

Linear Time Illumination Invariant Stereo Matching

Jintao Xu^{1,2} · Qingxiong Yang²  · Jinhui Tang³ · Zuren Feng¹

Received: 12 September 2014 / Accepted: 4 February 2016 / Published online: 15 February 2016
© Springer Science+Business Media New York 2016

Abstract This paper proposes a new similarity measure that is invariant to global and local affine illumination changes. Unlike existing methods, its computational complexity is very low. When used for stereo correspondence estimation, its computational complexity is linear in the number of image pixels and disparity searching range. It also outperforms the current state of the art similarity measures in terms of accuracy on the Middlebury benchmark (with radiometric differences).

Keywords Stereo Matching · Minimum spanning tree · Radiometric variation

1 Introduction

Computational stereo continues to be an active area of intense research interest (Brown et al. 2003; Scharstein and Szeliski 2002b). The community has made significant progress in terms of both accuracy (Yang et al. 2009; Klaus et al. 2006; Wang and Yang 2011) and efficiency (Yang et al. 2010; Yang 2015) in the past decade. However, most of these

achievements were obtained based on the intensity consistency assumption, i.e., the intensity should be the same in both images if it corresponds to the same world point in the scene.

In practice, there will be radiometric differences caused by exposure and illumination changes, especially on community photo collections (Snaveley et al. 2006; Goesele et al. 2007). The most popular affine illumination invariant similarity measure is normalized cross-correlation (NCC). It is used in many other vision tasks besides stereo matching. However, it suffers from the fattening effect around occlusions. Census transform (Zabih and Woodfill 1994) is more robust to occlusion. It is indeed an image transform that captures the local order of intensities rather than the raw intensity value, and thus the computational complexity of the transform is independent of the disparity search range. However, it can be operated only on a single image channel and reconstruction accuracy is not comparable to the current state-of-the-art around depth edges.

The first edge-preserving affine illumination invariant similarity measure—ANCC was proposed by Heo et al. (2011). It is essentially an integration of the joint bilateral filter with NCC for maintaining depth edges, which are normally also the color edges. In Yoon and Kweon (2006), the joint bilateral filter is used with the reference camera image as guidance to smooth the matching cost for noise reduction without blurring across the color edges which normally corresponds to the depth edges. Mutual Information (MI) (Egnal 2000; Kim et al. 2003; Hirschmuller 2008) is also used as a similarity measure in stereo matching but is only robust to global illumination changes (Hirschmuller and Scharstein 2009). Heo et al. (2009) later combined MI with the SIFT descriptor to handle local radiometric variations. Heo et al. (2013) also developed an iterative framework that infers both accurate depth maps and color-consistent stereo images for

Communicated by Masatoshi Okutomi.

✉ Qingxiong Yang
qiyang@cityu.edu.hk

Jinhui Tang
jinhuitang@njust.edu.cn

- ¹ Xi'an Jiaotong University, Xi'an, China
- ² City University of Hong Kong, Kowloon Tong, Hong Kong, China
- ³ School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, Jiangsu, China

radiometrically varying stereo images. All the above similarity measure metrics were developed under the assumption of Lambertian scenes. Using as few as two stereo image pairs captured under different illumination conditions, Wang et al. (2007) proposed a new invariant measure called *light transport constancy* based on a rank constraint for non-Lambertian surfaces.

This paper proposes a new similarity measure that outperforms the state of the art in terms of both accuracy and speed. It is invariant to global and local affine illumination changes and is valid only for Lambertian scenes. The similarity measure is derived as follows: (1) finding an optimal local affine transform at every correspondence based on the local linear model; (2) computing the similarity according to the affine transform. The computation of the affine transform is formulated as an image filtering problem that can be solved efficiently and effectively using edge-preserving filters (e.g., the tree filter Yang (2015) whose computational complexity is the same as the box filter). The experimental results are evaluated on the Middlebury data set (Scharstein and Szeliski 2002a), showing that our similarity measure is the top performer.

2 Illumination Invariant Stereo Correspondence

In this section, we give a detailed description of the proposed similarity measure and its usage in stereo matching. Section 2.1 gives a brief overview of the linear local model and Sect. 2.2 presents the detailed extension of this model for a new affine illumination invariant similarity measure and formulates correspondence estimation as an image filtering problem.

The derivation of the affine illumination invariant similarity measure is closely related to the *Matting Laplacian* proposed in Levin et al. (2008) and its approximate - *Guided Image Filter* in He et al. (2013). The detailed relationship is discussed in Sect. 2.2.1. Section 2.3 extends the framework in Sect. 2.2 to color image. Section 2.4 gives a detailed discussion of the potential image filters that are suitable for the proposed framework and Sect. 2.5 presents a new filtering method for the proposed similarity measure.

2.1 Linear Local Model

This section gives a brief introduction of the linear local model in computer vision. The linear local model is normally used to find a locally affine projection/mapping between two images by minimizing the projection error between the two. It has been demonstrated to be very effective for many computer vision applications including super resolution (Zomet and Peleg 2002), natural image matting (Levin et al. 2008), haze removal (He et al. 2011), and image filtering (He et al.

2013). The fast edge-aware filtering technique proposed in He et al. (2013) is a great success and has been widely used as a basic tool in many computer vision and computer graphics tasks like image matting (He et al. 2011), stereo matching (Hosni et al. 2013; De-Maeztu et al. 2011; Zhu et al. 2012), image retargeting (Ding et al. 2011), and image colorization (Chia et al. 2011), etc. The affine illumination invariant similarity measure proposed in Sect. 2.2 is an extension of the linear local model used in Levin et al. (2008) and its fast solutions presented in Sect. 2.4 and 2.5 are inspired by He et al. (2013).

2.2 Affine Transform for Gray-Scale Images

Let I_L and I_R denote the left and right image of a gray-scale stereo image pair, respectively. The traditional similarity measure metric for stereo matching is based on the intensity/color consistency assumption. That is, the corresponding pixels in I_L and I_R should have the same intensities:

$$I_L(p) = I_R(p'), \quad (1)$$

where p and p' denote the corresponding pixels in I_L and I_R , respectively.

However, there will be radiometric differences caused by exposure and illumination changes in practice. In this case, the corresponding pixels will have different brightness values, and the intensity of the image cannot be used as the matching invariant any more. Under the assumption of global affine illumination changes, we can linearly transform I_R so that the resulting image intensity can be directly used for matching. This assumption is, however, too restrictive and not practical. This paper relaxes it by assuming that the affine illumination changes are locally smooth so that the affine transform will be the same or very close inside a local region:

$$I_L(p) = \mathcal{A}_{p,p'}(I_R(p')) = a_{p,p'} \cdot I_R(p') + b_{p,p'}, \quad (2)$$

where $a_{p,p'}$ and $b_{p,p'}$ are two parameters of the affine transform $\mathcal{A}_{p,p'}$ between pixel p and p' . Let $\Delta = p - p'$ denote the motion vector between p and p' , and $I_{R,\Delta}$ be the shifted version of image I_R according to motion vector Δ so that $I_{R,\Delta}(p) = I_R(p')$, Eq. (2) can be rewritten according to each possible motion vector candidate Δ as follows:

$$I_L(p) = a_p \cdot I_{R,\Delta}(p) + b_p, \quad (3)$$

where $a_p = a_{p,p'}$ and $b_p = b_{p,p'}$.

Note that in practice, nonlinear gamma correction will be normally applied to the response of the camera which is linear in the flux of the incident light; thus the local affine transform assumption will be violated. A potential solution is using the

logarithm of the image but the dissimilarity between images will be reduced. As a result, ANCC (Heo et al. 2011) combines the similarity measure obtained from both the original image and its logarithm. However, we use only the original image to demonstrate that the proposed similarity measure is robust to the violation of the local affine transform assumption.

The illumination invariant stereo matching problem is then formulated as the computation of the affine transform at each pixel location using Eq. (3). Let Φ_p denote all pixels that have the same affine transform as pixel p in I_L , Eq. (3) can be extended as

$$\begin{pmatrix} \vdots & \vdots \\ I_{R,\Delta}(q) & 1 \\ \vdots & \vdots \end{pmatrix} \cdot \begin{pmatrix} a_p \\ b_p \end{pmatrix} = \begin{pmatrix} \vdots \\ I_L(q) \\ \vdots \end{pmatrix}, \tag{4}$$

where $q \in \Phi_p$.

Let

$$A = \begin{pmatrix} \vdots & \vdots \\ \mathcal{W}(p, q) \cdot I_{R,\Delta}(q) & \mathcal{W}(p, q) \cdot 1 \\ \vdots & \vdots \end{pmatrix} \tag{5}$$

and

$$B = \begin{pmatrix} \vdots \\ \mathcal{W}(p, q) \cdot I_L(q) \\ \vdots \end{pmatrix}, \tag{6}$$

Eq. (4) can be rewritten as

$$A \cdot \begin{pmatrix} a_p \\ b_p \end{pmatrix} = B \tag{7}$$

for any pixel $q \in I_L$ under the assumption that there exists a function \mathcal{W} that can perfectly decide whether q has the same affine transform as p :

$$\mathcal{W}(p, q) = \begin{cases} 1 & a_p = a_q \text{ and } b_p = b_q, \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

The linear system presented in Eq. (7) can be solved efficiently using Conjugated Gradient (Hestenes and Stiefel 1952) because the order of the linear system is very low with

$$A^T A = \begin{pmatrix} \sum_{q \in I_{R,\Delta}} \mathcal{W}(p, q)^2 \cdot I_{R,\Delta}^2(q) & \sum_{q \in I_{R,\Delta}} \mathcal{W}(p, q)^2 \cdot I_{R,\Delta}(q) \\ \sum_{q \in I_{R,\Delta}} \mathcal{W}(p, q)^2 \cdot I_{R,\Delta}(q) & \sum_{q \in I_{R,\Delta}} \mathcal{W}(p, q)^2 \end{pmatrix} \tag{9}$$

and

$$A^T B = \begin{pmatrix} \sum_{q \in I_L} \mathcal{W}(p, q)^2 \cdot I_L(q) \cdot I_{R,\Delta}(q) \\ \sum_{q \in I_L} \mathcal{W}(p, q)^2 \cdot I_L(q) \end{pmatrix}. \tag{10}$$

Note that each element of $A^T A$ and $A^T B$ can be treated as the response of a spatial filter (without a normalization step) weighted by $\mathcal{W}(p, q)^2 = \mathcal{W}(p, q)$. Let \mathcal{F} denote this filter, $\mathcal{F}_X(p)$ denote the corresponding response of an image X at a pixel location p , and $I \cdot J$ denote the element-by-element multiplication result of image I and another image J ; the linear system in Eq. (7) can be rewritten as:

$$\begin{pmatrix} \mathcal{F}_{I_{R,\Delta} \cdot I_{R,\Delta}}(p) & \mathcal{F}_{I_{R,\Delta}}(p) \\ \mathcal{F}_{I_{R,\Delta}}(p) & 1 \end{pmatrix} \cdot \mathcal{X} = \begin{pmatrix} \mathcal{F}_{I_L \cdot I_{R,\Delta}}(p) \\ \mathcal{F}_{I_L}(p) \end{pmatrix}, \tag{11}$$

where $\mathcal{X} = \begin{pmatrix} a_p \\ b_p \end{pmatrix}$. The matching cost obtained at pixel p with motion vector Δ is

$$\begin{aligned} \sum_{q \in I_L} \mathcal{W}(p, q) \cdot (I_L(q) - a_p \cdot I_{R,\Delta}(q) - b_p)^2 \\ = \mathcal{F}_{I_L \cdot I_L}(p) + a_p^2 \cdot \mathcal{F}_{I_{R,\Delta} \cdot I_{R,\Delta}}(p) + b_p^2 - 2b_p \cdot \mathcal{F}_{I_L}(p) \\ - 2a_p \cdot \mathcal{F}_{I_L \cdot I_{R,\Delta}}(p) + 2a_p b_p \cdot \mathcal{F}_{I_{R,\Delta}}(p). \end{aligned} \tag{12}$$

The detailed derivation of Eq. (12) is presented in Appendix 1.

