Quality Assessment of Non-dense Image Correspondences

Anita Sellent and Jochen Wingbermühle

Robert Bosch GmbH, CV Research Lab, Hildesheim, Germany
sellent@iam.unibe.ch, jochen.wingbermuehle@de.bosch.com

Abstract. Non-dense image correspondence estimation algorithms are known for their speed, robustness and accuracy. However, current evaluation methods evaluate correspondences point-wise and consider only correspondences that are actually estimated. They cannot evaluate the fact that some algorithms might leave important scene correspondences undetected - correspondences which might be vital for succeeding applications. Additionally, often the reference correspondences for real world scenes are also sparse. Outliers that do not hit a reference measurement can remain undetected with the current, point-wise evaluation methods. To assess the quality of correspondence fields we propose a histogram based evaluation metric that does not rely on point-wise comparison and is therefore robust to sparsity in estimate as well as reference.

1 Introduction

Image correspondence algorithms such as optical flow and stereo disparity estimation do not always return dense correspondences [4], especially if confidence measures [5] or consistency checks [4] are applied, Fig. 1. The results of both dense and non-dense algorithms are usually evaluated by point-wise comparison to dense ground truth fields [3, 6]. This evaluation, however, does not take into account the danger of sparse image correspondences to entirely miss complete objects. For correct evaluation and comparison of non-dense image correspondences, the evaluation measure should take into account the sparseness of a correspondence field as well as the distribution of correspondences over the objects in the scene. Many applications in computer vision and image processing are designed to deal also with sparse image correspondences, for example ego-motion estimation [7]. Therefore, direct evaluation of non-dense image correspondences is desirable.

A general problem in full-reference quality assessment is the availability of ground truth data. Dense ground truth correspondences are available only for a very small set of test sequences that are either synthetic, cf. Ref. [8, 9] or acquired in a controlled environment, cf. Ref. [1, 10]. For real world scenarios, laser scanners are often used to generate ground truth data, since they provide the easiest way to obtain reference measurements in realistic environments and to evaluate the performance of correspondence algorithms also outside of laboratories [11, 12]. Even carefully calibrated reference devices always contain a residual

A. Fusiello et al. (Eds.): ECCV 2012 Ws/Demos, Part II, LNCS 7584, pp. 114–123, 2012.

(a) First input frame (b) Dense flow estimation (c) Reliable flow estimates

Fig. 1. (a) Input image from scene taken from [1]. (b) Dense flow estimation with [2]. (c) Confidence measures [3] can eliminate spurious correspondences, yielding a non-dense but more accurate flow field.

calibration error [13], so that in these set-ups point-wise evaluation might not be a fair measurement.

We propose a histogram based error metric that compares dense as well as non-dense correspondences to given ground truth and does not depend on any additional algorithms such as interpolation or warping [14]. By design, our histogram based evaluation method can deal competently with alignment errors and is robust to reference measurements with a coarse sampling grid. In the evaluation of our error measure we focus on two aspects: **Distribution** (i.e. evaluate, how well correspondence are distributed between all different scene entities) and **Outliers** (i.e. indicate the presence and frequency of outliers).

1.1 Related Work

Due to its importance, a variety of optical flow and disparity estimation methods exist, c.f. Ref. [1, 6, 8, 9]. Some of these algorithms estimate correspondences for every pixel in the image [2, 15], while others focus on salient points [16] or apply confidence measures [3] and consistency checks [4] resulting in non-dense correspondence fields. In spite of the estimated correspondence fields being dense or non-dense, comparison is usually performed by using a point-wise error measure [9]. Point-wise error measures can also be evaluated for different image regions, e.g., highly textured or occluded regions [1,6] and so indicate where the results can be improved.

