Testing Stereo Data – Not the Matching Algorithms

Ralf Haeusler and Reinhard Klette

The *.enpeda..* Project
The University of Auckland

Abstract. Current research in stereo image analysis focuses on improving matching algorithms in terms of accuracy, computational costs, and robustness towards real-time applicability for complex image data and 3D scenes. Interestingly, performance testing takes place for a huge number of algorithms, but, typically, on very small sets of image data only. Even worse, there is little reasoning whether data as commonly applied is actually suitable to prove robustness or even correctness of a particular algorithm. We argue for the need of testing stereo algorithms on a much broader variety of image data then done so far by proposing a simple measure for putting image stereo data of different quality into relation to each other. Potential applications include purpose-directed decisions for the selection of image stereo data for testing the applicability of matching techniques under particular situations, or for realtime estimation of stereo performance (without any need for providing ground truth) in cases where techniques should be selected depending on the given situation.

1 Introduction

Performance evaluation of stereo algorithms became increasingly popular since the availability of various test sites such as [18] at Middlebury University. Such evaluations were speeding up progress in the design of stereo matching algorithms. Ranking is typically done by comparing a few error measures, calculated with respect to given ground truth and a relatively small number of images. Evaluations lead to particular insights, for example about the role of used cost functions [8], or of image preprocessing methods.

Necessity and limitation of such evaluations have been extensively discussed in the literature. Issues often treated are missing ground truth for real-world scenes [5] and a lack in theoretical understanding that prevents from making intelligent predictions of stereo performance on yet unseen imagery [19].

Stereo image data, depending on recorded scenes, sensor quality and so forth, can be of very different characteristics and origin (e.g., synthetic, controlled indoor, real-world outdoor). The question arises: Given a stereo image pair, what is the minimum error we may expect? This question should be answered for a wide range of different types of stereo image data, ultimately allowing to quantify this material in terms of quality. However, for the most interesting scenarios

– outdoor real-world, highly dynamic and complex scenes with potentially very poor image quality – the common evaluation approach of stereo matching techniques is not feasible due to the lack of ground truth. Previous work [11, 14] that does not require ground truth needs at least three time-synchronous views of a scene. We propose an alternative approach that only needs binocular imagery.

The objective of the paper is to demonstrate that it might be possible to quantify the quality of recorded stereo images with respect to some measures. We also suggest that those measures may be used to indicate domains of relevant imaging scenarios when performing evaluations for some particular test data.

The proposed approach is based on Lowe's SIFT-descriptor [12], which in general outperforms other descriptors in terms of discriminative power [13]. SIFT-matching supports the definition of similarity measures that allow us to derive spatial relations between (e.g.) millions of images [15].

Such results suggest that SIFT-matching can be used to define a measure for establishing some relationships between different sets of stereo image data. There is space for more advanced proposals in future, but a simple SIFT-based measure of matching counts (Figure 2 illustrates SIFT matches in four stereo pairs) is sufficient to initiate a discussion about this type of data evaluation.

To human viewers, it is immediately clear, whether a stereo pair is of good quality for extracting depth information. For example, stereo photos taken under insufficient lighting conditions (such as outdoors during the night), very high contrast images with poor texturing or stereo pairs with contrast differing with a factor of more than two between left and right image cannot be matched properly by the human visual system. Similarly, semi-occluded objects or vertical parallax lead to retinal rivalry and therefore to strong eyestrain.

The construction of SIFT descriptors itself is inspired by the functioning of primate V1 cortical neurons. Such biological models have been successfully applied to the task object recognition [2]. Hence, there is another intuitive justification for using these in assessing the quality of stereo data in terms of retrievable depth information.

We envision four major benefits of assessing stereo image data independently from geometric ground truth. First, it can guide the selection of applied methods as already mentioned above. Second, it may make processing of real-world stereo images more tractable by providing an additional measure of confidence. Third, we can identify "problematic" situations in real-time; this gives a chance to identify unexpected problems when doing an on-line stereo analysis of real-world stereo image sequences, and to be aware of those when further improving stereo matching. Fourth, it may advance theoretical knowledge about stereo matching by implementing performance evaluation on sophisticated synthetic scenes (i.e., using progress in physics-based rendering) and showing its conclusiveness regarding relevance to real-world scenarios.

The paper is structured as follows. Section 2 introduces two measures based on SIFT matching counts. Section 3 provides details of data used in this study and presents results from experiments to point out the feasibility of the approach. Section 4 explains potential applications in more detail and concludes.

2 Our Method

SIFT features are defined by extrema in a difference-of-Gaussians scale-space, also applying sub-pixel accuracy and rejection of poorly defined locations. Together with a well-constructed descriptor, these are called *distinctive features*.