Apparently, the computational complexity of the matching cost at each correspondence/disparity candidate mainly depends on the computational complexity of filter \mathcal{F} . According to Eqs. (11) and (12), filter \mathcal{F} will be applied to a total of five images including I_L , $I_L \cdot I_L$, $I_{R,\Delta}$, $I_L \cdot I_{R,\Delta}$ and $I_{R,\Delta} \cdot I_{R,\Delta}$. Nevertheless, image I_L and $I_L \cdot I_L$ is independent of motion vector Δ , and thus is only filtered once. As a result, the computational complexity of the proposed invariant matching cost can be very efficient - $O(1)$ at every pixel given a motion vector candidate Δ , as long as filter \mathcal{F} has a $O(1)$ solution.

2.2.1 Relation to the Matting Laplacian (Levin et al. 2008) and Guided Image Filter (Scharstein and Szeliski 2002a)

The derivation of the affine illumination invariant similarity measure in Sect. 2.2 is closely related to the solutions of the linear systems in Levin et al. (2008) and He et al. (2013). Let β denote a coarse alpha matte (e.g., a trimap); Levin et al. (2008) expresses the alpha matte α of a grayscale image I as a locally linear function

$$\alpha_p = a_p I_p + b_p, \tag{13}$$

with the constraint that

$$\alpha_p \approx \beta_p \tag{14}$$

at each pixel location p , for instance, $\alpha_p = \beta_p$ for confident pixels in the coarse alpha matte.

Note that Eq. (13) is the same as Eq. (3) in Sect. 2.2, and Eq. (14) gives a weight to each linear equation in the linear system in Eq. (13). This is indeed very close to function \mathcal{W} defined in Eq. (8).

In practice, Levin et al. (2008) computes an $N \times N$ matrix L named *Matting Laplacian* at each pixel location to eliminate the linear mapping functions $\{a_p, b_p\}$ and formulates the computation of the unknown alpha matte as

$$(L + \Lambda)\alpha = \Lambda\beta, \tag{15}$$

where Λ is a diagonal matrix encoded with the weights of the constraints. The fast filtering technique presented in He et al. (2013) can be used as a good approximation of the global optimized solution in Levin et al. (2008) when β is reasonably good. Based on the following linear system

$$\beta_p = a_p I_p + b_p, \tag{16}$$

it computes linear mapping functions $\{a_p, b_p\}$ via linear regression:

$$a_p = \frac{\frac{1}{|w_p|} \sum_{q \in w_p} I_q \beta_q - \mu_p \bar{\beta}_p}{\sigma_p^2 + \epsilon}, \tag{17}$$

$$b_p = \bar{\beta}_p - a_p \mu_p, \tag{18}$$

where μ_p and σ_p^2 are the mean and variance of I in a local patch w_p around pixel p , $|w_p|$ is the number of pixels in w_p , and $\bar{\beta}_p$ is the mean of β in w_p . The mean of the mapping function ($\bar{a}_p = \frac{1}{|w_p|} \sum_{q \in w_p} a_p$ and $\bar{b}_p = \frac{1}{|w_p|} \sum_{q \in w_p} b_p$) is used to compute the alpha matte

$$\alpha_p = \bar{a}_p I_p + \bar{b}_p. \tag{19}$$

The computational complexity of this approximation is very low. The main computation is the box filtering/mean operations in Eq. (17) which can be computed in time linear in the number of pixels. It is similar to our solution for Eq. (12) whose computational complexity is also $O(1)$ at every pixel given a motion vector candidate Δ , as long as filter \mathcal{F} has a $O(1)$ solution.

2.3 Extension to Color Images

The illumination invariant stereo correspondence framework in Sect. 2.2 is derived based on grayscale images. This section further extends it to color images. A color image has

three color channels: red, green, and blue channel. A simple extension is using the sum of the matching cost obtained from each color channel. However, this extension ignores the relationship between the color channels. Similar to Levin et al. (2008), this section extends the affine transform presented in Eq. (3) by assuming that there is a linear affine relationship between a color channel of an image (e.g., I_L) and all the color channels of the image to be matched (e.g., I_R). Let I^c denote the c -th channel of an image I , Eq. (3) can be extended for color images as follows:

$$\begin{aligned} I_L^1(p) &= a_p^{11} I_{R,\Delta}^1(p) + a_p^{12} I_{R,\Delta}^2(p) + a_p^{13} I_{R,\Delta}^3(p) + b_p^1, \\ I_L^2(p) &= a_p^{21} I_{R,\Delta}^1(p) + a_p^{22} I_{R,\Delta}^2(p) + a_p^{23} I_{R,\Delta}^3(p) + b_p^2, \\ I_L^3(p) &= a_p^{31} I_{R,\Delta}^1(p) + a_p^{32} I_{R,\Delta}^2(p) + a_p^{33} I_{R,\Delta}^3(p) + b_p^3. \end{aligned} \tag{20}$$

Let $I^c \cdot I^{c'}$ denote the element-by-element multiplication result of the c -th channel of a color image I and c' -th channel of another color image I' , and

$$\tilde{\mathcal{X}} = \begin{pmatrix} a_p^{11} & a_p^{12} & a_p^{13} & b_p^1 \\ a_p^{21} & a_p^{22} & a_p^{23} & b_p^2 \\ a_p^{31} & a_p^{32} & a_p^{33} & b_p^3 \end{pmatrix}^T \tag{21}$$

Eq. (11) can be extended for color images as follows:

$$\begin{aligned} &\begin{pmatrix} \mathcal{F}_{I_{R,\Delta}^1 \cdot I_{R,\Delta}^1}(p) & \mathcal{F}_{I_{R,\Delta}^1 \cdot I_{R,\Delta}^2}(p) & \mathcal{F}_{I_{R,\Delta}^1 \cdot I_{R,\Delta}^3}(p) & \mathcal{F}_{I_{R,\Delta}^1}(p) \\ \mathcal{F}_{I_{R,\Delta}^2 \cdot I_{R,\Delta}^1}(p) & \mathcal{F}_{I_{R,\Delta}^2 \cdot I_{R,\Delta}^2}(p) & \mathcal{F}_{I_{R,\Delta}^2 \cdot I_{R,\Delta}^3}(p) & \mathcal{F}_{I_{R,\Delta}^2}(p) \\ \mathcal{F}_{I_{R,\Delta}^3 \cdot I_{R,\Delta}^1}(p) & \mathcal{F}_{I_{R,\Delta}^3 \cdot I_{R,\Delta}^2}(p) & \mathcal{F}_{I_{R,\Delta}^3 \cdot I_{R,\Delta}^3}(p) & \mathcal{F}_{I_{R,\Delta}^3}(p) \\ \mathcal{F}_{I_{R,\Delta}^1}(p) & \mathcal{F}_{I_{R,\Delta}^2}(p) & \mathcal{F}_{I_{R,\Delta}^3}(p) & 1 \end{pmatrix} \cdot \tilde{\mathcal{X}} \\ &= \begin{pmatrix} \mathcal{F}_{I_{R,\Delta}^1 \cdot I_L^1}(p) & \mathcal{F}_{I_{R,\Delta}^1 \cdot I_L^2}(p) & \mathcal{F}_{I_{R,\Delta}^1 \cdot I_L^3}(p) \\ \mathcal{F}_{I_{R,\Delta}^2 \cdot I_L^1}(p) & \mathcal{F}_{I_{R,\Delta}^2 \cdot I_L^2}(p) & \mathcal{F}_{I_{R,\Delta}^2 \cdot I_L^3}(p) \\ \mathcal{F}_{I_{R,\Delta}^3 \cdot I_L^1}(p) & \mathcal{F}_{I_{R,\Delta}^3 \cdot I_L^2}(p) & \mathcal{F}_{I_{R,\Delta}^3 \cdot I_L^3}(p) \\ \mathcal{F}_{I_L^1}(p) & \mathcal{F}_{I_L^2}(p) & \mathcal{F}_{I_L^3}(p) \end{pmatrix}, \end{aligned} \tag{22}$$

and the matching cost at pixel p is

$$\begin{aligned} &\sum_{c1=1}^3 \sum_{q \in I_L} \mathcal{W}(p, q) \cdot (I_L^{c1}(q) \\ &\quad - \sum_{c2=1}^3 a_p^{c1c2} I_{R,\Delta}^{c2}(q) - b_p^{c1})^2 = \sum_{c1=1}^3 \mathcal{F}_{I_L^{c1} \cdot I_L^{c1}}(p) \\ &\quad + \sum_{c3=1}^3 \sum_{c2=1}^3 (\sum_{c1=1}^3 a_p^{c1c2} \cdot a_p^{c1c3}) \mathcal{F}_{I_{R,\Delta}^{c2} \cdot I_{R,\Delta}^{c3}}(p) \\ &\quad + \sum_{c1=1}^3 (b_p^{c1})^2 - 2 \sum_{c1=1}^3 b_p^{c1} \mathcal{F}_{I_L^{c1}}(p) \end{aligned}$$