A further common measure is the ratio of wrong correspondences [6, 13, 17]. On the basis of the point-wise differences between test and reference fields, the correspondence is labeled wrong if a fixed threshold is exceeded. This approach takes the sparseness of the algorithm into account by normalizing with the total number of valid correspondences but evaluates accuracy only where both estimate and reference are defined. To obtain independence from reference measurements, Steingrube et al. [18] consider the free space in front of an object that moves without collisions and count the ratio of correspondence predicting spurious collisions. However, wrong pixel ratios cannot evaluate whether some objects remain completely undetected in non-dense correspondence estimation.

Szeliski [14] introduces a different metric in form of the prediction error for additional frames. Assuming that correspondences can be used to extrapolate or interpolate an additional frame, the prediction error towards this additional frame gives rise to an error measure that does not require known ground truth correspondences. Given these additional frames [1,6], this metric depends heavily on the warping scheme, is sensitive to locally cast shadows and most of all, it requires dense correspondences to warp the images.

While for synthetic scenes, e.g. in Ref. [1, 9], ground truth is known exactly, this assumption does not hold for all references. Correspondence established with fluorescent color [1] or structured light [10] are restricted to controlled indoor scenarios. In outdoor or real world dynamic scenarios, the need to approximate reference correspondences e.g. on known planar surfaces [17] arises. Relying on estimated correspondences, the approximated surface is not sure to either reflect the actual location of the surface, nor does it give any indication of the accuracy. Or, reference devices such as laser scanners are used to establish at least sparse measurements [12, 13]. This requires careful registration of the reference measurements to the images. Additionally, the resolution of laser-scanners is considerably lower than of images, so that references are given only for a sparse set of pixels which, however, is usually reasonabley distributed between the objects in the scene.

2 Evaluation Measures

We will use the same notational framework for stereo imagery and video imagery here. For a pair of rectified images from a stereo camera and for two temporally adjacent images from a monocular camera we use the notations of mappings from the image plane to gray- or color-values $I_1 : \Omega_1 \to \mathbb{R}^3$ and $I_2 : \Omega_2 \to \mathbb{R}^3$ with $\Omega_1, \Omega_2 \subset \mathbb{R}^2$. Likewise, image correspondence $w : \Omega \to \Omega_2$ with $\Omega \subset \Omega_1$ designates optical flow with its two independent components as well as stereo disparity with only one independent component.

We distinguish between two types of correspondences: the output of a correspondence algorithm is designated with $w_{est} : \Omega_{est} \to \Omega_2$ while the reference correspondence is designated with $w_{ref} : \Omega_{ref} \to \Omega_2$ where $\Omega_{est}, \Omega_{ref} \subset \Omega_1$ are the subsets for which correspondence are estimated or given, respectively. Note that the set $\Omega_0 = \Omega_{est} \cap \Omega_{ref}$ on which both estimate and reference are defined might be the empty set.

2.1 Point-Wise Evaluation Measures

For $\Omega_0 \neq \emptyset$ the usual error measure to compare correspondences is the point-wise difference, also known as the point-wise endpoint error [6]

$$EE(z) = \|w_{est}(z) - w_{ref}(z)\|_2 \quad \forall z \in \Omega_0 \tag{1}$$

together with the ratio of pixels that exceed a fixed threshold on the EE

$$R_\tau = \frac{1}{|\Omega_0|} |\{z \in \Omega_0 | EE(z) > \tau\}| \tag{2}$$

where $|\cdot|$ returns the number of elements in a set. For optical flow evaluation, also the point-wise angular error

$$AE(z) = \frac{(w_{est}(z) - z; 1)^\top (w_{ref}(z) - z; 1)}{\|(w_{est}(z) - z; 1)\|_2 \, \|(w_{ref}(z) - z; 1)\|_2} \tag{3}$$

and the ratio of wrong pixels in reference to this error, A_τ, are considered.

