Visual Odometry based on Transitivity Error Analysis in Disparity Space - A Third-eye Approach

Hsiang-Jen Chien, Haokun Geng, and Reinhard Klette
The .enpeda.. Project*, Department of Computer Science
The University of Auckland, New Zealand
hchi509@aucklanduni.ac.nz

ABSTRACT

Accurate estimation of ego-motion heavily relies on correct point correspondences in the context of visual odometry. In order to ensure a metric reconstruction of camera motion, we can refer to the 3D structure of the scene. In this paper we present an indicator for evaluating the accuracy of stereo-based 3D point measurements as well as for filtering out low-confidence correspondences for ego-motion estimation. In a typical binocular system, the left and right images are matched to produce a disparity map. For a trinocular system, however, the map can be derived indirectly via disparity maps of both cameras with respect to the third camera. The difference between an explicitly matched disparity map and its indirect construction defines a *transitivity error in disparity space* (TED).

We evaluate the effectiveness of TED from different perspectives, using a trinocular vehicle-mounted vision system. Results presented in 3D Euclidean space, or in 2D images show improvements of more than 7.5%, indicating that, by taking TED into account, more consistency is ensured for ego-motion estimation.

Categories and Subject Descriptors

I.4.8 [**Computer Vision**]: Scene Analysis—*motion, range data, stereo analysis, feature tracking*

1. INTRODUCTION

Visual odometry (VO) is an active research topic in the field of computer vision and robotics. Rovers on Mars demonstrate its functionality. The development of VO is closely related to *structure-from-motion* (SfM) and *simultaneous localisation and mapping* (SLAM) techniques. VO, SfM, and SLAM support applications such as autonomous vehicles, driver-assistance system, unmanned navigation, or robotics.

*Environment Perception and Driver Assistance

VO relies on the fact that the motion of a camera can be recovered from a set of corresponding points successfully identified in images taken at different camera poses. Depending on the configuration of a vision system, the type of available correspondences is either 3D-to-3D, 3D-to-2D, or 2D-to-2D [1]. In the purely 2D case, there is no need to know about the 3D structure of the scene at all, making it a preferable choice among monocular systems such as mobile phones or hand-held devices. In general, the Euclidean metric cannot be recovered in the projective space defined by one pinhole-type camera [2].[1] Thus, the estimation of camera motion will benefit in general from a reconstruction of some 3D scene structure.

Accuracy of 3D measurement has a considerable influence on the performance of ego-motion estimation. The robustness of a VO framework is decided by its outlier rejection scheme and how well it works when a certain percentage of noise is present. The random sampling consensus (RANSAC) approach has been widely deployed in the context of VO to remove wrongly associated 2D matches [1, 4]. Despite its effectiveness, some inliers could still possess significant 3D measurement noise, which will in turn deteriorate the recovered motion.

A binocular vision system measures scene depth using disparity values. This is typically done by applying a stereo matching algorithm (e.g. [5, 6, 7, 8]) on images taken by a left and right camera, possibly preprocessed in some way [9]. Since noisy 3D points have a considerable impact to a VO system, it is very important to identify unreliable disparity values before they are transformed to 3D space and used in ego-motion estimation.

In the literature a few criteria have been proposed to evaluate a disparity map (e.g. [11, 12]). One approach [13] is to warp the left image using the left-right disparity map (see Fig. 1 for an example). By studying the difference between the warped image and the real one, disparity values are assessed. Another approach is to calculate an inverse disparity map (i.e. right-to-left mapping) and compare the obtained disparity values. In this work we generalise such an idea to a multi-camera system having more than two cameras.

The rest of this paper is organised as follows. In Section 2

[1]Distance information can be derived from monocular information to some degree using calibrated bird's-eye views or other techniques; for example, see [3].

Figure 1: A captured view (*top*) and its disparity-warped reconstruction obtained from another captured image and calculated disparities (*bottom*).

we briefly describe the typical VO framework and motion recovery algorithms used to test the proposed method. In Section 3 we propose a consistency indicator for disparity evaluation. Experimental results are shown in Section 4. Section 5 concludes.

