A Measure for Accuracy Disparity Maps Evaluation

Ivan Cabezas, Victor Padilla, and Maria Trujillo

Escuela de Ingeniería de Sistemas y Computación, Universidad del Valle,
Ciudadela Universitaria Melendez, Cali, Colombia
{ivan.cabezas,victor.padilla,maria.trujillo}@correounivalle.edu.co

Abstract. The quantitative evaluation of disparity maps is based on error measures. Among the existing measures, the percentage of Bad Matched Pixels (BMP) is widely adopted. Nevertheless, the BMP does not consider the magnitude of the errors and the inherent error of stereo systems, in regard to the inverse relation between depth and disparity. Consequently, different disparity maps, with quite similar percentages of BMP, may produce 3D reconstructions of largely different qualities. In this paper, a ground-truth based measure of errors in estimated disparity maps is presented. It offers advantages over the BMP, since it takes into account the magnitude of the errors and the inverse relation between depth and disparity. Experimental validations of the proposed measure are conducted by using two state-of-the-art quantitative evaluation methodologies. Obtained results show that the proposed measure is more suited than BMP to evaluate the depth accuracy of the estimated disparity map.

Keywords: Computer vision, corresponding points, disparity maps, quantitative evaluation, error measures.

1 Introduction

A stereo image set captures a 3D scene from slightly different viewpoints. A disparity estimation algorithm takes as input a stereo image set, and produces a set of disparity maps (DM) as output. Disparity is the shift between stereo corresponding points. The 3D structure of the captured scene can be recovered based on estimated disparities. The estimation of DM is a fundamental problem in computer vision, which has to be addressed in several applications domains, such as: robotics, unmanned vehicles, entertainment and telecommunications, among others [6], [12], [16]. The evaluation of DM, in terms of estimation accuracy, is quite important since small inaccuracies may have a large impact on the results of the 3D final reconstruction. Moreover, the objective comparison of different disparity estimation algorithms is based on the quantitative evaluation of DM [10], [15]. This evaluation allows also for the tuning of parameters of an algorithm within a particular context [7], determining the impact of specific components and procedures [5], and decision taking for researchers and practitioners

C. San Martin and S.-W. Kim (Eds.): CIARP 2011, LNCS 7042, pp. 223–231, 2011.

among others. In fact, a quantitative evaluation approach must be supported by a quantitative evaluation methodology [2]. Among the different components that a quantitative evaluation methodology may involve, the set of error measures is a fundamental one.

In some scenarios, the quantitative evaluation of DM has to be conducted in the absence of ground-truth data. In this case, a prediction error approach can be used to perform the evaluation [14]. This approach consists in comparing a set of third views of the scene, against a set of rendered views computed from reference images and their associated DM.

Image quality measures such as the Mean Squared Error (MSE), the Root Median Squared Error (RMSE), the Peak Signal-to-Noise Ratio (PSNR), and the Structural Similarity Index Measure (SSIM) [19] can be used for quantitative evaluation under a prediction error approach [15]. Although, the MSE, the RMSE, and the PSNR are widely adopted and have a clear physical meaning, they are not closely related to the perceived visual quality by the human visual system [18], [19].

The disparity gradient and the disparity acceleration indices are presented in [20] to measure the smoothness of the DM. These indices require the use of thresholds. However, no information is provided about how the threshold can be fixed. Moreover, the capability of these indices to distinguish between an inaccurate estimation and a true depth discontinuity is not discussed. On the other hand, the fact that the DM may vary smoothly but, at the same time, they may be totally inaccurate is ignored.

The comparison of results using the SSIM and the PSNR measures on noisy DM by adding salt and pepper is addressed in [13]. Although it is concluded in [13] that obtained PSNR values are closer to the scores assigned by subjective evaluation, this conclusion does not coincide with the well-known drawback of the PSNR [18], [19]. Additionally, there is not a clear relation between the type and the level of noise introduced, and the artifacts that a disparity estimation algorithm may produce. Consequently, the considered evaluation scenario lacks of realism.

Ground-truth based error measures can be computed by comparing estimated DM against disparity ground-truth data. Measures such as, the Mean Absolute Error (MAE), the MSE, and the RMSE are considered in [10], [16] for ground-truth based evaluation. A modification of SSIM, termed R-SSIM, and designed for range images, is proposed in [8]. The modification consists in the introduction of the capability to handle missing data in both, the ground-truth disparity map, and in the estimated DM. It is shown in [8] that there exists a strong linear association between the BMP and the R-SSIM.

