Evaluation of Close-range Stereo Matching Algorithms Using Stereoscopic Measurements

Description

This paper describes a new assessment protocol for close range stereo matching algorithms using a stereo display and presents evaluation results of three stereo processing pipelines used in current and future Mars rover operations.

Abstract

The performance of binocular stereo reconstruction is highly dependent on the quality of the stereo matching result. In order to evaluate the performance of different stereo matchers, several quality metrics have been developed based on quantifying error statistics with respect to a set of independent measurements usually referred to as ground truth data. However, such data are frequently not available, particularly in practical applications or planetary data processing. To address this, we propose a ground truth independent evaluation protocol based on manual measurements. A stereo visualization tool has been specifically developed to evaluate the quality of the computed correspondences. We compare the quality of disparity maps calculated from three stereo matching algorithms, developed based on a variation of GOTCHA, which has been used in planetary robotic rover image reconstruction at UCL-MSSL (Otto and Chau, 1989). From our evaluation tests with the images pairs from Mars Exploration Rover (MER) Pancam and the field data collected in PRoViScout 2012, it has been found that all three processing pipelines used in our test (NASA-JPL, JR, UCL-MSSL) trade off matching accuracy and completeness differently. NASA-JPL's stereo pipeline produces the most accurate but less complete disparity map, whilst JR's pipeline performs best in terms of the reconstruction completeness.

Keywords

Stereo matching, Stereoscopic visualization, Rover image processing, 3D reconstruction, Stereo matching evaluation

1. Introduction

Stereo matching has long been a fundamental and challenging research topic in computer vision. A large number of fully automated stereo matching algorithms have been developed since the earliest approach made by Hannah (Hannah, 1974) and further variations of local algorithms, which rely on the computation of correlations of local patches, developed in the 1990s. Follow-on optimisation and statistical machine learning techniques including dynamic programming (Birchfield and Tomasi, 1998), Markov random field (Geman, 1984), graph cuts (Boykov, 2001), belief propagation (Sun et al., 2003), semi-global matching (Hirschmuller, 2008), and seed-growing algorithms (Lhuillier and Quan, 2002), have been shown to be able to produce high quality disparity maps, but it is getting difficult to evaluate various matching algorithms developed for different purposes.

To our best knowledge, the Middlebury test is the most influential work on recent stereo evaluation (Scharstein and Szeliski, 2002). In this test, the authors propose a new taxonomy of comprehensive stereo algorithms and a C++ test bed for the quantitative evaluation of dense two-frame stereo correspondence algorithms. The Middlebury test basically performs an evaluation based on the error metrics computed from sparse "ground truth" point pairs or by synthesizing a warped image from pre-computed dense disparity maps. Therefore, the reference data plays an important role in the evaluation process.

When the algorithms were not strong enough to process complicated scenes, the 3D geometry of reference data does not need to be complex, but it needs to be dense enough to evaluate a sparse set of point correspondences produced by test algorithms. For this reason, Scharstein et al. configured a test scene with a set of slanted 2D planes. Since a 2D homography of a planar object can be easily defined by 4 point correspondences, this approach can produce a virtually complete disparity map of two images from a few manual correspondences (Scharstein et al., 2001). However, as stereo algorithms evolve, a simple geometry is no longer able to differentiate advanced algorithms and people need more complex geometry at higher pixel resolution.

Synthetic images can be an option to improve the scene complexity (Morales and Klette, 2011) but they are generally insufficient to synthesize practical scenes affected by a range of noise and various lighting conditions. Alternatively, an active 3D sensor can be used to produce reference data. For example, a special structured light system was developed in the 2003 Middlebury test, where one or two projectors are used with a translating camera to create a dense reference disparity map for a stereo pair (Scharstein and Szeliski, 2003). This approach is particularly useful as we can have control over the spatial resolution of a disparity map with higher depth accuracy. However, a structured light is more suitable for capturing small objects in a controlled indoor environment. Geiger et. al. also pointed out this limitation, mentioning that higher ranking algorithms from the Middlebury reference data can go below average when it is tested against the images from outside the laboratory (Geiger et al., 2012).

Creating reference data for multiview stereo algorithms could be even more challenging. In addition to classic stereo matching, estimating external transforms between image pairs and locating the position of a camera in a previously reconstructed scene are other imperative features of a multiview stereo algorithm (e.g. visual odometry or SLAM). Therefore, the reference data should be registered with correct positional information. This normally requires combining multiple heterogeneous sensors and more complicated calibration steps.

For example, the Middlebury test images for multiview stereo algorithms were obtained using a robotic arm that can move on the surface of one-metre radius sphere with high precision (Seitz, 2006). In addition, to improve the accuracy of a 3D model, the initial point cloud from multiple images was registered with a more refined laser scanning result using an ICP method. Jensen et. al. recently published a data set containing 80 scenes for large scale multiview stereo evaluation using a similar approach but with a structured light (Jensen et al., 2014). For outdoor scenes, Strecha et. al. proposed a method that can combine multiple Lidar scans with images based on physical markers placed on a test scene (Strecha et al., 2008). Later, Geiger et al. proposed more automated method which combines Lidar and two video cameras with accurate localisation systems (e.g., GPS and IMU) to cover a wider area from a long-distance drive (39.2 km) (Geiger et al., 2012).

It is possible to produce a good quality of reference data for outdoor scene by registering active sensors to stereo cameras as mentioned above, and in fact it is widely used in the orbital sensor calibration process in many remote sensing applications. For example, the performance of the SIMBIO-SYS imagining suite employed in ESA BepoColombo mission was assessed during a pre-flight calibration process, where laser scans of a small target object are used to validate a stereo reconstruction result of the sensor (Simioni et al., 2014). Also, the high-resolution stereo camera (HRSC) on Mars Express was validated based on various outdoor scenes captured during on-ground and airborne test (Jaumann et al., 2007). However, this approach is not always available, especially, when performing planetary 3D reconstruction using robotic vision systems. Also, creating reference data using multiple sensors would be a very expensive process in terms of computation complexity and labour, even though a new set of test data is frequently required to evaluate advanced algorithms. To address this, we introduce a new accuracy evaluation method to assess stereo matching results when there is no prior knowledge about the depth of points within a scene. This "ground truth" independent evaluation criteria were inspired by the use of manual measurements in stereo photogrammetry, originally performed using film media and optic mechanical instrumentation but since the early 2000s using so-called softcopy stereo workstations based on stereoscopic displays. An early example of the use of these manual photogrammetric measurements using an analytical stereoplotter is discussed by Day and Muller, 1989. A recent paper also showed that the use of 3D stereoscopic display can improve human performance in locating objects and inferring depths of surfaces within a scene (Mcintire, 2014), so that this approach is not only more effective than the manual point selection used by the computer vision community in early days (Nakamura et al., 1996), but also closely related to the local cross-correlation process inspired by a biological vision system (Fleet et al. 1996).

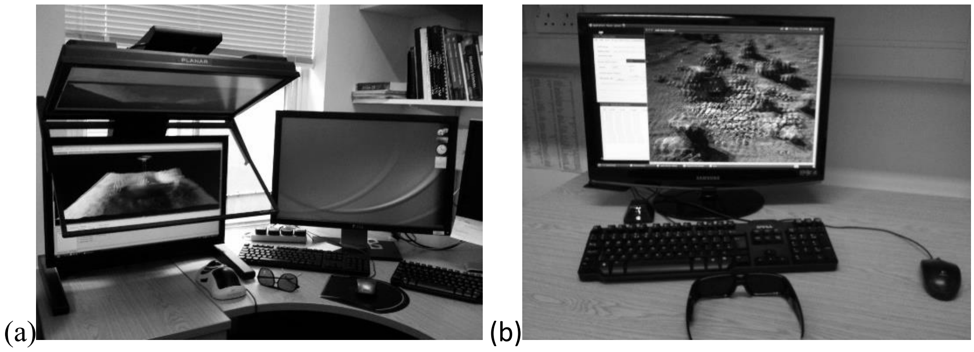


Figure 1: Example of stereoscopic visualisation with a passive stereo display where images from upper and lower displays are reflected on a polarised beamspliter in the middle (a), whereas an active stereo display uses a high refreshing LCD screen (120 HZ) with synchronised NVIDA shutter glasses (b).