2. VISUAL ODOMETRY

Visual odometry approaches use inter-frame point correspondences to derive camera motion. A generic binocular VO framework is depicted in Fig. 2. The detection and tracking of key points are the building blocks of a VO system. The 3D coordinates of a tracked key point are acquired using stereo matching and triangulation techniques. The locations of traced key points and their 3D coordinates are then taken as input to an ego-motion estimator, possibly followed by bundle adjustment (e.g. [14]) which is optionally carried out to optimise the solved motion. Depending on the type of point correspondences, a variety of ego-motion estimators have been developed. In this section two popular motion recovery algorithms are reviewed.

2.1 3D-to-3D: Horn's Analytical Solution

Given a set of point correspondences $x_i^j \to x_i^{j+1}$ where x_i^j and x_i^{j+1} in $\mathbb{R}^2$ denote the 2D image coordinates of a tracked feature $\mathcal{F}_i$ in frame j and $j+1$, respectively, and a 3D measurement function $g : \mathbb{R}^2 \to \mathbb{R}^3$. The motion of the camera can be estimated by solving for the rigid transformation (R, t) which minimises the functional

$$\sum_{1 \le i \le N} \| R \cdot g(x_i^j) + t - x_i^{j+1} \|^2 \tag{1}$$

The analytical solution of Equ. (1) is given by B. K. Horn in [15].

2.2 3D-to-2D: A Perspective-n-Point Solver

The error term measured in Euclidean space as formulated in (1) tends to result in highly unstable estimation. The impact is significant especially in the case when g relies on disparity values, as a slight error in disparity space could lead to a great difference in Euclidean space. A more robust solution is to estimate the motion (R, t) in projective space.

Given a projection function $\Pi : \mathbb{R}^3 \to \mathbb{P}^2$, with

$$\Pi(x) \mapsto \begin{pmatrix} f & 0 & c_x \\ 0 & f & c_y \\ 0 & 0 & 1 \end{pmatrix} x \tag{2}$$

parametrised over f, the focal length, and (c_x, c_y), the principal point. Mapping $\mapsto$ denotes equality up to a scale. The error term can be modeled as follows as the sum of squares of reprojection errors:

$$\sum_{1 \le i \le N} \| \Pi(R \cdot g(x_i^j) + t) - x_i^{j+1} \|^2 \tag{3}$$

Finding pose parameters (R, t) that minimise (3) has been well studied in the literature on solving the *perspective-n-point* (PnP) problem. In this work we adopt an efficient PnP solver proposed in [17]. The algorithm first selects four arbitrary reference points in 3D space to transform the Cartesian coordinates $g(x)$ to barycentric coordinates. Given that applying a rigid transformation to $g(x)$ does not change its barycentric representation, a linear system is constructed

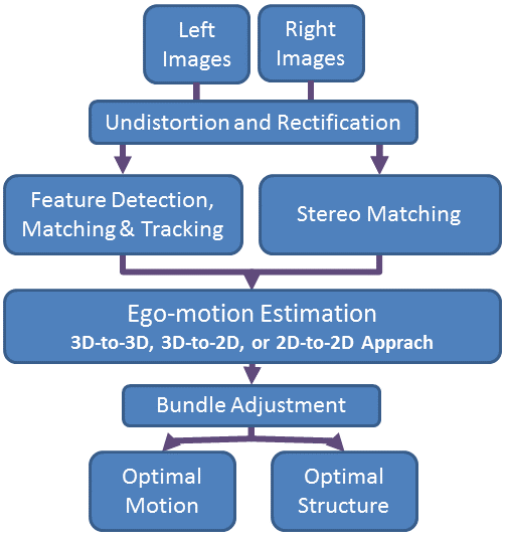

Figure 2: Pipeline of a binocular VO system.

under such invariances. By solving an over-determined system, the optimal solution (R, t) is obtained. The algorithm scales well when the number of point correspondences grows.

3. TRANSITIVE DISPARITY ERRORS

In this section we define the *transitivity error in disparity space* (TED) of a generalised multi-camera system, and then study its property in practice using a trinocular setup. (Following the proposed idea, the approach can be generalised to four or more cameras.)