A modification of the Mean Absolute Percentage Error (MAPE) is presented in [16]. The modification consists in the capability to handle the absence of estimations in the evaluated DM. Although MAPE considers the inverse relation between depth and disparity, it is designed in the context of forecasting [3]. Additionally, the use of the mean, which is sensitive to outliers, may introduce bias in the evaluation.

The BMP, was introduced in [10] as a component of the Middlebury's evaluation methodology [9], [11]. It is formulated in Equation (1).

$$\text{BMP} = \frac{1}{N} \sum_{(x,y)} \varepsilon_{(x,y)}; \;\; \varepsilon_{(x,y)} = \begin{cases} 1 & \text{if } |D_{true}(x,y) - D_{estimated}(x,y)| > \delta \\ 0 & \text{if } |D_{true}(x,y) - D_{estimated}(x,y)| \leq \delta \end{cases}, \quad (1)$$

where, D_{true} is the disparity ground-truth data, $D_{estimated}$ is the disparity map under evaluation, and δ is the error tolerance threshold (commonly, $\delta = 1$).

The error tolerance threshold δ is considered by the BMP in order to determine if there is a disparity estimation error. The BMP can be gathered on different image regions, related to different image phenomena, such as occluded, near to depth discontinuities, and areas lacking of texture, among others [10].

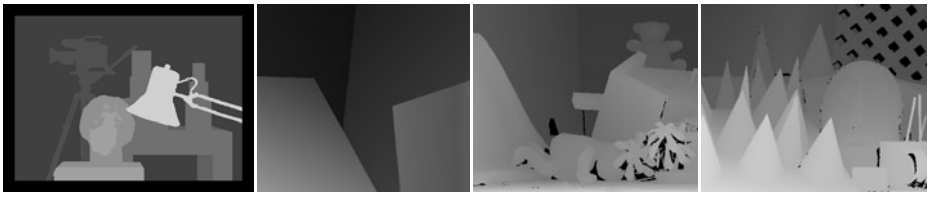
Among the existing quantitative measures, the BMP is widely used. However, it is a measure of the quantity of errors occurring in DM. Moreover, such a quantity may do not indicate how accurately a particular disparity map fulfils the task for which it was estimated: to recover the depth of the scene captured in the stereo image set [4]. In fact, the BMP can be seen as a binary function by the using of a threshold, which selection may impact on the evaluation results.

In this paper, a ground-truth based measure is presented. The proposed measure is supported by the inverse relation between depth and disparity. It computes a global error measure with a physical interpretation and without thresholds intervention.

2 Problem Statement

The BMP is commonly used as a measure of disparity errors evaluation. Nevertheless, in practice, the estimation of the DM is an intermediate step on a process, which the ultimate goal is to achieve depth accuracy. In fact, the BMP has drawbacks such as: it may be sensitive to the selection of δ, since small changes on this value, may lead to obtain significantly different percentages. Moreover, the magnitude of the difference between the estimated disparity and the ground-truth value is ignored. Thus, the BMP may conceal disparity estimation errors of large magnitude, and at the same time, it may penalise errors of low impact on the final 3D reconstruction. On the other hand, disparity estimation errors of the same magnitude may cause depth errors of different magnitude. However, the BMP does not consider this fact. Consequently, the BMP measure is not suited to measure the depth accuracy of a disparity estimation process.

The DM of the Tsukuba, Venus, Teddy and Cones stereo images [10], [11] are used in Fig. 1 for illustrating the stated problem. These maps are varying smoothly, and their percentages of BMP are equals to zero. However, Table 1 shows that the values of other ground-truth based measures, as well as image quality values of rendered views, computed from the DM, are contradicting to the values reported by the BMP. It can be observed, that although the BMP is reporting a perfect accuracy on the entire image, the other ground-truth based error measures, the MSE and the MAPE, are indicating error presence. Additionally, the MSSIM, the MSE and the PSNR of rendered views are indicating



(a) Tsukuba (b) Venus (c) Teddy (d) Cones

Fig. 1. DM, varying smoothly but being totally wrong

Table 1. Obtained values of the BMP ($\delta = 1$), the MSE, the MAPE, based on ground-truth; and obtained values of the MSSIM, the MSE and the PSNR, based on rendered views, by using the DM in Fig. 1

Disparity Map	BMP all	MSE all	MAPE all	MSSIM	MSE	PSNR
Fig. 1(a)	0.000	1.000	16.474	0.758	187.091	25.410
Fig. 1(b)	0.000	1.000	14.316	0.792	173.636	25.734
Fig. 1(c)	0.000	1.000	4.117	0.831	128.543	27.040
Fig. 1(d)	0.000	1.000	3.380	0.744	184.313	25.475

a low quality. This exemplifies the sensitivity of the BMP to the selection of δ, and the fact that obtaining a low percentage of BMP does not imply, necessarily, that the DM under evaluation are accurate in terms of 3D scene reconstruction.