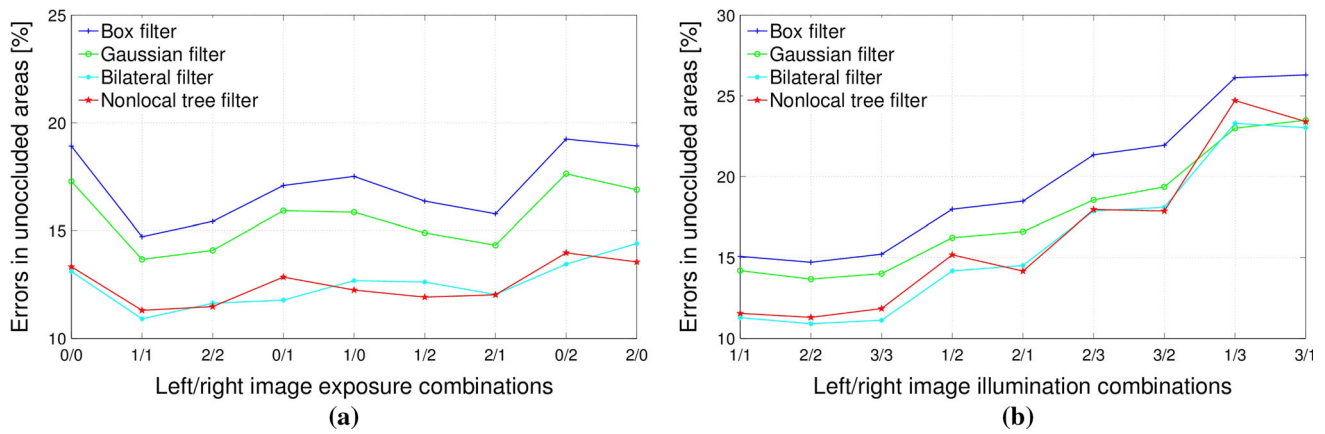


Fig. 1 Numerical comparison of different approximations of filter \mathcal{F} on stereo correspondence estimation under different radiometric variations. **a** Presents the average errors under different exposure combinations and **b** presents the average errors under different illumination

combinations. The edge preserving bilateral and nonlocal tree filters perform better than the others under both exposure and illumination changes

$$\begin{aligned}
 & -2 \sum_{c1=1}^3 \sum_{c2=1}^3 a_p^{c1c2} \mathcal{F}_{I_L^{c1}, I_R^{c2}}(p) \\
 & + 2 \sum_{c2=1}^3 \left(\left(\sum_{c1=1}^3 a_p^{c1c2} \cdot b_p^{c1} \right) \cdot \mathcal{F}_{I_R^{c2}}(p) \right) \quad (23)
 \end{aligned}$$

The detailed derivation of Eq. (22) and (23) is presented in Appendix 2 and 3.

2.4 Discussion of Filter \mathcal{F}

Sections 2.2 and 2.3 reformulate the correspondence estimation problem as an image filtering problem with the assumption of a perfect filter \mathcal{F} defined in Eq. 8. However, this filter does not exist in practice. This section presents a discussion of the potential approximations of the perfect \mathcal{F} using existing image filters.

An optimal \mathcal{F} filter can perfectly determine whether two pixels have undergone the same illumination changes or not by giving out a binary response, where “0” means different illumination changes and “1” means the same illumination changes. The first class of approximations of \mathcal{F} are popular low pass filters including Box filter and Gaussian filter. The basic assumption is that when two pixels are spatially close to each other, then they are likely to have the same illumination changes. The detailed numerical comparison of the performance of the two spatial filters is presented as the blue and green curves in Fig. 1. The parameters for each filter (e.g., filter kernel size) are all trained using the *Aloe* data set (Scharstein and Szeliski 2002a) with no illumination or exposure changes. The trained parameters are used for all the other data sets and other illumination/exposure settings. Note that although the box filter also gives a binary response by giving all neighboring pixels all equal weights and rejects

all pixels outside the filter, its performance is indeed lower than the Gaussian filter that weights contribution of neighborhood pixels according to their closeness to the center. In the Gaussian filter, pixels further away from the center will be weighted less as they are more likely to undergo different illumination changes, and this is a better approximation of \mathcal{F} in Eq. 8.

Besides spatial distance, the color distance can also be used to measure the similarity between two pixels. If two pixels have different colors, they are also likely to have different illumination changes. Figure 2 presents the illumination change maps with the ground-truth disparity maps of the *Aloe* and *Art* data sets under different lighting and exposure changes between left and right images. This shows that two pixels are likely to have similar illumination variance if they are close to each other and have similar color. Low pass filters that take into account both spatial and color distance are normally edge-preserving filters, and the bilateral filter is the most popular one. A bilateral filter is similar to the Gaussian filter but has an additional filter kernel named range filter kernel that respond with respect to the color similarity, and a joint bilateral filter uses an additional image as the guidance to compute the range filter kernel. The joint bilateral filter outperforms the Gaussian filter on average as shown in Fig. 1 when the reference camera image is used as the guidance image to compute the range filter kernel. Note that the cyan curves (corresponding to the joint bilateral filter) are normally below the green curves.

The main limitation of the use of the (joint) bilateral filter is that the computational complexity of its brute-force implementation is high. The recent tree filter proposed by Yang (2015) is indeed also a special type of bilateral filter. Unlike the standard bilateral filter, it performs by filtering along a minimum spanning tree (MST) and thus the computational

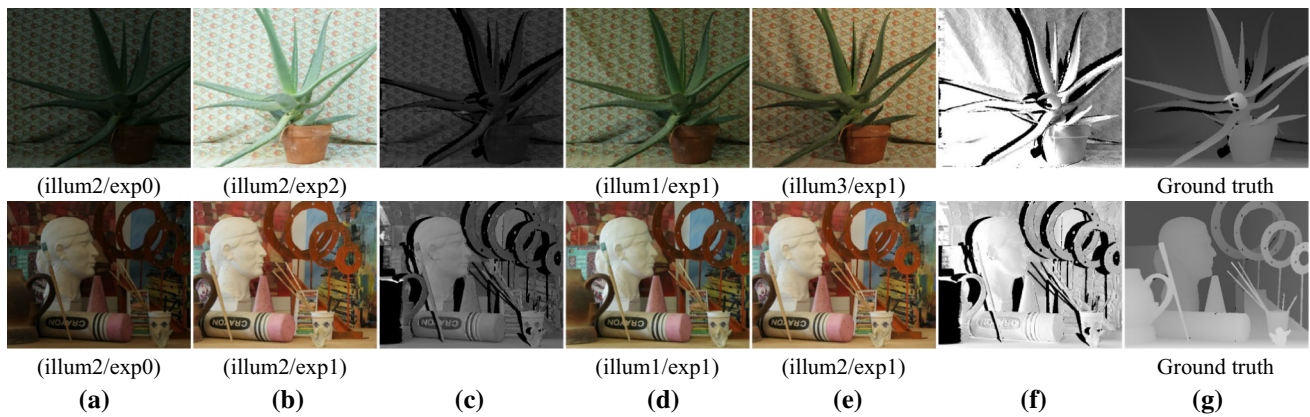


Fig. 2 Illumination change maps on *Aloe* and *Art* data sets under different illumination and exposure combinations. **a, b** Input *left* and *right* images under exposure changes and **d, e** input *left* and *right* images under illumination changes. **c, f** Corresponding illumination change maps computed from **a, b** and **d, e**, respectively. **g** Ground-truth dis-

parity maps. **c, f** demonstrate that the illumination change between two pixels is closely related to their spatial distance and color differences: if two pixels are spatially close to each other and have similar color, they are likely to have similar illumination change

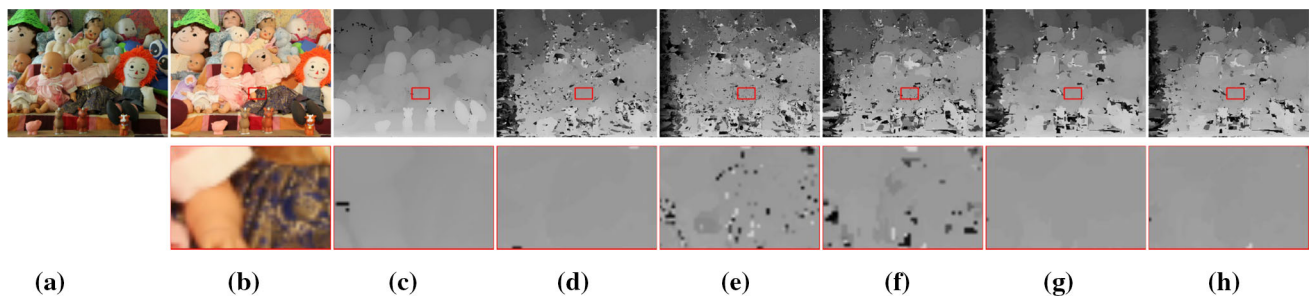


Fig. 3 Visual comparison of different approximations of filter \mathcal{F} on stereo correspondence estimation under different illumination combinations. **a, b** Input *right* image (with illum1/exp1) and *left* image (with illum3/exp1) of *Dolls* data set. **c** Ground-truth disparity map of **b**. **d–h** are disparity maps obtained from the proposed similarity measure with filter \mathcal{F} being **d** Gaussian filter, **e** bilateral filter, **f** nonlocal tree filter, **g** local tree filter, and **h** combined tree filter, respectively. Note

that the Gaussian filter outperforms standard edge-preserving filters (e.g., the bilateral filter and nonlocal tree filter) around textured regions. The numbers presented under each disparity map are the corresponding percentage of error pixels, which demonstrate that the proposed combination of the local and nonlocal tree filter is the best approximation of filter \mathcal{F}

complexity is extremely low—it is independent of the filter kernel size. The performance of this nonlocal tree filter is close to the joint bilateral filter as can be seen in Fig. 1.

Nevertheless, the performance of most edge-preserving filters are low around highly-textured regions as shown in Fig. 3e, f. Figure 3d, f present visual comparisons of disparity maps obtained from the proposed similarity measure method by approximating filter \mathcal{F} with Gaussian, bilateral, and nonlocal tree filters, respectively. The disparity maps obtained from the joint bilateral filter and nonlocal tree filter (Fig. 3e, f) appear be more noisy than the disparity map obtained from the Gaussian filter (Fig. 3d) around textured regions. A simple combination of an edge-preserving filter and a Gaussian filter seems to be more reliable. However, it will definitely decrease the reconstruction accuracy around depth edges. As a result, a combined filtering framework that is more robust to textures is presented in Sect. 2.5.

2.5 Combination of Local and Nonlocal Tree Filter

As discussed in Sect. 2.4, the nonlocal tree filter (Yang 2015) is very suitable for the proposed correspondence estimation framework except around highly-textured regions. The nonlocal tree filter performs on an MST derived from the reference camera image. The similarity between any two pixels is decided by their shortest distance on the MST, and the distance between every two neighboring pixels (on MST) was originally defined in Yang (2015) as their color difference, and thus the distance between any two pixels on MST will be the sum of the color differences on MST. As a result, the distance of two pixels inside a flat region will be always zero. However, the distance between two pixels inside a highly-textured region is likely to always be very large and thus the original spatial information is ignored.