In this work, a Java-based stereo workstation has been developed based on work performed at JPL on being able to display stereo data on different stereo displays (Pariser and Deen, 2009). We trained a group of research participants to make repeat measurements of the three-dimensional position of fixed points in the same scenes using a stereo cursor on a stereo workstation display (Azari et al., 2009; Shin et al. 2011). A stereo display is afforded either using anaglyptic fusion of stereo-pairs on a colour display or by using different specialist stereo display devices [Fig. 1(a) and (b)] of increasing sophistication and cost. These tie-points are then used to compute error metrics of different stereo matching algorithms by comparing the computed disparity map with the corresponding manual measurements under three different manual selection scenarios. A 2D Gaussian function based scoring metrics have also been introduced for a quantitative evaluation.

The proposed evaluation method can be used to complement the Middlebury test when we need new test images from more complex scene at higher image resolution. More importantly, it can complement the missing evaluation work of stereo matching of rover imagery from planetary robotic missions, such as the NASA Mars Exploration Rover (MER) or Mars Science Laboratory

(MSL), where obviously we do not have either any “ground truth” 3D data nor any prior knowledge of the scene.

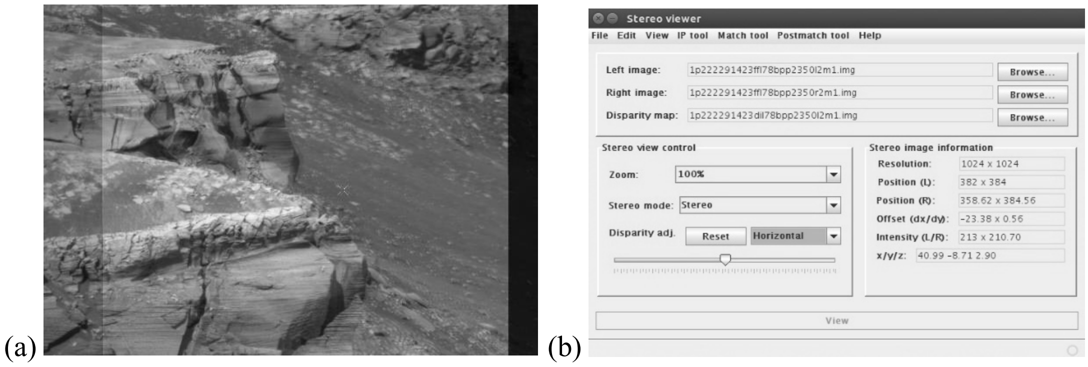
This evaluation method was proposed within the EU FP-7 Planetary Robotics Vision Ground Processing (PRoVisG: EU FP-7 PRoVisG project, http://provisg.eu/), and has been applied to evaluate the accuracy of disparity maps computed from stereo pairs in the PRoVisG Mars 3D challenge campaign (http://cmp.felk.cvut.cz/mars/) as well as additional stereo-pairs captured in the ExoMars Pancam test campaign at Clarach Bay in Aberystwyth (ExoMars test campaign: https://www.youtube.com/watch?v=6gRo8QSXX5c), using state-of-art planetary stereo technologies from NASA-JPL (USA), Joanneum Research Institute (Austria) and UCL-MSSL (UK).

We explain more details of the proposed evaluation protocol in the following section. Based on which, we present the evaluation results of a couple of disparity maps produced by JPL, JR, and UCL in Sec. 3, followed by our discussion in Sec. 4.

2. Method

2.1 Stereo Workstation

Most stereo matching algorithms used in the remote sensing community employ an automated workflow that has been built based on different mathematical definitions of image features (e.g. corners and edges) and/or matching (dis-)similarity of corresponding points on a stereo pair. However, this often neglects the impact of different detection errors from various imaging conditions such as image noise, viewing angle, resolution, and scale difference. In addition, there is normally no proper visual validation of the detected point pairs.



Left image: 1p222291423ffl78bpp2350l2m1.img
Right image: 1p222291423ffl78bpp2350r2m1.img
Disparity map: 1p222291423dil78bpp2350l2m1.img

Stereo view control

Zoom: 100%
Stereo mode: Stereo
Disparity adj. Horizontal

Stereo image information

Resolution: 1024 x 1024
Position (L): 382 x 384
Position (R): 358.62 x 384.56
Offset (dx/dy): -23.38 x 0.56
Intensity (L/R): 213 x 210.70
x/y/z: 40.99 -8.71 2.90

Figure 2: Example of a stereo anaglyph showing a stereo cursor: (a) the offset of a stereo cursor is automatically set according to the supplied disparity map; (b) triangulated 3D position of a corresponding pair is also displayed when there is calibration data based on the use of the CAHVOR calibration model employed by NASA for MER and MSL cameras (Di and Li, 2004).

To address these issues, we developed a Stereo WorkStation (StereoWS) under the PRoVisG project. The proposed system is capable of visualizing tie-points on a stereo pair in a hardware-independent manner, e.g. with a conventional colour display, it will automatically switch the rendering mode to stereo anaglyphs [see Fig. 2(a)].

We also developed intuitive user interfaces to facilitate the tie-point validation and selection process. For example, provided there is no pre-existing disparity map, users can make measurements using a floating 3D cursor, or fix the cursor in the left image at a pre-defined point and only allow the right image cursor to move in 3D (i.e. by changing the disparity of the stereo cursor) in order to be able to place the 3D cursor onto a visually perceived surface. When there is an initial disparity map available, however, the offset of the stereo cursor will be automatically adjusted to speed up the tie-point selection process.

Information on each collected tie-point such as tie-point ID, coordinates, can also be displayed in a separate window [see Fig. 2(b)], so that a user can easily edit the incorrect tie-point as well as monitor progress. To assist a user to select a tie-point more efficiently, a range of basic image processing tools are also included, and our in-house stereo matching algorithm, i.e. Adaptive Least Squares Correlation (ALSC) (Gruen, 1985) and Region growing (GOTCHA) (Otto and Chau, 1989) have been integrated into the software to produce a denser disparity map from the collected manual tie-points, if required.

2.2 Selection of tie-points

In this work, we define three types of tie-points and employ slightly different selection procedures to prepare a sub-pixel reference tie-points:

(a) **Feature based**: Irregularly distributed tie-points.

(b) **Regular grid**: Regularly distributed tie-points.

(c) **Discontinuities**: Tie-points around depth discontinuities.

Type (a) (i.e. feature-based) tie-points are collected to generate highly detectable reference tie-points from standard feature matching algorithms. Since many detectable image features are found around highly textured areas, we can easily select feature-based tie-points from visual identification. The selection procedure initially defines a number of "interesting" points from the left image using generic feature extraction algorithms, and then ask participants to identify the corresponding right point by adjusting the offset of a stereo cursor. Corresponding tie-points in the right image are, therefore, defined at integer resolution initially. However, an average is taken of a set of manual selections that result in sub-pixel selection. Alternatively, ALSC can be applied to the right tie-point to refine the pixel position.

Type (b) (i.e. regular grid) tie-points are proposed to collect regularly distributed reference tie-points across the whole image. This will improve the chance of getting reference tie-points from small depth discontinuity or from less-textured areas. Unlike the feature based selection, it will be a bit more challenging to pick a correct tie-point from visual identification. Therefore, participants are asked to collect tie-point from visual validation, i.e. an initial guess for a right tie-point is given at the beginning. To provide good starting points to participants, a dense disparity map is generated using an in-house stereo processing pipeline and sampled at regular grid points. These initial tie-points are then visually inspected, e.g. moving the stereo cursor around the grid points and check if there is any abnormality, or adjusting the disparity offset of a stereo cursor at the point to check whether the estimation appears to be the best solution, and/or do both with 1.5 or 2 times scaled-up images, which will increase the chance of getting correct correspondences (Chan et al., 2003). Finally, collect the resulting tie-points that pass the validation test.

Type (c) tie-points (i.e. discontinuities) aims to collect reference tie-points from the places that general automated matchers may fail (so-called pathological cases). These areas are normally resulted from occlusions, insufficient texture, and strong depth discontinuities, i.e. pixels whose neighbouring disparities differ by more than a threshold (refer to the Middlebury stereo evaluation (Scharstein et al., 2001)). Amongst these, we are particularly interested in matching performance around depth discontinuity, since some algorithms deliberately enforce the local smoothness around depth discontinuities in order to densify a disparity map. We manually select two pairs of tie-points around this area, i.e. one tie-point from background and another one from foreground and evaluate how correctly an algorithm can handle the scene occlusions (see Fig. 3 and Sec. 2.4). The scene occlusion is a well-known issue in classic stereo matching, therefore it might be interesting to see if it is possible to design an automated pipeline for populating type (c) tie-points (i.e. discontinuities) with conventional feature detectors. However, without knowing true foreground and background segmentation, we found this would be difficult to make it fully automated.

