Illumination Invariant Cost Functions in Semi-Global Matching

Simon Hermann*, Sandino Morales, Tobi Vaudrey, and Reinhard Klette

.enpeda.. group, Dept. Computer Science, University of Auckland, New Zealand

Abstract. The paper evaluates three categories of similarity measures: ordering-based (census), gradient-based, and illumination-based cost functions. The performance of those functions is evaluated especially with respect to illumination changes using two different sets of data, also including real world driving sequences of hundreds of stereo frames with strong illumination differences. The overall result is that there are cost functions in all three categories that can perform well on a quantitative and qualitative level. This leads to the assumption that those cost functions are in fact closely related at a qualitative level, and we provide our explanation.

Keywords: cost functions, stereo matching, illumination invariance.

1 Introduction and Related Literature

Stereo algorithms typically solve the correspondence problem by using some cost function (usually called the *data cost*) to determine a good match between pixels, and a discontinuity condition (usually called the *smoothness cost*) to handle outliers and homogeneous areas of the data. The combined cost is then minimized using an optimisation strategy that yields either scan-line or global consistency. At the moment, state-of-the-art optimisation strategies can be split into four major groups: belief propagation [7], graph-cuts [3], dynamic programming [17] which has been extended to a semi-global-matching technique (SGM) [10], and variational techniques [4,23].

One major problem in stereo matching that affects, primarily, the data cost are illumination differences (between stereo images). This effect can have a major influence on the image data and therefore on the quality of the matching cost itself. This is especially prominent when it comes to real world image sequences [5]. One approach [1,20,21] to handle illumination changes is to decompose the input images into a structure and a texture component. The texture component tends to be robust against illumination changes.

Recent studies [11,12] evaluated the performance of cost functions under illumination changes and found the census [22] cost function to be very robust against lighting differences. However, a gradient-based measure was unfortunately not part of those evaluations. In [13] the gradient was employed as a

* The author thanks the German Academic Exchange Service (DAAD) for financial support.

R. Koch et al. (Eds.): ACCV 2010 Workshops, Part II, LNCS 6469, pp. 245–254, 2011.

similarity measure that was additively incorporated into the sum of absolute difference (SAD) cost function (accumulated over a 3×3 window). Another study used the same two cost functions (SAD and gradient) when creating a similarity measure, but used a multiplicative contribution along with the normalized cross correlation cost function [6]. The contribution of the gradient was shown to provide a more reliable cost function when analysed under different lighting conditions. However, none of those studies were using or analysing the gradient concept isolated from other cost functions to determine the performance contribution of the gradient.

In this paper, the performances of three cost functions are compared: SAD applied to residual images (RSAD), the census cost function (both were previously identified as being robust against illumination differences) and a gradient-based measure, each being a representative of different categories of cost functions. Census belongs to the non-parametric ordering-based cost functions. Distances of central differences as approximation of image derivatives define a gradient-based similarity measure. RSAD is used as an example of illumination-based cost functions. We also use the regular version, the SAD cost function, to evaluate a metric that purely relies on the assumption of intensity consistency. In the methodology presented below, the four cost functions are evaluated under differing illumination and exposure settings on data sets where ground truth is available. The performance comparison is done using SGM [10] as the optimization technique. This method has proven to be computationally efficient [8] and of high quality [14].

The main goal of the presented research is to improve the robustness of stereo algorithms when used in real-world applications. Therefore, a comparison of the performance of the selected cost functions using two real-world sequences is performed. The sequences were recorded using three synchronized (trinocular) cameras, so evaluation is possible using the prediction error technique as described in [15].

The following two sections introduce the matching costs, as well as the semi-global matching technique with implementation details used in the experiments. This is followed by the methodology, data sets, and testing measures used for evaluation. This leads onto a discussion of the experimental results, which is then finalised by conclusions.