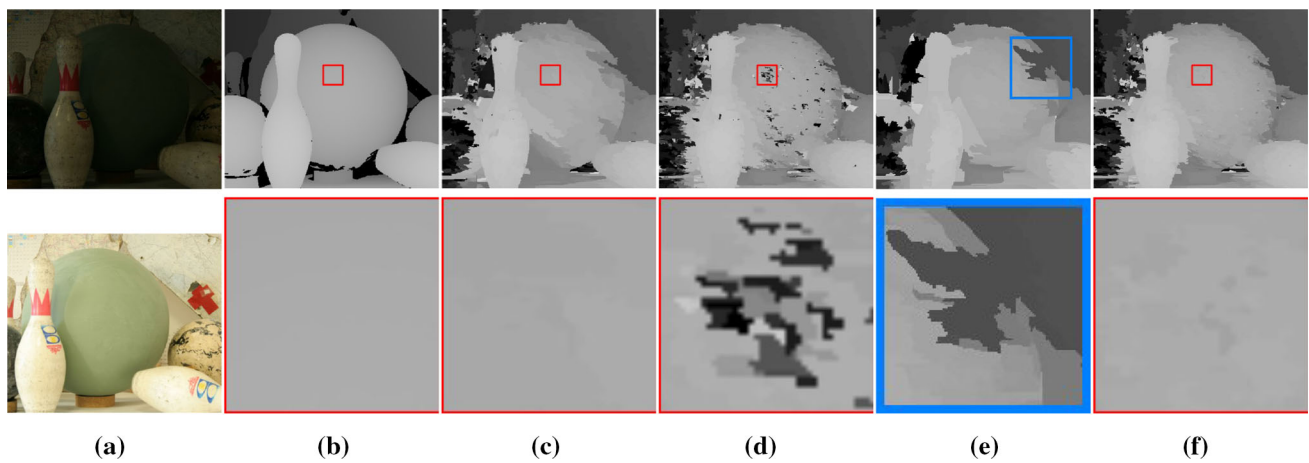


Fig. 4 Limitation of the local tree filter. **a** Input *left* (with illum2/exp0) and *right* (with illum2/exp2) images of *Bowling2* data set, respectively. **b** Ground-truth disparity map of **a**. **c–f** are disparity maps obtained from the proposed method with \mathcal{F} being the nonlocal tree filter Yang (2015), local tree filters, and the combined tree filter, respectively. Note that the local tree filter cannot handle low texture regions (where the nonlocal

filter performs well) when the filter kernel is relatively small. On the other hand, it cannot preserve all the edges when the filter kernel is large. The combination of the local and nonlocal tree filters is more reliable. The *numbers* presented under each disparity map are the corresponding percentage of error pixels

To be effective around textured regions, a low-pass filter can be performed on the same MST by defining the distance between two neighboring pixels (on MST) to be a constant value (e.g., “1”). In this case, the behavior of the resulting filter will be similar to the Gaussian filter around textured regions. Meanwhile, it still preserves the dominant color edges because MST has already automatically dragged away two dissimilar pixels that are close to each other in the spatial domain but indeed belong to two homogeneous regions. This filter is referred to as the *local tree filter* in this paper. It performs aggregation in two sequential steps. Let I denote the input image and $I^{A\uparrow}$ and I^A denote the aggregated values after the first and second steps, respectively. $I^{A\uparrow}$ is computed via recursive aggregation from the leaf nodes to the root node on the original image I while I^A is aggregated from the root node to the leaf nodes on $I^{A\uparrow}$:

$$I^{A\uparrow}(v) = I(v) + \sum_{P(v_c)=v} \exp(-\frac{0.5}{\sigma^2}) \cdot I^{A\uparrow}(v_c), \quad (24)$$

$$I^A(v) = (1 - \exp(-\frac{1}{\sigma^2})) \cdot I^{A\uparrow}(v) + \exp(-\frac{0.5}{\sigma^2}) \cdot I^A(P(v)), \quad (25)$$

where $\sigma = 2.3$ is a constant used to control the filter kernel size and $P(v)$ is the parent of a node v . If node v is a leaf node (that has no child), then $I^{A\uparrow}(v) = I(v)$. $I^A(v) = I^{A\uparrow}(v)$ if v is the root node.

The disparity map obtained from the local tree filter is presented in Fig. 3g. Note that the noise in Fig. 3e, f is removed from (g).

As a local filter, its performance is low around large textureless regions (when the filter kernel is small) as can be seen in Fig. 4d. The nonlocal tree filter on the other hand is robust to the lack of textures as can be seen in Fig. 4c. We thus synergistically combine the matching cost obtained from the local and nonlocal tree filters with a simple summation operation:

$$C(p) = C_{local}(p) + C_{nonlocal}(p), \quad (26)$$

where $C_{local}(p)$ and $C_{nonlocal}(p)$ are the matching costs computed from Eq. (23) with \mathcal{F} being the local tree filter and nonlocal tree filter, respectively.

The corresponding disparity map is presented in Fig. 4f. Figure 5 quantitatively compares the performance of the local tree filter, nonlocal tree filter, and their combination. Note that the performance of the combined similarity measure always outperforms the local and nonlocal tree filters (Yang 2015). This is the main reason why the proposed similarity measure outperforms other bilateral filter-based similarity measures like ANCC (Heo et al. 2011). On the other hand, a direct combination of the nonlocal tree filter and the Gaussian filter will be more robust to textured regions and can potentially improve the performance under illumination changes as can be seen from Fig. 5b. However, as discussed in Sect. 2.4, it will blur the depth edges and thus degrade the reconstruction accuracy near discontinuities. The local tree filter can better preserve edges and thus can obtain higher reconstruction accuracy near discontinuities when integrated with the original nonlocal tree filter.

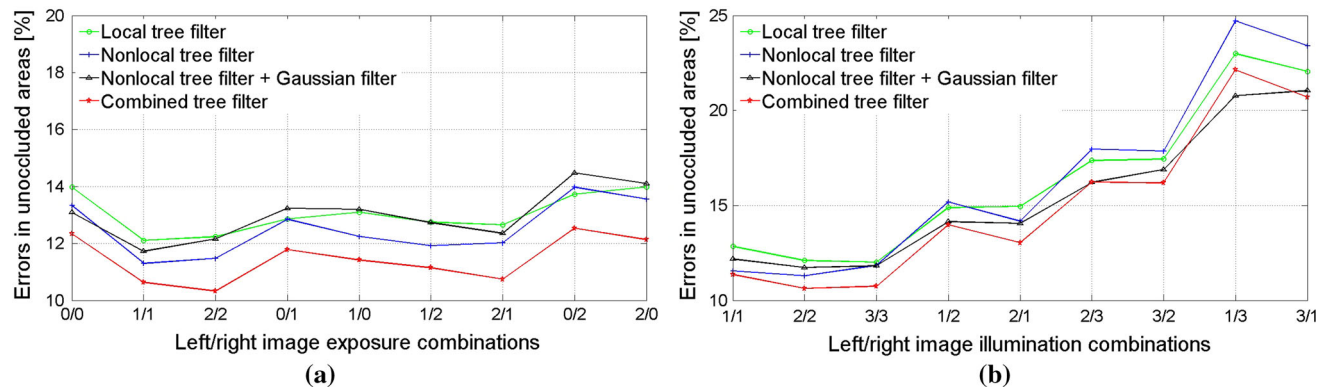


Fig. 5 Comparison of proposed similarity measure with different \mathcal{F} : nonlocal, local, nonlocal+Gaussian, and combined tree filters. Notice that nine stereo images (*Aloe*, *Art*, *Bowling2*, *Cloth4*, *Dolls*, *Lampshade1*, *Laundry*, *Moebius*, and *Rocks1*) are used with different

radiometric variations. **a, b** Present the average errors under different exposure combination and different lighting source combination, respectively. As can be seen, the combined tree filter can improve the performance on radiometric variations

3 Experimental Results

We evaluated the proposed method using Middlebury test data sets (Scharstein and Szeliski 2002a). Similar to Hirschmuller and Scharstein (2009) and Heo et al. (2011), a total of nine data sets were used (including *Aloe*, *Art*, *Bowling2*, *Cloth4*, *Dolls*, *Lampshade1*, *Laundry*, *Moebius* and *Rocks1* data sets). These various data sets can provide different texture conditions. Each of them consists of three different illuminations (indexed by “illum 1”, “illum 2”, and “illum 3”) and three different exposures (indexed by “exp 0”, “exp 1”, and “exp 2”), as well as the ground-truth disparity maps. In our experiments, the exposure was set to “exp 1” when testing on illumination changes, and the illumination was set to “illum 2” when testing different exposure changes. As a result, we had a total of nine experiment settings for both illumination and exposure changes. The same as Hirschmuller and Scharstein (2009) and Heo et al. (2011), the downsampled versions of all data sets (obtained using a downsampling factor of 3) were used and the parameters were all trained using the *Aloe* data set with no illumination or exposure changes. The trained parameters were used for all the other data sets and other illumination/exposure settings.

We quantitatively compared our similarity measure with the state-of-the-art measures including NCC, Census (Zabih and Woodfill 1994), ANCC (Heo et al. 2011), and MI+SIFT (Heo et al. (2013)) on stereo correspondence estimation. The source codes provided by the authors of ANCC (Heo et al. 2011) and MI+SIFT (Heo et al. 2013) were used to compute the corresponding disparity maps and measure the computational cost. MI+SIFT (Heo et al. (2013)) used the boosting strategy between depth estimation with color consistency. To make sure that the conducted experiments evaluated only the performance on the proposed similarity measures, cost aggregation, disparity optimization, and refinement steps that are

normally required in a stereo algorithm were excluded. This means that a winner-takes-all disparity selection scheme was directly applied to the matching cost volume obtained from different similarity measures to compute the corresponding disparity maps. The errors presented in this paper were only measured in unoccluded areas as percentages of error pixels. The same as the Middlebury benchmark (Scharstein and Szeliski 2002a), the disparity error threshold was set to one pixel throughout the experiments. To evaluate only the performance of the similarity measure presented in Heo et al. (2013), we used the disparity map computed directly from the initial matching costs (obtained from MI combined with the SIFT operator) and denoted it *MI+SIFT*. However, this matching cost is not the major contribution of Heo et al. (2013). As a result, we also compared with the complete solution published by Heo et al. (2013) (using the default parameters) but explicitly excluded the MRF optimization step (which is not included in the other similarity measures). We denoted this method *MI+SIFT(+)*. It normally converges in 3–4 iterations and thus we fixed the number of iterations to four. Unlike *MI+SIFT*, *MI+SIFT(+)* includes the SCHE stereo images generation, plane-fitting, and occlusion handling strategies.