3.1 TED

Consider a disparity-defined coordinate transformation $\delta : \mathbb{R}^2 \to \mathbb{R}^2$ and a function $M : \mathbb{R}^2 \to \mathbb{R}^n$. We define the warping of M via δ as

$$\phi(M, \delta)(x) = M(x + \delta(x))(x) \tag{4}$$

Function ϕ generates the warping of an image (see Fig. 1 for an example) when $M : \mathbb{R}^2 \to \mathbb{R}$. Furthermore, it can be applied to construct the concatenation of two transformations

$$\tau(\delta_{01}, \delta_{12}) = \delta_{01} + \phi(\delta_{12}, \delta_{01}) \tag{5}$$

where δ_{01} and δ_{12} are coordinate transformations defined by disparity maps of image pairs $(0, 1)$ and $(1, 2)$, respectively.

Due to the transitivity in the disparity space, one may use τ to synthesise the disparity map δ_{ij} of an image pair (i, j) where $i \neq j$ via intermediate maps $\delta_{ik_1}, \delta_{k_1k_2}, ..., \delta_{k_mj}$, without explicit stereo matching. In particular, given a construction sequence $S = (i, k_1, k_2, ..., k_m, j)$ where $s \in S$ denotes a referenced image, the synthetic disparity map $\delta'_{ij} = \Delta(1, m + 2))$ is recursively constructed where

$$\Delta(p, q) = \begin{cases} \tau[\Delta(p, q-1), \Delta(q-1, q)], & \text{if } p < q - 1 \\ \delta_{S_pS_q}, & \text{if } p = q - 1 \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

The construction of the derived disparity map δ'_{ij} starts from concatenating δ_{ik_1} to $\delta_{k_1k_2}$, as formulated by Equ (6). The concatenated map is in turn combined with $\delta_{k_2k_3}$. The integration continues until the last disparity map δ_{k_mj} is incorporated into the concatenation.

The constructed map δ'_{ij} can be further compared with δ_{ij}, the explicitly established one. Their differences should, to some degree, indicate the confidence of the estimated disparity values. We denote the absolute difference $|\delta_{ij} - \delta'_{ij}|$ as the *transitivity error in disparity space* (TED).

In the binocular case, TED is defined reflectively. Consider $S = (0, 1, 0)$, it yields

$$\delta'_{00}(x) = \Delta(1, 3)(x) = \delta_{01}(x) + \delta_{10}(x + \delta_{01}(x))(x) \tag{7}$$

whereas the normalised indicator

$$\epsilon_{00}(x) = \frac{1}{|\delta'_{00}(x) - \delta_{00}| + 1} = \frac{1}{|\delta'_{00}(x)| + 1} \tag{8}$$

is known as *left-right consistency* [8]. Figure 3 visualises the described consistency indicator applied to a trinocular system, with the third camera as being the agent for the synthesis of a left-right disparity map.

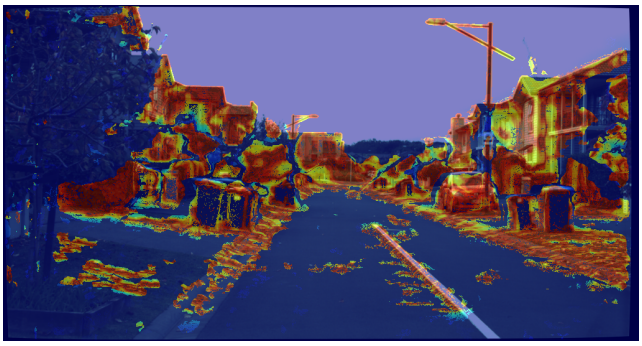

Figure 3: TED-based disparity consistency indicator fused with a captured image. Red pixels indicate a high consistency value up to 1.0, while blue pixels show low consistency close to 0.

3.2 Evaluation of Effectiveness

The effectiveness of TED determines its value in identifying unreliable stereo matches. In this work we evaluate TED by using it as a criterion in correspondence selection for egomotion estimation.

Assuming that image key points are perfectly tracked, the remaining dominating factor for motion estimation accuracy is the 3D structure. If the removal of measurements with higher TED leads to improved motion recovery, it is reasonable that the shifted 3D data have better quality, and so do their disparity values. This way, the effectiveness of TED is assessed.

Cautions have to be taken in the context of VO when adopting TED as a point selection criterion. Motion estimation could be spatially biased if the used 2D features are detected locally. Also, if strong correlation exists between TED and disparity values, this leads to the removal of either near or far points, which also introduces a spatial bias in 3D space.

To study these issues, we ran several tests on 1.8 millions of tracked key points of a real dataset. The dataset was recorded using a vehicle-mounted trinocular vision system. Observations are provided in the following subsections.

