LETTER

An Efficient Wide-Baseline Dense Matching Descriptor

Yanli WAN[†a)], **Zhenjiang MIAO**[†], **Zhen TANG**[†], **Lili WAN**[†], ***Nonmembers***, ***and*** **Zhe WANG**[††], ***Student Member***

SUMMARY This letter proposes an efficient local descriptor for wide-baseline dense matching. It improves the existing Daisy descriptor by combining intensity-based Haar wavelet response with a new color-based ratio model. The color ratio model is invariant to changes of viewing direction, object geometry, and the direction, intensity and spectral power distribution of the illumination. The experiments show that our descriptor has high discriminative power and robustness.

key words: *dense matching, wide-baseline, DAISY, Haar wavelet, photometric color invariants*

1. Introduction

Dense matching is one of the most active research areas in computer vision. Over the last few years, a number of excellent short-baseline dense matching algorithms has been proposed. However, the wide-baseline dense matching faces much more challenging due to large perspective distortions. It is worth addressing in many pattern recognition and computer vision tasks.

Tola et al. proposed a fast descriptor (DAISY) [1] for wind-baseline dense matching. It significantly reduces computational cost by convolving gradient map to compute the bin values. It not only retains the robustness of existing descriptors, such as SIFT [2] and GLOH [3] which were designed for robustness to perspective and lighting changes in sparse wide-baseline matching, but also can be computed quickly at every single image pixel. Bay et al. [4] proposed a descriptor (SURF) based on an integral image to compute the histogram bins. Although this method was also computationally effective, all pixels in a regular region contribute equally to their respective bins which does away SIFT's spatial weighting scheme.

The above excellent descriptors are all based on intensity by transferring color to grey images. However, color also provides powerful information for matching tasks. If it is neglected, a very important source of distinction may be lost. Abdel-Hakin et al. [5] proposed a colored local invariant feature descriptor (CSIFT) with Gaussian invariance color model. It is more robust than the conventional SIFT with respect to color and photometrical variations. However, it depends on the changes of illumination color. Gevers et al. [6] proposed a color constant model $m_1 m_2 m_3$. It is invariant to the changes of viewing direction, surface orientation, illumination direction, illumination intensity, illumination color, except highlights.

In this paper, we propose an efficient local descriptor for wide-baseline dense matching. It has the following characteristics:

- The descriptor combines two different sub-descriptors based on the Haar wavelet response and a color invariant model respectively to improve the robustness and distinctiveness.
- In the first sub-descriptor, the advantages of DAISY and SURF are combined to improve the speed in the stage of pixel description.
- In the second sub-descriptor, a color model is proposed. It is invariant to the changes of viewing direction, highlights, illumination direction, illumination intensity, and illumination color.

The remainder of this paper is organized as follows. The detail of our descriptor is described in Sect. 2. Experimental results are presented and discussed in Sect. 3. Section 4 concludes the paper.

2. Our Local Descriptor

Our local descriptor is built based on two sub-descriptors: Haar wavelet response sub-descriptor, and color sub-descriptor. It is defined as:

$$F = [\omega H, \quad (1-\omega)C] \tag{1}$$

where H is a 100-dimension Haar wavelet response sub-descriptor, which improves DAISY descriptor by Haar wavelet response instead of the oriented gradient. C is a 75-dimension color sub-descriptor. ω is a weighting factor. Thus, Our descriptor has 175-dimension in total.

2.1 Sub-Descriptor Based on Haar Wavelet Response

The sub-descriptor based on the Haar wavelet response is built with the grey images. The Haar wavelet filter is a $t \times t$ box type convolution filters (Fig. 1 (a)). The two responses at pixel $m(x, y)$, which are written as d_x and d_y, are respectively convolved in x and y directions with the integral image $I_\Sigma(m)$ (Fig. 1 (b)). The integral image is used to improve computational effectiveness,and it is the sum of all pixels in

Manuscript received July 26, 2011.
Manuscript revised January 12, 2012.