3.1 Exposure Changes

Figure 6a quantitatively evaluates the performance of different similarity measures under different exposure combinations. As the exposure variations are more likely to change the image intensity/color globally, the performance of all the evaluated methods is relatively high. Census transform (blue curve) uses the ordering relationship between a pixel of interest and the other pixels in a neighborhood, and thus is robust to both linear and nonlinear illumination changes. However, it is not robust to noise. NCC (cyan curve) is more

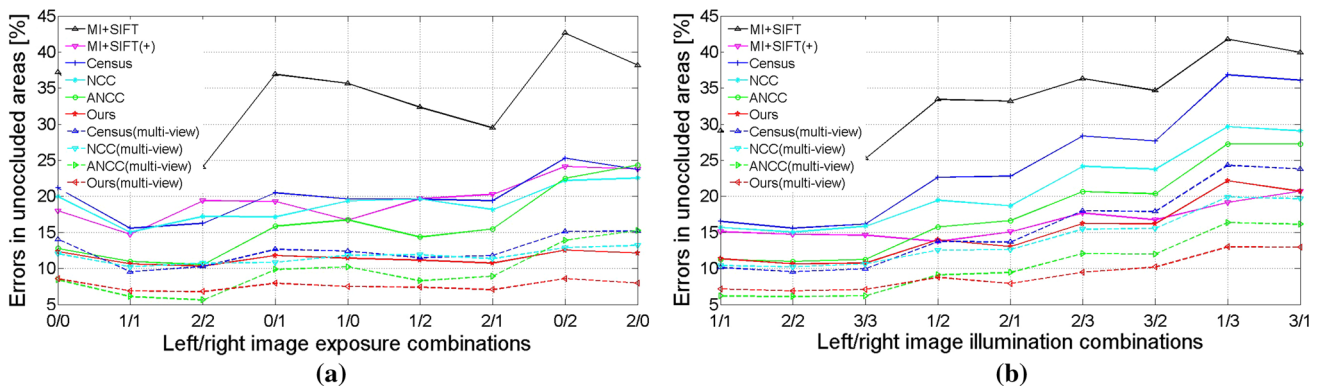


Fig. 6 Quantitative evaluation of different similarity measures under different exposure and illumination combinations on nine Middlebury data sets. The average percentages of error pixels are presented. The

red curves demonstrate that the proposed method outperforms the state-of-the-art under exposure/illumination changes in both binocular and multiview stereo matching (Color figure online)

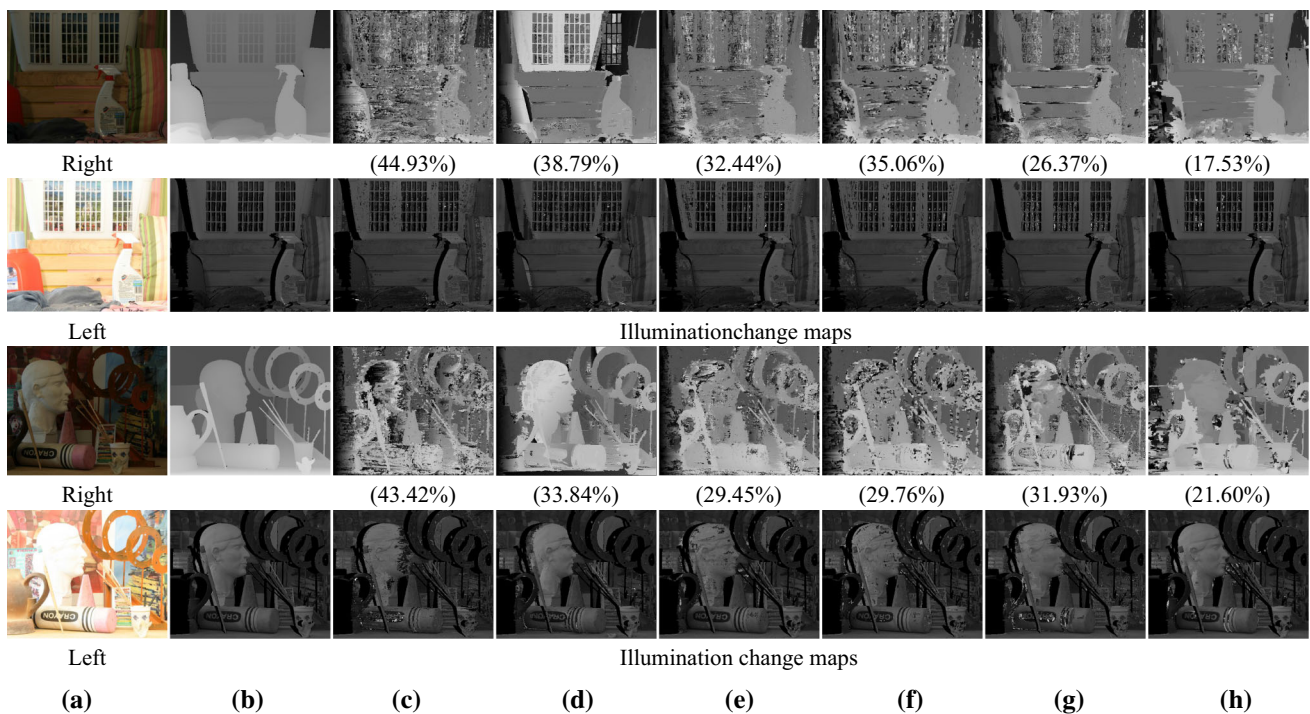


Fig. 7 Visual evaluation using *Laundry* and *Art* data sets under exposure changes. **a** Left image (with illum2/exp2) and right image (with illum2/exp0). **b** is the ground-truth disparity map and the corresponding illumination change map between left and right images. **c–h** are the disparity maps (top) and the corresponding illumination change

maps (bottom) obtained from MI+SIFT, MI+SIFT(+) (Heo et al. 2013), Census, NCC, ANCC (Heo et al. 2011), and the proposed measure, respectively. The numbers presented under the disparity maps are the corresponding percentages of error pixels

robust to noise but suffers from the fattening effect (which blurs depth edges). As a result, it does not have higher performance (comparing to Census) under global illumination changes. By introducing the bilateral weights, ANCC Heo et al. (2011) can significantly reduce the reconstruction errors around depth edges under relatively small or no exposure changes. However, its performance is relatively low under severe exposure differences (e.g., “exp 0/2” or “exp 2/0” in Fig. 6a). MI+SIFT (Heo et al. 2013) is also robust to global

radiometric changes, but the disparity maps obtained from this measure are noisy especially around low texture regions as shown in Fig. 7d. This is due to the use of SIFT which is vulnerable to the lack of textures. As a result, the reconstruction error is relatively high on average as shown in the black curve in Fig. 6a. Unlike MI+SIFT, MI+SIFT(+) includes the SCHE stereo images generation, plane-fitting, and occlusion handling strategies which are effective to noise. As a result, MI+SIFT(+) clearly outperformed MI+SIFT. The proposed

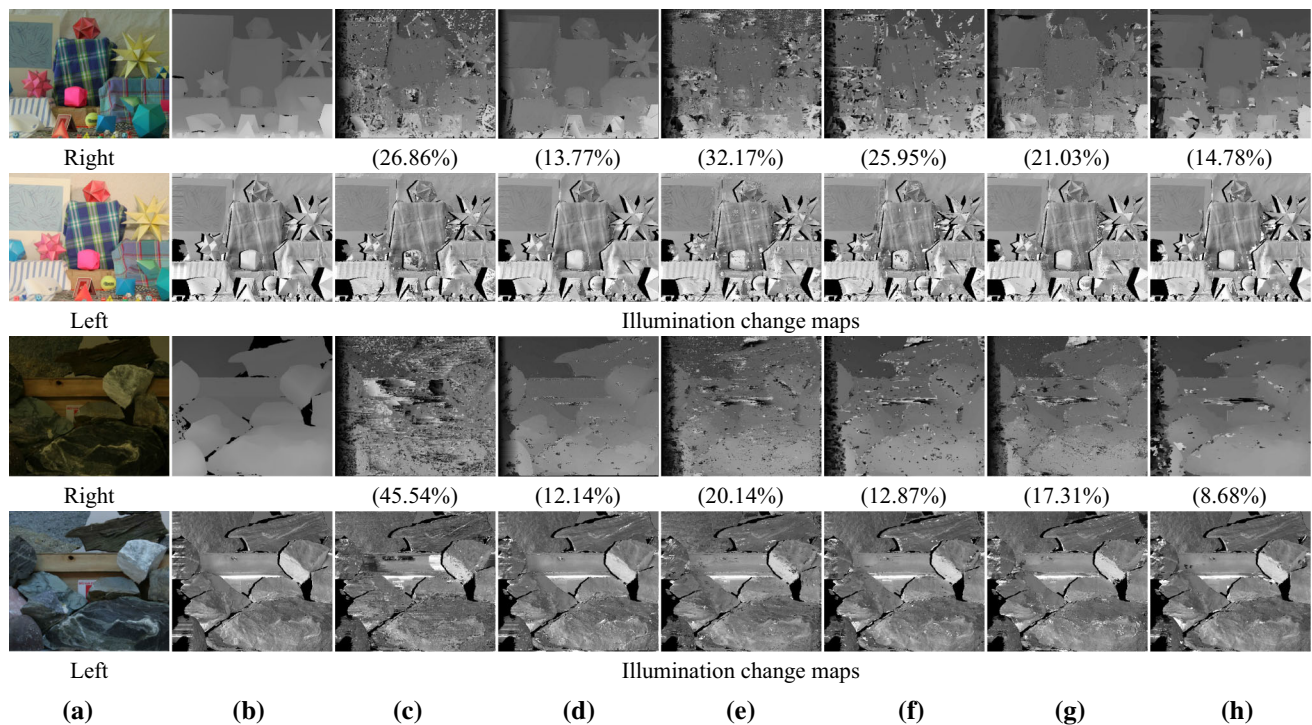


Fig. 8 Visual evaluation using *Moebius* and *Rocks1* data sets under lighting source changes. **a** Left image (with illum3/exp1) and right image (with illum1/exp1). **b** Ground-truth disparity map and the corresponding illumination change map between left and right images. **c–h** Disparity maps (top) and the corresponding illumination change

maps (bottom) obtained from MI+SIFT, MI+SIFT+ (Heo et al. 2013), Census, NCC, ANCC (Heo et al. 2011), and the proposed measure, respectively. The numbers presented under the disparity maps are the corresponding percentages of error pixels

similarity measure outperformed the others under all exposure combinations as can be seen in Fig. 6a. Unlike the other measures, the proposed similarity measure maintains its performance with the increase of the exposure difference, and thus is clearly more robust than the others. The corresponding disparity maps (of the *Laundry* and *Art* data sets) obtained from different similarity measures are presented in Fig. 7 for visual evaluation.