To select type (c) tie-points, an expert manually chooses a set of challenging tie-points around a typical problematic area, and participants are asked to validate them. The validation process is quite similar to the regular grid selection, except that this time no clues are given around tentative tie-points.

2.3. Error metrics

The next step is to estimate the error bounds according to the statistics recorded in the three types of manual tie-point selection process. Suppose that T^k is a set of left tie-points from type (k) dataset, i.e.$T^k = \{\boldsymbol{t}_0^k, \cdots, \boldsymbol{t}_M^k\}$, where $k \in \{a, b, c\}$ and M is the number of left tie-point defined in type (k). Similarly, we can define a set of right tie-points corresponding to $\boldsymbol{t}_i^k$ from manual selections as $S_i^k = \{\boldsymbol{s}_{0i}^k, \cdots, \boldsymbol{s}_{Ni}^k\}$, where N is the number of participants performing manual measurement.

Although it is not always true that some of the measurements in S_i^k happen to be identical to ground truth, it is highly likely that a true correspondence of $\boldsymbol{t}_i^k$ can be found within a cluster of selected points. Thus, our scoring method basically defines a local cluster of S_i^k based on the mean $\boldsymbol{m}_i^k$ and the standard deviation $\boldsymbol{\sigma}_i^k$ and evaluates final matching score.

When estimating the statistics from manual measurements, it should be considered that not everyone is good at fusing a stereo pair and few people are not even capable of perceiving depth difference from the stereo fusion. Therefore, the outliers need to be identified and removed before evaluating statistics of the tie-point positions from a large group of manual selections.

To identify outliers, we define a simple error function using a pre-computed disparity map D. For example, a selection error of a tie-point $(\boldsymbol{t}_i^k,\ \boldsymbol{s}_{mi}^k)$, can be defined as the pixel difference between the manual measurement and computed disparity map for a point, i.e.

$$e\left(\boldsymbol{t}_i^k, \boldsymbol{s}_{mi}^k \colon D\right) = d\left(\boldsymbol{t}_i^k, \boldsymbol{s}_{mi}^k\right) - d(\boldsymbol{t}_i^k, D(\boldsymbol{t}_i^k)), \tag{1}$$

where $d\left(\boldsymbol{t}_i^k, \boldsymbol{s}_i^k\right) = \boldsymbol{s}_i^k - \boldsymbol{t}_i^k$ and $D\left(\boldsymbol{t}_i^k\right) = \tilde{\boldsymbol{s}}_i^k$ is a corresponding point of $\boldsymbol{t}_i^k$ defined by a pre-computed disparity map D.

With this error metric (1), we can define an inlier set $\hat{S}_i^k$ containing all reliable right tie-points,

$$\hat{S}_i^k = \left\{\boldsymbol{s}_{mi}^k \middle| \boldsymbol{s}_{mi}^k \in S_i^k, \boldsymbol{s}_{mi}^k \in C_m^k, \left\| e\left(\boldsymbol{t}_i^k, \boldsymbol{s}_{li}^k \colon D\right) \right\| < \delta\ \forall \boldsymbol{s}_{li}^k \in C_m^k\right\}, \tag{2}$$

where δ is an error threshold which is normally set to around 10 pixels, and C_m^k is a set of right tie-points collected by the m-th participant. Thus, an error bound of $\boldsymbol{t}_i^k$ (denoted by $\boldsymbol{b}_i^k$ in this paper) can be defined as

$$\boldsymbol{b}_i^k = \begin{bmatrix} \boldsymbol{m}_i^k \\ \boldsymbol{\sigma}_i^k \end{bmatrix} = \frac{1}{|\hat{S}_i^k|} \begin{bmatrix} \sum_i \boldsymbol{s}_{mi}^k \\ \sqrt{\sum_i (\boldsymbol{s}_{mi}^k - \boldsymbol{m}_i^k)^2} \end{bmatrix}. \quad (3)$$

As a general quality metric of a set of stereo measurements, we can also define a total measurement error as

$$e_{tot}(T^k, S^k : D) = \frac{1}{MN} \sum_i^M \sum_j^N \left\| d\left(D(\boldsymbol{t}_i^k), \boldsymbol{s}_{ji}^k\right) \right\|, \quad (4)$$

where S^k represents all measurements, i.e. $S^k = \cup_i S_i^k$. Similarly, we can also define a measurement error of an inlier set and an outlier set, i.e. $e_{in}(T^k, \hat{S}^k : D)$ and $e_{out}(T^k, S^k - \hat{S}^k : D)$, respectively.

2.4 Assessment criteria

The proposed evaluation method basically assesses a disparity map in terms of matching score (M) and rewarding score (R). A matching score is similar to the classic quality metric used in stereo evaluation but the main difference is that our method evaluates it based on a set of error bounds rather than ground truth. The proposed method also introduced a rewarding score. The main purpose of this is to award more scores when an algorithm can cope well with challenging matching problem defined in the discontinuous point selection.

In order to compute matching score, we define a 2D Gaussian function from an error bound. For example, a scoring function for $\tilde{\boldsymbol{s}}_i^k$ (i.e. the right pixel position of $\boldsymbol{t}_i^k$ obtained from an input disparity map for evaluation) is

$$g(\tilde{\boldsymbol{s}}_i^k, \boldsymbol{b}_i^k) = exp\left\{-0.5\left(\tilde{\boldsymbol{s}}_i^k - \boldsymbol{m}_i^k\right)^T \begin{bmatrix}\sigma_{xi}^2 & 0\\ 0 & \beta\sigma_{xi}^2\end{bmatrix}^{-1} \left(\tilde{\boldsymbol{s}}_i^k - \boldsymbol{m}_i^k\right)\right\}, \tag{5}$$

where $\boldsymbol{b}_i^k$ is the error bound of $\boldsymbol{t}_i^k$, σ_{xi}^2 is the variance of the *x* values of the *i*-th tie-points in type (k) data set, and $0 < \beta < 1$.

This means that we give a higher matching score when an input disparity is closer to the mean of inlier measurements. If a stereo selection is not confident (i.e. σ_x is high), then we penalise less even if a tie-point is further away from the mean. Another thing to note is that the covariance matrix in (5) is defined by a horizontal standard variance only, i.e. σ_{xi}. This is because σ_{yi} of manual measurements is nil as we rectify an input stereo pair for stereo measurement. However, to allow a little variation in *y* direction as some algorithms do refine vertical positions even if an input stereo pair is rectified, we have used $\sigma_{yi} = 0.2\sigma_{xi}$ in our test. Please note that this weighting value was selected empirically based on our ALSC refinement results of the manual measurements.

A matching score of a set of right points from a disparity map is then defined as a weighted sum of (5), i.e.

$$M(D,B) = \tfrac{1}{L}\textstyle\sum_k \sum_i^{|T^k|} w_i g(\tilde{\boldsymbol{s}}_i^k, \boldsymbol{b}_i^k), \tag{6}$$

where $L = |T^a| + |T^b| + |T^c|$, B^k is a set of all error bounds, *D* is a disparity map for evaluation which defines $\tilde{\boldsymbol{s}}_i^k$, and $w_i = 1 - \frac{\sigma_{xi}}{2\max(\sigma_{x0,}\cdots,\sigma_{xk,})}$, i.e. a higher weight is given to a more confident measurement.

