Photogrammetric Week '13 Dieter Fritsch (Ed.) Wichmann/VDE Verlag, Belin & Offenbach, 2013

How Cars Learned to See

UWE FRANKE, STEFAN GEHRIG, Daimler AG Research & Development, Böblingen

ABSTRACT

Stereo Vision is a key technology for modern Driver Assistance Systems aiming at increased traffic safety and driving comfort. The opportunity to directly measure the depth of pixels and objects is crucial for understanding the current traffic situation and to detect risky situations. Inspired by results obtained in photogrammetry, we developed a stereo vision system that has recently been introduced to the market in the Mercedes-Benz S-class and E-class.

1. INTRODUCTION

In 2010, around 1.24 Million people died in traffic accidents worldwide. Roughly 50% are vulnerable road users, i.e. pedestrians, bicyclists and motorcyclists. This number demands for active driver assistance systems that help the driver to avoid accidents with these highly endangered traffic participants. A prerequisite is a deep understanding of the current traffic situation, in particular precise measurements of position and motion of potential victims.

Since 1996, we have been working on stereo vision based scene understanding as the basis for future intelligent vehicles. In 2013, Mercedes-Benz introduced a stereo camera system in their S- and E-class vehicles that, in combination with radar sensors, offers

- An active braking assistant reacting to crossing traffic of any type.
- Pedestrian collision avoidance up to 50km/h by *autonomous* braking.
- Lane Keeping up to 200km/h even under adverse weather conditions.
- Low speed *autonomous* driving in traffic jams.
- Magic Body Control (active body control utilizing the measured road profile).

In Fig. 1 the new camera system mounted behind the windshield is clearly visible. It has a baseline of 22cm and a horizontal angle of view of about 50deg. All electronic parts are integrated in the camera module. The heart of these systems is the depth estimation based on Semi-Global Matching (SGM) introduced by Hirschmüller in 2005 [HH05]. An improved variant of this scheme was successfully implemented on automotive compliant hardware (FPGA) in 2008 [SG09]. A fruitful coop-

Figure 1: Mercedes-Benz S-class vehicle with stereo camera system behind the wind shield.

eration between Daimler and DLR allowed the exchange of ideas and further improvements. The current FPGA is used by DLR to guide an unmanned aerial vehicle.

In this paper, we highlight the similarities between photogrammetry and our area of computer vision. At the same time, we sketch the specific challenges of stereo vision in vehicles. In addition, we present the state-of-the art in computer vision for intelligent vehicles as implemented in the new cars, in particular the estimation of pose and motion of independently moving objects by means of 6D-Vision.

2. STEREO VISION FOR INTELLIGENT VEHICLES

Stereo vision has been an active area of research for many decades. It consists of mainly two tasks: Calibrating two or more cameras in the first step, and finding corresponding points within different cameras in the second step. Applications in close-range photogrammetry prefer convergent camera setups that converge at the region of interest using photogrammetric markers. This camera setup changes for natural environments when the views of corresponding points should be as similar as possible, leading to parallel stereo setups with moderate baselines. This aligned stereo geometry is also used for stereo-based driver assistance.

The task of finding corresponding points in stereo vision is reduced to a one-dimensional search along the epipolar line when the standard stereo geometry is obtained after a rectification step. Classic correlation-based approaches have been used frequently to solve this task, e.g. [UF96]. Real-time solutions are available since every correspondence is computed independently. Recent years show a trend towards using global stereo algorithms that optimize all disparities in the image jointly, exploiting the fact that the scene consists of smooth 3D structures with few depth discontinuities. [MB13] hosts a list of the current state-of-the-art stereo algorithms where global stereo algorithms dominate the top 10 ranking.

Stereo vision for intelligent vehicles shares many requirements also common in photogrammetry, e.g. precise localization of object edges, trying to minimize the foreground fattening effect of classical correlation approaches. The assumption of mainly smooth 3D structure holds for both applications and low-contrast areas should also be measured, rendering global stereo algorithms attractive.

However, there are several requirements in stereo vision that are to a certain extent unique to the intelligent vehicles field:

- The range of measurement covers a large area, currently typical from 2m to 50m.
- Precise sub-pixel interpolation is required, since the stereo baseline is constrained by design and packaging issues. Unfortunately, camera resolution and imager size is limited too due to night-time performance requirements and costs.
- The so-called pixel-locking effect limits the available sub-pixel accuracy [MS01].
- Above requirements lead to high accuracy demands for the calibration process while the cameras are being exposed to strong temperature changes and vibrations.
- An online calibration is necessary that works without signalized points and measures the relative orientations of the camera continuously. Calibration has to be guaranteed for the whole lifetime of the car. At the same time, disparity estimation should be insensitive to small errors [HH09].