[†]The authors are with Institute of Information Science, Beijing Jiaotong University, Beijing 100044, P.R. China.
[††]The author is with the School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, P.R. China.
a) E-mail: 07112067@bjtu.edu.cn

DOI: 10.1587/transinf.E95.D.2021

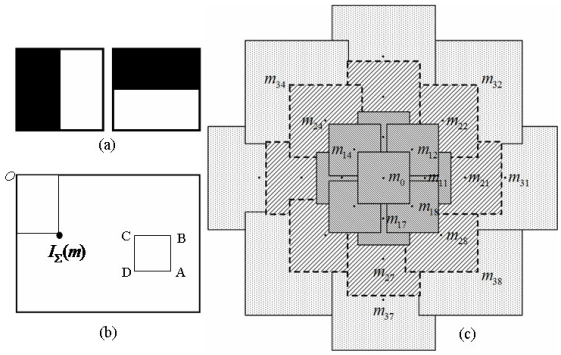

Fig. 1 (a) Haar wavelet filters. (b) Diagram for integral image. (c) Our description region.

the grey image $I(m)$ within a rectangular region formed by the origin O and point m:

$$I_\Sigma(m) = \sum_{i=0}^{i\leqslant x}\sum_{j=0}^{j\leqslant y} I(i,j) \tag{2}$$

The neighboring region of our descriptor is similar to DAISY structure (Fig. 1 (c)). The difference is that each circle sub-region is replaced by a square. It can greatly improve computational effectiveness of Haar wavelet response histogram in each squared sub-region.

Similar to the construction process of DAISY, Haar wavelet response sub-descriptor is also constructed by three steps: computing Haar wavelet response maps, convolving response maps with different Gaussian kernels, and constructing sub-descriptor with convolved response maps.

2.1.1 Haar Wavelet Response Maps Computation

We first compute 4 *Haar wavelet response maps*, which are written as $G_{r1}, G_{r2}, G_{r3}, G_{r4}$. These response maps are respectively acquired by response values $\{d_x, |d_x|, d_y, |d_y|\}$, where $|d_x|$ and $|d_y|$ are the absolute values of the Haar wavelet responses at x and y directions.

2.1.2 Response Maps Convolution

Each response map is then convolved several times with Gaussian kernels of different Σ values to obtain convolved response maps G_r^Σ for different sized regions. Supposing $G_r^{\Sigma_1}$ is a response map generated by convolving G_r with the smallest Gaussian kernel G_{Σ_1}, then $G_r^{\Sigma_2}$ can be obtained by a larger Gaussian kernel as:

$$G_r^{\Sigma_2} = G_{\Sigma_2} * G_r = G_\Sigma * G_{\Sigma_1} * G_r = G_\Sigma * G_r^{\Sigma_1} \tag{3}$$

where $\Sigma = \sqrt{\Sigma_2^2 - \Sigma_1^2}$. This consecutive convolutions can greatly reduce computational cost. Figure 2 shows the process of constructing those *convolved response maps*.

2.1.3 Sub-Descriptor Construction

As depicted by Fig. 1 (c), the sub-descriptor $H(m_0)$ based on

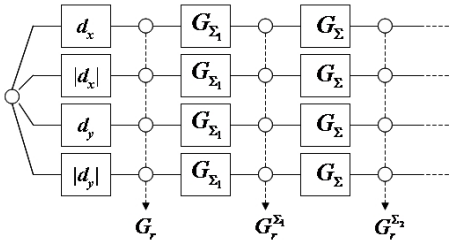

Fig. 2 Process of convolving Haar response maps.

