Features for Recognition: Viewpoint Invariance for Non-Planar Scenes

Andrea Vedaldi and Stefano Soatto

University of California, Los Angeles
Computer Science Department
Technical Report # TR040049

November 29, 2004

Abstract

We present a technique for local image representation that is invariant to viewpoint for scenes with arbitrary non-planar shape. We show that generic viewpoint invariance can be achieved, under suitable conditions, although the resulting invariant is not shape-discriminative. Our results serve both to validate existing approaches to local feature detection and description, as well as to complement them where they are not applicable. We illustrate our approach on images of 3-D corners where existing approaches fail.

Contents

1 Introduction

Visual classification plays a key role in a number of applications and has received considerable attention in the community during the last decade. The fundamental question is easy to state, albeit harder to formalize analytically: when do two or more images "belong to the same class"? A class reflects some commonality among scenes being portrayed by the images in question [14, 17, 30]. Classes that contain only one element are often called "objects," in which case the only variability in the images is due to extrinsic factors – the imaging process – but there is no intrinsic variability in the scene. Extrinsic factors include illumination, viewpoint, and so-called clutter, or more generally visibility effects. Classification in this case corresponds to recognition of a particular scene (object) in two or more images. In this manuscript we restrict ourselves to object recognition. While this is considerably simpler than classification in the presence of intrinsic variability, there are some fundamental questions yet unanswered: What is the "best" representation for recognition? Is it possible to construct features that are viewpoint-invariant for scenes with arbitrary (non-planar) shape? If so, are these discriminative? Under what conditions can illumination invariance be achieved? In fact, do we even need a notion of "feature" to perform recognition? We wish to contribute to formalizing these questions, and where possible give precise answers. Our contributions are highlighted as (a)-(e) in Section 1.3.

1.1 Generalized correspondence

The simplest instance of our problem can be stated as follows: *When do two (or more) images portray (portions of) the same scene?* Naturally, in order to answer the question we need to specify what is an image, what is a scene, and how the two are related. We will make this precise later; for now, we just use a formal notation for the *image* I and the *scene* ξ. An image I is obtained from a scene ξ via a certain function(al) h, that also depends on certain *nuisances* ν of the image formation process, namely viewpoint, illumination, and visibility effects. With this notation we can formalize the question above: we say that two images are in *correspondence*[1] if there exists a scene that generates them

$$I_1 \leftrightarrow I_2 \;\Leftrightarrow\; \exists\, \xi \mid \begin{cases} I_1 = h(\xi, \nu_1) \\ I_2 = h(\xi, \nu_2) \end{cases} \tag{1}$$

for some nuisances ν_1, ν_2. *Matching*, or deciding whether two or more images are in correspondence, is equivalent to finding a scene ξ that generates them all, for some nuisances $\nu_i, i = 1, 2, \ldots$. These (viewpoint, illumination, occusions, cast shadows) could be *estimated explicitly* as part of the matching procedure, akin to "recognition by reconstruction," or they could be *factored out* in the *representation*, as in "recognition

[1]Note that there is no locality implied in this definition, so correspondence here should not be confused with point-correspondence.

using features" (see Appendix A for further details). But what is a *feature*? and why do we need it? We will address these questions in Section 2.2.

In the definition of correspondence the "=" sign may seem a bit strong, and it could certainly be relaxed by allowing a probabilistic notion of correspondence. However, even with such a strong requirement, it is trivial to show that any two images can be put in correspondence. For instance, choose the scene to be a mirror surface (e.g. a sphere), the viewpoint to be arbitrary, and the illumination be a larger sphere where each image is back-projected via the mirror surface. This example should convince the reader that the notion of correspondence is meaningless without additional knowledge on the scene and the nuisance. Such knowledge could come in *probabilistic* form, e.g. as a prior distribution on "likely" scenes and nuisances, or in *physical* form, e.g. via assumptions on reflectance and illumination. In the former case, assuming that we have prior distributions $dP(\nu), dP(\xi)$, we could tune up the definition of correspondence as a Bayesian decision.[2] While this is formally easy, in practice this is unfeasible because even for the simplest scenes we do not know how to endow the space of shapes, reflectance functions, illumination functions with a metric and a probabilistic structure, let alone actually computing the likelihood ratio. Therefore, we choose to make *physical assumptions*, all and only those that allow us to give a meaningful answer to the correspondence problem.

