DASC: Dense Adaptive Self-Correlation Descriptor for Multi-modal and Multi-spectral Correspondence

Seungryong Kim[1], Dongbo Min[2,3], Bumsub Ham[4,*], Seungchul Ryu[1], Minh N. Do[5], Kwanghoon Sohn[1,†]
[1]Yonsei University [2]Chungnam Nat. University [3]ADSC [4]Inria [5]UIUC
http://seungryong.github.io/DASC/

Abstract

Establishing dense visual correspondence between multiple images is a fundamental task in many applications of computer vision and computational photography. Classical approaches, which aim to estimate dense stereo and optical flow fields for images adjacent in viewpoint or in time, have been dramatically advanced in recent studies. However, finding reliable visual correspondence in multi-modal or multi-spectral images still remains unsolved. In this paper, we propose a novel dense matching descriptor, called dense adaptive self-correlation (DASC), to effectively address this kind of matching scenarios. Based on the observation that a self-similarity existing within images is less sensitive to modality variations, we define the descriptor with a series of an adaptive self-correlation similarity for patches within a local support window. To further improve the matching quality and runtime efficiency, we propose a randomized receptive field pooling, in which a sampling pattern is optimized with a discriminative learning. Moreover, the computational redundancy that arises when computing densely sampled descriptor over an entire image is dramatically reduced by applying fast edge-aware filtering. Experiments demonstrate the outstanding performance of the DASC descriptor in many cases of multi-modal and multi-spectral correspondence.

1. Introduction

Recently, many computer vision and computational photography problems have been reformulated to overcome an inherent limitation by leveraging multi-modal and multi-spectral images such as RGB and near-infrared (NIR) image pairs [6, 40], flash and no-flash images [27], color and dark flash images [21], blurred images [14, 23], and images taken under different radiometric conditions [35].

Figure 1. Some challenging multi-modal and multi-spectral images such as (from top to bottom) RGB-NIR, flash-noflash images, two images with different exposures, and blur-sharp images. The images in the third column are the results obtained by warping images in the second column to images in the first column with dense correspondence maps estimated by using our DASC descriptor.

Estimating dense visual correspondence for given multi-spectral and multi-modal images is a key enabler for realizing such tasks. In general, the performance of correspondence estimation algorithms relies primarily on two components: appearance descriptor and optimization algorithm. Traditional dense correspondence approaches for estimating depth [33] or optical flow [8, 37] fields, in which input images are acquired in a similar imaging condition, have been dramatically advanced in recent studies. To define a matching fidelity term, they typically assume that multiple images share a similar visual pattern, *e.g.*, color, gradient, and structural similarity. However, when it comes to multi-modal and multi-spectral images, such properties do

*WILLOW project-team, Département d'Informatique de l'Ecole Normale Supérieure, ENS/Inria/CNRS UMR 8548.
†Corresponding author: khsohn@yonsei.ac.kr.

not hold as shown in Fig. 1. In these cases, conventional descriptors or similarity measures often fail to capture reliable matching evidence, leading to a poor matching quality. Although employing powerful optimization techniques could help estimate a reliable solution with a spatial context [13, 24, 20], an optimizer itself cannot address an inherent limitation without suitable matching descriptors for challenging multi-spectral and multi-modal images [28].

Our method starts from the observation that the local internal layout of self-similarities is less sensitive to photometric distortions, even when an intensity distribution of an anatomical structure is not maintained across different imaging modalities [34]. The local self-similarity (LSS) descriptor enables overcoming many inherent limitations of existing descriptors in establishing correspondence between multi-modal or multi-spectral images. It is worth noting that although several approaches based on the LSS have been presented for multi-modal and multi-spectral image registration [16, 39], they do not scale well to estimating dense correspondence for multi-modal and multi-spectral images, and thus their matching performance is still poor.

In this paper, we propose a novel local descriptor, called dense adaptive self-correlation (DASC), designed for establishing dense multi-modal and multi-spectral correspondence. It is defined with a series of patch-wise similarities within a local support window. The similarity between patch-wise receptive fields is computed with an adaptive self-correlation measure, which encodes intrinsic structure while providing the robustness against modality variations. To further improve the matching quality and runtime efficiency, we also propose a randomized receptive field pooling strategy with sampling patterns that selects two patches within the local support window, rather than using a center patch and a patch of a neighboring pixel. A linear discriminative learning is employed for obtaining an optimal sampling pattern. Moreover, the computational redundancy that arises when computing densely sampled descriptors over an entire image is dramatically reduced by applying fast edge-aware filtering [15]. Experimental results show that our DASC descriptor outperforms conventional area-based approaches and feature-based approaches (including LSS descriptor [34]) on various benchmarks; 1) Middlebury stereo benchmark [1] consisting of images with varying illumination and exposure conditions, 2) multi-modal and multi-spectral dataset including RGB-NIR images [36, 6], different exposure [36, 35], flash-noflash images [35], and blurry images [14, 23], and 3) MPI optical flow benchmark [8] containing motion blur and illumination changes.

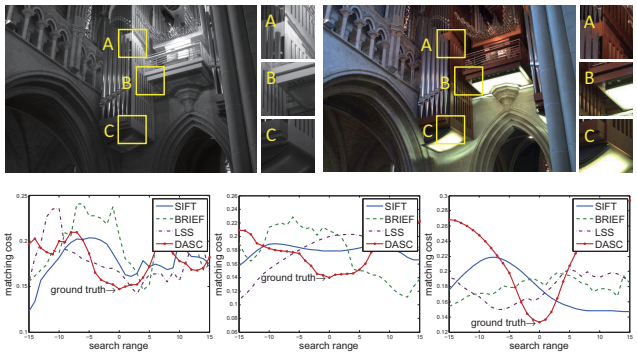
The contributions of this paper can be summarized as follows. First, to the best of our knowledge, our approach is the first attempt to design an efficient, dense descriptor for matching multi-modal and multi-spectral images. Second, unlike a center-biased dense max pooling, we propose a randomized receptive field pooling with sampling patterns optimized via a discriminative learning, making the descriptor more robust against matching outliers incurred by different imaging modalities. Third, we propose an efficient computational scheme that significantly improves the runtime efficiency of the proposed dense descriptor. Finally, we provide an intensive comparative study with state-of-the-art methods using various datasets.