The proposed rewarding score is defined for the tie-points at discontinuities (i.e. type (c)). As briefly explained earlier in Sec. 2.2. we have defined a pair of tie-points around object boundary. Supposing that P_i is the *i*-th pair of the discontinuous tie-points obtained around object boundary, we can define the *i*-th pair $P_i = \{(\boldsymbol{t}_{2i}^c, \tilde{\boldsymbol{s}}_{2i}^c), (\boldsymbol{t}_{2i+1}^c, \tilde{\boldsymbol{s}}_{2i+1}^c)\}$ and an example of a pair can be found

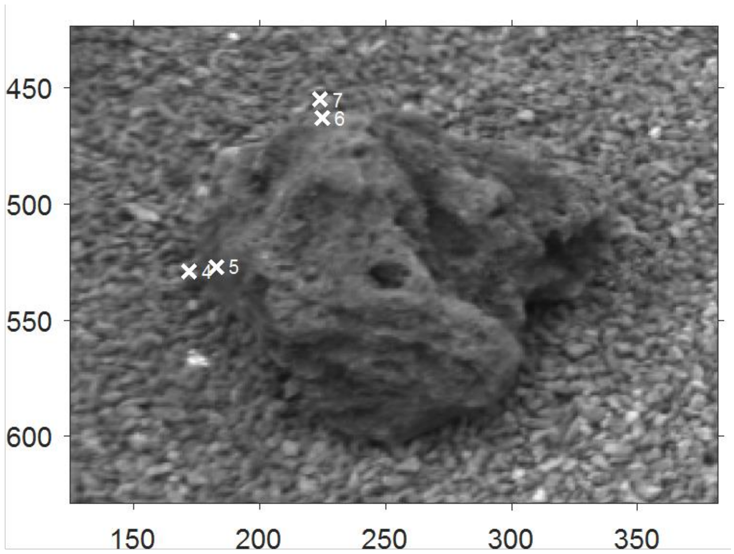


Figure 3: Example of a pair of tie-points around object boundary, e.g. $\boldsymbol{t}_5$ and $\boldsymbol{t}_6$ are a pair of left tie-points collected from background and foreground to evaluate rewarding score.

in Fig. 3. In this case, our rewarding function is defined as an averaged sum of sigmoid function values, i.e.

$$R(D,B,P) = \frac{1}{|P|}\sum_{i=0}^{|P|}\gamma(-|d(\boldsymbol{t}_{2i+1}^{c},\boldsymbol{t}_{2i}^{c}) - d(\tilde{\boldsymbol{s}}_{2i+1}^{c},\tilde{\boldsymbol{s}}_{2i}^{c})|), \tag{7}$$

where $\gamma(x)$ is a sigmoid function, $\frac{2}{1+\exp(-x)}$, and P is a set of all pairs of tie-points, $P = \cup_i P_i$. Thus, (7) gives additional scores when a disparity map can give a similar estimation to the average of manual measurements around a depth discontinuity.

Finally, a total score (TS) is defined as a weighted sum of the matching score and the rewarding score, i.e.

$$TS(D,B,P) = (1-\alpha)M(D,B) + \alpha R(D,B,P), \tag{8}$$

where $0 < \alpha < 1$. The weighting coefficient in (8) can be set up differently depending on applications, e.g. a higher weight (e.g. $0.5 < \alpha$) could be given to put the matching score ahead over the rewarding score of a disparity map.

3. Experiment results

The evaluation work described in this paper is based on the stereo matching results from UCL-MSSL, NASA-JPL, and the Joanneum Research Institute (JR hereafter) with respect to the datasets from the PRoVisG Mars 3D challenge and the ExoMars PanCam test campaigns. The PRoVisG Mars 3D challenge 2011, aimed at testing and improving the state of the art algorithms of visual odometry and 3D terrain reconstruction in planetary exploration.

The task of the PRoVisG Mars 3D challenge was to reconstruct depth, camera trajectory and 3D maps of Mars landscapes observed by MER. The ExoMars PanCam test campaign also focused on the 3D processing results, as they are an essential component of mission planning and scientific data analysis for the ESA's ExoMars Rover mission, planned for launch in 2020.

We demonstrate the evaluation with 3 test sequences, taken from one of the PRoVisG Mars 3D challenge I datasets (sets C33) and the ExoMars PanCam test campaign ("65246" and "70000"). Examples of the images from each of these 3 test sequences are shown in Fig. 4. The evaluation work demonstrated in this paper was achieved through a workshop hosted at UCL-MSSL with 15 participants including 9 students and 6 trainers.

3.1 Test datasets

During this stereo matching evaluation workshop, the students were trained on how to use the StereoWS tool including the stereo display, manual measurements, and statistical analysis procedure. In this workshop, we have collected manual measurements, which were selected by different members of the workshop.

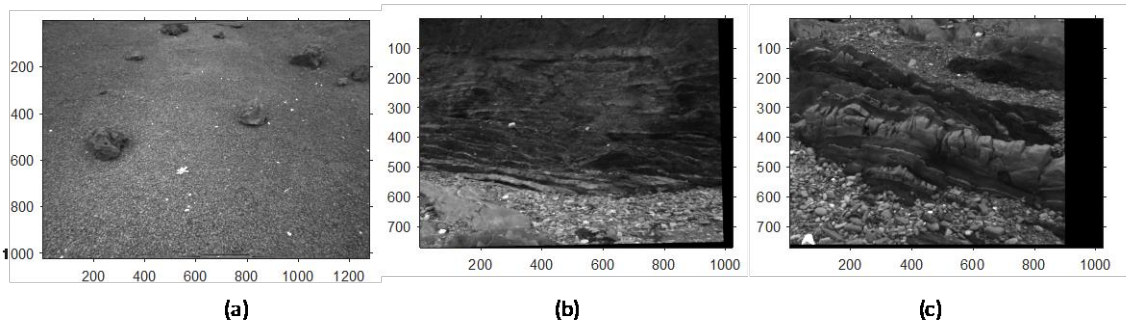
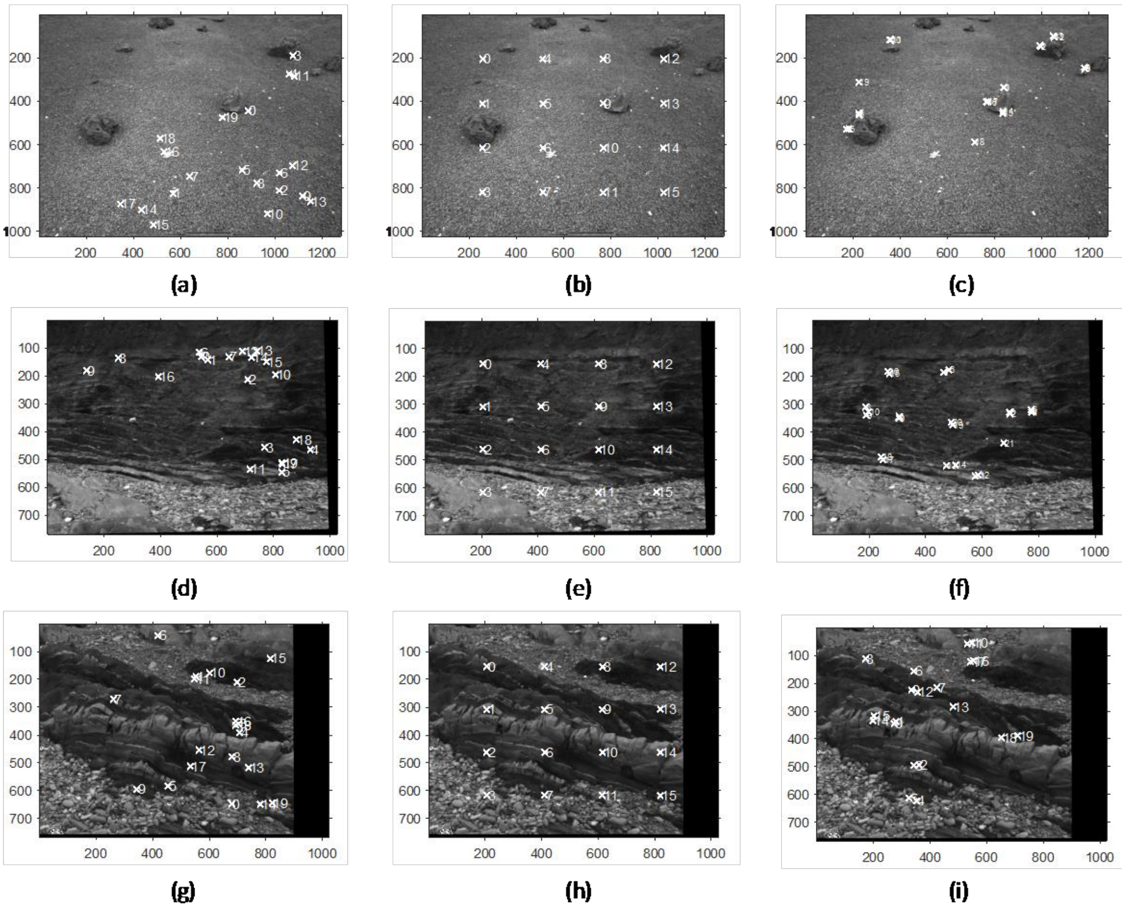


Figure 4: Test datasets from PRoVisG Mars 3D Challenge and ExoMars PanCam Test Campaign, showing left-eye images randomly picked from each test dataset; (a) C33 (b) 65246 (c) 70000.

