Road Surface Reconstruction by Stereo Vision

Hauke Brunken¹ · Clemens Gühmann¹

Received: 8 January 2019 / Accepted: 25 September 2020 / Published online: 18 November 2020
© The Author(s) 2020

Abstract

This paper covers the problem of road surface reconstruction by stereo vision with cameras placed behind the windshield of a moving vehicle. An algorithm was developed that employs a plane-sweep approach and uses semi-global matching for optimization. Different similarity measures were evaluated for the task of matching pixels, namely mutual information, background subtraction by bilateral filtering, and Census. The chosen sweeping direction is the plane normal of the mean road surface. Since the cameras' position in relation to the base plane is continuously changing due to the suspension of the vehicle, the search for the base plane was integrated into the stereo algorithm. Experiments were conducted for different types of pavement and different lighting conditions. Results are presented for the target application of road surface reconstruction, and they show high correspondence to laser scan reference measurements. The method handles motion blur well, and elevation maps are reconstructed on a millimeter-scale, while images are captured at driving speed.

Keywords Stereo vision · Semi-global matching · Plane-sweep · Pavement distress · Surface reconstruction

Zusammenfassung

Rekonstruktion von Straßenoberflächen aus stereoskopischen Bildern. Diese Arbeit behandelt die Rekonstruktion von Straßenoberflächen aus stereoskopischen Bildern, welche von hinter der Windschutzscheibe eines Fahrzeugs montierten Kameras während der Fahrt aufgenommen wurden. Basierend auf dem Plane-Sweep-Ansatz und Semi-global Matching wurde dazu ein spezieller Algorithmus entwickelt. Für den Vergleich von Pixeln wurden drei verschiedene Ähnlichkeitsmaße ausgewertet: Mutual Information, Hintergrundsubtraktion durch bilaterale Filterung und die Census-Transformation. Der Plane-Sweep wird in Richtung der Flächennormale der mittleren Straßenoberfläche durchgeführt. Da sich die Position der Kameras in Bezug zu dieser aufgrund der gefederten Radaufhängung des Fahrzeugs ständig ändert, wurde deren Suche in den Algorithmus integriert. Zur Evaluation wurden Experimente mit unterschiedlichen Straßenbelägen unter variierenden Lichtverhältnissen durchgeführt. Die Ergebnisse zeigen ein hohes Maß an Übereinstimmung mit Laserscan-Referenzmessungen. Höhenunterschiede der Straßenoberfläche werden bei normaler Fahrgeschwindigkeit im Millimetermaßstab aufgelöst.

1 Introduction

For road maintenance, it is important to have current information about road conditions. A common approach to acquire such information is to employ mobile mapping vehicles, which are equipped with LIDAR and laser triangulation

devices (Coenen and Golroo 2017; Eisenbach et al. 2017; Bundesanstalt für Straßenwesen 2018). The disadvantages of specially equipped vehicles are the high costs and the difficulty of keeping information about an entire road network up to date with a reasonable number of vehicles. This work aims to provide a method that is easily integrated into a large number of vehicles and does not require any external installations.

For surface defects, monocular cameras are suitable, as they are able to detect cracks, patches, and potholes. For surface deformations like rutting and shoving, however, 3D information is required (Coenen and Golroo 2017). Although stereo cameras can capture depth information,

✉ Hauke Brunken
hauke.brunken@tu-berlin.de

Clemens Gühmann
clemens.guehmann@tu-berlin.de

¹ Chair of Electronic Measurement and Diagnostic Technology, Technische Universität Berlin, Straße des 17. Juni 135, 10623 Berlin, Germany

few papers have been published about the reconstruction of road surfaces.

Wang and Gong (2002) show how a small patch of a road surface is reconstructed by a classic stereo vision approach. The stereo algorithm does not use any regularization, and only qualitative results are shown. Ahmed et al. (2011) perform a basic study about the suitability of structure-from-motion with close-range pictures of road surfaces and use commercial software packages. They provide promising results, but the approach is not suitable for use in a moving vehicle. The structure from motion approach requires control points on the surface, and the images are ideal in terms of very high resolution and the lack of motion blur. The cameras are manually pointed directly at the region of interest. El Gendy et al. (2011) develop a system based on stereo vision for close-range pictures where the distance to the surface is well below 1 m. They use a box to shield the scene from ambient light. Zhang et al. (2014) describe a stereo vision system that can detect potholes. More recently, Fan et al. (2018) proposed a method based on a stereo camera capable of reconstructing a road's surface. Qualitative results for close-up images are shown. However, it is unclear how well the method works at longer distances while driving.

The publications above show that stereo vision is suitable in principle for road surface reconstruction. In this contribution, the possibilities of stereo vision for this purpose are further investigated. The images are taken from a vehicle at normal driving speed, so that the roads do not have to be closed. For easy integrability, the cameras are mounted behind the vehicle's windshield, and for a high depth precision, the baseline is chosen to be as wide as possible. A method is proposed that makes it possible to reconstruct the road surface with high precision despite these difficult circumstances. Results are validated against a laser scan reference. The developed method offers a cost-effective alternative to mobile mapping vehicles equipped with laser scanning devices.

2 Previous Work

2.1 Stereo Vision

If intrinsic camera parameters and the relative orientation of the individual cameras are known, the problem of depth estimation from stereo images can be broken down into matching pixels in a pair of images. Most stereo algorithms follow four basic steps to solve it:

1. The images of a stereo camera are rectified so that corresponding pixels are located on the same horizontal line.

2. For each pixel of the reference image, a similarity measure for every pixel in a specified disparity range on the same horizontal line of the target image is calculated. The similarity measure is often calculated on a window around the pixel of interest.
3. A smoothing term is introduced, which penalizes jumps in the disparity image.
4. The similarity is maximized while the smoothing term is minimized.

That procedure creates several problems in the application of road surface reconstruction:

1. To cover the lane width with both cameras, they must have convergent viewing directions. Rectification stretches the resulting images and reduces their quality.
2. The disparity is directly linked to the distance between an object and the cameras. A perfectly flat tilted plane has a broad range of disparity values. As a result, a broad range of disparity values has to be searched.
3. The pixels that correspond to a rectangular patch in an image are generally not arranged rectangularly in the other image. It rather depends on the underlying geometry. The compared patches, therefore, do not show the same area.
4. By penalizing jumps in disparity space, a fronto-parallel scene is implied. In the target application, it is known that the underlying geometry is a plane. Since the cameras are located behind the windshield, the plane is not fronto-parallel, but tilted to the cameras.