Our hypothesis is that sparse matching of distinctive features provides measures strongly correlated to the outcome of a dense stereo matching process. The chosen implementation [20] together with the method outlined below seems to be sufficient to illustrate this correlation according to our experiments.

For a rectified stereo pair and known ground truth, a match between a feature location (i_l, j_l) in the left image and a feature location (i_r, j_r) in the right image is *correct up to known constraints* if

$$(i_r - \varepsilon_i < i_l < i_r + \varepsilon_i) \wedge (j_l + d_{ij} - \varepsilon_j \leq j_r \leq j_l + d_{ij} + \varepsilon_j) \tag{1}$$

for small $\varepsilon_i, \varepsilon_j > 0$ with known disparity values d_{ij} (i.e., the ground truth). If ground truth is not available, then we evaluate by testing for

$$(i_r - \varepsilon_i < i_l < i_r + \varepsilon_i) \wedge (j_l \leq j_r \leq j_l + d_{max}) \ , \tag{2}$$

where d_{max} is the maximum disparity between both stereo views. We choose $\varepsilon_i = \varepsilon_j = 1$.

Equation (2) appears to be very much "forgiving". However, note that in this case of modeled unavailable ground truth, the probability of the event that "a mismatch is wrongly classified as being correct up to known constraints" has a very small probability of at most $(2\varepsilon_i d_{max} - 1)/(I \cdot J)$ in an image of size $I \times J$. This assumption can be violated, for example, for images with repetitive textures in some areas.

We analyze the counts of correct matches (up to known constraints) and the ratio between detected and matched features for given stereo pairs. Assume that the feature detector identifies n features in the base image that lead to m matches in the match image, and that from those m matches, o are classified as being incorrect. In this case, we define that

$$x = n/m \qquad \text{and} \qquad y = 100 \times o/m \ , \tag{3}$$

where x is the *matching rate* and y the *mismatch rate*. Thus, $x \geq 1$ and $y \leq 100\%$.

The matching rate x expresses how many features on average lead to one match (no matter whether correct or not), while the mismatch rate y identifies the percentage of incorrect matches.

3 Experiments

Our experiments are designed to demonstrate that the information provided by a selective stereo matching process of distinctive features may be suitable to label stereo image data with an expected quality of disparity calculations, without requiring any ground truth except the value of d_{max}.

Comparison of SIFT-matches on various data sets. In particular, we compare recorded stereo pairs, both "engineered" and real-world, to synthetic pairs, and we attempt to modify the synthetic stereo pairs in a way such that they quantify similar to the recorded pairs for the proposed measures. We compare values of our measures with values of the normalized cross correlation (NCC) derived from prediction errors following [14], on stereo sequences showing "problematic" situations. We use the following stereo image data:

(1) Synthetic data:
   - `EISATS 2, Sequence 1`, see [3]: a sequence of 100 frames with low object complexity showing views from a simulated moving vehicle.
   - `EISATS 2, Sequence 2`, same source: a more complex sequence of 300 frames, containing vegetation modelled with L-Systems.
   - Synthetic stereo data of high complexity, rendered by the authors with physically correct simulation of the light distribution, using path-tracing [17]. Different image sensor distortion effects are applied to study their effect, including blooming and chromatic aberration.

(2) Engineered test images (i.e., photos taken under controlled lighting):
   - `Middlebury 2001 and 2003`, see [18], in particular the stereo sets named `Tsukuba`, `Venus`, `Cones`, and `Teddy`.
   - `Middlebury 2006`, see [8], a more extensive stereo test set, containing 21 images; each image is available for three different illuminations and three different exposures (normal, two f-stops under- and overexposed).

(3) Real-world sequences (150 - 200 stereo frames per sequence) of public road scenarios captured with industrial b/w cameras from a moving vehicle:
   - `EISATS 1, "Construction site"`, see [3]: a 10-bit stereo recording.
   - `NZ Road 1-3`, traffic scenes on New Zealand roads, 8-bit trinocular recordings as made available by [14]; these sequences support an error estimation without ground truth based on calculating the third view.

Figure 3 illustrates feature matching relationships between the listed stereo image data when always applying the same SIFT matcher and measures x and y as defined in (3). The test set Middlebury 2001 evaluates similar to synthetic images of medium complexity, but is significantly different from real world scenes captured with industrial cameras. The quantization resolution (8-, 10- or 12-bit) is of minor relevance (see Fig. 3), optimal exposure and contrast provided.

We also see that the more extensive (and somehow "closer" to uncontrolled scenarios) dataset Middlebury 2006, which is not yet widely used for testing, spans a much wider region in our xy-space. However, our xy-space still shows a clear separation of this dataset from real-world outdoor scenes.