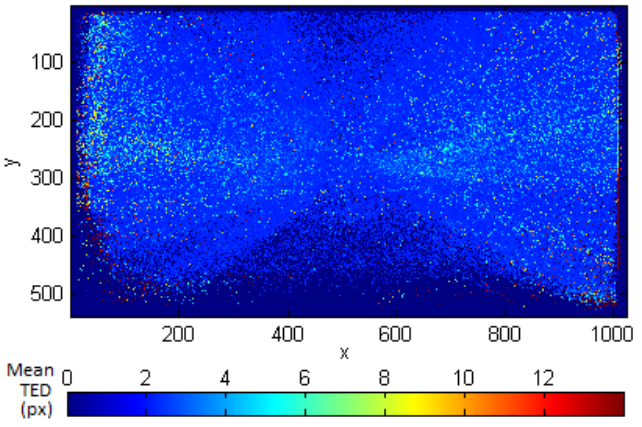

Figure 4: Mean TED values of tracked features and their distribution in the image.

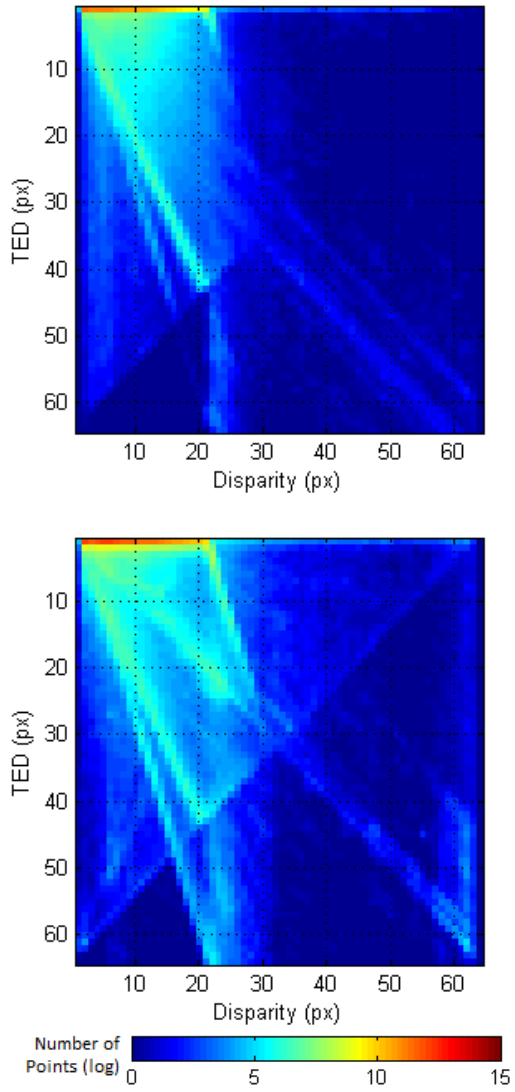

Figure 5: Point density with respect to TED and disparity values obtained by using a block matcher (*top*) and a semi-global matcher (*bottom*). The entries are colour-coded in a logarithmic scale.

3.3 Spatial Distribution of TED

To show the spatial correlation between key points and their TED, we accumulate and average TED values found on each pixel. The result is shown in Fig. 4.

It appears that the evenly distributed TED presents no strong locality in the image. Note that some pixels close to the left and right borders show very high TED ($\geq 2\sigma$). These pixels, however, present only roughly 1% of all the tested key points.

3.4 Distribution in Disparity Space

It is very important to understand the distribution of TED in disparity space. A significant correlation between TED and disparity values leads to truncated 3D structures after applying TED as a filtering criterion. Since noise at far points in disparity space is magnified nonlinearly in Euclidean space [10], the truncation will directly affect the motion-estimation error. In this case, it is difficult to tell whether the improvement (or the deterioration) is caused by the removal of bad measurements according to TED, or whether it is a consequence of range-data truncation.

