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
{dong,soatto}@cs.ucla.edu, {nikarianakis,jbalzer}@ucla.edu, {damek,jheez}@math.ucla.edu

Abstract

We frame the problem of local representation of imaging data as the computation of minimal sufficient statistics that are invariant to nuisance variability induced by viewpoint and illumination. We show that, under very stringent conditions, these are related to "feature descriptors" commonly used in Computer Vision. Such conditions can be relaxed if multiple views of the same scene are available. We propose a sampling-based and a point-estimate based approximation of such a representation, compared empirically on image-to-(multiple)image matching, for which we introduce a multi-view wide-baseline matching benchmark, consisting of a mixture of real and synthetic objects with ground truth camera motion and dense three-dimensional geometry.

1. Introduction

For visual data, a "feature descriptor" is a function of images designed to be "insensitive" to nuisance variability and yet "discriminative" with respect to intrinsic properties of the scene or object of interest. Nuisance variability may be due to changes of viewpoint and illumination, and intrinsic properties include three-dimensional shape and material properties of the scene, or object-specific deformations. The best-known local descriptors are SIFT [20], HOG [7] and their variants [3], which we refer to collectively as HoG (histogram of gradient): For an image region centered at a point, they are histograms of the orientation of its gradient in that region, variously normalized.

On the other hand, representation learning via neural networks [18] constructs functions that are insensitive to nuisance variability by training a convolutional architecture supported on the entire image domain. There have been several studies of the empirical performance of local feature descriptors, including their comparison [24], and their generative abilities [37, 29]. However, efforts to elucidate their relationships have only recently begun to appear [5, 4]. But what is an ideal representation? In terms of being "discriminative" of the intrinsic properties of the scene, such as its shape and reflectance, one could do no better than a (minimal) sufficient statistic, for instance the likelihood function [30]. In terms of being "insensitive" to nuisance factors, such as viewpoint and illumination, one could do no better than a (maximal) invariant to their action on the data. So, *an ideal representation would be a minimal sufficient statistic that is maximally invariant to nuisance factors* [30].
Does such a representation exist? If so, can it be computed? If not, can it be approximated? Can existing descriptors be related to it? If so, under what conditions? If not, how can we construct better approximations of an ideal representation?

1.1. Related Work

There are many engineered descriptors of *one* image [20, 7, 3, 35], that differ on where and how the local histograms are aggregated and normalized, with many implementation details affecting performance [6]. Some entail learning [39, 19] to minimize classification (correspondence) error. Relatively few local descriptors aggregate multiple views: [9] combines spatial (averaged SIFT) and temporal statistics; [16] performs feature selection from trajectories of key points. Deformable parts models [14] are also learned from multiple views to capture intrinsic variability.

One could also learn away nuisance variability through a neural network architecture [18, 27]. This approach has been steadily improving performance in large-scale pattern recognition [10], but not in correspondence, where it is outperformed by engineered descriptors, even some built using a single image [12]. Rather than performing direct comparison between different descriptors, we instantiate an *ideal local representation* relative to a simple image-formation (Lambert-Ambient, or LA) model, and relate various descriptors to it.

1.2. Summary

To quantify how "discriminative" a descriptor is, we characterize its dependency on intrinsic properties of the scene, namely shape S and reflectance[1] ρ. To quantify how "insensitive" it is, we describe its dependency on nuisance factors such as viewpoint and illumination. In [22] the LA model is described as the simplest to capture the phenomenology of image formation for the purpose of correspondence. Local illumination changes are modeled, to first-order approximation, as monotonic continuous transformations of the range of the image, also known as *contrast transformations*. They form a group[2], and under certain conditions [31] the gradient orientation is a maximal invariant. So we can eliminate first-order dependency on illumination by replacing the image[3] I with its gradient orientation $\theta(x) = \angle\nabla I(x) \doteq \nabla I(x)/\|\nabla I(x)\|$, at locations x where $\nabla I(x) \neq 0$. For a local neighborhood $\mathcal{B} \subset \mathbb{R}^2$, the *likelihood function,* computed at a location $x \in \mathcal{B}$ and conditioned on a given shape S and reflectance ρ, is a minimal sufficient statistic [30], and can be thought of as a probability density on θ, $p_{\mathcal{B}}(\theta|\rho, S)$ with marginals[4] $p_x(\theta|\rho, S)$. If there are additional groups G acting on the scene (for instance changes of spatial position and orientation, $G = SE(3)$) they can be marginalized, thus obtaining a density

$$p_{x,G}(\theta|\rho, S). \tag{1}$$

The marginalized likelihood is a maximal contrast-invariant that is also G-invariant. With respect to this *ideal representation*, our goals are to: (i) Instantiate the formal notation above using the LA model and derive an expression for (1) suitable for computation (Sec. 2.1). (ii) Show that HoG approximates an ideal descriptor when the scene is planar and the viewer is constrained to translating parallel to it (Sec. 2.1). (iii) Derive a sampling approximation of (1), which we call MV-HoG, where the scene (S, ρ) is replaced with a collection of images of it, captured from multiple viewpoints $\{I_t\}_{t=1}^T$ (Sec. 3.1). (iv) Derive a point-estimate based approximation of (1), which we call R-HoG, where the scene (S, ρ) is replaced with a point estimate $(\hat{S}, \hat{\rho})$ reconstructed from a finite sample $\{I_t\}_{t=1}^T$, possibly using structured illumination (Sec. 3.2).

[1] In the LA model $S \subset \mathbb{R}^3$ is a multiply-connected piecewise smooth surface in Euclidean space, and $\rho : S \to \mathbb{R}^+$ is a positive-valued scalar function called "albedo." As we model illumination via contrast transformations of the albedo, we interpret ρ modulo contrast changes as the *reflectance* of the surface S.

[2] If strictly monotonic, lest they form a monoid.

[3] Here $I : D \subset \mathbb{R}^2 \to \mathbb{R}^+;\ x \mapsto I(x)$ is a gray-scale image, $x \in D$ is a point on the plane. In practice, I takes a finite number of values on a quantized domain, extended to the entire plane by zero-padding.

[4] If we knew the viewpoint, under the assumptions of the LA Model, the conditional density would be spatially independent, (10); otherwise, marginalizing viewpoint introduces spatial dependency, so the product of the marginals is only an approximation, (12).

