Dense Stereomatching Algorithm Performance for View Prediction and Structure Reconstruction

Jana Kostková, Jan Čech, and Radim Šára

Center for Machine Perception
Czech Technical University, Prague, Czech Republic
{kostkova,cechj,sara}@cmp.felk.cvut.cz
http://cmp.felk.cvut.cz

Abstract. The knowledge of stereo matching algorithm properties and behaviour under varying conditions is crucial for the selection of a proper method for the desired application. In this paper we study the behaviour of four representative matching algorithms under varying signal-to-noise ratio in six types of error statistics. The errors are focused on basic matching failure mechanisms and their definition observes the principles of independence, symmetry and completeness. A ground truth experiment shows that the best choice for view prediction is the Graph Cuts algorithm and for structure reconstruction it is the Confidently Stable Matching.

1 Introduction

Dense stereoscopic matching can be used for two broad classes of applications: *view prediction*, where the task is to generate previously unseen images and *structure reconstruction*, where the task is to recover 3D scene structure. Both tasks impose vastly different requirements on the matching: the view prediction requires no *false negatives*, the structure reconstruction no *false positives and mismatches*. Both classes must typically work under a *wide range of conditions:* namely varying texture contrast and varying level of geometric detail in the scene.

There exist two main classes of matching algorithm evaluation methods. Class 1 methods are based on image prediction: Matches are evaluated for their image self-consistency [8] or validated in independent images [12]. The disadvantage of this approach is that good image prediction does not guarantee good disparity maps [5]. The class 2 methods are based on ground-truth obtained from independent measurement [9], ground control points [4], digital elevation model [3], or semi-manually using a strong prior model, e.g. piecewise planarity.

There have been several medium to large-scale systematic stereo algorithm evaluations. The largest are the JISCT and the Middlebury studies. The JISCT [1] used a large set of stereo images of real complex scenes and evaluated three types of error: false negatives, false positives and mismatches. Ground-truth in a few selected points was provided manually. A recent study by Scharstein et al. [11] used four scenes as the test set. Ground truth was provided semi-manually. Mismatches in texture-less, non-occluded and depth-discontinuity regions are counted. The contribution of the Scharstein's study is that it considers *multiple evaluation criteria*. Their idea to measure mismatches in

J. Bigun and T. Gustavsson (Eds.): SCIA 2003, LNCS 2749, pp. 101–107, 2003.

texture-less regions was the first step to understanding the behaviour of algorithms under varying signal-to-noise ratio. Unfortunately, the selected performance criteria are still too aggregate and so unable to discriminate among the *mechanisms* of matching failure. The test scenes are not designed to any specific performance evaluation goal. Moreover, half-occluded regions are excluded from the evaluation altogether.

Previous experimental evaluations did not pay much attention to distinguishing among the algorithms, based on their purpose. Moreover, their performance under varying conditions has not been studied. The goal of this work is to fill this gap. In this paper we will study strong-model and weak-model matching algorithms. The former typically use a strong continuity prior, the latter use weak constraints like ordering.

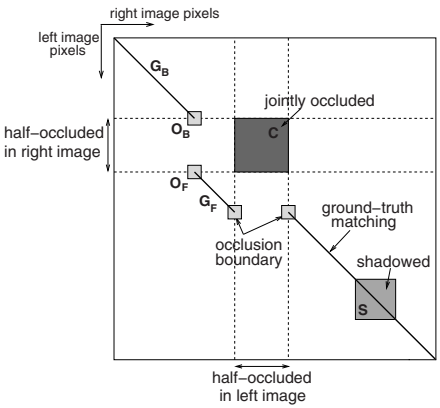
To this end, several new, highly focused performance measures are proposed in Sec. 2. In Sec. 3 we describe our ground-truth experimental setup. The performance of the selected algorithms is evaluated under varying texture quality in Sec. 4. We limited our study to binocular matching. Sec. 5 then concludes the paper. The full version of this paper has been published as a technical report [7]. The report gives all details on experimental setup, matching error definition, ground-truth properties, algorithm parameter choice, and the evaluation procedure.

