VISUAL SCENE REPRESENTATIONS:
SUFFICIENCY, MINIMALITY, INVARIANCE AND APPROXIMATIONS

Stefano Soatto *
UCLA Vision Lab
University of California, Los Angeles
soatto@ucla.edu

ABSTRACT

Visual representations are functions of visual data that are minimal sufficient statistics for a class of tasks and maximally invariant to nuisance variability. Minimal sufficiency guarantees that we can store the statistic in lieu of the raw data with no performance loss and smallest complexity. Maximal invariance guarantees that the statistic is constant with respect to unwanted transformations of the data, and nothing else. We derive an expression for such representations and show that, under certain restrictive assumptions, they are related to “feature descriptors” commonly in use in the computer vision community, as well as to increasingly popular convolutional architectures. This link highlights the conditions tacitly assumed by these descriptors and networks, under which they can be expected to perform well, and also suggests ways to improve and generalize them, by lifting such assumptions. 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 the images that is “useful” for a task. The task is the inference of properties of the “scene”. Such properties could be geometric (shape), photometric (reflectance), dynamic (motion) and semantic (identities, relations). In addition to such properties, the data also depends on a variety of nuisance variables that are irrelevant to the task. They include unknown characteristics of the sensor (intrinsic calibration), its inter-play with the scene (viewpoint, occlusion), and properties of the scene that are not directly of interest (e.g., illumination).

We are interested in the analysis and design of visual representations: How can we measure “useful” for functions of the data (i.e. statistics)? What guidelines or principles should inform their design? Is there such a thing as an optimal representation? Can it be computed? If not, approximated? Can it be learned?

When restricted to subsets of the images, image-based representations are known as local descriptors, typically involving weighted histograms of gradient orientation, computed at selected scales and locations, pooled over spatial regions, thought of as vectors, and variously normalized Lowe (2004). This process can be stacked to yield a convolutional architecture Bruna & Mallat (2011). But why a histogram? Why computed at extremal scales? Why aggregated spatially? Why weighted by the norm of the gradient? Are these choices sensible? Necessary? Optimal? Can analysis point the way to improvements, as opposed to just numerical end-to-end comparison of alternatives in empirical tests?

1.1 RELATED WORK, CONTRIBUTIONS AND CAVEATS

There are many local descriptors in the literature, more than we can review here, and empirical comparisons (e.g., Mikolajczyk et al. (2004)) have recently expanded to include convolutional networks Fischer et al. (2014). Unfortunately, there are so many implementation details and parameters that it is hard to draw general conclusions Chatfield et al. (2011), let alone insight on why a given descriptor works, under what assumptions it does, and how it could be improved.

We take a different approach, and derive a formal expression of optimal representations from basic principles: Sufficiency, minimality, invariance, completeness. Then, we show how existing descriptors are related to optimal repre-
sentations, highlighting the (often tacit) underlying assumptions, that give insight on why a given descriptor can be expected to work (or not), and point the way to improvements.

This paper is organized as follows:
1) We formalize optimal representations as maximal invariants and minimal sufficient statistics of nuisance transformations of a basic data formation model informed by the task. Image matching is taken as an elementary task.
2) We explicitly characterize optimal descriptors when such nuisances include contrast transformations and planar similarities. The optimal descriptor is interpreted as a likelihood function, a novel interpretation that, while unintuitive, offers insight to extensions and improvement of existing descriptors.
3) We interpret the construction of invariant statistics in terms of sampling and anti-aliasing as taught in traditional sampling theory. This interpretation allows clear specification of space and time complexity, and of their tradeoff with fidelity. It also explains “pooling,” as ordinarily performed in convolutional network architectures, and its relation to invariance.
4) We explicitly consider occlusion as show how its marginalization, otherwise intractable, can be approximated. This provides a justification for the use of “patches” or “receptive fields,” through a link to the optimal correspondence hypothesis test. 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 harmonic analysis and scale-space theory, another unintuitive consequence of the analysis.

Our work relates to Bruna & Mallat (2011) in its aim to construct and analyze complete representations of image data for classification tasks. However, we care to represent the scene, rather than the image, so occlusion and locality play a key role in our models. Our work is also related to Dong et al. (2013) where multiple view local descriptors are derived from first principles, and to Morel & Yu (2011) in its aim to characterize the maximal invariant to certain nuisance transformations. Our models are more general, involving transformations of both the range and the domain of the data. It also relates to Tishby et al. (2000) in its attempt to formalize a notion of optimal representation, albeit we operate in a more restricted setting, and to the growing literature on representation learning Ranzato et al. (2007); LeCun (2012); Simonyan et al. (2014); Serre et al. (2007); Bouvier et al. (2009); Susskind et al. (2011), that however imposes a certain architecture by fiat, whereas we are interested in understanding what architecture would be best under explicitly stated conditions. We present an alternate interpretation of pooling, in the context of classical sampling theory, that differs from other analyses Gong et al. (2014); Boureau et al. (2010).

A number of background concepts used in this paper are listed in the appendix for reference, and assumed known to the reader. Our hope is that, for those with the necessary background, this paper will offer insight into the design, analysis and learning of visual representations that would enable improving them.

## 2 Designing and Learning Visual Representations

Because there is uncertainty in the mechanisms that generate images, we treat them as random vectors \( y \). They have high but finite dimension, say in the millions. On the other hand, the scene that generates them is infinitely more complex.\(^1\) Thus, even if we restrict our attention to a single static scene, and think of it as a “model” or “parameter” \( z \), then \( z \) is in general infinite-dimensional. As a result, we will never know the “true” scene. It is also technically difficult to endow the space of scenes with metric and probabilistic structures. Nevertheless, we can ask questions about the scene, by interrogating the data Geman & Geman (2014). The number of questions (“classes”) \( K \) can be large, but is assumed finite. Correspondingly, there is a partition of the scene, which we will indicate with \( \{z_1, \ldots, z_K\} \), the class “centers.”

### 2.1 Desiderata for Representations

Among (measurable) functions of the data (a.k.a. “statistics”), a representation should be “useful” to the task. The data itself is useful, as it informs answers to questions about the scene. This is captured by the notion of mutual information between the data \( y \) and the object of interest \( z \) (Cover & Thomas (1991) eq. 2.28, page 18). While a representation can be no more informative than the data (Data Processing Inequality, page 88 of Shao (1998)), ideally it should be no less. This is captured by the notion of sufficient statistic (Sec. 6.7 of DeGroot (1989), page 356). It should also

\(^1\)Scenes are made of surfaces, that support reflectance functions, that interact with illumination, etc. No matter how many images of a scene we already have, even a single static scene can generate images that are different, even simply by changing changing vantage point.
be “simpler” than the data. This is captured by the notion of minimal sufficient statistic (DeGroot (1989), page 368). Finally, the representation should discount the effects of nuisance variables \( g \), and ideally be invariant to them (Shao (1998), Sec. 4.2, page 213). So, the ideal representation of the image(s) \( y \) for answering questions about the scene \( z \) should be a minimal sufficient statistic that is invariant to nuisance variables \( g \).

