VISUAL SCENE REPRESENTATIONS:
SCALING AND OCCLUSION IN CONVOLUTIONAL ARCHITECTURES

Stefano Soatto  Jingming Dong  Nikolaos Karianakis

ABSTRACT

We study the structure of representations, defined as approximations of minimal sufficient statistics that are maximal invariants to nuisance factors, for visual data subject to scaling and occlusion of line-of-sight. We derive analytical expressions for such representations and show that, under certain restrictive assumptions, they are related to features commonly in use in the computer vision community. This link highlights the condition tacitly assumed by these descriptors, and also suggests ways to improve and generalize them. This new interpretation draws connections to the classical theories of sampling, hypothesis testing and group invariance.

1 INTRODUCTION

A visual representation is a function of past images that is useful to answer questions about the scene given future images from it, regardless of nuisance variability that will affect them. Soatto & Chiuso (2014) define an optimal representation as a minimal sufficient statistic (of past data for the scene) and a maximal invariant (of future data to nuisance factors), and propose a measure of how “useful” (informative) a representation is, via the uncertainty of the prediction density. What is a nuisance depends on the task, that includes decision and control actions about the surrounding environment, or scene and its geometry (shape, pose), photometry (reflectance), dynamics (motion) and semantics (identities, relations of “objects” within). Depending on the task, nuisance variables may include viewpoint, illumination, sensor calibration, and occlusion of line of sight. In this paper we focus on the latter and its impact in the design and learning of representations.

1.1 RELATED WORK AND CONTRIBUTIONS

This paper builds on Soatto & Chiuso (2014) by focusing on occlusion and scaling phenomena. There, a representation is seen as an approximation of the likelihood function, with nuisance factors either marginalized or profiled. Most work in low-level vision handles occlusions by restricting the attention to local regions of the image, resulting in representations known as local descriptors – too many to review here, with SIFT a prototypical representative Lowe (2004). Scale changes are handled by performing computation in scale-space Lindeberg (1998). Empirical comparisons abound (e.g., Mikolajczyk et al. (2004)) and recently expanded to include convolutional networks Fischer et al. (2014). What is missing is a framework to relate various descriptors to each other, so the assumptions on which they rely become patent, and to an “ideal” representation, so one can see how to improve them, not just compare them on any given dataset.

We show how removing nuisance variability due to occlusions is generally intractable, but can be approximated leading to a composite (correspondence) hypothesis test, which provides grounding for the use of “patches” or “receptive fields,” ubiquitous in practice (Sect. 3.2.1). The analysis reveals that the size of the domain of the filters should be decoupled from spectral characteristics of the image, unlike traditionally taught in scale-space theory, an unintuitive consequence of the analysis (Sect. 3.2.2). In Sect. 3.2.3, as a way of example, we show how the theory can be exploited to approximate the optimal descriptor of a single image, under an explicit model of image formation (the Lambert-Ambient, or LA, model), and nuisance variability. This turns out to be related to SIFT Lowe (2004) and the Scattering Transform Bruna & Mallat (2011), except for an important modification and known as domain-scale pooling (DSP). This suggests a way to improve both: DSP-SIFT was introduced in Dong & Soatto (2014). An extension of domain-size pooling to the scattering transform (DSP-SC) and to a convolutional neural network (DSP-CNN) are described in Sect. 3.5 and 3.6 respectively.
Of course, the restriction to a single training image is rather limiting, so in Sect. 3.4 we show how to exploit multiple images to construct local representations, either by sampling (an approach known as multi-view (MV)-HOG) or by explicitly reconstructing a point-estimate of the underlying geometry (reconstructive (R)-HOG), both introduced in Dong et al. (2013).

2 BACKGROUND

We treat images as random vectors $x, y$ and the scene $\theta$ as an (infinite-dimensional) parameter. A representation is a function $\phi$ of past images $x' = \{x_1, \ldots, x_t\}$ that maximally reduces uncertainty on the scene, given images from it and regardless of nuisances $g \in G$. Unfortunately, even defining the uncertainty on the scene is challenging, let alone computing it and optimizing it with respect to all possible measurable functions to find optimal representations. Mutual information is defined for random variables that can take a countable number of values, and can be extended to the continuum, but there are technical problems with the extension to infinite-dimensional objects such as $\theta$. Its entropy would be infinite no matter how large the sample on which we are conditioning Tishby et al. (2000). Soatto & Chiuso (2014) propose a characterization in terms of the likelihood function, which summarized in Sect. 2.4 after some nomenclature.

2.1 Nomenclature

A statistic $T$ is a function of the sample; it is sufficient (of $x$ for $\theta$) if $X \mid T = t$ does not depend on $\theta$ (Def. 3.1 of Pawitan (2001)); it is minimal if it is a function of all other sufficient statistics (DeGroot (1989), page 368). Thus all "information" about $\theta$ in the sample $x$ is contained in $T$. If $T$ is minimal, any smaller $U$ (in the sense of inclusion of sigma algebras) entails "information loss."

A nuisance $g \in G$ is an unknown parameter that is not of interest and yet appears in the likelihood. A function $\phi_g(x)$ is $G$-invariant if $\phi_g(gx) = \phi_g(x)$ for all $g \in G$. It is a maximal invariant if $\phi_g(\tilde{x}) = \phi_g(x) \Rightarrow \tilde{x} = gx$ for some $g \in G$ (Def. 2.4 of Eaton (1989)). A statistic $T$ is complete if for any function $g$ we have $E_\theta(g(T)) = 0 \Rightarrow g(t) = 0$. The likelihood approach to eliminating nuisance parameters is to maximize, or max-out: If $L(\theta, g; x) = p_{\theta,g}(x)$ is the joint likelihood, omitting reference to the data $x$,

$$L(\theta) = \max_{g \in G} L(\theta, g) = \max_{g \in G} p_{\theta,g}(x)$$

is the profile likelihood. If $g$ is treated as a random variable with prior $dP(g)$, it can be eliminated by marginalization from the conditional likelihood $L(\theta, g; x) = p_{\theta}(x|g)$:

$$L_G(\theta) = \int_G p_{\theta}(x|g)dP(g)$$

(2)

is the marginal likelihood. The prior may be uninformative, $dP(g) = d\mu(g)$ the base (normalized Haar) measure on $G$. If $G$ is a group, the orbit traced for all possible values of $g \in G$ is the orbit likelihood:

$$L(\theta|G) = \{L(\theta, g)\}_{g \in G}.$$ 

(3)