2. Engineered Features Revisited

A "cell" of the HOG/SIFT descriptor[5] h of an image I in a region centered at a pixel x is a histogram of the orientation of its gradient, θ, around x. If the histogram is not normalized, we call it uHoG (un-normalized HoG) and indicate it with

$$h_x(\theta|I) \qquad \text{uHoG}. \tag{2}$$

Given one image I, this un-normalized histogram returns a positive number for each orientation θ, related to the number of pixels around x where the image gradient orientation is close to θ. Variants of HoG differ in where they compute and how they aggregate and normalize such histograms. For instance, SIFT [7] evaluates the histogram above on a 4×4 grid $\mathcal{B} = \{x_i,\ i = 1, \dots, 16\}$, and concatenates the result into a vector $[h_{x_1}, \dots, h_{x_{16}}]$, that is then normalized, clamped, and re-normalized. Discrete bins are computed using a bilinear interpolation kernel κ_ϵ with $\epsilon = 2\pi/\#\text{bins}$, and a linear spatial weighting kernel κ_σ with σ the area of each cell in the 4×4 grid, further weighted by the magnitude of the image gradient $\|\nabla I\|$. If we extend the sum to the continuum, we can write the histogram in each cell as [36, 11]

$$h_x(\theta|I) = \int \kappa_\epsilon\big(\theta - \angle\nabla I(y)\big)\kappa_\sigma(x - y)\|\nabla I(y)\| dy \tag{3}$$

where the argument of the orientation kernel is intended modulo 2π. Alternatively, histograms can be normalized independently at each location x:

$$\bar{h}_x(\theta|I) = \frac{h_x(\theta|I)}{\int_{\mathbb{S}^1} h_x(\theta|I) d\theta}, \qquad h = [h_{x_1}, h_{x_2}, \dots, h_{x_i}, \dots]. \tag{4}$$

Note that in HoG, described above, the nuisance group G is absent. We introduce it next.

2.1. Ideal descriptor of one view and its HoG

As a preliminary step to computing the minimal sufficient invariant statistic (1), and to understand its relation to single-view descriptors, consider a special case obtained by assuming that the scene is a plane parallel to the image plane, with albedo equal to the image irradiance. Then, conditioning on the image I, we have $p_{x,G}(\theta|I)$, which we wish to relate to uHoG (2).

To guarantee contrast-invariance, one could replace the intensity $I(x) \in \mathbb{R}^+$ with the curvature of the isocontours [1], or with its dual, the orientation of the gradient, $\angle\nabla I(x) \in \mathbb{S}^1$ where $\nabla I(x) \neq 0$. Let (G, P) be a probability space, with G a group and P a probability distribution on the group, and suppose that to each $g \in G$ we can associate a "transformed" image I_g. For each pixel $x \in \mathbb{R}^2$ where

[5] Here $\theta \in \mathbb{S}^1$ is an angle (the free variable) and $h : D \times \mathbb{S}^1 \to \mathbb{R}^+; (x, \theta) \mapsto h_x(\theta)$ for a fixed image I.

$\nabla I_g(x) \neq 0$, we can then define a (marginal) probability density function over θ, for instance:

$$p_{x,G}(\theta|I,g) \doteq \mathcal{N}_\varepsilon(\theta - \angle\nabla I_g(x)) \quad (5)$$

where the difference is intended in $\mathbb{S}^1$, and correspondingly $\mathcal{N}_\epsilon$ denotes an angular Gaussian [38]. Kernels κ other than Gaussian can also be considered without significant changes to the arguments that follow. Using P, we can marginalize[6] this distribution to eliminate its dependency on $g \in G$:

$$p_{x,G}(\theta|I) \doteq \int_G p_{x,G}(\theta|I,g)dP(g). \quad (6)$$

To understand the relationship with uHoG, we restrict G to be the group of planar translations, $G = \mathbb{R}^2$, and choose a particular measure for $\mathbb{R}^2$, $d\mu(v|I) \doteq \|\nabla I_v(x)\|dv$ where, if $v \in G$, $I_v(x) = I(x+v)$ is the transformed image. We then marginalize with respect to the (un-normalized) distribution $dP(v) = \mathcal{N}_\sigma(v)d\mu(v|I_v)$. This corresponds to assuming that the scene is *flat*, parallel to the image-plane (fronto-parallel) and constrained to translate parallel to it. The likelihood function is given by $p_{x,G}(\theta|I,v) = \mathcal{N}_\varepsilon(\theta - \angle\nabla I_v(x))$. Integrating against $dP(v)$, we obtain

$$\begin{aligned} h_x(\theta|I) &= \int_G p_{x,G}(\theta|I,v)dP(v) = \\ &= \int_{\mathbb{R}^2} \mathcal{N}_\varepsilon(\theta - \angle\nabla I_v(x))\mathcal{N}_\sigma(v)d\mu(v|I_v) \\ &= \int_{\mathbb{R}^2} \mathcal{N}_\varepsilon(\theta - \angle\nabla I(y))\mathcal{N}_\sigma(y-x)\|\nabla I(y)\|dy, \end{aligned} \quad (7)$$

which is one cell of uHoG (3) once we restrict to the discrete lattice and replace the Gaussian kernels with (bi-)linear ones. The full descriptor is just the concatenation of a number of cells, suitably normalized; for the case of a single cell,

$$p_{x,G}(\theta|I) = \frac{h_{x,G}(\theta|I)}{\int h_{x,G}(\theta|I)d\theta} \quad (8)$$

which leads us to conclude that HOG/SIFT approximates the ideal representation at a point under the assumption that the scene is flat and fronto-parallel, undergoing purely translational motion parallel to the image plane.

