methods is the detection of interest points, i.e., well-localizable image patches such as corners or blobs, for which correspondences are estimated. These locally most significant points are usually selected via non-maxima-suppression. The subsequent stage—computation of

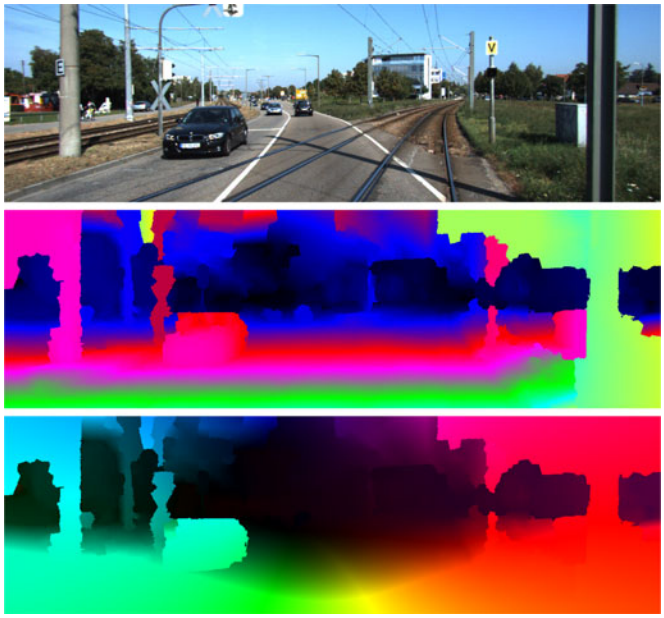

Fig. 2. Sample left input image (top) from public *KITTI* benchmark [26], [27], dense stereo vision (middle) and optical flow results obtained with [28] (bottom).

descriptors—is again common and similar among sparse and dense methods: A descriptor ideally represents a pixel and its proximity in a way which allows for robust and efficient matching in the other image. Since direct matching of raw image intensities lack robustness, descriptors are engineered based on, e.g., gradients [29], histograms thereof [30], [31] or binary comparisons [32], [33]. Some are invariant to illumination, scale, and/or rotation, which enables matches to be found even if lighting/exposure, distance or orientation vary between images. Interested readers are referred to [34] for a comprehensive overview. Automated learning rather than manual engineering enabled Lategahn *et al.*'s [35] descriptor to produce robust matches even across different time-of-day, weather and season.

Finding the best match for a given image point involves comparing and minimizing a dissimilarity measure, which is usually quantified by the Hamming distance for binary descriptors or the sum of absolute/squared differences for vector descriptors. Mutual information [36] is a more robust but less efficient alternative. Generally, due to the large amount of data processed computational complexity is an important aspect for real-time applications, and hence small descriptors are favorable in this respect. The latter—even as scalars or 16-bit-strings—enable most dense matching methods to accumulate their limited information content across a rectangular window with almost no overhead by running sum tables algorithms [37].

2) Stereo Vision: For estimating the depths indicated by color in Fig. 2, specifically stereo vision takes advantage of the fixed arrangement of predominantly two cameras which capture images at the same point in time: Rectification transforms both images such that the search space of matches can be limited to a one-dimensional disparity along a common image row. This helps to resolve ambiguities. Furthermore, it makes the more data-intensive dense methods feasible and explains why sparse methods [29] are relatively rare for this task. Even though many state-of-research algorithms are not real-time capable, several methods have been implemented in commercial products [38]–[40].

Stereo vision methods can be categorized by the data taken into account during the optimization of each pixel's best match: Local methods only minimize the dissimilarity of small image patches [41], and are therefore susceptible to ambiguities, e.g., due to regions without unique texture. This so called *aperture effect* poses a main challenge of stereo vision. It has been tackled by global methods, which minimize an energy term over the whole image that additionally includes the smoothness of resulting depth images [42], [43]. Dynamic programming-based "semi-global" approaches offer a good compromise between the formers' efficiency and the latters' quality [36].

By matching image regions with a constant displacement most methods implicitly assumed locally constant depth, corresponding to surfaces viewed at 90°. Since this assumption is not met in most parts of the image, the matching results exhibit severe artifacts. In particular the road surface is often poorly matched by such methods, because it is typically viewed at an acute angle and is poorly textured. This challenge has recently been solved by methods that relax the locally constant depth assumption to locally slanted planar surface patches viewed at an acute angle. A wide variety of corresponding approaches has been developed: Gallup *et al.* [44] refines planar surfaces only after an initial 3-D reconstruction. Local real-time methods directly integrate few image-wide plane models, either pre-defined [45], [46] or adaptively [47]. In contrast, Yamaguchi *et al.* [48] models hundreds of smaller local planes—one per superpixel segment [49]—and co-optimizes disparities and the relations of neighboring superpixels in a Markov random field.

Yet another challenge for stereo vision methods are specular surfaces. In an intelligent vehicles context these regularly appear on other cars and often result in deformed 3-D reconstructions. Guney and Geiger [50] proposes to integrate their typical shape through CAD models into a method similar to [48], and thereby further improves matching results.

3) Optical Flow: Optical flow represents the velocity and direction of image motion as indicated by brightness and hue in Fig. 2. Its estimation of optical flow is a more general problem than stereo vision, because objects in the viewed scene as well as the camera itself may have moved arbitrarily between consecutive frames. Similarly to stereo rectification, their epipolar geometry can still be exploited for increasing a method's runtime efficiency [51], [52] while in exchange limiting its scope to static scenes without moving objects. Such approaches enable 3-D reconstruction through structure-from-motion, and require the camera's ego-motion as an input.

When including moving objects, the generic 2-D search space of matches explains why for real-time applications sparse optical flow methods have remained more popular than with stereo vision. At this, corresponding pairs of interest point descriptors are usually found through an exhaustive nearest-neighbor search; rejecting implausible matches in advance commonly optimizes performance, while requiring consistent results when matching

from current to previous image and vice versa improves robustness. Another reference method [53] does not require any descriptors but instead finds a point's match through applying the brightness constancy constraint—an important fundamental concept of the following dense methods as well.

A popular approach to dense optical flow estimation is minimizing a global energy which contains a data term and a regularizer to enforce brightness constancy and spatial smoothness of the flow field, respectively. The original method [54] algorithmically depends on the L2 norm of deviations from both objectives, and therefore was susceptible to noise and abrupt flow changes, e.g., at object boundaries. A detailed analysis of derived methods and their relevant enhancements is given by [55]. The reference method [56] successfully mitigates the above issue through a different solver compatible with the L1 norm. Exemplarily, a more recent method does not penalize locally affine flow vectors as caused by—again—planar surface patches [57].

Alternative methods with different foundations, e.g., apply the semi-global stereo matching approach to optical flow [58], or use sparse matches and further image features to initialize the solution of a final variational pass [59], [60].

C. Multi-Frame

Multi-frame methods generally use more images than either a stereo or optical flow pair; an important subset of such methods evaluates a quadruple of images from two cameras (stereo) at two points in time. This enables the estimation of scene flow, a motion field representing the 3-D velocities of reconstructed 3-D points. The closely related "6-D vision" [61] additionally estimates and compensates ego-motion to offer results w. r. t. a world coordinate frame rather than the cameras', which is very useful for moving object detection and tracking. Sparse scene flow estimation can perform a more robust consistency check than described above by requiring a closed loop of pair-wise matches from current-left over current-right, previous-right, previous-left back to current-left image [29]. Dense methods vary in the extent of coupling of the four images: Independently estimated initial disparities and flow vectors can be merged on the level of rigid moving objects [28]. Hornacek *et al.* [62] incorporates stereo depths into flow estimation, i.a., to improve quality. A joint estimation of position and motion of again planar segments is introduced by [63], and achieves a similar quality as the aforementioned methods at run-times of several seconds rather than minutes.

Multiframe analysis as has been outlined for the above example of quadmatches can readily be extended to longer sequences of mono- or binocular frames. Such multiframe correspondence search inherently incorporates redundancy that can be exploited to resolve ambiguities and to remove outliers in optical flow and disparity estimation. Furthermore, feature tracks over more than two frames enhance 3-D reconstruction [64]. While frame-wise filter approaches require the lowest computational effort, bundle adjustment methods working on a sliding window of multiple frames possess the highest potential from an information-theoretic point of view [65], [66].