1.2 Lambertian scenes in ambient light

While global correspondence can be computed for scenes with complex reflectance, under suitable assumptions [26], local correspondence cannot be established in the strict sense defined by (1) unless the scene is Lambertian, and even then, it is necessary to make assumptions on illumination to guarantee uniqueness [7]. In particular, one can easily verify following [7] that if the illumination is assumed to be constant (ambient) then local correspondence can be established. We therefore adopt such assumptions and relegate all non-Lambertian effects to "noise."

We can now make the formal notation above more precise: We represent an image as an array of positive numbers; for simplicity we neglect quantization in the pixels and gray levels and represent images in a continuum: $I : D \subset \mathbb{R}^2 \longrightarrow \mathbb{R}_+;\ x \mapsto I(x)$. A Lambertian scene is represented by a collection of (piecewise smooth) surfaces embedded in $\mathbb{R}^3$, which we indicate collectively by $S \subset \mathbb{R}^3$, that support a positive-valued function $\rho : S \to \mathbb{R}_+$ with bounded variation, called *albedo*. So, the scene is described by $\xi = \{S, \rho\}$ where both shape and albedo are infinite-dimensional objects (functions).

The scene and the image are related by an image formation model. This requires specifying a *viewpoint*, i.e. a moving reference frame $g_t \in SE(3)$, where $SE(3)$ denotes a Euclidean reference frame (rotation and translation relative to a fixed reference frame), and an *illumination*. In case of ambient illumination, we have a linear scaling of the image α_t and an offset β_t, as one can easily verify. The overall model can thus be written as

$$\begin{cases} I_t(x_t) = \alpha_t \rho(p) + \beta_t + n_t(x) \\ x_t = \pi(g_t p), \quad p \in S \end{cases} \tag{2}$$

where $\pi : \mathbb{R}^3 \to \mathbb{R}^2$ is the perspective projection and n_t is a "noise" term that includes all the nuisances that are not explicitly modeled. The nuisance proper here is limited to viewpoint and illumination, $\nu = \{g_t, \alpha_t, \beta_t\}$. We have so far neglected visibility effects (occlusions and cast shadows), but we will come back to it. Note that the equation above is reminiscent of deformable templates [50, 9, 18], although here we do not know the templates. Correspondence is also naturally related to wide-baseline matching [43, 15, 15, 10, 28]. From now on we will restrict our attention to the model (2).

1.3 State of the art and our contributions

One of the questions addressed in this manuscript is whether it is possible to construct viewpoint and illumination invariants, and whether such invariants are discriminative. Belhumeur and coworkers [7] showed that even for Lambertian scenes there exist no discriminative illumination invariant. We show that, if one

[2] The likelihood ratio is $L(I_1, I_2) = p(I_1, I_2|H_0)/p(I_1, I_2|H_1) \geq \tau$ where H_0 is the null hypothesis (correspondence) and H_1 is the alternative (different scenes); τ is a threshold that depends on the cost of wrong decisions and priors on each hypothesis. The component densities can be written as $p(I_1, I_2|H_j) = \int p(I_1 - h(\xi_1, \nu_1))p(I_2 - h(\xi_2, \nu_2))dP(\xi_1, \xi_2|H_j)dP(\nu_1)dP(\nu_2)$, assuming independent nuisances.

assume ambient illumination, then **(a)** *illumination invariants can be constructed,* and indeed this is often done in practical (local) correspondence algorithms (Section 2.3).