Other authors have addressed some of these problems. Gang Li and Zucker (2006) integrate prior knowledge about surface normals into the optimization procedure, with surface normals being extracted directly from intensity images (Devnery and Faugeras 1994). Woodford et al. (2009) use second-order smoothness priors to account for slanted surfaces. Second-order smoothness can also be encoded by using 3D-labels as described by Olsson et al. (2013) and applied by Li et al. (2018). Although these methods generally allow for slanted surfaces, they do not favour any particular surface. In the task of surface reconstruction, the surface can be assumed to be a single plane. A drawback of 3D-labels is the enlarged label search space. Ivanchenko et al. (2009) describe the task of surface reconstruction. They convert disparity values to elevation from a ground plane and penalize a change in neighbouring pixels' height. The location of the ground plane has to be known in advance. Sinha et al. (2014) use plane priors in order to pre-estimate disparity values for an arbitrary scene. Afterward, only a small range around that value is searched, but the smoothing term penalizes jumps in disparity space. The algorithm is not optimized for a single plane. A similar approach is described by Fan

et al. (2018), where road surface reconstruction is targeted. They use a seed and grow algorithm, where the disparity is first calculated at the bottom of the image and then propagated to the lines above. Their smoothing term penalizes jumps in disparity space. Both latter methods use sparse image features to find the ground plane. Scharstein et al. (2017) use a smoothing term that is dependent on a local plane hypothesis. That makes it possible to favour slanted surfaces. However, their method works with discrete disparity values and cannot fully account for fractional surface slants. Zhao et al. (2018) do not search through disparity space but discretized heights of a base plane. As a result, they can penalize jumps in height. The location of the base plane is considered prior knowledge.

Except for Li et al. (2018), none of the methods above account for non-corresponding rectangular patches. Recently, Roth and Mayer (2019) proposed a method that addresses that issue by first searching for dominant planes in a scene and then transforming one image to the other's space. They use a smoothing term similar to the one introduced by Ivanchenko et al. (2009). Addressing non-corresponding rectangular patches is especially important if the underlying geometry is a highly slanted plane or if motion blur occurs, as motion blur requires the comparison of large patches to make matching robust.

2.2 Plane-Sweep

In the application of road surface reconstruction, it is known that the underlying geometry resembles a flat surface. In this case, the plane-sweep approach is a natural choice to limit the search space and to take into account the similarity in the height of adjacent pixels. It was first described by Collins (1996). The basic idea is to sweep a virtual plane through 3D space, where it is hit by rays of back-projected feature points that belong to images from at least two different perspectives. The plane is segmented, and the number of rays hitting a segment is counted. If many rays hit a segment, it indicates that the plane is at the corresponding feature's position in 3D space. The result is a 3D point cloud of the feature points.

It is also possible to warp an entire image according to virtual planes to do dense reconstruction. Gallup et al. (2007) test multiple sweeping directions and assign pixels to the best matching one. Their method is especially useful for scenes that consist of multiple elementary planes. Yang et al. (2002) use the plane-sweep approach for dense reconstruction of 3D space from images captured by an array of coplanar cameras. However, the focus lies on real-time implementation, and a smoothing term is not implemented. The sweeping direction is in the cameras' z -direction, which, in the case of a stereo camera, corresponds to a search through disparity space. Irschara et al. (2012) use the plane-sweep approach for multi-image matching of

aerial images in the second step of a structure from motion pipeline and choose the z -direction of the reference camera for plane-sweeping. Total variation is used as a smoothing term, which enables an efficient global optimization algorithm. Bulatov (2015) also uses the plane-sweep approach for multi-image matching and chooses the z -direction of the reference camera for plane-sweeping. He implements a smoothing term and uses a global optimization algorithm.

2.3 Similarity Measure

To match pixels or patches between the left and right images, a measure describing their similarity is required. Hirschmüller and Scharstein (2009) compare different measures on images with radiometric differences. These are expected in the current application due to the baseline of more than 1 m. A large baseline results in different angles between the light source, the road, and the two cameras and, thus, to radiometric differences. According to the results of Hirschmüller and Scharstein (2009), background subtraction by bilateral filtering (BilSub) (Ansar et al. 2004) and hierarchical mutual information (HMI) are good pixel-wise matching costs, but they found that Census as a window-based matching cost outperforms all other measures. These three similarity measures will be tested in the presented application (Sects. 3.3, 5.2 and 5.3).

Recently, convolutional neural networks (CNN) have become a popular choice for estimating the similarity of pixel patches (Zbontar and LeCun 2016). At the time of starting this work, all top-ranking stereo algorithms on the Middlebury stereo evaluation benchmark were using CNNs. Nevertheless, they are not used in this work, because they require a large training set of stereo camera images and ground truth depth information. Creating the training set is time-consuming and costly. Furthermore, the advantage of a CNN in comparing low-textured patches seems limited compared to traditional methods, and as our results show, the traditional methods work sufficiently well.

Current research has shifted to using CNN architectures capable of processing stereo images in a single step (Chang and Chen 2018). In contrast to the CNNs mentioned above, they do not require any subsequent optimization procedure but include this step. They can be adapted for road surface reconstruction (Brunken and Gühmann 2019), but they have the limitation of a reduced image resolution. The proposed method, however, scales very well to large images. The image size influences both the spatial resolution of the final elevation map and its height resolution. Especially if a large part of the road is covered in a single stereo pair, a large image is necessary. Furthermore, these networks also require a large training set.

2.4 Smoothing Term and Optimization

The similarity measure describes how well pixels or patches match. Due to repeating patterns, low textured surfaces, or noise, the measure is ambiguous. If the disparity is chosen based on similarity values alone, noisy disparity maps will be the result. That problem can be addressed by introducing a smoothing term. It encourages neighbouring pixels to have similar disparity values.

Mathematically, the problem is formulated as an energy minimization problem (Szeliski 2011)

$$E(l) = E_{\text{data}}(l) + E_{\text{smooth}}(l). \quad (1)$$

l is the set of disparity values, and E_{smooth} is an energy that measures the negative smoothness of the disparity image. E_{data} describes how well l fits the data. E_{data} is composed by the sum of the dissimilarity measure $D_p(l_p)$ of all pixels P , where l_p is the disparity of the pixel p :

$$E_{\text{data}}(l) = \sum_{p \in P} D_p(l_p). \quad (2)$$

The smoothing term is defined by

$$E_{\text{smooth}}(l) = \sum_{p \in P} \left(\sum_{q \in N_p} f(l_p, l_q) \right). \quad (3)$$

N_p is a neighbourhood of the pixel of interest p . $f(l_p, l_q)$ defines a penalty for differing disparities of neighbouring pixels. Because all pixels indirectly interact with each other, a global optimization problem is defined.

For solving such problems different global optimization methods exist (Szeliski et al. 2008). Some important ones are based on minimization via graph cuts (Boykov et al. 2001). Another popular algorithm is semi-global matching (SGM) (Hirschmüller 2008). The former express the problem as a Markov random field (MRF) and search for a global optimum. The latter accumulates and minimizes costs in various search paths across the disparity image. It is therefore an approximation method for optimizing an MRF. Semi-global matching has the benefit of being easy to implement and parallelize, achieving comparable performance to the former method. For that reason, SGM is used in this paper.

3 Method

The proposed method uses a plane-sweep approach, as this solves the difficulties listed in Sect. 2.1:

1. The images do not need to be rectified.

2. With a sweeping direction orthogonal to the road surface, the volume that has to be searched can be limited to a few centimeters above and below the mean road surface. The reduction of the search space is a key aspect of the proposed algorithm because it reduces the ambiguity when matching pixels between the left and right images.
3. Assuming that compared image patches lie on one of the plane hypotheses, the similarity measure is calculated on correctly transformed patches. That is because the image from one camera is transformed into the space of the other camera.
4. A smoothing term that penalizes jumps in disparity can easily be adjusted to penalize jumps in elevation.