2. Related Work

Feature-based approaches As a pioneering work, the scale invariant feature transform (SIFT) was first introduced by Lowe [25] to estimate robust sparse correspondence under geometric and photometric variations. Recently, based on the simple intensity comparison, fast binary descriptors, such as binary robust independent elementary features (BRIEF) [9] and fast retina keypoint (FREAK) [3], have been popularly proposed. Unlike these sparse descriptors, Tola *et al.* developed a dense descriptor, called DAISY [38], which re-designs conventional sparse descriptors, *i.e.*, SIFT, to efficiently compute densely sampled descriptors over an entire image. Although these conventional gradient-based and intensity comparison-based descriptors show satisfactory performance for small deformation, they cannot properly describe the multi-modal and multi-spectral images including non-linear deformation frequently.

To estimate correspondences in multi-modal and multi-spectral images, some variants of SIFT have been developed [32], but these gradient-based descriptors have an inherent limitation similar to the SIFT, especially when an image gradient varies across different modality images. Schechtman and Irani introduced the LSS descriptor [34] for the purpose of template matching, and achieved impressive results in object detection and retrieval. Torabi *et al.* employed the LSS as a multi-spectral similarity metric to register human region of interests (ROIs) [39]. The LSS has also been applied to the registration of multi-spectral remote sensing images [42]. For multi-modal medical image registration, Heinrich *et al.* proposed a modality independent neighborhood descriptor (MIND) [16] inspired by the LSS. However, none of these approaches scale very well to dense matching tasks for multi-modal and multi-spectral images due to a low discriminative power and a huge complexity.

Area-based approaches As surveyed in [29], the mutual information (MI), leveraging the entropy of the joint probability distribution function (PDF), has been popularly applied to multi-modal medical image alignment. However, the MI is sensitive to local variation since it is assumed that there exists a global transformation [18]. Although cross-correlation based methods such as an adaptive normalized cross-correlation (ANCC) [17] show satisfactory results for locally linear variations, they provide limited performances



(a) Matching cost in A (b) Matching cost in B (c) Matching cost in C

Figure 2. Examples of matching cost comparison. Multi-spectral RGB and NIR images have locally non-linear deformation as depicted in A, B, and C. Matching costs computed with different descriptors along A, B, and C's scanelines were plotted in (a)-(c). Unlike conventional descriptors, the proposed DASC descriptor yields a global minimum.

under severe modality variations. Irani *et al.* [19] employed cross-correlation on the Laplacian energy map for measuring multi-sensor image similarity. Recently, a robust selective normalized cross-correlation (RSNCC) [36] was proposed for the dense alignment between multi-modal images, but its performance is still unsatisfactory due to an inherent limitation of similarity measure based on intensity value.

3. Background

Given an image $f_i : \mathcal{I} \rightarrow \mathbb{R}$ or $\mathbb{R}^3$, a dense descriptor $\mathcal{D}_i : \mathcal{I} \rightarrow \mathbb{R}^L$ is defined on a local support window centered at each pixel i, where $\mathcal{I} = \{i = (x_i, y_i)\} \subset \mathbb{N}^2$ is a discrete image domain. Conventionally, local descriptors were computed based on the assumption that there is a common underlying visual pattern which is shared by two images. However, as shown in Fig. 2, multi-spectral images such as a pair of RGB-NIR have a nonlinear photometric deformation even within a small window, *e.g.*, gradient reverses and intensity order variation. More seriously, there are outliers including structure divergence caused by shadow or highlight. In these cases, conventional descriptors using an image gradient (SIFT [25]) or an intensity comparison (BRIEF [9]) cannot capture coherent matching descriptors, which induce erroneous local minima in estimating dense correspondences. Without suitable descriptors, a matching quality has an inherent matching ambiguity even with a spatial context by leveraging a powerful optimization,

Unlike these conventional descriptors, the LSS descriptor $\mathcal{D}_i^{\text{LSS}}$ measures a correlation between two patches $\mathcal{F}_i$ and $\mathcal{F}_j$ centered at pixel i and j within a local support window $\mathcal{R}_i$ [34]. As shown in Fig. 3, it discretizes the correlation surface on a log-polar grid, generates a set of bins, and then stores a maximum correlation value within each bin. Formally, $\mathcal{D}_i^{\text{LSS}} = \bigcup_l d_{i,l}^{\text{LSS}}$ for $l = 1, ..., L^{\text{LSS}}$ is a $L^{\text{LSS}} \times 1$ feature vector, and can be written as follows:

$$d_{i,l}^{\text{LSS}} = \max_{j \in \text{bin}_i(l)} \{\mathcal{C}(i,j)\}, \tag{1}$$

where $\text{bin}_i(l) = \{j | j \in \mathcal{R}_i, \rho_{r-1} < |i-j| \leq \rho_r, \theta_{a-1} < \angle(i-j) \leq \theta_a\}$ with a log radius ρ_r for $r \in \{1, \cdots, N_\rho\}$ and a quantized angle θ_a for $a \in \{1, \cdots, N_\theta\}$ with $\rho_0 = 0$ and $\theta_0 = 0$. The correlation surface $\mathcal{C}(i,j)$ is typically computed using simple similarity metric such as the sum of square difference (SSD) with a normalization factor σ_s:

$$\mathcal{C}(i,j) = \exp\left(-\text{SSD}(\mathcal{F}_i, \mathcal{F}_j)/\sigma_s\right). \tag{2}$$

The LSS descriptor has been shown to be robust in cross-domain object detection [34], but it provides unsatisfactory results in densely matching multi-modal images as shown in Fig. 2. It is because the max pooling strategy performed in each $\text{bin}_i(l)$ lose matching details, leading to a poor discriminative power. Furthermore, the center-biased correlation measure cannot handle severe outliers effectively, which frequently exist in multi-modal and multi-spectral images. In terms of a computational complexity, there exists no efficient computational scheme designed for dense matching descriptor.

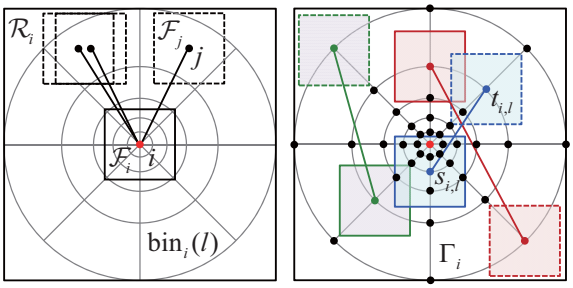
4. The DASC Descriptor

Our objective is to design a dense descriptor for multi-modal correspondence, while a computational complexity is kept low. Our descriptor is built with a series of adaptive self-correlation for patch-wise receptive fields, which is efficiently computed with fast edge-aware filtering.