2.2 Sampling, anti-aliasing and the SOA likelihood

Sampling refers to a restriction of a function $f(g), g \in G$ to a discrete subset $\{g_i\}_{i=1}^N$ generated by a (deterministic or stochastic) sampling mechanism $\psi$, possibly depending on the data. A co-variant detector is a functional $\psi$ (Morse in $g$, i.e., having isolated extrema $\{g_i(x)\}_{i=1}^N = \{g \mid \psi(x, g) = 0\}$ that equi-vary, $\psi(\tilde{g}x, g) = 0 \Rightarrow \psi(x, \tilde{g}g) = 0$. Anti-aliasing refers to the weighted average of the function $f(\cdot)$ in a (possibly unbounded) neighborhood of a sample $g_i$ with weight function $w(\cdot)$:

$$\hat{f}_i = \int_G f(g, \tilde{g})w(g^{-1})d\mu(g),$$

normalized with $\int w(g)d\mu(g) = 1$. When the weights are positive we indicate $w(g^{-1})d\mu(g) = dP(g)$. In this case, anti-aliasing is equivalent to local marginalization around the sample $g_i$. Although classical sampling theory only considers the translation group of the real line, the definitions apply to other groups. The sampled-orbit anti-aliased (SOA) likelihood is defined in Soatto & Chiuso (2014) as

$$\hat{L}_{G,\epsilon(N)}(\theta; x) = \{\hat{L}(\theta, g_i; x)\}_{i=1}^N,$$

where $\hat{L}(\theta, g_i; x) = \int_G L(\theta, g_i g; x)dP(g)$

(4)
where the integral with respect to $dP(g) = w(g^{-1})d\mu(g)$ is anti-aliasing. The point maximum of the SOA is the profile SOA likelihood:

$$L_{G,\epsilon}(\theta; x) = \max_i \hat{L}(\theta, g_i; x), \ i = 1, \ldots, N(\epsilon)$$  \hspace{1cm} (5)

Under suitable regularity conditions on the map $g \mapsto L(\theta, g; x)$, when samples are generated by a co-variant detector, the SOA likelihoods arbitrarily approximate a maximal $G$-invariant, even when marginalization is not relative to the base measure (and therefore marginalization does not yield invariance) Soatto & Chiuso (2014).

2.3 The Lambert-Ambient (LA) Model

The Lambert-Ambient (LA) model Dong & Soatto (2014) is the simplest to capture the phenomenology of image formation including scaling, occlusion, and rudimentary illumination. For us, what matters of the LA model are three facts: First, the scene separates $x^t \perp y \mid \theta$, meaning that $p_\theta(x^t, y) = p_\theta(x^t)p_\theta(y)$. Second, conditioning on viewpoint factorizes the likelihood: If $g \in G = SE(3)$ is the position and orientation of the camera in the reference frame of the scene $\theta$ and the image $y$ is made of pixels $y_i$, then

$$p_\theta(y|g) = \prod_i p_\theta(y_i|g)$$  \hspace{1cm} (6)

Third, the action of restricted groups $G \subset SE(3)_i$, for instance planar translations, rotations, scalings, affine and projective transformations, contrast transformations, etc. is approximately equivariant, in the sense that for a sufficiently small domain,

$$p_\theta(g_1y|g_2) = p_\theta(y|g_1g_2)$$  \hspace{1cm} (7)

where the product $g_1g_2$ denotes group composition and the bar (omitted henceforth) denotes the embedding of the group action on the (2-D) plane into (3-D) Euclidean space. In Sect. 3.2 we will motivate these assumptions by restricting the representation to local spatial domains, and use it in Sect. 3.3 to achieve invariance to arbitrary vantage points.

2.4 Optimal Representation for the LA Model

Under the assumptions of the LA model, an optimal representation Soatto & Chiuso (2014)

$$\max_{g \in G} p_\theta(gy, x^t) = \phi_{\theta, G}(y)\phi_\theta(x^t)$$  \hspace{1cm} (8)

for nuisance variables $G$ including contrast (Sect. 3.1), viewpoint (Sect. 3.3), and occlusions (Sect. 3.2), where $\phi_{\theta, G}(y) = \max_{y \in G} \phi_\theta(gy)$, can be approximated from a training set $x^t \sim p_\theta(\cdot)$, via

$$\phi_\theta(\cdot) \doteq p_\theta(\cdot) \simeq \hat{p}_{X^t}(\cdot), \ X^t \sim p_\theta(\cdot)$$  \hspace{1cm} (empirical past likelihood)  \hspace{1cm} (9)

$$\phi_{\theta, G}(\cdot) \doteq p_{\theta, G}(\cdot) \simeq \hat{p}_{X^t, G}(\cdot)$$  \hspace{1cm} (marginal future likelihood)  \hspace{1cm} (10)

$$\phi_{\theta, G}(y) \doteq \phi_{\theta, G}(y)\phi_\theta(x^t) \simeq \hat{p}_{X^t, G}(y)\hat{p}_{X^t}(x^t) \propto \hat{p}_{X^t, G}(y)$$  \hspace{1cm} (learned representation)  \hspace{1cm} (11)

Alternatively, invariance to $G$ can be achieved by marginalization of the base measure, in which case

$$\phi_{\theta, G}(y) \doteq \int p_\theta(y|g)dP(g).$$  \hspace{1cm} (12)

Remark 1 (What you lose if you use lousy view(s) – Active Sensing). A representation, informative as it may be, can be no more informative than the data itself, uninformative as it may be. This is irrelevant in our context, for we are seeking statistics that are as informative as the (training) data (sufficient), however good or bad that is. For the representation to (asymptotically) approach the informative content of the scene, it is necessary to design the experiment $E$ so that the data collected $x^t$, with $t \to \infty$, yields statistics that are asymptotically complete Fedorov (1972). Such active learning or active sensing is beyond the scope of this paper.

When a single training datum is given, $x^t = x$, no intrinsic (intra-class) variability can be learned, and the variability in the data is ascribed to the nuisances. The representation for $t = 1$ thus reduces to

$$\phi_{X, G}(y) = p_G(y|x).$$  \hspace{1cm} (13)

We now illustrate how to approximate optimal representations explicitly.
3 Learning Visual Representations

In Soatto & Chiuso (2014), it is shown that the orbit likelihood of the LA model is maximally invariant and minimally sufficient. The advantage of retaining the orbit is that conditioning on the vantage points makes each pixel independent (6). This greatly simplifies the construction of the representation as the joint histogram is the product of one-dimensional marginals. The disadvantage is that, to evaluate a test datum $y$ we have to solve an optimization problem (1), a search that cannot be performed until the test datum $y$ becomes available. On the other hand, evaluation of the marginal likelihood (2) is straightforward and does not entail any search. However, groups acting on the domain of the data introduce spatial dependencies, and therefore the marginal likelihood is difficult to construct. Furthermore, although the marginal likelihood is $G$-invariant, it is not maximal. The SOA likelihood (Sect. 2.2) trades off the two approaches (marginalization and max-out), within the framework of sampling theory. Thus, visual representations can be learned or designed to compute the SOA likelihood with respect to nuisances that include illumination, viewpoint (with the associated scale changes), and partial occlusions.

3.1 Contrast invariance

Contrast is a monotonic continuous transformation of the (range space of the) data, and it is well-known that the curvature of the level sets is a maximal invariant Alvarez et al. (1993). Since the gradient orientation is everywhere orthogonal to the level sets, it is also a maximal contrast invariant. The following expression for the invariant is obtained via marginalization of the norm of the gradient orientation of $\nabla y$.

\[ \phi_x(\alpha) = \prod_i N_{\mathbb{S}^1}(\alpha_i - \angle \nabla x_i; \epsilon_\alpha)\| \nabla x_i \| \]

In the rest of the paper, we use the symbol $\alpha$ to denote the orientation of the image gradient relative to one of the coordinate axes, and omit the subscript $G$ when referring to contrast (since the use of the argument $\alpha$ makes it unambiguous). The width of the kernel $\epsilon_\alpha$ is a design (regularization) parameter.