It is assumed that the intrinsic camera parameters and the relative orientation and translation of both individual cameras of the stereo setup are known, and that possible lens distortions have been removed from the images.

Two coordinate systems are defined, which are shown in Fig. 1. The road coordinate system x – y – z is placed such that the x – y -plane coincides with the average road surface. The object coordinate system is set to the point in-between camera centers, such that the x^0 -axis points to the right, the y^0 -axis down, and the z^0 -axis in the viewing direction of the stereo rig. This helps to create an elevation map later on. It is accomplished by splitting the rotation matrix R_C and translation vector T_C , which relate the stereo camera heads to each other and which are a result of stereo camera calibration, into two parts by the square root of a matrix:

$$\begin{pmatrix} R_{C,R} & T_{C,R} \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} R_C & T_C \\ 0 & 1 \end{pmatrix}^{\frac{1}{2}}, \quad (4)$$

where $R_{C,R}$ is the rotation matrix from the object coordinate system to the right camera, and $T_{C,R}$ is the corresponding translation vector. The location and orientation of the left camera are found by

Fig. 1 Virtual planes (unfilled rectangles) are swept in z -direction around the average road surface (grey rectangle) located in the x – y -plane of the road coordinate system. The cameras are inclined towards each other and tilted downwards. The baseline B is the connecting line between the camera centers, which are located at height H

$$R_{C,L} = R_{C,R}^{-1} \text{ and } T_{C,L} = -T_{C,R}. \quad (5)$$

3.1 Road Surface Initialization

To perform the plane-sweep orthogonally to the mean road surface, its location must first be determined. It is roughly approximated and later refined. The first approximate location can be found in two ways:

1. Since the cameras are fixed in the vehicle, the relationship between cameras and the road surface can be easily measured, e. g., by using a tape measure and calculating the rotation angles between the stereo camera and the road.
2. A sparse point cloud can be estimated from image features by triangulation. The best-fitting plane is then found in the point cloud as described in Sect. 3.6. This method often fails because good features are hard to find on road surfaces, but as the vehicle moves forward, an image sequence is captured, and eventually, an image pair with rich features will appear. This approach is used in the experiments (Sect. 5).

The refinement of the surface location is accomplished in two steps that are repeated until convergence:

1. With the approximate location, dense reconstruction, as described in this section, is performed, and a 3D point cloud is generated.
2. A new best-fitting plane is found in the point cloud, as described in Sect. 3.6.

As a coarse approximation of the initial location is sufficient, the approximate location only needs to be found once per stereo camera setup. That also applies to the cameras mounted on the moving vehicle. However, the refinement of the surface's location must be repeated for every image pair. Due to the suspension of the vehicle, the location of the road in relation to the cameras changes constantly.

The refinement of the exact plane location is an essential step of the proposed method. The space that is searched by plane-sweeping (Sect. 3.2) as well as the smoothing term (Sect. 3.4) depend on it.

3.2 Plane-Sweep

Assuming there is a plane parallel to the x - y -plane at height z_i , a point from this plane is projected into the left and right cameras by Collins (1996)

$$\mathbf{H}_{L,i} = \mathbf{K}_L [\mathbf{r}_{L,1} \ \mathbf{r}_{L,2} \ z_i \mathbf{r}_{L,3} + \mathbf{t}_L] \quad (6)$$

and

$$\mathbf{H}_{R,i} = \mathbf{K}_R [\mathbf{r}_{R,1} \ \mathbf{r}_{R,2} \ z_i \mathbf{r}_{R,3} + \mathbf{t}_R]. \quad (7)$$

$\mathbf{K}_L, \mathbf{K}_R$ are the camera matrices of the left and right cameras. The locations and orientations of the cameras are given by the columns of the rotation matrices $\mathbf{r}_{L,\{1,2,3\}}, \mathbf{r}_{R,\{1,2,3\}}$ and the translation vectors $\mathbf{t}_L, \mathbf{t}_R$.

The left image is warped onto the plane with index i and into the geometry of the right camera by applying the mapping that is represented by the plane induced homography:

$$\mathbf{H}_i = \mathbf{H}_{R,i} \cdot \mathbf{H}_{L,i}^{-1}. \quad (8)$$

Warping of the left image is performed for every plane hypothesis in question $I_L \rightarrow I_{LW,i}$. If the virtual plane is at the true location in 3D space for parts of the images, these parts match in the warped left and unchanged right image. The right camera image therefore is the reference image. For every pixel $\mathbf{p}$, a virtual plane must be identified, for which this is the case. The identified plane index i for a specific pixel $\mathbf{p}$ is the label $l_{\mathbf{p}}$, and the set of all $l_{\mathbf{p}}$ is the label image l . With the true l , a new image I_{LW}^* can be assembled from $I_{LW,i}$, which perfectly resembles I_R and hence minimizes $E_{\text{data}}(l)$. The task is to find a label image, such that the total energy $E(l)$ is minimized, i. e. to find a label image that is in accordance with the images and the smoothing term.

To reconstruct the road surface, the x - y -plane is placed so that it coincides with the average road surface. Therefore, the virtual planes are swept in the normal direction of the average surface, and the heights z_i are measured from the average surface. This is shown in Fig. 1. The label image l consists of plane indices for the pixels in the coordinate frame of the right camera image. By converting the indices i to the corresponding heights z_i , l is converted to an elevation image. Later, the elevation image is transformed into an elevation map, i. e. the elevation in the road coordinate system. Its estimation is the objective of the proposed method.

3.3 Similarity Measures

Three different (dis-)similarity measures are evaluated for this task: BilSub, HMI, and Census. Since motion blur causes noise when comparing pixels, the similarity measures are summed on a window around every pixel. That effectively leads to comparing the mean similarity of those windows. The result is a pixel-wise energy, which has to be minimized.

3.3.1 BilSub

BilSub (Ansar et al. 2004) works by subtracting the result of a bilateral filter from both images to reduce radiometric

differences. Afterward, patches of both images are compared by the sum of absolute differences (SAD).

The bilateral filter weights pixels depending on their neighbourhood (Szeliski 2011):

$$g(j, k) = \frac{\sum_{l,m} f(l, m) w(j, k, l, m)}{\sum_{l,m} w(j, k, l, m)}. \quad (9)$$

$f(l, m)$ is the pixel intensity at coordinates (l, m) of an input image. The weighting factor w is the product of a domain kernel $d(j, k, l, m)$, which depends on the pixel distance, and a range kernel $r(j, k, l, m)$, which depends on pixel intensities. In the case of Gaussian kernels these are (Szeliski 2011)

$$d(j, k, l, m) = \exp\left(-\frac{(j-l)^2 + (k-m)^2}{2\sigma_d^2}\right) \quad (10)$$

$$r(j, k, l, m) = \exp\left(-\frac{|f(j, k) - f(l, m)|}{2\sigma_r^2}\right). \quad (11)$$

In Eq. (9), $g(j, k)$ is the pixel intensity of the filtered output image at coordinates (j, k) . σ_d and σ_r are constants.