Usually the spatial means of the errors over all valid pixels are reported [3], i.e. $MEE = \frac{1}{|\Omega_0|} \sum_{z \in \Omega_0} EE(z)$ and $MAE = \frac{1}{|\Omega_0|} \sum_{z \in \Omega_0} AE(z)$ or - for stereo disparities [6] - the root-mean-squared-error

$$RMSE = \sqrt{\frac{1}{|\Omega_0|} \sum_{z \in \Omega_0} \|w_{est}(z) - w_{ref}(z)\|_2^2} \, .$$

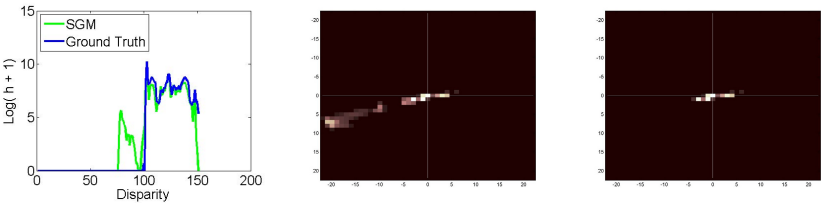
Given exactly aligned ground truth correspondence fields, point-wise measures are highly suited to find where estimated correspondences are accurate and where they are inaccurate. While outliers are clearly distinguishable in the differences between estimate and reference, Eqs. (1), (3), averaging over all valid pixels for $RMSE$, MEE and MAE mixes outliers with the accuracy of correct estimates and therefore these values are found to be of limited significance [1]. Considering the percentage of pixels with an error larger than a threshold τ, outliers can be identified more clearly, requiring however to set this threshold appropriately.

These measures are not sufficient for non-dense image correspondences: A scene can contain objects with properties that inhibit confident estimation of image correspondences on this object, e.g. low texture or changing illumination. However, the undetected object might be of significant importance in the scene. Additionally, if reference correspondences are sparse, the set of joint correspondences Ω_0 might contain only few, possibly non-representative correspondences.

2.2 Histogram-Based Evaluation Measures

We propose to consider the normalized distribution of flow vectors, i.e. the normalized histogram $h_w(u)$. For stereo disparity, the histogram is one dimensional (u denotes the disparity value), while for optical flow fields the histogram has two dimensions (u denotes the vertical and horizontal flow component), Figs. 2b, 2c. If the correspondences for every object in the scene are detected reliably, the normalized histograms h_{ref} and h_{est} are similar in shape and amplitude. If, in contrast, an entire object with independent correspondences remains undetected, the corresponding bins have smaller amplitude or might even remain empty, Fig.2. In the definition of the histograms the set Ω_0 of jointly defined correspondences is not relevant. Thus, the comparison of normalized histograms is independent towards sparseness as long as the correspondences are evenly distributed over all objects.

Histograms are expressed in a discretized form by choosing a suitable bin size b. Discretized histograms can be compared bin-wise [19]. Here, however, the choice of the bin size b as well as slight inaccuracies of the image correspondences



(a) Disparities (b) Full Flow Field (c) Object Motion Removed

Fig. 2. Histograms of disparity fields (a) are 1-dimensional. Histograms of flow fields are 2-dimensional. The difference between the full field (b) and flow fields with object-wise removed correspondences (c) is clearly visible.

have a high influence on the result of the comparison. A more robust method to compare normalized histograms has been proposed by Rubner et al. [20], known as the *Earth Mover's Distance* (EMD). The metric $EMD(h_1, h_2)$ between two histograms h_1 and h_2 gives the minimal cost required to match both histograms. With the choice of the EMD for histogram comparison, dependency on bin size can be reduced.