3. Ideal Descriptor Approximations

To move one step closer to the ideal representation, and to relax the stringent assumptions implicit in HOG/SIFT, suppose for now that we have complete knowledge of the underlying scene (S,ρ). A pinhole camera projects each point on the scene to the image plane via[7] $\pi : S \to D \subset \mathbb{R}^2$ and its associated inverse $\pi_S^{-1} : D \to S$, where $\pi_S^{-1}(x)$ is the point of the first intersection of the pre-image (a line) of x with the scene S. Under the assumptions of the LA model, there exists an open subset $G_0 \subseteq SE(3)$ with compact closure and – after a suitable change of reference frame – containing the identity, such that each $g \in G_0$, with the action

$$I_g(x) = \rho \circ g \circ \pi_S^{-1}(x) \quad (9)$$

can be associated with a domain diffeomorphism $w_g : \mathbb{R}^2 \to \mathbb{R}^2$, with $I_g(x) = I(w_g(x))$. Here "$\circ$" denotes function composition. When emphasizing the dependency of w_g on shape, we indicate it with $w_g(x|S)$. Let P be a probability measure on G_0, *e.g.,* the normalized restriction of the Haar measure on $SE(3)$ to G_0, which is no longer a group, but a subset of G, where the probability of actions outside G_0 is assigned to zero. Then the marginalized descriptor, for a *known* scene, is given by

$$\begin{aligned} p_{x,G_0}(\theta|\rho,S) &= \int_{G_0} \mathcal{N}_\varepsilon(\theta - \angle\nabla\rho \circ g \circ \pi_S^{-1}(x))dP_{G_0}(g) \\ &= \int_{G_0} \mathcal{N}_\varepsilon(\theta - \angle\nabla I(w_g(x|S)))dP_{G_0}(g). \end{aligned} \quad (10)$$

The first approximation step is to reduce the dimensionality of $G_0 \subset SE(3) = SO(3) \times \mathbb{R}^3$ to simplify marginalization. This can be done locally around a point $\pi_S^{-1}(x)$ through the use of a *co-variant detector,* a function of the image that returns multiple isolated elements of subsets of G_0 that co-vary with g. For instance, a translation-scale detector [20] returns isolated locations on the image plane, x_i, and their corresponding scales σ_i, which can be used to define a local reference frame centered at x_i with unit σ_i. To first approximation, as we qualify in the next paragraph, these co-vary with the translation component of G_0: A spatial translation parallel to the image plane induces a planar translation of x_i, and a spatial translation orthogonal to the image plane induces a change of scale σ_i. Thus, locally around $\pi_S^{-1}(x_i)$, we can annihilate the effects of spatial translation simply by *canonizing* the location-scale group, *i.e.* imposing $x_i = 0, \sigma_i = 1$, by applying the inverse transformation of that determined by the co-variant detector. This procedure can be applied to any planar group transformation, including the entire group of diffeomorphisms [32]. In particular, planar rotation can be canonized using the direction of gravity as a reference [17], leaving only "out-of-plane" rotations to be marginalized in (10).

In reality, spatial translations do not co-vary with planar translation-scale transformations, for the former in-

[6] The integral is well defined by Fubini's theorem; $p_{x,G}(\theta|I,g)$ is a measurable function of g and bounded so the marginalization converges. Thus, we can integrate over θ and exchange the integrals. But while marginalization guarantees invariance to $g \in G$, it does not yield a maximal invariant, which is instead described in [30].

[7] π incorporates the projection by dividing the coordinates of a point in S by the third component and applying a planar affine transformation depending on the intrinsic calibration of the camera [22].

duces (shape-dependent) deformations of the image domain (9) in addition to non-invertible transformations due to *occlusions*, which are absent in the latter. Such shape-dependent image variability is lost in any descriptor computed from a single image: Any finite-dimensional planar group-covariant detector co-varies with spatial translations only when the scene is flat and the neighborhood of size σ_i centered in x_i, $\mathcal{B}_{\sigma_i}(x_i)$, does not straddle occluding boundaries. Fortunately, we are not constrained to building descriptors using a single image; instead, we can capture residual deformations after canonization by marginalizing with respect to out-of-plane rotations in $SO(3)$. In addition, we can also marginalize small residual changes in translation v and scale σ using some prior $P_{\mathcal{N}_\sigma} \times P_{\mathcal{E}_s}$, where[8] $dP_{\mathcal{N}_\sigma}(v) = \mathcal{N}_\sigma(v)d\mu(v)$ and $dP_{\mathcal{N}_s}(\sigma) = \mathcal{E}_s(\sigma)d\sigma$ with $\mathcal{E}$ a unilateral density (*e.g.,* exponential) to ensure $\sigma > 0$. Thus, our un-normalized conditional distribution becomes:

$$h_{x,G}(\theta|\rho,S) = \int_{G_0} \mathcal{N}_\varepsilon(\theta - \angle\nabla I_g(x))dP_{G_0}(g) \simeq \quad (11)$$

$$\int \mathcal{N}_\varepsilon(\theta - \angle\nabla I(w_g(y)))dP_{SO(3)}(g)\mathcal{N}_\sigma(y-x))\mathcal{E}_s(\sigma)d\mu(y)d\sigma.$$

If out-of-plane rotations are neglected, or if the scene is planar, one image is sufficient to construct an idea descriptor, which then reduces to DSP-SIFT, recently introduced in [12]. To obtain the ideal descriptor of *a region* $\mathcal{B}$, we must consider the joint distribution of all pixels within: $h_{x_1,\dots,x_k,G}(\theta_1,\dots,\theta_k|\rho,S)$. Aggregating histograms in high dimensions is challenging but the joint distribution can be approximated by a collection of one-dimensional marginals. The simplest approximation is to neglect spatial correlations altogether: From (10),

$$\begin{aligned} & p_{x_1,\dots,x_k,G_0}(\theta_1,\dots,\theta_k|\rho,S) = \\ = & \int_{G_0} \prod_{i=1}^{k} \mathcal{N}_\varepsilon(\theta_i - \angle\nabla I(w_g(x_i|S)))dP_{G_0}(g) \\ \simeq & \prod_{i=1}^{k} h_{x_i,G}(\theta_i|\rho,S). \end{aligned} \quad (12)$$

As already pointed out[4], under the assumptions of the LA model, if the vantage point $g \in SE(3)$ was known, then the conditional density above would indeed factorize into the product of marginals computed independently at each pixel. However, marginalizing viewpoint introduces spatial dependencies, so the above is just an approximation.[9]

[8]It should be noted that this approximation step does not reduce the generality of the approach: In practice, one would have to discretize the group G_0 anyway in order to perform the marginalization in (10), and covariant detectors are just an adaptive discretization mechanism. A trivial detector is one that returns regular samples of the group, for instance a discretization of planar translations and scales as customary in "dense SIFT." Indeed, this discretization is necessary also to compactify the translational component of G_0, that otherwise would have to be marginalized with respect to an improper measure.