In our implementation bilateral filtering and subtraction are performed before warping of the left image:

$$I'_{L,R} = I_{L,R} - \text{bilateralFilter}(I_{L,R}). \quad (12)$$

The background subtracted left image is warped using all plane hypotheses i , $I'_L \rightarrow I'_{LW,i}$, and a window around every pixel of every $I'_{LW,i}$ is compared to the window around the corresponding pixel of I'_R by

$$D_{\mathbf{p}}^{\text{BilSub}}(i) = \sum_{\mathbf{q} \in N_{\mathbf{p}}} |I'_{LW,i}(\mathbf{q}) - I'_R(\mathbf{q})|, \quad (13)$$

where $N_{\mathbf{p}}$ is a patch around the pixel $\mathbf{p}$.

3.3.2 Census

The Census transform (Zabih and Woodfill 1994) transforms a set of pixels surrounding a pixel into a bit string by comparing their intensities. A pixel is represented by a binary 0 if its intensity is less than that of the central pixel. Otherwise, it is represented by a binary 1. First, warping of the left image is performed. Then, the Census transform is applied to a patch around every pixel of every $I_{LW,i}$ and on a patch around every pixel of I_R :

$$I_{LW,i}^{\text{CS}} = \text{censusTransform}(I_{LW,i}) \quad (14)$$

$$I_R^{\text{CS}} = \text{censusTransform}(I_R). \quad (15)$$

It is applied after warping to account for the different distortion of the patches. The Hamming distance (denoted by Δ) is the number of bits that differ between two bit strings. It is used to measure the dissimilarity between the bit strings of every pixel of every $I_{LW,i}^{\text{CS}}$ and every corresponding pixel of I_R^{CS} :

$$D_{\mathbf{p}}^{\text{Census}'}(i) = \Delta(I_{LW,i}^{\text{CS}}(\mathbf{p}), I_R^{\text{CS}}(\mathbf{p})). \quad (16)$$

To make the similarity measure more robust against noise, the matching costs for each pixel is determined as the sum of the Hamming distance of pixels within a window centered on that pixel

$$D_{\mathbf{p}}^{\text{Census}}(i) = \sum_{\mathbf{q} \in N_{\mathbf{p}}} D_{\mathbf{p}}^{\text{Census}'}(i). \quad (17)$$

3.3.3 HMI

Mutual information can be utilized as a measure for image similarity (Hirschmüller 2008; Kim et al. 2003):

$$MI_{I_1, I_2} = H_{I_1} + H_{I_2} - H_{I_1, I_2}, \quad (18)$$

where I_1, I_2 are intensity images, H_{I_1, I_2} is their joint entropy and H_{I_1}, H_{I_2} are the entropies of the individual images. The larger the MI is, the more similar I_1 and I_2 are. MI can also be expressed as a sum over all pixels:

$$MI_{I_1, I_2} = \sum_{\mathbf{p}} [h_{I_1}(I_1(\mathbf{p})) + h_{I_2}(I_2(\mathbf{p})) - h_{I_1, I_2}(I_1(\mathbf{p}), I_2(\mathbf{p}))]. \quad (19)$$

The terms h_{I_1}, h_{I_2} and h_{I_1, I_2} are calculated from the estimated probability distributions of pixel intensities. I_1 is I_R in this case. I_2 is assembled from $I_{LW,i}$ by

$$I_2(\mathbf{p}) = I_{LW,i}(\mathbf{p}) \forall \mathbf{p}, \quad (20)$$

such that MI_{I_1, I_2} is maximized. That means the assembled I_2 resembles I_1 . The individual terms in Eq. (19) can be used as pixelwise dissimilarity measures:

$$D_{\mathbf{p}}^{\text{mi}'}(i) = -h_{I_1}(I_{LW,i}(\mathbf{p})) - h_{I_2}(I_R(\mathbf{p})) + h_{I_1, I_2}(I_{LW,i}(\mathbf{p}), I_R(\mathbf{p})). \quad (21)$$

We also take the sum over a rectangular window around each pixel to reduce noise:

$$D_{\mathbf{p}}^{\text{mi}}(i) = \sum_{\mathbf{q} \in N_{\mathbf{p}}} D_{\mathbf{p}}^{\text{mi}'}(i). \quad (22)$$

The problem is that h_{I_1}, h_{I_2} and h_{I_1, I_2} have to be known to calculate and maximize MI_{I_1, I_2} , but h_{I_2} and h_{I_1, I_2} are a result of the maximization. The solution is to assemble I_2 using an

initial label image l_{ini} that corresponds to a flat plane. Then, optimization of MI_{I_1, I_2} , and assembly of I_2 with the new label image l are alternated until convergence.

For computational efficiency, hierarchical mutual information (HMI) uses an image pyramid of downscaled input images. First, images with the lowest resolution are processed. Labels for every pixel are calculated and the resulting label image is upsampled for use in the next level of the pyramid. In this work, the refinement of the road surface location (Sect. 3.6) is integrated into the process. Therefore, the label image is warped according to the new location and upsampled afterward.

3.4 Smoothing Term

The result of the three methods above is a 3D array that saves a matching cost for every pixel of the reference image and every plane hypothesis. The total matching cost is a sum of the pixel-wise matching cost and the smoothing term:

$$E_{\text{smooth}}(l) = \sum_{\mathbf{p} \in P} \sum_{\mathbf{q} \in N_{\mathbf{p}}} K |l_{\mathbf{p}} - l_{\mathbf{q}}|. \quad (23)$$

The label image is a plane index for every pixel and corresponds to a height, or elevation, measured from the average road surface. The smoothing term favours small changes between neighbouring labels and therefore favours a smooth elevation map. As jumps in elevation are not expected on road surfaces, the penalty of differing neighbouring labels is proportional to their difference. The factor K is not changed across the picture as is commonly done [e. g. by Li et al. (2018)], because a change in colour or intensity does not necessarily relate to a change in elevation in the presented application. That can be seen in Fig. 5, where some leaves are squeezed on the surface and are perfectly flat.

3.5 Optimization

Semi-global matching (SGM) (Hirschmüller 2008) is utilized to minimize the total energy $E(l)$. SGM minimizes several search paths across the image, instead of solving the global optimization problem directly. The cost that is summed in a search path is recursively defined by

$$L_{\mathbf{r}}(\mathbf{p}, i) = D_{\mathbf{p}}(i) + \min_{ii} (L_{\mathbf{r}}(\mathbf{p} - \mathbf{r}, ii) + K|i - ii|), \quad (24)$$

where $\mathbf{p}$ is a pixel coordinate, $\mathbf{r}$ is a search path direction, i and ii are the virtual plane indices, and $D_{\mathbf{p}}(i)$ is one of the dissimilarity measures from Sect. 3.3. The reader is referred to Hirschmüller (2008) for the details of SGM. In

our implementation, 16 search path directions are used. The costs of all search paths are summed:

$$S(\mathbf{p}, i) = \sum_{\mathbf{r}} L_{\mathbf{r}}(\mathbf{p}, i), \quad (25)$$

and the final label for a pixel is found by

$$l_{\mathbf{p}} = \arg \min_i S(\mathbf{p}, i). \quad (26)$$

3.6 Road Surface Location Refinement

To perform the plane-sweep, the location of the cameras in relation to the mean road surface is needed (Sect. 3.2). A coarse estimate is used initially, and the dense elevation image estimation is performed, as described in the previous section. The virtual planes must cross at least parts of the real road surface. The parts of the road that are out of reach of the plane-sweep appear as noisy regions in the elevation map and can be filtered by a local variance filter with a threshold. The valid pixels of the depth map are back-projected into 3D space and create a point cloud in which a plane is searched for with a random sample consensus (RANSAC) algorithm (Fischler and Bolles 1981):