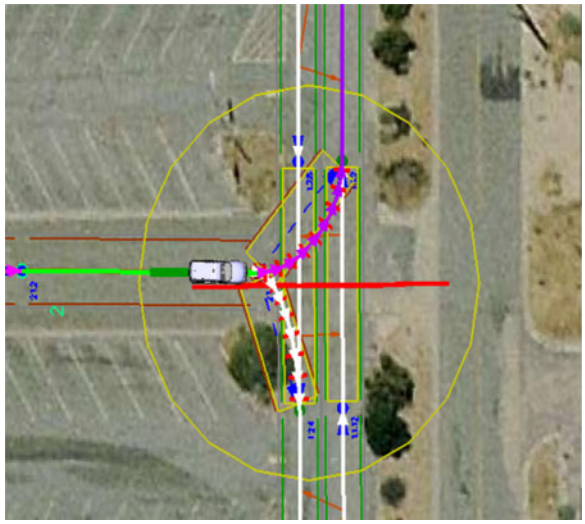

Fig. 3. The route network definition file (RNDF) map represents road center lines by polygons of geopositions yielding a directed graph for the road network.

III. Maps and Localization

Digital maps have been widely used in driver assistance systems and have become a "virtual sensor." Maps available in the market today have typically been composed by culminating satellite imagery with street level information yielding an accuracy of several meters. The information in these maps is sufficient for comfort and information applications such as, e.g., navigation. Highly accurate digital maps with an accuracy in the range of a few cm enable applications that conduct vehicle control. In automated driving map usage has led to a major breakthrough. While automated vehicles successfully travelled some 95% of a route purely relying on onboard sensor information [9], [11], detailed map information enabled automation of the full course over long mileage [67]–[69]. Fig. 3 depicts the 2-D map of the route network definition file (RNDF) that had been provided to the participants in the DARPA Urban Challenge 2007 [68]. In order to employ such a map for behavior and trajectory planning the vehicle must localize itself within this map. For 2-D maps localization is the estimation of a pose (x, y, Ψ) that is composed of a 2-D position x, y and a yaw angle Ψ. While usage of high precision global navigation satellite system (GNSS) receivers with an inertial measurement unit (IMU) has proven sufficient for lateral control of the vehicle in the environments of this competition, the reliability and accuracy of GNSS localization is insufficient in general traffic environments due to satellite occlusion and multipath propagation caused by tunnels, larger buildings, etc.

A remedy to GNSS denial is offered by visual localization. Matching features of available aerial imagery with onboard camera images has yielded localization accuracies in the range of 50–150 cm [70], [71]. While this accuracy is already sufficient for many driver assistance functions, it does not fulfill the requirements of automated vehicle control.

Therefore, modern 3-D maps extend the RNDF information by inclusion of dedicated map layers that facilitate localization with onboard sensors, such as video, RADAR, or

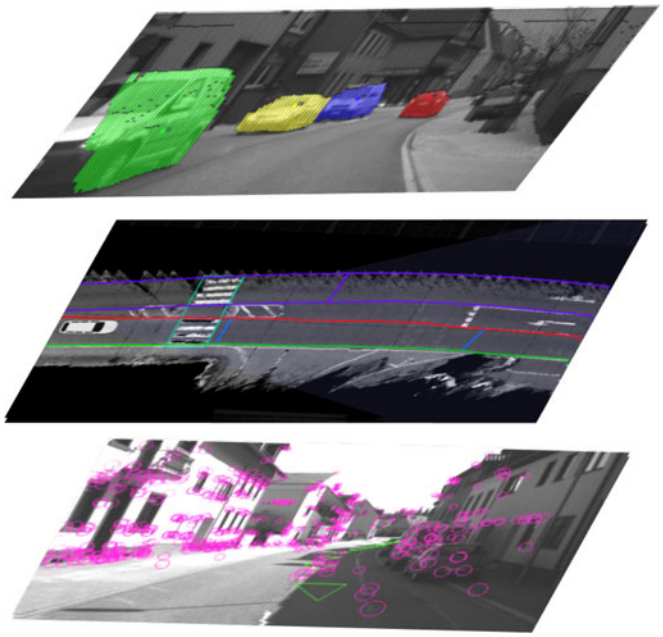

Fig. 4. Layered 3-D map with localization layer (bottom), tactical layer (middle), and dynamic layer (top).

LIDAR. Fig. 4 depicts such a layered map that had successfully been used for automated driving on the historic Bertha Benz route over some 100 km through German cities, villages, and countryside regions [69]. The bottom localization layer includes a six-dimensional (6-D) pose (3-D position and 3-D orientation) sequence that had been driven during the map acquisition process along with distinctive 3-D landmarks in the scene. The localization process, i.e., determination of the 6-D pose in an automated drive, requires the identification and association of at least three landmarks per frame [72], [73].

A combination of GNSS and surface maps are proposed for a set-based localization approach in [74]. LIDAR reflectance maps have been proposed to improve GPS localization in [75].

While visual localization methods have significantly matured in the recent years and commercial solutions are readily offered [76]. Nevertheless, many challenges remain open. Localization under significantly different illumination or weather conditions as compared to the mapping process as well as localization across seasonal or other changes have been studied in [77]. The quantification of the maximal amount of change that can be coped with by localization as well as assessment and procurement of the reliability of a given map and localization under any reasonable conditions are yet open fields of research.

The information that makes a map useful for vehicle automation is incorporated in the tactical layer. As depicted in Fig. 4 this information foremost includes lane geometry that is often represented by lanelets [78]. Furthermore, the tactical layer includes information on local traffic rules and traffic sign information such as speed limits. Last not least, tactical behavior at intersections can be derived to a large extend from the tactical map. This information includes the geometry and the transitions between lanelets, the right of way as well as the accurate position of traffic lights and their associations with individual lanelets.

Map information in these two layers can be considered to stem from a virtual sensor with a huge delay and a huge range. Hence procuring up-to-dateness of this information is a crucial issue for safety critical functions. Crowd mapping approaches aim to offer solutions in this open field of research.

The upper layer of the map depicted in Fig. 4 is the dynamic layer. This layer does not represent map information in the traditional sense, but it aligns information on static and moving objects acquired by onboard sensors, such as LIDAR, RADAR or video. The illustration in the figure shows stixels and objects of a stereo camera from [79].

One last yet important building block in this context is odometry, as it can cover limited localization outages through dead reckoning and provide a vehicle's relative poses during mapping or SLAM. Like with "early vision," such functions can well be based on monocular or stereo cameras and evaluate two or more frames at a time: While a stereo camera can inherently measure the actual length of its 3-D translation vector, a monocular camera cannot directly estimate the scale of the scene in which it is moving. This scalar factor can either be obtained from other sensors such as GNSS and tachometer, or from scene knowledge like the camera's height above the ground plane [80]. Two-frame methods require this information between every pair of frames; the more complex multi-frame methods employ techniques, i.a., bundle adjustment to track the scale for varying amounts of time [81].

IV. Recognition and Classification

A particular strength of human drivers is the seamless visual recognition of any object of interest. Even from a still photograph, i.e., in the absence of disparity or optical flow cues, humans can recognize and classify objects.

In machine vision typically a first processing step aims at selecting areas of interest. The major objective of the detector is to discard large portions of the image with a low computational effort. In the context of vehicular vision, such detectors may employ appearance cues that are sensitive to symmetry, shadows, local texture or colour gradients, as these encode characteristics of vehicles and other traffic objects [82]–[84].

Three-dimensional scene geometry provides another strong cue for the presence of objects. In particular, disparity serves to detect, localize and reconstruct arbitrarily shaped objects in the scene and hence has extensively been used for object detection (see, e.g., [85], [86]). Optical flow is another cue for object detection as it jointly incorporates information on geometry and motion of objects. Clustering techniques group image regions with similar motion vectors. In order to reduce the false alarm rate object hypotheses may be tracked over time and only stable detections are passed on to subsequent processing stages. The standard uncertainty of distance measurements from disparity and optical flow grows quadratically with distance. Hence, such features are well suited for object detection close to the vehicle whereas appearance cues are applied to recognize specific objects at larger distances.