[9]Coarse as it seems, this is nevertheless the approximation implicit in most single-view descriptors, that consider the concatenation of (independently aggregated, scalar) histograms. Some single-view descriptors attempt to recapture some of the lost spatial correlations by joint (re)-normalization [7].

Even this approximation, however, requires knowledge of the scene (S,ρ) to be computed. We now address how to cope with absence of such knowledge.

3.1. Sampling approximation: MV-HoG

If we do not have complete knowledge of the scene, (S,ρ), but we have a collection of images of it $\{I_t\}_{t=1}^T$, we can approximate (11) by Monte-Carlo sampling, after noticing that $I_t(x) = \rho \circ g_t \circ \pi_S^{-1}(x) = I(w_{g_t}(x))$ with $\{w_{g_t}|t = 1,\cdots,T\}$ and $g_t \sim P_{G_0}$ with the restriction G_0 determined by visibility. Under sufficient excitation conditions on the sample $\{I_t\}_{t=1}^T$, asymptotically for $T \to \infty$, we can approximate the integral with

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

Scale σ can also be marginalized as in (11). Sufficient excitation conditions mean that the orbit in $SE(3)$ is sampled along all directions (in the Lie Algebra), which is a tall order, as it requires every surface element to be seen from all vantage points, at all distances, while g_t remains in G_0. This requirement can be mitigated by restricting the marginalization to $SO(3)$ or even to just out-of-plane rotations, using (11) in conjunction with a co-variant detector or other sampling mechanism.

Alternatively, we can use whatever data is available to reconstruct a model (a point estimate) of the scene, which can then be used to render synthetic samples from the orbits of $SE(3)$.

3.2. Point-estimate approximation: R-HoG

Samples $\{I_t\}$ can be used to compute an approximation of ρ, S, for instance in the sense of maximum-likelihood, with suitable regularization [13, 15]

$$\hat{\rho},\hat{S} = \arg\max_{\rho,S,g_t} p(\{I_t\}|\rho,S) + \lambda R(S)$$
$$\text{subject to } I_t = \rho \circ g_t \circ \pi_S^{-1} + n_t \quad (13)$$

where $R(S)$ is, for instance, surface area $\int_S dA$, n_t is white and Gaussian, and λ is a scalar multiplier, and then compute (10) restricted to out-of-plane rotations:

$$h_{x,G}(\theta|\hat{\rho},\hat{S}) = \int_{SO(3)} \mathcal{N}_\varepsilon(\theta - \angle\nabla\hat{\rho}\circ g\circ\pi_{\hat{S}}^{-1}(x))dP_{SO(3)}(g) \quad (14)$$

or its spatially regularized version:

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

or its scale-marginalized version as in (11). Convergence and unbiasedness of the maximum-likelihood estimator ensures convergence of R-HoG to (11). Note that it is possible for the reconstruction to be significantly different from S and yet R-HoG be similar to the ideal descriptor, so long as the re-projections $\hat{\rho} \circ g \circ \pi_{\hat{S}}^{-1}(x)$ are compatible with $w_{g_t}(x|S)$. This can happen, for instance, when $\hat{S}$ differs from S in regions where ρ is constant. Also note that, in theory, two views with non-trivial baseline are sufficient to reconstruct an approximation of $\hat{S}$ and $\hat{\rho}$, locally in the co-visible region. Therefore, R-HoG is preferable when T is small and the sample I_t is unlikely to be sufficiently exciting. Normalized versions of each descriptor are obtained as

$$p(\theta|X) = \frac{h_{x,G}(\theta|X)}{\int h_{x,G}(\theta|X)d\theta}, \tag{15}$$

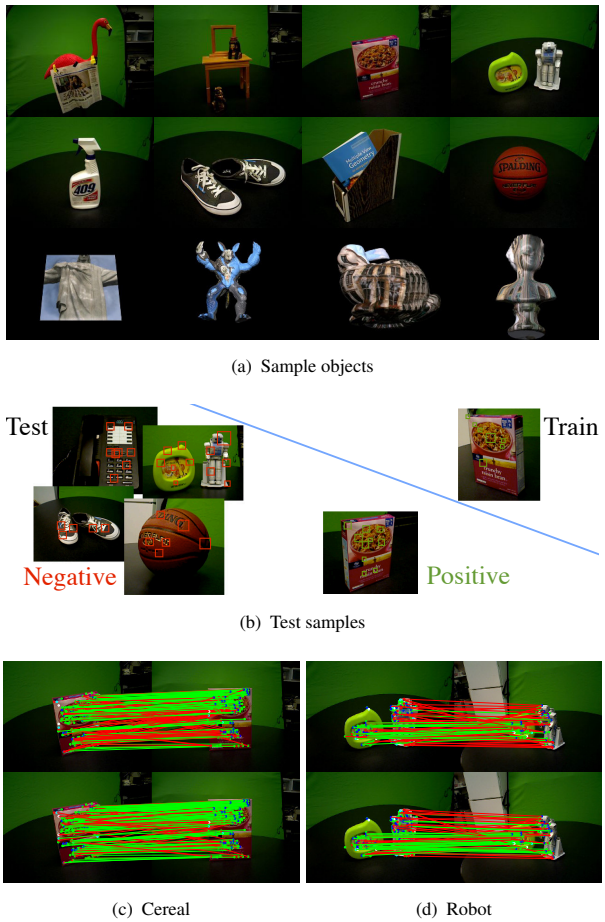
where $X = I$ for HOG, $X = \{I_t\}$ for MV-HoG, $X = \{\hat{\rho}, \hat{S}\}$ for R-HoG, and $X = \{\rho, S\}$ for the ideal descriptor that marginalizes the nuisance assuming a known scene.
While MV-HoG had a stringent sampling requirement, R-HoG has its own challenges, in that obtaining a reliable, dense reconstruction of a scene and its photometry can be a tall order. However, an estimate of the surface is only needed locally, where smooth surfaces can be approximated with parametric models of low order. Also, calibrated reconstruction is not necessary, so a projective reconstruction can be obtained through solving systems of linear equations [22]. Alternatively, a structured model can be inferred through factorization methods such as principal component analysis or sparse coding, whereby S is represented by the coefficients of a linear combination of a collection of "basis elements" $\{S_i\}$.

