Domain-Size Pooling in Local Descriptors: DSP-SIFT

Jingming Dong, Stefano Soatto
UCLA Vision Lab, University of California, Los Angeles, CA 90095

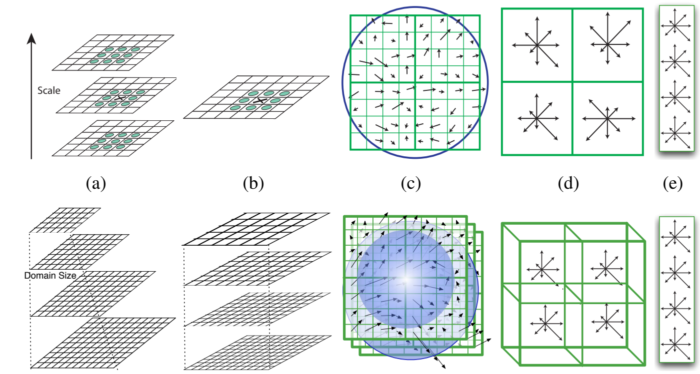

Figure 1: In SIFT (top) isolated scales are selected (a) and the descriptor constructed from the image at the selected scale (b) by computing gradient orientations (c) and pooling them in spatial neighborhoods (d) yielding histograms that are concatenated and normalized to form the descriptor (e). In DSP-SIFT (bottom), pooling occurs across different domain sizes (a): Patches of different sizes are re-scaled (b), gradient orientation computed (c) and pooled across locations *and* scales (d), and concatenated yielding a descriptor (e) of the same dimension of ordinary SIFT.

A "cell" of a SIFT descriptor, computed on an image I in a region of size $\hat{\sigma}$ around a location x, can be written as

$$h_{\text{SIFT}}(\theta|I,\hat{\sigma})[x] = \int \mathcal{N}_{\varepsilon}(\theta - \angle\nabla I(y))\mathcal{N}_{\hat{\sigma}}(y-x)d\mu(y) \qquad (1)$$

where $d\mu(y) \doteq \|\nabla I(y)\|dy$, θ is the independent variable, ranging from 0 to 2π, corresponding to an orientation histogram bin of size ε, and $\hat{\sigma}$ is the *spatial pooling scale*. The kernel $\mathcal{N}_{\varepsilon}$ is bilinear of size ε and $\mathcal{N}_{\hat{\sigma}}$ separable-bilinear of size $\hat{\sigma}$. Both the location x and the scale $\hat{\sigma}$ are typically sampled by a co-variant detector, or regularly as in "dense SIFT." Spatial pooling, interpreted as local marginalization against the kernel $\mathcal{N}_{\hat{\sigma}}(y-x)$, affords insensitivity to small translations around the sampled location x.

But while translations are locally marginalized around the sample x, changes of scale around the sampled $\hat{\sigma}$ are not.

DSP-SIFT is designed to obviate this asymmetry of treatment, by locally marginalizing scale, in addition to translation. If $s > 0$ and $\mathcal{E}$ is an exponential or other unilateral density function, the process can be written as

$$h_{\text{DSP}}(\theta|I)[x] = \int h_{\text{SIFT}}(\theta|I,\sigma)[x]\mathcal{E}_s(\sigma)d\sigma \quad x \in \Lambda \qquad (2)$$

as illustrated in Fig. 1 and implemented in few lines of code. DSP-SIFT has the same dimension of SIFT and improves its performance by 10% to 40% mean-average precision (mAP) on the datasets we tested. It also outperforms a deep convolutional architecture (CNN) in wide-baseline matching tasks, despite having a considerably smaller size and requiring no training (Fig. 2).

DSP-SIFT pools gradient orientations in regions of different size, hence the name *domain-size pooling*, in apparent violation of the principles of scale-space theory and scale selection. There, the *size* of the region where statistics are computed is *tied* to the spatial frequencies of the image, which facilitates correspondence under changes of scale or distance. However, this does not take into account occlusions: How large a portion of a scene is visible in each corresponding image(s) depends on the *shape* of the scene, the *pose* of the two cameras, and the resulting visibility (*occlusion*) relations, not on the spatial frequencies of the image. Therefore, we *untie* the size of the domain where the descriptor is computed ("scale") from photometric characteristics of the image (Fig. 3). While somewhat unintuitive, as histogram bins mix different regions of the same image, this procedure is rooted in classical sampling theory and the practice of *anti-aliasing*.

Domain-size pooling can be applied to a range of other low-level vision operations, such as in other histogram-based representations, including the lower layers of convolutional neural networks [2]. A more detailed derivation of DSP-SIFT and its relation to sampling theory is described in [1], and the implementation is available at `vision.ucla.edu/dsp-sift`.

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

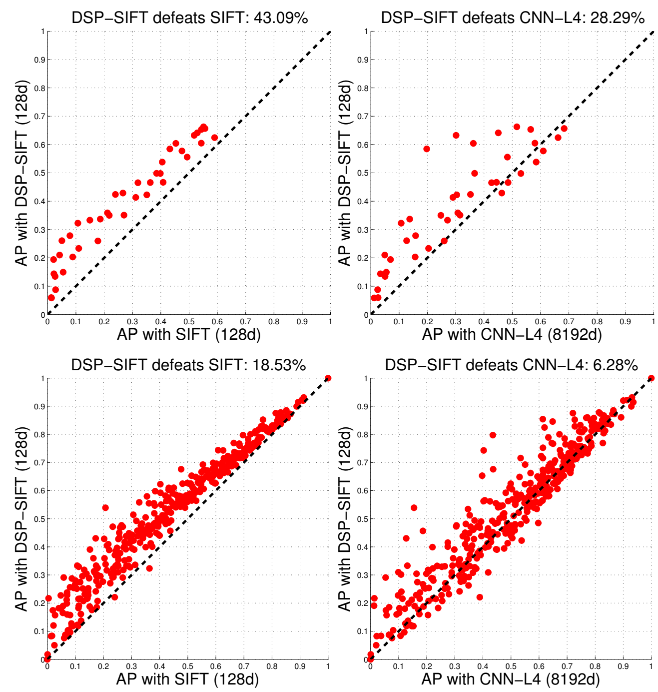

Figure 2: Each point represents a pair of images from two benchmarks. Left: DSP-SIFT consistently improves the original SIFT (relative improvement shown in title). Right: DSP-SIFT also outperforms the CNN descriptors without increase in dimension (shown in axis).

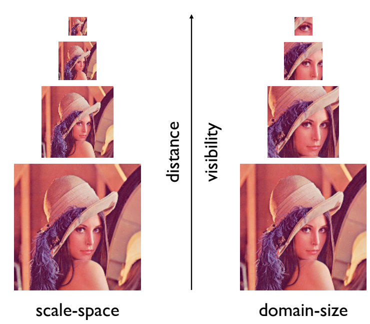

Figure 3: *Scale-space* refers to a continuum of images obtained by smoothing and downsampling a base image. It is relevant to searching for correspondence when the distance to the scene changes. *Size-space* refers to a scale-space obtained by maintaining the same scale of the base image, but considering subsets of it of variable size. It is relevant to searching for correspondence in the presence of occlusions, where the size (and shape) of co-visible domains are not known.

[1] J. Dong and S. Soatto. Domain-size pooling in local descriptors: DSP-SIFT (extended version of this paper), *ArXiv preprint:1412.8556*, 2014.

[2] S. Soatto, J. Dong, and K. Karianakis. Visual scene representations: Scaling and occlusion in convolutional architectures. *ArXiv preprint:1412.6607*, 2014.