On viewpoint invariance, Burns et al. [5] showed that there do not exist generic viewpoint invariants. This statement, however, is misleading, since it refers to collections of points in space, with no photometric signature associated to them. We show that **(b)** *viewpoint invariance can be achieved for scenes with arbitrary shape, regardless of their albedo,* under suitable conditions (Section 2.1). As a corollary, however, we show that **(c)** *any viewpoint invariant necessarily "kills" shape information,* and therefore discrimination has to occur based solely on the photometric signature (Section 2.1). In deriving our results (a)-(c), we will lay out a general framework for designing detector/descriptor pairs that **(d)** allows *comparison of existing algorithms on analytical grounds,* in addition to experimental as done in [37]. Finally, we illustrate our theory on a test case, by introducing **(e)** *a 3-D corner detector/descriptor,* and test its performance on scenes where existing approaches fail (Section 3). So, our work serves to validate existing methods where appropriate, and to complement them where their applicability is limited.

The topic of this manuscript relates to a vast body of work in low-level image representation, recognition, wide-baseline matching, segmentation. We will therefore point out relationship throughout the manuscript.

2 Recognition using features

We define a *feature* to be any image statistic, that is a known vector-valued function(al) of the image: $\phi(I) \in \mathbb{R}^k$. In particular, the image itself is a (trivial) feature, and so is the function $\phi(I) = 0 \; \forall \; I$. A feature $\phi(I) = \psi(\{I(x), \; x \in \Omega \subset D\})$ where D is the domain of the image, is called a *local feature*. Obviously, of all features, we are interested in those that facilitate correspondence between two images I_1, I_2, or equivalently recognition of the scene ξ. This requires handling the nuisance ν, either in the correspondence process (expensive) or by designing features that are invariant with respect to the nuisance. A feature is *invariant*[3] if its value does not depend on the nuisance: $\phi(I) = \phi \circ h(\xi, \nu) = \phi \circ h(\xi, \mu) \; \forall \; \nu, \mu$.

As we have mentioned, $\phi(I) = 0 \; \forall \; I$ is a feature, and indeed it is an invariant one. Alas, it is not very helpful in the correspondence process. Therefore, one can introduce the notion of *discriminative feature* when two different scenes yield different statistics:[4] $\xi_1 \neq \xi_2 \Rightarrow \phi \circ h(\xi_1, \mu) \neq \phi \circ h(\xi_2, \nu) \; \forall \; \mu, \nu$. In particular, we say that a feature is *shape-discriminant* if scenes with different shape (but possibly identical albedo) result in different statistics, and similarly for *albedo-discriminant*.

2.1 Viewpoint invariant features

In the image formation model (2) the nuisance ν comprises viewpoint and illumination. Because of the assumptions of Lambertian reflection and ambient illumination, a change in the illumination results in an affine transformation of the albedo. We will discuss in Section 2.3 how to achieve illumination invariance by normalization. For now, we assume that this has been done, and therefore let $\alpha_t = 1$ and $\beta_t = 0$ in (2). In general, invariance to viewpoint and illumination will have to be determined simultaneously.

If we neglect self-occlusion (see below on this issue), we can parametrize the surface S as $\Omega \subset \mathbb{R}^2 \rightarrow \mathbb{R}^3, \; x \mapsto S(x)$ for some choice of local coordinates, for instance $x = \pi(p)$, the perspective projection of $p \in S$ onto the image plane from the viewpoint $g = Id$ (the group identity, see eq. (2)). Since both ρ and S are unknown, and we only measure their composition through $I_t(x_t) = \rho \circ S(x)$, with an abuse of notation we can rename the function $\rho \doteq \rho \circ S$. Similarly, we call $w_t \doteq \pi \circ g_t \circ S : \Omega \subset \mathbb{R}^2 \rightarrow \mathbb{R}^2$ the function that maps the point x to the point x_t. This yields the following simplified model:

$$\begin{cases} I(x_t) = \rho(x), & x \in \Omega \\ x_t = w_t(x). \end{cases} \tag{3}$$

We have dropped the generic "noise" term n_t since that will only affect the inference technique, not the general modeling paradigm and invariance considerations. Excluding self-occlusion and the other visiblity

[3]The "=" sign can be relaxed to yield a probabilistic notion of invariance (insensitivity relative to the nuisance distribution). However, since in general we do not have manageable ways to define distributions on the spaces of nuisances, this pursuit is well beyond our scope.