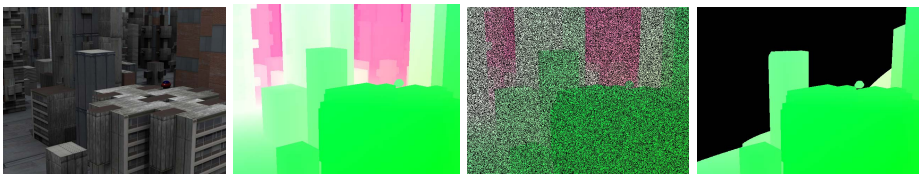
The advantage of histogram based evaluation is that it is robust towards different densities of reference and estimated correspondences, and considers distribution of the correspondences over different scene entities. Building histograms over the entire image, however, looses the property to distinguish between different regions of error. Thus, noise in one part of the image can compensate for erroneous correspondences in other parts of the image. In addition to the histogram over the entire image, we therefore consider histograms $h_w^{A_i}$ on subregions A_i of the image domain. As the location of objects is in general unknown, we simply partition the image domain in 2^n equally sized tiles and average the histogram distances

$$H^n = \frac{1}{2^n} \sum_{i=1}^{2^n} EMD(h_{est}^{A_i}, h_{ref}^{A_i}) \ . \tag{4}$$

Note that for $|A_i| = 1$ this corresponds to a version of the mean endpoint error where the error is discretized by the bin size. If a correspondence field contains numerous motions and a high degree of noise, we expect H^n to increase with increasing n, as on small regions noise can no longer be balanced. We noticed however, that usually consideration up to $n = 2$, i.e. on four tiles, suffices.

3 Implementation Details

We have chosen a bin size $b = 1$ for the discretized histograms in all our experiments. In the EMD calculation, the Euclidean l^2 norm between the two histograms has been chosen as the *ground distance*. For simplicity and reproducibility, we use the MATLAB implementation of [21] for the calculation of the



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

Fig. 3. (a) From the synthetic scene *Urban2* [1] with known ground-truth correspondences, (b), we remove 50% of the correspondences either (c) randomly or (d) contiguously, simulating flow estimates that are rejected by a confidence measure.

EMD. The number of non-empty bins influences the size of the problem and on the speed with which a solution can be found. For example, the determination of H^1 and H^2 for the *Urban 2* scene with 64 non-empty ground-truth bins and 193 non-empty bins in the estimated flow field requires $0.58s$ using a MATLAB implementation on a $3.40GHz$ CPU while the scenes *Venus* with 18 and 48 non-empty bins requires only $0.11s$.

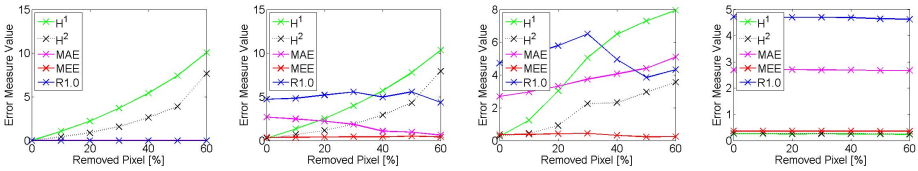
4 Experiments

We evaluate the proposed histogram measure to show its sensitivity to missed objects, its robustness to misalignments and its sensitivity to outliers. For comparable visualizations we use the color encoding as proposed in Ref. [22].

4.1 Non-densely Estimated Correspondences

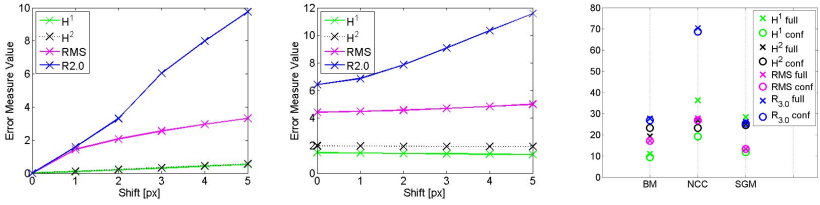
One motivation for our metric is the desire to evaluate whether a sparse algorithm misses a complete object. Figure 3 shows an exemplary scene, from which we successively remove the motion of the back-ground objects. As reference and test field differ only in missed correspondences, point-wise error metrics report zero differences. The value of the histogram distances between the fields, however, differs from zero, Fig. 4a. Considering correspondences estimated with the algorithm by Chambolle and Pock with default parameters [2] a similar behavior can be observed, Fig. 4. Histogram based measures detect the removal of objects while MEE and $R_{1.0}$ do not show a significant response. As the removed pixels have small motions, the MAE shows a tendency to decrease. However, if the fast moving objects in the foreground are removed, the MAE increases, Fig. 4c. Fig. 4d shows that the actual density does not have a significant effect on any error measure, as long as the distribution is approximately uniform.