Remark 2 (No invariance for $x$), Note that (15) is invariant to contrast transformations of $y$, but not of $x$. For a single training image, the latter can be handled by normalization as we will see next. For multiple images, the factor $G$ can in principle be different for each training image.

Remark 3 (Bayesian invariant). In the proof of Theorem 1, the gradient magnitude is marginalized with respect to the base measure. With a different prior, for instance arising from bounds on the gradient or from statistics of natural images, marginalization yields a factor other than $\| \nabla x \|$. Clamping, described next, can be understood as a particular choice of prior for marginalization of the gradient magnitude.

Invariance to contrast transformations in the (single) training image can be performed by normalizing the likelihood, which in turn can be done in a number of ways. If contrast transformations are globally affine, then the joint likelihood can be normalized by simply dividing by the integral over $\alpha$, which is the $\ell^1$ norm of the histogram across the entire image/patch.

\[ \frac{\phi_x(\alpha)}{\| \phi_x(\alpha) \|_{\ell^1}} = \frac{p(\alpha|x)\| \nabla x \|}{\int p(\alpha|x)d\alpha\| \nabla x \|} = p(\alpha|x) \]

that should be used instead of the customary $\ell^2$ Lowe (2004). If the contrast transformation is non-linear, it cannot be eliminated by global normalization.
Remark 4 (Clamping). When the joint distribution is approximated by the product of marginals, as in Lowe (2004), joint normalization is still favored in practice as it introduces some correlations among marginal histograms Dalal & Triggs (2005). However, cells with large gradients tend to dominate the histogram, pushing all other peaks lower. Alternatively, one could independently normalize each of the histograms, \( \phi_{x_i}(\alpha) \) and then concatenate them. But this has the opposite effect: Cells with faint peaks, once re-normalized, are given undue importance and relative intensity difference between different cells are discarded. A common trick consisting of joint normalization (so faint cells do not prevail) followed by “clamping” (saturation of the maximum to a fraction of the value of the highest peak, so large gradients do not dominate), and then re-normalization, seems to achieve a tradeoff between the two Lowe (2004). This process can also be understood as a way of marginalizing \( \rho \), with respect to a different measure \( dP(\rho) \), as described in Rem. 3 while assuming that, within each region, contrast transformations are affine.

Once invariance to contrast transformations is achieved, which can be done on a single image \( x \), we are left with nuisances \( G \) that include general viewpoint changes, including the occlusions they induce. This can be handled by computing the SOA likelihood with respect to the product \( G \) of \( SE(3) \) (the group of general rigid motions, Sect. 3.3) and the scale semi-group, which we discuss next, from a training sample \( x^n \), leading to

\[
\hat{L}(\theta, g_i; x^i) = \left\{ \int_G \phi_{x^i}(\alpha | g_i \circ g) dP(g) \right\}^N_{i=1}
\]  

(17)

In the next section we show how to handle occlusions, and in the following one general viewpoint changes.

3.2 Occlusions

We do not know ahead of time what portion of an object or scene, seen in training images, will be visible in a test image. Occlusion, or visibility, is arguably the single most critical aspect of visual representations. It enforces locality, as dealing with occlusion nuisances entails searching through, or marginalizing, all possible (multiply-connected) subsets of the test image. This power set is clearly intractable even for very small images.

3.2.1 Bypassing Shape and Justifying “patches” or “Receptive Fields”

We illustrate a principle to bypass combinatorial explosion for a single training image, absent all other nuisances. A training \( x \) and a test image \( y \) correspond (hypothesis \( H_0 \)) if there exist subsets of \( x, \Omega_x \), and of \( y, \Omega_y \), such that the restrictions come from the same scene, \( i.e. \) in this setting they differ by a white (zero-mean, uninformative) residual.\(^1\) Under this simplistic model, the subsets \( \Omega_x = \Omega_y = \Omega \) are the same, and \( y = x + n \) where \( n \) is either a white (spatially i.i.d) zero-mean process with a small covariance, \( n_{ij} \sim N(0, \epsilon^2) \) in the corresponding region, or something else, for instance uniform with a mean in the order of magnitude of the intensity range, assumed normalized to one, \( n_{ij} \sim U \). Hypothesis \( H_1 \) is that there exists no such region, and \( n \sim U \) on the entire domain. Since we do not know the region \( \Omega \), this is a composite hypothesis testing problem, where the likelihood ratio is given by

\[
\frac{p(y|x, H_0)}{p(y|x, H_1)} = \frac{\max_{\Omega} \left( p(y_{ix} | x_{ix}, H_0) \right) \cdot p(y_{ic} | x_{ic}, H_0)}{\max_{\Omega} \left( p(y_{ic} | x_{ic}, H_1) \right) \cdot p(y_{ic} | x_{ic}, H_1)} = \frac{\max_{\Omega} N(y_{ix} - x_{ix}; \epsilon^2)}{\max_{\Omega} U(y_{ic} - x_{ic})} \quad (18)
\]

Missed detections (treating a co-visible pixel as occluded) and false alarms (treating an occluded pixel as visible) have different costs: Omitting a co-visible pixel from \( \Omega \) decreases the likelihood by a factor corresponding to multiplication by a Gaussian for samples drawn from the same distribution; vice-versa, including a pixel from \( \Omega^c \) (false alarm) decreases the log-likelihood by a factor equal to multiplying by a Gaussian evaluated at points drawn from another distribution, such as uniform. So, testing for correspondence on subsets of the co-visible regions, assuming the region is sufficiently large, reduces the power, but not the validity, of the test. This observation can be used to fix the

\(^1\)Of course, absent all other nuisances, all pixels are independent so corresponding regions can be determined by “background subtraction” techniques. This of course does not work in the presence of other nuisances, so the example serves just to illustrate the principle.
3.2.2 INVARIANCE TO OCCLUSION

Much of the literature on local descriptors further simplifies correspondence by selecting scales 
Lowe (2004), using the appearance of the scene (really, an image x) to determine the size of the 
region \( \Omega \) where the descriptor is computed. This is motivated by Lindeberg (1998), where the tying of appearance (e.g., spatial frequencies) and size is known as the “uncertainty principle.” But while the tie makes sense when the goal is to compress the image, it does not for correspondence, as the size of a region that will be visible in a test image has nothing to do with the appearance of the scene within. Therefore, the size of the domain where the descriptor is computed must be untied from the appearance of the scene within, and instead treated as an independent nuisance and included among those to be managed by max-out or marginalization. Surprisingly, no existing descriptor or convolutional architecture did so, until recently.