During the manual measurement process, each participant was asked to collect 20 feature based points, 16 regular grid points, and 10 discontinuity tie-points for each pair of test images shown in Fig. 4. Figure 5 illustrates an example of left tie-points of some of the test images (i.e. C33, 65246, 7000) prepared for measurement.

For the feature based tie-points (see the first column of Fig. 5), participants only needed to identify the corresponding right points using the stereo display. 20 left points are selected from the extracted Scale Invariant Feature Transform (SIFT) key-points (Lowe, 2004) with the highest matching similarity values. For the regular gird tie-points (see the second column of Fig. 5), we collected 16 left points from the dense disparity map generated by our in-house GOTCHA matcher. Participants were then asked to validate their matching correctness based on visual clues by moving the stereo 3D cursor around the grid points to check if there were any abnormalities and adjusting the disparity offset of the stereo cursor at certain points to seek for better solutions. Results in this case that passed the validation were collected and averaged. For discontinuity tie-points (see the last column of Fig. 5), an expert user from the workshop manually selected 10 pairs of left points around the object edge and other problematic areas. 9 pairs of discontinuity tie-points are defined around an object boundary in C33, whilst the last two tie-points are selected



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 5: Example of left tie-points used in the stereo workshop: (a), (d), and (g) show 20 left tie-points defined on the test images shown in Fig. 4(a), (b), and (c), respectively; (b), (e), and (h) show 16 regular grid tie-points for the same test images; (c), (f), and (i) are for 20 tie-points around discontinuities.

from a relatively smooth and less-textured area. [see Fig. 5(c)]. Other workshop participants then defined the correspondences on the right image.

3.2 Evaluation of collected tie-points

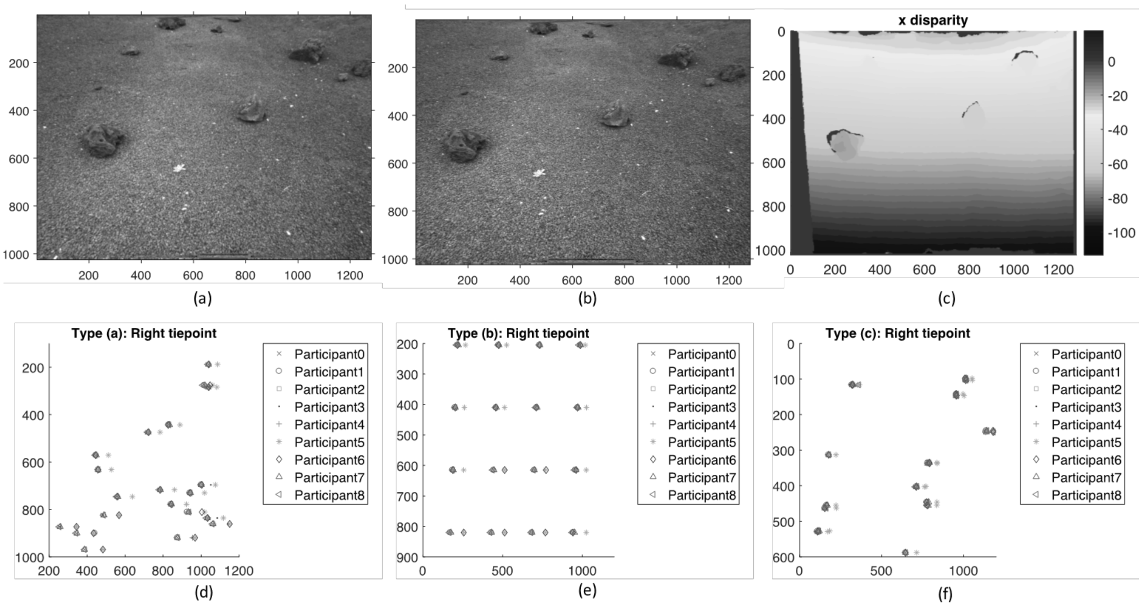


Figure 6: Example evaluation results of manual selection: (a) and (b) left and right input image of C33; (c) a disparity map of (a) produced by UCL which was used to identify outliers in manual measurements; (d), (e), and (f) show all measured right tie-points for type (a), (b), and (c), respectively.

The manual selection results from the 9 workshop participants are presented in Fig. 6, where input data are shown in the first row, whilst the positions of measured right tie-points are presented in the second row.

It appears that some of the workshop participants can perform good visual identification and visual validation with all three types of tie-points. On the other hand, a few workshop participants were not good at fusing the stereo images. For example, participant 1, participant 3, participant 5, participant 6 were not able to select good right points for the feature based tie-points [see Fig. 7(a)], and the performance of participant 5, participant 6, participant 8 was particularly worse with discontinuity tie-points [see Fig. 7(e)]. Their average measurement error (i.e. e_{out}) is 16.65 pixels which is significantly above the error bounds from a normal visual identification and validation results. Their performance was improved when a pre-computed disparity map is given

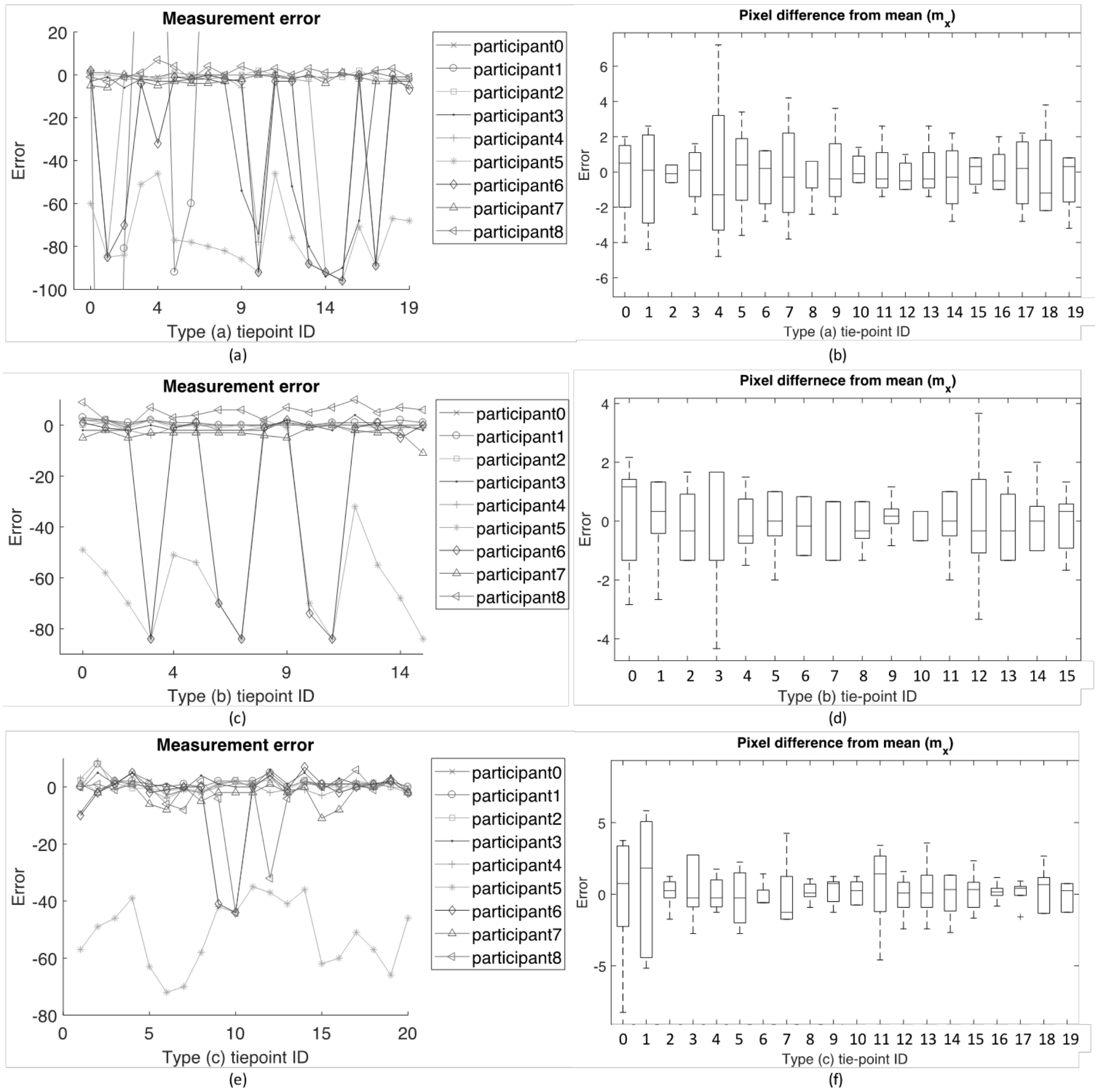


Figure 7: Example of evaluation of manual measurements of C33: (a), (c), and (e) simple measurement error from (1) of type (a)-(c) tie-points, respectively; (b), (d), and (f) show bar charts of difference between inlier measurements and m_x

although two participants still cannot visualise the tie-points in 3D, i.e. Participants 5 and 6 [see Fig. 7(c)]. These outliers were then removed before calculating the error bounds.