2 Performance measures

In defining our errors, we will observe the following four principles, which we believe are essential for the complex algorithm's evaluation:

Orthogonality to make the errors mutually independent;
Symmetry to make the errors invariant to the reference image selection;
Completeness to make the errors valid in scenes of an arbitrary geometry;
Algorithm independence to allow for disparity maps that are sparse or not one-to-one.

Traditionally, experimenters measure false positive, false negative and mismatch errors directly in a disparity map. But this error definition is too closely related to a reference image selection and thus violates the principles.



We define the errors by the means of various events in the matching table according to the figure shown left. The table, i.e. the set of all possible matches P (on one rectified image row), is covered by five regions: the ground-truth $G = G_B \cup G_F$ (shown as a solid line), the jointly occluded region C (dark gray), the occlusion boundary neighborhood $O = O_B \cup O_F$ (light gray), the texture-less region S (gray), and the complement $T = P \setminus G \setminus C \setminus O \setminus S$ (white). In G and O we distinguish their background $(\cdot)_B$ and foreground $(\cdot)_F$ parts. For the sake of error orthogonality we define the set of *false negatives* (FN) as unmatched ground-truth correspondences which are not induced by a mismatch.

Let M, $M \subset P$ be a matching obtained from a tested algorithm. The following errors will be considered here (more errors are defined in [7]):

Primary errors *Mismatch Rate (MIR)* measures the accuracy of matching M, *False Positive Rate (FPR)* the inability to detect jointly occluded regions, *False Negative Rate (FNR)* the disparity map sparsity, *Failure Rate (FR)* the overall image prediction error:

$$MIR = \frac{|M \cap T|}{|M \setminus C \setminus O \setminus S|}, \qquad FPR = \frac{|M \cap C|}{|M \setminus O \setminus S|},$$
$$FNR = \frac{|FN \setminus O \setminus S|}{|G \setminus O \setminus S|}, \qquad FR = \frac{|M \cap (T \cup C)| + |FN \setminus O \setminus S|}{D(P)},$$

where $D(P)$ is the length of the main diagonal in matching table P. Note that the MIR and FPR errors are normalized by matching *size*.

Secondary errors *Occlusion Boundary Inaccuracy (OBI)* measures the precision of occlusion boundary detection and *Bias (B)* measures the difference between unmatched pixels in the background object and the foreground object:

$$OBI = \frac{|(G \setminus M) \cap O|}{D(O)}, \qquad B = \frac{|G_B \setminus M \setminus S \setminus O|}{|G_B \setminus S \setminus O|} - \frac{|G_F \setminus M \setminus O|}{|G_F \setminus O|},$$

where $D(O)$ is the diagonal length of region O. Zero B implies unbiased matching, positive B means that foreground is matched more often than the background, and negative B means the opposite.

3 Experimental setup

The test scene consists of five long thin stripes (the foreground) in front of a flat panel (the background), see Fig. 1. The scene was set up to preserve ordering. A random texture is projected onto the scene using a projector.

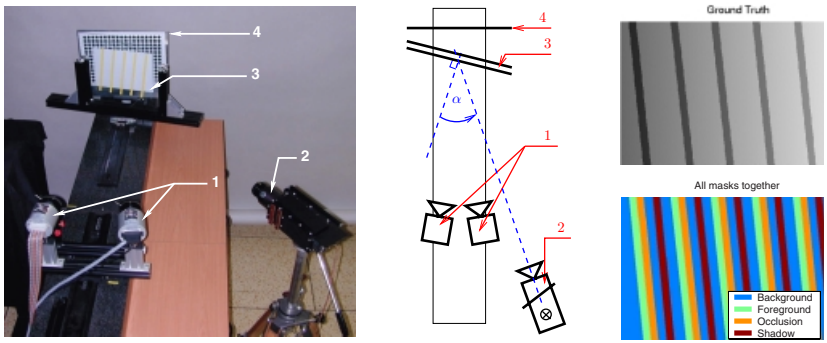


Fig. 1. The experimental setup: 1–cameras, 2–texture projector, 3–test scene, 4–calibration target (left and center). Ground truth shown as disparity map and the corresponding masks (right).

Cameras were calibrated including radial distortion and the images were rectified. The ground truth was generated in two steps: first, the stripes and shadows were seg-

mented from the background (with the help of special illumination) and second, a disparity plane was fit to segmented regions using disparity map obtained from a selected matching algorithm. The result of this procedure is a composite disparity map and a set of masks (both shown in Fig. 1).