3.2.3 LOCAL DESCRIPTORS REVISITED

Co-variant scale selection can be thought as a way to sample scale, as opposed to selecting corresponding scales between training and test sets. But to approximate the SOA likelihood, co-variant sampling must be coupled with anti-aliasing, which corresponds to domain-size pooling (DS pooling), a concept introduced in Dong & Soatto (2014) and illustrated below for the case of a descriptor built on a single image. The difference between scale-space and “size-space” is illustrated in Fig. 1. If a translation-scale co-variant detector generates a (sufficient) number of locations \((u_i, v_i)\) and scales \(\hat{\sigma}\), then the SOA likelihood with respect to the translation-scale group of the plane is given by

\[
\phi_{x_i}(\alpha) = \int \kappa_\epsilon(\alpha - \nabla I(x_j)) \kappa_\sigma(i - j) ||\nabla x_j|| \mu(j) dP(\sigma)
\]

where the location prior is captured by an isotropic kernel \(\kappa\) with dispersion parameter \(\sigma\), and the scale prior is captured by a measure \(dP(\sigma)\) (for instance an exponential unilateral). If orientation is also sampled in a co-variant manner, or canonized with gravity as described in Ex. 1, then orientation anti-aliasing is already implicit in the histogram binning, represented by a kernel with dispersion parameter \(\epsilon\). The above is precisely one cell of DSP-SIFT as introduced in Dong & Soatto (2014). We therefore have the following:

\footnote{Alternatively, the sampling can be framed as a sequential hypothesis test for joint matching and domain size estimation, as in region-growing or quickest setpoint change detection.}
Corollary 1 (DSP-SIFT). The DSP-SIFT descriptor Dong & Soatto (2014) \((19)\) approximates the ideal representation \((13), (17)\) for \(G\) the group of planar similarities and local contrast transformations, when the scene is a single training image, and the test image is restricted to an unknown subset of its domain.

Since the standard SIFT/HOG and its variants do not anti-alias domain size, invariance to partial occlusion (and therefore scale) is lost.

Corollary 2 (Assumptions implicit in SIFT/HOG). The SIFT descriptor Lowe (2004) and its variants are a special case of DSP-SIFT \((19)\) with \(dP(\sigma) = \delta(\sigma - \bar{\sigma})\), and therefore approximate the ideal representation \((13)\) only at a fixed scale \(\bar{\sigma}\) for a scene that is flat, fronto-parallel, undergoing purely translational motion along the image plane.

SIFT as designed violates the sampling principles described here, as sampling occurs with respect to the full similarity group (positions, scales and rotations are selected using a co-variant detector), but anti-aliasing is only performed in position (spatial pooling) and orientation (histogram smoothing), \emph{not in scale}, which in SIFT is tied to domain size.

In both SIFT and DSP-SIFT, and most other local descriptors, samples are aggregated independently at each neighborhood (although there may be overlap) and concatenated. In the interpretation of SIFT as a likelihood function, this corresponds to assuming that the joint likelihood factorizes as the product of the marginals, which in turn corresponds to assuming that the random variables \(x_i\) are (spatially) independent. Each local descriptor is then considered independently, as in a “bag-of-word” approach. Deep convolutional architectures relax this assumption, as shown in Soatto & Chiuso (2014). In Sect. 3.5 we show how a scattering network Bruna & Mallat (2011) fits in our framework.

Finally, a single image does not afford the ability to separate nuisance from intrinsic variability. This issue becomes patent when attempting to achieve invariance to general viewpoint changes.

3.3 General viewpoint changes

If a co-variant translation-scale \emph{and size} sampling/anti-aliasing mechanism is employed, then around each sample the only residual variability to viewpoint \(SE(3) = \mathbb{R}^3 \times SO(3)\) is reduced to \(SO(3)\). That can be further factored into a rotation of the image plane (“in-plane” rotation), and its complement (“out-of-plane” rotation). We next show how in-plane rotations can be eliminated, leaving only out-of-plane rotations.

Example 1 (Rotation Invariance). Canonization is particularly well suited to deal with planar rotation, since the statistics of natural images ensure that with high probability orientation-co-variant detectors have few isolated extrema. An example is the local maximum of the norm of the gradient \(\hat{\alpha}(x)\).\(^4\) Invariance to \(G = SO(2)\) can be achieved by retaining the samples

\[
p_\theta(\alpha|G) = \{ p_\theta(\alpha|\hat{\alpha}_i) \}_{i=1}^L
\]

Rotation anti-aliasing is performed by regularizing the orientation histogram. Note that, as it was for contrast, planar rotations can affect both the training \(x\) and the test image \(y\). In some cases, a consistent reference (canonical element) is available for both when scenes or objects are geo-referenced: The projection of the gravity vector onto the image plane. In this case, \(L = 1\), and \(\hat{\alpha}\) is the angle of the projection of gravity onto the image plane (well defined unless they are orthogonal):

\[
p_\theta(\alpha|G) = p_\theta(\alpha|\hat{\alpha}).
\]

In reality, rotation canonization should contend with spatial quantization, neglected here since rotation errors are absorbed by the binning of gradient orientation \(\epsilon_\alpha\).

\(^3\)In reality, translation in space is not equivalent to translation and scaling of the image plane, for the former induces deformations of the image domain due to parallax effects and occlusions, which are absent in the latter. However, locally and away from occlusions, one is a first-order approximation of the other, so the derivation is valid for each local region that does not straddle an occluding boundary, justified by our handling of occlusions via the restriction to receptive fields in Sect. 3.2.

\(^4\)Here \(g\) acts on \(x\) via \(gx(u, v) = x(u', v')\) where \(u'' = u \cos \alpha - v \sin \alpha\) and \(v'' = u \sin \alpha + v \cos \alpha\), and a canonical element \(\hat{g}_i(x) = \hat{\alpha}\) can be obtained as \(\hat{\alpha} = \arg \max_{\alpha} \| \nabla x(u_i', v_i') \| \). The corresponding rotation invariant \(\hat{g}^{-1}(x)x\) is \(\hat{\alpha}\). Where \(u' = u \cos \alpha + v \sin \alpha\) and \(v' = -u \sin \alpha + v \cos \alpha \approx \hat{\alpha}'\).
This leaves out-of-plane rotations to be profiled or marginalized. Unfortunately, the effects of such rotations on future images depend on the shape of the underlying scene, which is unknown, and that cannot be determined from a single image. Therefore, the only way in which true viewpoint changes can be factored out of the representation is if multiple training images of the same scene are available. In the next section we show how such multi-view representations can be constructed.

### 3.4 Extension to multiple views

Out-of-plane rotations induce a scene shape-dependent deformation on the domain of the image that cannot be determined from a single image. Dong et al. (2013) have proposed extensions based on a sampling approximation of the likelihood function, $p_{\hat{\theta}}$, or on a point estimate of the scene $p_{\hat{\theta}}$, multi-view HOG and reconstructive HOG respectively. The estimated scene has a geometric component (shape) $\hat{S}$ and a photometric component (radiance) $\hat{\rho}$, inferred from the LA model as described in Dong & Soatto (2014). These in turn enable the approximation of the predictive likelihood $p_{\hat{\theta},G}$, and hence the representation:

$$\phi_{\hat{\theta},G}(\alpha_i) = \int_{SO(3)} N_{\hat{\theta}}(\alpha_i - \hat{\rho} \circ g \circ \pi^{-1}_S(u_j, v_j); \epsilon_\alpha) \|\nabla \hat{p}\| N_{\sigma}(i - j) d\mu(j) dP_{SO(3)}(g) \quad (22)$$

where $\hat{\theta} = (\hat{S}, \hat{\rho})$, $\nabla \hat{y} = \alpha$ and $\pi^{-1}$ is the pre-image of a perspective projection (the point of first intersection of the ray through the pixel $(u_j, v_j)$ with the surface $\hat{S}$). Alternatively, a sampling approximation of the likelihood function $\hat{p}_{\theta}(a^t)$ yields “multi-view HOG”

$$\phi_G(\alpha_i | a^t) \doteq 1 \sum_l \int_{\mathbb{R}^2} N_{\hat{\theta}}(\alpha_i - \angle \nabla x_{\tau_j}; \epsilon_\alpha) N_{\sigma}(i - j) d\mu(j) dP(\sigma) \quad (23)$$