Appearance-based methods detect characteristics of an object type that may be trained on the basis of a data set with

Fig. 5. Semantic labeling ground truth from the *Cityscapes* dataset [99] with classes for road, sidewalk, persons, buildings, poles, trees and sky.

known classification results. This data set includes representative patterns of appearances of the object type under consideration as positive training samples. As negative training samples the dataset includes patterns that somewhat resemble object appearances but do not depict the considered object. These training samples are used to tune parameters of a classifier such that it performs with as many as possible correct detections and with as few as possible false detections. Typically an expressive feature vector is automatically selected during the training procedure from a huge set of candidate features. Simultaneously, a classifier is trained or the probability distribution of the features is modelled. Candidate features often include the coefficients of a linear transform, such as, e.g., Haar or Gabor wavelets [87], [88]. Part-based models decompose an object appearance into characteristic components to cope with partly occlusions [89]–[91]. Best results are achieved by combining several of the above detection techniques followed by a temporal aggregation and consistency checking [92], [93].

Two trends in this area have gained popularity during the last few years: Semantic labeling combines image segmentation and classification in order to assign labels such as road, person, vehicle, building, vegetation or sky with pixel resolution as presented in Fig. 5. Corresponding approaches employ classical methods such as support vector machines [94] or conditional random fields [95], as well as convolutional neural networks (CNN) as the second aforementioned trend [96]. These are a specific type of deep artificial neural networks, an umbrella category in which depth indicates a greater number of layers compared with previous shallow networks. The recent availability of suitable processors and implementations [97] has contributed to the increasing application of deep learning. CNN in particular increase efficiency by reducing the space of learned parameters through the use of specific types of layers: Convolutional layers only require kernel coefficients within a local receptive field, which are applied identically to all inputs. Pooling layers reduce the amount of data and create a local translation invariance, e.g., by only passing a spatial neighborhood's maximum value. A trained CNN may be relatively descriptive by first convolving with simple shapes such as tires or taillights, followed by pooling and another convolution to allow some deviation while still enforcing their typical spatial arrangement on a car. A CNN is typically trained using back-propagation [98] based on the same data as mentioned above. One notable difference is that feature engineering, i.e., the partly manual selection and parametrization of a descriptor, is replaced by feeding raw input pixels into the CNN.

In addition to the above overview of recognition and classification methods, selected application-specific surveys are provided in the following: Mukhtar *et al.* [100] and Sivaraman and Trivedi [101] present an overview of vehicle detection based on monocular or stereo cameras and supplemental active sensors. For pedestrian detection, Benenson *et al.* [102] includes a retrospect and categorization while Geronimo *et al.* [103] performs a survey on a per-processing step level. Similarly, common building blocks are analyzed for lane [104] and traffic sign detection [105] respectively. The state of research, evaluation and future directions of traffic light recognition systems are exemplarily presented by [106]. Although all previous applications process images of an intelligent vehicle's outside environment, driver monitoring will also remain important for advanced driver assistance systems (ADAS) and partial automation to ensure the driver's availability if required. The surveys [107] and [108] distinguish between the conditions drowsiness, fatigue and distraction, and present both visual and non-visual features for their detection. The former include eye closing/blinking and yawning, which are usually recognized with classical techniques, as well as general facial expressions classified by neural networks.

V. Driving Scene Abstraction

As stated in the initial outline, the gap between sensor-specific perception and vehicle behavior generation is bridged by a common environment model. It usually abstracts sensor-specific and data-intensive "early vision" results into a more generic and condensed representation, and hereby often supports a fusion of data from multiple and diverse sensors. In return for abstracting from particular sensors, the model may instead be tailored towards its use by specific methods for scene understanding, trajectory planning and others. The following paragraphs will focus on such abstractions of the current driving scene which can be created and updated from vision-based methods, and will categorize them into static and dynamic environment elements wherever applicable.

Segmentation of free space can serve as a prerequisite for determining where an intelligent vehicle could possibly drive without collision. It usually involves a geometric model of the road surface with varying complexity, e.g., a flat plane [80] or longitudinal B-splines [109], whose parameters are estimated based on 3-D reconstructions from monocular [80] or stereo images [109]. The free space boundary can subsequently be found through the 3-D reconstruction's height w. r. t. the road surface model. Furthermore, this forms a basis for representing obstacles as an opposite and complementary information to free space: At this, stixels [110]—thin vertical rectangles on the ground—have become a widely used model. Their tracking enables application not only to static but also to dynamic elements in the driving scene. In addition, they can be semantically

labeled [111] and grouped into dynamically moving objects [112].

While stixels are still detected and often visualized in the image domain, changing the reference frame to a top view of an intelligent vehicle's environment abstracts from the camera(s) as a specific sensor type and is thereby favorable for functions like multi-sensor fusion, mapping or trajectory planning. At this, the most common environment model are occupancy grids which rasterize the environment at a fixed cell size. The grid is usually flat for intelligent vehicles applications, but may contain a third vertical dimension [113]. The concept of a cell's occupancy is typically implemented based on Bayesian probability [114] or Dempster–Shafer evidence theory [115]; the latter's advantage is the capability of modeling that a cell has neither been observed as free nor occupied. Occupancy grids can be extended towards explicitly modeling dynamically moving cells by supplementing each of them with an estimated motion vector [116].

For dynamic scene elements, a further abstraction and data-reduction often leads to an object list, whose entries can not only be found from stixels as described above or by appearance as described in Section IV, but also, e.g., from scene flow [117]. Each entry's attributes may include position, velocity and acceleration, shape and extent (with bounding boxes as a common and simple model), existence probability, and object class (such as car, cyclist, etc.). Tracking and motion models can be used for filtering velocity or acceleration, and thereby enable short term prediction. While the just outlined representation is generic, others can be tailored towards a specific use: Exemplarily, a continuous trajectory planner places acute trapezoids around obstacles as depicted in Fig. 6. Each trapezoid's legs guide the optimized trajectory around its corresponding obstacle smoothly on a previously determined side.

Determining this side to pass an obstacle on is only one motivation for assigning objects to lanes, and to consequently include lanes in the environment model. They are obviously also needed for trajectory planning, and useful for scene understanding [118] and longer-term dynamic object prediction [119]. An intelligent vehicle often detects its own and neighboring lanes online [120], and applies one of various geometric models [121] to fill gaps between lane marking measurements and to increase robustness. Most models assume a flat 2-D ground plane, but including 3-D height information improves the representation of hills, dips and inclined curves [122]. As an alternative to online detection, a tactical map layer as described in Section III can provide more distant lanes with additional annotations such as traffic rules and intersection topologies.

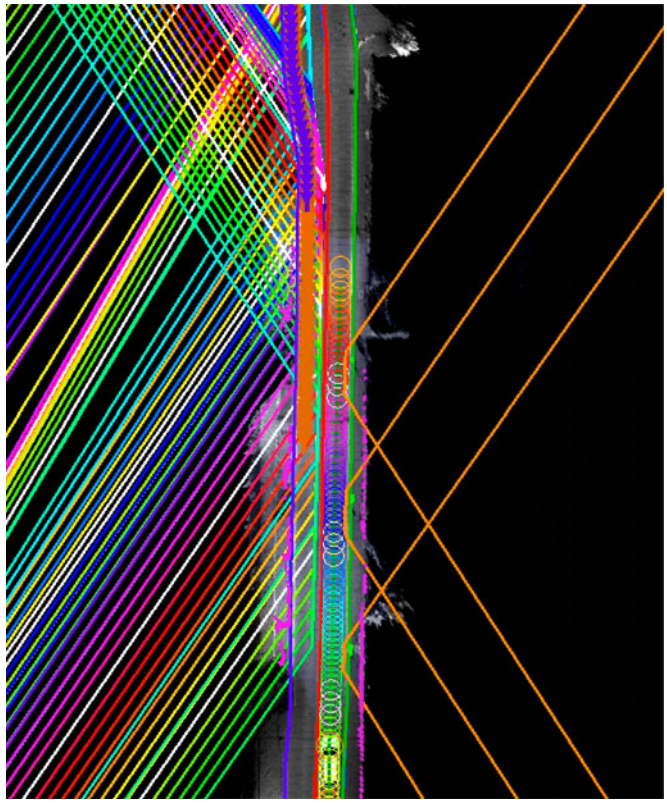

Fig. 6. Visualization of the obstacle layer of Ziegler *et al.*'s [69] environment model: The left/right boundary of the white ego-vehicle's allowed driving corridor is indicated by a red/green line respectively. Obstacles are contained within the "tip" of trapezoids, which extend away from the driving corridor in order to guide the trajectory planner around their correct side. The orange trapezoids on the right are placed around static obstacles and are therefore valid at any point in time. The rainbow-colored groups of trapezoids on the left contain past and predicted future positions of dynamic obstacles, with each color representing a specific point in time. The ego-vehicles planned trajectory is indicated by circles with the same time-to-color mapping, and therefore—in order to be collision-free—must not intersect trapezoids with identical color but may well intersect those with different color.