3 The Sigma-Z-Error

The proposed measure in this paper is termed Sigma-Z-Error (SZE). It is based on the inverse relation between depth and disparity using the error magnitude. In this sense, it aims to measure the final impact of a disparity estimation error, which depends on the true distance between the stereo camera system and the captured point, and on the disparity error magnitude. The SZE is described as follows.

The distance between a point of the captured scene and the camera system can be computed, without loss of generality, based on the information of the stereo rig and the estimated disparity as is formulated in Equation (2).

$$Z_{true} = \frac{f * B}{d_{true}} \,, \tag{2}$$

where f is the focal length in pixels, B is the baseline in meters (i.e. the distance between optical centres), d_{true} is the true disparity value in pixels, and Z_{true} is the distance along the camera Z axis in meters.

However, in practice, an inaccurate Z distance is generated due to a disparity estimation error, as is formulated in Equation (3).

$$Z_{false} = \frac{f * B}{d_{false}} \,, \tag{3}$$

where Z_{false} is the inaccurate distance estimation, and d_{false} is the falsely estimated disparity.

The proposed SZE measure consists in summing the difference between Z_{true} and Z_{false}, over the entire estimated disparity map (or in a particular image region) based on the information provided by disparity ground-truth data. The SZE is formulated in Equation (4).

$$SZE = \sum_{(x,y)} \left| \frac{f * B}{D_{true}(x,y) + \mu} - \frac{f * B}{D_{estimated}(x,y) + \mu} \right| \,, \tag{4}$$

where, μ is a small constant which avoids the instability caused by missing disparity estimations. The SZE fulfils the properties of a metric. However, it is unbounded.

Table 2 shows the values of the SZE and the BMP, as well as the PSNR and the MSSIM of the rendered views using different DM (i.e. the ground-truth, an inaccurate map varying smoothly, and a map containing streaking artefacts) of the Cones stereo image. It can be observed that despite of the low values of the MSSIM and the PSNR, the BMP values are indicating that there is no estimation error.

Table 3 shows obtained values of the BMP, the MAE, the MSE, and the MAPE, based on disparity ground-truth data by three different disparity estimation algorithms [9], and using the Venus stereo image. It can be observed that the values of the BMP are quite similar. On the other hand, the values of the SZE and the MAPE are indicating that there exists a difference in the accuracy of the considered algorithms.

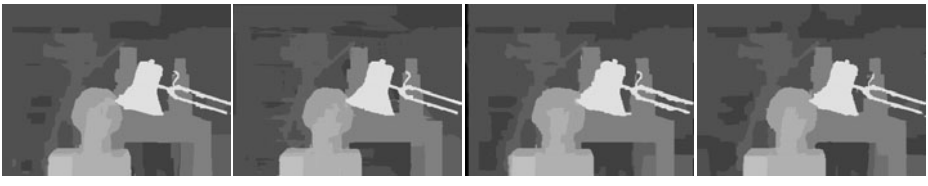
Fig. 2 illustrates estimated DM calculated using the Tsukuba stereo image, and four disparity estimation algorithms [9], which have similar results of the percentage of the BMP on the non-occluded region. Table 4 shows obtained values on the non-occluded region, in relation to DM in Fig. 2, of the SZE, the BMP, the MAE, the MSE, and the MAPE, as well as the MSSIM using rendered views. In this case, the obtained values of the SZE are consistent with

Table 2. Obtained ground-truth based error measures and rendered image quality measures considering DM of the Cones stereo image

Disparity Map	SZE nonocc	SZE all	SZE disc	BMP nonocc	BMP all	BMP disc	PSNR	MSSIM
Ground-truth	0.000	0.000	0.000	0.000	0.000	0.000	29.948	0.903
Inaccurate	193.703	218.905	66.945	0.000	0.000	0.000	25.475	0.774
Artefacts	11.088	11.800	3.524	0.000	0.000	0.000	24.639	0.729

Table 3. Obtained values of the SZE, the BMP, the MAE, the MSE and the MAPE, considering different algorithms, and using the Venus image