Note that the gradient weight $\|\nabla x_{\tau}\|$ is absent, since individual samples of past data do not enable separating nuisance from intrinsic variability, and each sample image $x_{\tau}$ has different contrast, so the factor cannot be simply eliminated by normalization as done in Rem. 3 for a single image. Therefore, in MV-HOG it is necessary to assume that training images are captured under the same illumination conditions. In MV-HOG, regularization is implicit in the kernel, and the predictive likelihood is based on simple planar transformations. In R-HOG, the estimated scene (which requires regularization to be inferred) acts as the regularizer Dong et al. (2013). Once the effects of occlusions are considered (which force the representation to be local), and the effects of general viewpoint changes are accounted for (which creates the necessity for multiple training images of the same scene), a maximal contrast/viewpoint/occlusion invariant can be approximated via the SOA likelihood. Using (19) and (22), the SOA likelihood (17) becomes:

$$\hat{L}_{SE(3),\epsilon(N)}(\alpha_i) = \max_k \left\{ \int_{SO(3)} N_{\hat{\theta}}(\alpha_i - \hat{\rho} \circ g_k \circ \pi^{-1}_S(x_j); \epsilon_\alpha) \kappa_\sigma(i - j) d\mu(j) dP(\sigma) dP_{SO(3)}(g) \right\}^N_{k=1} \quad (24)$$

The assumption that all existing multiple-view extensions of SIFT do not overcome is the conditional independence of the intensity of different pixels (6). This is discussed in Soatto & Chiuso (2014) for the case of convolutional deep architectures, and in the next section for Scattering Networks.

### 3.5 DSP-Scattering Networks

The scattering transform Bruna & Mallat (2011) convolves an image (or patch) with a Gabor filter bank at different rotations and dilations, takes the modulus of the responses, and applies an average operator to yield coefficients. This is repeated to produce coefficients at different layers in a scattering network. The first layer is equivalent to SIFT Bruna & Mallat (2011), in the sense that (14) can be implemented via convolution with a Gabor element with orientation $\alpha$ then taking the modulus of the response. This is the convolution-modulus step in the scattering transform. Then the local marginalization in (19), except for the integral with respect to $dP(\sigma)$, corresponds to a low-pass filter applied to the histogram, yielding a spatially-pooled (regularized) histogram of gradient orientations. This is (19) for a fixed $\sigma$. The same operator exists in the scattering network where the modulus of the filter response goes through a low-pass filter to generate the final coefficients.

Of course there could be implementational difference depending on the choices of kernels, their parameters, and number of samples or filters in the bank. Nevertheless, one could conjecture that
domain-size pooling (DSP) applied to a scattering network would improve performance in tasks that involve changes of scale and visibility. We call the resulting method DSP Scattering Transform (DSP-SC). Indeed, this is the case, as we show in the Supplementary Material, where we compare DSP-SC to the single-scale scattering transform (SC) on both the Oxford Mikolajczyk & Schmid (2003) and Fischer’s Fischer et al. (2014) datasets.

Stacked architectures, such as the SC, offer the promise to lift the strong spatial independence assumption implicit in SIFT and its variant, as well as DSP-SIFT Dong & Soatto (2014). However, the empirical results in Bruna & Mallat (2011) suggest that this value is exhausted after $M = 2$ layers. In the next section we discuss an extension to deep architectures.

3.6 DSP-CNN

Deep convolutional architectures can be understood as implementing successive approximations of the optimal representation (8), by stacking layers of (conditionally) independent local representations of the form (24), which have been shown by Soatto & Chiuso (2014) to increasingly achieve invariance to large deformations, despite locally marginalizing only affine (or similarity) transformations. As Dong & Soatto (2014) did for SIFT, and as we did for the Scattering Transform above, we conjectured that pooling over domain size would improve the performance of a convolutional network. In the Supplementary material we report experiments to test the conjecture using a pre-trained and fine-tuned network on benchmark datasets. For computational reasons, we limited the domain-size pooling to 3 sizes (including the base size), and only to the first two layers. Still, the experiments show marginal improvement. We conjecture that more thorough incorporation of domain-size pooling, as described in the appendix, would further improve performance after proper training (rather than a pre-trained network) is employed.

4 CONCLUSIONS

We have derived an expression (24) for a minimal sufficient statistic of past data when the test image is restricted to a neighborhood of $y$ where $\alpha_i$ is computed, corresponding to sampled locations around $(u_k, v_k)$, with scales $\sigma$ pooled according to the prior $dP(\sigma)$ around the samples $\sigma_k$.

If a sufficiently exciting training set is available, spanning variability due to out-of-plane rotations, marginalization of $SO(3)$ can be replaced by temporal averaging of the training images (23). The joint distribution of local descriptors can be captured by a stacked architectures, as shown in Soatto & Chiuso (2014) and illustrated for Scattering Networks, Deformable Parts Models, and general deep convolutional architectures.

ACKNOWLEDGMENTS

We acknowledge discussions with Alessandro Chiuso, Jingming Dong, Joshua Hernandez, Arash Amini, Ying-Nian Wu, Taco Cohen, Virginia Estellers, Jonathan Balzer.

REFERENCES


Dong, J. and Soatto, S. Machine Learning for Computer Vision, chapter Visual Correspondence, the Lambert-Ambient Shape Space and the Systematic Design of Feature Descriptors. R. Cipolla, S. Battiato, G.-M. Farinella (Eds), Springer Verlag, 2014.
Under review as a conference paper at ICRL 2015


SUPPLEMENTARY MATERIAL

PROOF OF THEOREM 1

Note that in the claim we do not assume that the gradient norm is independent of its orientation, just that the error in the gradient norm is independent of the error in its orientation. We model the first as a Gaussian with \( \epsilon_p \), sufficiently small to ensure that the samples are positive with high probability, and the second as an angular Gaussian (Watson, 1983):

\[
\begin{align*}
p(\|\nabla y\| | \|\nabla x\|) &= \mathcal{N}(\|\nabla y\| - \|\nabla x\|; \epsilon_p) \\
p(\langle \nabla y, \nabla x \rangle) &= \mathcal{N}_S(\langle \nabla y - \nabla x; \epsilon_{\alpha}\rangle)
\end{align*}
\] (25)