4.1. Randomized Receptive Field Pooling

Instead of using a center-biased max pooling of LSS descriptor in Fig. 3(a), our DASC descriptor incorporates a randomized receptive field pooling with sampling patterns in such a way that a pair of two patches are randomly selected within a local support window. It is motivated by three observations; 1) In multi-spectral and multi-modal images, there frequently exist non-informative regions which are locally degraded, *e.g.*, shadows or outliers. 2) Center-biased pooling is very sensitive to a degradation of a center patch, and cannot deal with a homogeneous or salient center pixel which does not contain self-similarities [34]. 3) From the relationship between Census transform [43] and BRIEF [9] descriptor, it is shown that the randomness enables a descriptor to encode structural information more robustly.

Our approach encodes a similarity between patch-wise receptive fields sampled from log-polar circular point set Γ_i as shown in Fig. 3(b). It is defined as $\Gamma_i = \{j | j \in \mathcal{R}_i, |i-j| = \rho_r, \angle(i-j) = \theta_a\}$ where the number of points is defined as $N_c = N_\rho \times N_\theta + 1$, and has a higher density of points near a center pixel, similar to



(a) (b)

Figure 3. Demonstration of (a) LSS descriptor [34] and (b) DASC descriptor. Within a support window, solid and dotted line box depict source and target patch, respectively. Unlike a center-biased dense max pooling on each $\text{bin}_i(l)$, the DASC descriptor incorporates a randomized pooling using sampling pattern $(s_{i,l}, t_{i,l}) \in \cup_i^L$ on Γ_i optimized by a discriminative learning.

DAISY [38] descriptor. Given N_c points in Γ_i, there exist $N_{pc} = \{N_c \times (N_c - 1)\}/2$ candidate sampling patterns, leading to a dramatically high-dimension descriptor. However, many of the sampling pattern pairs might not be useful in describing a local support window. Therefore, we employ a randomized approach to extract L sampling patterns from N_{pc} pattern candidates. Our descriptor $\mathcal{D}_i = \bigcup_l d_{i,l}$ for $l = 1, ..., L$ is encoded with a set of patch similarity between two patches based on sampling patterns that are selected from Γ_i:

$$d_{i,l} = \mathcal{C}(s_{i,l}, t_{i,l}), \quad s_{i,l}, t_{i,l} \in \Gamma_i, \tag{3}$$

where s_l and t_l are l^{th} selected sampling patterns. Note that the sampling patterns are fixed for all pixels in an image. Namely, all pixels share the same offset vectors, enabling a fast computation of dense descriptors, which will be detailed in Sec. 4.3. Although the DASC descriptor uses only sparse patch-wise pairs in a local support window, many of patches are overlapped when computing patch similarities between the sparse pairs, allowing the descriptor to consider the majority of pixels in the support window and to reflect original image attributes effectively.

Sampling pattern learning Finding a optimal randomized sampling pattern is a critical issue in our descriptor. With the assumption that there is no single hand-craft feature that always provides the robustness to all circumstances [11], we employ a discriminative learning to optimal sampling patterns describe a local support window. Given candidate sampling patterns $\cup_i = \{(s_{i,l}, t_{i,l}) | l = 1, ..., N_{pc}\}$, our goal is to select the best sampling patterns which derive an important spatial layout.

Our approach exploits support vector machines (SVMs) with a linear kernel [10]. For learning, we build a dataset $\mathcal{P} = \{(\mathcal{R}_m^1, \mathcal{R}_m^2, y_m) | m = 1, ..., N_t\}$ where $(\mathcal{R}^1, \mathcal{R}^2)$ are support window pairs in multi-modal or multi-spectral images, and N_t is the number of training samples. y is a binary label that becomes 1 if two patches are matched or 0 otherwise. The training data set $\mathcal{P}$ was built from ground truth dense correspondence maps for images captured under varying illumination conditions and/or with imaging devices [1, 8, 36]. First, the feature $\mathbf{r}_m = \bigcup_l r_{m,l}$ that describes two support window pairs $\mathcal{R}_m^1$ and $\mathcal{R}_m^2$ is defined

$$r_{m,l} = \exp\left(-(d_{m,l}^1 - d_{m,l}^2)^2 / 2\sigma_r^2\right), \tag{4}$$

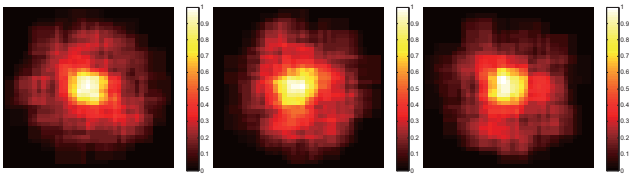
where σ_r is a bandwidth for Gaussian kernel and $d_{m,l}$ is an adaptive self-correlation, which will be explained in Sec. 4.2. The decision function to classify training dataset $\mathcal{P}$ into matching and non-matching can be represented as

$$\rho(\mathbf{r}_m) = \mathbf{v}^T \mathbf{r}_m + \mathbf{b}, \tag{5}$$

where $\mathbf{v}$ indicates an amount of contribution of each candidate sampling pattern and $\mathbf{b}$ is a bias. Learning $\mathbf{v}$ can be formulated as minimizing the following objective function

$$\mathcal{L}(\mathbf{v}) = \lambda ||\mathbf{v}||^2 + \sum\nolimits_m \max(0, 1 - y_m \rho(\mathbf{r}_m)), \tag{6}$$

where λ represents a regularization parameter. We use LIBSVM [10] to minimize this objective function. The weight $|\mathbf{v}_l|$ encodes the importance of corresponding sampling pattern towards the final decision [22]. Therefore, we rank top L sampling patterns based on $|\mathbf{v}_l|$ value, and use them in our descriptor, which is denoted as $\cup_i^L$. Fig. 4 shows visualizations of learned receptive fields of the DASC descriptor.



(a) (b) (c)

Figure 4. Visualization of patch-wise receptive fields of the DASC descriptor which are learned from (a) Middlebury benchmark [1], (b) multi-spectral and multi-modal benchmark [36], and (c) MPI SINTEL benchmark [8].