[4]This definition can be relaxed as $\exists \; \xi_1 \neq \xi_2 \mid \phi \circ h(\xi_1, \mu) \neq \phi \circ h(\xi_2, \nu)$ as proposed in [7].

effect, w_t is an homeomorphism. Since homeomorphisms form a transformation group, they induce a partition of the set of images $\{I(x)\}$ in equivalence classes. Any function that maps $I(x)$ to a unique representative $\hat{I}(x)$ of its equivalence class $[I(x)]$ provides a viewpoint invariant. Moreover, since the resulting invariants are in one-to-one correspondence with the equivalence classes $[I(x)]$, any other invariant can be expressed as a function of them. Hence, we can state the following result, that clarifies a fact that is implicitly exploited by many existing work in the literature, specifically [36, 27, 35, 46] for affine invariance, and [2, 39, 15] for more general transformations:

Theorem 1 (Viewpoint invariants exist ...) *Given an image I of a Lambertian scene ξ with continuous (not necessarily smooth) surfaces with no self-occlusions, viewed under ambient light, there always exist non-trivial viewpoint invariants.*

This result is at the base of our approach to design 3-D viewpoint-invariant features, and its application will be illustrated in steps.

Unfortunately, invariant features can never be fully "discriminative". Indeed, let $\phi(I)$ be any viewpoint invariant statistic of the image $I(x)$. Any image $I_1(x_1)$ can be obtained by the image formation model (2) from an object that is either a plane S_a or a curved surface S_b. At the same time, by fixing one surface and changing the viewpoint, we get a (generally) different image $I_2(x_2)$. Since ϕ is viewpoint invariant, we must have $\phi(I_1) = \phi(I_2)$. Therefore I_1 and I_2 are two different images obtained from two different scenes S_a and S_b that have the same value of the feature. Formally

Corollary 1 (... but are not shape-discriminant) *If ϕ is a viewpoint-invariant feature, then for any scene ξ yielding an image I there exists a scene ξ' with different shape yielding a different image I' such that $\phi(I) = \phi(I')$.*

This does not mean that an invariant feature is useless! The albedo "information" is still present, warped together with shape information, in ρ and Ω.

Visibility and local features. The technical assumption to prove Theorem 1 requires the domain deformation w_t to be invertible, which in turn implies that there are no visibility effects such as self-occlusion or clutter. Clutter is an "adversarial" nuisance (one can always make object A look like object B by placing object B in front of it), and no analytical results can be proven that will guarantee (worst-case) invariance to generic clutter. Therefore, we can relax the notion of correspondence by requiring that a given scene ξ generates *at least a (non-empty) subset* of each image I_1, I_2. That is, with an abuse of notation: $I_1 \leftrightarrow I_2 \;\Leftrightarrow\; \exists\, \Omega \subset D, \xi \mid \forall\, x \in \Omega : I_1(x) = h(\xi(x), \nu_1),\; I_2(x) = h(\xi(x), \nu_2)$.

This brings us to the notion of *local feature* which is what we will use from now on. The extent of the domain Ω depends on the visibility boundaries and will be determined by a *detector*, which is itself a feature (i.e. a function of the image), as we discuss in Section 2.3.

2.2 Why features?

Before we marry to the notion of feature it is useful to pause: In fact, Rao-Blackwell's theorem ([47], page 87), adapted to our context, claims that there is no advantage in using features, as opposed to using the entire data I_1, I_2. That is, unless we could eliminate the nuisance ν without "throwing away information" on the scene ξ.[5] Unfortunately, Corollary 1 says that this is not possible: *in order to achieve viewpoint invariance, shape information has to be sacrificed.* In light of this result, then, *does it still make sense to use features?*

Posing the correspondence problem as an optimal decision requires marginalizing nuisances, that are infinite-dimensional unknowns living in spaces that are not easily endowed with a metric (let alone probabilistic) structure. Therefore, unless we are willing to perform recognition by reconstructing the entire observable component of the scene and its nuisances, the use of invariant statistics seems to be the only computationally viable option. However, by choosing a viewpoint invariant we are agreeing to give up some discriminative power, and therefore accept some degradation of recognition performance relative to the optimal (Bayes) risk.