**Proof.** Since the spatial gradient cannot be computed if only the scalar value \( y_i \) is given, we first map the data \( y \in \mathbb{R} \) to the spatial gradient \( \nabla y \in \mathbb{R}^2 \), then re-parametrize it in polar coordinates as \( (\|\nabla y\|, \|\nabla y\|) \in \mathbb{S}^2 \times \mathbb{R}^+ \), finally marginalizing the magnitude to arrive at a distribution on the circle: The first map is linear and has the effect of zeroing the mean and doubling the variance. Applying a change of coordinates \( Z \equiv (\|\nabla y\|, \|\nabla y\|) = \phi(Y) \), and conditioning on \( X = \nabla x \), we obtain

\[
p_{Y|X}(y|x) = p_{\phi^{-1}(Z)|X}(y|x) = p_{Z|X}(\phi(y)|x)|J_{\phi}(y)|^{-1}
\] (26)

Using Bayes’ rule, and transforming it to \( W = \phi(X) = (\nabla x, \|\nabla y\|) \) we obtain

\[
p_{Y|X}(y|x) = p_{Z|W}(\phi(y)|\phi(x))|J_{\phi}(y)|^{-1}
\] (27)

which is written explicitly as

\[
p_{\nabla y|\nabla x}(\alpha|\alpha') p_{|\nabla y| | |\nabla x|}(\rho|\rho') |\rho|
\] (28)

where \( \alpha \) is a realization of \( \nabla y \) and \( \rho \) is a realization of \( \|\nabla y\| \). Finally, by marginalizing \( \rho \) we have, neglecting the subscripts for simplicity, and leveraging on the fact that \( \rho \geq 0 \):

\[
\int p(\alpha|\alpha')p(\rho|\rho')|\rho|d\rho = p(\alpha|\alpha') \int p(\rho|\rho')|\rho|d\rho = p(\alpha|\alpha')|\rho'|
\] (29)

which is \( p_{\alpha}(y|x) = p(\nabla y|\nabla x, \|\nabla x\|) \). From which the statement (14) follows. \( \square \)

SCATTERING TRANSFORM AND DSP-SC

We test the conjecture that adding domain-scale pooling to the scattering transform, yielding DSP-SC, improves performance in tasks that entail scale and occlusion nuisances. We use the online toolbox ScatNet \(^1\) to compute scattering coefficients on the MSER-normalized \( 91 \times 91 \) image patch. Since the scattering transform is not contrast invariant, we normalize the raw patch to zero mean and unit variance before feeding it to the scattering network. We use the best parameters from Bruna & Mallat (2011), \( J = 3 \) for the number of scales in the filter bank and \( L = 8 \) number of orientations. We set the depth of the scattering network \( M = 2 \) so that the coefficients of the first three \(^2\) layers are used. The oversampling option is disabled. By concatenating all the coefficients output by the network, we get a scattering descriptor of dimension \( 217 \times 12 \times 12 = 31,248 \) for a \( 91 \times 91 \) image patch. DSP-Scattering is computed by sampling \( 3^2 \) scales around the scale detection \((2/3\sigma, 4/3\sigma)\), re-normalizing image patches to the same size \( 91 \times 91 \), computing scattering coefficients individually, and averaging the resulting outputs to generate the descriptor of the same size as that of the standard scattering descriptor. Fig. 2 shows the average precision as a function of the transformation magnitude. Fig. 3 shows the head-to-head comparison between these methods. Despite the fact that DSP-SC aggregates only 3 scales, it outperforms the single scale scattering transform by a sizable margin. The improvements are consistent across almost all the samples from the two evaluation datasets except for only 6 out of 400 pairs from Fischer dataset.

\(^1\)From http://www.di.ens.fr/data/software/scatnet/

\(^2\)The scattering network starts at layer 0.

\(^3\)Due to the computational inefficiency of scattering transform, we only select 3 scales to illustrate the concept of DS-pooling. In practice, one should use a much higher sampling rate and pool over a larger scale neighborhood. As a matter of comparison, DSP-SIFT uses 15 scales sampled from \((1/6\tilde{\sigma}, 4/3\tilde{\sigma})\) around the detection \( \tilde{\sigma} \).
Figure 2: Average Precision for different magnitude of transformations. The left 9 panels show (AP) for increasing magnitude of the 8 transformations in the Oxford dataset Mikolajczyk & Schmid (2003). The mean AP over all pairs with corresponding amount of transformation are shown in the middle of the third row. The right 6 panels show the same for Fischer’s dataset Fischer et al. (2014).

Figure 3: Head-to-head comparisons. Similarly to Fischer et al. (2014), each point represents one pair of images in the Oxford (left) and Fischer’s (right) datasets. The coordinates indicate average precision for each of the two methods under comparison. DSP-SC clearly outperforms SC in both datasets. The relative performance improvement of the winner is shown in the title of each panel. The mean AP for each method is shown in the axes.

DOMINO-SIZE POOLED CONVOLUTIONAL NETWORK

To the best of our knowledge, HMAX Serre et al. (2007) is the only network that pools across different scales, but yet no approach that we know of pools across different domain sizes. The representation is built in successive stages (Fig. 4), where at each stage \( n \) the representation \( \theta_n \) is represented by a distribution of images \( x \), from which one can sample. In the next subsection we illustrate the construction of one layer, shown visually in Fig. 4. In the following subsection we describe a reference implementation, together with its testing on PASCAL VOC 2007.
Figure 4: Stages of a layer (Level) of a deep convolutional architecture. Domain-size pooling is indicated by the scaled images in stages 3-4.

DERIVATION

**Stage 0: Embedding** $\theta_0 = x_0 : D \to [0, 255]; (u_i, v_i) \mapsto x_i$. Given only one image, the representation is the image:

$$\phi_0(y) = p(y|\theta_0) = \delta(y - \theta_0) = \prod_i p(y_i|x_i) = \prod_i \delta(y_i - \theta_0(u_i, v_i))$$

(30)

If $\theta_0$ can take, at each point $(u_i, v_i)$, a value $k \in [0, \ldots, K = 255]$, the above can be interpreted as a filter map, where each filter $F_k$, at location $(u_i, v_i)$ gives the likelihood of the scene taking value $k$ there. It is one if $\theta_0(u_i, v_i) = k$ and zero otherwise. In other words, $F_k(u_i, v_i) = \delta(k - \theta_0(u_i, v_i))$. Once a test image $y$ is given, $p(y_i|\theta_0) = p(y_i = k|\theta_0) = F_k(u_i, v_i) y_i$, and therefore $p(y_i|\theta_0) = F_k \cdot y$, the point-wise product. Considering all possible values of $k$, we obtain, stacking the filters $F = [F_0, \ldots, F_K]$,

$$\phi_0(y) = p(y|\theta_0) = [F_0, \ldots, F_{255}] \cdot y.$$  

(31)

Note that $\phi_0$ is just a different encoding of $\theta_0$: Instead of $\theta_0 : D \to [0, 255]$, we represent it as $\phi_0 : D \times [0, 255] \to \{0, 1\}; (u_i, v_i, k) \mapsto p_{\theta_0(u_i, v_i)}(k) = \phi_0(k)$. This representation is low-entropy, dense, i.i.d., unimodal, and exhibits no invariance properties. The only “scene” to have non-zero likelihood is one that is identical to the test image (Fig. 4).