2 Cost Functions

In a rectified stereo image pair we consider a *base* and a *match* image. The base image is assumed to be the left image L. The match image R is usually the right image. The images are of same size within the image domain Ω. We only consider intensity images (ignoring colour) in this paper with values in the range $[0, I_{\max}] \subset \mathbb{N}$. Any cost function Γ defines a global mapping $\Gamma(L, R) = C$ that takes rectified stereo images L and R as input, and outputs a 3D cost matrix C with elements $C(i, j, d)$. The cost matrix represents the cost when matching a pixel (i, j) in L with a pixel $(i - d, j)$ in R, for any relevant disparity

d in the range $[1, d_{\max}] \subset \mathbb{N}$ (zero is used for an "invalid" disparity, such as an occlusion). The ranges for i and j are $[0, n] \subset \mathbb{N}$ and $[0, m] \subset \mathbb{N}$, respectively. We simplify notation as we are working with rectified images (epipolar lines are aligned horizontally), and we consider a fixed image row j in both the base and match image. Let p_i denote a pixel location in L at column i. Let L_i be the value at this location in the base image; q_{i-d} denotes the pixel location $(i-d, j)$ in the match image R with intensity R_{i-d}. The cost can be abbreviated to omit the row $C(i, d)$.

We identify three different categories of cost functions: ordering-based, gradient-based, and intensity-based. We now introduce one representative of each function category that we evaluate in this paper.

Non parametric or ordering-based cost function. The census [22] cost function was identified to be a very robust measure when it comes to illumination changes [12]. Its performance serves as a reference when compared to the other two cost functions. We use it based on the following definition:

$$C_{\text{census}}(i, d) = \sum_{(x,y) \in \mathcal{N} + \{p_i\}} \rho(x, y, d) \text{ with} \tag{1}$$

$$\rho(x, y, d) = \begin{cases} 0 & \text{if } L_{x,y} > L_i \text{ and } R_{x-d,y} > R_{i-d} \\ 0 & \text{if } L_{x,y} < L_i \text{ and } R_{x-d,y} < R_{i-d} \\ 1 & \text{otherwise} \end{cases} \tag{2}$$

where $\mathcal{N}$ denotes the set of all nine pixel locations of the used 3×3 window when centred at reference point $(0, 0)$.

Gradient-based cost function. This cost function employs the spatial distance of the end points of the gradient vectors as the similarity measure. It is defined as:

$$C_{\text{GRAD}}(i, d) = |\nabla L_i - \nabla R_{i-d}|_1 \tag{3}$$

where ∇ is estimated using central differences[1] and $|*|_1$ is the L_1-norm. Using central differences also keeps the neighbourhood influence within a 3×3 window.

Intensity-based cost function. The *absolute difference* (AD) of the base and match pixel is the simplest and cheapest (in terms of computational cost) intensity-based measure:

$$C_{\text{AD}}(i, d) = |L_i - R_{i-d}| \tag{4}$$

In order to make a comparison more fair to census and the gradient, which use information from a 3×3 neighborhood, we choose to sum the absolute difference over a 3×3 window. This extension is known to be the *sum of absolute differences* (SAD) cost function. We define this intensity-based representative as:

[1] Our experiments use central differences. However, other gradient operators may possibly provide even better results depending on given image data.

$$C_{\mathrm{SAD}} = \frac{1}{|\mathcal{N}|} \sum_{(x,y)\in\mathcal{N}+\{p_i\}} |L_{x,y} - R_{x-d,y}| \tag{5}$$

with cardinality $|\mathcal{N}| = 9$. However, since SAD is known to perform bad when it comes to illumination differences, we also apply this cost function on the texture component of the input images. We calculate the residual image (texture component) T of an image I as

$$T(I) = I - S(I) \tag{6}$$

where $S(I)$ denotes the smoothed image (in this case, a 3×3 mean image) of I. We refer to this version as RSAD.

3 Semi-Global Matching

This paper uses the *semi-global matching* technique (SGM) [10]. The SGM algorithm approximates the minimum of a 2D energy function by minimizing multiple 1D energies, and employing a dynamic programming scheme. The energy function consists of a data term and a smoothness term. The smoothness term penalizes small disparity changes of neighbouring pixels with a rather low penalty c_1 to allow slanted surfaces. A second penalty is applied for larger disparity changes with a higher penalty c_2. This second penalty is independent of the actual disparity change in order to preserve depth discontinuities. The previously mentioned 1D energies are defined as minimum cost paths $L_{\mathbf{a}}$ that start at each border pixel of the image and are traversed in direction $\mathbf{a}$.