The attempt to synthesize stereo image data using physics-based rendering, also including physics-based imaging distortions, leads to distributions of xy-values which are very close to those of uncontrolled image data. Interestingly, applying further distortions does not produce the results we might expect: Chromatic aberration increases both, matching rate and mismatch rate. Adding sensor blooming slightly increases the mismatch rate but improves matching.

This indicates to us that either our xy-space is somehow incomplete, or the applied model for the simulation of blooming and chromatic aberration is still "too simple".

Comparison with results based on third-view analysis. This subsection discusses how our simple SIFT-based evaluation relates to a particular kind of "ground truth-based evaluation" when testing stereo matching techniques on real-world data. In fact, providing theoretical evidence for this relationship is rather difficult, as there seems to be no common underlying model. However, there are some strong statistical dependencies between our SIFT-based evaluation on two views, and the third-eye approach for stereo algorithm evaluation as proposed in [14]. For illustrating those, we use real-world stereo sequences as provided in Set 5 of [3].

We process each stereo sequence with five different stereo matching algorithms, namely belief propagation (BP) [4], semi-global matching (SGM) [7] using either the Birchfield-Tomasi (BT) or a mutual information (MI) cost function, graph cut (GC) [10], or dynamic programming (DP) [16].[1] Each of those algorithms run either on the original stereo sequence, a Sobel operator preprocessed stereo sequence, or on residual images [1] with respect to 40-times repeated 3×3 mean filtering (or one run of a comparable large smoothing kernel). In uncontrolled image data, suitable preprocessing often has a dramatic effect on the quality of stereo results. We compare altogether 15 different matching results, used for generating a virtual third view, compared by normalized cross correlation with the recorded third view. For a simple comparison to the proposed SIFT test, we use the mean $(x + y)/2$ of matching and mismatch rate. The distribution of observed values in Fig. 2 suggests that this even more simplified measure is sufficiently discriminative. See Fig. 3 and Fig. 4 for results on real-world stereo sequences with 150 stereo frames. For this visual comparison, all these error measures are normalized as follows: if T is the number of frames and $\mathrm{NCC}(t)$ is the error measure for a particular frame t with $1 \leq t \leq T$, we display $(\mathrm{NCC}(t) - \mu_T)/\sigma_T$, where $\mu_T = 1/T \sum_{t=1}^{T} \mathrm{NCC}(t)$ and $\sigma_T^2 = 1/(T-1) \sum_{t=1}^{T} (\mathrm{NCC}(t) - \mu_T)^2$. The same normalization is applied to the results of SIFT matching.

Statistical relation between error measures. Normalized cross correlation was used again to examine the relationship between error measures of all stereo matching algorithms and between stereo and SIFT-matching counts, for long sequences as illustrated by the previous two figures. The correlation coefficients and p-values were computed. Table 1 summarizes the correlation between those stereo algorithms. Due to space limitations, we only display results for the mean of NCC values for the three different preprocessing options of the two sequences already illustrated in Figs. 3 and 4.

[1] Sources of used matching programs are as acknowledged in [14].

Table 1 indicates a moderate correlation between errors of the SIFT-measure and all stereo algorithms except DP. Strongest correlations are mostly found between global algorithms, but all measures in the highway sequence are significantly correlated ($p < 0{\cdot}001$) except the combinations SIFT – DP ($p = 0{\cdot}59$), and SGM(MI) – DP ($p = 0{\cdot}17$). In the night-time sequence, all measures are significantly correlated($p < 0{\cdot}001$). Reasons for outliers in particular frames leading to weaker correlation between BP and SIFT based measures are as follows:

In the highway sequence, the most obvious deviation is in Frames 39 and 40. In these images, a large area (a big truck on the highway) is coming close (less than ten times the baseline) to the camera, resulting in semi-occluded areas at the image border. Many stereo algorithms do not cope with this situation. However, for the method described above this simply results in no matches being found in this area, thus no mismatches can occur. Frames 120 to 150 are subject to major brightness differences, where belief propagation stereo performs poorly.

For the night-time sequence, significant deviations occur in Frames 1 to 16, and 50 to 60. Outliers in Frames 111, 128 and 136 are caused by time-synchronization problems. Frames 50 to 60 are big objects coming closer and becoming increasingly semi-occluded by the image border. Of interest are Frames 1 to 16, where strong blooming (caused by strong light sources nearby) is present. This is not very well detected by counting matches.

We see that for many artifacts in uncontrolled image data there is no correlation between matching statistics and stereo performance. These need to be addressed by different methods.

4 Future work and conclusions

It is certainly of general interest in computer vision to have some evaluation of stereo data at hand, for judging its complexity, or qualitative relation to other sequences of stereo data (also covering the common case that ground truth is not available). This evaluation is of interest for the following:

Table 1. Pearson correlation between error measures.