**Stage 1: Smoothing (noise)** $\theta_0 = \theta_1 + n$ with $n \sim \mathcal{N}(0; \epsilon^2)$ i.i.d.. Assuming conditional independence across layers:

$$\phi_1(y) = p(y|\theta_1) = \int p(y|\theta_0)p(\theta_0|\theta_1)d\theta_0 = \int \delta(x-\theta_0)\mathcal{N}(\theta_0|\theta_1; \epsilon^2)d\theta_0 = \mathcal{N}(x-\theta_1; \epsilon^2).$$

(32)

Again, this can be interpreted as the response of a bank of filters, $G_k$, where now $G_k(u_i, v_i) = \mathcal{N}(k - \theta_0(u_i, v_i), \epsilon^2)$. We then have $\phi_1(y_i|\theta) = \mathcal{N}_k * \phi_0(y_i|\theta)$ point-wise, or

$$\phi_1(y|\theta) = p(y|\theta_1) = [G_1, \ldots, G_K] \cdot \phi_0(y|\theta)$$

(33)

This representation is still low-entropy, dense, i.i.d., unimodal, but it exhibits reduced sensitivity to additive perturbations (“noise”), so scenes that are “noisy versions” of the image can still have high likelihood.

**Stage 2: Invariant analysis (contrast)** $\theta_1 = \kappa(\theta_2)$, where $\kappa$ is a contrast transformation. Marginalizing it yields

$$\phi_2(y|\theta) = p(y|\theta_2) = \int p(y|\theta_1)p(\theta_1|\theta_2)d\theta_1 = \int p(y|\theta_2, \kappa)d\mu(\kappa) = \prod_i \mathcal{N}(\nabla y|\theta_2)\nabla \theta_2(u_i, v_i))$$

(34)

As usual, we can interpret this as the response of a filter bank, in this case a (Gabor) oriented set of filters on a spatial domain of size $\sigma_2$, corresponding of the domain where the regularized gradient operator is computed:

$$\phi_2(y|\theta) = p(\nabla y = \alpha_k|\theta_2) = [\ldots, \nabla \sigma_2 \mathcal{G}(\alpha_k), \ldots] \cdot \phi_1(y|\theta)$$

(35)

Note that both $F$ and $G$ can be incorporated into $G$, and therefore we can also write $\phi_2(y|\theta) = \nabla \sigma_2 \mathcal{G}(\alpha_k) \cdot \phi_0(y|\theta)$. This representation is still low-entropy, unimodal, but is now sparse (the gradient is almost zero at most locations) and no longer i.i.d. as the gradient operator captures local correlations. The representation is, by construction, invariant to contrast transformations of the test data. Scenes that are equivalent to the image up to contrast transformations and additive noise have high likelihood under this model.
Remark 4. locally should be normalized. This can be done point-wise, thus reducing discriminative power. Instead, (16). In one layer of a convolutional architecture, the scene is represented by a sample from it, invariant to contrast transformations of the sample the group \( g \) point) and \( k \) (domain size). This corresponds to the following model:

\[
\phi_3(y_i|\theta) = \frac{\phi_2(y_i|\theta)}{\int_{B_{d,2}} \phi_2(y(u, v)|\theta)dudv} \tag{36}
\]

This operation is written in the same fashion regardless of whether it is interpreted as a likelihood function or as a filter map. In the latter case, the denominator is the \( \ell^1 \) norm of the feature map in a local neighborhood. The resulting representation is, as in stage 2, low-entropy, sparse, unimodal, but contrast-invariant and no longer i.i.d because correlations within a spatial range \( \sigma_2 \) has been introduced by normalization.

Stage 4: Marginalization (mean pooling, anti-aliasing) A nuisance group \( g \), which can include translation, rotation, scale, and even domain size (a semi-group), can be anti-aliased by spatial pooling, relative to “weights” (prior) \( dP(g) \). This could be a concentrated density around the identity of the group \( c \in G \). Note that we already have two independent indices \( i \) (spatial location of the point) and \( k \) (index of the filter, or free argument of the likelihood function). Now we have to sample the group \( g_i \). Translations are already represented in the index \( i \), so sampling with respect to translations can be accomplished simply by sub-sampling the index \( i \). The free variable \( k \) remains, which can be considered a sampling necessary to manage the contrast transformation group. We now need to sample the rotation group (but this can be canonized if the scene is geo-referenced, using gravity, Example 1), and the scale and domain size variables. This introduced two additional variables \( s \) (scale) and \( d \) (domain size). This corresponds to the following model:

\[
\theta_4(u_i, v_i) = W(u_i, v_i; s_j, d)\theta_4 \tag{37}
\]

where \( W \) is a linear operator of the kind

\[
W(u_i, v_i; s, d)\theta = \int_{B_{d,2}^{(u_i, v_i)}} \delta\left(\frac{u_i - u}{s_u}, \frac{v_i - v}{s_v}\right) \theta(u, v)dP(u, v)dP(d)dP(s) \tag{38}
\]

so

\[
\theta_3(u_i, v_i) = \theta_4\left(\frac{u_i - u}{s_u}, \frac{v_i - v}{s_v}\right) \tag{39}
\]

and

\[
\phi_4(y_i|\theta) = p(y_i|\theta_4) = \int_G \phi_3(y|\theta, g)dP(g) = \mathcal{W}(G)\phi_3(y) \tag{40}
\]

where the linear operator \( \mathcal{W} \) can be interpreted as filtering with respect to a deformed dictionary (referred to as deformation hypercolumns in Sect. 6.5.1, page 73 of Soatto (2010)), or simply as pooling or averaging in a neighborhood of each location, orientation, scale and receptive field size. This representation is now not i.i.d., it is invariant to planar similarity (local affine transformations), contrast, and it is now multi-modal. Thus the representation is high entropy and not sparse.

Stage 5: Joint neighborhood description \( \theta_4 = \pi(\theta_5) \), where \( \pi \) is a projection that includes visibility, for instance the selection of one of the components of \( \theta_5 \). In particular, if

\[
\phi_5(y_i|\theta) = p(y_i|\theta_5) = [\phi(y_i|\theta_4), \ldots, \phi(y_n|\theta_4)], \{y_j\}_{j=1}^n \in B_{d,2}^{(u_i, v_i)} \tag{41}
\]

then \( \pi \) corresponds to the selection of subsets of the \( y_j \). Following Sect. 3.2, these subsets (“receptive fields”) can be constrained to a fixed shape, but of course can have variable size. The representation is now high entropy, non-i.i.d., dense. This stage captures spatial correlations by concatenating independently-pooled histograms at fixed spatial locations.

Sub-sampling (or max-pooling) \( y_i \in S \subset D \). Co-variant sampling of the anti-aliased representation. This could be preceded or followed by additional normalization. The next stage re-establishes a representation that is low-entropy, i.i.d., and sparse.