- Besides accuracy of the disparity estimate, the robustness of the result is of uttermost importance, since the system should also operate under adverse weather conditions such as snow, rain, fog, and any combination thereof.
- Real-time performance is mandatory, while frame rates above 15Hz are required.
- Since the processing should be done in the camera box for cost reasons, power dissipation turns out to be a serious problem since it heats the imagers. Cooling is not permitted and – unfortunately – the camera is mounted behind the wind shield where heating by sun is maximal.
- Rolling shutter cameras are used to obtain a more sensitive sensor, leading to very high accuracy demands for camera synchronization.
- The ultimate goal is to detect relevant objects; the precise 3D-reconstruction is rarely needed. Different applications have different demands, e.g. the exact shape outline of a vehicle is not needed while the height of the street curb is an important quantity.
- Additional constraints such as expected orientation of objects [DP11] or expected scene content [SG07] can be exploited in the disparity estimation and interpretation process.
- Besides the actual measurement, a reliability estimate of confidence of the measurement is also important.

We focus on the topics sub-pixel estimation, robustness, and real-time performance in the following.

2.1. Disparity Estimation and Challenges

2.1.1. Sub-pixel Estimation

Due to limitations in baseline and camera-resolution, sub-pixel estimation is mandatory for intelligent vehicles to achieve the desired measurement range. Classic sub-pixel interpolation based on correlation exhibits the pixel-locking effect, first described by [MS01]. The effect is visualized in Figure 2a, red depicts the ground truth (a rendered scene), and blue shows the estimated disparity with correlation. For global stereo algorithms that accumulate correlation costs, the subsequent sub-pixel interpolation suffers from even more severe pixel-locking (see Figure 2b). This is due to fact that the accumulated costs also reflect smoothness penalties. Since precise motion estimation suffers from this deficiency, object tracking requires additional depth estimation e.g. using iterative sub-pixel interpolation schemes similar to the Kanade-Lucas-Tomasi tracker [JS94].

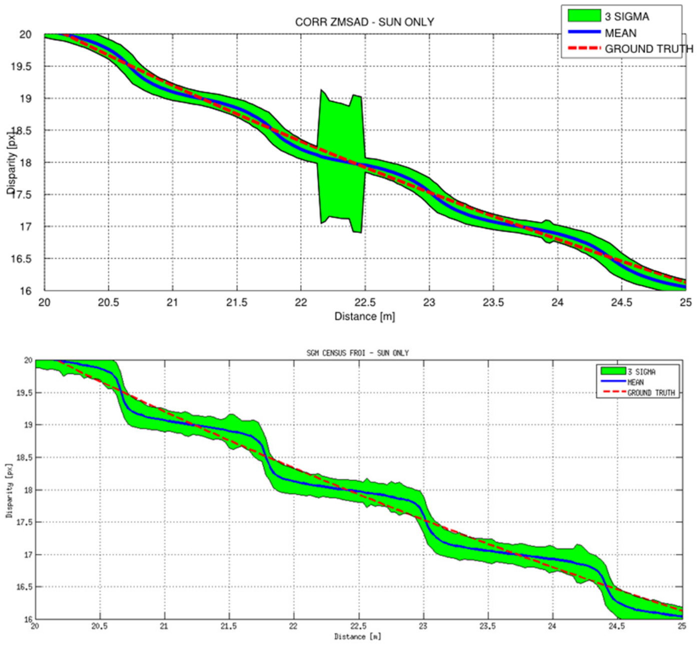

Figure 2: Pixel locking effect for a) a pyramidal correlation stereo and b) SGM with Census costs.

2.1.2. Robustness

The disparity estimation is directly used to determine distance to objects and especially gross errors in this estimation may lead to unnecessary emergency braking maneuvers, which must be avoided under all circumstances. Therefore, a robust disparity estimation is needed. Here, our selected stereo algorithm semi-global matching (SGM) [HHr05] and other global algorithms have advantages since they perform a global disparity optimization whereas local correlation-based approaches estimate the disparity independently. A performance evaluation comparing two local stereo approaches with SGM showed a clear advantage in accuracy and detection rate of the leader vehicle. In a data set of 21000 different situations taken under various lightning conditions, a stereo based vehicle detection scheme optimized to work on depth data of a local correlation scheme missed 5% of all vehicles, whereas the same detector missed only 0.5% if SGM was in use.

