Multi-View Feature Engineering and Learning

Jingming Dong, Nikolaos Karianakis, Damek Davis, Joshua Hernandez, Jonathan Balzer, Stefano Soatto
UCLA Vision Lab, University of California, Los Angeles, CA 90095

We frame the design of feature descriptors as the approximation of an *"ideal representation"* consisting of a minimal sufficient statistic (the likelihood function of the scene given imaging data) that is made invariant to nuisance factors such as vantage point and local contrast transformations via marginalization. We employ the Lambert-Ambient (LA) Model, the simplest known to capture the phenomenology of image formation for the purpose of correspondence, including scaling and occlusion phenomena. We then seek to relate existing descriptors, such as SIFT and HOG, to the ideal representation, as well as to develop better approximations of it.

The first consequence of our derivation is that existing descriptors computed from *a single image* can only approximate the ideal representation under very restrictive conditions. These include assuming that the scene is flat, parallel to the image plane, and only allowed to translate parallel to it. The likelihood function of (a scene that is flat and translating parallel to the image plane, which can then be though of as) an image I is then approximated at each point by a "cell" of SIFT/HOG:

$$\begin{aligned} h_x(\theta|I) &= \int_G p_{x,G}(\theta|I,v)dP(v) = \\ &= \int_{\mathbb{R}^2} \mathcal{N}_\varepsilon(\theta - \angle\nabla I_v(x))\mathcal{N}_\sigma(v)d\mu(v|I_v) \\ &= \int_{\mathbb{R}^2} \mathcal{N}_\varepsilon(\theta - \angle\nabla I(y))\mathcal{N}_\sigma(y-x)\|\nabla I(y)\|dy \end{aligned} \quad (1)$$

where $v \in G$ is the group of planar translations, θ is the free variable, and $\mathcal{N}_\alpha$ is an isotropic Gaussian or bilinear kernel with parameter $\alpha = \varepsilon, \sigma$. The assumptions underlying this approximation break in the presence of scaling and occlusion phenomena.

The second observation is that, by leveraging on *multiple images* of the same underlying scene, we can develop better approximations of an ideal (local) representation. Multiple images afford the ability of separating *nuisance variability* (due to vantage point, illumination, partial occlusions) from *intrinsic variability*, and therefore better trade off insensitivity to the former with discriminative power.

We thus introduce two multi-view local representations. The first is based on a sampling approximation of the ideal descriptor, and is named *multi-view HoG* (MV-HoG). It is a natural extension of single-view descriptors based on histogram of gradient orientations such as SIFT, HOG and their variants:

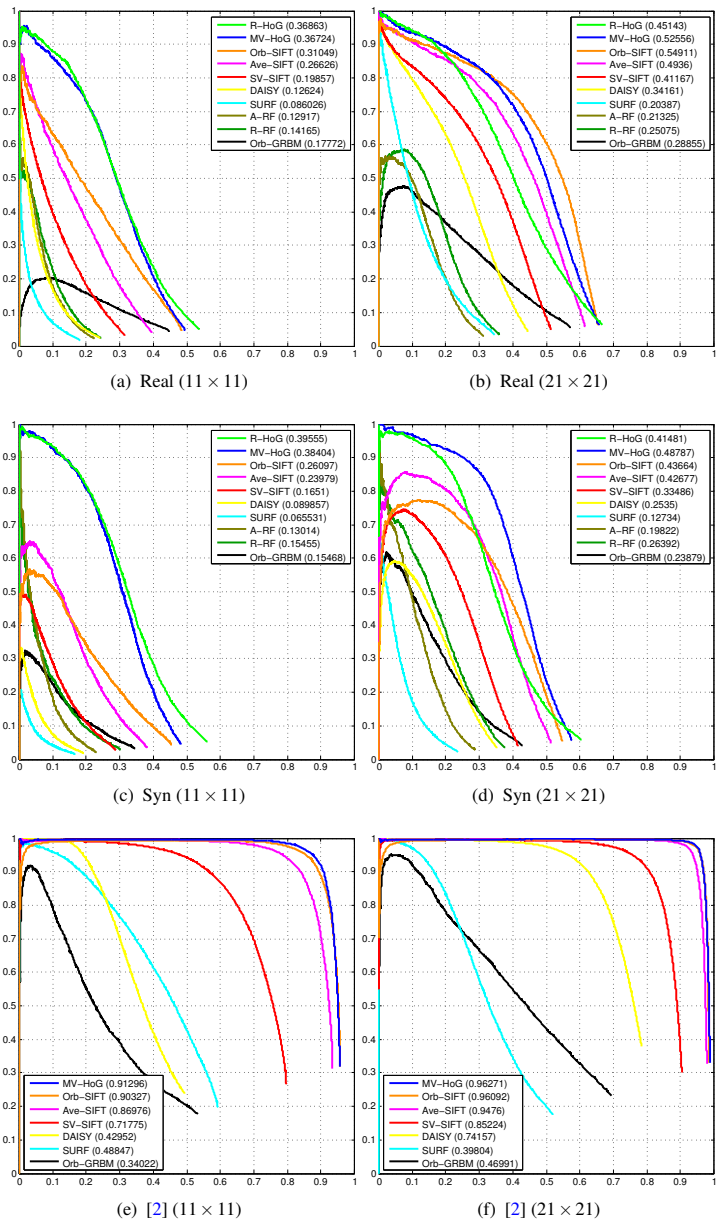
$$h_{x,G}(\theta|\{I_t\}_{t=1}^T) \doteq \frac{1}{T}\sum_{t=1}^{T}\int_{\mathbb{R}^2} \mathcal{N}_\varepsilon(\theta - \angle\nabla I_t(y))\mathcal{N}_\sigma(y-x)\mathcal{E}_s(\sigma)d\mu(y)d\sigma.$$

where the images are assumed to represent a *sufficiently exciting* sample of vantage points. Note that, in addition to small translations parallel to the image plane, translations orthogonal to it, resulting in a change of scale σ, are also marginalized relative to a density $\mathcal{E}$ with range parameter s. The temporal averaging is then charged with marginalizing rotations of the surface, by sampling the view sphere.

Alternatively, one can use the samples to infer a *point-estimate* of the scene's geometry, $\hat{S} \subset \mathbb{R}^3$ and reflectance (albedo) $\hat{\rho} : S \to \mathbb{R}^+$, and then marginalize rotations ($SO(3)$) explicitly, giving raise to *reconstructive HoG*, or R-HoG:

$$\begin{aligned} &h_{x,G}(\theta|\hat{\rho},\hat{S}) = \\ &\int_{SO(3)\times\mathbb{R}^2} \mathcal{N}_\varepsilon(\theta - \angle\nabla\hat{\rho}\circ g\circ\pi_{\hat{S}}^{-1}(y))dP_{SO(3)}(g)\mathcal{N}_\sigma(y-x)\mathcal{E}_s(\sigma)d\mu(y)d\sigma \end{aligned}$$

Although these multi-view descriptors are very crude approximations of an ideal representation, they still out-perform descriptors computed in a single-view such as single-view SIFT (SV-SIFT), SURF, DAISY, as well as a gated restricted Boltzmann machine (GRBM) developed for correspondence purposes. In theory, they should not outperform an orbit of descriptors computed independently on each of the images $\{I_t\}_{t=1}^T$ (Orb-SIFT), although in practice they often do because of the unusual way in which many descriptors are normalized (for instance using the ℓ^2 norm). Convolutional architectures can also be related to an ideal representation, as we further explore in [1]. A restriction of MV-HoG to a single image is presented in [3].



(a) Real (11×11) (b) Real (21×21)

(c) Syn (11×11) (d) Syn (21×21)

(e) [2] (11×11) (f) [2] (21×21)

Figure 1: *Precision-Recall Curves.* Precisions (ordinate) over recall rates (abscissa) with F1-scores in the legends.

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

[2] S. A. Winder and M. Brown. Learning local image descriptors. In *Proceedings of IEEE Conference on Computer Vision & Pattern Recognition*, 2007.

[3] J. Dong and S. Soatto. Domain-size pooling in local descriptors: DSP-SIFT. In *Proc. IEEE Conf. on Comp. Vision and Pattern Recogn.*, 2015.

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