### 2.2 Impossibilities

Unfortunately, even defining some of the desired properties above is challenging, let alone computing them and optimizing them 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 \( z \). Even if it was possible, sufficient statistics with finite dimension and defined on a constant domain exist only for the exponential family of distributions (Pitman-Koopman-Darmois’ theorem Diaconis (1988)). The Information Bottleneck Tishby et al. (2000) was devised as an alternative to, and approximation of, minimal sufficient statistics, with an explicit complexity cost or bound. However, it is defined in terms of mutual information and therefore it is subject to the same limitations thus described.

### 2.3 The likelihood function

Despite the difficulties above, we can certainly collect samples of data \( \{y_1, \ldots, y_t\} \sim y^t \) generated by a particular scene \( z \), \( y^t \sim p(y|z) \), given which the likelihood function

\[
L^t(z) = p(y^t|z)
\]

is a minimal sufficient statistic.\(^2\) In practice, for each question after is a statistic that is function, or “descriptor” of the scene that the representation as defined may be quite uninformative of the scene \( z \). The descriptor is thus to be thought of as a minimal sufficient statistic.

\[\hat{p}_{z_k}(y) \doteq p(y|y^t), \quad y^t \sim p(y|z_k) \]

where \( p(y|y^t) \) can be approximated by a density estimator with kernel \( \kappa_\sigma \) of width \( \sigma > 0 \):

\[
\hat{p}_{z_k}(y) \simeq \sum_{y_i \sim p(y|z_k)} \kappa_\sigma (y - y_i)
\]

or otherwise a regularized histogram with \( \sigma \) the width of the bins. Then the collection

\[
\hat{p}_{z_k}(y) = \{\hat{p}_{z_k}(y), \ldots, \hat{p}_{z_k}(y)\}
\]

is a minimal sufficient statistic of \( y \) for the purpose of answering questions about \( z \).

The last statement can be confusing, as \( \hat{p} \) is an infinite-dimensional function of the “test” image \( y \), which we have yet to measure, and dependent on the scene \( z \), which is also infinite-dimensional and unknown. However, note that \( \hat{p} \) depends on \( z \) only indirectly, through the sample \( y^t \sim p(y|z) \).\(^3\) As \( K \) grows to infinity and classes become closer, the entire likelihood function \( \hat{p}_{z_k}(y) \) has to be estimated. So, from now on we drop the hat (and the superscript \( t \) when the size of the training set is not critical to the discussion) and refer to the approximation \( \hat{p}_{z_k}(y) \) as simply the likelihood function, or “descriptor” of the scene \( z \) computed from the training sample \( y^t \), which is a function of the generic test sample \( y \), which we have yet to measure

\[
\phi(|z) \doteq \hat{p}_{z}(\cdot) \simeq \{p(\cdot|y^t), \ y^t \sim p(y|z)\}
\]

The descriptor is thus to be thought of not as a vector but as a function of the test datum \( y \), computed or learned from the training set \( y^t \), with the result of learning being an approximation \( \phi(\cdot|z) \) of the likelihood function \( p(\cdot|z) \). Note that the representation as defined may be quite uninformative of the scene \( z \), but that is irrelevant here, for what we are after is a statistic that is as informative as the (training) data, however good or bad that is. Once invariance is brought into the picture, this is what is called Actionable Information in Soatto (2009). To obtain a complete representation, which would be equivalent to knowing the scene, we would have to ensure that the sample is sufficiently exciting, an

\(^2\) The sense in which \( L^t(z) \) is minimal sufficient is explained by Bahadur (1954) in Theorem 6.1 and the corollary that follows (see also Barndorff-Nielsen et al. (1976); Fraser & Naderi (2007)).

\(^3\) In probabilistic generative modeling, we would reduce \( z \) to finite dimensions, treat it as a latent variable and compute its posterior. But we know of no viable ways of discretizing the scene \( z \) or putting a prior on it. Instead, we discretize the likelihood \( p \) and compute a finite-complexity approximation of it, non-parametrically or through a regularized estimator such as a mixture density, which would entail some clustering (unsupervised learning).
asymptotic notion, that generally requires experiment design Fedorov (1972), that is the design or control of the data acquisition process so as to ensure that the sample is maximally informative. This is also known as active learning and active sensing in other contexts.

The main challenge in computing the representation above is approximating the density \( p(y|y^t) \) in a space of the dimension of \( y \). Even restricting the attention to small images, say \( 10 \times 10 \), estimating densities in 100 dimensions is hard. If nuisances were known, the problem would be greatly simplified by conditioning on them; for instance, conditioning on viewpoint \( g \in SE(3) \) in the Lambert-Ambient model makes pixels independent:

\[
p(y|z, g) = \prod_i p(y_i|z, g)
\]

(6)

Naturally, the nuisance \( g \) is unknown.

### 3 Dealing with Nuisances

In this section we derive complete invariants that remove nuisance variability with no loss of discriminative power. Note we do not seek “approximate” or “local invariance” by “blurring,” which is at the expense of discriminative power. Interpreting the descriptor as a likelihood function, rather than a “vector,” means that that integration along the observation becomes available. On the other hand, evaluation of the marginalized density is hard. If nuisances were known, the problem would be greatly simplified by blurring. We perform the derivation for the actual likelihood \( p(y|z) \), although all derivations translate to its approximation \( \phi(y|z) \).

#### 3.1 Marginalization and max-out

If the data depends on an unknown nuisance variable \( g \in G \) and a prior \( dP(g) \) is available, it can be marginalized:

\[
p_o(y|z) \doteq \int_G p(y|z, g) dP(g).
\]

(7)

For instance, marginalization of the group of contrast transformations of the range of the image is derived in Sect. 3.2. Alternatively, if no prior is available, one could consider the entire “orbit”

\[
[p(y|z, g)] = \{p(y|z, g), \ g \in G\}
\]

(8)

which is a maximal invariant of the likelihood function under the action if \( G \) is a group. The advantage of retaining the orbit is that conditioning on the vantage points makes each pixel independent (6), so 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 (a search referred to as “max-out”):

\[
\phi_G(y|z) \doteq \arg \max_{g \in G} p(y|z, g).
\]

(9)

that cannot be performed until the test datum \( y \) becomes available. On the other hand, evaluation of the marginalized likelihood is straightforward and does not entail any search. The downside is, because of the spatial dependencies, this representation is difficult to construct. In Sect. 4 we show how to trade off the two, which can be cast in the framework of classical sampling theory. Before then, we illustrate marginalization for the case of contrast transformations.