3.2 Light Source Changes

Figure 6b quantitatively evaluates the performance of different similarity measures under different illumination combinations. As can be seen, the errors of all similarity measures increased rapidly with increasing illumination difference. This is because existing similarity measures are invariant to only some specific illumination changes. These illumination assumptions are normally valid under small illumination changes but are likely to be violated under severe illumination changes.

MI is only effective for global radiometric differences. The combination of MI and SIFT (Heo et al. 2013) increased the robustness to local radiometric differences, but the reconstruction error was still high because SIFT is vulnerable to

the lack of textures. Meanwhile, the local order of intensities is likely to be distorted under local radiometric changes; and thus the performance of Census decreased rapidly under severe illumination changes. NCC and ANCC (Heo et al. 2011) had relatively larger errors under severe local radiometric differences as well. The performance of the proposed method was again the highest under different illumination combinations as can be seen in Fig. 6b. It clearly outperformed the others especially under severe illumination changes. Figure 8 presents the disparity maps (of *Moebius* and *Rocks1* data sets) obtained from different similarity measures for visual evaluation.

3.3 Comparison of Runtime

Table 1 presents the runtime of stereo algorithms using different similarity measures. The experiments were conducted on a 3.2 GHz Intel Core i7 CPU. As can be seen, the proposed similarity measure is the fastest. Also note that Census¹ and ANCC all are based on local image patches; and thus the computational complexity is linear in patch size and increases with respect to the image resolution. On the other

¹ The trained patch size for Census transform is 19×19 .

Table 1 The exact runtime (in s) on the *Aloe* data set

	Resolution	Census	MI+SIFT (Heo et al. 2013)	ANCC (Heo et al. 2011)	Proposed
Runtime	ThirdSize	10.1	18.8	161	2.94
	Full	2504			89.7

The runtime of ANCC and MI+SIFT was computed using the source codes published by the authors. The memory cost of ANCC and MI+SIFT is not affordable on a standard PC (and is estimated to be around 27 GB) when the full-resolution *Aloe* images (1282×1110×210, where 210 is the disparity search range) are used. As a result, the corresponding runtime is unknown and is not presented in the table

hand, the computational complexity of the proposed measure is independent of the filter kernel size. Given the trend toward higher-resolution images, which will correspondingly require higher filter kernel sizes, the $O(1)$ computational complexity makes the described similarity measure future-proof.

3.4 Quantitative Evaluation of Multiview Stereo

The same nine Middlebury test data sets (Scharstein and Szeliski 2002a) were used to evaluate the performance of the proposed measure for multi-view stereo matching.

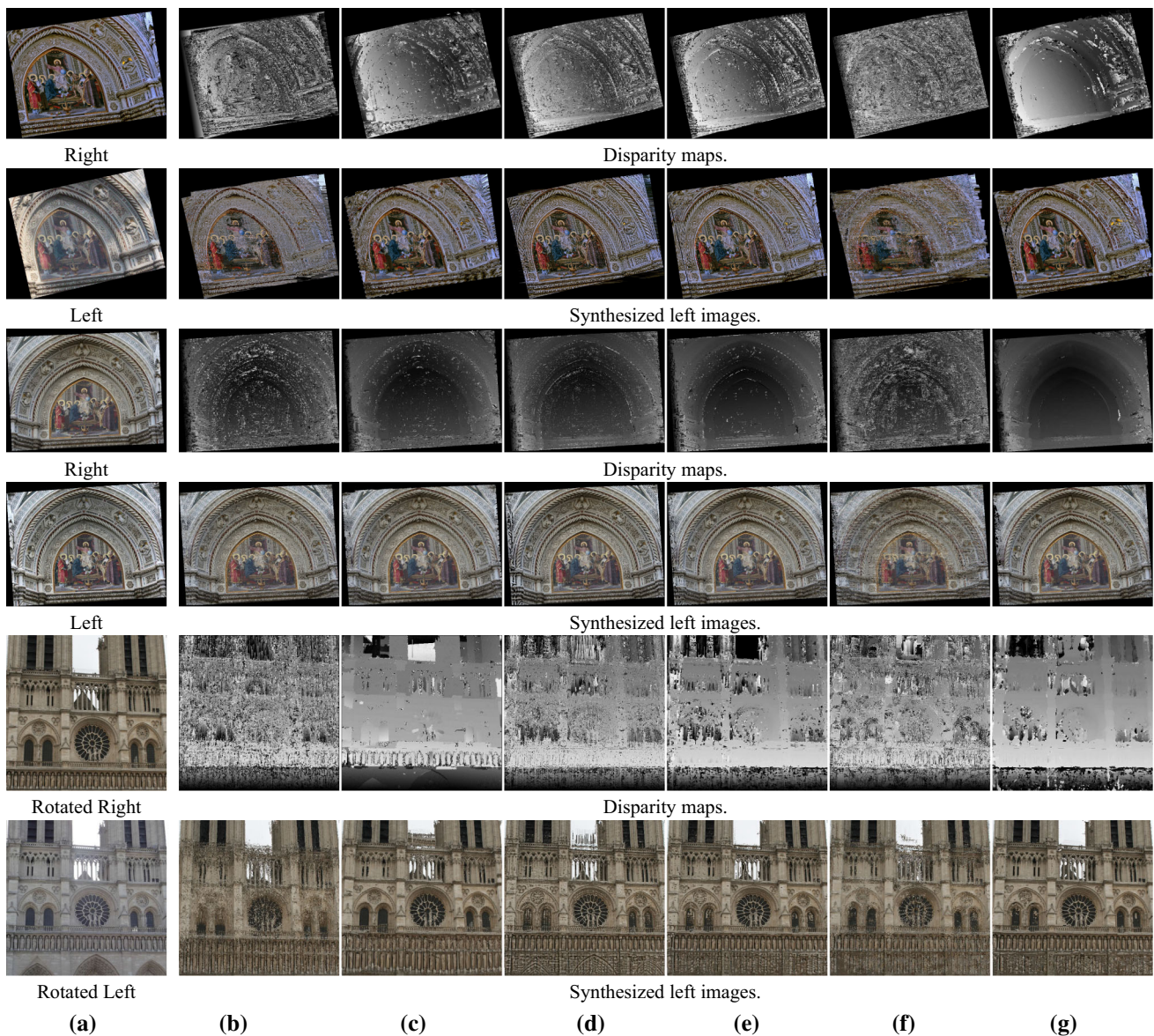


Fig. 9 Visual evaluation using two stereo pairs downloaded from the Flickr website. **a** Rectified input stereo image pairs. **b–g** are the disparity maps and the corresponding synthesized left images obtained from

MI+SIFT, MI+SIFT(+) (Heo et al. 2013), Census, NCC, ANCC (Heo et al. 2011), and the proposed measure, respectively

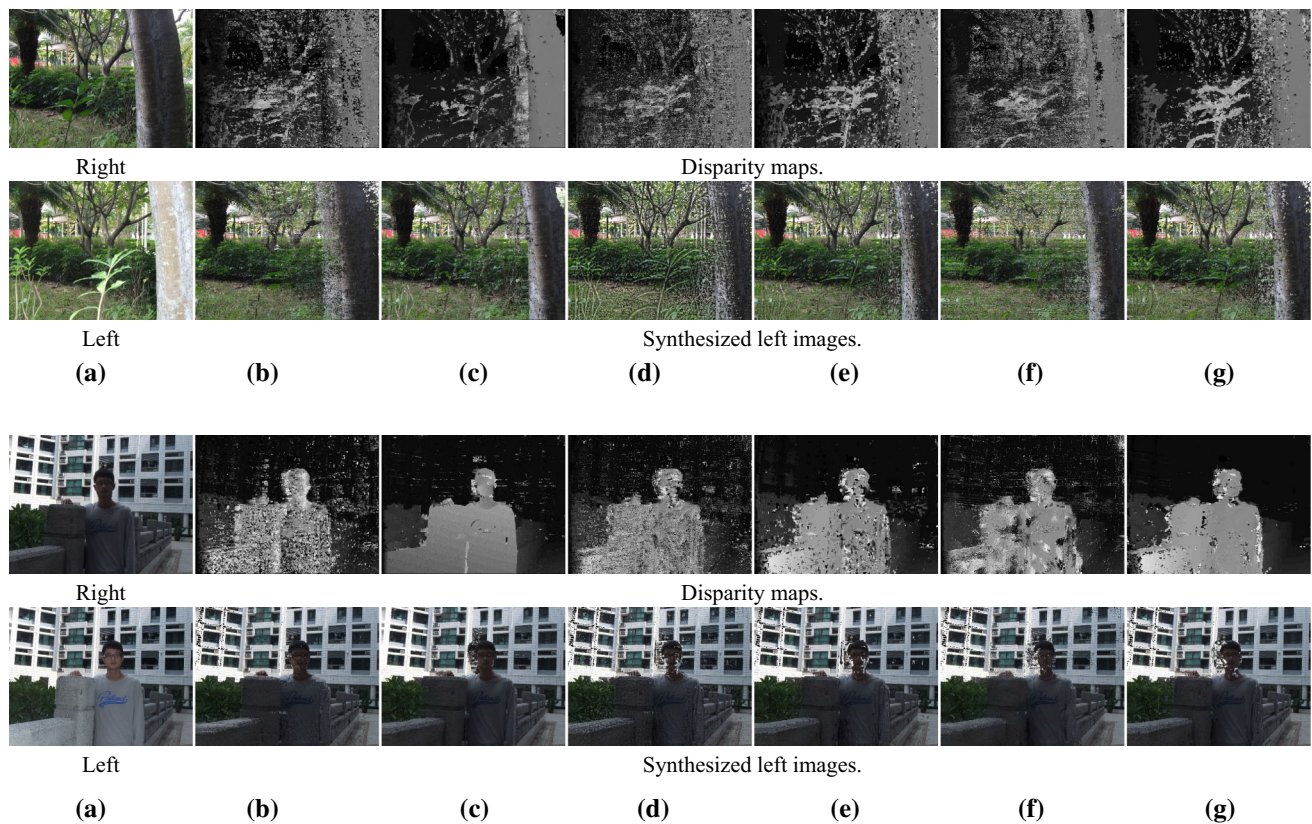


Fig. 10 Visual evaluation using images captured with and without flash. **a** *Left* and *right* images. The *left* images were captured with flash and the *right* images were without flash. **b–g** are the disparity maps

and the corresponding synthesized *left* images obtained from MI+SIFT, MI+SIFT(+) (Heo et al. 2013), Census, NCC, ANCC (Heo et al. 2011), and the proposed measure, respectively

The same as the binocular stereo evaluation, the “view1” image was used as the reference/left image, and “view2” – “view6” were mapped to “view1” to compute the matching cost. The sum of the matching cost computed from every image pair was used to compute the final disparity map. The same lighting/exposure conditions were used for “view2”–“view6”. Similarly, the exposure was set to “exp 1” when testing illumination changes, and the illumination was set to “illum 2” when testing different exposure changes. Due to memory issues, the comparison with MI+SIFT (Heo et al. 2013) in multiview stereo was ignored. The quantitative comparison of different similarity measures is presented in Fig. 6. It demonstrates that the multiview stereo consistently outperformed the binocular stereo and the proposed measure had the highest performance with radiometric changes.