4. Dataset and Ground Truth

Since our focus here is to leverage on *multiple views* to build better descriptors, which can then be matched to single-images in wide-baseline tests, to perform comparisons we need a dataset where *multiple* training images (of the same scene) are available, whereas correspondence testing can be performed on single images.

Many datasets are available to test image-to-image matching, *e.g.,* [25], where both training and test sets are individual images, each of a different scene. Testing our approach on such datasets would require forgoing marginalization of out-of-plane rotation, thus reducing our approach to DSP-SIFT, which has been tested on [25] by [12].

Fewer datasets are available for testing multi-view descriptors [26, 39]. The latter contains three scenes: Trevi, Half Dome and Notre Dame and provides pixel-level correspondence by back-projecting 3D reconstructed keypoints onto images, which can be used for evaluation. To enable the comparison, we extract a subset containing only features having more than 10 samples. We randomly hold out



(a) Sample objects

(b) Test samples

(c) Cereal

(d) Robot

Figure 1. *Dataset, Test Samples and Qualitative Match Visualization.* (a): Samples from the real and synthetic object dataset. (b): Positive test samples from the object; negative samples are ten-fold more numerous. (c), (d) show correct (green) and wrong (red) matches claimed by SV-SIFT (Top) and MV-HoG (Bottom). The latter yields many more correct matches, similar to R-HoG.

5 samples for testing and use the rest for descriptor aggregation. Negative samples are randomly selected from the other scenes.

Almost perfect results are obtained on [39] (Fig. 2), thus limiting the value of the dataset; we have therefore constructed a new dataset, similar in spirit to [26], but with a *separate* test set and dense ground truth for validation, using a combination of 31 real and 15 synthetic objects. The latter are generated by texture-mapping random images onto surface models available in MeshLab. The former are household objects of the kind seen in Fig. 1. Some with significant texture variability, others with little; some with complex shape and topology, others simple. In each case, a sequence of (training) images per object is obtained by moving around the objects in a closed trajectory. For real objects, a 400-frame-trajectory circumnavigates them to re-

veal most visible surfaces; for synthetic ones, 100 frames span a smaller orbit.

Ground Truth: We compare descriptors built from the (training) video and test single frames, by first selecting test images where a sufficient co-visible area is present. To establish ground truth, we reconstruct a dense model of each (real) object using an RGB-D (structured light) range sensor with YAS [2]. The reconstructed surface enables dense correspondence between co-visible regions in different images by back-projection. This is further validated with standard tools from multiple-view geometry by epipolar RANSAC. Occlusions are determined using the range map. Further implementation details are described in [11].

Detection and Tracking: We use FAST [28] as a mechanism to (conservatively) eliminate regions that are expected to have non-discriminative descriptors, but this step could be forgone. Scale changes are handled in a discrete scale-space, *i.e.* images are downsampled by half up to 4 times and FAST is computed at each level. Short-baseline correspondence is established with standard MLK [21]. A sequence of image locations is returned by the tracker for each region, which is then sampled in a rectangular neighborhood at the scale of the detector. We report experiments on two window sizes, 11×11 and 21×21, illustrative of a range of experiments conducted. The sequence of such windows is then used to compute the descriptors.

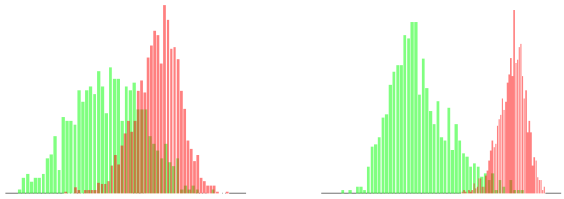
5. Evaluation and Comparison

We briefly describe the descriptors and classifiers involved in the evaluation and refer to [11] for the implementation details, parameter selections and training procedures.

Single-View Descriptors: We use SIFT from [36] as baseline (SV-SIFT), computed on each patch at each frame as determined by the detector and tracker. We also compare single-view descriptor representatives DAISY [35] and SURF-128 [3] computed on the individual images.

Multiple-View Descriptors: MV-HoG is implemented according to Sect. 3.1 using the tracks returned by the MLK tracker. We also tested Random Forest [19] as an alternative way of utilizing multiple samples. We present to the RFs the training samples, and refer to this as A-RF. Deformable parts models would be too slow to test on our dataset, so we forgo that comparison.

Reconstructive Descriptors: To compute an approximation of R-HoG in Sect. 3.2, we compute dense 3-D reconstructions both from some tracked sequences and using a structured-light sensor. Where visual reconstruction was successful, performance was similar, but dense reconstruction was laborious and the quality was not consistent across samples, so to make the evaluation independent of reconstruction methods, we report the results using a structured light sensor only. We use the keyframe where features are first extracted, and sample a viewing hemisphere with 576 vantage points. The R-HoG is built upon these synthesized samples. As in the multiple view case, we also feed synthesized patches to the Random Forest (R-RF).