Algorithm	Sequence	BP	SGM(BT)	DP	GC	SGM(MI)	SIFT
Belief propagation	Highway	1	0·95	0·30	0·81	0·69	0·63
	Night	1	0·97	0·85	0·97	0·96	0·57
Semi-global matching (BT)	Highway		1	0·35	0·88	0·60	0·64
	Night		1	0·83	0·94	0·94	0·52
Dynamic programming	Highway			1	0·55	0·11	0·05
	Night			1	0·88	0·82	0·40
Graph cut	Highway				1	0·50	0·56
	Night				1	0·97	0·62
Semi-global matching (MI)	Highway					1	0·43
	Night					1	0·66

Identification of crucial scenarios in large datasets of stereo images: Crucial 3D scenarios, defined by special events in the real world, need to be identified when testing stereo matching in the real world. Such events have to be isolated from a sufficiently diversified database of real world data (e.g., when running a stereo analysis system for days or weeks in real-world traffic). As ground truth is generally not available, our approach helps in identification of these critical datasets.

Realtime check of stereo data in real world applications: In our method, computing feature descriptors and matching depends on the number of detected interest points, which are numerous in highly structured images. Ensuring realtime here requires to limit their number to a fixed upper bound. For SIFT-features such kind of pruning is described in [6]. In its application to image-database retrieval, an insignificant decline in performance was reported even if the number of features is very small. Such realtime checks may be crucial for reliable safety-relevant decisions in, for example, driver assistance systems.

Purposeful design of synthetic sequences for stereo testing: Synthetic data will remain important for testing stereo matching, especially due to having full control about the image formation process. Simulations of interesting situations (rarely appearing in reality, but possible) such as for weather, poor light conditions, or deficiencies in cameras systems, need to come with some evidence of its adequacy for testing stereo vision algorithms.

We have shown that even a simple measure, such as the matching count based on SIFT-features, can provide error measures significantly correlated to a third-view error measure. We pointed out the necessity to benchmark a fairly "huge amount" of stereo image data, and to put those data into qualitative relation to each other.

Future research may aim at more complex measures, allowing to analyze more detailed quality aspects of stereo images. In continuation of the simple count measure as presented here, this could be based on statistics of spatial distributions of matches or mismatches in stereo image pairs. (Note that a simple root-mean square or NCC error value in relation to ground truth does not yet give any information about the spatial distribution of errors.)

Models as presented in [9] may be of very high interest, yet their use is limited due to prohibitive computational costs.

References

1. Aujol, J.F., Gilboa, G., Chan, T., Osher, S.: Structure-texture image decomposition - modeling, algorithms, and parameter selection. *Int. J. Computer Vision*, **67**:111–136 (2006)
2. Edelman, S., Intrator, N., and Poggio, T. 1997. Complex cells and object recognition. http://kybele.psych.cornell.edu/ edelman/Archive/nips97.pdf
3. EISATS: *.enpeda..* image sequence analysis test site. See www.mi.auckland.ac.nz/EISATS
4. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient belief propagation for early vision. *Int. J. Computer Vision*, **70**:41–54 (2006)

5. Förstner, W.: 10 pros and cons against performance characterization of vision algorithms. *Machine Vision Applications*, **9**:215–218 (1997)
6. Foo, J.J., Sinha, R.: Pruning SIFT for scalable near-duplicate image matching. In Proc. *Australasian Database Conf.*, pages 63–71 (2007)
7. Hirschmüller, H.: Stereo processing by semiglobal matching and mutual information. *IEEE Trans. Pattern Analysis Machine Intelligence*, **30**:328–341 (2007)
8. Hirschmüller, H., Scharstein, D.: Evaluation of stereo matching costs on images with radiometric differences. *IEEE Trans. Pattern Analysis Machine Intelligence*, **31**:1582–1599 (2009)
9. Hyvärinen, A., Hurri, J., Hoyer, P.: *Natural Image Statistics*. Springer, Amsterdam (2009)
10. Kolmogorov, V., Zabih, R.: Computing visual correspondence with occlusions via graph cuts. In Proc. *Int. Conf. Computer Vision*, pages 508–515 (2001)
11. Leclerc, Y. G. and Luong, Q.-T. and Fua, P. V.: Self-consistency, Stereo, MDL, and Change detection. In Proc. *IJCV*, (2002)
12. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. *Int. J. Computer Vision*, **60**:91–110 (2004)
13. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. *IEEE Trans. Pattern Analysis Machine Intelligence*, **10**:1615–1630 (2005)
14. Morales, S., Klette, R.: A third eye for performance evaluation in stereo sequence analysis. In Proc. *Computer Analysis Images Patterns*, LNCS 5702, pages 1078–1086 (2009)
15. Nistér, D., Stewénius, H.: Scalable recognition with a vocabulary tree. In Proc. *IEEE Conf. Computer Vision Pattern Recognition*, volume 2, pages 2161–2168 (2006)
16. Ohta, Y., Kanade, T.: Stereo by two-level dynamic programming. In Proc. *Int. Joint Conf. Artificial Intelligence*, pages 1120–1126 (1985)
17. Pharr, M., Humphreys, G.: *Physically Based Rendering: From Theory to Implementation*. Morgan Kaufmann, San Francisco (2004)
18. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. J. Computer Vision*, **47**:7–42 (2002)
19. Thacker, N.A., Clark, A.F., Barron, J.L., Beveridge, J.R., Courtney, P., Crum, W.R., Ramesh, V., Clark, C.: Performance characterization in computer vision: A guide to best practices. *Computer Vision Image Understanding*, **109**:305–334 (2008)
20. Vedaldi, A., Fulkerson, B.: VLFeat: An open and portable library of computer vision algorithms. See www.vlfeat.org (2008)