3.5 Evaluation of Community Photo Collections

Community photo collections have emerged as a powerful new type of image dataset in recent years (Snaveely et al. 2006; Goesele et al. 2007). The images were collected from Internet photo sharing websites (e.g., Flickr and Google) and

captured by many photographers from a variety of different cameras, under varying / uncontrollable illumination and weather conditions. Normally, the lighting conditions and camera parameters of these corresponding images are much more disparate than Middlebury images.

The proposed measure was evaluated on two stereo pairs obtained from community photo collections, and the visual comparisons are presented in Fig. 9. The images in Fig. 9a were downloaded from Flickr. Each pair of images was rectified in advance before stereo matching.

The disparity maps obtained from different similarity measures are presented in Fig. 9b–f. State-of-the-art similarity measures are less robust to the tremendous local radiometric variations in these outdoor community collected photos; and thus there are large amounts of visible bad pixels in the corresponding disparity maps in Fig. 9b–e. The proposed measure can significantly reduce these errors as shown in Fig. 9f. The ground-truth disparities are not available, and thus the synthesized images were presented below the corresponding disparity maps for visual evaluation. They were computed from the right image based on the left disparity maps obtained from different similarity measures. Visual comparison with the rectified left image shows that most of

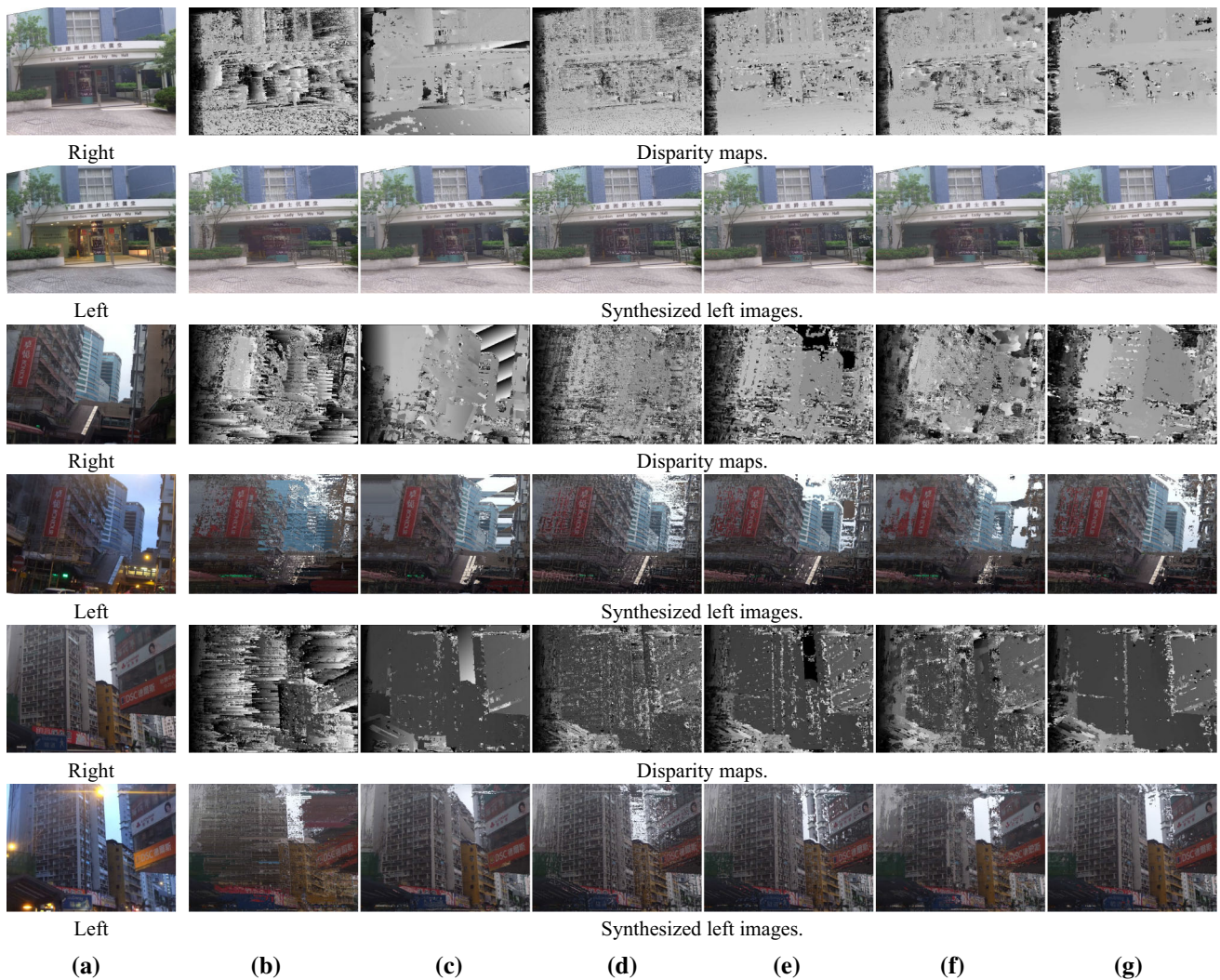


Fig. 11 Visual evaluation using stereo pairs captured under different times during the day. **a** Rectified input stereo image pairs. **b–g** are the disparity maps and the corresponding synthesized *left* images obtained

from MI+SIFT, MI+SIFT(+) (Heo et al. 2013), Census, NCC, ANCC (Heo et al. 2011), and the proposed measure, respectively

the disparity values obtained from the proposed measure are accurate.

Figure 10 presents the experimental results of real outdoor scenes captured with and without flash. The left images were captured with additional flash conditions, while the right images were without flash. The foreground objects are relatively far away from the background; and thus the flash only affects the foregrounds (e.g., leaves in *Plant* dataset). The disparity maps and synthesized left images generated from different measures are presented in Fig. 10b–f. Note that the proposed metric is more robust to these LOCAL illumination changes.

The left and right images in Fig. 11 were captured at different times of day when the sunlight is different. Similar to the previous observations, the proposed measure was more robust to the radiometric variations in these compli-

cated scenes and had the ability to reduce the errors caused by large local illumination changes.

4 Conclusion

A similarity measure for stereo correspondence estimation is proposed. It is invariant to both global and local affine illumination changes. It outperforms the state of the art in terms of both accuracy and speed on the Middlebury benchmark. The proposed similarity measure can be further extended for illumination changes undergoing more complex transforms. A simple extension is changing the affine transform in Eq. (3) to polynomial transform. We leave this problem for future research. We are also planning to combine the similarity measures computed from both the original image

and its logarithm in a way like ANCC (Heo et al. 2011) to obtain a more robust similarity measure in the near future. Preprocessing techniques like dynamic histogram warping (Cox et al. 1995) are likely to improve the quality of our method and will be analyzed as well.

Acknowledgments This work was supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CityU 21201914).

Appendix

Appendix 1: Derivation of Eq. 12

The matching cost measured from two corresponding pixels p and p' in two grayscale images I_L and I_R is:

$$\begin{aligned} & \sum_{q \in I_L} \mathcal{W}(p, q) \cdot (I_L(q) - a_p \cdot I_{R,\Delta}(q) - b_p)^2 \\ &= \sum_{q \in I_L} \mathcal{W}(p, q) (a_p^2 I_{R,\Delta}(q)^2 + 2a_p b_p I_{R,\Delta}(q) \\ &\quad - 2a_p I_L(q) I_{R,\Delta}(q) - 2b_p I_L(q) + b_p^2 + I_L(q)^2) \\ &= a_p^2 \sum_{q \in I_L} \mathcal{W}(p, q) I_{R,\Delta}(q)^2 + 2a_p b_p \sum_{q \in I_L} \mathcal{W}(p, q) I_{R,\Delta}(q) \\ &\quad - 2a_p \sum_{q \in I_L} \mathcal{W}(p, q) I_L(q) I_{R,\Delta}(q) \\ &\quad - 2b_p \sum_{q \in I_L} \mathcal{W}(p, q) I_L(q) + b_p^2 + \sum_{q \in I_L} \mathcal{W}(p, q) I_L(q)^2 \\ &= a_p^2 \cdot \mathcal{F}_{I_{R,\Delta} \cdot I_{R,\Delta}}(p) + 2a_p b_p \cdot \mathcal{F}_{I_{R,\Delta}}(p) \\ &\quad - 2a_p \cdot \mathcal{F}_{I_L \cdot I_{R,\Delta}}(p) - 2b_p \cdot \mathcal{F}_{I_L}(p) + b_p^2 + \mathcal{F}_{I_L \cdot I_L}(p). \end{aligned} \tag{27}$$

Appendix 2: Derivation of Eq. 22

Similar to Eq. 4, we can extend Eq. 20 for color images as follows:

$$\tilde{A} \cdot \tilde{\mathcal{X}} = \tilde{B}. \tag{28}$$

$\tilde{\mathcal{X}}$ is defined in Eq. 21, and

$$\tilde{A} = \begin{pmatrix} \cdots \mathcal{W}(p, q) \cdot I_{R,\Delta}^1(q) \cdots \\ \cdots \mathcal{W}(p, q) \cdot I_{R,\Delta}^2(q) \cdots \\ \cdots \mathcal{W}(p, q) \cdot I_{R,\Delta}^3(q) \cdots \\ \cdots \mathcal{W}(p, q) \cdot 1 \cdots \end{pmatrix}^T, \tag{29}$$

and

$$\tilde{B} = \begin{pmatrix} \cdots \mathcal{W}(p, q) \cdot I_L^1(q) \cdots \\ \cdots \mathcal{W}(p, q) \cdot I_L^2(q) \cdots \\ \cdots \mathcal{W}(p, q) \cdot I_L^3(q) \cdots \end{pmatrix}^T. \tag{30}$$