As neither point-wise nor histogram-based evaluation distinguishes between reference and estimated correspondences, the considerations from above also hold for sparse reference measurements. We note that a disadvantage of the histogram based evaluation is that reference correspondences need to be evenly distributed over the objects in the scene.



(a) Reject slow obj. (b) Reject slow obj. (c) Reject fast obj. (d) Random

Fig. 4. Error metric response to missing pixel correspondences. Contiguous removal of correspondences from ground truth (a) and from estimation results using [2] (b), (c), as well as random removal from estimation results (d).



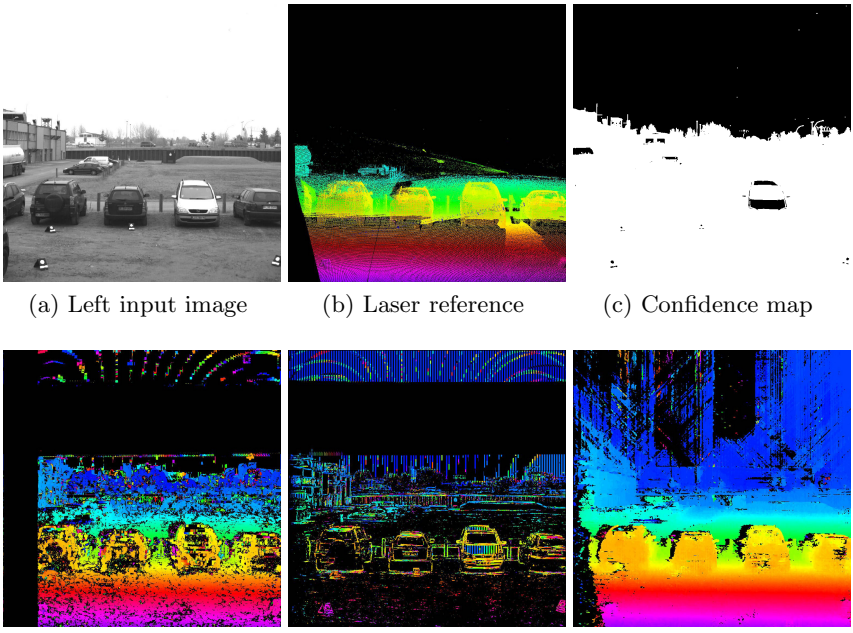
(a) Ground truth (b) SGM estimated (c) Error Metrics

Fig. 5. (a) Shifting ground truth correspondences [10] or (b) correspondences estimated with SGM [4] relative to dense ground truth. (c) Influence of rejection of disparities by confidence measure on error metric response.

4.2 Measured Reference Correspondences

Reference fields are usually acquired with a different device that needs to be calibrated to the input images. We compare the reference fields to a shifted version of themselves using the data from Ref. [10]. Also, we estimate disparities with an implementation of the semi-global matching (SGM) algorithm [4] and compare to shifted references. We note, Fig. 5, that in both cases the histogram-based error measure does not show any significant changes while the point-wise RMS error and the ratio of wrong correspondences shows larger deviation.

We simulate the influence of confidence measures on stereo correspondences by removing disparities for overexposed pixels which have an overexposed 4-neighborhood. Due to the - correct - absence of reference measurements in the overexposed sky, the RMS and ratio of wrong pixels only react to the rejection of overexposed regions in the cars. They are approximately indifferent to the impact of the confidence measure, utilizing block matching (BM), cross correlation (NCC), and SGM estimation, Fig. 5c. In contrast, the histogram-based error measure H^1 clearly indicates the improvement in the correspondence fields. The sub-region measure H^2 is of limited significance in this example as half the sub-regions contain less than 50 pixels of the reference field.