The mean of the point cloud is subtracted. A subset of three points is randomly chosen, and a singular value decomposition (SVD) of their coordinates is performed. The result is a rotation matrix R that relates the coordinate system to the plane defined by those points. The rotation matrix is checked for consistency with the road's possible location by converting it into Tait-Bryan angles, which should be within reasonable boundaries. All remaining points are checked to be consistent with the plane found this way and are marked as inliers if the distance to the plane is within a pre-defined threshold. The root-mean-square (RMS) value of distances between the surface and the inliers is calculated, and the plane with the lowest value that has a sufficient ratio of inliers (i.e., more than a pre-defined threshold) is chosen. Then, SVD is repeated with the coordinates of the inliers, and a new plane is found. If the z -axis of the plane points downwards, an additional rotation of 180° around the x -axis is applied to the rotation matrix. Otherwise, the next plane-sweep would take the wrong direction.

As the plane that is found by SVD can be arbitrarily aligned, the rotation matrix is converted into Tait-Bryan angles, the rotation around the z -axis is set to zero, and the angles are transformed into a rotation matrix. That ensures that the elevation map, as shown in Fig. 2, is approximately aligned in the direction of travel.

Fig. 2 Elevation map of dynamic scene 2 from Fig. 10b

The rotation matrix and translation vector in Eq. (6) are a combination of $R_{C,\{L,R\}}$, $T_{C,\{L,R\}}$, the found mean vector T and the rotation matrix R found by SVD.

3.7 Iterations

As described in Sect. 3.1, the steps of refining the mean surface location and performing the plane-sweep are iterated. The iteration starts with images downscaled by a factor of $s = 5$ in both dimensions and is repeated with images downscaled by a factor of $s = 4$ and so forth. At the same time, the search space around the mean surface is reduced in each iteration. This allows a large space to be searched for the correct road surface and makes the algorithm robust against misalignment of the initial surface location.

3.8 Algorithm Overview

Figure 3 gives an overview of the algorithm with hierarchical mutual information as the similarity measure. The input images are downscaled by the factor s in both dimensions (“downscale by s ”), and the heights of the plane hypotheses are chosen according to the reduced resolution (“Choose acc. to s ”). The location of the mean road surface in relation to the stereo camera system is given by R and T . The virtual planes are placed below and above of it, and plane homographies are calculated (“calc. homographies”). The plane-sweep is performed with the left camera image (“plane-sweep”) using the plane homographies. In order to calculate the MI similarity, the final label image is required, i. e. the plane index for every pixel. It is a result of the following SGM optimization (“SGM”) and is found iteratively by alternating the SGM optimization and the MI calculation. At the same time, a point grid is extracted from the label image (“extract point grid”) and the position of the road is refined by RANSAC with SVD (“RANSAC w/SVD”). Since the scaling of the image is adjusted and the estimated location of the mean road surface changes, the label image used to calculate the MI must be adjusted accordingly. This is done by warping and upscaling the label image by a homography (“warp and upscale”). The MI calculation is followed by summing over rectangular windows (“MI and mean filter”). By converting the indices to heights (“convert”), the label image is converted to an elevation image.

4 System Setup

To capture the road in front of the vehicle and create a setup that is easily integrated into any vehicle, the cameras are mounted behind the windshield. A large baseline increases the resolution of a depth measurement. Hence, the baseline is chosen as wide as possible, which is $B = 1.10$ m in the test vehicle (a van). To fix the cameras’ relative position to each other, they are mounted onto a bar, which is, in turn, mounted above the dashboard. The height above ground is $H = 1.40$ m. The cameras are tilted downwards by 12° ,

Fig. 3 Overview of the stereo algorithm with hierarchical mutual information as the similarity measure. For details refer to Sect. 3.8

which is limited by the hood of the vehicle. The cameras could have been mounted at the top of the windshield, but the height is supposed to be comparable to a height that is possible in a sedan.

Two Basler acA1920-150uc global shutter colour cameras are employed. The sensor has an optical size of $2/3''$ with a size of $1920 \text{ pixel} \times 1200 \text{ pixel}$. The pixel size is $4.8 \mu\text{m} \times 4.8 \mu\text{m}$. The lane width, which should be covered by the stereo cameras, is chosen to be 2 m. That requires a 25 mm lens. To synchronize the cameras, they are triggered externally. Due to the small overlapping area of two cameras with a wide baseline, they are both inclined towards each other by 5° . In this way, the viewing directions intersect at a point about 6 m away from the cameras.

Some thought was put into determining the aperture that should be used. The aperture greatly influences the depth of field. The smaller the aperture, the higher the depth of field will be. Typically, this is limited by other effects such as diffraction and depends on the hardware. In still images, the aperture with the highest depth of field can be used. A small aperture requires a higher exposure time, which will cause blurred moving objects. By examining the total area size that emits light which hits a single pixel, it is found that the aperture should be as large as possible for regular driving speeds, which is $f/1.4$ for the lens used. As motion blur cannot be prevented, we took care to have the same amount of motion blur in both images of the stereo pair by using the same exposure time in both cameras. However, it is changed dynamically between frames.

4.1 Elevation Resolution

A change in elevation of a pixel in the reference image I_R can be detected if it changes the coordinate of the corresponding pixel in the left image by a sufficient amount. The change in elevation per shift in pixels is embraced as the elevation map's resolution and is shown in Fig. 2 as contour lines in mm/pixel. It is around 1 mm/pixel up to a distance of 4.5 m from the cameras and reduces to 2.5 mm/pixel at a distance

of 10.5 m. The resolution depends on the plane-sweep direction in 3D space. For comparison, a plane-sweep in viewing direction of the cameras, which is similar to a search through disparity space, yields a resolution of 6 mm/pixel at a distance of 6.0 m. The acute angle between the camera's viewing direction and road surface thus helps increase the elevation map resolution.

These calculations rely on perfectly calibrated pinhole cameras. In the presented application, the cameras are calibrated by OpenCV's stereoCalibrate function, which returns the RMS reprojection error as a quality criterion, which was 0.4 pixels in the experiments. Assuming the stereo algorithm is able to perfectly match pixels and ignoring motion blur, the uncertainty of the predicted elevation is calculated by multiplying the resolution with the RMS re-projecting error (applying linear propagation of uncertainty), resulting in an elevation uncertainty between 0.4 mm and 1.0 mm, depending on camera distance. Since a neighbourhood of pixels is involved in predicting each pixel's elevation, the total uncertainty is further reduced. The stereo method is able to reach subpixel precision by placing plane hypothesis at arbitrary positions.