The test scene was captured under 20 different exposures corresponding to 20 different contrast values, which emulates varying signal-to-noise ratio. The left rectified images of contrast 1, 8, 14 (the T-level, see later), and 20 (the maximum level) are shown in Fig. 3.

4 Results

We have chosen four area-based algorithms employing strong and weak continuity priors: The former represented by MAP matching via graph cuts [6] (GC) and ML matching via dynamic programming [2] (DP), and the later by Confidently Stable Matching [10] (CSM) and Winner-Take-All [11] (WTA) algorithms, respectively.

All tested algorithms have several parameters: (1) those common for all the algorithms are fixed; (2) the fundamental for each particular algorithm are tuned to optimal performance; and (3) the auxiliary are set to default values. The setting is kept constant for all contrast levels. In the DP and GC the parameters are set to minimize the FR error under the middle texture contrast level (the *T-level*). The parameters of CSM represent the matching confidence level and thus directly determine the MIR over a very wide range [10] (low MIR implies high FNR and vice versa). The parameters of CSM are set so that the MIR becomes equal to that of the GC under the maximum contrast. The WTA has no parameters. A detailed description of parameter tuning is found in [7].

The results of the tested algorithms are shown as plots in Fig. 2 and as disparity maps in Fig. 3. The respective error rates are plotted against texture contrast in all plots. We will now discuss the individual algorithm behaviour in all errors in turn:

Mismatch Rate The GC gives very poor results below the T-level contrast as confirmed in Fig. 3. The MIR improves about $100\times$ in the maximum contrast. The difference between the DP and the GC above the T-level is explained as follows: The DP minimizes the sum of the SSD error and the λ-multiples of disparity jumps along the epipolar lines. Higher contrast increases the SSD in all potential matches which decreases the influence of λ. The MIR thus levels off on the value of texture self-similarity (which is constant). The GC does not show this behaviour because the isotropic continuity prior is stronger than the directional one used in DP. The CSM reaches a low level very fast and then stays constant. The WTA gives the MIR consistently bad.

False Negative Rate In the GC and DP the FNR generally grows with increasing contrast. In the GC the FNR is induced by disparity jumps: as the disparity resolution improves the frequency of jumps increases faster than their decreasing magnitude, see Fig. 3. In the DP the increasing FNR is due to the relative decrease of λ with increasing contrast: Skipping a match near a disparity jump is cheaper than including it with a non-zero SSD. The CSM has a high and constant FNR over all contrasts. The WTA has zero FNR because it does not model occlusion.

False Positive Rate The GC and DP both have a large FPR below the T-level, since the strong continuity prior results in constant disparity solution passing through