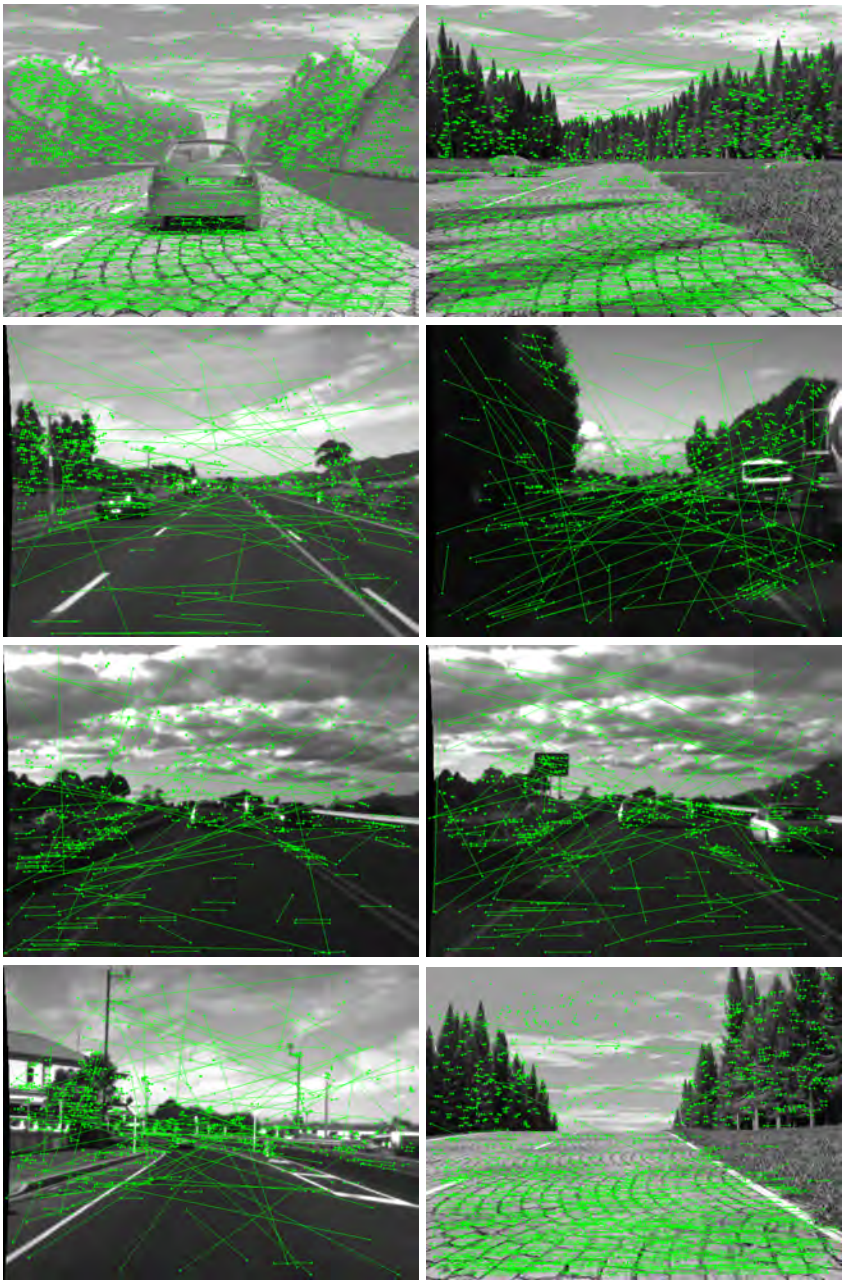

Fig. 1. Illustration of sparse stereo matching with SIFT-features (not constrained by epipolar geometry, but on rectified images) applied to stereo pairs of different characteristics (first three images: synthetic EISATS stereo pairs; rest: real-world scenes of suboptimal quality). Straight connectors of locations of matched features are overlaid to the left image of the used image pair. Synthetic or engineered images generally show a majority of same-row matches.