Figure 7(a), (c), and (e) summarise the errors from the inlier means $d(\boldsymbol{t}_i^k, \boldsymbol{m}_i^k)$ of all tie-points from 9 participants. It is observed that tie-points from the indistinctive textures are generally difficult to select, for example, $\boldsymbol{t}_1^a$, $\boldsymbol{t}_4^a$, $\boldsymbol{t}_5^a$, $\boldsymbol{t}_7^a$, and $\boldsymbol{t}_9^a$ in the feature based tie-points have larger measurement variation and more outliers [see Fig. 7(b)]. This reconfirms our understanding that a stereo visualisation can help us detect correct tie-points better around the object boundary than within plain/repetitive texture.

One interesting observation from the error graph is that the performance of participant 5, who consistently produced a large measurement error regardless of the type of dataset, deteriorates when a tie-point is closer to a camera (i.e. a larger x disparity). For example, the measurement errors for $\boldsymbol{t}_3^b$, $\boldsymbol{t}_7^b$, $\boldsymbol{t}_{11}^b$, and $\boldsymbol{t}_{15}^b$ (which is the bottom row of the grid in Fig. 5(b)) are getting worse than the rest and we can see this pattern in Fig. 7(c).

The error metrics of measurements are evaluated and summarised in table 1. Without the removal of outliers, the total measurement error increases significantly. The maximum of e_{tot} was recorded with the feature based tie-points (20.83), whereas the minimum (8.39) was obtained from the discontinuity tie-points. However, after removing obvious outliers (i.e. $\delta > 10$ in (2)), the measurement errors drop sharply to less than 2 pixels with small standard variation (see e_{in} and avg. σ_x in table 1). As mentioned earlier, we believe this happens because of the outliers introduced by a few participants who fuse a stereo pair differently than the rest.

Table 1 Measurement errors of C33 (N.B. the Type (a) results of participant 2 was excluded due to the incomplete of measurements.)

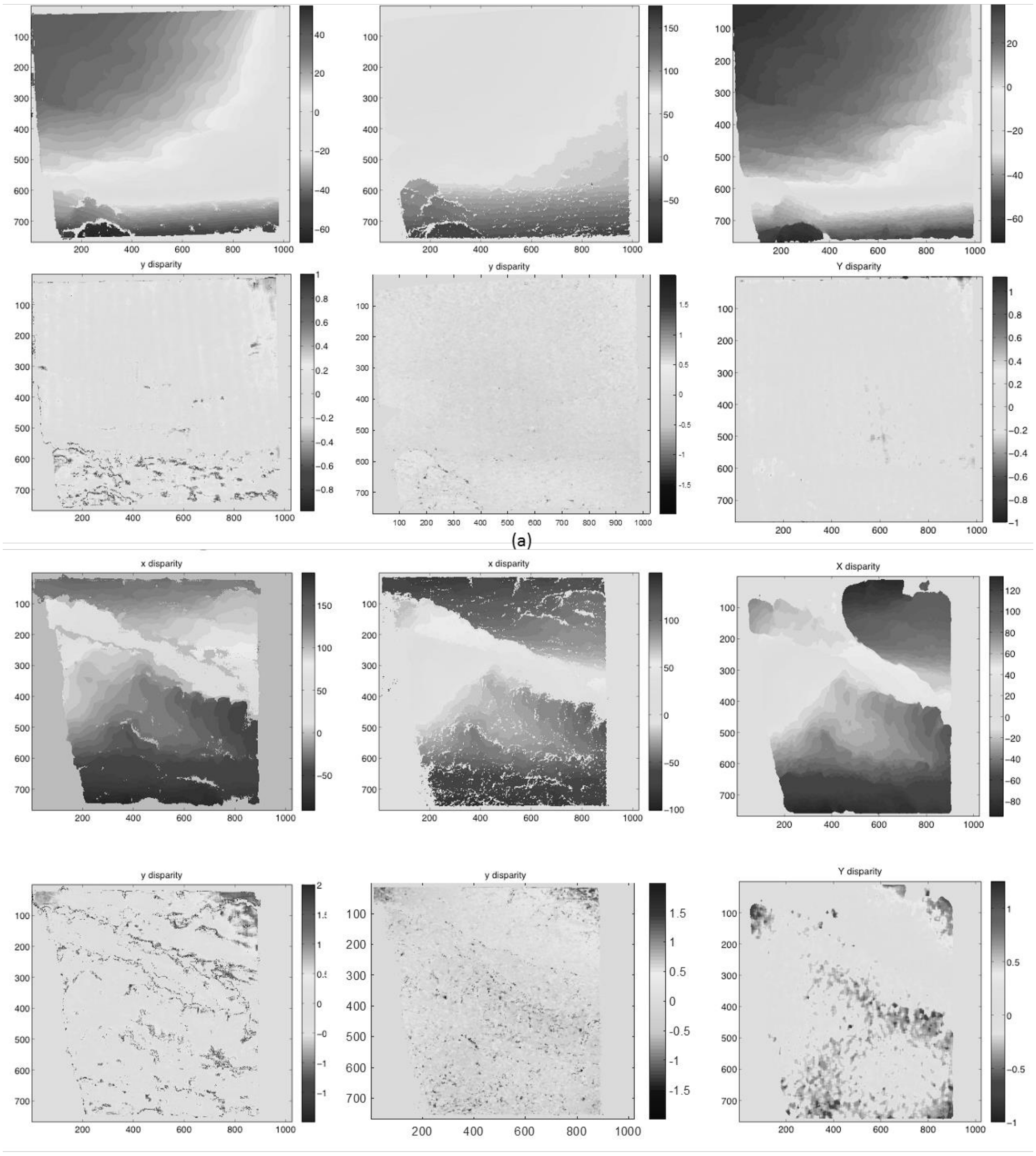
Type	e_{tot}	e_{in}	e_{out}	avg. σ_x
a	20.83	1.61	40.04	0.92
b	10.83	1.10	22.98	1.71
c	8.39	1.78	16.65	0.93
avg.	13.35	1.50	26.56	1.19

The bar charts of the inlier measurements for 3 datasets are shown in the second column of Fig. 7. Each bar chart summarises the differences between the inlier measurements and the mean of the inlier measurements. Type (b) tie-point selection appears to be more difficult as participants are often required to fuse the stereo cursor around textureless or smooth (i.e. small depth separation) areas. As a consequence, the inlier measurements of regular grid tie-points are generally inconsistent (i.e. avg. $\sigma_x = 1.71$) compared to the others [see Fig. 7(d)]. On the other hand, strong depth discontinuity around an object boundary from type (c) tie-points improve the consistency of the measurements [see Fig. 7(f)]. We have found that the maximum standard deviation is 2.56 pixels, the minimum standard deviation is 0.37 pixels, and the average is 0.93 pixels.

It is also interesting to see that SIFT keypoints performs the best for stereo fusion. Its average standard deviation is 0.92 which is marginally better than the second best but the left tie-points of type (a) were selected simply based on the texture information [see Fig. 7(b)]. We think that the distinctive gradient information around a keypoint can improve the performance of stereo measurements.

3.3. Results of automated stereo matching

In our evaluation, we have collected two sets of processing results (i.e. a x and y disparity map) from UCL, JPL, and JR. Fig. 8(a) and (b) respectively represent these disparity maps of dataset 65246 and 70000 from ExoMars PanCam Test Campaign, and each column of the figure represents the results from different organisations. To our best knowledge all three algorithms have been developed based on a variation of a correlation based stereo matching algorithm with an adaptive least square fitting technique (Deen and Lorre, 2005; Otto and Chau, 1989), but all results seem to be slightly different in terms of the completeness and the estimated values of a disparity map. All three results were able to produce a relatively denser disparity map with dataset 65246. However, the results seem different with the other dataset, e.g. the JR result shown in the last column of Fig. 8(b) looks overly smoothed and its density is more incomplete than the other two (but this does not mean it is sparse). Please also note that both y disparity maps from JPL (see the second column of Fig. 8) contains a few spikes which are removed for visualisation.