the Haar wavelet responses can be defined as:

$$\begin{aligned} H(m_0) = [&\tilde{h}_{\Sigma_1}^T(m_0),\\ &\tilde{h}_{\Sigma_1}^T(m_{11}), \tilde{h}_{\Sigma_1}^T(m_{12}), \cdots, \tilde{h}_{\Sigma_1}^T(m_{18}),\\ &\tilde{h}_{\Sigma_2}^T(m_{21}), \tilde{h}_{\Sigma_2}^T(m_{22}), \cdots, \tilde{h}_{\Sigma_2}^T(m_{28}),\\ &\tilde{h}_{\Sigma_3}^T(m_{31}), \tilde{h}_{\Sigma_3}^T(m_{32}), \cdots, \tilde{h}_{\Sigma_3}^T(m_{38})] \end{aligned} \tag{4}$$

where $\tilde{h}_{\Sigma_1}^T(m_0)$ is a normalized vector in each histogram of $h_{\Sigma_1}^T(m_0)$ that represents a vector composed of the values at location m_0 in the response maps after convolution by a Gaussian kernel of standard deviation Σ_1.

$$h_{\Sigma_1}^T(m_0) = [G_{r1}^{\Sigma_1}(m_0), G_{r2}^{\Sigma_1}(m_0), G_{r3}^{\Sigma_1}(m_0), G_{r4}^{\Sigma_1}(m_0)] \tag{5}$$

where $G_{r1}^{\Sigma_1}$, $G_{r2}^{\Sigma_1}$, $G_{r3}^{\Sigma_1}$ and $G_{r4}^{\Sigma_1}$ denote the convolved Haar wavelet responses. The normalization is performed in each histogram independently, and it corresponds to each square sub-region in Fig. 1 (c). The sub-descriptor H at each pixel is composed of $4 * 25 = 100$ values, which are extracted from 25 locations and 4 responses.

2.2 Sub-Descriptor Based on Color Invariant Model

Since the raw color recorded by a camera is not reliable because of many factors, Therefore, the selected criteria of color model should be robust to varying illumination across the scene, and the changes in surface orientation of the object.

2.2.1 Reflectance Model

The changes in the illumination can be modeled by a diagonal-offset model [7]:

$$\begin{pmatrix} R^c \\ G^c \\ B^c \end{pmatrix} = \begin{pmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{pmatrix} \begin{pmatrix} R^u \\ G^u \\ B^u \end{pmatrix} + \begin{pmatrix} o_1 \\ o_2 \\ o_3 \end{pmatrix} \tag{6}$$

Based on the diagonal-offset model, five types of changes are categorized: light intensity changes ($a = b = c, o_1 = o_2 = o_3 = 0$), light intensity shifts($a = b = c = 1, o_1 = o_2 = o_3$), light intensity changes and shifts ($a = b = c, o_1 = o_2 = o_3$), light color changes ($a \neq b \neq c, o_1 = o_2 = o_3 = 0$), and light color changes and shifts ($a \neq b \neq c, o_1 \neq o_2 \neq o_3$). The surface reflectance $s(x, \lambda_C)$ can be given by [8]:

$$R^c = e^c(\lambda_R)s(x,\lambda_R) = ae^u(\lambda_R)s(x,\lambda_R) + A(\lambda_R),$$

$$G^c = e^c(\lambda_G)s(x,\lambda_G) = be^u(\lambda_G)s(x,\lambda_G) + A(\lambda_G), \quad (7)$$
$$B^c = e^c(\lambda_B)s(x,\lambda_B) = ce^u(\lambda_B)s(x,\lambda_B) + A(\lambda_B).$$

where $e^u(\lambda_C)$ is the color of the unknown light source, $e^c(\lambda_C)$ is the transformed color, $s(x,\lambda_C)$ is the surface reflectance, and $A(\lambda_C)$ is the term that models the diffuse light ($C \in \{R,G,B\}$).

2.2.2 Color Invariant Model

In order to make the color invariant to above five types of changes, we present the following color ratio model:

$$f_1 = \frac{(R_{x_o} - \mu_{R_{x_o}})(G_{x_i} - \mu_{G_{x_i}})}{(R_{x_i} - \mu_{R_{x_i}})(G_{x_o} - \mu_{G_{x_o}})},$$
$$f_2 = \frac{(B_{x_o} - \mu_{B_{x_o}})(R_{x_i} - \mu_{R_{x_i}})}{(B_{x_i} - \mu_{B_{x_i}})(R_{x_o} - \mu_{R_{x_o}})}, \quad (8)$$
$$f_3 = \frac{(G_{x_o} - \mu_{G_{x_o}})(B_{x_i} - \mu_{B_{x_i}})}{(G_{x_i} - \mu_{G_{x_i}})(B_{x_o} - \mu_{B_{x_o}})}.$$

where x_o and x_i denote the locations of the two neighboring pixels, $\mu_{C_{x_o}}$ is the mean in channel C over the rectangle area centered at x_o.