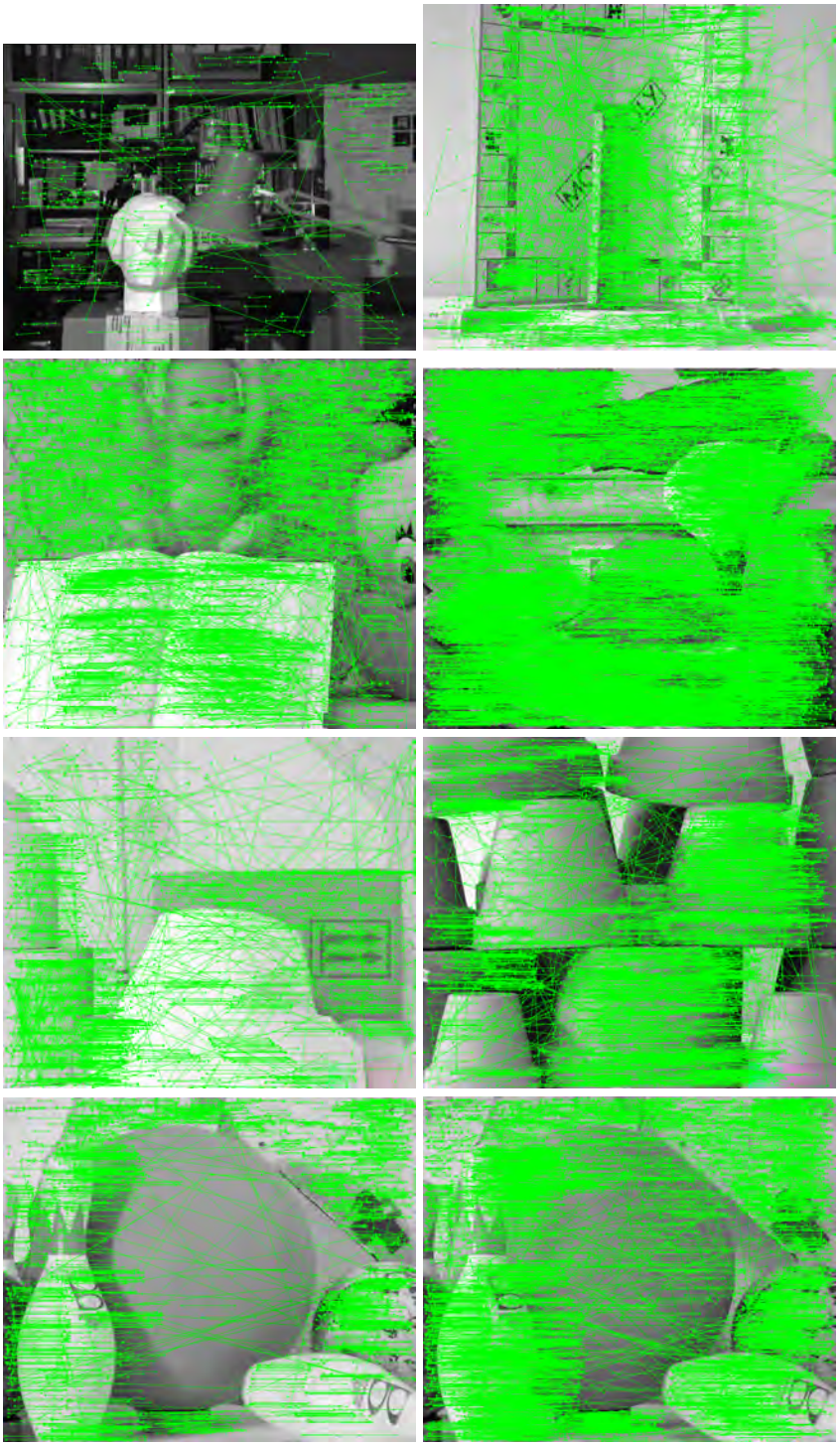

Fig. 2. Same as in previous figure on image data from the Middlebury stereo vision website.