4.2. Adaptive Self-Correlation Measure

With the sampling patterns $(s_{i,l}, t_{i,l})$ estimated, our descriptor measures a patch similarity with an adaptive self-correlation measure in order to robustly encode a local internal layout of self-similarities. For the sake of simplicity, we omit (i, l) in the correlation metric from here on, as it is repeatedly computed for all (i, l). For $(s, t) \in \cup_i^L$, we compute the adaptive self-correlation $\Psi(s, t)$ between two patches $\mathcal{F}_s$ and $\mathcal{F}_t$ as follows:

$$\Psi(s,t) = \frac{\sum_{s',t'} \omega_{s,s'} \omega_{t,t'} (f_{s'} - \mathcal{G}_s)(f_{t'} - \mathcal{G}_t)}{\sqrt{\sum_{s'} \{\omega_{s,s'} (f_{s'} - \mathcal{G}_s)\}^2} \sqrt{\sum_{t'} \{\omega_{t,t'} (f_{t'} - \mathcal{G}_t)\}^2}}, \tag{7}$$

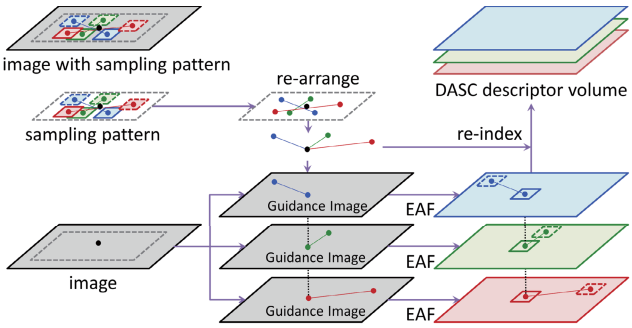


Figure 5. Efficient computation framework of the DASC descriptor. In order to reduce a computational load in computing the adaptive self-correlation, the DASC descriptor re-arranges the sampling pattern and employs fast EAF scheme.

where $s' \in \mathcal{F}_s$ and $t' \in \mathcal{F}_t$, and $\mathcal{G}_s = \sum_{s'} \omega_{s,s'} f_{s'}$.

The weight $\omega_{s,s'}$ represents how similar two pixels s and s' are, and is normalized, *i.e.*, $\sum_{s'} \omega_{s,s'} = 1$. It can be defined with any kind of edge-aware weights [41, 15, 12]. This weighted sum better handles outliers and local variations in patches compared to other patch-wise similarity metrics. It is worth noting that the adaptive self-correlation used here is conceptually similar to the ANCC [17], but our descriptor employs the correlation metric for measuring *self-similarity* within a single image (which is used for matching two images later), while the ANCC is used to directly measure *inter-similarity* between two different intensity (or color) images.

Finally, our patch-wise similarity between $\mathcal{F}_s$ and $\mathcal{F}_t$ is computed with a truncated exponential function, which has been widely used in robust estimator [4]:

$$\mathcal{C}(s,t) = \max(\exp(-(1-|\Psi(s,t)|)/\sigma), \tau), \qquad (8)$$

where σ is a bandwidth of Gaussian kernel and τ is a truncation parameter. Here, a absolute value of $\Psi(s,t)$ is used for mitigating the effect of intensity reverses. The correlation $\mathcal{C}(s_{i,l}, t_{i,l})$ is normalized with unit norm of all l.

4.3. Efficient Computation for Dense Description

For densely constructing our descriptor on an entire image, we should compute $\mathcal{C}(s_{i,l}, t_{i,l})$ for all patch pairs belonging to $(s_{i,l}, t_{i,l}) \in \cup_i^L$ for each pixel i. Thus, a straightforward computation can be extremely time-consuming. In specific, the computational complexity becomes $O(INL)$, where I, N, and L represent an image size, a patch size, and the number of sampling patterns, respectively.

In this section, we present an efficient method for computing the DASC descriptor. To compute all weighted sums in (7) for $(s_{i,l}, t_{i,l})$ efficiently, we apply a constant-time edge-aware filter (EAF), *e.g.*, the guided filter (GF) [15]. However, the symmetric weight $w_{s,s'}w_{t,t'}$ varies for each l, and thus computing the numerator in (7) is still very time-consuming. Moreover, $w_{s,s'}w_{t,t'}$ is computed with a

Algorithm 1: Dense Adaptive Self-Correlation (DASC)

Input : image f_i, candidate sampling patterns $\cup_i$, training patch pairs $\mathcal{P}$.
Output : the DASC descriptor volume $\mathcal{D}_i$.
Parameters and Notatation :
L: descriptor dimension.
$\omega_{i,i'}$: weight between pixel i and $i' \in \mathcal{F}_i$.

/ Offline Procedure */*
1 : Compute $\mathbf{r}_m$ using (4) for possible candidate sampling patterns $\cup_i$ on training support window pairs $\mathcal{P}$.
2 : Learn a weight $\mathbf{v}_l$ by optimizing (6).
3 : Select the maximal L sampling patterns $(s_{i,l}, t_{i,l})$ in terms of $|\mathbf{v}_l|$, denoted as $\cup_i^L$.

/ Online Procedure */*
4 : Compute $\mathcal{G}_i = \sum_{i'} \omega_{i,i'} f_{i'}$ for all pixel i.
5 : Compute $\mathcal{G}_{i^2} = \sum_{i'} \omega_{i,i'} f_{i'}^2$.
for $l = 1 : L$ **do**
  6 : Re-arrange $(s_{i,l}, t_{i,l})$ as $(i,j) = (i, i + t_{i,l} - s_{i,l})$.
  7 : Compute $\mathcal{G}_{ij} = \sum_{i',j'} \omega_{i,i'} f_{i'} f_{j'}$.
  8 : Compute $\mathcal{G}_{i,j} = \sum_{i',j'} \omega_{i,i'} f_{j'}$.
  9 : Compute $\mathcal{G}_{i,j^2} = \sum_{i',j'} \omega_{i,i'} f_{j'}^2$.
  10 : Estimate $\tilde{\Psi}(i,i')$ and $\mathcal{C}(i,i')$ using (10) and (8).
  11 : Re-index $d_{i,l} = \mathcal{C}(s_{i,l}, t_{i,l})$.
end for

range distance using 6-D vector (or 2-D vector), when an input is a color image (or a greyscale image). It significantly increases a computational burden needed for employing constant-time EAFs [15, 26].