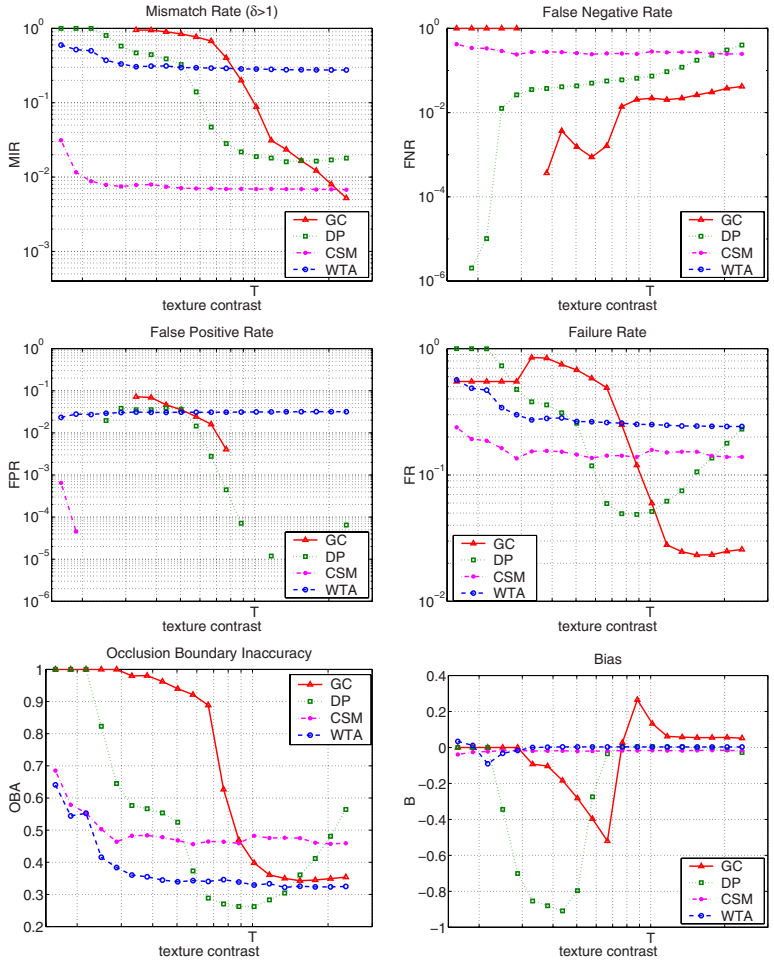


Fig. 2. Matching results for the tested algorithms.

the C region, cf. Fig. 3. Above the T-level the FPR is zero in both the GC and DP. In the CSM the FPR is zero except for the lowest two contrasts. As the WTA has no occlusion model its FPR is large.

Failure Rate The GC shows the best results above the T-level. In the DP the FR increases above the T-level. The CSM shows stable results across the whole contrast range, it is by about five times worse than GC under the maximum contrast. The performance of WTA is similar to that of the CSM in the FR.

Occlusion Boundary Inaccuracy In the GC the OBI is low and constant above the T-level and very high below it as confirmed in Fig. 3. In the DP the T-level performance is about $1.5\times$ better than in the GC. This is due to the difference in strength of the isotropic and directional priors. In the DP the OBI grows above the T-level for the same reasons as the FNR. The CSM has almost constant OBI above the level of 4. The

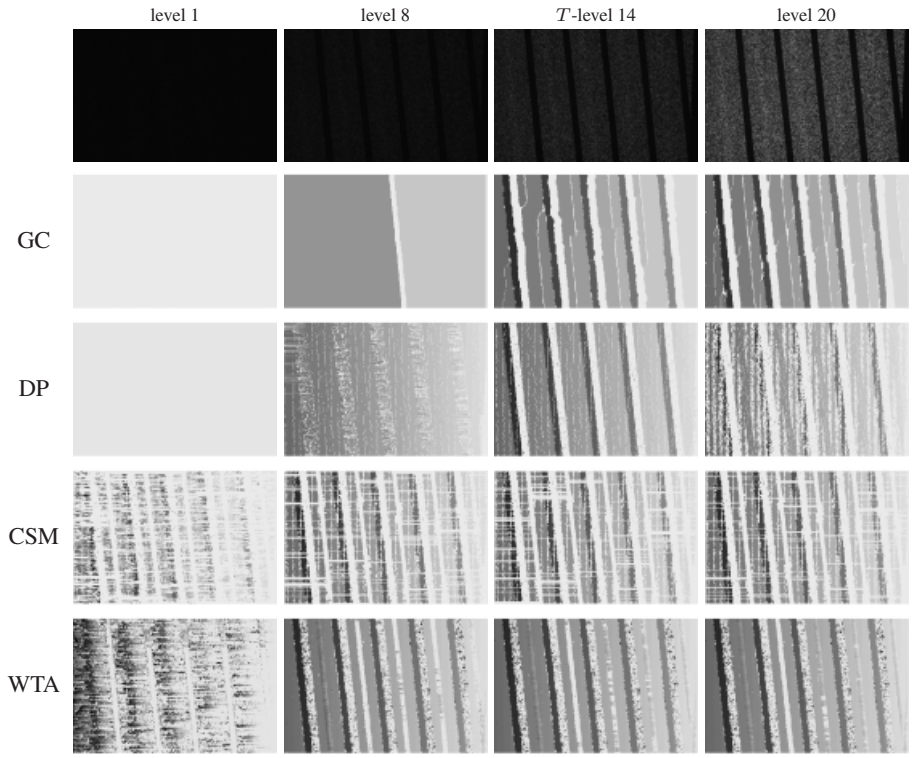


Fig. 3. Disparity maps by the tested algorithms under four texture contrasts (1st row).