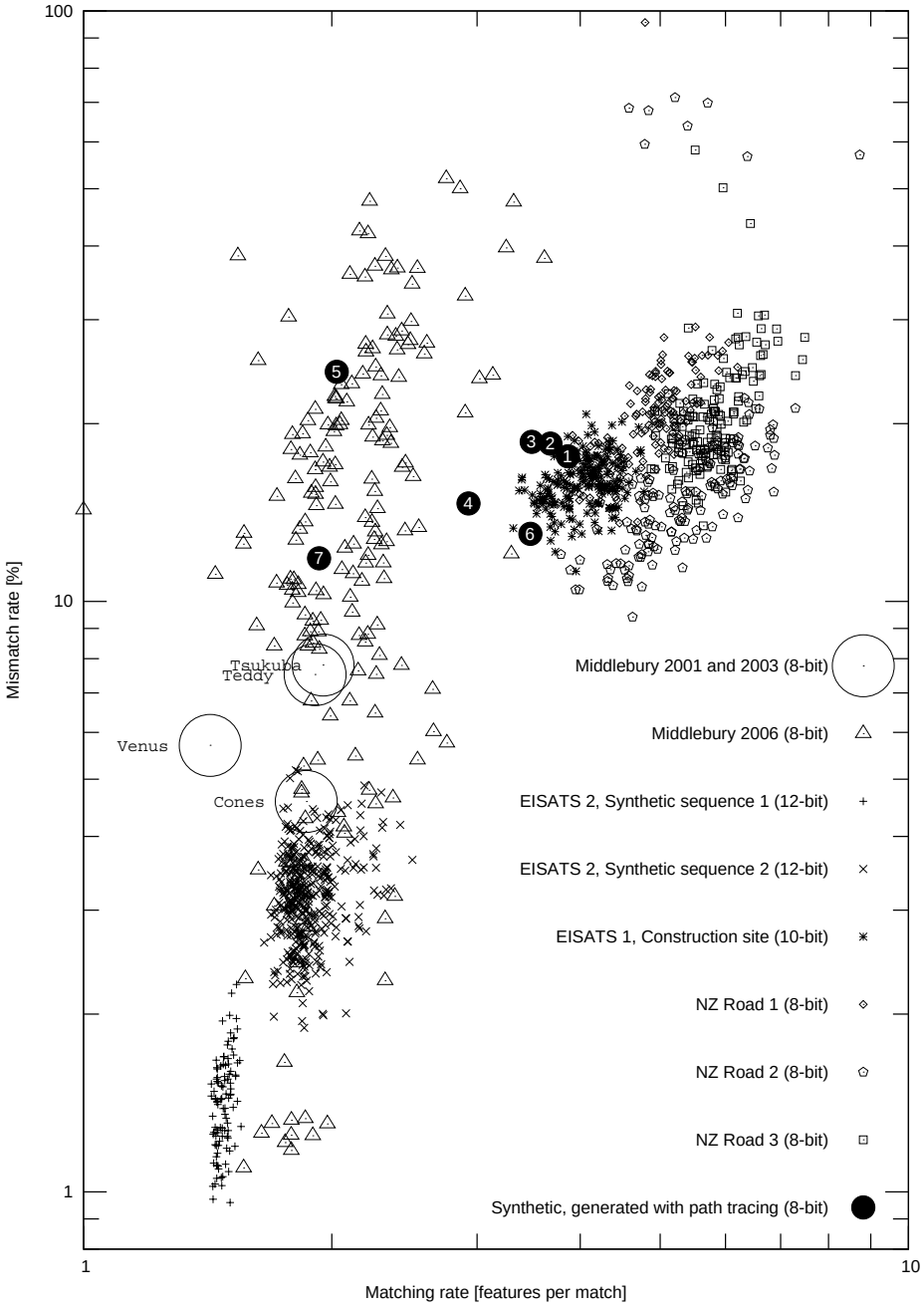

Fig. 3. Mismatch rate y (in percent) and matching rate x in logarithmic scales. Symbols show how stereo data of different origin and quality is discriminated by the proposed measures. Filled black disks for physics-based synthetic data are numbered as follows: 1 (original), 2,3,5 (low, moderate, or severe blooming), 4,6 (moderate or strong chromatic aberration), and 7 (comparison to ordinary raytracing).