An example comparison on a typical street scene shows errors on top of the bridge and a clean dense reconstruction for SGM (see Figure 3 left). Here, many post-processing checks described in the literature for correlation were applied resulting in fewer 3D measurements and still erroneous measurements occur. Figure 3 on the right shows the result on the same scene with SGM, a dense reconstruction without any obvious measurement errors.

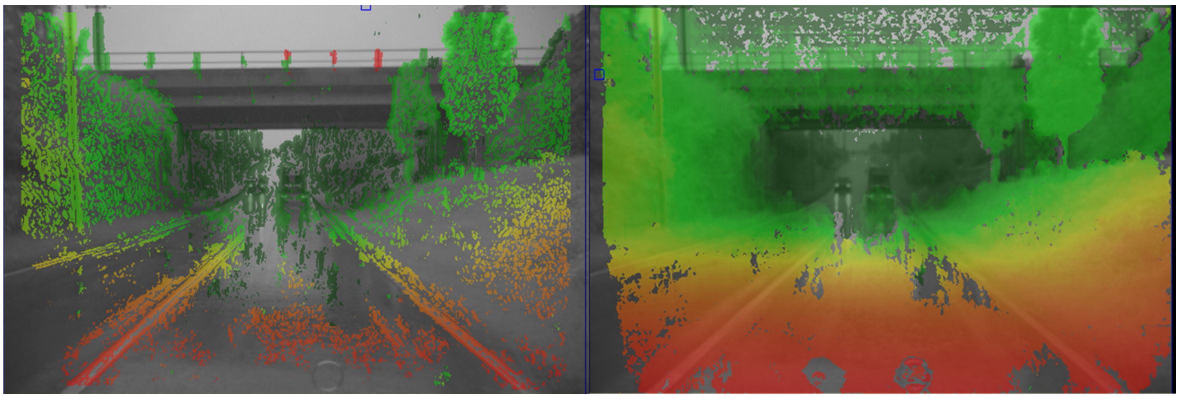

Figure 3: Disparity estimation with correlation (left) and SGM (right). Red pixels are nearby, green are far away, “no color” means that the disparity estimation was not successful.

2.1.3. Real-time Stereo Vision

Early real-time stereo algorithms (using a standard single core 400MHz Pentium) employed correlation-based approaches (e.g. [UF96]). Besides correlation, stereo algorithms based on dynamic programming were published that achieved near real-time performance, however, at the expense of streaking artifacts since consecutive lines are optimized independently. Semi-global matching performs disparity optimization among multiple 1D paths avoiding such artifacts and results in robust disparity estimates. The first implementation of SGM in real-time was published in 2009 running on reconfigurable hardware (FPGA) [SG09]. The actual implementation operates at half-resolution and computes parts of the image with full resolution in a second pass. A typical result is shown in Fig. 3b.

2.2. 6D-Vision

Emergency braking systems require detection of moving objects within a few frames and estimating their motion state. While Radar based systems reliably detect longitudinally moving objects and are common already available in mid-class vehicles, lateral moving objects are the domain for stereo vision. The basic principle is to track single pixel over time and to recursively determinate image position and disparity. Assuming linear motion, an Extended Kalman Filter is able to estimate position and motion of tracked points simultaneously. The 6-dimensional state vector leads to the name "6D-Vision" [UF05].

Fig. 4 illustrates the performance of the approach that forms the basis for the emergency systems of the new vehicles. The arrows indicate the predicted position of tracked points within 500msec. The time between the first and last image is 240msec. In practice, 200msec are sufficient to detect dangerous situations.

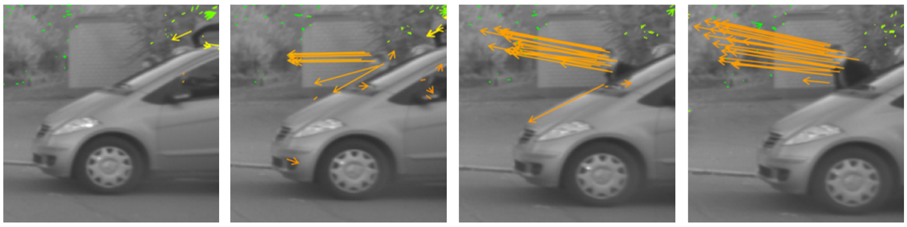

Figure 4: Image sequence with 6D-Vision result. The shown arrows point to the predicted position in 0.5s.

One prerequisite of 6D-Vision is a module that is able to find correspondences in an image sequence. Such modules are referred to as Optical Flow algorithms in the literature. One specific challenge in the automotive domain is the need to detect very large displacements that occur when fast moving objects cross the street. We developed an Optical Flow algorithm that is able to compute arbitrary large displacements in image sequences while being computationally efficient [FS04]. This algorithm is used for 6D-Vision in the new S- and E-class.