The linear system presented in Eq. 28 can be rewritten as:

$$\tilde{A}^T \tilde{A} \cdot \tilde{\mathcal{X}} = \tilde{A}^T \tilde{B}, \tag{31}$$

where

$$\begin{aligned} & \tilde{A}^T \tilde{A} \\ &= \begin{pmatrix} \mathcal{F}_{I_{R,\Delta}^1 \cdot I_{R,\Delta}^1}(p) & \mathcal{F}_{I_{R,\Delta}^1 \cdot I_{R,\Delta}^2}(p) & \mathcal{F}_{I_{R,\Delta}^1 \cdot I_{R,\Delta}^3}(p) & \mathcal{F}_{I_{R,\Delta}^1}(p) \\ \mathcal{F}_{I_{R,\Delta}^2 \cdot I_{R,\Delta}^1}(p) & \mathcal{F}_{I_{R,\Delta}^2 \cdot I_{R,\Delta}^2}(p) & \mathcal{F}_{I_{R,\Delta}^2 \cdot I_{R,\Delta}^3}(p) & \mathcal{F}_{I_{R,\Delta}^2}(p) \\ \mathcal{F}_{I_{R,\Delta}^3 \cdot I_{R,\Delta}^1}(p) & \mathcal{F}_{I_{R,\Delta}^3 \cdot I_{R,\Delta}^2}(p) & \mathcal{F}_{I_{R,\Delta}^3 \cdot I_{R,\Delta}^3}(p) & \mathcal{F}_{I_{R,\Delta}^3}(p) \\ \mathcal{F}_{I_{R,\Delta}^1}(p) & \mathcal{F}_{I_{R,\Delta}^2}(p) & \mathcal{F}_{I_{R,\Delta}^3}(p) & 1 \end{pmatrix} \end{aligned} \tag{32}$$

and

$$\tilde{A}^T \tilde{B} = \begin{pmatrix} \mathcal{F}_{I_{R,\Delta}^1 \cdot I_L^1}(p) & \mathcal{F}_{I_{R,\Delta}^1 \cdot I_L^2}(p) & \mathcal{F}_{I_{R,\Delta}^1 \cdot I_L^3}(p) \\ \mathcal{F}_{I_{R,\Delta}^2 \cdot I_L^1}(p) & \mathcal{F}_{I_{R,\Delta}^2 \cdot I_L^2}(p) & \mathcal{F}_{I_{R,\Delta}^2 \cdot I_L^3}(p) \\ \mathcal{F}_{I_{R,\Delta}^3 \cdot I_L^1}(p) & \mathcal{F}_{I_{R,\Delta}^3 \cdot I_L^2}(p) & \mathcal{F}_{I_{R,\Delta}^3 \cdot I_L^3}(p) \\ \mathcal{F}_{I_L^1}(p) & \mathcal{F}_{I_L^2}(p) & \mathcal{F}_{I_L^3}(p) \end{pmatrix}. \tag{33}$$

Appendix 3: Derivation of Eq. 23

The matching cost for color images is:

$$\begin{aligned} & \sum_{c1=1}^3 \sum_{q \in I_L} \mathcal{W}(p, q) \cdot (I_L^{c1}(q) - \sum_{c2=1}^3 a_p^{c1c2} I_{R,\Delta}^{c2}(q) - b_p^{c1})^2 \\ &= \sum_{c1=1}^3 \sum_{q \in I_L} \mathcal{W}(p, q) \cdot ((I_L^{c1}(q))^2 + (\sum_{c2=1}^3 a_p^{c1c2} I_{R,\Delta}^{c2}(q))^2 \\ &\quad + (b_p^{c1})^2 - 2b_p^{c1} I_L^{c1}(q) - 2 \sum_{c2=1}^3 a_p^{c1c2} I_{R,\Delta}^{c2}(q) I_L^{c1}(q) \\ &\quad + 2 \sum_{c2=1}^3 a_p^{c1c2} b_p^{c1} I_{R,\Delta}^{c2}(q)) \\ &= \sum_{c1=1}^3 (\sum_{q \in I_L} \mathcal{W}(p, q) \cdot I_L^{c1}(p) I_L^{c1}(p)) \\ &\quad + \sum_{c3=1}^3 \sum_{c2=1}^3 (\sum_{c1=1}^3 a_p^{c1c2} \cdot a_p^{c1c3}) (\sum_{q \in I_L} \mathcal{W}(p, q) \\ &\quad \cdot I_{R,\Delta}^{c2}(p) I_{R,\Delta}^{c3}(p)) \\ &\quad + \sum_{c1=1}^3 (b_p^{c1})^2 - 2 \sum_{c1=1}^3 b_p^{c1} (\sum_{q \in I_L} \mathcal{W}(p, q) \cdot I_L^{c1}(p)) \end{aligned}$$

$$\begin{aligned}
& -2 \sum_{c1=1}^3 \sum_{c2=1}^3 a_p^{c1c2} \left(\sum_{q \in I_L} \mathcal{W}(p, q) \cdot I_L^{c1}(p) I_{R,\Delta}^{c2}(p) \right) \\
& + 2 \sum_{c2=1}^3 \left(\left(\sum_{c1=1}^3 a_p^{c1c2} \cdot b_p^{c1} \right) \cdot \left(\sum_{q \in I_L} \mathcal{W}(p, q) \cdot I_{R,\Delta}^{c2}(p) \right) \right) \\
& = \sum_{c1=1}^3 \mathcal{F}_{I_L^{c1}, I_L^{c1}}(p) + \sum_{c3=1}^3 \sum_{c2=1}^3 \sum_{c1=1}^3 \left(\sum_{p} a_p^{c1c2} \cdot a_p^{c1c3} \right) \mathcal{F}_{I_{R,\Delta}^{c2}, I_{R,\Delta}^{c3}}(p) \\
& + \sum_{c1=1}^3 (b_p^{c1})^2 - 2 \sum_{c1=1}^3 b_p^{c1} \mathcal{F}_{I_L^{c1}}(p) \\
& - 2 \sum_{c1=1}^3 \sum_{c2=1}^3 a_p^{c1c2} \mathcal{F}_{I_L^{c1}, I_{R,\Delta}^{c2}}(p) \\
& + 2 \sum_{c2=1}^3 \left(\left(\sum_{c1=1}^3 a_p^{c1c2} \cdot b_p^{c1} \right) \cdot \mathcal{F}_{I_{R,\Delta}^{c2}}(p) \right) \quad (34)
\end{aligned}$$

References

- Brown, M., Burschka, D., & Hager, G. D. (2003). Advances in computational stereo. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 25(8), 993–1008.
- Chia, A., Zhuo, S., Gupta, R., Tai, Y., Cho, S., Tan, P., et al. (2011). Semantic colorization with internet images. *ACM Transaction on Graphics*, 30(6), 156:1–156:8.
- Cox, I., Roy, S., & Hingorani, S. (1995). Dynamic histogram warping of image pairs for constant image brightness. In *ICIP*.
- De-Maezdu, L., Mattocchia, S., Villanueva, A., & Cabeza, R. (2011). Linear stereo matching. In *ICCV, Nov 2011* (pp. 1708–1715).
- Ding, Y., Xiao, J., & Yu, J. (2011). Importance filtering for image retargeting. In *CVPR*.
- Egnal, G. (2000). Mutual information as a stereo correspondence measure. Technical Report MS-CIS-00-20. Computer and Information Science, University of Pennsylvania.
- Goesele, M., Snavely, N., Curless, B., Hoppe, H., & Seitz, S. (2007). Multi-view stereo for community photo collections. In *ICCV*.
- He, K., Rhemann, C., Rother, C., Tang, X., & Sun, J. (2011). A global sampling method for alpha matting. In *CVPR, June 2011* (pp. 2049–2056).
- He, K., Sun, J., & Tang, X. (2011). Single image haze removal using dark channel prior. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 33(12), 2341–2353.
- He, K., Sun, J., & Tang, X. (2013). Guided image filtering. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 35(6), 1397–1409.
- Heo, Y., Lee, K., & Lee, S. (2009). Mutual information-based stereo matching combined with sift descriptor in log-chromaticity color space. In *CVPR, 2009* (pp. 445–452).
- Heo, Y., Lee, K., & Lee, S. (2011). Robust stereo matching using adaptive normalized cross-correlation. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 33(4), 807–822.
- Heo, Y., Lee, K., & Lee, S. (2013). Joint depth map and color consistency estimation for stereo images with different illuminations and cameras. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 35(5), 1094–1106.
- Hestenes, M., & Stiefel, D. (1952). Methods of conjugate gradients for solving linear systems. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 49, 409–436.
- Hirschmuller, H. (2008). Stereo processing by semiglobal matching and mutual information. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 30(2), 328–341.
- Hirschmuller, H., & Scharstein, D. (2009). Evaluation of stereo matching costs on images with radiometric differences. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 31(9), 1582–1599.
- Hosni, A., Rhemann, C., Bleyer, M., Rother, C., & Gelautz, M. (2013). Fast cost-volume filtering for visual correspondence and beyond. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 35(2), 504–511.
- Kim, J., Kolmogorov, V., & Zabih, R. (2003). Visual correspondence using energy minimization and mutual information. In *ICCV*.
- Klaus, A., Sormann, M., & Karner, K. (2006). Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure. In *ICPR, 2006* (pp. 15–18).
- Levin, A., Lischinski, D., & Weiss, Y. (2008). A closed-form solution to natural image matting. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 30(2), 228–242.
- Scharstein, D., & Szeliski, R. (2002a). Middlebury stereo datasets. <http://vision.middlebury.edu/stereo/data/>.
- Scharstein, D., & Szeliski, R. (2002b). A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47, 7–42.
- Snavely, N., Seitz, S. M., & Szeliski, R. (2006). Photo tourism: Exploring photo collections in 3d. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 25(3), 835–846.
- Wang, L., & Yang, R. (2011). Global stereo matching leveraged by sparse ground control points. In *CVPR, June 2011* (pp. 3033–3040).
- Wang, L., Yang, R., & Davis, J. (2007). Brdf invariant stereo using light transport constancy. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 29(9), 1616–1626.
- Yang, Q., Wang, L., & Ahuja, N. (2010). A constant-space belief propagation algorithm for stereo matching. In *CVPR, 2010* (pp. 1458–1465).
- Yang, Q., Wang, L., Yang, R., Stewenius, H., & Nister, D. (2009). Stereo matching with color-weighted correlation, hierarchical belief propagation and occlusion handling. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 31(3), 492–504.
- Yang, Q. (2015). Stereo matching using tree filtering. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 37(4), 834–846.
- Yoon, K.-J., & Kweon, I.-S. (2006). Adaptive support-weight approach for correspondence search. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 28(4), 650–656.
- Zabih, R., & Woodfill, J. (1994). Non-parametric local transforms for computing visual correspondence. In *ECCV, 1994* (pp. 151–158).
- Zhu, S., Zhang, L., & Jin, H. (2012). A locally linear regression model for boundary preserving regularization in stereo matching. In *Proceedings of the 12th European conference on computer vision—Volume Part V, ser. ECCV'12* (pp. 101–115). Berlin: Springer-Verlag.
- Zomet, A., & Peleg, S. (2002). Multi-sensor super resolution. In *WACV*.