5 Experiments

In order to validate the proposed method, two static images of different pavements were taken. In addition, two sections of a motorway were filmed several times during a test drive. The results are compared to laser scans. For still images, the aperture is set to $f/8$, and the cameras are mounted on a tripod. For dynamic images taken while driving, the largest possible aperture of $f/1.4$ is selected, and the cameras are mounted behind the windshield. The exposure time is set dynamically. In all cases, in the last iteration, 128 plane hypotheses are used, and the search range is from $z_{\min} = -50 \text{ mm}$ to $z_{\max} = +50 \text{ mm}$. This choice corresponds to a spacing of $\approx 0.8 \text{ mm}$ between the virtual planes and is guided by the theoretical analysis of the elevation image

resolution in Sect. 4.1. The compared window size N_p is 5×5 pixels. The Census transform is calculated on 9×9 pixels, and the parameter K is chosen depending on the similarity measure:

$$K = \begin{cases} 10^{-6} & \text{for HMI} \\ 10^{-2} & \text{for BilSub} \\ 5 & \text{for Census} \end{cases} \quad (27)$$

Both the patch size and the parameter K have been chosen by increasing them until obvious mismatches no longer occur.

5.1 Calculation of the Difference to a Laser Scan

In order to compare the stereo method to a laser scan, 3D coordinates of every pixel of the elevation image are found. As the precise relation between the stereo camera coordinate system and the laser scan coordinate system is unknown, the generated point cloud and the laser scan point cloud are aligned using the software "Cloud Compare" (GPL software 2017). It uses the iterative closest point algorithm (Besl and McKay 1992), in which the compared point cloud is rotated and shifted until the mean square distance between the nearest neighbours of both point clouds is minimal. This approach is problematic if the compared point clouds belong to a planar surface. However, in the experiments, only road surfaces with defects are compared, so the planes are not flat. As both point clouds' resolutions are different and the individual points of both do not lie on top of each other, only the z -components of the distances to the nearest neighbour is considered the differences between the laser scan and the stereo method. The nearest neighbour is searched for only with respect to the distance orthogonal to the plane normal of the mean road surface.

The laser scans of static scenes were conducted using a Z+F IMAGER 5006 h. The scanner has a root-mean-square (RMS) range uncertainty between 0.4 mm and 0.7 mm, depending on the distance to the object and its reflectivity, and uncertainty in the vertical and horizontal direction of 0.007° . The uncertainties are combined to a total uncertainty RMS_{LS} in the direction perpendicular to the road surface. The laser scans of dynamic scenes were conducted with a laser line scanner mounted on a mobile mapping vehicle. The obtained data have an RMS uncertainty in the height direction of 1.4 mm, which was found by repeated measurements.

As the laser scan error and the stereo method error both are in the same order of magnitude, the laser scan cannot be regarded as the ground truth. Besides, the position of the point clouds relative to each other is unknown. Assuming that both methods' errors follow a Gaussian distribution and that the mean of both is zero, the difference between the

laser scan point cloud and the stereo vision point cloud (in the z -direction) also follows a Gaussian distribution:

$$d_z \sim \mathcal{N}(0, \sigma_{LS}^2) - \mathcal{N}(0, \sigma_{SV}^2) \sim \mathcal{N}(0, \sigma_{LS}^2 + \sigma_{SV}^2). \quad (28)$$

As a zero mean value is assumed, the standard deviations σ correspond to the RMS values. Therefore, the RMS of the stereo vision point cloud is recovered by

$$\text{RMS}_{SV} = \sqrt{\text{RMS}_{d_z}^2 - \text{RMS}_{LS}^2}. \quad (29)$$

The mean difference is assumed to be zero because the point clouds are aligned to best match each other, and the point cloud of the laser scan is expected to have zero bias.

Due to the flat perspective, the densities of the extracted point clouds decrease over distance. As points close to the cameras have higher precision than points that are far away, the RMS calculated on all points appears to be very small. Therefore, the RMS is calculated on bins, each of which encloses 50 mm in the direction of travel. In addition, the RMS_{SV} of each bin is divided by the range Δz in which points from the laser scan measurements are encountered in that bin. This relates the accuracy to the underlying geometry and is performed because the proposed method favours flat surfaces. It is therefore a more comparable measure for different surfaces. In Fig. 4 the RMS of the laser scanner RMS_{LS} , the derived RMS of the stereo vision method RMS_{SV} , and the $\text{RMS}_{SV}/\Delta z$ are shown. Furthermore, the interval that encloses 90% of the differences d_z is shown. In Table 1 mean values of RMS_{SV} and $\text{RMS}_{SV}/\Delta z$ are reported, which are denoted by mRMS_{SV} and $\text{mRMS}_{SV}/\Delta z$.

5.2 Static Scenes

Figure 5 shows the results of an asphalt concrete roadway in the foreground and a cobblestone pavement in the background. Figure 5e shows the right camera image and Fig. 5a the reference elevation image obtained by the laser scanner. For an easy comparison, it is shown from the right camera's perspective. In the center of the elevation image, a bump accompanied by a crack can be seen. In the background of the scene, a surface depression can be spotted on the cobblestone pavement. The elevation images obtained by the proposed method with the three different similarity measures are shown in Fig. 5b–d. The difference to the laser scan (Sect. 5.1) is shown in Fig. 5f–h. The bump and the depression are clearly visible, and the difference images indicate a high accuracy of the proposed method.

Using BilSub and HMI as pixel-wise similarity measures, it is possible to reconstruct small objects, whereas Census, as a window-based method, smoothes them out. This can be seen in Fig. 5, where some leaves are reconstructed, and in Fig. 6, where the joints between the tiles are resolved

Fig. 4 Surfaces of the still images are viewed from above (as in Fig. 2) and the RMS values are calculated on bins, each of which encloses 50 mm in the direction of travel. The RMS_{SV} is divided by

the range Δz in which points from the laser scan measurements are encountered in that bin. Furthermore, the interval that encloses 90% of the differences d_z and the elevation resolution are shown

Table 1 Results for different scenes with different settings, v —vehicle speed, t_E —exposure time

Scene	v	f/l	t_E	HMI		BilSub		Census	
				mRMS _{SV}	mRMS _{SV} /Δz	mRMS _{SV}	mRMS _{SV} /Δz	mRMS _{SV}	mRMS _{SV} /Δz
	in km/h		in ms	in mm		in mm		in mm	
Static 1	0	8	40.0	2.1	0.10	2.1	0.10	1.9	0.09
Static 2	0	8	40.0	2.2	0.07	1.7	0.06	1.9	0.06
Dynamic 1	65	1.4	2.89	3.2	0.18	2.0	0.11	2.0	0.11
	70		0.37	1.9	0.11	1.3	0.07	1.3	0.07
	68		0.13	2.1	0.12	1.5	0.08	1.4	0.08
Dynamic 2	67	1.4	2.34	3.6	0.15	1.7	0.08	1.7	0.07
	70		0.38	2.0	0.09	1.2	0.06	1.3	0.06
	71		0.13	2.5	0.12	1.7	0.08	1.6	0.08

Values in bold highlight the best value per row (per test case)

Fig. 5 Static scene 1: **a** ref. elevation image, **b, f** HMI elevation/difference images **c, g** BilSub elevation/difference images, **d, h** Census elevation/difference images, **e** right camera image

Fig. 6 Static scene 2: **a** reference elevation image, **b, f** HMI elevation/difference images **c, g** BilSub elevation/difference images, **d, h** Census elevation/difference images, **e** right camera image

in much more detail. That is caused by the larger windows that are compared using Census. Besides, it is noticeable that shadows and shiny surfaces do not impact the elevation images.