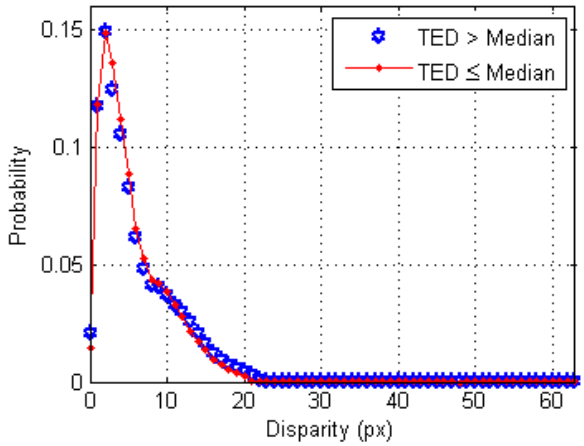

Figure 6: Probability density functions of TED in disparity space. Key points with TED greater than the median show a similar distribution as the remaining.

The density of key points in disparity space with respect to TED is shown in Fig. 5. Tests has been conducted for different matching algorithms, namely a block matcher (BM) and a semi-global matcher (SGM [5]), among which similar results are observed. The majority of key points have low TED ($\leq$ 1.5 pixels) and lie in the first 30% of the disparity space that ranges from 0 to 64.

To further study how the TED-based filtering impacts the distribution of key points in disparity space, we cut on the median TED to divide the tracked key points into two equally sized groups. The distributions, depicted in Fig. 6, are nearly identical, indicating that using the median TED as the threshold does not cause a depth-truncation effect.

4. EXPERIMENTS

A series of experiments had been designed to evaluate the effectiveness of applying TED as a disparity quality indicator to improve ego-motion estimation. The tested video sequence was recorded by HAKA1, an experimental vehicle with trinocular setup of real-time high-resolution image sensors. The sensors synchronised capture chromatic image data of 2046×1080 pixels at 30 fps. *Speeded-Up Robust Features* (SURF [18]) key points are detected and tracked through 1,926 collected frames. We implemented an SGM algorithm to calculate dense 3D structures from disparity maps with respect to the first and the second camera. Disparity maps from these two cameras to the third are also generated to build TED. The median of TED decides the threshold that separates key points into two sets, and the

camera motion is estimated respectively using each point set. At the evaluation stage, all of the key points are included. The effectiveness of TED is evaluated in both projective and Euclidean space. Neither global nor local bundle adjustment has been deployed.

4.1 Evaluation in Projective Space

The 3D coordinates of each tracked key point are projected onto the image using estimated motion, and the differences between the reprojected and the detected locations are calculated. Both 3D-to-3D and 3D-to-2D ego-motion estimation are carried out. The reprojection error (RPE) plots are shown in Fig. 7.

The mean RPE drops from 0.4 to 0.37 pixels with an improvement of 7.5%, in the 3D-to-2D case where the motion is solved using the EPnP algorithm. The suppression of projective errors is not obvious because it is explicitly modeled in the PnP problem. On the other hand, in the 3D-to-3D case the errors significantly reduce from 8.55 to 5.45 pixels. An improvement of 36.3% is achieved.

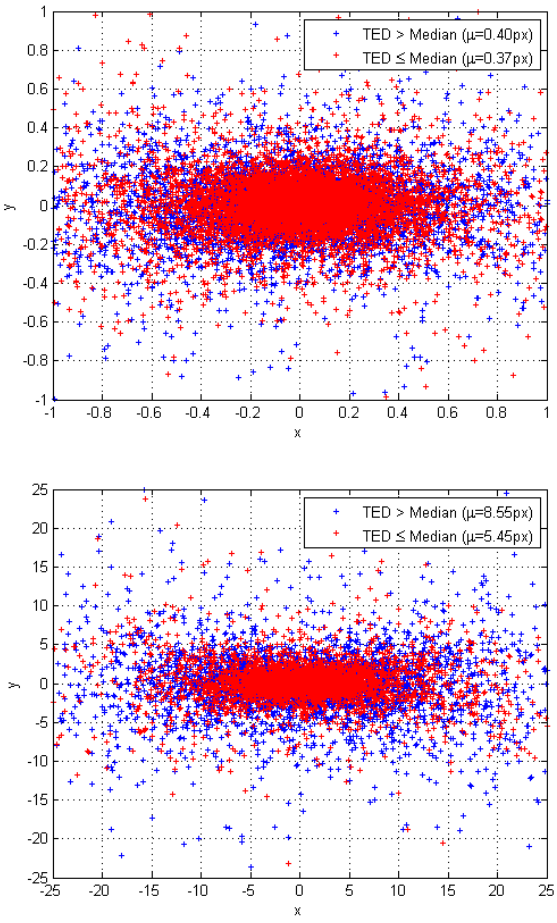

Figure 7: Reprojection error plots of keypoints transformed by motions estimated using 3D-to-2D (*top*) and 3D-to-3D (*bottom*) correspondences, respectively.