#### 3.2 Contrast Invariance via Marginalization

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.

However, it is not sufficient to replace \( p(y|z) \) with the density of the gradient orientation, \( p(\angle \nabla y|z) \), since contrast changes can affect both the images used for training, \( y^t \), and the test image \( y \). The following theorem derives an expression for the maximal invariant, which is obtained via marginalization of the norm of the gradient.

**Theorem 1.** Given a training image \( z \) and a test image \( y \), assuming that the latter is affected by noise that is independent in the gradient direction and magnitude, then the maximal invariant of \( y \) to the group \( G \) of contrast transformations is given by

\[
p_G(y|z) = p(\angle \nabla y|z) \parallel \nabla z\parallel.
\]

(10)
The independence assumption above is equivalent to assuming that the gradient magnitude and orientation of \( y \) are related to the gradient magnitude and orientation of \( z \) by a simple additive model: \( \| \nabla y \| = \| \nabla z \| + n_\rho \) and \( \angle \nabla y = \angle \nabla z \oplus n_\theta \), where \( \oplus \) denotes addition modulo \( 2\pi \), and \( n_\rho \) and \( n_\theta \) are independent.\(^4\) These are all modeling assumptions, clearly not strictly satisfied in practice, but reasonable first-order approximations. The theorem is proven in the appendix. Note that, other than for the gradient, the computations above can be performed point-wise, so we could write (10) at each pixel \( y_i \). To simplify the notation, we call \( \theta \equiv \angle \nabla y_i \), and write

\[
\phi_{y_i}(\theta|z) = N_{G} (\theta - \angle \nabla z; \epsilon_\theta) \| \nabla z_i \| \tag{12}
\]

Since we always require contrast-invariance, we omit the subscript \( G \) when referring to contrast (which is clear from the argument \( \theta \)), and also the subscript of the point \( y_i \).

**Remark 1 (Bayesian descriptor).** 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 \( \| \nabla z \| \). Clamping, described next, can be understood as a particular choice of prior for marginalization of the gradient magnitude.

**Remark 2 (No invariance for \( z \)).** Note that (12) is a maximal invariant to contrast transformations of \( y \), but it is not invariant to contrast transformations of \( z \). Since the minimal sufficient statistic is a likelihood function, not a density on \( z \), the latter cannot be meaningfully marginalized. Nevertheless, contrast changes of \( z \) can be handled by normalization as we will see next. Also note that the normalization factor cannot be neglected as common in likelihood ratio tests, as different hypotheses yield different factors.

3.3 **Normalization and “clamping”**

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

\[
\bar{\phi}(\theta|z) \doteq \frac{\phi(\theta|z)}{\|\phi(\theta|z)\|_{\ell^1}} = \frac{p(\theta|z)\|\nabla z\|}{\int p(\theta|z)d\theta\|\nabla z\|} = p(\theta|z) \tag{13}
\]

instead of the customary \( \ell^2 \) Lowe (2004). 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, \( \phi_{y_i}(\theta|z) \) 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. 1.

4 **Sampling, pooling and canonization**

As seen in Sect. 3.1 achieving invariance by retaining the orbit (8), instead of marginalizing the group (7), has the advantage of reducing correlations, but at the expense of forcing a search at decision time (9). The damage can be limited by selecting as few samples as possible on the orbit. Characterizing the tradeoff between the number of samples (complexity) and the approximation error (fidelity) is a goal of sampling theory Marvasti (2001). In our case, the “original” (continuous) function is the orbit \( f(g) = [p(y|z, g)] \), which we must sample at \( \{g_j\}_{j=1}^L \), trading off fidelity, measured by classification error.

\(^4\)Note that we are not assuming 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 \( \epsilon_\rho \) sufficiently small to ensure that the samples are positive with high probability, and the second as an angular Gaussian Watson (1983):

\[
p(\|\nabla y\|, \|\nabla z\|) = N(\|\nabla y\| + \|\nabla z\|; \epsilon_\rho) \]

\[
p(\angle \nabla y, \angle \nabla z) = N_{G} (\angle \nabla y - \angle \nabla z; \epsilon_\theta) \tag{11}
\]
4.1 Pooling = Anti-Aliasing

When the number of classes grows to infinity, as discussed in Sect. 2.3, sampling error reduces to the approximation error between the orbit and its reconstruction using the samples \( \{ f(g_l) \}_{l=1}^L \) = \( \{ p(y | z, g_l) \}_{l=1}^L \). Sampling below Nyquist’s rate introduces aliasing artifacts, that can be detrimental in our context because such artifacts represent “structures” that do not exist in the real scene (aliases). Traditional sampling theory teaches that aliasing can be reduced by anti-aliasing, which is the (weighted\(^5\)) averaging of the original signal around the samples:

\[
\tilde{f}(g_l) = \int_G p(y | z, g_l) dP(g)
\]

where \( g \in G \) is a group action and \( dP \) is a weight or measure, for instance uniform in a small ball around the origin. This anti-aliasing, restricted to planar translation, is known as spatial pooling in most descriptors as well as convolutional architectures. However, there is no reason to restrict oneself to the translation group.

4.2 Canonization = Sampling

Now that we know how to anti-alias, we have to decide how to sample. In the context of achieving invariance to group actions, the samples should co-vary with the group. Canonization is an adaptive sampling mechanism that can also be thought of as a vehicle to perform co-variant sampling when the nuisance acts in data space (rather than on the hidden variable \( z \)); we indicate the action as \( gy \).

Let \( \psi \) be a (co-variant detector) functional such that \( \psi(y, g) = 0 \) locally uniquely determines functions \( \tilde{g}_l(\cdot) \) via the implicit function theorem Guillemin & Pollack (1974), \( \psi(y, \tilde{g}_l(y)) = 0 \), in a way that co-varies, \( \tilde{g}_l(gy) = g\tilde{g}_l(y) \). Then the statistic \( \phi(y) \doteq \tilde{g}_l^{-1}(y)y \) is a maximal \( G \)-invariant since from it one can reconstruct the entire orbit \( g\phi(y) \).

As there are multiple isolated extrema, \( \tilde{g}_l \), matching entails a search (max-out), but now with a smaller number of samples. In between samples, we can perform anti-aliasing (marginalization), as shown above.