2.3. Pedestrian recognition

In addition to stereo vision, a classifier is employed to determine whether the considered object is a pedestrian or not. This independent step decreases the risk of detecting a false positive object significantly. As a consequence, the current E-class and S-class models apply the brake autonomously if a pedestrian is detected in the critical path.

3. DEVELOPMENTS FOR FUTURE VEHICLES

3.1. The Stixel-World

Given the disparity image, the task is to analyze the depth data in order to extract moving objects, to estimate their pose, size and motion, to determine the free-space, to detect curbs, to recognize pedestrians etc. In order to save bandwidth and to reduce the computational burden of subsequent processing steps, we developed the so called Stixel World [DP11].

As shown in Fig. 5, the Stixel World is an efficient super-pixel representation of 3D traffic scenes using vertically oriented rectangles with a fixed width and a variable height. The Stixel World tries to approximate the disparity image column-wise with a minimum number of Stixels taking into account different physically motivated world model priors, such as gravity and ordering constraints. The optimal solution is is achieved through the use of dynamic programming.

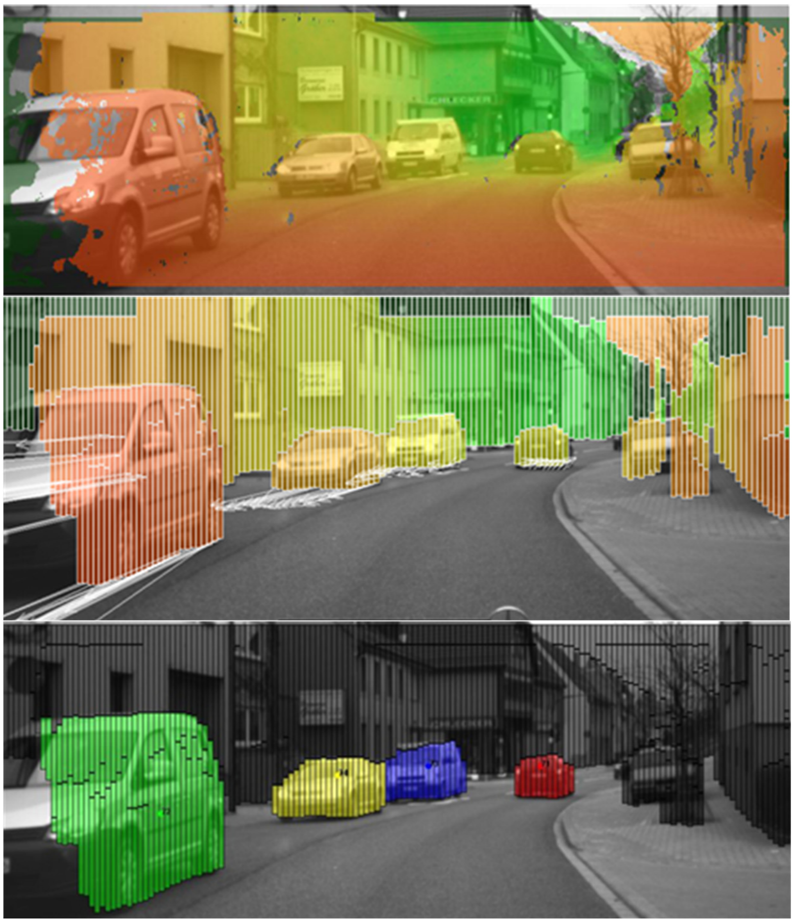

Figure 5 Top: Disparity image. Middle: Visualization of the Stixel-World, the white arrows indicate the motion state obtained by 6D-Vision. Below: Result of the optimal grouping, black indicates static background; different colors encode individual moving objects.

The relevant information in the scene is represented with a few hundreds of Stixels instead of half a million individual stereo depth measurements. At the same time, Stixels give easy access to the most task-relevant information such as freespace and obstacles and thus bridge the gap between low-level (pixel-based) and high-level (object-based) vision.

If the Stixels are tracked over time, 6D-Vision as sketched above allows to estimate lateral and longitudinal motion of each Stixel. This is the basis for subsequent grouping of the Stixels into independently moving objects and the static background. Fig. 5 shows the result obtained by GraphCuts [FE13]. As the graph usually consists of less than a thousand Stixels, real-time performance of this algorithm is no problem.

The described processing pipeline with stixels is also augmented by using confidences from the disparity estimation step. Recent work has shown that a reduction of false positive stixels by a factor of six is obtained maintaining almost the same detection rate [DP13].