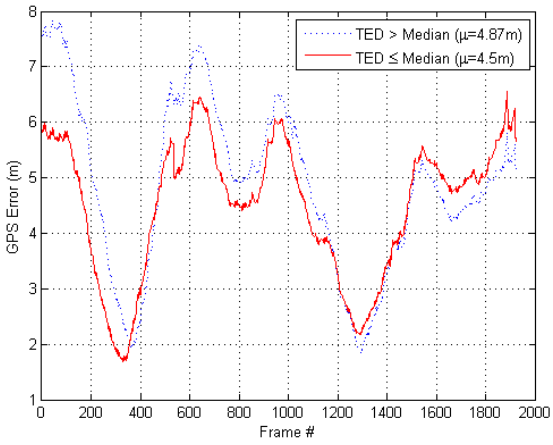

Figure 8: Deviations between GPS trajectory and estimated ego-motion.

4.2 Evaluation using GPS Data

The recorded geodetic coordinates (latitude-longitude-altitude, LLA) are transformed to Earth-centered Earth-fixed (ECEF) coordinates using the WGS84 model. Since inertial data are not recorded by the GPS unit, the rotation part of complete motion data is not available. We instead perform a trajectory registration technique to analyse the difference between GPS-derived motion and the ego-motion estimated from 3D-to-2D correspondences. The optimal registration is calculated using the closed-form solution given by [15].

The errors between registered trajectories are plotted in Fig. 8. Along the GPS trace of 370 metres, it shows that the motion estimated using points with higher TED has slightly larger deviations. The error drops by 7.6%, from 4.87 m to 4.5 m, when points with lower TED are incorporated into the estimation of ego-motion.

We also perform 3D reconstruction of the tested sequence using ego-motion estimates; see [16] for earlier work on 3D roadside reconstruction. By visual inspection, it has been found that misalignments of point clouds are significantly reduced when TED is taken into account. Figure 9, for example, shows multiple scans of road surfaces aligned consistently using the improved ego-motion estimate.

5. CONCLUSIONS AND FUTURE WORK

In this paper we proposed a quality indicator of 3D measurements based on analysing transitive errors in disparity space. It evaluates the consistency in disparity space by integrating multiple disparity maps from different camera pairs. The experiments show that, in a trinocular setup, the ego-motion estimate achieves improvements of more than 7.5% in both projective and Euclidean space, when TED is used as the point selection criterion through the pipeline of visual odometry. The results have also been visually assessed, showing that misaligned point clouds are amended.

The experimental results show that there exists a linkage between TED and disparity errors. For explaining such link-