If the illuminant color is assumed to be locally constant (i.e. $e^u_{x_1}(\lambda_C) = e^u_{x_2}(\lambda_C) = \bar{e}^u_{x_1}(\lambda_C) = \bar{e}^u_{x_2}(\lambda_C)$), our model $f_1 f_2 f_3$ is independent to the above five types of changes by substituting Eq. (7) (R^c, G^c, B^c) into Eq. (8).

$$f_1 = \frac{(s(x_1,\lambda_R) - \bar{s}(x_1,\lambda_R))(s(x_2,\lambda_G) - \bar{s}(x_2,\lambda_G))}{(s(x_2,\lambda_R) - \bar{s}(x_2,\lambda_R))(s(x_1,\lambda_G) - \bar{s}(x_1,\lambda_G))},$$
$$f_2 = \frac{(s(x_1,\lambda_B) - \bar{s}(x_1,\lambda_B))(s(x_2,\lambda_R) - \bar{s}(x_2,\lambda_R))}{(s(x_2,\lambda_B) - \bar{s}(x_2,\lambda_B))(s(x_1,\lambda_R) - \bar{s}(x_1,\lambda_R))}, \quad (9)$$
$$f_3 = \frac{(s(x_1,\lambda_G) - \bar{s}(x_1,\lambda_G))(s(x_2,\lambda_B) - \bar{s}(x_2,\lambda_B))}{(s(x_2,\lambda_G) - \bar{s}(x_2,\lambda_G))(s(x_1,\lambda_B) - \bar{s}(x_1,\lambda_B))}.$$

From above equation it can be seen that f_1, f_2, f_3 only depend on the sensors and the surface albedo.

2.2.3 Color Invariant Sub-Descriptor Construction

Just as Haar wavelet response sub-descriptor, we first compute 3 *color ratio maps* G_{d1}, G_{d2}, G_{d3}. These color ratio maps are respectively acquired by our invariant model $|f_1|, |f_2|, |f_3|$ (Eq. 8) at center point and its neighboring points in a rectangle area. Each color ratio map G_d is then convolved several times with Gaussian kernels of different Σ values to obtain *convolved color ratio maps* for regions of different size. These Gaussian kernels are determined the same as Eq. 3. Then, the sub-descriptor $C(m_0)$ is defined as:

$$\begin{aligned} C(m_0) = [&\tilde{c}^T_{\Sigma_1}(m_0), \\ &\tilde{c}^T_{\Sigma_1}(m_{11}), \tilde{c}^T_{\Sigma_1}(m_{12}), \cdots, \tilde{c}^T_{\Sigma_1}(m_{18}), \\ &\tilde{c}^T_{\Sigma_2}(m_{21}), \tilde{c}^T_{\Sigma_2}(m_{22}), \cdots, \tilde{c}^T_{\Sigma_2}(m_{28}), \\ &\tilde{c}^T_{\Sigma_3}(m_{31}), \tilde{c}^T_{\Sigma_3}(m_{32}), \cdots, \tilde{c}^T_{\Sigma_3}(m_{38})] \end{aligned} \quad (10)$$

where $\tilde{c}^T_{\Sigma_1}(m_0)$ is a normalized vector in each histogram of $c^T_{\Sigma_1}(m_0)$, and $c^T_{\Sigma_1}(m_0) = [G^{\Sigma_1}_{d1}(m_0), G^{\Sigma_1}_{d2}(m_0), G^{\Sigma_1}_{d3}(m_0)]$ denote the vector composed of the values at location m_0 in the convolved color ratio maps. The color invariant sub-descriptor at each pixel is composed of $3 * 25 = 75$ values.