To alleviate these limitations, we simplify (7) by considering only the weight $w_{s,s'}$ from the source patch $\mathcal{F}_s$ so that a fast computation of (7) using fast edge-aware filter is feasible. It should be noted that such an asymmetric weight approximation has also been used in cost aggregation for stereo matching [31]. We also found that in our descriptor, a performance gap between using the asymmetric weight $w_{s,s'}$ and the symmetric weight $w_{s,s'}w_{t,t'}$ is negligible. Furthermore, for efficient description, we also rearrange the sampling pattern $(s_{i,l}, t_{i,l})$ to referenced-biased pairs $(i,j) = (i, i+t_{i,l}-s_{i,l})$. The adaptive self-correlation in (7) is then approximated as follows:

$$\tilde{\Psi}(i,j) = \frac{\sum\limits_{i',j'} \omega_{i,i'}(f_{i'} - \mathcal{G}_i)(f_{j'} - \mathcal{G}_{i,j})}{\sqrt{\sum\limits_{i'} \omega_{i,i'}(f_{i'} - \mathcal{G}_i)^2} \sqrt{\sum\limits_{i',j'} \omega_{i,i'}(f_{j'} - \mathcal{G}_{i,j})^2}}, \qquad (9)$$

where $\mathcal{G}_i = \sum_{i'} \omega_{i,i'} f_{i'}$. $\mathcal{G}_{i,j} = \sum_{i',j'} \omega_{i,i'} f_{j'}$ means weighted average of $\mathcal{F}_j$ with a guidance $\mathcal{F}_i$ with our approximation.

We then decompose denominator and numerator in (9)

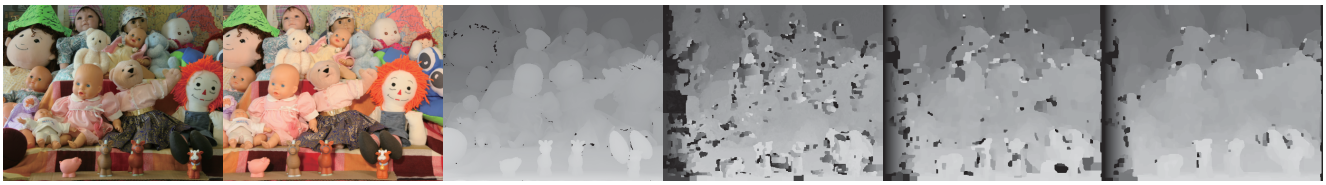


Figure 6. Comparison of disparity estimation for *Dolls* image pairs taken under illumination combination '1/3'. (from left to right) Left color image, right color image, and disparity maps for the ground truth, ANCC [17], SIFT [25], and DASC+LRP. Conventional approaches cannot estimate a reliable disparity map. In contrast, the DASC+LRP descriptor estimates accurate and edge-preserved disparity maps while reducing artifacts.

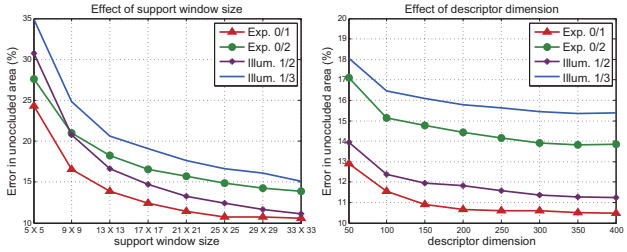


Figure 7. Average bad-pixel error rate on Middlebury benchmark of DASC+LRP descriptor with WTA optimization as varying support window size and descriptor dimension.

image size	DAISY	LSS	DASC†	DASC‡
463×370	$2.5s$	$31s$	$128s$	$2.7s$

Table 1. Evaluation of computational time. The brute-force and efficient computation of DASC is denoted as †and ‡, respectively.

after some arithmetic derivations:

$$\frac{\mathcal{G}_{i,ij} - \mathcal{G}_i \cdot \mathcal{G}_{i,j}}{\sqrt{\mathcal{G}_{i^2} - {\mathcal{G}_i}^2} \cdot \sqrt{\mathcal{G}_{i,j^2} - {\mathcal{G}_{i,j}}^2}}, \tag{10}$$

where $\mathcal{G}_{i^2} = \sum_{i'} \omega_{i,i'} f_{i'}^2$, $\mathcal{G}_{i,ij} = \sum_{i',j'} \omega_{i,i'} f_{i'} f_{j'}$ and $\mathcal{G}_{i,j^2} = \sum_{i',j'} \omega_{i,i'} f_{j'}^2$. While the $\mathcal{G}_i$ and $\mathcal{G}_{i^2}$ can be computed on image domain once, $\mathcal{G}_{i,ij}$, $\mathcal{G}_{i,j}$, and $\mathcal{G}_{i,j^2}$ should be computed on each offset. All these components can be efficiently computed using a constant-time edge-aware filter (EAF). Thus, our approach removes the complexity dependency on the patch size N, *i.e.*, $O(IL)$. Furthermore, since there exist repeated offsets, the complexity is further reduced as $O(I\tilde{L})$ for $\tilde{L} < L$. Finally, the dense descriptor $\mathcal{D}_i$ is re-indexed as $d_{i,l} = \mathcal{C}(s_{i,l}, t_{i,l})$ though the robust function in (8). Fig. 5 describes our efficient method for computing the DASC descriptor. Algorithm 1 summarized the DASC descriptor computation.