(a) SV-HoG (b) MV-HoG

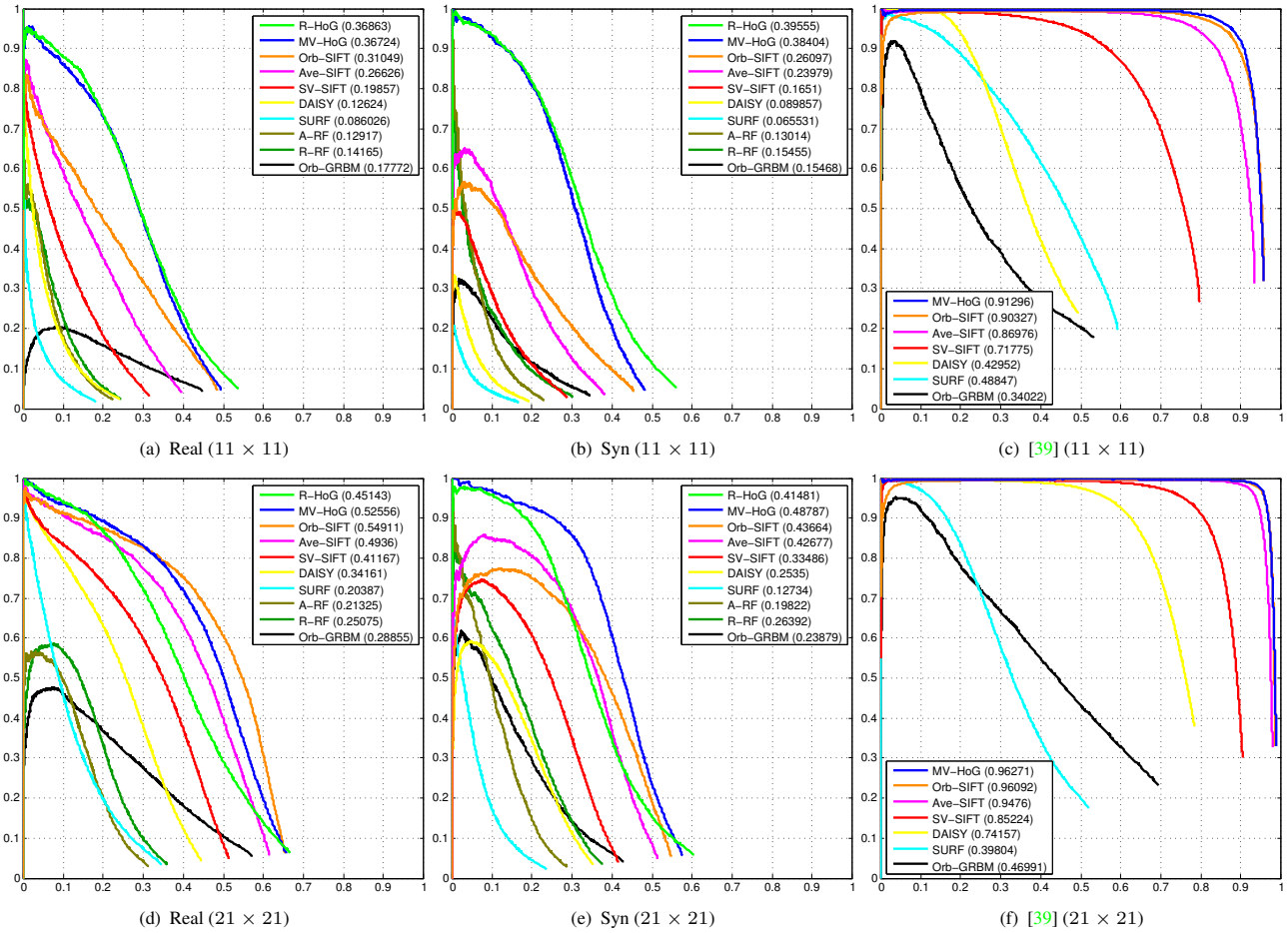
Figure 3. *Distance Distribution.* The horizontal axis indicates the distance between two descriptors in increasing order from left to right. The distribution of distances between corresponding features are shown in green and that of mismatches in red. The error (overlapping area) in 3(b) is considerably smaller than 3(a). This leads to a lower risk of misclassification in MV-HoG.

Classifier and Strategies: Given a descriptor database, the simplest method to match a test query is via *nearest neighbor* (NN) search. We compare five combinations using the same NN search method: (i) single view SV-SIFT, SURF and DAISY – computed on a random image from the training sequence, (ii) Ave-SIFT [9] – averaged SIFT of all frames, (iii) Orb-SIFT – all of the SV-SIFTs stored to represent the orbit which includes the best possible exemplar for each feature [16], (iv) MV-HoG and (v) R-HoG.

Network Architecture: We also compare our methods with a simple network architecture in the form of a gated restricted Boltzmann machine (G-RBM) [23, 34, 33], employed by the authors in correspondence tasks similar to those considered in this paper. We use the same matching strategy as Orb-SIFT, so we call the network Orb-GRBM. Details of the G-RBMs are in [11].

5.1. Metrics

We use precision-recall curves (PR-curves) to quantitatively evaluate the descriptors proposed and compare them to existing methods. For each query patch, nearest neighbor search returns a predicted label and its associated distance. By changing a distance thresh τ_d, a precision-recall curve can be generated. Precision and recall are defined as $p = \frac{\#\text{true matches}}{\#\text{false matches}+\#\text{true matches}}$, $r = \frac{\#\text{true matches}}{\#\text{positive samples}}$. The *positive samples* are the test queries that have correspondences in the training databases as opposed to the *negative* samples which are never seen in training. The *matches* are the queries that pass the distance threshold test. A match is considered to be a *true match* if the predicted label is correct according to the ground truth. As only one predicted label is obtained for each query, r could remain < 1 once any predicted label is wrong. We report the F1-score $\left(\frac{2pr}{p+r}\right)$ for each PR curve. Similarly, random forests (A-RF and R-RF)



(a) Real (11 × 11) (b) Syn (11 × 11) (c) [39] (11 × 11)

(d) Real (21 × 21) (e) Syn (21 × 21) (f) [39] (21 × 21)

Figure 2. *Precision-Recall Curves.* Precisions (ordinate) over recall rates (abscissa) with F1-scores in the legends.

return an averaged probability as a confidence score for the predicted label. A precision-recall curve can be generated by changing a belief threshold τ_p.

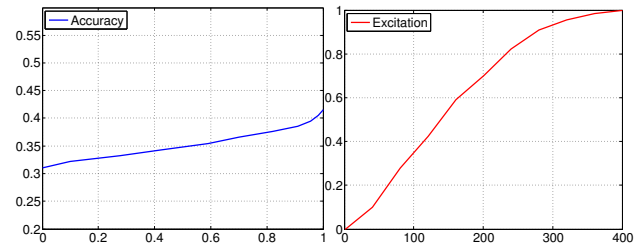
5.2. Empirical Results

Qualitative results are shown in Fig. 1 and [11]. In Fig. 2, PR curves are shown for all the datasets on two different patch sizes. R-HoG and MV-HoG are comparable on 11×11 patches and outperform other methods. On 21×21 patches, the 3D-reconstruction generates artifacts in the view-set generation, so the performance of R-HoG decreases below that of MV-HoG in both the real and synthetic datasets. It should not be surprising that Orb-SIFT performs the best among all the other methods, as it entails exhaustive search over the orbit of transformed views. However, its precision drops sharply when the number of negatives is large, as it inherits the vulnerability of SV-SIFT to outliers. Also, MV-HoG is consistently better than Ave-SIFT across all datasets. Note that both involve averaging histograms, but Ave-SIFT averages *normalized* descriptors computed in each frame, and then re-normalized, whereas MV-HoG aggregates gradient orientation over time, and only normalizes the descriptor at the end, using the same procedure and clamping threshold as Ave-SIFT. This shows that temporal aggregation improves performance compared to simply averaging single-view descriptors computed independently.