2.3 Matching Cost

After all the descriptors are constructed in two images, we find the best matches by the following matching cost.

$$D = \omega D_H + (1-\omega)D_C \quad (11)$$

where the matching cost D_H of Haar wavelet response sub-descriptor is computed by Euclidean distance:

$$D_H = |H_i - H_j| = \sqrt{\sum_{k=1}^{100}(H_{i,k} - H_{j,k})^2} \quad (12)$$

The matching cost D_C of color invariant sub-descriptor is computed by χ^2 distances:

$$D_C = \chi^2 = \frac{1}{2}\sum_{k=1}^{75}\frac{(C_{i,k} - C_{j,k})^2}{C_{i,k} + C_{j,k}} \quad (13)$$

The χ^2 measure is very useful since it can normalize larger bins.

3. Experimental Results

3.1 Demonstration of Our Color Model

Figure 3 demonstrates our color invariant model. Although the *RGB* color distributions of pixels in the two 20*20 corresponding windows (shown in (a) and (b) with red and blue rectangle) is very different in the two images [9] with illumination changes, the color distributions of pixels are similar after the color transformation with our color invariant model.

3.2 Comparisons of Descriptors in Performance

In our dense matching algorithm, the sparse features are first extracted in two uncalibrated images, and then the epipolar constraint and homography constraint are estimated based

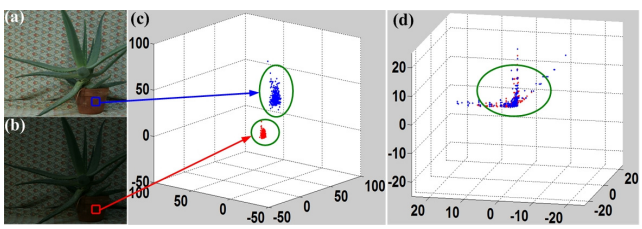

Fig. 3 Demonstration of our color invariant model. (a) and (b) show two images with sudden illumination changes, and two 20*20 corresponding windows; (c) and (d) show the *RGB* color distributions and the absolute value of color ratio $|f_1||f_2||f_3|$ distributions of pixels in two windows.

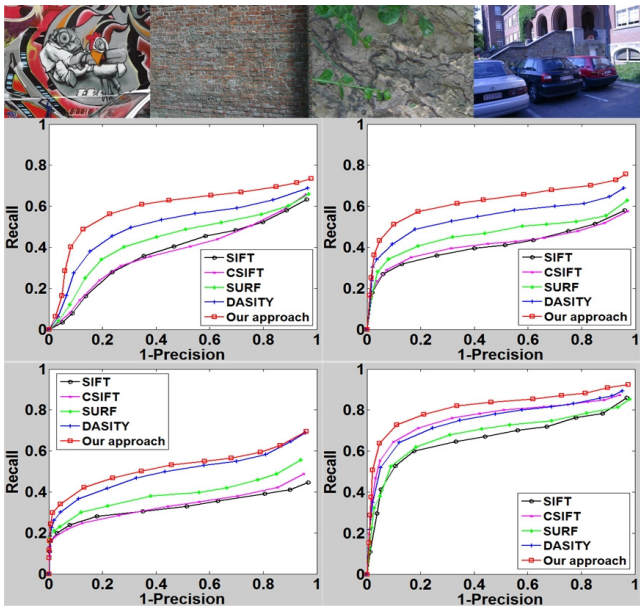

Fig. 4 Performance evaluation of five descriptors (Daisy, SIFT, SURF, CSIFT and our approach) in the 4 image sets with three changes from INRIA: (1), (2) viewpoint changes, (3) zoom+rotation, (4) light changes.

on the self-adaptive RANSAC algorithm. These two constraints can acquire a small searching region which is required in the dense matching. For each point in the left image, we compare descriptors with a simple nearest neighbor distance (Eq. 11) with a threshold T_d on the match, and keep the pair with the best match in the small searching region of the right image. In these experiments, the weighting factor $\omega = 0.5$, and T_d is also set to be 0.5.