(a) Left input image (b) Laser reference (c) Confidence map

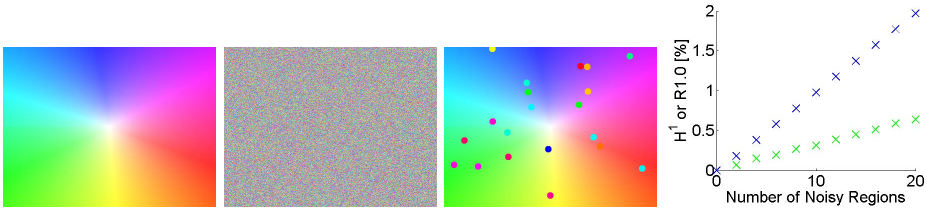
(d) Block matching (e) Cross Correlation (f) SGM

Fig. 6. For the scene *parking lot 3* [12] we compare estimated disparities to laser scanned reference data and evaluate the impact of a confidence map that eliminates correspondences with overexposed pixels

4.3 The Influence of Noise

Due to the loss of locality, the histogram-based metric cannot distinguish between noise and reliable estimates. Considering e.g. a purely diverging flow field and a flow field with random values in the same range, Fig. 7, only the point-based error measure can identify the poor quality of the noise field with $MEE = 10.44$ and $R_{1.0} = 99.79\%$ while the overall histogram distance $H^1 = 0.03$ is small. Here only the consideration of sub-level histogram distances H^n with $n > 1$, e.g. $H^2 = 7.65$ and $H^3 = 9.07$ that are considerably larger than H^1 hints at a noisy quality of the flow.

However, purely random correspondence fields can generally already be dismissed by visual inspection [22]. More important is the detection of small regions which are assigned a random correspondence field, Fig. 7c. To the purely diverging flow field we therefore add regions of random correspondences: randomly chosen image regions with a diameter of 10 pixels are replaced with an arbitrary correspondence somewhere in the image. As show in Fig. 7d, both point-wise and histogram based measures are able to detect the disturbance as both errors increase in a strictly monotone way with the number of outlier regions.



(a) Div. Motion (b) Noise (c) Outlier regions (d) Error Measures

Fig. 7. Histogram-based error measures cannot distinguish the diverging motion in (a) from random values in the same range (b). However, if an increasing number of regions is substituted with arbitrary matches within the image (c), point-based and histogram-based measures increase in a monotone fashion (d).

5 Conclusion and Future Work

Usual point-wise error measures can evaluate accuracy of a correspondence field only at those points where estimated and reference correspondence are defined. We have shown that this can be misleading in the evaluation of non-dense correspondences where entire objects can be undetected. Our histogram based measure can reliably detect missed objects in correspondence fields. Additionally, it is robust to misalignments between reference and estimates that occur due to reference sensor calibration errors. We have shown in our experiments that our histogram based measure is also suitable to evaluate the frequency of random outliers in correspondence fields. However, we believe no single metric should be used to evaluate correspondence algorithms. We strongly encourage authors of algorithms also to evaluate the accuracy of their correspondences, e.g. via point-wise error measures on suitable data set, the robustness of their algorithm to noise in the input images or the drift of their results when concatenated over long sequences.