Figure 4a, b show the values RMS_{SV} , $RMS_{SV}/\Delta z$, RMS_{LS} , the interval enclosing 90% of all d_z , and the elevation map resolution over distance for the stereo images from Figs. 5 and 6. At a distance of 9 m (Fig. 4a), the asphalt concrete pavement ends and the cobblestone pavement begins (Fig. 5). As the asphalt concrete pavement has a much smoother surface, the elevation image better fits the plane prior and results in a lower RMS value. Thus, a smooth surface is easier to reconstruct by the proposed method. Occlusions that occur on the cobblestone pavement with the stereo camera and the laser scanner also increase the RMS value.

The average values $mRMS_{SV}$ and $mRMS_{SV}/\Delta z$ are reported in Table 1. BilSub and Census achieve a slightly better performance than HMI, although they are all comparable.

5.3 Dynamic Scenes

Three test drives were carried out on two sections of a motorway. The first trip was made early in the morning with little sunlight. That made long exposure times necessary, with $t_E = 2.89$ ms (Table 1). At a driving speed $v = 65$ km h⁻¹ this corresponds to a motion blur of more than 50 mm, and this in turn to a motion blur of 15 pixels at the lower image edge and to 5 pixels at the upper edge. For the shortest exposure time $t_E = 0.13$ ms it corresponds to a motion blur between

Fig. 7 Dynamic scene 1: long exposure time: **a** reference elevation image, **b, f** HMI elevation/difference images **c, g** BilSub elevation/difference images, **d, h** Census elevation/difference images, **e** right camera image

Fig. 8 Scene 1: medium exposure time: **a** reference elevation image, **b, f** HMI elevation/difference images **c, g** BilSub elevation/difference images, **d, h** Census elevation/difference images, **e** right camera image

0.7 pixel and 0.2 pixel. Figures 7, 8 and 9 show the results for the first section of the motorway with different exposure times (Table 1) that were achieved with different similarity measures.

Even though motion blur is absolutely not negligible at poor lighting conditions, it has a relatively low impact on the results. HMI is more affected by this than the other measures, as can be seen in Fig. 7f and Table 1. If the surface is flat, the direction and amount of motion blur is the same on the warped left and unchanged right image because it is transformed perspectively correct. Thus, blurred textures appear in the same way in both images. It effectively results in a reduced vertical spatial resolution of the input images, as smaller structures can no longer be distinguished. It has a

low impact in the horizontal direction. The horizontal direction, however, is the direction that is primarily searched by plane-sweeping. Thus, the motion blur has little effect on the resolution of the elevation image.

The results for the second section of the motorway are shown in Fig. 10 for Census only. The elevation images are accurate over large regions of the road surface. In this example, the best result is obtained with medium exposure time. Note that the shown sections are not exactly the same across Fig. 10. In Fig. 10c, d the vehicle drove a little further to the left.

For all dynamic scenes, BilSub and Census outperform HMI in terms of the $mRMS_{SV}$ value, as can be seen in Table 1.

Fig. 9 Dynamic scene 1: short exposure time: **a** reference elevation image, **b, f** HMI elevation/difference images **c, g** BilSub elevation/difference images, **d, h** Census elevation/difference images, **e** right camera image

Fig. 10 Dynamic scene 2 evaluated with Census: **a, e** ref. elevation image and corresponding right camera image, **b, f** elevation/difference images, long exposure time (ET) **c, g** elev./diff. images, medium ET, **d, h** elev./diff. images, short ET

6 Conclusion

The application of stereo vision for road surface reconstruction by using a newly proposed algorithm has been presented. The images were taken from a vehicle at normal driving speed with cameras mounted behind the windshield. The method is capable of reconstructing the road surface up to several meters ahead of the vehicle. Three different similarity measures for matching pixels were evaluated for this task. The pixel-wise similarity measures HMI and BilSub show more detailed elevation maps than Census as a window-based measure, and HMI is more

sensitive to motion blur than BilSub and Census. Therefore Bilsub has the best overall performance.

By using the plane-sweep approach, motion blur is handled very well by the proposed method. The generated depth maps can be used for road maintenance. The next steps are developing a classifier for the automatic detection of damages and the investigation of the real-time capability of the algorithm.

Acknowledgements We would like to thank Daniel Wujanz for performing the laser scans of static scenes and helping with its evaluation. The laser scans of dynamic scenes were performed within the BMBF project “Cooperative cloudbased road condition monitoring – Street-Probe”, Funding reference number 01MD16006D.