VI. Benchmarks and Validation

A large variety of methods has been developed for each of the aforementioned tasks and research for robust methods is still ongoing. Their quantitative validation and comparison is enabled by public data sets and benchmarks, which may aid and guide future research. A benchmark's key characteristics ideally include realistic and representative data, the availability of ground truth or reference results (for a training subset), and a popular ranking list of methods. The following paragraphs will briefly present selected data sets for different intelligent vehicle applications:

At "early vision," most benchmarks focus on dense stereo vision and optical flow. The prominent *Middlebury* benchmark [123], [124] provides precise ground truth obtained from a structured light method, and is therefore limited to indoor scenes. As a traffic-focused alternative, the *KITTI* benchmark [26], [27] also includes a scene flow ranking and is based on reference results from a 64-beam LIDAR and CAD vehicle models. Synthesized data like in [125] allow calculating perfect ground truth results. The *Robust Vision Challenge* of 2012 focused on adverse conditions such as night, rain and glare, in which a jury of experts ranked the participating stereo and optical flow methods. Aside from "early vision," Pfeiffer *et al.*[126] offers reference results for dynamic stixels on highways during both good and bad weather.

For vision- or LIDAR-based localization and odometry, one [26], [127] or more [128] real-time kinematic GNSS/IMU sensors are the typical source of reference results. Specific data sets

exist for the related tasks of lane [26] or lane marking detection [129], both based on manually annotated ground truth.

To support evaluating the detection and tracking of pedestrians by appearance, [130] provides images with temporally connected bounding boxes. Pixel-precise segmentations of pedestrians [131] and tagged events about their behavior [132] are also available. Caraffi *et al.* [133] offers bounding boxes for cars and trucks, while [26] differentiates between pedestrians, cyclists and cars. Estimates of the full 6-D relative position between car and camera can be evaluated on [134].

Creating ground truth for per-pixel semantic labeling is a labor-intensive process, which nevertheless has been done [135], [136]. A benchmark including a ranking [137] is announced followed by one with finer-grained labels and stereo images from various cities [99].

VII. Implementation in Vehicles

The realization of any machine vision application begins with one or more cameras. Selection criteria include color versus grayscale, resolution and frame rate, as well as shutter type: Most color cameras use a *Bayer* filter to expose each pixel to either red, green or blue light only. In comparison to grayscale cameras this requires longer exposure times or higher signal amplification, causing more motion blur or noise respectively. A high-bandwidth connection to the camera enables high frame rates and/or resolutions—the latter are essential, e.g., for pedestrian gaze estimation [138]. Global shutter cameras capture a whole image at one point in time while rolling shutter cameras capture row after row, which needs to be mitigated algorithmically. Implementing a global shutter on CMOS image sensors is more complex than on the alternative CCDs. The former however offer a higher dynamic range between saturated black and white pixels, while the latter suffer from blooming at which the charge of overexposed pixels spreads into their neighbors. Since the illuminance in traffic scenes varies greatly, i.a., with time of day, weather and casted shadows, exposure time and signal gain need to be continuously adapted. A suitable feedback control is available on most cameras, but may also be implemented by the user.

As stated initially, cameras are a very data-intensive sensor for intelligent vehicles, which often requires an efficient use of computational resources—not only in production vehicles but also in research prototypes. Here a fundamental change throughout the last decade has been the growing relevance of parallelism: While transistor density still keeps increasing according to Moore's law, the performance of a single processor core is more and more constrained by power, memory and instruction-level parallelism walls [139]. This encourages manufacturers to rather built additional cores from additional transistors, and requires programmers to make efficient use of them.

Luckily, not every researcher needs to consider this even when targeting real-time applications: Libraries such as *OpenCV*, *Point Cloud Library*, *Intel Integrated Performance Primitives* or *NVIDIA Performance Primitives* offer implementations of numerous machine vision methods and image processing building blocks which make efficient use of modern processors. Doing the same with own implementations requires taking into account the target processor's characteristics: Today's CPUs from servers to portable devices contain several cores, each of which is capable of single-instruction-multiple-data (SIMD) operations. The cores can perform independent or cooperative tasks, while SIMD depends on applying identical operations to adjacent data. Not all algorithms allow the latter, but if so can be implemented very efficiently. Independent tasks within an intelligent vehicle can exploit multiple cores through frameworks such as *Robot Operating System*, *Automotive Data and Time-Triggered Framework* or *Real-Time Database for Cognitive Automobiles (KogMo-RTDB)*. In contrast, multi-core as well as SIMD cooperation on a single task may be implemented automatically by an optimizing compiler, or via explicit compiler directives such as *OpenMP*. Additionally, the scope of graphics processing units (GPU) has extended from image synthesis to image analysis and general-purpose computations throughout the last decade. Well-suited algorithms need to offer thousands of independent yet similar computations; if so, the quantitative advantage of GPUs w. r. t. performance and energy efficiency is still very application-dependant [140]. The efforts for GPU programming were greatly reduced via the aforementioned libraries or *OpenACC* compiler directives, compared to the specific programming languages *CUDA* or *OpenCL*. An alternative to the above general-purpose CPUs or GPUs exists in the form of field-programmable gate arrays, which trade increased programming efforts for improved power efficiency: This makes them most relevant for production vehicles and research towards this goal [141], [142].

A complete ADAS or autonomous vehicle is usually constituted by various functions best-suited for different types of processors, which favors heterogeneous hardware platforms: Many simple GPU cores are used, e.g., for "early vision" on a per-pixel level, while more complex CPU cores perform, i.a., situation understanding or trajectory planning. Such configurations are not only found in research prototypes with discrete CPU(s) and GPU(s), but—at lower electrical and therefore processing power—also in systems-on-chip for production vehicles. Exemplarily, the *Tegra K1* and *X1* combine 4-8 CPU- with 192-256 GPU-cores, the latter for both general-purpose computation and visualization. The widely-used and vertically-integrated *EyeQ* series features, i.a., application-specific "vision computing engines," while the upcoming *S32V234* will employ general CPU- and GPU- as well as dedicated machine vision cores. The recently announced *Drive PX 2* platform will be one of the first to provide enough processing power for real-time 360° perception by CNNs, but requires a liquid cooling circuit to dissipate its power of up to 250 W.

VIII. Conclusion and Outlook

This overview has outlined the present and the potential future role of machine vision for driver assistance and automated driving. Like no other sensor data, images comprise a rich variety of information on the traffic scene which makes cameras the dominant sensors for vehicular perception. This information comprises metric knowledge, such as the geometry of the

static and dynamic environment, symbolic knowledge, such as class information and conceptual knowledge such as the interrelation of objects in the scene. Numerous image analysis techniques have largely been inspired by the machine vision community. However, the requirements on performance in terms of robustness w. r. t. environmental conditions, true and false detection rates and admissible computational effort for real-time operation, and, last not least, the reliability level required in the automotive domain are hardly comparable with other fields of application. Indeed, safety-critical functions require the fusion of the information from a set of different sensors and maps forming a diverse perception system. Typical image processing pipelines successively condense the huge image raw data in several consecutive steps. "Early vision" obtains image features that are further condensed to objects in the scene. Decisions of intelligent vehicles often rely on data stored in digital maps. High precision visual mapping and localization has significantly extended the performance of autonomous driving.

Recognition and classification schemes identify symbolic knowledge on specific objects of interest. Learning methods often based on CNN have been successfully applied to traffic scene labeling. The information from vision and other sensors may be gathered in a 2-D scene representation that serves as the basis to a subsequent planning stage. Vision benchmarks have been discussed that allow to quantitatively compare the huge variety of available methods specifically for driver assistance and automated driving. The advances in computing hardware provide numerous implementation options on real-time hardware onboard experimental vehicles.

Naturally, such an overview cannot be exhaustive but only shed light on selected issues and work in the field. We have restricted this overview to perception of the vehicle's exterior environment. For a discussion of the potential of simultaneous interior and exterior vision we refer to [143].

Despite the exciting progress of camera-based driver assistance functions in the market and vision-based automated driving in experimental vehicles, vehicular vision is still in its infancy. As extended capabilities like recognition, detection, tracking and classification with larger ranges and fields of view or novel capabilities like scene understanding emerge, we will witness a boosting of new vision-based applications in our automobiles culminating in door to door automated driving.

