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, Korea [2]Chungnam Nat. University, Korea [3]ADSC, Singapore [4]Inria, France [5]UIUC, USA
http://seungryong.github.io/DASC/

Recently, many computer vision and computational photography problems have been reformulated to overcome an inherent limitation by leveraging multi-modal and multi-spectral images. Estimating dense visual correspondence for these images is a key enabler for realizing such tasks. To define a matching fidelity term, conventional descriptors 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 not hold. In these cases, conventional descriptors or similarity measures often fail to capture reliable matching evidence.

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.

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. The LSS descriptor $\mathcal{D}_i^{\text{LSS}}$ [1] measures a correlation between two patches $\mathcal{F}_i$ and $\mathcal{F}_j$ centered at pixel i and j within a local support window $\mathcal{R}_i$. As shown in Fig. 1, 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. However, it provides unsatisfactory results in densely matching multi-modal images. 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.

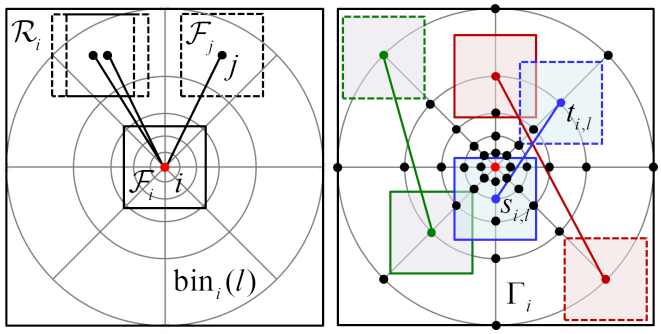
Instead of using a center-biased max pooling of LSS descriptor [1] in Fig. 1(a), our DASC descriptor incorporates a randomized receptive field pooling with sampling patterns in such a way that a pair of two patches are randomly selected within a local support window. Our approach encodes a similarity between patch-wise receptive fields sampled from log-polar circular point set Γ_i as shown in Fig. 1(b). It is defined as $\Gamma_i = \{j | j \in \mathcal{R}_i, |i-j| = \rho_r, \angle(i-j) = \theta_a\}$, and has a higher density of points near a center pixel. 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:

$$d_{i,l} = \mathcal{C}(s_{i,l}, t_{i,l}), \quad s_{i,l}, t_{i,l} \in \Gamma_i, \tag{1}$$

where s_l and t_l are l^{th} selected sampling patterns.

Finding a optimal randomized sampling pattern is a critical issue in our descriptor. 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.

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 $(s,t) \in \cup_i^L$, we compute the adaptive self-correlation $\Psi(s,t)$ between two patches $\mathcal{F}_s$ and $\mathcal{F}_t$ as follows:



(a) (b)

Figure 1: Demonstration of (a) LSS descriptor [1] 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.

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

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

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:

$$\mathcal{C}(s,t) = \max(\exp(-(1 - |\Psi(s,t)|)/\sigma), \tau), \tag{3}$$

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.

For densely constructing our descriptor on an entire image, a straightforward computation can be extremely time-consuming. To alleviate these limitations, we simplify (2) by considering only the weight $w_{s,s'}$ from the source patch $\mathcal{F}_s$ so that a fast computation of (2) using fast edge-aware filter is feasible. Furthermore, for efficient description, we also re-arrange 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 (2) is then approximated as follows:

$$\tilde{\Psi}(i,j) = \frac{\sum_{i',j'} \omega_{i,i'} (f_{i'} - \mathcal{G}_i)(f_{j'} - \mathcal{G}_{i,j})}{\sqrt{\sum_{i'} \omega_{i,i'} (f_{i'} - \mathcal{G}_i)^2} \sqrt{\sum_{i',j'} \omega_{i,i'} (f_{j'} - \mathcal{G}_{i,j})^2}}, \tag{4}$$

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. All these components can be efficiently computed using a constant-time edge-aware filter (EAF).

Experimental results show that our DASC descriptor outperforms conventional area-based approaches and feature-based approaches on various benchmarks; 1) Middlebury stereo benchmark consisting of images with varying illumination and exposure conditions, 2) multi-modal and multi-spectral dataset including RGB-NIR images, different exposure, flash-noflash images, and blurry images, and 3) MPI optical flow benchmark containing motion blur and illumination changes.

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

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.