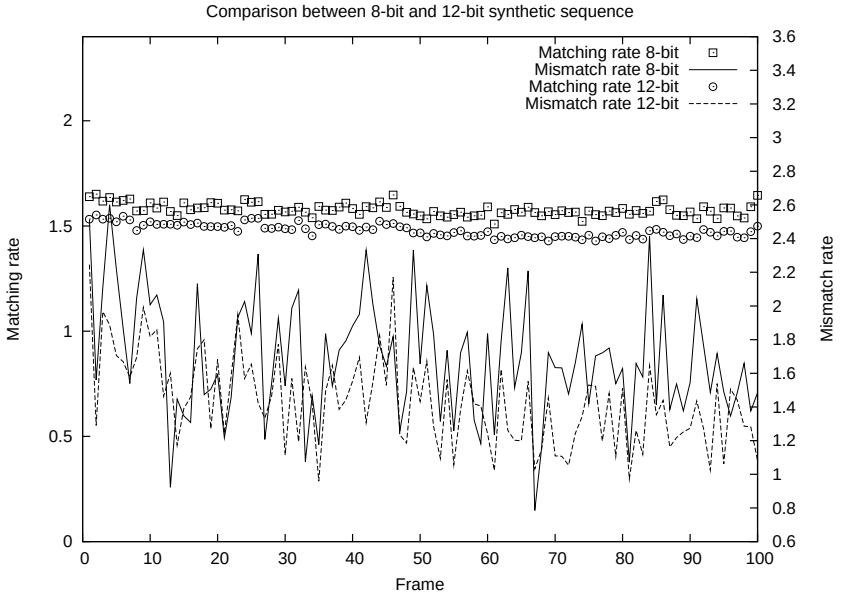

Fig. 4. Matching rate and mismatch rate differ only by about 10% between 8-bit and 12-bit quantization.

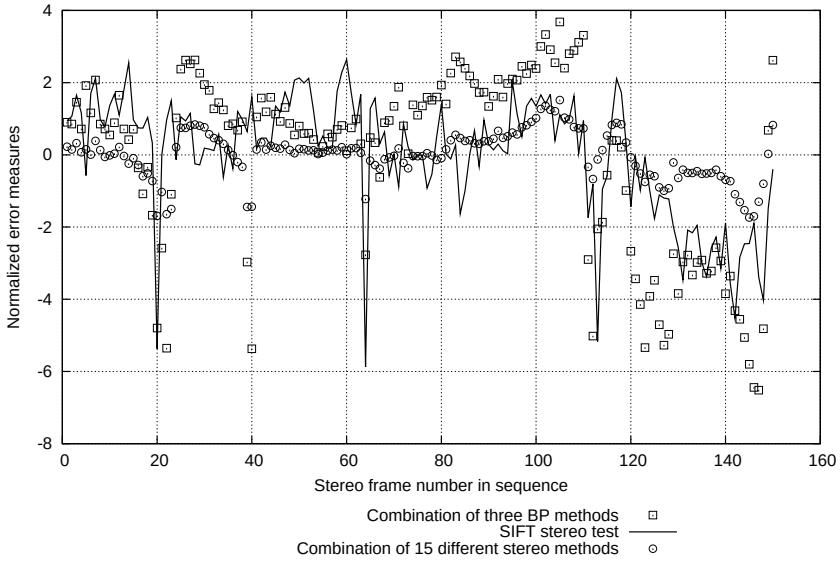

Fig. 5. Normalized error measures for stereo frames 1 to 150 of a day-light highway sequence. We compare results of the proposed SIFT-measure with prediction error based on third-view-synthesis using 15 different stereo matching schemes. For clarity of presentation, only the mean of selected values is displayed.

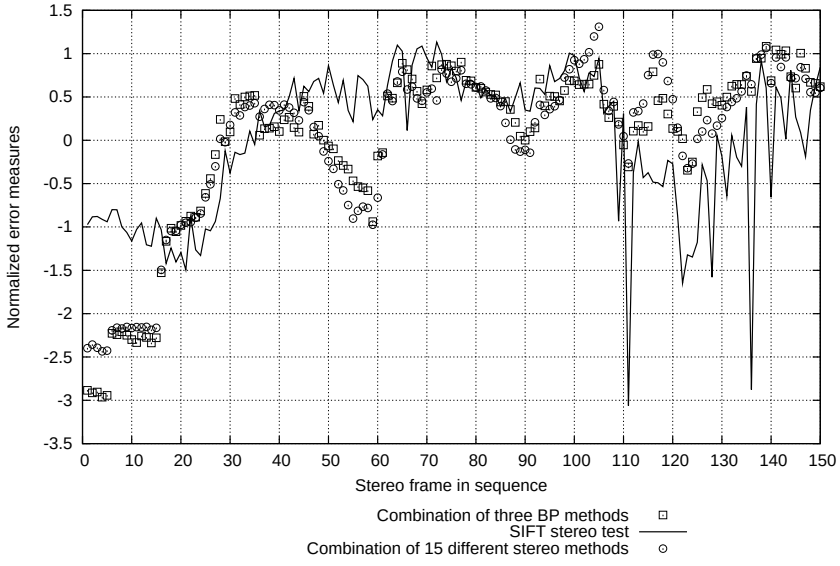

Fig. 6. Same measures as in Fig. 3 but on a night-time sequence of 150 stereo frames.