REFERENCES

[1] K. Bengler, K. Dietmayer, B. Färber, M. Maurer, C. Stiller, and H. Winner, "Three decades of driver assistance systems - review and future perspectives," *IEEE Intell. Transp. Syst. Mag.*, vol. 6, no. 4, pp. 6–22, Winter 2014.

[2] S. Ishida and J. Gayko, "Development, evaluation and introduction of a lane keeping assistance system," in *Proc. IEEE Intell. Vehicles Symp.*, Parma, Italy, Jun. 2004, pp. 943–944.

[3] S. Estable, J. Schick, F. Stein, R. Janssen, R. Ott, W. Ritter, and Y.-J. Zheng, "A real-time traffic sign recognition system," in *Proc. Intell. Vehicles Symp.*, 1994, pp. 213–218.

[4] T. Dang, J. Desens, U. Franke, D. Gavrila, L. Schäfers, and W. Ziegler, "Steering and evasion assist," in *Handbook of Intelligent Vehicles*, A. Eskandarian, Ed. London, U.K.: Springer, 2012, pp. 759–782.

[5] M. Maurer, "Forward collision warning and avoidance," in *Handbook of Intelligent Vehicles*, A. Eskandarian, Ed. London, U.K.: Springer, 2012, pp. 657–687.

[6] P. Stein, G. O. Mano, and A. Shashua, "Vision-based ACC with a single camera: Bounds on range and range rate accuracy," in *Proc. Intell. Vehicles Symp.*, 2003, pp. 120–125.

[7] L. Ulrich, "Top ten tech cars," *IEEE Spectr.*, vol. 51, no. 4, pp. 38–47, Apr. 2014.

[8] U. Franke, S. Mehring, A. Suissa, and S. Hahn, "The daimler-benz steering assistant: A spin-off from autonomous driving," in *Proc. Intell. Vehicles Symp.*, Oct. 1994, pp. 120–124.

[9] E. Dickmanns, R. Behringer, D. Dickmanns, T. Hildebrandt, M. Maurer, F. Thomanek, and J. Schiehlen, "The seeing passenger car'VaMoRs-P," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, 1994, pp. 68–73.

[10] E. Dickmanns and B. Mysliwetz, "Recursive 3-D road and relative ego-state recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 2, pp. 199–213, Feb. 1992.

[11] M. Maurer, R. Behringer, S. Fürst, F. Thomanek, and E. Dickmanns, "A compact vision system for road vehicle guidance," in *Proc. 13th Int. Conf. Pattern Recognition*, 1996, pp. 313–317.

[12] H.-H. Nagel, W. Enkelmann, and G. Struck, "FhG-Co-Driver: From map-guided automatic driving by machine vision to a cooperative driver support," *J. Math. Comput. Modeling*, vol. 22, pp. 101–108, 1995.

[13] D. Pomerleau and T. Jochem, "Rapidly adapting machine vision for automated vehicle steering," *IEEE Expert*, vol. 11, no. 2, pp. 19–27, Apr. 1996.

[14] C. Stiller and J. Ziegler, "Situation assessment and trajectory planning for annieway," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. Workshop Perception Navigat. Auton. Vehicles Human Environ.*, San Francisco, CA, USA, Sep. 2011.

[15] T. V. Papathomas, *Early Vision and Beyond*. Cambridge, MA, USA: MIT Press, 1995.

[16] C. Tomasi, "Early vision," in *Encyclopedia of Cognitive Science*, Hoboken, New Jersey, USA: John Wiley & Sons Ltd, 2006.

[17] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[18] T. Strauss, J. Ziegler, and J. Beck, "Calibrating multiple cameras with non-overlapping views using coded checkerboard targets," in *Proc. Intell. Transp. Syst. Conf.*, 2014, pp. 2623–2628.

[19] T. Dang, C. Hoffmann, and C. Stiller, "Continuous stereo self-calibration by camera parameter tracking," *Trans. Image Process.*, vol. 18, no. 7, pp. 1536–1550, 2009.

[20] S. Schneider, T. Luettel, and H.-J. Wünsche, "Odometry-based online extrinsic sensor calibration," in *Proc. Int. Conf. Intell. Robots Syst.*, 2013, pp. 1287–1292.

[21] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-8, no. 6, pp. 679–698, Nov. 1986.

[22] R. O. Duda and P. E. Hart, "Use of the hough transformation to detect lines and curves in pictures," AI Center, SRI Int., Menlo Park, CA, USA, Tech. Rep. no. 36, Apr. 1971.

[23] T. Schwarze and M. Lauer, "Minimizing odometry drift by vanishing direction references," presented at the Int. Conf. Indoor Positioning Indoor Navigation, Banff, AB, Canada, 2015.

[24] M. Á. García-Garrido, M. Á. Sotelo, and E. Martín-Gorostiza, "Fast road sign detection using hough transform for assisted driving of road vehicles," in *Computer Aided Systems Theory – EUROCAST 2005*, R. M. Díaz, F. Pichler, and A. Q. Arencibia, Eds. Berlin, Germany: Springer, 2005, pp. 543–548.

[25] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Proc. Int. Conf. Comput. Vision*, 1998, pp. 839–846.

[26] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. Conf. Comput. Vision Pattern Recognition*, 2012, pp. 3354–3361.

[27] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. (2013 Sep.) 'Vision meets robotics: The KITTI dataset. *Int. J. Robot. Res.*[Online]. *32*, pp. 1229–1235. Available: http://www.mrt.kit.edu/z/publ/download/2013/GeigerAl2013IJRR.pdf

[28] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *Proc. Conf. Comput. Vision Pattern Recognition*, 2015, pp. 3061–3070.

[29] A. Geiger, J. Ziegler, and C. Stiller, "Stereoscan: Dense 3D reconstruction in real-time," in *Proc. Intell. Vehicles Symp.*, 2011, pp. 963–968.

[30] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. Int. Conf. Comput. Vision*, 1999, pp. 1150–1157.

[31] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. Conf. Comput. Vision Pattern Recognition*, 2005, pp. 886–893.

[32] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "Brief: Binary robust independent elementary features," in *Proc. Eur. Conf. Comput. Vision*, 2010, pp. 778–792.

[33] A. Alahi, R. Ortiz, and P. Vandergheynst, "Freak: Fast retina keypoint," in *Proc. Conf. Comput. Vision Pattern Recognition*, 2012, pp. 510–517.

[34] S. Krig, *Computer Vision Metrics – Survey, Taxonomy, and Analysis*. New York, NY, USA: Apress, 2014.

[35] H. Lategahn, J. Beck, and C. Stiller, "DIRD is an illumination robust descriptor," in *Proc. Intell. Vehicles Symp.*, 2014, pp. 756–761.

[36] H. Hirschmüller, "Accurate and efficient stereo processing by semi-global matching and mutual information," in *Proc. Conf. Comput. Vision Pattern Recognition*, 2005, pp. 807–814.

[37] B. Ranft, T. Schönwald, and B. Kitt, "Parallel matching-based estimation—A case study on three different hardware architectures," in *Proc. Intell. Vehicles Symp.*, 2011, pp. 1060–1067.

[38] *Zed—3D Camera for AR/VR and Autonomous Navigation*, Stereolabs Inc., San Francisco, CA, USA. (2016). [Online]. Available: https://www.stereolabs.com/zed/specs/

[39] *SGMStereovisionFPGAIP*, Supercomputing Systems AG, Zurich, Switzerland. (2013). [Online]. Available: http://www.scs.ch/blog/en/2013/01/sgm-stereovision-fpga-ip-2/

[40] *3DV-ESystem*, VisLab, Parma, Italy. (2015). [Online]. Available: http://vislab.it/products/3dv-e-system/

[41] A. Geiger, M. Roser, and R. Urtasun, "Efficient large-scale stereo matching," in *Proc. Asian Conf. Comput. Vision*, 2010, pp. 25–38.

[42] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 11, pp. 1222–1239, 2001.

[43] P. F. Felzenszwalb, and D. P. Huttenlocher, "Efficient belief propagation for early vision," *Int. J. Comput. Vision*, vol. 70, no. 1, pp. 41–54, 2006.

[44] D. Gallup, J.-M. Frahm, P. Mordohai, Q. Yang, and M. Pollefeys, "Real-time plane-sweeping stereo with multiple sweeping directions," in *Proc. Conf. Comput. Vision Pattern Recognition*, 2007, pp. 1–8.