Algorithm	SZE all	BMP all	MAE all	MSE all	MAPE all
CoopRegion	552.274	0.206	0.106	0.076	1.849
Undr+OvrSeg	735.384	0.224	0.199	0.097	2.815
AdaptingBP	929.368	0.212	0.165	0.104	3.069



(a) (b) (c) (d)

Fig. 2. Disparity maps of the Tsukuba image, estimated by: (a) DoubleBP, (b) CoopRegion, (c) GlobalGCP, (d) OutlierConf

Table 4. Obtained ground-truth based error measures and rendered image quality measures for disparity maps in Fig. 2

Algorithm	SZE nonocc	BMP nonocc	MAE nonocc	MSE nonocc	MAPE nonocc	MSSIM
DoubleBP	658.867	0.880	0.223	0.475	3.764	0.908
CoopRegion	662.485	0.872	0.228	0.507	3.780	0.905
GlobalGCP	817.656	0.868	0.263	0.530	4.560	0.908
OutlierConf	915.254	0.879	0.284	0.550	4.921	0.908

the obtained values of the MAE, the MSE, and the MAPE, and contradictories with the percentage of the BMP. On the other hand, the MSSIM values may be indicating that the quality of the rendered views may appear quite similar for a human observer.

4 Experimental Evaluation

In order to assess the impact of the proposal on evaluation results, the SZE and the BMP are considered as the error measures during an evaluation process. The top fifteen ranked algorithms in [9] (May, 2011) are selected as the algorithms under evaluation, and Tsukuba, Venus, Teddy and Cones stereo images are selected as the test-bed [11].

Two evaluation methodologies are used: the Middlebury methodology [9], [10], [11], and the $\mathbf{A}^*$ methodology [2]. The $\mathbf{A}^*$ methodology is a non-linear evaluation

Table 5. Quantitative evaluation of algorithms considering the SZE and the BMP as error measures, using the Middlebury and the $\mathbf{A}^*$ evaluation methodologies, respectively

Algorithm	SZE Avg. Rank	SZE Rank	SZE Algorithm $\in A^*_{(\mathrm{SZE})}$	BMP Avg. Rank	BMP Rank	BMP Algorithm $\in A^*_{(\mathrm{BMP})}$
GC+SegmBorder	1.17	1	Yes	9.58	11	Yes
SubPixDoubleBP	5.25	2	No	8.50	9	Yes
CoopRegion	5.92	3	No	5.33	3	Yes
SurfaceStereo	7.00	4	No	8.00	8	Yes
FeatureGC	7.67	5	Yes	8.75	10	Yes
CostFilter	7.83	6	No	11.25	15	No
ObjectStereo	8.08	7	No	7.92	7	Yes
AdaptingBP	8.42	8	No	4.83	2	Yes
Undr+OvrSeg	8.50	9	No	10.08	13	Yes
DoubleBP	8.83	10	Yes	6.33	4	Yes
WarpMat	9.75	11	No	9.75	12	Yes
GlobalGCP	9.83	12	No	10.92	14	Yes
OutlierConf	10.00	13	No	7.25	5	Yes
RDP	10.42	14	No	7.42	6	Yes
ADCensus	11.33	15	No	4.08	1	Yes

methodology. It computes the Pareto optimal set (denoted as A^*) from the set of algorithms under evaluation (denoted as A), by considering vectors of error measures [1], [17]. In this way, the set A^* contains those algorithms of comparable performance among them, and at the same time, of superior performance to $A \backslash A^*$.

Table 5 shows evaluation results of the error measures and the evaluation methodologies considered. It can be observed that using the SZE the evaluation results are significantly different, in both methodologies, to the results obtained by using the BMP as the error measure. Moreover, the smaller cardinality of the set A^*, when the SZE measure is used, can be attributed to a larger uniformity in the error measurements.

5 Conclusions

In this paper, the SZE is introduced as a measure for evaluating quantitatively estimated DM. It is based on the inverse relation between depth and disparity. The SZE offers advantages over the BMP, since it is focused on the impact of disparity estimation errors in terms of distance along the Z axis. In this way, it is related to an error value with a physical interpretation and meaning. Moreover, the SZE does not require the use of thresholds, which may introduce bias to the evaluation results. The analysis of different estimated DM shows that, under different circumstances, the BMP may not reflect properly the accuracy,

in terms of depth, of the estimated disparity map. On the other hand, the SZE is consistent with other measures.

Innovative results in relation to algorithms evaluation were obtained when the SZE was used to support the quantitative evaluation, since it leads to a different ranking, by using the Middlebury evaluation methodology and a different composition of the set A^* by using the $\mathbf{A}^*$ evaluation methodology. Thus, the algorithms that are reported as achieving the most accurate DM, based on the BMP measure, may not necessarily correspond to those allowing the most accurate 3D reconstruction.