WTA has the lowest *OBI* because of its ability to find one-to-two matches at disparity jumps.

Bias The GC and DP are unbiased above the T-level. The breakdown of the GC and DP below the T-level is clearly seen in Fig. 3: The strong continuity prior results in the background disparity interpolated over the foreground stripes causing the negative bias. The positive bias in the GC around the T-level is due to fewer disparity jumps in the stripes than in the background. The CSM and WTA are unbiased over all the contrasts.

5 Discussion and conclusions

From our experimental analysis we conclude that the best choice for view prediction is the GC algorithm and for structure reconstruction it is the CSM algorithm:

The GC has the best overall failure rate mainly because of good disparity map density (low false negative rate) and low mismatch rate that continuously improves with increasing texture contrast. The GC has the ability to detect half-occlusions as long as they have a good contrast. The GC parameters must be tuned to the worst-contrast texture in the scene, since the method fails in low contrasts. In other words it is prone

to *illusions*: its inability to reject unreliable data would be a serious drawback if the method was used in structure reconstruction.

The CSM results are sparser than in the other methods (especially under low contrasts) but they are accurate independently on texture contrast. The low mismatch rate, zero false positive rate, and the unbiasedness make the CSM suitable for structure reconstruction in complex scenes of varying texture contrast. The high false negative rate renders this method unsuitable for view prediction, however.

The WTA always produces very erroneous results, and therefore is not suitable for both the structure reconstruction and the view prediction except when the speed is a strong concern. The main disadvantage of DP is its high sensitivity to parameter setting, since the performance decreases both below and above the optimal contrast level.

Acknowledgement

This work has been supported by the Grant Agency of the Czech Republic under project GACR 102/01/1371, by the Grant Agency of the Czech Tecnical University under project CTU 8306413, and by the Czech Ministry of Education under project MSM 212300013.

References

1. R. C. Bolles, H. H. Baker, and M. J. Hannah. The JISCT stereo evaluation. In *Proc. DARPA Image Understanding Workshop*, pages 263–274, 1993.
2. I. J. Cox, S. L. Higorani, S. B. Rao, and B. M. Maggs. A maximum likelihood stereo algorithm. *Computer Vision and Image Understanding*, 63(3):542–567, 1996.
3. T. Day and J.-P. Muller. Digital elevation model production by stereo-matching spot image-pairs: a comparison of algorithms. *Image and Vision Computing*, 7(2):95–101, 1989.
4. G. Gimel'farb. Pros and cons of using ground control points to validate stereo and multiview terrain reconstruction. *Evaluation and Validation of Computer Vision Algorithms*, 1998.
5. G. Gimel'farb and H. Li. Probabilistic regularisation in symmetric dynamic programming stereo. In *Proc. Image and Vision Computing New Zealand 2000*, pages 144–149, 2000.
6. V. Kolmogorov and R. Zabih. Computing visual correspondence with occlusions using graph cuts. In *Proc. International Conf. on Computer Vision*, 2001.
7. J. Kostková, J. Čech, and R. Šára. The CMP evaluation of stereo algorithms. Tech. Report CTU–CMP–2003–01, Center for Machine Perception, Czech Technical University, 2003.
8. Y. G. Leclerc, Q.-T. Luong, and P. Fua. Measuring the self-consistency of stereo algorithms. In *Proc. European Conf. on Computer Vision*, volume 2, pages 282–298, 2000.
9. J. Mulligan, V. Isler, and K. Daniilidis. Performance evaluation of stereo for tele-presence. In *Proc. of International Conf. on Computer Vision*, 2001.
10. R. Šára. Finding the largest unambiguous component of stereo matching. In *Proc. European Conf. on Computer Vision*, volume 3, pages 900–914, 2002.
11. D. Scharstein, R. Szeliski, and R. Zabih. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 47(1):7–42, 2002. url: www.middlebury.edu/stereo.
12. R. Szeliski. Prediction error as a quality metric for motion and stereo. In *Proc. International Conf. on Computer Vision*, volume 2, pages 781–788, 1999.