Stage 6: Coding (sparsification) \( \theta_5 = \theta_5(k) + n \). Each location \( y_i \) with \( i \in S \) has attached to it a “vector” \( \phi_5(y_i|\theta) = p(y_i|\theta_5) \) that is high-entropy (many components of \( y_i \) are non-zero).
dense (the vector is non-zero for many locations \((u_i, v_i)\)), and high-dimensional \((\sigma_5)\), even though it has been sub-sampled. We can retrieve a representation that is low-entropy/sparse by encoding \(\phi_5\) relative to an over-complete dictionary \(\{M_1, \ldots, M_K\}\), for instance obtained using \(K\)-means, so that \(F_k(u_i, v_i) = \delta(\theta_5(u_i, v_i) - M_k)\). Then

\[
\phi_6(y_i|\theta) = p(y_i|\theta_6) = \int p(y_i|\theta_5)p(\theta_5 - \theta_6)d\theta_5 = \int \phi_5(y_i|\theta_5)N(\theta_5 - M_k)d\theta_5 = \mathcal{M} \cdot \phi_5(y|\theta_5)
\]

In the most extreme case, we simply have

\[
\phi_6(y|\theta) = \delta(\phi_5(y|\theta) - M_k)
\]

point-wise, which we can interpret as new data already embedded as in Stage 0: \(\phi_6 : S \subset D \times [0, \ldots, K] \rightarrow \{0, 1\}; (u_i, v_i, k) \rightarrow p(y_i|\theta_5(u_i, v_i)) = \phi_6(y_i|\theta)\) which can be fed to Stage 0 of the next layer.

The likelihood, for instance assuming 5 layers, can be computed via

\[
\phi_5(y) = p(y|\theta_5) = \int \phi_0(y)\phi_1(\theta_0)\phi_2(\theta_1) \ldots \phi_5(\theta_5)d\theta_0 \ldots d\theta_5
\]

where the marginalization is computed by the network through a series of steps including encoding (matrix multiplication), normalization, mean and max pooling. Then localization can be solved by

\[
\omega = \arg \max\limits_{(u, v) \in \omega} p(y(u, v)|\theta_5)
\]

Note that the same maximum, restricted to other layers \(\theta_n\), provides activation maps that could be used for other purposes.

The derivation shows that a CNN represents a scene \(\theta\) as a sequence of selections of regions of deformed images, rather than explicitly attempting to reconstruct shape, reflectance and occlusions. Shape (modulo nuisance variability) is encoded in the deformations, and reflectance in the gradient orientation distribution. Depending on the task, either one can be a nuisance.

### IMPLEMENTATION

To compare a domain-size pooled CNN (DSP-CNN) to an ordinary CNN, we use discriminatively pretrained models (ILSVRC-2012 Russakovsky et al. (2014)) fine-tuned on Pascal 2007 train-validation data, tested on the VOC test set. Considering that Pascal VOC is a multi-label dataset, the softmax regression loss can be replaced with either one-vs-rest classification hinge loss or a ranking hinge loss. To keep things simple (with an expected small performance loss) we keep softmax and augment the training set so that images with multiple labels are entered in the pipeline at each epoch as many times as labels they have.

The pretrained on ILSVRC-2012 model which is described as CNN-S in Chatfield et al. (2014) is used in our experiments. Then similar fine-tuning protocol with the one suggested by the authors is followed. In order to control overfitting, we use the following learning rate schedule: \(10^{-3} / 10^{-4}\) (epochs \(1 - 10\)), \(10^{-4} / 10^{-5}\) (epochs \(11 - 20\)), \(10^{-5} / 10^{-6}\) (epochs \(21 - 100\)) (first/second number pertain to last/hidden layers correspondingly). Average-pooling over scale is deployed for conv1 and conv2 layers (3 scales; ratios 0.6, 1, 1.4). No data augmentation is applied for either training or testing, which explains the lower reported mAP statistics compared to the numbers that are reported in Chatfield et al. (2014). Our interest is only to evaluate the relative merit of DSP, so the absolute numbers are not as relevant.

All trainings were performed using the open source MatConvNet toolbox. Representative results are shown in Fig. 5 and 6. The improvement, also reported in Table 1 is marginal but nevertheless present, as expected considering that the number of domain-size pooled, in addition to the base size, is two, and domain-size pooling is restricted to the first two layers.

---

8 http://www.vlfeat.org/matconvnet
Figure 5: Head to head comparison between CNN and DSP-CNN. Each point corresponds to the Average Precisions for one class from Pascal 2007 VOC Everingham et al. (20 categories in total). The coordinates indicate average precision for each method and the relative performance improvement of the winner is shown in the title. DSP-CNN consistently outperforms its CNN counterpart.

Figure 6: Precision-recall curves over 20 classes in the Pascal 2007 dataset. DSP-CNN is plotted in blue, while the original CNN in red.

### Table 1: PASCAL VOC 2007 results.

<table>
<thead>
<tr>
<th>Class</th>
<th>CNN</th>
<th>DSP-CNN</th>
<th>Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aeroplane</td>
<td>0.8323</td>
<td>0.7554</td>
<td>-0.0769</td>
</tr>
<tr>
<td>Bicycle</td>
<td>0.7754</td>
<td>0.7828</td>
<td>0.0074</td>
</tr>
<tr>
<td>Bird</td>
<td>0.8312</td>
<td>0.8282</td>
<td>-0.0030</td>
</tr>
<tr>
<td>Boat</td>
<td>0.3700</td>
<td>0.3700</td>
<td>0.0000</td>
</tr>
<tr>
<td>Bottle</td>
<td>0.6401</td>
<td>0.6401</td>
<td>0.0000</td>
</tr>
<tr>
<td>Bus</td>
<td>0.6244</td>
<td>0.6244</td>
<td>0.0000</td>
</tr>
<tr>
<td>Car</td>
<td>0.7544</td>
<td>0.7544</td>
<td>0.0000</td>
</tr>
<tr>
<td>Cat</td>
<td>0.6091</td>
<td>0.6147</td>
<td>0.0056</td>
</tr>
<tr>
<td>Chair</td>
<td>0.6464</td>
<td>0.6464</td>
<td>0.0000</td>
</tr>
<tr>
<td>Cow</td>
<td>0.5420</td>
<td>0.5420</td>
<td>0.0000</td>
</tr>
<tr>
<td>Table</td>
<td>0.8073</td>
<td>0.8073</td>
<td>0.0000</td>
</tr>
<tr>
<td>Dog</td>
<td>0.7724</td>
<td>0.7724</td>
<td>0.0000</td>
</tr>
<tr>
<td>Horse</td>
<td>0.7043</td>
<td>0.7094</td>
<td>0.0051</td>
</tr>
<tr>
<td>Motorbike</td>
<td>0.6935</td>
<td>0.6935</td>
<td>0.0000</td>
</tr>
<tr>
<td>TV Monitor</td>
<td>0.5835</td>
<td>0.5835</td>
<td>0.0000</td>
</tr>
<tr>
<td>TV Monitor</td>
<td>0.6785</td>
<td>0.6785</td>
<td>0.0000</td>
</tr>
<tr>
<td>mAP</td>
<td>0.6824</td>
<td>0.6673</td>
<td>-0.0151</td>
</tr>
<tr>
<td>mAP</td>
<td>0.6824</td>
<td>0.6673</td>
<td>-0.0151</td>
</tr>
<tr>
<td>mAP</td>
<td>0.4935</td>
<td>0.5328</td>
<td>0.0393</td>
</tr>
<tr>
<td>mAP</td>
<td>0.6940</td>
<td>0.7039</td>
<td>0.0099</td>
</tr>
</tbody>
</table>

DSP−CNN defeats CNN: 2.72%