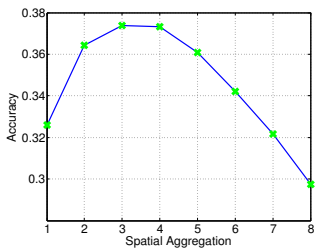
Fig. 3 shows the distance distributions between descriptors of corresponding and non-corresponding patches. SV-HoG is computed from a random single sample from each track, and MV-HoG is aggregated over the whole track. The overlapping area between the two distributions indicates the probability of making a classification error in descriptor matching. The distributions in Fig. 3(b) have much less overlapping area than that in Fig. 3(a). It shows that the discriminative power of the descriptor is improved by aggregating over multiple views.

5.3. Support Region, Spatial Aggregation, Sample Sufficiency and Complexity

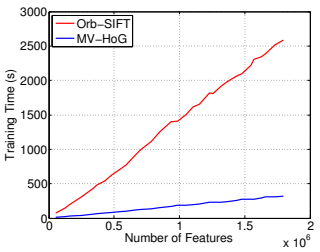
The size of the domain where descriptors are computed impacts performance (Fig. 2): the larger, the better, so long as the domain remains co-visible (*i.e.* $g_t \in G_0$). Fig. 4(b)



(a) Sufficient excitation



(b) Spatial σ



(c) Time complexity

Figure 4. (a) *Sufficient excitation.* Left: Accuracy (maximum recall) as a function of a proxy of sufficient excitation (see text). Right: Excitation as a function of the number of frames. All results are averaged over multiple runs using frames $i, \ldots, i+k-1$ where i is selected at random. (b) F1-score varies with spatial aggregation parameter σ. (c) Time complexity as a function of the number of features with FLANN precision at 0.7. Higher precision will further increase computational load.

shows the effect of the spatial parameter σ in MV-HoG (Sect. 3.1). A slight spatial aggregation enhances robustness until σ reaches a critical value, beyond which discriminative power drops. Multiple view descriptors perform scene-dependent blurring, and therefore remain more discriminative, as long as sufficient excitation conditions are met. Clearly, if a sequence of identical patches is given (video with no motion), the descriptor will fail to capture the representative variability of images generated by the underlying scene. In this case, MV-HoG reduces to DSP-SIFT [12], which differs from SV-SIFT because of domain-size aggregation (averaging over σ). In Fig. 4(a) we explore the relation between performance gain and excitation level of the training sequence. As a proxy of the latter, we measure the variance of the intensity relative to the mean using the ℓ_2 distance. The right plot shows that the variance reaches the maximum when most frames are seen. We normalize the variance so that 1 means maximum excitation. The left plot shows accuracy increases with excitation. The fact that accuracy does not saturate is due to the fact that the sufficient excitation is only reachable asymptotically. At test time, all descriptors of n features have the same storage complexity $O(n)$ except that Orb-SIFT stores every instance ($O(kn)$). The search can be done in approximate form using *approximate nearest neighbors* [8]. Fig. 4(c) shows the training time using the *fast library for approximate nearest neighbors* (FLANN) vs MV-HoG on a commodity PC with 8GB memory and Xeon E3-1200 processor. MV-HoG scales well and is more memory-efficient while Orb-SIFT requires more training time and occupies more than 60% of the available memory. Another advantage of MV-HoG is that the descriptor can be updated incrementally, and does not require storing processed samples.

6. Discussion

By interpreting the SIFT/HOG family as the probability density of sample images conditioned on the underlying scene, with nuisances marginalized, and observing that a single image does not afford proper marginalization, we have been able to extend it using nuisance distributions learned from multiple training samples of the same underlying scene. The result is a multi-view extension of HoG that has the same memory and run-time complexity as its single-view counterpart, but better trades off sensitivity with discriminative power, as shown empirically, even with the classifier trivialized.

Our method has several limitations: It is restricted to static (or slowly-deforming) objects; it requires correspondence in multiple views to be assembled (although it reduces to DSP-SIFT if only one image is available), and is therefore sensitive to the performance of the tracking (MV-HoG) or reconstruction (R-HoG) algorithm. The former also requires sufficient excitation conditions to be satisfied, and the latter requires sufficiently informative data for multi-view stereo to operate, although if this is not the case (for instance in textureless scenes), then by definition the resulting descriptor is insensitive to nuisance factors; it is also, of course, uninformative, as it describes a constant image, and therefore this case is of no interest. It also requires the camera to be calibrated, but for the same reason, this is irrelevant as what matters is not that the reconstruction be correct in the Euclidean sense, but that it yields consistent reprojections.

Our empirical evaluation of R-HoG yields a performance upper bound, as we use a better approximation of the reconstruction (from a structured light sensor or ground truth) rather than multi-view stereo that, while possible, yielded inconsistent results across different samples. As the quality (and speed) of the latter improve, the difference between the two will shrink. We have also neglected the effects of sampling artifacts in the approximation of the ideal descriptor. However, in practice we have found them to be of second-order, compared to the approximation implicit in the spatial independence of the locally-aggregated histograms. Also, we wish to point out that ideal representations, in the sense of sufficient statistics that are (maximally) invariant, are not unique. However, they are equivalent from the informational standpoint [30]. Analytical evaluation of our approach is forthcoming [31].

Acknowledgments

Research supported by ONR N000141110863, NSF RI-1422669, NGA HM02101310004, ARO W911NF-11-1-0391.