**Example 1** (Rotation Invariance). Canonization is particularly well suited to deal with planar rotation, since it is possible to design co-variant detectors with few isolated extrema. An example is the local maxima of the norm of the gradient along the direction \( \theta \), \( \tilde{h}_l = \tilde{h}_l(y) \).\(^6\) Invariance to \( G = SO(2) \) can be achieved by retaining the samples

\[
p_G(\theta | z) = \{ p(\theta | z, \hat{\theta}_l) \}_{l=1}^L
\]

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

\[
p_G(\theta | z) = p(\theta | z, \hat{\theta}).
\]

In reality, rotation canonization should contend with spatial quantization, neglected here since rotation errors are absorbed by the binning of gradient orientation in the descriptor.

4.3 Complexity

As we have pointed out in Sect. 4, conditioning on samples of the nuisances reduces dependencies, whereas marginalization introduces correlations. On the other hand, more samples increase complexity at decision time. The trade-off between the two is a standard sampling/aliasing problem, where the theory suggests local fine marginalization (anti-aliasing, or “mean pooling”) around global co-variant samples (“max pooling” or canonization). The maximal invariant and minimal sufficient statistic is then approximated by:

\[
\phi_G(\cdot | z) = \left\{ \int p(\cdot | z, g_l g) dP(g) \right\}_{l=1}^L
\]

---

\(^5\)In general anti-aliasing, the weights are allowed to take negative values whereas here, to be considered proper marginalization, they are constrained to be positive and normalized.

\(^6\)Here \( g \) acts on \( y \) via \( gy(u_i', v_i') = y(u_i', v_i') \) where \( u'' = u \cos \theta - v \sin \theta \) and \( v'' = u \sin \theta + v \cos \theta \), and a canonical element \( \tilde{g}_l(y) = \tilde{\theta} \) can be obtained as \( \tilde{\theta} = \arg \max u' \| \nabla y(u_i', v_i') \| \). The corresponding rotation invariant \( \tilde{g}_l^{-1}(y)y \) is \( \tilde{\theta} \) of \( y \).

\( \parallel \cdot \parallel \) averaging of the original signal around the samples:

\[
\tilde{f}(g_l) = \int_G p(y | z, g_l) dP(g)
\]

where \( g \in G \) is a group action and \( dP \) is a weight or measure, for instance uniform in a small ball around the origin. This anti-aliasing, restricted to planar translation, is known as spatial pooling in most descriptors as well as convolutional architectures. However, there is no reason to restrict oneself to the translation group.

4.2 Canonization = Sampling

Now that we know how to anti-alias, we have to decide how to sample. In the context of achieving invariance to group actions, the samples should co-vary with the group. Canonization is an adaptive sampling mechanism that can also be thought of as a vehicle to perform co-variant sampling when the nuisance acts in data space (rather than on the hidden variable \( z \)); we indicate the action as \( gy \).

Let \( \psi \) be a (co-variant detector) functional such that \( \psi(y, g) = 0 \) locally uniquely determines functions \( \tilde{g}_l(\cdot) \) via the implicit function theorem Guillemin & Pollack (1974), \( \psi(y, \tilde{g}_l(y)) = 0 \), in a way that co-varies, \( \tilde{g}_l(gy) = g\tilde{g}_l(y) \). Then the statistic \( \phi(y) \doteq \tilde{g}_l^{-1}(y)y \) is a maximal \( G \)-invariant since from it one can reconstruct the entire orbit \( g\phi(y) \).

As there are multiple isolated extrema, \( \tilde{g}_l \), matching entails a search (max-out), but now with a smaller number of samples. In between samples, we can perform anti-aliasing (marginalization), as shown above.

**Example 1** (Rotation Invariance). Canonization is particularly well suited to deal with planar rotation, since it is possible to design co-variant detectors with few isolated extrema. An example is the local maxima of the norm of the gradient along the direction \( \theta \), \( \tilde{h}_l = \tilde{h}_l(y) \).\(^6\) Invariance to \( G = SO(2) \) can be achieved by retaining the samples

\[
p_G(\theta | z) = \{ p(\theta | z, \hat{\theta}_l) \}_{l=1}^L
\]

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

\[
p_G(\theta | z) = p(\theta | z, \hat{\theta}).
\]

In reality, rotation canonization should contend with spatial quantization, neglected here since rotation errors are absorbed by the binning of gradient orientation in the descriptor.

4.3 Complexity

As we have pointed out in Sect. 4, conditioning on samples of the nuisances reduces dependencies, whereas marginalization introduces correlations. On the other hand, more samples increase complexity at decision time. The trade-off between the two is a standard sampling/aliasing problem, where the theory suggests local fine marginalization (anti-aliasing, or “mean pooling”) around global co-variant samples (“max pooling” or canonization). The maximal invariant and minimal sufficient statistic is then approximated by:

\[
\phi_G(\cdot | z) = \left\{ \int p(\cdot | z, g_l g) dP(g) \right\}_{l=1}^L
\]

---

\(^5\)In general anti-aliasing, the weights are allowed to take negative values whereas here, to be considered proper marginalization, they are constrained to be positive and normalized.

\(^6\)Here \( g \) acts on \( y \) via \( gy(u_i', v_i') = y(u_i', v_i') \) where \( u'' = u \cos \theta - v \sin \theta \) and \( v'' = u \sin \theta + v \cos \theta \), and a canonical element \( \tilde{g}_l(y) = \tilde{\theta} \) can be obtained as \( \tilde{\theta} = \arg \max u' \| \nabla y(u_i', v_i') \| \). The corresponding rotation invariant \( \tilde{g}_l^{-1}(y)y \) is \( \tilde{\theta} \) of \( y \).

\( \parallel \cdot \parallel \) averaging of the original signal around the samples:
This procedure highlights the complexity in the representation, that in (4) was first presented as an infinite-dimensional
object. The time complexity is simply given by the number of samples \( L \). The space complexity, or memory, is the
approximation complexity of the marginalized descriptor. In a non-parametric setting, this is related to the width of the
histogram bins for \( \theta, \epsilon_p \). In a parametric setting, it is the number of modes or cluster centers for the mixture density.
In any case, this is smaller than the number of samples, the trivial sufficient statistic.

5 Conclusions

So far we have discussed nuisances due to the action of groups. A qualitatively different source of variability is
occlusion: 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. In theory, then, correspondence entails searching through (9), or marginalizing (7), all possible subsets
of the test image. This power set is clearly intractable even for very small images.