[45] N. Einecke and J. Eggert, "Block-matching stereo with relaxed fronto-parallel assumption," in *Proc. Intell. Vehicles Symp.*, 2014, pp. 700–705.

[46] N. Einecke and J. Eggert, "A multi-block-matching approach for stereo," in *Proc. Intell. Vehicles Symp.*, 2015, pp. 585–592.

[47] B. Ranft and T. Strauss, "Modeling arbitrarily oriented slanted planes for efficient stereo vision based on block matching," in *Proc. IEEE 17th Int. Conf. Intell. Transp. Syst.*, 2014, pp. 1941–1947.

[48] K. Yamaguchi, D. McAllester, and R. Urtasun, "Efficient joint segmentation, occlusion labeling, stereo and flow estimation," in *Proc. Eur. Conf. Comput. Vision*, 2014, pp. 756–771.

[49] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. sstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, 2012.

[50] F. Guney and A. Geiger, "Displets: Resolving stereo ambiguities using object knowledge," in *Proc. Conf. Comput. Vision Pattern Recognition*, 2015, pp. 4165–4175.

[51] K. Yamaguchi, D. McAllester, and R. Urtasun, "Robust monocular epipolar flow estimation," in *Proc. Conf. Comput. Vision Pattern Recognition*, 2013, pp. 1862–1869.

[52] B. Kitt and H. Lategahn, "Trinocular optical flow estimation for intelligent vehicle applications," in *Proc. IEEE 15th Int. Conf. Intell. Transp. Syst.*, 2012, pp. 300–306.

[53] C. Tomasi and T. Kanade, *Detection and Tracking of Point Features*. Pittsburgh, PA, USA: School Comput. Sci., Carnegie Mellon Univ., 1991.

[54] B. K. P. Horn and B. G. Schunck, "Determining optical flow," *Artif. Intell.*, vol. 17, pp. 185–203, 1981.

[55] D. Sun, S. Roth, and M. J. Black, "A quantitative analysis of current practices in optical flow estimation and the principles behind them," *Int. J. Comput. Vision*, vol. 106, no. 2, pp. 115–137, 2014.

[56] C. Zach, T. Pock, and H. Bischof, "A duality based approach for realtime TV-l1 optical flow," in *Proc. German Conf. Pattern Recognition*, 2007, pp. 214–223.

[57] J. Braux-Zin, R. Dupont, and A. Bartoli, "A general dense image matching framework combining direct and feature-based costs," in *Proc. IEEE Int. Conf. Comput. Vision*, 2013, pp. 185–192.

[58] S. Hermann and R. Klette, "Hierarchical scan line dynamic programming for optical flow using semi-global matching," in *Proc. Asian Conf. Comput. Vision Workshops*, 2012, pp. 556–567.

[59] J. Revaud, P. Weinzaepfel, Z. Harchaoui, and C. Schmid, "Epicflow: Edge-preserving interpolation of correspondences for optical flow," in *Proc. Conf. Comput. Vision Pattern Recognition*, 2015, pp. 1164–1172.

[60] R. Timofte and L. Van Gool, "Sparseflow: Sparse matching for small to large displacement optical flow," in *Proc. IEEE Winter Conf. Appl. Comput. Vision*, 2015, pp. 1100–1106.

[61] C. Rabe, T. Müller, A. Wedel, and U. Franke, "Dense, robust and accurate 3D motion field estimation from stereo image sequences," in *Proc. Eur. Conf. Comput. Vision*, 2010, pp. 582–595.

[62] M. Hornacek, A. Fitzgibbon, and C. Rother, "Sphereflow: 6 DoF scene flow from RGB-D pairs," in *Proc. IEEE Conf. Comput. Vision Pattern Recognition*, 2014, pp. 3526–3533.

[63] C. Vogel, K. Schindler, and S. Roth, "Piecewise rigid scene flow," in *Proc. IEEE Int. Conf. Comput. Vision*, 2013, pp. 1377–1384.

[64] F. Fraundorfer, C. Wu, and M. Pollefeys, "Combining monocular and stereo cues for mobile robot localization using visual words," in *Proc. 20th Int. Conf. Pattern Recognition*, Aug. 2010, pp. 3927–3930.

[65] K. Konolige and J. Bowman, "Towards lifelong visual maps," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2009, pp. 1156–1163.

[66] G. Grisetti, R. Kümmerle, C. Stachniss, and W. Burgard, "A tutorial on graph-based slam," *IEEE Intell. Transp. Syst. Mag.*, vol. 2, no. 4, PP. 31–43, Winter 2010.

[67] S. Thrun, M. Montemerlo, H. Dahlkamp, D. Stavens, A. Aron, J. Diebel, P. Fong, J. Gale, M. Halpenny, G. Hoffmann, K. Lau, C. Oakley, M. Palatucci, V. Pratt, P. Stang, S. Strohband, C. Dupont, L.-E. Jendrossek, C. Koelen, C. Markey, C. Rummel, J. van Niekerk, E. Jensen, P. Alessandrini, G. Bradski, B. Davies, S. Ettinger, A. Kaehler, A. Nefian, and P. Mahoney, "Stanley: The robot that won the darpa grand challenge," *J. Field Robot.*, vol. 23, no. 9, pp. 661–692, 2006.

[68] C. Urmson, J. Anhalt, D. Bagnell, C. Baker, R. Bittner, M. N. Clark, J. Dolan, D. Duggins, M. Gittleman, S. Harbaugh, Z. Wolkowicki, J. Ziglar, H. Bae, T. Brown, D. Demitrish, V. Sadekar, W. Zhang, J. Struble, M. Taylor, M. Darms, and D. Ferguson, "Autonomous driving in urban environments: Boss and the urban challenge," *J. Field Robot.*, vol. 25, no. 8, 2008, pp. 425–466.

[69] J. Ziegler, P. Bender, M. Schreiber, H. Lategahn, T. Strauss, C. Stiller, T. Dang, U. Franke, N. Appenrodt, C. Keller, E. Kaus, R. Herrtwich, C. Rabe, D. Pfeiffer, F. Lindner, F. Stein,F. Erbs, M. Enzweiler, C. Knöppel, J. Hipp, M. Haueis, M. Trepte, C. Brenk, A. Tamke, M. Ghanaat, M. Braun, A. Joos, H. Fritz, H. Mock, M. Hein, and E. Zeeb, "Making Bertha drive—An autonomous journey on a historic route," *IEEE Intell. Transp. Syst. Mag.*, vol. 6, no. 2, pp. 8–20, Summer 2014.

[70] O. Pink and C. Stiller, "Automated map generation from aerial images for precise vehicle localization," in *Proc. IEEE Int. Conf. Intell. Transp. Syst.*, Madeira, Portugal, Sep..2010, pp. 1517–1522.

[71] N. Mattern and G. Wanielik, "Vehicle localization in urban environments using feature maps and aerial images," in *Proc. 14th Int. IEEE Conf. Intell. Transp. Syst.*, Oct.. 2011, pp. 1027–1032.

[72] H. Lategahn and C. Stiller, "Vision-only localization," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 3, pp. 1246–1257, Jun., 2014.

[73] J. Ziegler, H. Lategahn, M. Schreiber, C. G. Keller, C. Knöppel, J. Hipp, M. Haueis, and C. Stiller, "Video based localization for Bertha," in *Proc. IEEE Intell. Vehicles Symp.*, Jun. 2014, pp. 1231–1238.

[74] V. Devrelle and P. Bonnifait, "iGPS: Global positioning in urban canyons with road surface maps," *IEEE Intell. Transp. Syst. Mag.*, vol. 4, no. 3, pp. 6–18, Fall 2012.

[75] J. Levinson and S. Thrun, "Robust vehicle localization in urban environments using probabilistic maps," in *Proc. IEEE Int. Conf. Robot. Autom.*, Anchorage, AK, USA, May 2010, pp. 4372–4378.

[76] Atlatec localization and mapping sytems, Atlatec GmbH, Karlsruhe, Germany. (2015) [Online]. Available: http://atlatec.de

[77] P. Nelson, W. Churchill, I. Posner, and P. Newman, "From dusk till dawn: Localisation at night using artificial light sources," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2015, pp. 5245–5252.

[78] P. Bender, J. Ziegler, and C. Stiller, "Lanelets: Efficient map representation for autonomous driving," in *Proc. IEEE Intell. Vehicles Symp.*, Jun. 2014, pp. 420–425.