A direction is basically a digitized line, and all digital lines of identical slopes are considered to be equivalent. Usually eight directions (up, down, left, right, and the in-between angles) are sufficient in SGM to obtain high-quality results. For a digital line in direction $\mathbf{a}$, processed between image border and pixel p, we only consider the segment $p_0, p_1, \ldots, p_n$ of that digital line, with p_0 on the image border, and $p_n = p$. The cost at pixel position p (for a disparity d) on the path $L_{\mathbf{a}}$ is recursively defined as follows (for $i = 1, 2, \ldots, n$):

$$L_{\mathbf{a}}(p_i, d) = C(p_i, d) + \min \left\{ \begin{array}{c} L_{\mathbf{a}}(p_{i-1}, d) \\ L_{\mathbf{a}}(p_{i-1}, d-1) + c_1 \\ L_{\mathbf{a}}(p_{i-1}, d+1) + c_1 \\ \min_{\Delta} L_{\mathbf{a}}(p_{i-1}, \Delta) + c_2 \end{array} \right\} - \min_{\Delta} L_{\mathbf{a}}(p_{i-1}, \Delta) \tag{7}$$

where $C(p, d)$ corresponds to the data cost term and is the similarity cost of pixel p for disparity d. The costs of paths $L_{\mathbf{a}}$, for all (say, eight) directions $\mathbf{a}$, are accumulated at a pixel p, for all disparities d in the range $[1, d_{max}] \subset \mathbb{N}$, and the disparity d_{opt} with the lowest cost is finally selected. To adjust the second penalty, the magnitude of the forward difference is calculated at each pixel p_i in direction $\mathbf{a}$. The magnitude of the forward difference scales the penalty for each p_i with

$$c_2(p_i) = \frac{c_2}{|I(p_{i-1}) - I(p_i)|} \tag{8}$$

To enforce the uniqueness of a disparity map (for a given stereo pair), roles of base and match images are swapped, which allows the calculation of a second disparity image. In a final consistency check, a pixel is labelled valid only if the corresponding disparities are identical; otherwise the pixel is labelled invalid. This is often referred to as a left-right consistency check.

The implementation used in this paper follows the SGM description from the original paper, as outlined above. However, it deviates in the following three points. To achieve sub-pixel accuracy the original paper proposes the standard procedure to fit a quadratic curve through costs of disparities $d_{opt} - 1$, d_{opt}, and $d_{opt} + 1$, and to take the disparity position of the minimum. Since a comparison of cost functions is the objective of this paper, generating disparities with sub-pixel accuracy is omitted, as results may differ depending on the nature of the cost function.

The second difference is omitting the use of median filtering to remove outliers, because this is considered a post processing technique to improve performance, and raw performance of the cost functions are of interest in this paper.

The third difference is that costs are not scaled to 11-bit. The intention of this scaling is to have similar settings of penalties when cost functions are exchanged. However, simple scaling may not be descriptive enough to have a fair parameter setting between cost functions. For example, consider the census function that produces discrete costs in the range of $[0, 8] \subset \mathbb{N}$. There are basically no outliers possible, because of the nature of this function, while one outlier in SAD may result in an unfortunate scaling. However, this is an interesting topic and with a deeper understanding of the characteristics of cost functions w.r.t. the data, it should be possible to derive parameter settings for the optimization techniques. This will be discussion for future work. – Our implementation uses eight accumulation paths with $c_1 = 30$ and $c_2 = 150$.

4 Methodologies and Datasets

Illumination issues have been proven to cause major issues when it comes to stereo matching and may, in fact, be the worst type of noise for stereo matching [16]. The first methodology uses a data set where ground truth is available. It tests the presented cost functions under normal lighting conditions, as well as with different exposures and illuminations between the left and right camera. The calculated costs are then evaluated when applied to the SGM optimisation approach. The second methodology examines the behaviour of the analyzed cost functions in combination with SGM using real-world image sequences. To overcome the lack of ground truth correspondence, we evaluate the output of the stereo algorithm using a prediction error technique [15]; which is similar to the approach reported in [19] to evaluate optical flow techniques.

Synthetic or engineered test data. Such stereo images provide a way to obtain ground truth, but come with their specifica [9]. Stereo images may be recorded under different lighting and exposure settings, to provide test data where illumination/exposure could cause issues. Figure 1 shows an example from

Fig. 1. Illumination and exposure differences for the *Art* [14] input pair. Left to right: left (base) reference image, and right (match) image with identical illumination/exposure, right image with illumination change, right image with exposure change.

the data set [14] used in this paper; the cost functions are tested against the following images from this dataset: *Art, Books, Dolls, Laundry, Moebius,* and *Reindeer*. For each image pair used, the base image is using the exposure setting of 1 and illumination setting of 2, as defined on [14]. The left image is kept at this setting, but both illumination and exposure are varied in the right hand image. For each measure (outlined below) three tests are performed using different right hand images:

1. *Reference:* Identical lighting conditions (exp. 1, illum. 2)
2. *Illumination* Illumination difference (exp. 1, illum. 1)
3. *Exposure:* Exposure difference (exp. 0, illum. 2)

We calculate the *good pixel percentage* (GPP) for all datasets. The GPP is defined as follows. Let G be the ground truth image of the corresponding data set where G_i encodes the *true disparity* at pixel p_i. The *good pixel percentage* is defined below:

$$GPP = 100\% \times \frac{1}{|\Omega|} \sum_{(i,j)\in\Omega} \begin{cases} 1, & \text{if } |d_{opt} - G_i| \leq 1 \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

where Ω is the set of all pixels where $G_i \neq 0$, as 0 is used to identify occlusions.

In other words, if the optimal disparity d_{opt} is within one disparity distance of the ground truth, it is a good pixel. We take the mean GPP over all data sets for each illumination setting as quality measure for the cost functions. Results are shown in Figure 3 and discussed in Section 5.

Fig. 2. Sample stereo pairs from a real world data set on [5]. The first two images from the left are a stereo pair form the bird sequence. The last to images from the left are a sample stereo pair from the driving straight sequence.

Real world test data. We analysed two sequences (400 trinocular frames each) as available on [5], recorded within an urban scenario; both sequences, were recorded the same day with only a few minutes of difference. See Figure 2 for sample frames of both sequences. The first sequence, *bird*, was chosen due to the strong brightness difference between the stereo pairs and varies throughout the sequence. The second sequence, *driving straight*, was recording while driving on a straight road. It is a traffic sequence in which the brightness in both input images varies only slightly.

Trinocular stereo evaluation. The prediction error technique of [15] for stereo sequences requires at least three different images of the same scene (from different perspectives at the same time instance). The objective is to generate a *virtual* image V from the output of a stereo matching algorithm, and to compare this with an image recorded by an additional *control* camera, that was not used to generate the disparity map. We generate the virtual image by mapping (warping) each pixel of the reference image into the position in which it would be located in the *control image* N (image recorded with the control camera). Then, N and V are compared by calculating the *normalized cross correlation* (NCC) index between them as follows:

$$NCC(N,V) = \frac{1}{|\Omega|} \sum_{(i,j)\in\Omega} \frac{[N(i,j) - \mu_N][V(i,j) - \mu_V]}{\sigma_N \sigma_V} \tag{10}$$

where μ_N and μ_V denote the means, and σ_N and σ_V the standard deviations of the control and virtual images, respectively. The domain Ω is only for non-occluded pixels (i.e., pixels visible in the three images).

5 Results and Discussion

Figure 3 shows the mean GPP over the evaluated engineered test data for different illumination settings applied to SGM. From this evaluation, all cost functions seem to perform equally well, except for the pure SAD function, which is not surprising because the intensity consistency assumption is violated. It appears though that census is slightly more robust especially when looking at exposure changes, as the mean GPP does not change significantly when looking at different illumination settings. Conversely, RSAD and gradient both seem to be robust to illumination change, but not as much to exposure change.

Figure 4 shows the NCC percentage for all 400 frames of both real-world driving sequences. The overall performance on the driving straight (left) sequence is better than in the bird sequence (right). This may be explained because of the higher illumination variance between stereo frames in the bird sequence. We see that the overall quality of all cost functions is lower when illumination differences are strong; this is seen when we compare the driving straight (low changes) with the bird (high changes) sequence.

The gradient seems to outperform the census cost function for all but a few frames when looking at the driving straight scene (left). This may be due because illumination differences are not that strong. Otherwise performance appears to be almost identical.

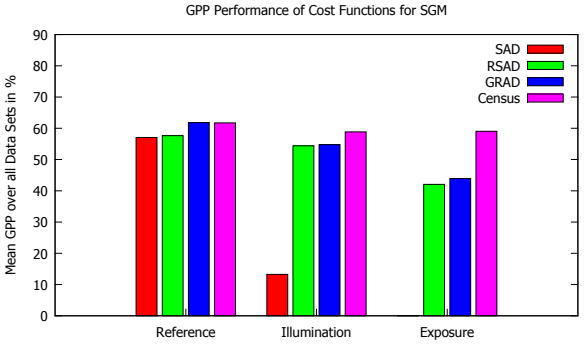

Fig. 3. Results for ground truth evaluation on engineered test data. The GPP is a mean value over all six datasets evaluated.