5.1 Bypassing shape and justifying “patches” or “receptive fields”

We illustrate a principle to bypass combinatorial explosion for the case of a single training image, absent all other
nuisances. A training image \( z \) and a test image \( y \) correspond (null hypothesis, \( H_0 \)) if there exists a subset of \( z, \Omega_z \),
and a subset of \( y, \Omega_y \), such that, restricted to these subsets, the two differ by a white (zero-mean, uninformative)
residual.\(^7\) Under this simplistic model, the subsets \( \Omega_z = \Omega_y = \Omega \) are the same, and \( y = z + n \) where \( n \) is either
a white (spatially i.i.d) zero-mean process with a small covariance, \( n_{\Omega} \sim N(0, \epsilon^2) \) 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, \( n_{\Omega} \sim \mathcal{U} \). The alternate hypothesis \( H_1 \) is that there exists no such region, and \( n \sim \mathcal{U} \) 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

\[
L(y, z) = \frac{p(y|z, H_0)}{p(y|z, H_1)} = \frac{\max_{\Omega} p(y_{\Omega}|z_{\Omega}, H_0) p(y_{\Omega}|z_{\Omega}, H_0)}{\max_{\Omega} p(y_{\Omega}|z_{\Omega}, H_1) p(y_{\Omega}|z_{\Omega}, H_1)} = \frac{\max_{\Omega} \mathcal{N}(y_{\Omega} - z_{\Omega}; \epsilon^2)}{\max_{\Omega} \mathcal{U}(y_{\Omega} - z_{\Omega})}
\]

The numerator is the product of Gaussians, one for each pixel, and the log likelihood is the \( \ell^2 \) norm of the residual.
Reasoning in terms of pixels (i.e. treating each pixel as an independent decision), missed detections (treating a co-
visible pixel as occluded) and false alarms (treating an occluded pixel as visible) have different cost: Omitting a pixel
from the correct region \( \Omega \) decreases the likelihood by a factor corresponding to forgoing multiplication by a Gaussians
of samples are drawn from the same distribution; vice-versa, including a pixel from \( \Omega^C \) decreases the log-likelihood
by a factor equal to multiplying by a Gaussian evaluated at points drawn from another distribution.

Therefore, 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 shape of the regions, leaving
only their size to be marginalized, or searched over.\(^8\) This reasoning justifies the use of “patches” or “receptive fields”
to seed image matching.

5.2 Marginalizing visibility

Much of the literature on local descriptors further simplifies correspondence by selecting scales (Lowe, 2004), using
the appearance of the scene \( z \) to determines the size of the region \( \Omega \) where the descriptor is computed. This is

\(^7\) Of course, absent all other nuisances, \( p(y|z) \) factorizes into the product of the marginals (all pixels are independent), so corresponding regions can be determined by “background subtraction” techniques. In our case, determination of corresponding regions has to be wrapped around the management of other nuisances, so we cannot treat pixels as independent, but this assumption serves just to illustrate the principles.

\(^8\) 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.

7

8
motivated by Lindeberg (1998), and Mallat (1989); Bruna & Mallat (2011), where the tying of appearance (e.g., spatial frequencies) and size is known as the “uncertainty principle”. But while the tie makes sense to reconstruct 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 included among nuisance factors to be managed by max-out or marginalization. This suggests a way to improve existing descriptors by pooling domain sizes. In other words, while (co-variant) scale selection can be thought as a way to sample scale, as opposed to selecting corresponding scales, the co-variant sampling procedure must be coupled with anti-aliasing, which in this context consist of domain-size pooling (DS pooling), a concept introduced in Dong & Soatto (November 12, 2014) and illustrated below for the case of a descriptor built on a single image.

6 SIFT revisited

A single image does not afford the ability to separate nuisance from intrinsic variability. Nevertheless, if one image is all one has, it is legitimate to ask what is the best function of that image (or part of it) for any task that requires viewpoint and contrast invariance. SIFT aggregates samples spatially in a neighborhood, independently at a collection of (sixteen) neighborhoods. In Dong et al. (2013); Vedaldi & Fulkerson (2010) SIFT is written explicitly as a histogram, which shows that only small translations are marginalized, samples are aggregated in a weighted fashion in a neighborhood, and the histograms computed independently at each neighborhood (although there is some overlap) are 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 \( y \) are (spatially) independent.

Moreover, for the spatial samples to be related to actual samples, it is necessary to assume that the image is (locally) stationary and ergodic, another assumption that is patently violated. Finally, 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), not in scale, which in SIFT corresponds to domain size. These considerations lead to the following claim:

**Theorem 2** (Assumptions implicit in SIFT/HOG). The SIFT descriptor approximates the ideal representation at a point under the assumption of a flat and fronto-parallel scene undergoing purely translational motion parallel to the image plane, when the underlying radiance is a sample of a process that is locally stationary and ergodic.

Note that we are referring to pixels within a cell, since multiple cells are stored in order, together with their coordinates \((1, 1)\) through \((4, 4)\) on the lattice of 16 pixels that are concatenated in the overall descriptor.

7 Extensions

The assumptions under which SIFT operates are highly restrictive. Some can be easily lifted.

7.1 Extensions to multiple views

The derivation of the maximal invariant/minimal sufficient statistic applies to any number of images, so the restriction to a single one appears counter-productive. Indeed, recently extensions of SIFT to multiple views have begun appearing Dong et al. (2013). Such extensions are based either on a direct point estimate of the representation \( p_x(y) \), which is called “reconstructive HOG” in Dong et al. (2013), or on a sample approximation of the likelihood function \( \hat{p}(y|z) \) called “multi-view HOG”.

When multiple views are present, one no longer relies on the assumptions of stationarity and ergodicity implicit in single-view descriptors, nor on the restriction of the marginalization to planar similarity. Indeed, Dong et al. (2013) marginalizes (approximately) the entire group of rigid motions.

Intermediate solutions where the planar affine or projective group is marginalized result in increased complexity at a modest performance gain. Solutions beyond spatial rigid motion, to arbitrary homeomorphisms of the plane, result in a loss of discriminative power, for in that case co-occurrence is the only spatial property that is preserved and one ends up with a “bag of features” model.
The assumption that all existing multiple-view extensions do not overcome is the conditional independence of the intensity of different pixels (6). This is the major step forward in convolutional architectures, described next.

7.2 Extensions to capture spatial correlations

It is well known that SIFT can be interpreted as one stage, or layer, of a convolutional neural network (CNN). To see that, it is sufficient to notice that the contrast-invariant term of the likelihood can be written as the projection of the “scene” $z$ (which in this case consists of a single image) onto a basis of Gabor filters at different orientation $\theta$ and scale $\sigma$, $G_{\sigma}(\theta)$:

$$\phi_{y_i}(\theta|z) \simeq p_{y_i}(\theta|z)|\nabla z| = \langle \nabla y_i G_{\sigma}(\theta), z \rangle |\nabla z|$$

(20)

Spatial pooling is then the marginalization of the translation group, customary in convolutional architectures, like the subsequent normalization. This is a layer of a convolutional network Ranzato et al. (2007), also related to the scattering transform Bruna & Mallat (2011).