5. Experimental Results and Discussion

In experiments, our DASC descriptor is implemented with the following same parameter settings for all datasets: $\{\sigma, \tau, N, M, L\} = \{0.5, 0.03, 5\times5, 31\times31, 128\}$ where M is a local support window size. We implemented the DASC descriptor in C++ on Intel Core i7-3770 CPU at 3.40 GHz,

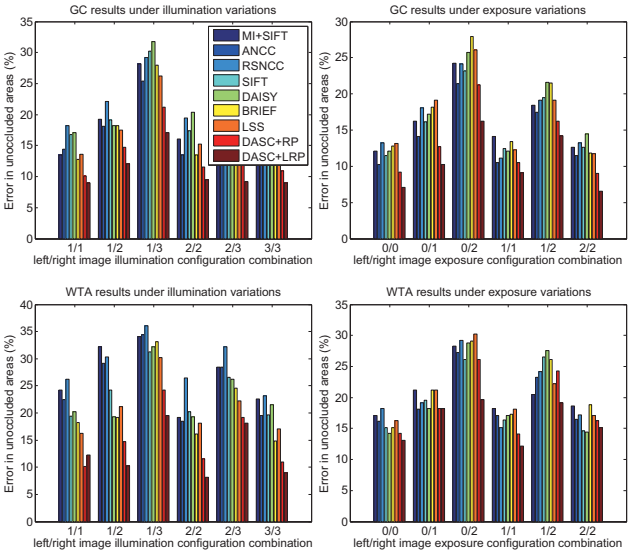
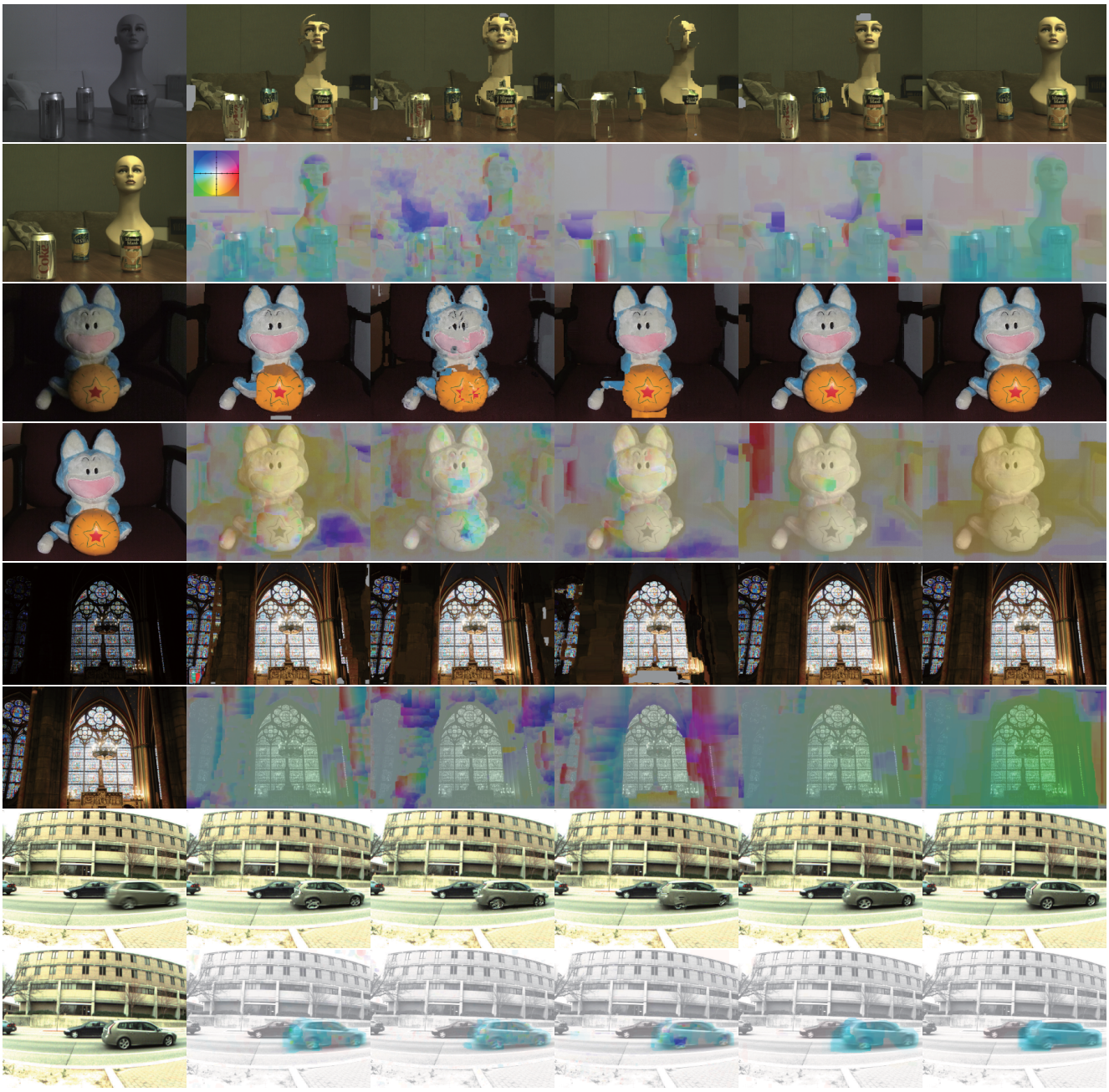


Figure 8. Average bad-pixel error rate on Middlebury benchmark with illumination variations and exposure variations. The GC (first row) and WTA (second row) were used for optimization, respectively. Our DASC+LRP shows the best performance.

and measured the runtime on a single CPU core without further code optimizations and parallel implementation using multi-core CPUs/GPU. The DASC descriptor was evaluated with other state-of-the-art descriptors, *e.g.*, SIFT [25], DAISY [38], BRIEF [9], and LSS [34], and other area-based approaches, *e.g.*, ANCC [17], MI+SIFT [18], and RSNCC [36]. Specifically, we evaluated the DASC using a randomized pooling (DASC+RP) and the DASC using a learned randomized pooling (DASC+LRP) that is our final descriptor. It is worth noting that we built three training sets from benchmark databases for each following experiments, and each training set was excluded from experiments. Fig. 7 shows the effects of a support window size M and the number of sampling patterns L in DASC descriptor. The computational complexity of DASC descriptor compared with other descriptors was evaluated in Table 1.