Funding Open Access funding enabled and organized by Projekt DEAL.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Ahmed M, Haas C, Haas R (2011) Toward low-cost 3D automatic pavement distress surveying: the close range photogrammetry approach. *Can J Civ Eng* 38(12):1301–1313
- Ansar A, Castano A, Matthies L (2004) Enhanced real-time stereo using bilateral filtering. In: *Proceedings. 2nd international symposium on 3D data processing, visualization and transmission, 2004. 3DPVT 2004*, pp 455–462. <https://doi.org/10.1109/TDPVT.2004.1335273>
- Besl PJ, McKay ND (1992) A method for registration of 3-D shapes. *IEEE Trans Pattern Anal Mach Intell* 14(2):239–256. <https://doi.org/10.1109/34.121791>
- Boykov Y, Veksler O, Zabih R (2001) Fast approximate energy minimization via graph cuts. *IEEE Trans Pattern Anal Mach Intell* 23(11):1222–1239
- Brunken H, Gühmann C (2019) Incorporating plane-sweep in convolutional neural network stereo imaging for road surface reconstruction. In: *Proceedings of the 14th international joint conference on computer vision, imaging and computer graphics theory and applications—volume 5: VISAPP, SciTePress*, pp 784–791. <https://doi.org/10.5220/0007352107840791>
- Bulatov D (2015) Temporal selection of images for a fast algorithm for depth-map extraction in multi-baseline configurations. In: *Proceedings of the 10th international conference on computer vision theory and applications, SciTePress, Berlin, Germany*, pp 395–402. <https://doi.org/10.5220/0005239503950402>. <http://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0005239503950402>
- Bundesanstalt für Straßenwesen (2018) Measuring vehicles and systems. https://www.bast.de/durabast/DE/durabast/Messfahrzeuge/fahrzeuge_node.html. Accessed 10 Nov 2018
- Chang JR, Chen YS (2018) Pyramid stereo matching network, p 9. [arXiv:180308669](https://arxiv.org/abs/180308669)
- Coenen TBJ, Golroo A (2017) A review on automated pavement distress detection methods. *Cogent Eng* 4(1):1374822. <https://doi.org/10.1080/23311916.2017.1374822>
- Collins RT (1996) A space-sweep approach to true multi-image matching. In: *Proceedings CVPR IEEE computer society conference on computer vision and pattern recognition, IEEE*, pp 358–363. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=517097
- Devernay F (1994) Computing differential properties of 3-D shapes from stereoscopic images without 3-D models. In: *Proceedings of IEEE conference on computer vision and pattern recognition CVPR-94, IEEE Computer Society Press, Seattle, WA, USA*, pp 208–213. <https://doi.org/10.1109/CVPR.1994.323831>. <http://ieeexplore.ieee.org/document/323831/>
- Eisenbach M, Stricker R, Seichter D, Amende K, Debes K, Sesselmann M, Ebersbach D, Stoeckert U, Gross HM (2017) How to get pavement distress detection ready for deep learning? A systematic approach. In: *2017 international joint conference on neural networks (IJCNN), IEEE, Anchorage, AK, USA*, pp 2039–2047. <https://doi.org/10.1109/IJCNN.2017.7966101>. <http://ieeexplore.ieee.org/document/7966101/>
- El Gendy A, Shalaby A, Saleh M, Flintsch GW (2011) Stereo-vision applications to reconstruct the 3D texture of pavement surface. *Int J Pavement Eng* 12(3):263–273. <https://doi.org/10.1080/10298436.2010.546858>
- Fan R, Ai X, Dahmoun N (2018) Road surface 3D reconstruction based on dense subpixel disparity map estimation. *IEEE Trans Image Process* 27(6):3025–3035. <https://doi.org/10.1109/TIP.2018.2808770>
- Fischler MA, Bolles RC (1981) Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun ACM* 24(6):381–395
- Gallup D, Frahm JM, Mordohai P, Yang Q, Pollefeys M (2007) Real-time plane-sweeping stereo with multiple sweeping directions. In: *2007 IEEE conference on computer vision and pattern recognition, IEEE*, pp 1–8. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4270270
- Gang L, Zucker S (2006) Surface geometric constraints for stereo in belief propagation. In: *2006 IEEE computer society conference on computer vision and pattern recognition—volume 2 (CVPR'06), IEEE, New York, NY, USA*, pp 2355–2362. <https://doi.org/10.1109/CVPR.2006.299>. <http://ieeexplore.ieee.org/document/1641042/>. Accessed 30 April 2020
- GPL software (2017) Cloudcompare 2.8. <http://www.cloudcompare.org>. Accessed 01 Feb 2018
- Hirschmüller H (2008) Stereo processing by semiglobal matching and mutual information. *IEEE Trans Pattern Anal Mach Intell* 30(2):328–341. <https://doi.org/10.1109/TPAMI.2007.1166>
- Hirschmüller H, Scharstein D (2009) Evaluation of stereo matching costs on images with radiometric differences. *IEEE Trans Pattern Anal Mach Intell* 31(9):1582–1599. <https://doi.org/10.1109/TPAMI.2008.221>
- Irschara A, Rumpler M, Meixner P, Pock T, Bischof H (2012) Efficient and globally optimal multi view dense matching for aerial images. In: *ISPRS annals of photogrammetry, remote sensing and spatial information sciences, vol 1-3*, pp 227–232. <https://doi.org/10.5194/isprsannals-I-3-227-2012>. <http://www.isprs-ann-photogramm-remote-sens-spatial-inf-sci.net/I-3/227/2012/>. Accessed 16 Oct 2018
- Ivanenko V, Shen H, Coughlan J (2009) Elevation-based MRF stereo implemented in real-time on a GPU. In: *2009 workshop on applications of computer vision (WACV), IEEE*, pp 1–8.
- Kim J, Kolmogorov V, Zabih R (2003) Visual correspondence using energy minimization and mutual information. In: *Proceedings ninth IEEE international conference on computer vision, IEEE*, vol 2, pp 1033–1040. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1238463. Accessed 18 Aug 2016
- Li L, Zhang S, Yu X, Zhang L (2018) PMSC: PatchMatch-based superpixel cut for accurate stereo matching. *IEEE Trans Circuits Syst Video Technol* 28(3):679–692
- Olsson C, Ulén J, Boykov Y (2013) In defense of 3D-label stereo. In: *2013 IEEE conference on computer vision and pattern recognition*, pp 1730–1737. <https://doi.org/10.1109/CVPR.2013.226>
- Roth L, Mayer H (2019) Reduction of the fronto-parallel bias for wide-baseline semi-global matching. In: *ISPRS annals of photogrammetry, remote sensing and spatial information sciences, vol IV-2/W5*, pp 69–76. <https://doi.org/10.5194/isprs-annals-IV-2-W5-69-2019>. <https://www.isprs-ann-photogramm-remote-sens-spatial-inf-sci.net/IV-2-W5/69/2019/>. Accessed 23 Dec 2019

- Scharstein D, Tanaii T, Sinha SN (2017) Semi-global stereo matching with surface orientation priors. In: 2017 international conference on 3D vision (3DV), IEEE, Qingdao, pp 215–224. <https://doi.org/10.1109/3DV.2017.00033>. <https://ieeexplore.ieee.org/document/8374574/>. Accessed 16 Oct 2018
- Sinha SN, Scharstein D, Szeliski R (2014) Efficient high-resolution stereo matching using local plane sweeps. In: 2014 IEEE conference on computer vision and pattern recognition, pp 1582–1589. <https://doi.org/10.1109/CVPR.2014.205>
- Szeliski R (2011) Computer vision. Texts in computer science. Springer, London. <http://link.springer.com/10.1007/978-1-84882-935-0>
- Szeliski R, Zabih R, Scharstein D, Veksler O, Kolmogorov V, Agarwala A, Tappen M, Rother C (2008) A comparative study of energy minimization methods for Markov random fields with smoothness-based priors. *IEEE Trans Pattern Anal Mach Intell* 30(6):1068–1080. <https://doi.org/10.1109/TPAMI.2007.70844>
- Wang KC, Gong W (2002) Automated pavement distress survey: a review and a new direction. In: Pavement evaluation conference, pp 21–25
- Woodford O, Torr P, Reid I, Fitzgibbon A (2009) Global stereo reconstruction under second-order smoothness priors. *IEEE Trans Pattern Anal Mach Intell* 31(12):2115–2128. <https://doi.org/10.1109/TPAMI.2009.131>
- Yang R, Welch G, Bishop G (2002) Real-time consensus-based scene reconstruction using commodity graphics hardware. In: 10th Pacific conference on computer graphics and applications, 2002. Proceedings., vol 22, pp 225–234
- Zabih R, Woodfill J (1994) Non-parametric local transforms for computing visual correspondence. European conference on computer vision. Springer, Berlin, pp 151–158
- Zbontar J, LeCun Y (2016) Stereo matching by training a convolutional neural network to compare image patches. *J Mach Learn Res* 17(1–32):2
- Zhang Z, Ai X, Chan CK, Dahnoun N (2014) An efficient algorithm for pothole detection using stereo vision. In: 2014 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, Florence, Italy, pp 564–568. <https://doi.org/10.1109/ICASSP.2014.6853659>. <http://ieeexplore.ieee.org/document/6853659/>. Accessed 16 Oct 2018
- Zhao W, Yan L, Zhang Y (2018) Geometric-constrained multi-view image matching method based on semi-global optimization. *Geo-spat Inf Sci* 21(2):115–126. <https://doi.org/10.1080/10095020.2018.1441754>