[79] U. Franke, D. Pfeiffer, C. Rabe, C. Knoeppel, M. Enzweiler, F. Stein, and R. Herrtwich, "Making bertha see," in *Proc. IEEE Int. Conf. Comput. Vision Workshops*, Dec. 2013, pp. 214–221.

[80] J. Gräter, T. Schwarze, and M. Lauer, "Robust scale estimation for monocular visual odometry using structure from motion and vanishing points," in *Proc. Intell. Vehicles Symp.*, 2015, pp. 475–480.

[81] C. Golban, P. Cobarzan, and S. Nedevschi, "Direct formulas for stereo-based visual odometry error modeling," in *Proc. IEEE Int. Conf. Intell. Comput. Commun. Process.*, Sep. 2015, pp. 197–202.

[82] M. Bertozzi, A. Broggi, and S. Castelluccio, "A real-time oriented system for vehicle detection," *J. Syst. Archit.*, vol. 43, no. 1, pp. 317–325, 1997.

[83] T. Kalinke, C. Tzomakas, and W. V. Seelen, "A texture-based object detection detection and an adaptive model-based classification," in *Proc. IEEE Int. Conf. Intell. Vehicles*, vol. 1, 1998, pp. 341–346.

[84] C. Hoffmann, T. Dang, and C. Stiller, "Vehicle detection fusing 2D visual features," in *Proc. IEEE Intell. Vehicles Symp.*, Parma, Italy, Jun. 2004, pp. 280–285.

[85] M. Bertozzi and A. Broggi, "GOLD: A parallel real-time stereo vision system for generic obstacle and lane detection," *IEEE Trans. Image Process.*, vol. 7, no. 1, pp. 62–81, Jan. 1998.

[86] F. Flohr and D. Gavrila, "PedCut: An iterative framework for pedestrian segmentation combining shape models and multiple data cues," in *Proc. Brit. Mach. Vision Conf.*, 2013, pp. 66.1–66.11.

[87] C. Papageorgiou and T. Poggio, "A trainable system for object detection in images and video sequences," *Int. J. Comput. Vision*, vol. 38, no. 1, pp. 15–33, 2000.

[88] Z. Sun, G. Bebis, and R. Miller, "On-road vehicle detection using gabor filters and support vector machines," in *Proc. 14th Int. Conf. Digital Signal Process.*, vol. 2, 2002, pp. 1019–1022.

[89] A. Mohan, C. Papageorgiou, and T. Poggio, "Example-based object detection in images by components," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 4, pp. 349–361, Apr. 2001.

[90] M. Pedersoli, J. Gonzalez, X. Hu, and X. Roca, "Toward real-time pedestrian detection based on a deformable template model," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 1, pp. 355–364, Feb. 2014.

[91] P. Bhattacharya and M. L. Gavrilova, "Combining dense features with interest regions for efficient part-based image matching," in *Proc. Int. Conf. Comput. Vision Theory Appl.*, vol. 2, Jan. 2014, pp. 68–75.

[92] S. Sivaraman and M. Trivedi, "Looking at vehicles on the road: A survey of vision-based vehicle detection, tracking, and behavior analysis," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 4, pp. 1773–1795, Dec. 2013.

[93] T. Gandhi and M. Trivedi, "Pedestrian protection systems: Issues, survey, and challenges," *IEEE Trans. Intell. Transp. Syst.*, vol. 8, no. 3, pp. 413–430, Sep. 2007.

[94] T. Scharwächter, M. Enzweiler, U. Franke, and S. Roth, "Efficient multi-cue scene segmentation," in *Pattern Recognition*. New York, NY, USA: Springer, 2013, pp. 435–445.

[95] J. Verbeek and W. Triggs, "Scene segmentation with CRFs learned from partially labeled images," in *Proc. Adv. Neural Inform. Process. Syst.*, 2008, pp. 1553–1560.

[96] J. M. Alvarez, T. Gevers, Y. LeCun, and A. M. Lopez, "Road scene segmentation from a single image," in *Proc. Eur. Conf. Comput. Vision*, 2012, pp. 376–389.

[97] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. 22nd ACM Int. Conf. Multimedia*, Orlando, FL, USA, 2014, pp. 675–678.

[98] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 2nd ed. Upper Saddle River, NJ, USA: Pearson Education, 2003.

[99] M. Cordts, M. Omran, S. Ramos, T. Scharwächter, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset," in presented at the Conf. Comput. Vision Pattern Recognition, Workshop Future Datasets Vision, Boston, MA, USA, 2015.

[100] A. Mukhtar, L. Xia, and T. B. Tang, "Vehicle detection techniques for collision avoidance systems: A review," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 5, pp. 2318–2338, Oct. 2015.

[101] S. Sivaraman and M. M. Trivedi, "Looking at vehicles on the road: A survey of vision-based vehicle detection, tracking, and behavior analysis," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 4, pp. 1773–1795, Dec. 2013.

[102] R. Benenson, M. Omran, J. Hosang, and B. Schiele, "Ten years of pedestrian detection, what have we learned?" in *Proc. Eur. Conf. Comput. Vision Workshops*, 2015, pp. 613–627.

[103] D. Geronimo, A. M. Lopez, A. D. Sappa, and T. Graf, "Survey of pedestrian detection for advanced driver assistance systems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 7, pp. 1239–1258, Jul. 2010.

[104] A. Bar Hillel, R. Lerner, D. Levi, and G. Raz, "Recent progress in road and lane detection: A survey," *Mach. Vision Appl.*, vol. 25, no. 3, pp. 727–745, 2012.

[105] A. Mogelmose, M. M. Trivedi, and T. B. Moeslund, "Vision-based traffic sign detection and analysis for intelligent driver assistance systems: Perspectives and survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 4, pp. 1484–1497, Dec. 2012.

[106] M. B. Jensen, M. P. Philipsen, A. Mogelmose, T. B. Moeslund, and M. M. Trivedi, "Vision for looking at traffic lights: Issues, survey, and perspectives," *IEEE Trans. Intell. Transp. Syst.*, vol. PP, no. 99, pp. 1–16, 2016.

[107] S. Kaplan, A. Guvensan, M. G. Yavuz, and Y. Karalurt, "Driver behavior analysis for safe driving: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 6, pp. 3017–3032, Dec. 2015.

[108] Y. Dong, Z. Hu, K. Uchimura, and N. Murayama, "Driver inattention monitoring system for intelligent vehicles: A review," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 2, pp. 596–614, Jun. 2011.

[109] A. Wedel, H. Badino, C. Rabe, H. Loose, U. Franke, and D. Cremers, "B-spline modeling of road surfaces with an application to free-space estimation," *IEEE Trans. Intell. Transp. Syst.*, vol. 10, no. 4, pp. 572–583, Dec. 2009.

[110] D. Pfeiffer and U. Franke, "Efficient representation of traffic scenes by means of dynamic stixels," in *Proc. Intell. Vehicles Symp.*, 2010, pp. 217–224.

[111] T. Scharwächter, M. Enzweiler, S. Roth, and U. Franke, "Stixmantics: A medium-level model for real-time semantic scene understanding," in *Proc. Eur. Conf. Comput. Vision*, 2014, pp. 533–548.

[112] U. Franke, D. Pfeiffer, C. Rabe, C. Knoeppel, M. Enzweiler, F. Stein, and R. Herrtwich, "Making bertha see," in *Proc. Int. Conf. Comput. Vision Workshops*, 2013, pp. 214–221.

[113] A. Azim and O. Aycard, "Detection, classification and tracking of moving objects in a 3D environment," in *Proc. Intell. Vehicles Symp.*, 2012, pp. 802–807.

[114] T. Gindele, S. Brechtel, J. Schröder, and R. Dillmann, "Bayesian occupancy grid filter for dynamic environments using prior map knowledge," in *Proc. Intell. Vehicles Symp.*, 2009, pp. 669–676.

[115] J. Moras, V. Cherfaoui, and P. Bonnifait, "Credibilist occupancy grids for vehicle perception in dynamic environments," in *Proc. Int. Conf. Robot. Autom.*, 2011, pp. 84–89.

[116] D. Nuss, T. Yuan, G. Krehl, M. Stübler, S. Reuter, and K. Dietmayer, "Fusion of laser and radar sensor data with a sequential monte carlo Bayesian occupancy filter," in *Proc. Intell. Vehicles Symp.*, 2015, pp. 1074–1081.