In order to capture longer-range spatial correlations, the gradient filters can be replaced with other filters, so long as there is a topology in the basis that enables meaningful pooling. Max-pooling across each layer can then be interpreted as canonization or co-variant sampling as before. 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. Improvements in performance, without significant changes in the architecture or learning algorithms, may follow the marginalization of larger classes of groups.

This may be less critical for stacked networks than for local descriptors, however, since stacked networks represent spatial transformations hierarchically, where the composition of translations at each scale/stage can approximate arbitrary diffeomorphisms (in the limit, given sufficient layers and sufficient scales and transformations).

8 Discussion

We have formalized the problem of designing and learning visual representations using basic principles from statistical inference and decision theory. This enables us to relate existing descriptors to the ideal representation, highlighting the assumptions under which the former approximate the latter, and therefore can be expected to operate. Hopefully this will eventually result in quantitative performance bounds, whereby one can guarantee specifications under given assumptions, rather than relying on the interpretation of empirical test and the hope they will generalize to other datasets.

The principles described in this manuscript apply to neural network-based distributed representations just as they apply in local representations. The only difference is that local representations, or “patches” are designed to marginalize the effects of occlusions. These are not explicitly modeled in a convolutional architecture. Perhaps this analysis will inspire better models that can achieve invariance to larger classes of transformations without the need to augment the dataset with additional label-preserving transformations, keeping complexity at bay.
REFERENCES

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

Bahadur, R. R. Sufficiency and statistical decision functions. Annals of Mathematical Statistics, 25(3):423–462, 1954.

Barndorff-Nielsen, O., Hoffmann-Jørgensen, J., and Pedersen, K. On the minimal sufficiency of the likelihood function. Scandinavian Journal of Statistics, 3(1):37–38, 1976.

Boureau, Y.-L., Ponce, J., and LeCun, Y. A theoretical analysis of feature pooling in visual recognition. In Proc. Intl Conf. on Mach. Learning, pp. 111–118, 2010.

Bouvier, J. V., Rosasco, L., and Poggio, T. On invariance in hierarchical models. In NIPS, pp. 162–170, 2009.

Bruna, J. and Mallat, S. Classification with scattering operators. In Proc. IEEE Conf. on Comp. Vision and Pattern Recogn., 2011.

Chatfield, K., Lempitsky, V., Vedaldi, A., and Zisserman, A. The devil is in the details: an evaluation of recent feature encoding methods. 2011.

Cover, T. M. and Thomas, J. A. Elements of Information Theory. Wiley, 1991.

Dalal, N. and Triggs, B. Histograms of oriented gradients for human detection. In Proc. IEEE Conf. on Comp. Vision and Pattern Recogn., 2005.

DeGroot, M. H. Probability and statistics. Addison-Wesley, 1989.

Diaconis, P. Sufficiency as statistical symmetry. In Proceedings of the AMS Centennial Symposium, pp. 15–26, 1988.

Dong, J. and Soatto, S. Domain size pooling in local descriptors. Technical Report UCLA CSD140022, November 12, 2014.

Dong, J., Balzer, J., Davis, D., Hernandez, J., and Soatto, S. On the design and analysis of multiple view descriptors. ArXiv: 1311.6048, 2013.

Fedorov, V. V. Theory of optimal experiments. Elsevier, 1972.

Fischer, P., Dosovitskiy, A., and Brox, T. Descriptor matching with convolutional neural networks: a comparison to sift. ArXiv:1405.5769, 2014.

Fraser, D. and Naderi, A. Minimal sufficient statistics emerge from the observed likelihood function. International Journal of Statistical Science, 6:55–61, 2007.

Geman, D. and Geman, S. A restricted Turing test for vision. 2014.

Gong, Y., Wang, L., Guo, R., and Lazebnik, S. Multi-scale orderless pooling of deep convolutional activation features. arXiv preprint arXiv:1403.1840, 2014.

Guillemin, V. and Pollack, A. Differential Topology. Prentice-Hall, 1974.

Kadir, T. and Brady, M. Scale saliency: A novel approach to salient feature and scale selection. In International Conference Visual Information Engineering, 2003.

LeCun, Y. Learning invariant feature hierarchies. In ECCV, pp. 496–505, 2012.

Lindeberg, T. Principles for automatic scale selection. Technical report, KTH, Stockholm, CVAP, 1998.

Lowe, D. G. Distinctive image features from scale-invariant keypoints. IJCV, 2(60):91–110, 2004.

Mallat, S. A theory of multiresolution signal decomposition: the wavelet representation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 11:674–693, 1989.
A BACKGROUND CONCEPTS

We make extensive use of the following concepts and notation: **Random variables** $X$, their realizations $x$, distribution $dP_X(x)$, density $p_X(x)$ (which we assume exists and is related to the distribution by $dP_X(x) = p_X(x)dx$). When clear from the context, we drop the subscripts and indicate $p_X(x)$ as $p(x)$. **Implicit and inverse function** theorem. If $f : U \rightarrow V; u \mapsto v = f(u)$, then $f^{-1}(v) = \{u \mid f(u) = v\}$. **Change of measure** from $X$ to $Y = f(X)$: $dP_X(x) = dP_{f^{-1}(Y)}(x) = dP_Y(f(x))|J_f(x)|^{-1}$ where $J_f(x) = df/dx$. For the densities, we have $p_X(x) = p_Y(f(x))|J_f(x)|^{-1}$. **Mutual information, entropy** (beneficial, but not necessary): For a discrete random variable, $\mathbb{I}(X) = -\mathbb{E}_p(\log p(X)) = -\int \log p(x)dP(x)$. The concept can be extended for continuous random variables with technicalities, and is not properly defined for infinite-dimensional vectors (functions), but we will nevertheless appeal to the formal concept. Mutual information $\mathbb{I}(X,Y) = \mathbb{H}(X) + \mathbb{H}(Y) - \mathbb{H}(X,Y)$ with the latter the joint entropy of the random vector $(X,Y)$. Equivalently, $\mathbb{I}(X,Y) = \mathbb{H}(X) - \mathbb{H}(X|Y)$. **Random vectors** $X$, their realizations $x$, distribution $dP(X)$ and density $p(x)$. For a one-to-one mapping $f$, $J_f(x)$ is the (Jacobian) matrix of partial derivatives and $|J_f(x)|$ its determinant. The vectors could be infinite-dimensional, for instance functions on $[a,b]$. A **kernel density estimator** of $p(x)$ from a sample $x^n$ is a function of the form $\hat{p}(x) = \sum_{i=1}^{n} \kappa_{\sigma}(x - x_i)$ for some “kernel” function that satisfies certain conditions, and typically depends on at least one parameter, say $\sigma > 0$. A **sufficient statistic** (of $x$, for $y$), any statistic such that $p(x' | y) = p(\phi(x') | y)$. A
minimal sufficient statistic is one that is a function of all other sufficient statistics (so it is the smallest in the sense of inclusion of sigma algebras). A minimal sufficient statistic summarizes the sample with the smallest complexity. A group $G$ of transformations acts on $x$ via elements $g \in G$ in a way that composes (given $g_1, g_2 \in G$ there is a group element, called $g_2 \circ g_1 \in G$ such that $g_2(g_1(x)) = g_2 \circ g_1(x)$), has a null element, typically called $e \in G$, such that $e(x) = x$, and is invertible (there exists a group element, called $g^{-1}$, such that $g^{-1} \circ g = e$). We often write the composition simply as $g_2g_1 \equiv g_2 \circ g_1$ and the action as $gx \equiv g(x)$. The set $[x] = \{gx, g \in G\}$ is the orbit of $x$ under the action of the group. $G = SE(3)$ refers to the group of rigid motion in three-dimensional Euclidean space.