(a)

(b)

Figure 8: Example of disparity maps: (a) x and y disparity maps of dataset 65246; (b) and dataset 70000. UCL, JPL, and JR results are shown in the first, the second and the last column.

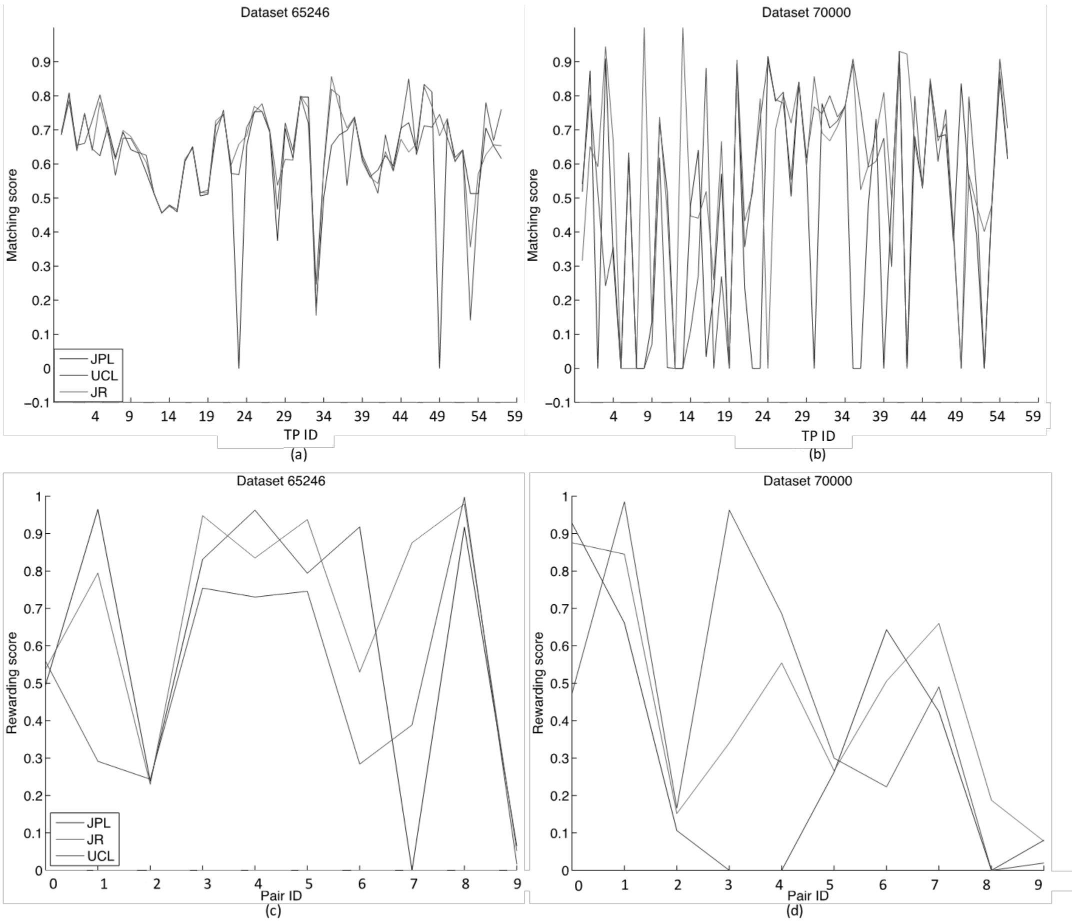
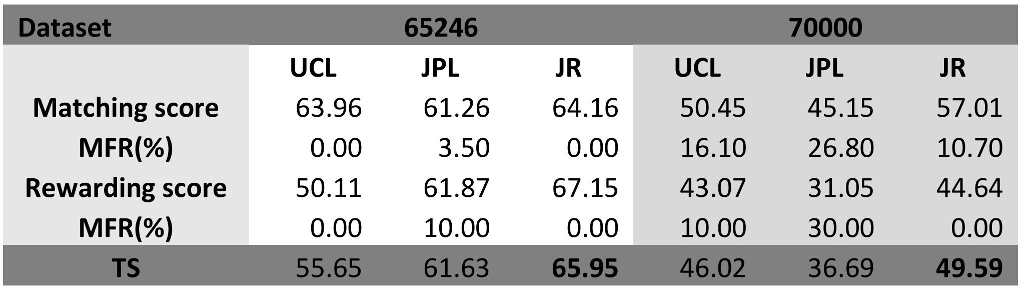


Figure 9: (a) Individual matching scores of the processing results of two datasets; (b) Rewarding scores from 10 tie-point pairs in two datasets.

Given the error bounds calculated from the manual measurements, the matching scores and rewarding scores of each tie-point are evaluated and the results are shown in Fig. 9. Matching scores of three algorithms are generally similar when they can define a tie-point, but when it fails to define a tie-point no score was awarded, e.g. see JPL matching scores of ID 23 and 49 in Fig. 9(a). The rewarding score of UCL's disparity map is generally lower than the other two with the dataset 65246 [see Fig. 9(c)]. However, it is improved with the other dataset having more depth discontinuities.

The total scores were calculated using an equal weight of the matching scores and rewarding scores, and the results are summarised in table 2, where the best scores for certain datasets are labelled in bold font. We can observe that for dataset 65246 that JR's stereo matching pipeline produced the best result for the overall area. To understand this result clearly, it is worth mentioning that the total score (TS) shown in (8) has been designed to award more scores if a disparity map defines all queried tie-points; in other words, no score is given if there is no corresponding tie-point in a disparity map. Thus, this metric is generally favoured for a dense and smooth disparity map, which we believe why JR's results perform best on both test datasets.

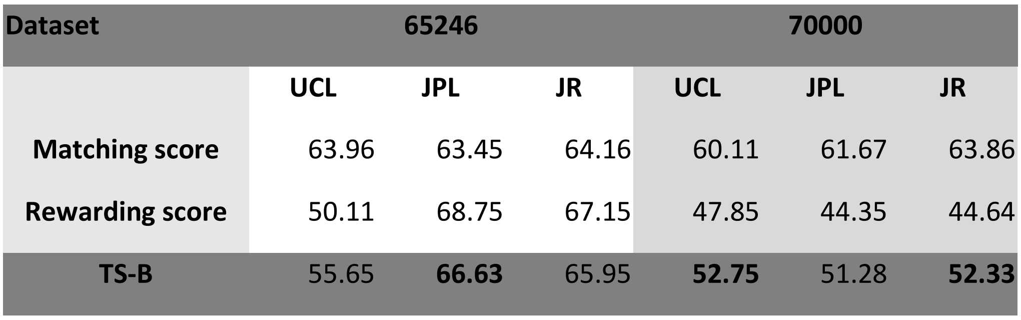
Table 2: Total score (TS) estimated from (8) with $\alpha = 0.5$



Dataset	**65246**			**70000**		
	UCL	**JPL**	**JR**	**UCL**	**JPL**	**JR**
Matching score	63.96	61.26	64.16	50.45	45.15	57.01
MFR(%)	0.00	3.50	0.00	16.10	26.80	10.70
Rewarding score	50.11	61.87	67.15	43.07	31.05	44.64
MFR(%)	0.00	10.00	0.00	10.00	30.00	0.00
TS	55.65	61.63	**65.95**	46.02	36.69	**49.59**

To give more weight on the accuracy of an algorithm, we modified (8) not to penalise when they failed to define a queried tie-point in a disparity map, and called this score, TS-B. The results of TS-B of both datasets are also presented in table 3.

Table 3: Total score B (TS-B) which is similar to TS but removes the effect of missing tie-points



Dataset	**65246**			**70000**		
	UCL	**JPL**	**JR**	**UCL**	**JPL**	**JR**
Matching score	63.96	63.45	64.16	60.11	61.67	63.86
Rewarding score	50.11	68.75	67.15	47.85	44.35	44.64
TS-B	55.65	**66.63**	65.95	**52.75**	51.28	**52.33**

We also introduce a new term MFR representing the Matching Failure Rate. MFR can be used as an indicator for either the incompleteness of a disparity map or how conservative the algorithm is. As shown in table 2, JPL's results have higher MFR, but without counting on the match failure area (i.e. using TS-B) JPL's pipeline produced the best result on the dataset 65246. For dataset 70000, JPL's pipeline gets the second best score whilst UCL's processing pipeline has produced the best accuracy.