[117] P. Lenz, J. Ziegler, A. Geiger, and M. Roser, "Sparse scene flow segmentation for moving object detection in urban environments," in *Proc. IEEE Intell. Vehicles Symp.*, 2011, pp. 926–932.

[118] F. Kuhnt, R. Kohlhaas, T. Schamm, and M. Zöllner, J. "Towards a unified traffic situation estimation model—Street-dependent behaviour and motion models," in *Proc. Int. Conf. Inform. Fusion*, 2015, pp. 1223–1229.

[119] S. Lefèvre, D. Vasquez, and C. Laugier, "A survey on motion prediction and risk assessment for intelligent vehicles," *ROBOMECH J.*, vol. 1, no. 1, pp. 1–14, 2014.

[120] A. Bar Hillel, R. Lerner, D. Levi, and G. Raz, "Recent progress in road and lane detection: A survey," *Mach. Vision Appl.*, vol. 25, no. 3, pp. 727–745, 2014.

[121] A. Chen, A. Ramanandan, and J. A. Farrell, "High-precision lane-level road map building for vehicle navigation," in *Proc. Position Location Navigat. Symp.*, 2010, pp. 1035–1042.

[122] H. Loose and U. Franke, "B-spline-based road model for 3D lane recognition," in *Proc. Intell. Transp. Syst. Conf.*, 2010, pp. 91–98.

[123] D. Scharstein and R. Szeliski, "High-accuracy stereo depth maps using structured light," in *Proc. Conf. Comput. Vision Pattern Recognition*, 2003, pp. 195–202.

[124] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nesic, X. Wang, and P. Westling, "High-resolution stereo datasets with subpixel-accurate ground truth," in *Proc. German Conf. Pattern Recognition*, 2014, pp. 31–42.

[125] A. Wedel, C. Rabe, T. Vaudrey, T. Brox, U. Franke, and D. Cremers, "Efficient dense scene flow from sparse or dense stereo data," in *Proc. Eur. Conf. Comput. Vision*, 2008, pp. 739–751.

[126] D. Pfeiffer, S. Gehrig, and N. Schneider, "Exploiting the power of stereo confidences," in *Proc. IEEE Conf. Comput. Vision Pattern Recognition*, 2013, pp. 297–304.

[127] G. Pandey, J. McBride, and R. Eustice, "Ford campus vision and lidar data set," *Int. J. Robot. Res.*, vol. 30, no. 13, pp. 1543–1552, 2011.

[128] J.-L. Blanco, F.-A. Moreno, and J. González-Jiménez, "The málaga urban dataset: High-rate stereo and lidars in a realistic urban scenario," *Int. J. Robot. Res.*, vol. 33, no. 2, pp. 207–214, 2014.
[129] T. Veit, J.-P. Tarel, P. Nicolle, and P. Charbonnier, "Evaluation of road marking feature extraction," in *Proc. Intell. Transp. Syst. Conf.*, 2008, pp. 174–181.
[130] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 743–761, Apr. 2012.
[131] F. Flohr and D. M. Gavrila, "PedCut: An iterative framework for pedestrian segmentation combining shape models and multiple data cues," in *Proc. Brit. Mach. Vision Conf.*, 2013.
[132] N. Schneider and D. M. Gavrila, "Pedestrian path prediction with recursive Bayesian filters: A comparative study," in *Proc. German Conf. Pattern Recognition*, 2013, pp. 174–183.
[133] C. Caraffi, T. Vojir, J. Trefny, J. Sochman, and J. Matas, "A system for real-time detection and tracking of vehicles from a single car-mounted camera," in *Proc. Intell. Transp. Syst. Conf.*, 2012, pp. 975–982.
[134] K. Matzen and N. Snavely, "Nyc3dcars: A dataset of 3D vehicles in geographic context," in *Proc. IEEE Int. Conf. Comput. Vision*, 2013, pp. 761–768.
[135] S. Morales and R. Klette, "Ground truth evaluation of stereo algorithms for real world applications," in *Proc. Asian Conf. Comput. Vision Workshops*, 2010, pp. 152–162.
[136] P. Xu, F. Davoine, J.-B. Bordes, H. Zhao, and T. Denoeux, "Information fusion on oversegmented images: An application for urban scene understanding," in *Proc. Int. Conf. Mach. Vision Appl.*, 2013, pp. 189–193.
[137] T. Scharwächter, M. Enzweiler, U. Franke, and S. Roth, "Efficient multi-cue scene segmentation," in *Proc. German Conf. Pattern Recognition*, 2013, pp. 435–445.
[138] E. Rehder, H. Klöden, and C. Stiller, "Head detection and orientation estimation for pedestrian safety," in *Proc. Intell. Transp. Syst. Conf.*, 2014, pp. 2292–2297.
[139] K. Asanovic, R. Bodik, B. C. Catanzaro, J. J. Gebis, P. Husbands, K. Keutzer, D. A. Patterson, W. L. Plishker, J. Shalf, S. W. Williams, and K. A. Yelick, "The landscape of parallel computing research: A view from Berkeley," EECS Dept., Univ. California, Berkeley, CA, USA, Tech. Rep. UCB/EECS-2006-183, 2006.
[140] B. Ranft, O. Denninger, and P. Pfaffe, "A stream processing framework for on-line optimization of performance and energy efficiency on heterogeneous systems," in *Proc. Int. Parallel Distrib. Process. Symp. Workshops*, 2014, pp. 1039–1048.
[141] S. K. Gehrig, F. Eberli, and T. Meyer, "A real-time low-power stereo vision engine using semi-global matching," in *Proc. Int. Conf. Comput. Vision Syst.*, 2009, pp. 134–143.
[142] J. Borrmann, F. Haxel, D. Nienhüser, A. Viehl, J. Zöllner, O. Bringmann, and W. Rosenstiel, "Stellar—A case-study on systematically embedding a traffic light recognition," in *Proc. Intell. Transp. Syst. Conf.*, 2014, pp. 1258–1265.
[143] A. Tawari, S. Sivaraman, M. Trivedi, T. Shannon, and M. Tippelhofer, "Looking-in and looking-out vision for urban intelligent assistance: Estimation of driver attentive state and dynamic surround for safe merging and braking," in *Proc. IEEE Intell. Vehicles Symp.*, Jun. 2014, pp. 115–120.

Benjamin Ranft studied mechatronics from Karlsruhe Institute of Technology, Karlsruhe, Germany, and Purdue University, West Lafayette, IN, USA. Since 2009, he has been working as a Research Scientist at the "Mobile Perception Systems" Department, FZI Research Center for Information Technology, Karlsruhe, and visited Institut Eurécom, France for four months. His research interests include automatic adaptation of stereo vision and other "early vision" applications to heterogeneous, parallel processors. Since 2014, he has been working as a Manager for the aforementioned department.

Christoph Stiller received the Diploma degree in electrical engineering from Aachen, Germany, and Trondheim, Norway. He received the Dr.-Ing. degree (Ph.D.) in 1994. In 1988, he became a Scientific Assistant at Aachen University of Technology. He spent a PostDoc year at INRS in Montreal, QC, Canada. In 1995, he joined the Corporate Research and Advanced Development of Robert Bosch GmbH, Hildesheim, Germany. In 2001, he became a Chaired Professor at Karlsruhe Institute of Technology, Karlsruhe, Germany. In 2010, he spent three months by invitation at CSIRO in Brisbane, Qld., Australia. In 2015, he spent a four month sabbatical with Bosch RTC and Stanford University in California.

Dr. Stiller was the President of the IEEE Intelligent Transportation Systems Society (2012–2013) and was the Vice-President since 2006. He served as the Editor-in-Chief of the IEEE INTELLIGENT TRANSPORTATION SYSTEMS MAGAZINE (2009–2011), as a Senior Editor of the IEEE TRANSACTIONS ON INTELLIGENT VEHICLES (2016–ongoing) and as an Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING (1999–2003), for the IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS (2004–ongoing), and for the IEEE INTELLIGENT TRANSPORTATION SYSTEMS MAGAZINE (2012–ongoing).

His autonomous vehicle AnnieWAY has been the Finalist in the Urban Challenge 2007, in the USA, and the Winner of the Grand Cooperative Driving Challenge 2011, in the Netherlands. In collaboration with Daimler AG, the team realized the automated Bertha-Benz-memorial-tour in 2013. He is the Coordinator of the German Science Foundation's focus programme on cooperatively interacting automobiles.