$G = SO(2)$ refers to the group of planar rotations, parametrized by an angle $\theta$ so that $R(\theta) = \left[ \begin{array}{cc} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{array} \right]$. An invariant statistic (of $x$ to $G$) is a statistic $\phi$ such that $\phi(gx) = \phi(x)$. A maximal invariant is an invariant that is equivalent to the entire orbit. Marginalization (of $x$ from $p(x, y)$) refers to the integration to eliminate $x$ to obtain the (marginal) density $p(y) = \int p(x, y) dP(x)$. The likelihood function (of a parameter $y$, given a sample $x$) is a function of the parameter $L(y) = p(x|y)$. The equality is formal, because the likelihood is not a conditional density (it is not a function of $x$, that is fixed), but is instead a function of the parameter $y$, that just happens to have the same expression of the conditional density of $x$ given $y$, except that $x$ is fixed.

The likelihood function is critical for hypothesis testing, in particular through likelihood ratios, and in composite hypothesis testing. Pooling (of $x$ in $G$) is a statistic of samples, for instance in a kernel density estimate, in a neighborhood of $G$. This can be represented by a “prior” $dP(g)$; for instance, pooling in a neighborhood of the origin $e$ of size $\epsilon > 0$, $B_\epsilon = \{g \in G \mid |g| \leq \epsilon\}$ is equivalent to assuming a uniform prior $dP(g)$ in $B_\epsilon$, and yields $\hat{p}(x) = \sum_{x_i \sim p(x), g_i \sim dP(g)} \kappa_\sigma(g_i x - x_i)$. A Gabor filter or wavelet is a family of functions indexed by an orientation and a scale parameter, in addition to an order parameters, often assumed to approximate the operation of early stages of visual processing in the primate neocortex. A representation (of a parameter $y$ given a sample $x$ subject to the action of a group $G$) is a minimal sufficient statistic of $x$ for $y$ that is a maximal invariant to $G$. The representation is complete if the invariant is complete. A digital image thought of as a random vector $y$ with components $y_i \in [0, 255]$ denoting the (gray-scale) intensity of a pixel at location $(u_i, v_i) \in \mathbb{R}^2$, so $y_i = x(u_i, v_i)$. The spatial gradient at the same location is $\nabla y_i = (\nabla u y_i, \nabla v y_i)$ where $\nabla u y_i = x(u_{i+1}, v_i) - x(u_i, v_i)$. The spatial gradient at the same location is $\nabla y_i = (\nabla u y_i, \nabla v y_i)$ where $\nabla u y_i = x(u_{i+1}, v_i) - x(u_i, v_i)$. The spatial gradient at the same location is $\nabla y_i = (\nabla u y_i, \nabla v y_i)$ where $\nabla u y_i = x(u_{i+1}, v_i) - x(u_i, v_i)$ in the simplest case, although more sophisticated and regularized gradient operators can be used. The vector of all gradients is indicated by $\nabla y$. An image descriptor is a just a statistic. An optimal descriptor for a task represented by $y$, given data $x$ subject to nuisance variability $G$, would be an ideal visual representation.

Proof of Theorem 1

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 $\nabla y = (\nabla u y, \nabla v y) \in \mathbb{S}^2 \times \mathbb{R}^+$, 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 from Cartesian $X = \nabla y$ to radial $Y \sim (\nabla u y, \nabla v y) = \phi(X)$, and conditioning on $Z = \nabla z$, we obtain

$$p_X|Z(x|z) = \rho^{a-1(Y)|Z(x|z)} = p_Y|Z(\phi(y)|z)|J_\phi(y)|^{-1} \quad (25)$$

9Recall that for a random variable $X$ with samples $x$ and density $p_X(x)$ and a function $\phi$ transforming it to another random variable $Y = \phi(X)$, correspondingly transforming the samples $y = \phi(x)$, with inverse function $\phi^{-1}(y) = \{x \mid \phi(x) = y\}$, and determinant of the Jacobian $J_\phi = \frac{\partial \phi}{\partial x}$ indicated by $|J_\phi|$, we have the change of measure formula

$$p_X(x) = p_{\phi^{-1}(Y)}(x) = p_Y(\phi(x))|J_\phi(x)|^{-1}. \quad (21)$$

9For a vector with Cartesian coordinates $v = (v_1, v_2)$, the change of coordinates to radial is a map

$$\phi(v) = \phi \left( \begin{array}{c} v_1 \\ v_2 \end{array} \right) = \left[ \begin{array}{c} \arctan \left( \frac{v_1}{\sqrt{v_1^2 + v_2^2}} \right) \\ \sqrt{v_1^2 + v_2^2} \end{array} \right] = \left[ \begin{array}{c} \theta \\ \rho \end{array} \right] \quad (22)$$

$$\phi^{-1} \left( \frac{\theta}{\rho} \right) = \left[ \begin{array}{c} \rho \cos \theta \\ \rho \sin \theta \end{array} \right] = \left[ \begin{array}{c} v_1 \\ v_2 \end{array} \right], \text{ and}$$

$$|J_\phi^{-1}| = \rho = \|v\| \quad (23)$$

10For a vector with Cartesian coordinates $v = (v_1, v_2)$, the change of coordinates to radial is a map