5.1. Middlebury Stereo Benchmark

We first evaluated our DASC+LRP descriptor in Middlebury stereo benchmark containing illumination and ex-



(a) Input image pairs (b) RSNCC [36] (c) BRIEF [9] (d) DAISY [38] (e) LSS [34] (f) DASC+LRP

Figure 9. Comparison of dense correspondence for (from top to bottom) RGB-NIR images, flash-noflash images, different exposure images, and blurred-sharpen images. The results consist of warped color images and correspondence flow fields overlaid with reference images. Compared to other conventional approaches, our DASC+LRP descriptor estimates reliable dense correspondence fields for challenging multi-modal and multi-spectral image pairs.

posure variations [1]. In experiments, the illumination (or exposure) combination '1/3' indicates that two images were captured under 1^{st} and 3^{rd} illumination (exposure) conditions, respectively [1]. Fig. 6 shows disparity maps for severe illumination variations obtained by varying cost functions with the winner-takes-all (WTA) optimization. Fig. 8 shows average bad matching errors in un-occluded areas of depth maps obtained under illumination or exposure variations with the graph-cut (GC) [5] and WTA optimization. Our DASC+LRP descriptor achieves the best results both quantitatively and qualitatively. Area-based approaches, *e.g.*, MI+SIFT [18], ANCC [17], and RSNCC

	RGB-NIR	Flash-noflash	Diff. Exp.	Blur-Sharp	Ave.
NRDC [13]	54.27	48.92	51.34	59.72	53.56
ANCC [17]	18.45	14.14	11.96	19.24	15.94
RSNCC [36]	13.41	15.87	9.15	18.21	14.16
SIFT [25]	18.51	11.06	14.87	20.78	16.35
DAISY [38]	20.42	10.84	12.71	22.91	16.72
BRIEF [9]	17.54	9.21	9.54	19.72	14.05
LSS [34]	16.14	11.88	9.11	18.51	13.91
DASC+RP	11.71	7.51	7.32	12.21	9.68
DASC+LRP	**8.10**	**5.41**	**6.24**	**10.81**	**7.64**

Table 2. Comparison of quantitative evaluation on multi-spectral and multi-modal images.

[36], are very sensitive to severe radiometric variations, especially when local variations frequently occur. Contrarily, the descriptor-based approaches perform better than the area-based approaches. Interestingly, the BRIEF [9] is better than other descriptor-based descriptors (SIFT [25] and DAISY [38]) thanks to an ordering robustness.

5.2. Multi-modal and Multi-spectral Image Pairs

Next, we evaluated our DASC+LRP descriptor with images under modality variations, *e.g.*, RGB-NIR [36, 6], different exposure [36, 35], flash-noflash [35], and blurred artifacts [14, 23]. Due to severe matching ambiguities and outliers that exist multi-modal and multi-spectral correspondence, the simple WTA method does not achieve excellent results. In experiments, we exploit the SIFT flow optimization based on the publicly available code, specifically hierarchical dual-layer belief propagation (BP) [24], as varying descriptors and similarity measures. Unlike the Middlebury stereo benchmark, these datasets have no ground truth correspondence maps, thus we manually obtained ground truth displacement vectors for 100 corner points for all images, and used them for an objective evaluation similar to [36]. Indeed, it is necessary to build up multi-modal databases including ground truth dense maps for more accurate assessment, but we reserve this task as future work. Table 2 shows an objective evaluation of DASC+LRP descriptor and other state-of-the-art methods on these datasets.

Area-based approaches, *e.g.*, ANCC [17] and RSNCC [36] are very sensitive to local variations. As already described in literatures [36], gradient-based approaches, SIFT [25] and DAISY [38], have shown limited performance in RGB-NIR pairs where the gradient reversal and inversion frequently appear. The BRIEF [9] cannot deal with the noisy and modality varying regions since it considers a pixel difference only. It should be noted that some efforts have been made to estimate reliable flow maps in the motion blur, *e.g.*, blur-flow [30], but they typically employ an iterative matching framework, which relies heavily on an initial estimate. Additionally, they do not scale well to general purpose matching scenarios. Unlike these approaches, the LSS [34] and our descriptor consider the local self-similarities, but the LSS still lacks a discriminative power for dense matching. Our DASC+RP descriptor combining patch-wise pooling with adaptive self-correlation provides satisfactory results under modality variations. By employing the optimal sampling pattern via discriminative learning (DASC+LRP), the matching accuracy was further improved. Fig. 9 shows subjective evaluation, clearly demonstrating the outstanding performance of our descriptor.

	Clean Pass		Final Pass	
	all	*unmat.*	*all*	*unmat.*
Classic-NL [37]	7.940	39.821	9.439	43.123
LDOF [7]	7.180	38.124	8.422	42.892
LDOF+BRIEF [9]	6.281	37.841	7.741	41.875
LDOF+LSS [34]	6.182	37.514	7.152	40.332
LDOF+DASC	**5.578**	**36.975**	**6.384**	**38.932**

Table 3. Comparison of average EPE on the MPI SINTEL [8].

5.3. MPI Optical Flow Benchmark

Our DASC descriptor can also be incorporated into variational optical flow approaches, which typically assume only a small displacement between consecutive frames. However, motion blur and illumination variation can degenerate the performance. In order to handle such challenging issues simultaneously, we applied the DASC descriptor to the large displacement optical flow (LDOF) [7] as an initial evidence in the MPI SINTEL database [8] containing specular reflections, motion blur, and defocus blur. The dataset consists of two kind of rendering frames, namely clean pass and final pass, and each contains 12 sequences with over 500 frames in total [8]. Table 3 shows average end-point error (EPE) results on MPI SINTEL clean and final passes. The DASC descriptor improves the performance of conventional LDOP method.

6. Conclusion

The robust novel local descriptor called the DASC has been proposed for dense multi-spectral and multi-modal correspondence. It leverages an adaptive self-correlation measure and a randomized receptive field pooling learned by the linear discriminative learning. Moreover, by making use of fast edge-aware filters, our DASC descriptor is capable of computing the dense descriptor very efficiently. The DASC demonstrated its robustness in establishing dense correspondence between challenging image pairs taken under different modalities, *e.g.*, RGB-NIR, different illumination and exposure, flash-noflash, blurring artifacts. We believe our method will serve as an essential tool for several applications using multi-modal and multi-spectral images. We made our code publicly available [2].

Acknowledgement. The authors would like to thank the anonymous reviewers for their valuable comments and suggestions. This research was supported by the MSIP (The Ministry of Science, ICT and Future Planning), Korea and Microsoft Research, under ICT/SW Creative research program (NIPA-2014-H0510-14-1019) supervised by the NIPA (National ICT Industry Promotion Agency). Bumsub Ham is supported by the research grant for the European Research Council, VideoWorld.).

References

[1] http://vision.middlebury.edu/stereo/.

[2] http://seungryong.github.io/DASC/.

[3] A. Alahi, R. Ortiz, and P. Vandergheynst. Freak : Fast retina keypoint. *In Proc. of CVPR*, 2012.

[4] M. J. Black, G. Sapiro, D. H. Marimont, and D. Heeger. Robust anisotropic diffusion. *IEEE Trans. IP*, 7(3):421–432, 1998.