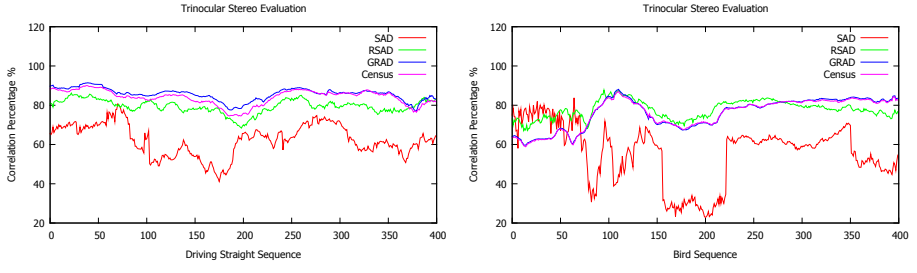

Fig. 4. Trinocular prediction error NCC analysis plots for the real-world data set. Left: Bird sequence. Right: Driving straight sequence.

However, all curves roughly seem to follow the same pattern (this is discussed later in this section), except for the pure SAD cost function.

The major difference is the RSAD cost function. While performance is consistently lower than for the gradient and the census function in the driving straight scene (left), it seem to respond slightly differently to the data in the bird sequence (right).

However, all cost functions seem to perform equally well (again except for the standard SAD). This may not be surprising because all of them respond to relative intensity jumps in the underlying image data. The left 3×3 window in Figure 5 shows a sample of a grey scale intensity image. The next window to the right shows the census transform when we choose 1 if the intensity increases from the centre pixel, and 0 otherwise. We gain from this transformation the signature vector $(1, 0, 1, 1, 1, 0, 1)$ when starting from the top left corner, and cycling clock-wise. However, if we compute forward differences in all eight directions of the 8-neighbourhood of the central pixel (look at the window labelled gradient) and write down the results in a vector (starting top left and cycling clock-wise), we get $(23, -41, 60, 47, 35, -10, 12, -31)$. If we just look at the signs and represent a positive value as a 1 and a negative value as 0, the resulting signature vector is identical to the census signature vector. This makes a close relation between

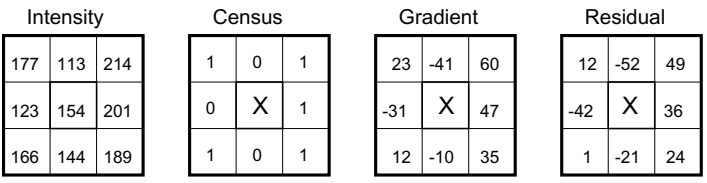

Fig. 5. *From left to right:* A 3×3 window of a intensity image. Followed by the corresponding census transformation. This is followed by forward differences when computed in all directions of a 8-neighbourhood. Finally the zero-mean calculation.

derivative-based (or gradient-based) and the census-based data descriptors which are employed for cost functions.

We can also compute the mean of this window (which is 165) and subtract it from the intensity of each neighbour we perform the zero-mean transformation. This is closely related to the residual computation we applied for the SAD cost function used in this paper. The resulting vector is the vector from the gradient shifted by an offset of 11 and would be identical if the mean happened to be 154.

This analysis shows that all three cost functions are related. All fit into a first order data term category, as each of those functions represent a relative intensity change in the image. But this is nothing else than the derivative in a distinctive direction; and this is closely related to edge detection.

6 Conclusions

This paper shows that the performance of a gradient based cost function competes with the performance of cost functions already identified as being robust to illumination changes. A potential relation between the categories of those cost functions is established. One conclusion of this analysis could be that finding a good and robust cost function for real world applications reduces to the problem of finding a cost function that describes intensity changes appropriately w.r.t the underlying data. All of the illumination robust cost functions seem, in effect, to be related to the gradient. The census function describes the gradient in a rough sense, which makes it robust to noise. The gradient and RSAD function adds intensity information on a relative scale to the cost, which adds more descriptive information to the cost. However, this makes those functions more affected by noise than the census. This may explain the better performance of census on the engineered data as noise has a bigger influence when exposure is changed.