The first row in Fig. 4 shows 4 image sets from INRIA used in our experiments. They have three different changes: viewpoint changes, zoom and rotation changes, and light changes. These image sets are related by homographies (plane projective transformations) which are regarded as ground truth to evaluate our algorithm. The Recall-Precision is used to evaluate our algorithm. The correct matches correspond to the same physical location which can be determined by image homography H. The number of the correct matches is determined by the criterion that the errors between real position and predicted position of matches are less than 3 pixels (i.e. $\|x'_i - Hx_i\| < 3$ and $\|x_i - H^{-1}x'_i\| < 3$). The total number of positive matches for the given dataset is known a priori. The Recall versus 1-Precision curves are generated by changing the different matching threshold.

$$\begin{aligned} &\text{Recall} = (\#\text{correct matches})/(\#\text{positive matches}) \\ &1 - \text{Precision} = (\#\text{false matches})/(\#\text{matches}) \end{aligned} \quad (14)$$

The last two rows in Fig. 4 shows the comparison results with four descriptors, including Daisy, SIFT, SURF, CSIFT, and our approach in 4 image sets. Although these image pairs have large viewpoint changes, affine variations or light changes, the performance of our algorithm is better than others. This is because first, although Haar wavelet uses 4 responses instead of 8 orientations, it still has enough information since Haar wavelet filter is based on a box type region. Second, our descriptor combines intensity information and color information together.

Table 1 Time complexity comparison in seconds.

Image Size	Daisy	SIFT	CSIFT	SURF	Our method	
965*726	8.01	344.55	320.51	104.62	4.81	6.75
832*553	5.79	228.65	182.15	68.25	3.36	4.88
930*598	6.32	265.48	223.56	82.54	3.85	5.02
1065*686	8.39	390.62	344.87	110.38	5.32	7.02

3.3 Comparisons of Descriptors in Efficiency

We compares the time complexity of our description algorithm with Daisy, SIFT, and SURF algorithms in the dense matching. Although our descriptor includes two subdescriptors, it has lower dimension than Daisy descriptor. Table 1 shows some comparison results. From the result we can see that our algorithm is faster than SIFT, CSIFT and SURF algorithms.

4. Conclusions

This letter presents an efficient descriptor combining Haar wavelet response and a new color invariant model. It provides powerful discrimination in wide-baseline dense matching. The color model is invariant to changes of viewing direction, and direction, intensity, and color of the illumination. Experiment results validate our descriptor.

Acknowledgments

This work is supported by National 973 Key Research Program of China (2011CB302203), National Natural Science Foundation of China (60973061) and (60803073).

References

[1] E. Tola, V. Lepetit, and P. Fua, "DAISY: An efficient dense descriptor applied to wide-baseline stereo," IEEE Trans. Pattern Anal. Mach. Intell., vol.32, no.5, pp.815–830, 2010.
[2] D. Lowe, "Distinctive image features from scale-invariant keypoints," IJCV, vol.20, no.2, pp.91–110, 2004.
[3] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool, "A comparison of affine region detectors," IJCV, vol.65, no.1/2, pp.43–72, 2005.
[4] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "SURF: Speed up robust features," CVIU, vol.110, no.3, pp.346–359, 2008.
[5] A.E. Abdel-Hakim and A.A. Farag, "CSIFT: A SIFT descriptor with color invariant characteristics," CVPR, vol.2, pp.1978–1983, 2006.
[6] T. Gevers and A.W.M. Smeulders, "Color-based object recognition," Pattern Recognit., vol.32, pp.453–464, 1999.
[7] F. Mindru, T. Tuytelaars, L. Van Gool, and T. Moons, "Moment invariants for recognition under changing viewpoint and illumination," CVIU, vol.94, no.1-3, pp.3–27, 2004.
[8] K.E.A. van de Sande, T. Gevers, and C.G.M. Snoek, "Evaluating color descriptors for object and scene recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol.32, no.9, pp.1582–1596, 2010.
[9] http://vision.middlebury.edu/stereo/data/scenes2006/