References

1. Ben Said, L., Bechikn, S., Ghedira, K.: The r-Dominance: A New Dominance Relation for Interactive Evolution Multi-criteria Decision Making. IEEE Trans. On Evolutionary Computation 14(5), 801–818 (2010)
2. Cabezas, I., Trujillo, M.: A Non-linear Quantitative Evaluation Approach for Disparity Estimation. In: Proc. Intl. Joint Conf. on Computer Vision and Computer Graphics Theory and Applications, pp. 704–709 (2011)
3. Chen, H., Wu, L.: A New Measure of Forecast Accuracy. In: Intl. Conf. on Information and Financial Engineering, pp. 710–712 (2010)
4. Gallup, D., Frahm, J., Mordohai, P., Pollefeys, M.: Variable Baseline/Resolution Stereo. In: Proc. Computer Vision and Pattern Recognition, pp. 1–8 (2008)
5. Hirschmuller, H., Scharstein, D.: Evaluation of Stereo Matching Costs on Images with Radiometric Differences. IEEE Trans. Pattern Analysis and Machine Intelligence, 1582–1599 (2009)
6. Isgro, F., Trucco, E., Xu, L.: Towards Teleconferencing by View Synthesis and Large-Baseline Stereo. In: Proc. Conf. on Image Processing, pp. 198–203 (2001)
7. Kostliva, J., Cech, J., Sara, R.: Feasibility Boundary in Dense and Semi-Dense Stereo Matching. In: Conf. on Comp. Vision and Pattern Recognition, pp. 1–8 (2007)
8. Malpica, W., Bovick, A.: Range Image Quality Assessment by Structural Similarity. In: IEEE Conf. on Acoustics, Speech and Signal Processing, pp. 1149–1152 (2009)
9. Scharstein, D.: Middlebury Stereo Evaluation, `http://vision.middlebury.edu/stereo/`
10. Scharstein, D., Szeliski, R.: A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms. Intl. Journal of Computer Vision 47, 7–42 (2002)
11. Scharstein, D., Szeliski, R.: High-accuracy Stereo Depth Maps using Structured Light. In: Computer Vision and Pattern Recognition, pp. 195–202 (2003)
12. Schreer, O., Fehn, C., Atzpadin, N., Muller, M., Smolic, A., Tanger, R., Kauff, P.: A Flexible 3D TV System for Different Multi-Baseline geometries. In: Proc. Conf. on Multimedia and Expo, pp. 1877–1880 (2006)
13. Shen, Y., Chaohui, L., Xu P., Xu, L.: Objective Quality Assessment of Noised Stereoscopic Image. In: Proc. Third Intl Conf. on Measuring Technology and Mechatronics Automation, pp. 745–747 (2011)
14. Szeliski, R.: Prediction Error as a Quality Metric for Motion and Stereo. In: Proc. Intl. Conf. on Computer Vision, vol. 2, pp. 781–788 (1999)

15. Szeliski, R., Zabih, R.: An Experimental Comparison of Stereo Algorithms. In: Triggs, B., Zisserman, A., Szeliski, R. (eds.) ICCV-WS 1999. LNCS, vol. 1883, pp. 1–19. Springer, Heidelberg (2000)
16. Van der Mark, W., Gavrila, D.: Real-time Dense Stereo for Intelligent Vehicles. IEEE Trans. on Intelligent Transportation Systems 7(1), 38–50 (2006)
17. Van Veldhuizen, D., Zydallis, D., Lamont, G.: Considerations in Engineering Parallel Multiobjective Evolutionary Algorithms. IEEE Trans. Evolutionary Computation 7(2), 144–173 (2003)
18. Wang, D., Ding, W., Man, Y., Cui, L.: A Joint Image Quality Assessment Method Based on Global Phase Coherence and Structural Similarity. In: Proc. Intl. Congress on Image and Signal Processing, pp. 2307–2311 (2010)
19. Wang, Z., Bovik, A., Sheikn, H., Simocell, E.: Image Quality Assessment: From Error visibility to Structural Similarity. IEEE Trans. on Image Processing 13(4), 600–612 (2004)
20. Zhang, Z., Hou, C., Shen, L., Yang, J.: An Objective Evaluation for Disparity map Based on the Disparity Gradient and Disparity Acceleration. In: Proc. Intl. Conf. on Information Technology and Computer Science, pp. 452–455 (2009)