$$\phi(v) = \phi \left( \begin{array}{c} v_1 \\ v_2 \end{array} \right) = \left[ \begin{array}{c} \arctan \left( \frac{v_1}{\sqrt{v_1^2 + v_2^2}} \right) \\ \sqrt{v_1^2 + v_2^2} \end{array} \right] = \left[ \begin{array}{c} \theta \\ \rho \end{array} \right] \quad (22)$$

$$\phi^{-1} \left( \frac{\theta}{\rho} \right) = \left[ \begin{array}{c} \rho \cos \theta \\ \rho \sin \theta \end{array} \right] = \left[ \begin{array}{c} v_1 \\ v_2 \end{array} \right], \text{ and}$$

$$|J_\phi^{-1}| = \rho = \|v\| \quad (23)$$
Formally treating $Z$ as a random variable, using Bayes’ rule, and transforming it to $W = \phi(Z) = (\nabla \nabla z, \|\nabla z\|)$ we obtain

$$p_{X|Z}(x|z) = p_Y(W(\phi(y)|\phi(z)) | J_\phi(y) )^{-1}$$  \hfill (26)

which written explicitly as

$$p_{|\nabla Y|<\nabla Z} (\theta | \theta') \ p_{|\nabla Y|>\|\nabla Z\|} (\rho | \rho') | \rho$$  \hfill (27)

where $\theta$ is a realization of $\nabla \nabla y$ 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 $\rho \geq 0$:

$$\int p(\theta | \theta') p(\rho | \rho') |\rho| d\rho = p(\theta | \theta') \int p(\rho | \rho') |\rho| d\rho = p(\theta | \theta') |\rho|$$  \hfill (28)

which, in the less pedantic notation, reads $p_c(y|z) = p(\nabla \nabla y | \nabla z) |\nabla z|$ from which the statement (10) follows.

B ADDITIONAL MATERIAL

B.1 TESTING FOR CORRESPONDENCE

Given training data, to classify a test datum $y$ as coming from a class $k$ we must perform a composite hypothesis test, where the null hypothesis $H_0$ corresponds to the datum coming from class $z_k$ and $H_1$ corresponds to some other $z_{-k} = z_{i \neq k}$. Since the scenes $z_k$ are not known, we must compose the hypothesis with the estimation of the likelihoods based on the (labeled) training data $y_k$. The test is then based on the likelihood ratio(s)

$$L(y) = \frac{p(y|H_0)}{p(y|H_1)} = \frac{\max_{y_k} p(y_k, y|z_k)p(y_{-k}|z_{-k})}{\max_{y_k} p(y_k, y|z_k)p(y_{-k}|z_{-k})}$$

$$= \frac{p_k(y_k, y|z_k)p_{-k}(y_{-k}|z_{-k})}{p_k(y_k, y|z_k)p_{-k}(y_{-k}, y|z_{-k})}$$  \hfill (29)

Now, assuming that $t \gg 1$ and therefore the estimate including the test datum $y$ is similar to that excluding it, we have, after simplifications

$$L(y) \simeq \frac{p_k(y|z)}{p_{-k}(y|z)}$$  \hfill (30)

A simplified test can be arrived at by considering, among all indices $-k$, only the “second-best” (the one with the second-highest likelihood). This is similar, but not identical, to the ratio test for local descriptors in SIFT [Lowe 2004]. SIFT considers descriptors as “vectors”, and compares descriptors computed in a (test) image $p(|z = y)$ to each descriptor computed in another (training) image $p_k(|z)$, $k = 1, \ldots, K$, where each descriptor is its own class (trained with a single sample), calling $k_1 = \text{arg} \min_{k \in \{1, \ldots, K\}} d(p(|y), p(|z_k))$ and $k_2 = \text{arg} \min_{k \in \{1, \ldots, k_1 -1, k_1+1, \ldots, K\}} d(p(|y), p(|z_k))$, then SIFT computes

$$d(p(|y), p(|z_{k_1})) \quad d(p(|y), p(|z_{k_2}))$$  \hfill (31)

Instead, the interpretation above suggests forgoing the computation of the descriptor in the test image $p(|y)$, and instead simply evaluate the ratio

$$\frac{p(y|z_{k_1})}{p(y|z_{k_2})}$$  \hfill (32)

where now $k_1 = \text{arg} \max_{k \in \{1, \ldots, K\}} p(y|z_k)$ and $k_2 = \text{arg} \max_{k \in \{1, \ldots, k_1 -1, k_1+1, \ldots, K\}} p(y|z_k)$. Note that this does not require computing the descriptor in the test image.

B.2 ENTROPY AND SELECTION

Co-variant detection only serves to sample the nuisance group, but provides no guarantee that the descriptors constructed on the corresponding frames are informative. Co-variant detectors aim to select regions that are localizable, so as to provide hypotheses for correspondence, but the resulting descriptors may not be optimal in terms of information they provide on the scene. For the purpose of constructing a representation, the most informative descriptors
are those that minimize the uncertainty of future “data.” In reality, the goal is not to be able to predict the data itself, but instead a maximal invariant to nuisance factors Soatto (2010), which is informed by the task. Note that this is the flip-side of active learning and experiment design, where one wants to select actions (including the collection of data) that maximize the uncertainty of future data. Instead, it is akin to inferring the state of a system, which is the statistic (function of past data) that minimizes the uncertainty of the prediction of the next measurement (innovation). This is measured by

\[ H(y|z) = \mathbb{H}(p_G(y|z)) \]  

(33)

where \( p_G \) is a maximal \( G \)-invariant. This is easy to see if \( p(\cdot|z) \) is finitely-parametrized, for instance with parameters \( \hat{z} \). So, selection of the most informative descriptor can be done by selecting those that have the smallest entropy. This appears to go against the task of compression, where the most “valuable” data is the most uncertain, as well as information-based feature detectors Kadir & Brady (2003) that select regions with the highest entropy. Here, the most informative regions are isolated structures such as edges, corners etc., whereas regions with uniform gradient orientation distribution are discarded as uninformative. Thus, as an alternative to using a detector to sample locations and scales, and then computing descriptors in the corresponding frames, we can compute the descriptor densely, and then discard the least informative ones. For instance, using a single image, instead of using the SIFT detector (DoG) to define regions where to compute the SIFT descriptor (HOG), one can use the descriptor as a detector, to decide which samples to retain.

B.3 Sampling and Visualization

Once SIFT is interpreted as a conditional density, it can be used for synthesis, as described in Dong et al. (2013). However, note that unlike Vondrick et al. (2013); Simonyan et al. (2012), there is no unique synthesis. Instead, one could sample different images (up to contrast transformations, resampling, and local deformations), or generate the maximum likelihood sample.