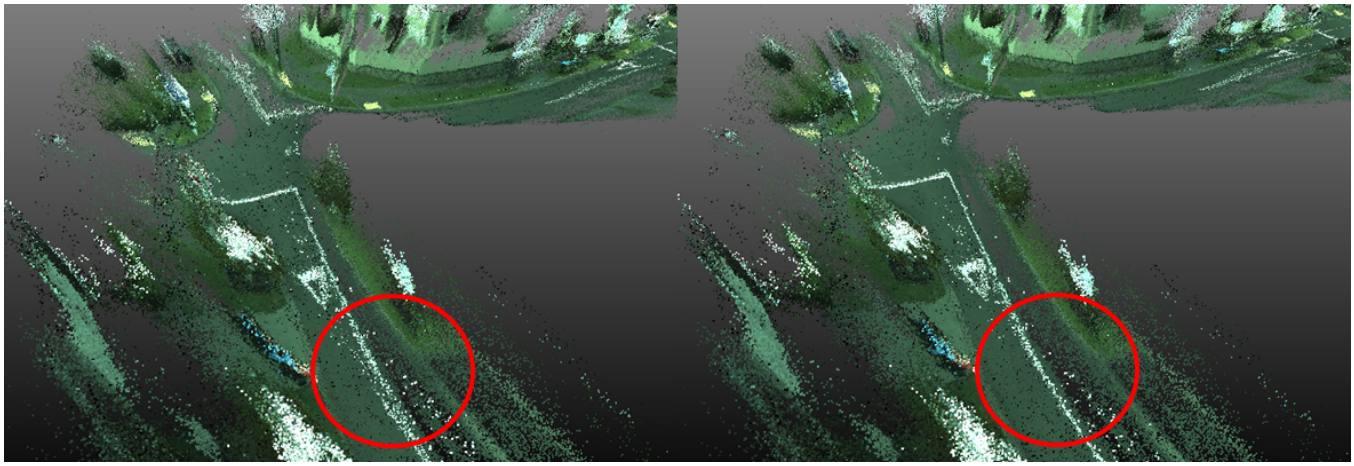

Figure 9: Reduced misalignment of the centre line as marked by the red circles is observed when disparities of lower TED are selected (*right*), compared with the result obtained using high TED disparities (*left*).

ages further, scene simulation could be helpful in providing ground truth analysis. Also, the proposed transitivity model is applicable to the combination of disparity maps and optical flows. The presented framework can be extended to include spatial and temporal consistency as a future work.

6. REFERENCES

[1] D. Scaramuzza and F. Fraundorfer. Visual odometry. *IEEE Robotics Automation Magazine*, **18**:80–92, 2011.

[2] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*, second edition. Cambridge University Press, Cambridge, 2004.

[3] M. Rezaei and R. Klette. Look at the driver, look at the road: No distraction! No accident! In Proc. *CVPR*, pages 129–136, 2014.

[4] B. Kitt, A. Geiger, and H. Lategahn. Visual odometry based on stereo image sequences with Ransac-based outlier rejection scheme. In Proc. *IEEE Symp. Intelligent Vehicles*, pages 486–492, 2010.

[5] H. Hirschmüller. Accurate and efficient stereo processing by semi-global matching and mutual information. In Proc. *CVPR*, volume 2, pages 807–814, 2005.

[6] S. Hermann and R. Klette. Iterative semi-global matching for robust driver assistance systems. In Proc. *ACCV*, LNCS 7726, pages 465–478, 2013.

[7] J. Sun, N.-N. Zheng, and H.-Y. Shum. Stereo matching using belief propagation. *IEEE Trans. Pattern Analysis Machine Intelligence*, **25**:787–800, 2003.

[8] R. Klette. *Concise Computer Vision*. Springer, London, 2014.

[9] S. Guan, and R. Klette. Belief-propagation on edge images for stereo analysis of image sequences. In Proc. *Robot Vision*, pages 291–302, 2008.

[10] W. Khan, J. Morris, and R. Klette. Stereo accuracy for collision avoidance. In Proc. *IVCNZ*, IEEE Catalogue Number CFP0967E, 2009.

[11] I. Cabezas, V. Padilla, and M. Trujillo. A measure for accuracy disparity maps evaluation. *Progress Pattern Recognition Image Analysis Computer Vision Applications*, LNCS 7042, pages 223–231, 2011.

[12] I. Cabezas, V. Padilla, and M. Trujillo. BMPRE: An error measure for evaluating disparity maps. In Proc. *IEEE Int. Conf. Signal Processing*, volume 2, pages 1051–1055, 2012.

[13] S. Morales, and R. Klette. A third eye for performance evaluation in stereo sequence analysis. In Proc. *Computer Analysis Images Patterns*, pages 1078–1086, 2009.

[14] M. I. A. Lourakis and A. A. Argyros, SBA: A software package for generic sparse bundle adjustment. *ACM Trans. Mathematical Software*, **36**:1–30, 2009.

[15] B. K. P. Horn. Closed-form solution of absolute orientation using unit quaternions. *J. Optical Society America A*, **4**:629–642, 1987.

[16] Y. Zeng and R. Klette. Multi-run 3D streetside reconstruction from a vehicle. In Proc. *CAIP*, LNCS 8047, pages 580–588, 2013.

[17] V. Lepetit, F. Moreno-Noguer, and P. Fua. EPnP: An accurate O(n) solution to the PnP problem. *Int. J. Computer Vision*, **81**: 155–166, 2009.

[18] B. Herbert, A. Ess, T. Tuytelaars, and L. Van Gool. SURF: "Speeded Up Robust Features" *Computer Vision Image Understanding*, **110**: 346–359, 2008.