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. Barnard, S.T., Fischler, M.A.: Computational stereo. ACM Computing Surveys 14, 553–572 (1982)

3. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. IEEE Trans. Pattern Analysis Machine Intelligence 23, 1222–1239 (2001)
4. Brox, T., Bruhn, A., Papenberg, N., Weickert, J.: High accuracy optical flow estimation based on a theory for warping. In: Pajdla, T., Matas, J(G.) (eds.) ECCV 2004. LNCS, vol. 3024, pp. 25–36. Springer, Heidelberg (2004)
5. .enpeda. image sequences analysis test site, `http://www.mi.auckland.ac.nz/EISATS`
6. El-Mahassani, E.D.: New robust matching cost functions for stereo vision. In: Proc. DICTA, pp. 144–150 (2007)
7. Felzenszwalb, P.F., Huttenlocher, D.: Efficient belief propagation for early vision. Int. J. Computer Vision 70, 41–54 (2006)
8. Gehrig, S.K., Eberli, F., Meyer, T.: A real-time low-power stereo vision engine using semi-global matching. In: Proc. ICCV, pp. 134–143 (2009)
9. Haeusler, R., Klette, R.: Benchmarking stereo data (Not the matching algorithms). In: Goesele, M., Roth, S., Kuijper, A., Schiele, B., Schindler, K. (eds.) Pattern Recognition. LNCS, vol. 6376, pp. 383–392. Springer, Heidelberg (2010)
10. Hirschmüller, H.: Accurate and efficient stereo processing by semi-global matching and mutual information. In: Proc. CVPR, vol. 2, pp. 807–814 (2005)
11. Hirschmüller, H., Scharstein, D.: Evaluation of cost functions for stereo matching. In: Proc. CVPR, pp. 1–8 (2007)
12. Hirschmüller, H., Scharstein, D.: Evaluation of stereo matching costs on images with radiometric differences. IEEE Trans. Pattern Analysis Machine Intelligence 31, 1582–1599 (2009)
13. Klaus, A., Sormann, M., Karner, K.: Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure. In: Proc. CVPR, vol. 3, pp. 15–18 (2006)
14. Middlebury College, stereo vision page, `http://vision.middlebury.edu/stereo/`
15. Morales, S., Vaudrey, T., Klette, R.: A third eye for performance evaluation in stereo sequence analysis. In: Jiang, X., Petkov, N. (eds.) CAIP 2009. LNCS, vol. 5702, pp. 1078–1086. Springer, Heidelberg (2009)
16. Morales, S., Woo, Y.W., Klette, R., Vaudrey, T.: A study on stereo and motion data accuracy for a moving platform. In: Kim, J.-H., Ge, S.S., Vadakkepat, P., Jesse, N., Al Manum, A., Puthusserypady, S.K., Rückert, U., Sitte, J., Witkowski, U., Nakatsu, R., Braunl, T., Baltes, J., Anderson, J., Wong, C.-C., Verner, I., Ahlgren, D. (eds.) Advances in Robotics. LNCS, vol. 5744, pp. 292–300. Springer, Heidelberg (2009)
17. Ohta, Y., Kanade, T.: Stereo by two-level dynamic programming. In: Proc. Int. Joint Conf. Artificial Intelligence, vol. 2, pp. 1120–1126 (1985)
18. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. In: Proc. ICCV, pp. 7–42 (2002)
19. Szeliski, R.: Prediction error as a quality metric for motion and stereo. In: Proc. ICCV, pp. 781–788 (1999)
20. Vaudrey, T., Klette, R.: Residual images remove illumination artifacts! In: Denzler, J., Notni, G., Süße, H. (eds.) Pattern Recognition. LNCS, vol. 5748, pp. 472–481. Springer, Heidelberg (2009)
21. Vaudrey, T., Wedel, A., Klette, R.: A methodology for evaluating illumination artifact removal for corresponding images. In: Jiang, X., Petkov, N. (eds.) CAIP 2009. LNCS, vol. 5702, pp. 1113–1121. Springer, Heidelberg (2009)
22. Zabih, R., Woodfill, J.: Non-parametric local transform for computing visual correspondence. In: Proc. ECCV, vol. 2, pp. 151–158 (1994)
23. Zach, C., Pock, T., Bischof, H.: A duality based approach for realtime TV-L1 optical flow. In: Hamprecht, F.A., Schnörr, C., Jähne, B. (eds.) DAGM 2007. LNCS, vol. 4713, pp. 214–223. Springer, Heidelberg (2007)