4. Discussions and Conclusions

In this paper, we introduced an accuracy evaluation method to assess the stereo matching results. The main motivation of this work is to provide a straightforward method which can be applied to the stereo matching evaluation work of planetary rover missions, where it is currently impossible to obtain ground truth data.

We have demonstrated the use of a generic portable stereo workstation including a stereo cursor from the open source StereoWS tool to produce visually correct tie-points of a stereo pair, i.e. manual tie-points, with the help of a softcopy stereo display. The manual tie-points from stereo measurements are not identical for all candidate tie-points, but our assumption is that the variation of multiple measurements can be used to estimate the confidence of a tie-point and this confidence values can quantitatively evaluate the quality of disparity maps from different algorithms. Based on this idea, we have defined useful evaluation metrics using the statistics of multiple measurements (such as means and variance). We also define three types of tie-points to test the performance at highly textured region, textureless region, and occluded region. The performance of textureless region is quite interesting for DTM construction from orbital imagery but this is left for the future work. Type (b) tie-points are related to the scene occlusion. At the moment, we populate these points manually but it is also possible to design a semi-automatic pipeline to collect these points, e.g. detect one tie-point by conventional feature detector and find adjacent feature from background manually.

It is worth noting that in these experiments, the number of tie-points, particularly for the discontinuities, may not be sufficient in some cases. It would have been better to add more tie-points. However, we erred on the side of setting an experiment which could be accomplished with a group of ten "citizen scientists" within a limited time period (a week). Other comparison results, e.g. disparity density or 3D accuracy, could also be employed in future experiments to improve the final matching score.

During the evaluation work, we implemented an open source stereo workstation with an integrated stereo matching method that is used to produce the UCL results shown in the evaluation. We have published the Java code of the Stereo Workstation on SourceForge under a BSD license (available from SourceForge, http://sourceforge.net/projects/stereows/) to encourage other stereo researchers to use and modify our system for their own evaluation.

The experiments reported in this paper focused on planetary images. It would be straightforward to apply this method and our StereoWS to any future stereo research projects when any quantitative evaluation is required, wherever it is on Mars or the Earth or anywhere else for that matter. In the future, we hope our efforts could also benefit the stereo correspondence evaluation work and include more datasets, in particular the results from a wider variety of general stereo. Also, we expect that the same idea behind StereoWS could be applied to develop a more intuitive and immersive stereo measurement system using recent virtual reality technologies. In conjunction with the stereo measurement workshop held in 2011, we can provide the possibility of evaluation of these stereo matching results including more methods from our collaborators.

As future work, it is also interesting to investigate the performance of manual measurements from different lighting conditions (Kirk et. Al., 2016). We could measure the variation of human depth perception under different illumination effects and reflect this on (5) to define more accurate metrics. However, this is currently beyond our research scope and left for the future work.

5. References

Azari, H., Cheng, I., & Basu, A. (2009). Stereo 3D mouse (S3D-mouse): Measuring ground truth for medical data in a virtual 3D space. In Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE (pp. 5744–5747). IEEE.

Birchfield, S., & Tomasi, C. (1998). Depth discontinuities by pixel-to-pixel stereo. In Computer Vision, 1998. Sixth International Conference on (pp. 1073–1080). IEEE.

Boykov, Y., Veksler, O., & Zabih, R. (2001). Fast approximate energy minimization via graph cuts. IEEE Transactions on pattern analysis and machine intelligence, 23, 1222–1239.

Chan, H.-P., Goodsitt, M. M., Hadjiiski, L. M., Bailey, J. E., Klein, K., Darner, K. L., & Sahiner, B. (2003). Effects of magnification and zooming on depth perception in digital stereomammography: an observer performance study. Physics in medicine and biology, 48, 3721.

Day, T., & Muller, J.-P. (1989). Digital elevation model production by stereo-matching spot image-pairs: a comparison of algorithms. Image and Vision Computing, 7, 95–101.

Deen, R. G., & Lorre, J. J. (2005). Seeing in three dimensions: correlation and triangulation of mars exploration rover imagery. In Systems, Man and Cybernetics, 2005 IEEE International Conference on (pp. 911–916). IEEE volume 1.

Di, K., & Li, R. (2004). CAHVOR camera model and its photogrammetric conversion for planetary applications. Journal of Geophysical Research: Planets, 109.

Fleet, D. J., Wagner, H., & Heeger, D. J. (1996). Neural encoding of binocular disparity: energy models, position shifts and phase shifts. Vision research, 36, 1839–1857.

Geiger, A., Lenz, P., & Urtasun, R. (2012). Are we ready for autonomous driving? the KITTI vision benchmark suite. In Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on (pp. 3354–3361). IEEE.

Geman, S., & Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. IEEE Transactions on pattern analysis and machine intelligence, 6, 721–741.

Gruen, A. (1985). Adaptive least squares correlation: a powerful image matching technique. South African Journal of Photogrammetry, Remote Sensing and Cartography, 14, 175–187.

Hannah, M. J. (1974). Computer matching of areas in stereo images. Ph.D. thesis Stanford university. 13

Hirschmuller, H. (2008). Stereo processing by semiglobal matching and mutual information. IEEE Transactions on pattern analysis and machine intelligence, 30, 328–341.

Jaumann, R., Neukum, G., Behnke, T., Duxbury, T. C., Eichentopf, K., Flohrer, J., Gasselt, S., Giese, B., Gwinner, K., Hauber, E. et al. (2007). The high-resolution stereo camera (HRSC) experiment on mars express: Instrument aspects and experiment conduct from interplanetary cruise through the nominal mission. Planetary and Space Science, 55, 928–952.

Jensen, R., Dahl, A., Vogiatzis, G., Tola, E., & Aanæs, H. (2014). Large scale multi-view stereopsis evaluation. In 2014 IEEE Conference on Computer Vision and Pattern Recognition (pp. 406–413).

Kirk, R., Howington-Kraus, E., Hare, T., & Jorda, L. (2016). The effect of illumination on stereo DTM quality: Simulations in support of europa exploration. ISPRS Annals of Photogrammetry, Remote Sensing &amp; Spatial Information Sciences, 3.

Lhuillier, M., & Quan, L. (2002). Match propagation for image-based modeling and rendering. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24, 1140–1146.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. International journal of computer vision, 60, 91–110.

Morales, S., & Klette, R. (2011). Ground truth evaluation of stereo algorithms for real world applications. In Computer Vision–ACCV 2010 Workshops (pp. 152–162). Springer.

Nakamura, Y., Matsuura, T., Satoh, K., & Ohta, Y. (1996). Occlusion detectable stereo-occlusion patterns in camera matrix. In Computer Vision and Pattern Recognition, 1996. Proceedings CVPR'96, 1996 IEEE Computer Society Conference on (pp. 371–378). IEEE.

Otto, G. P., & Chau, T. K. (1989). 'Region-growing' algorithm for matching of terrain images. Image and vision computing, 7, 83–94.

Pariser, O., & Deen, R. G. (2009). A common interface for stereo viewing in various environments. In IS&T SPIE Electronic Imaging (pp. 72371R–72371R). International Society for Optics and Photonics.

Scharstein, D., & Szeliski, R. (2002). A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. International Journal of Computer Vision, 47, 7–42.

Scharstein, D., & Szeliski, R. (2003). High-accuracy stereo depth maps using structured light. In Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on (pp. I–I). IEEE volume 1.

Scharstein, D., Szeliski, R., & Zabih, R. (2001). A taxonomy and evaluation of dense two-frame stereo correspondence algo- rithms. In Stereo and Multi-Baseline Vision, 2001.(SMBV 2001). Proceedings. IEEE Workshop on (pp. 131–140). IEEE.

Seitz, S. M., Curless, B., Diebel, J., Scharstein, D., & Szeliski, R. (2006). A comparison and evaluation of multi-view stereo reconstruction algorithms. In Computer vision and pattern recognition, 2006 IEEE Computer Society Conference on (pp. 519–528). IEEE volume 1.

Shin, D., Muller, J., & Poole, W. (2011). Open source software tools for joint ESA-NASA mars exploration developed in the EU-FP7 PRoVisG project. In Royal Astronomical Society Aurora meeting.

Simioni, E., Da Deppo, V., Naletto, G., Cremonese, G., & Re, C. (2014). Stereo camera for satellite application: A new testing method. In Metrology for Aerospace (MetroAeroSpace), 2014 IEEE (pp. 582–587). IEEE.

Sun, J., Zheng, N.-N., & Shum, H.-Y. (2003). Stereo matching using belief propagation. IEEE Transactions on pattern analysis and machine intelligence, 25, 787–800.