3.2. Future Systems

The new ability to precisely measure depth in front of the vehicle paves the way towards further and even more intelligent systems. On the roadmaps are:

- *Supporting the driver in narrow construction sites*. There, driving causes stress to the majority of drivers, as EEG-measurements and interviews clearly show. In particular, if the driving corridor is limited by concrete barriers on the one side and moving cars or trucks on the other side. Future cars will help the driver to keep safe distance to both sides.
- *Collision Avoidance by active steering*. At speeds over 40km/h it might be better to steer than to brake to avoid a collision by steering. In 2009, Daimler demonstrated a car that was

able to decide between both options in the case of a potential collision with suddenly appearing pedestrians.

- *Autonomous Driving*. This dream was in peoples mind in the 60th, pushed early work on driver assistance in the 90th and is on everyone's lips since Google published their work in 2010. Car manufacturers and suppliers work on highly automated driving on highways where the driver can temporarily give full control to the car. This includes automatic lane changes and fast reaction to sudden dangerous situations as the driver will not be able to react appropriately. Therefore, such systems need a high level of redundancy in environment sensing. Stereo vision will play an important role as it is the sensor with the highest spatial solution. However, its performance, in particular the look-ahead distance, has to be increased significantly. In addition, robustness will become a point of major concern.

4. SUMMARY

With the introduction of stereo vision in the current Mercedes Benz cars, stereo-based computer vision is becoming a mass product. Traditionally, Daimler democratizes safety systems. This is also planned for the stereo camera leading to ubiquitous stereo vision. Inspired by research work from photogrammetry, it was possible to implement a high-performance system that will increase road safety significantly already with its first generation. For the first time, especially the vulnerable road users will benefit from this innovation in the area of active safety.

At the same time, the expectations for future generations of camera systems rise, although the technology has just been introduced for the first time to the public. The dream of autonomous driving implies higher resolutions of the imagers, even higher computational power at a given thermal dissipation loss. Above all, those applications demand for a zero-error sensing system. Therefore, robustness of computer vision algorithms will become the key challenge for all of us.

5. ACKNOWLEDGEMENTS

We thank our colleagues Fridtjof Stein, David Pfeiffer, Clemens Rabe, Markus Enzweiler, Friedrich Erbs, and Pierre Greiner who contributed to this article.

6. REFERENCES

[DP10] D. Pfeiffer and U. Franke: Efficient representation of traffic scenes by means of dynamic Stixels. In: IEEE Intelligent Vehicles Symposium, pp. 217-224, San Diego, CA, USA, June 2010.

[DP11] D. Pfeiffer and U. Franke: Towards a global optimal multi-layer Stixel representation of dense 3D data. In: BMVC, Dundee, Scotland, August 2011.

[FE13] F. Erbs, U. Franke: From stixels to objects: A Markov random field approach.

[UF96] U. Franke, I. Kutzbach: Fast Stereo based Object Detection for Stop&Go. In: Intelligent Vehicles '96, Tokyo, 19./20. Sept. 1996, pp. 339-344.

[UF05] U. Franke, C. Rabe, H. Badino, and S. Gehrig: 6D-Vision: Fusion of stereo and motion for robust environment perception. In: Proceedings of the 27th DAGM Symposium, 2005, pp. 216-223.

[SG07] S. Gehrig, U. Franke: Improving Sub-pixel Accuracy for Long-Range Stereo. Workshop VRML, International Conference on Computer Vision, ICCV 2007.

[SG09] S. Gehrig, F. Eberli, and T. Meyer: A real-time low-power stereo vision engine using semi-global matching. In: ICVS 2009, pp. 134-143, October 2009.

[HH05] H. Hirschmüller: Accurate and efficient stereo processing by semi-global matching and mutual information. In: Proceedings of CVPR 05, San Diego, CA. Volume 2. (June 2005), pp. 807-814.

[HH09] H. Hirschmüller, S. Gehrig: Stereo Matching in the Presence of Sub-Pixel Calibration Errors. International Conference on Vision and Pattern Recognition, CVPR 2009.

[MB13] http://vision.middlebury.edu/stereo.

[JS94] J. Shi and C. Tomasi: Good features to track. CVPR 94.

[MS01] M. Shimizu and M. Okutomi: Precise sub-pixel estimation on area-based matching. In: ICCV, pp. 90-97, 2001.

[FS04] F. Stein: “Efficient Computation of Optical Flow Using the Census Transform”. 26th DAGM Symposium on Pattern Recognition, Tübingen, 2004.