References

[1] L. Alvarez, F. Guichard, P. L. Lions, and J. M. Morel. Axioms and fundamental equations of image processing. *Arch. Rational Mechanics*, 123, 1993. 2

[2] J. Balzer, M. Peters, and S. Soatto. Volumetric reconstruction applied to perceptual studies of size and weight. In *IEEE Winter Conference on Applications of Computer Vision*, 2014. 6

[3] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: speeded up robust features. In *Proceedings of European Conference on Computer Vision*, pages 404–417. Springer, 2006. 1, 6

[4] J. V. Bouvrie, L. Rosasco, and T. Poggio. On invariance in hierarchical models. In *Advances in Neural Information Processing Systems*, pages 162–170, 2009. 1

[5] J. Bruna and S. Mallat. Classification with scattering operators. In *Proceedings of IEEE Conference on Computer Vision & Pattern Recognition*, 2011. 1

[6] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In *British Machine Vision Conference*, 2011. 1

[7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proceedings of IEEE Conference on Computer Vision & Pattern Recognition*, 2005. 1, 2, 4

[8] D. Davis, J. Balzer, and S. Soatto. Asymmetric sparse kernel approximations for nearest neighbor search. In *Proceedings of IEEE Conference on Computer Vision & Pattern Recognition*, June 1, 2013. 8

[9] E. Delponte, N. Noceti, F. Odone, and A. Verri. The importance of continuous views for real-time 3d object recognition. In *ICCV Workshop on 3D Representation for Recognition*, 2007. 1, 6

[10] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei. Imagenet: a large-scale hierarchical image database. In *Proceedings of IEEE Conference on Computer Vision & Pattern Recognition*, pages 248–255, 2009. 1

[11] J. Dong, N. Karianakis, D. Davis, J. Hernandez, J. Balzer, and S. Soatto. Multi-view feature engineering and learning. *ArXiv preprint: 1311.6048*, 2013. 2, 6, 7

[12] J. Dong and S. Soatto. Domain-size pooling in local descriptors: DSP-SIFT. In *Proc. IEEE Conf. on Comp. Vision and Pattern Recogn.*, 2015; also *ArXiv preprint: 1412.8556*, 2014. 1, 4, 5, 8

[13] O. D. Faugeras and R. Keriven. Variational principles, surface evolution pdes, level set methods and the stereo problem. *INRIA TR*, 3021:1–37, 1996. 4

[14] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *Proceedings of IEEE Conference on Computer Vision & Pattern Recognition*, pages 1–8, 2008. 1

[15] G. Graber, J. Balzer, S. Soatto, and T. Pock. Efficient minimal surface regularization of perspective depth maps in variational stereo. In *Proc. IEEE Conf. on Comp. Vision and Pattern Recogn.*, 2015. 4

[16] M. Grabner and H. Bischof. Object recognition based on local feature trajectories. *I cognitive vision works*, 2, 2005. 1, 6

[17] E. Jones and S. Soatto. Visual-inertial navigation, localization and mapping: a scalable real-time large-scale approach. *International Journal of Robotics Research*, Apr. 2011. 3

[18] Y. LeCun, F. J. Huang, and L. Bottou. Learning methods for generic object recognition with invariance to pose and lighting. In *Proceedings of IEEE Conference on Computer Vision & Pattern Recognition*, volume 2, pages II–97, 2004. 1

[19] V. Lepetit, P. Lagger, and P. Fua. Randomized trees for real-time keypoint recognition. In *Proceedings of IEEE Conference on Computer Vision & Pattern Recognition*, volume 2, pages 775–781, 2005. 1, 6

[20] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. Journal of Computer Vision*, 60(2):91–110, 2004. 1, 3

[21] B. D. Lucas, T. Kanade, et al. An iterative image registration technique with an application to stereo vision. In *International Joint Conferences on Artificial Intelligence*, volume 81, pages 674–679, 1981. 6

[22] Y. Ma, S. Soatto, J. Kosecka, and S. Sastry. *An invitation to 3D vision, from images to geometric models*. Springer Verlag, 2003. 2, 3, 5

[23] R. Memisevic. Learning to relate images. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2013. 6

[24] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 27(10):1615–1630, 2005. 1

[25] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *Int. Journal of Computer Vision*, 1(60):63–86, 2004. 5

[26] P. Moreels and P. Perona. Evaluation of features detectors and descriptors based on 3d objects. *Int. Journal of Computer Vision*, 73(3):263–284, 2007. 5

[27] M. Ranzato, F. J. Huang, Y. L. Boureau, and Y. LeCun. Unsupervised learning of invariant feature hierarchies with applications to object recognition. In *Proceedings of IEEE Conference on Computer Vision & Pattern Recognition*, pages 1–8, 2007. 1

[28] E. Rosten and T. Drummond. Machine learning for high-speed corner detection. In *Proceedings of European Conference on Computer Vision*, volume 1, pages 430–443, May 2006. 6

[29] K. Simonyan, A. Vedaldi, and A. Zisserman. Learning local feature descriptors using convex optimisation. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2014. 1

[30] S. Soatto and A. Chiuso. Visual scene representations: sufficiency, minimality, invariance and deep approximation. *ArXiv: 1411.7676*, 2014. 1, 2, 3, 8

[31] S. Soatto, J. Dong, and N. Karianakis. Visual scene representation: scaling and occlusion in convolutional architectures. *ArXiv preprint: ArXiv:1412.6607*, 2014. 2, 8

[32] G. Sundaramoorthi, P. Petersen, V. S. Varadarajan, and S. Soatto. On the set of images modulo viewpoint and contrast changes. In *Proceedings of IEEE Conference on Computer Vision & Pattern Recognition*, June 2009. 3

[33] J. Susskind, R. Memisevic, G. E. Hinton, and M. Pollefeys. Modeling the joint density of two images under a variety of transformations. In *Proceedings of IEEE Conference on Computer Vision & Pattern Recognition*, pages 2793–2800, 2011. 6

[34] G. W. Taylor, R. Fergus, Y. LeCun, and C. Bregler. Convolutional learning of spatio-temporal features. In *Proceedings of European Conference on Computer Vision*, pages 140–153. Springer, 2010. 6

[35] E. Tola, V. Lepetit, and P. Fua. Daisy: an efficient dense descriptor applied to wide-baseline stereo. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 32(5):815–830, 2010. 1, 6

[36] A. Vedaldi and B. Fulkerson. Vlfeat: An open and portable library of computer vision algorithms. In *Proceedings of the international conference on Multimedia*, pages 1469–1472. ACM, 2010. 2, 6

[37] C. Vondrick, A. Khosla, T. Malisiewicz, and A. Torralba. Hoggles: visualizing object detection features. In *Proceedings of IEEE International Conference on Computer Vision*, 2013. 1

[38] G. S. Watson. *Statistics on spheres*. Wiley, 1983. 3

[39] S. A. Winder and M. Brown. Learning local image descriptors. In *Proceedings of IEEE Conference on Computer Vision & Pattern Recognition*, pages 1–8, 2007. 1, 5, 7