References

1. Baker, S., Scharstein, D., Lewis, J., Roth, S., Black, M., Szeliski, R.: A database and evaluation methodology for optical flow. IJCV 92, 1–31 (2011)
2. Chambolle, A., Pock, T.: A first-order primal-dual algorithm for convex problems with applications to imaging. JMIV 40, 120–145 (2011)
3. Bruhn, A., Weickert, J.: A confidence measure for variational optic flow methods. In: Geometric Properties for Incomplete Data, pp. 283–298 (2006)
4. Hirschmuller, H.: Accurate and efficient stereo processing by semi-global matching and mutual information. In: CVPR, vol. 2, pp. 807–814 (2005)
5. Kondermann, C., Kondermann, D., Jähne, B., Garbe, C.: An Adaptive Confidence Measure for Optical Flows Based on Linear Subspace Projections. In: Hamprecht, F.A., Schnörr, C., Jähne, B. (eds.) DAGM 2007. LNCS, vol. 4713, pp. 132–141. Springer, Heidelberg (2007)

6. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. IJCV 47, 7–42 (2002)
7. Schill, F., Mahony, R., Corke, P.: Estimating Ego-Motion in Panoramic Image Sequences with Inertial Measurements. In: Pradalier, C., Siegwart, R., Hirzinger, G. (eds.) Robotics Research. STAR, vol. 70, pp. 87–101. Springer, Heidelberg (2011)
8. Barron, J., Fleet, D., Beauchemin, S.: Performance of optical flow techniques. IJCV 12, 43–77 (1994)
9. Klette, R., Kruger, N., Vaudrey, T., Pauwels, K., Van Hulle, M., Morales, S., Kandil, F., Haeusler, R., Pugeault, N., Rabe, C.: Performance of correspondence algorithms in vision-based driver assistance using an online image sequence database. IEEE T-VT, 1 (2011)
10. Scharstein, D., Szeliski, R.: High-accuracy stereo depth maps using structured light. In: CVPR, vol. 1, p. I–195. IEEE (2003)
11. Morales, S., Klette, R.: Ground Truth Evaluation of Stereo Algorithms for Real World Applications. In: Koch, R., Huang, F. (eds.) ACCV 2010 Workshops, Part II. LNCS, vol. 6469, pp. 152–162. Springer, Heidelberg (2011)
12. Reulke, R., Luber, A., Haberjahn, M., Piltz, B.: Validierung von mobilen Stereokamerasystemen in einem 3D-Testfeld. In: 3D-NordOst, vol. 12 (2009)
13. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? In: CVPR, Providence, USA (2012)
14. Szeliski, R.: Prediction error as a quality metric for motion and stereo. In: ICCV, vol. 2, pp. 781–788. IEEE (1999)
15. Stein, F.J.: Efficient Computation of Optical Flow Using the Census Transform. In: Rasmussen, C.E., Bülthoff, H.H., Schölkopf, B., Giese, M.A. (eds.) DAGM 2004. LNCS, vol. 3175, pp. 79–86. Springer, Heidelberg (2004)
16. Lucas, B., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: Proc. of the Conf. on Art. Intelligence (1981)
17. Liu, Z., Klette, R.: Approximated Ground Truth for Stereo and Motion Analysis on Real-world Sequences. In: Wada, T., Huang, F., Lin, S. (eds.) PSIVT 2009. LNCS, vol. 5414, pp. 874–885. Springer, Heidelberg (2009)
18. Steingrube, P., Gehrig, S.K., Franke, U.: Performance Evaluation of Stereo Algorithms for Automotive Applications. In: Fritz, M., Schiele, B., Piater, J.H. (eds.) ICVS 2009. LNCS, vol. 5815, pp. 285–294. Springer, Heidelberg (2009)
19. Stricker, M., Orengo, M.: Similarity of color images. In: Proc. SPIE Storage and Retrieval for Image and Video Databases, vol. 2420, pp. 381–392 (1995)
20. Rubner, Y., Tomasi, C., Guibas, L.: A metric for distributions with applications to image databases. In: ICCV, pp. 59–66. IEEE (1998)
21. Zhang, Y.: Solving large-scale linear programs by interior-point methods under the matlab environment. Technical Report TR96-01, Department of Mathematics and Statistics, University of Maryland (1995)
22. Sellent, A., Lauer, P.S., Kondermann, D., Wingbermühle, J.: A toolbox to visualize dense image correspondences. Technical report, Heidelberg Collaboratory for Image Processing, HCI (2012)