[5] Y. Boykov, O. Yeksler, and R. Zabih. Fast approximation enermgy minimization via graph cuts. *IEEE Trans. PAMI*, 23(11):1222–1239, 2001.

[6] M. Brown and S. Susstrunk. Multispectral sift for scene category recognition. *In Proc. of CVPR*, 2011.

[7] T. Brox and J. Malik. Large displacement optical flow: Descriptor matching in variational motion estimation. *IEEE Trans. PAMI*, 33(3):500–513, 2011.

[8] D. Butler, J. Wulff, G. Stanley, and M. Black. A naturalistic open source movie for optical flow evaluation. *In Proc. of ECCV*, 2012.

[9] M. Calonder. Brief : Computing a local binary descriptor very fast. *IEEE Trans. PAMI*, 34(7):1281–1298, 2011.

[10] C. Chang and C. Lin. Libsvm: A library for support vector machines. *ACM Trans. IST*, 2(3):1–27, 2011.

[11] B. Fan, Q. Kong, T. Trzcinski, and Z. Wang. Receptive fields selection for binary feature description. *IEEE Trans. IP*, 23(6):2583–2595, 2014.

[12] E. Gastal and M. Oliveira. Domain transform for edge-aware image and video processing. *In Proc. of ACM SIGGRAGH*, 2011.

[13] Y. HaCohen, E. Shechtman, D. B. Goldman, and D. Lischinski. Non-rigid dense correspondence with applications for image enhancement. *In Proc. of ACM SIGGRAGH*, 2011.

[14] Y. HaCohen, E. Shechtman, and E. Lishchinski. Deblurring by example using dense correspondence. *In Proc. of ICCV*, 2013.

[15] K. He, J. Sun, and X. Tang. Guided image filtering. *IEEE Trans. PAMI*, 35(6):1397–1409, 2013.

[16] P. Heinrich, M. Jenkinson, M. Bhushan, T. Matin, V. Gleeson, S. Brady, and A. Schnabel. Mind: Modality indepdent neighbourhood descriptor for multi-modal deformable registration. *Medi. Image Anal.*, 16(3):1423–1435, 2012.

[17] Y. Heo, K. Lee, and S. Lee. Robust stereo matching using adaptive normalized cross-correlation. *IEEE Trans. PAMI*, 33(4):807–822, 2011.

[18] Y. Heo, K. Lee, and S. Lee. Joint depth map and color consistency estimation for stereo images with different illuminations and cameras. *IEEE Trans. PAMI*, 35(5):1094–1106, 2013.

[19] M. Irani and P. Anandan. Robust multi-sensor image alignment. *In Proc. of ICCV*, 1998.

[20] J. Kim, C. Liu, F. Sha, and K. Grauman. Deformable spatial pyramid matching for fast dense correspondences. *In Proc. of CVPR*, 2013.

[21] D. Krishnan and R. Fergus. Dark flash photography. *In Proc. of ACM SIGGRAGH*, 2009.

[22] C. Lee, A. Bhardwaj, V. Jagadeesh, and R. Piramuthu. Region-based discriminative feature pooling for scene text recognition. *In Proc. of CVPR*, 2014.

[23] H. Lee and K. Lee. Dense 3d reconstruction from severely blurred images using a single moving camera. *In Proc. of CVPR*, 2013.

[24] C. Liu, J. Yuen, and A. Torralba. Sift flow: Dense correspondence across scenes and its applications. *IEEE Trans. PAMI*, 33(5):815–830, 2011.

[25] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[26] S. Paris and F. Durand. A fast approximation of the bilateral filter using a signal processing approach. *IJCV*, 81(1):24–52, 2009.

[27] G. Petschnigg, M. Agrawals, and H. Hoppe. Digital photography with flash and no-flash iimage pairs. *In Proc. of ACM SIGGRAGH*, 2004.

[28] P. Pinggera, T. Breckon, and H. Bischof. On cross-spectral stereo matching using dense gradient features. *In Proc. of BMVC*, 2012.

[29] J. Pluim, J. Maintz, and M. Viergever. Mutual information based registration of medical images: A survey. *IEEE Trans. MI*, 22(8):986–1004, 2003.

[30] T. Portz, L. Zhang, and H. Jiang. Optical flow in the presence of spatially-varying motion blur. *In Proc. of CVPR*, 2012.

[31] C. Rhemann, A. Hosni, M. Bleyer, C. Rother, and M. Gelautz. Fast cost-volume filtering for visual correspondence and beyond. *In Proc. of CVPR*, 2011.

[32] S. Saleem and R. Sablatnig. A robust sift descriptor for multispectral images. *IEEE SPL*, 21(4):400–403, 2014.

[33] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 47(1):7–42, 2002.

[34] E. Schechtman and M. Irani. Matching local self-similarities across images and videos. *In Proc. of CVPR*, 2007.

[35] P. Sen, N. K. Kalantari, M. Yaesoubi, S. Darabi, D. B. Goldman, and E. Shechtman. Robust patch-based hdr reconstruction of dynamic scenes. *In Proc. of ACM SIGGRAGH*, 2012.

[36] X. Shen, L. Xu, Q. Zhang, and J. Jia. Multi-modal and multispectral registration for natural images. *In Proc. of ECCV*, 2014.

[37] D. Sun, S. Roth, and M. Black. Secret of optical flow estimation and their principles. *In Proc. of CVPR*, 2010.

[38] E. Tola, V. Lepetit, and P. Fua. Daisy: An efficient dense descriptor applied to wide-baseline stereo. *IEEE Trans. PAMI*, 32(5):815–830, 2010.

[39] A. Torabi and G. Bilodeau. Local self-similarity-based registration of human rois in pairs of stereo thermal-visible videos. *Pattern Recognition*, 46(2):578–589, 2013.

[40] Q. Yan, X. Shen, L. Xu, and S. Zhuo. Cross-field joint image restoration via scale map. *In Proc. of ICCV*, 2013.

[41] Q. Yang, K. Tan, and N. Ahuja. Real-time o(1) bilateral filtering. *In Proc. of CVPR*, 2009.

[42] Y. Ye and J. Shan. A local descriptor based registration method for multispectral remote sensing images with non-linear intensity differences. *ISPRS J. Photogram. Remote Sens.*, 90(7):83–95, 2014.

[43] R. Zabih and J. Woodfill. Non-parametric local transforms for computing visual correspondence. *In Proc. of ECCV*, 1994.