[5] "Throwing away information" in this context means lowering the Bayesian risk associated with the decision task of correspondence. A feature that maintains the Bayesian risk unaltered would be a sufficient statistic (with respect to the correspondence decision) for the scene ξ.

2.3 Invariance by canonization

It is immediate to see from (3) that $\{\rho(x),\ x \in \Omega\}$ is the "maximal" invariant feature, in the sense that any other invariant feature is a function of it (it only depends on the scene, and has no t subscript). Of course, we do not know ρ nor Ω. So, we start by expressing what we have in terms of what we want: $I_t(x_t) = \rho(w_t^{-1}(x_t)),\ x_t \in w_t(\Omega) \subset D$. Unfortunately, if we take any homeomorphism $v : \mathbb{R}^2 \to \mathbb{R}^2$ and replace $\rho(\cdot)$ with $\tilde{\rho}(\cdot) \doteq \rho \circ v(\cdot)$, $w_t(\cdot)$ with $\tilde{w}_t(\cdot) \doteq w_t \circ v(\cdot)$, and Ω with $\tilde{\Omega} \doteq v^{-1}(\Omega)$, we obtain the same images, and therefore we cannot distinguish $\{\rho(\cdot), \Omega\}$ from $\{\rho(v(\cdot)),\ v^{-1}(\Omega)\}$. In other words, what we can recover from $I_t(x_t),\ x_t \in D$ is *not* the invariant feature $\phi \doteq \{\rho(x),\ x \in \Omega\}$, but an entire *equivalence class* of invariant features: $[\phi] \doteq \{\rho(v(x)),\ x \in v^{-1}(\Omega),\ v : \mathbb{R}^2 \to \mathbb{R}^2 \text{ a homeo}\}$. Although we may still use the feature for classification by defining a distance $d([\phi_1], [\phi_2])$ between equivalence classes, this would be not easier than estimating the nuisance in the first place. The alternative is to identify, for each equivalence class, a *canonical representative*, that is a unique element of the class, $\hat{\phi}$, and then define a distance between feature elements, $d(\hat{\phi}_1, \hat{\phi}_2)$.

Feature detectors. A choice of the element $\hat{\phi}$ in the equivalence class $[\phi]$ must be made from the available data, that is $I_t(x_t),\ x_t \in D$. In other words, given an equivalence class $[\phi]$, we are looking for a canonical representative $\hat{\phi} = \{\hat{\rho}(x), x \in \hat{\Omega}\}$ where $\hat{\rho} \doteq \rho(\hat{v}),\ \hat{\Omega} \subset \hat{v}^{-1}(\Omega)$ for some diffeomorphism $\hat{v}$. More formally, from the pre-image theorem, we are looking for contra-variant functionals $F_i,\ i = 1, 2, \ldots$, such that $F_i([\phi]) = F_i(x, v, \Omega) = e_i$ uniquely determines $\hat{v}$, and therefore $\hat{\phi}$. Without loss of generality we can choose $e_i = 0$, since whatever value can be incorporated into the definition of F_i. Furthermore, in the presence of uncertainty, rather than looking for $\hat{\phi} \mid F_i(\hat{\phi}) = 0$, we can look for

$$\hat{\phi} \doteq \arg\min_{\phi} \|F_i(\phi)\| \tag{4}$$

for some choice of norm. For a suitable choice of the functionals F_i, one obtains all the existing methods (e.g. Harris [20] 2-D points, DoG [33] and Harris-Laplace [36] 3-D points, second order moments [32, 36], edge/intesity [15], saliency [27], level set [35] based affine regions, affine omogeneous-texture regions [46]).

Feature descriptors. Once $\hat{v}$ and $\hat{\Omega}$ have been determined, the statistic $I_t(v^{-1}(x)), x \in \hat{\Omega}$ becomes available. This is invariant by construction, and we therefore call it, or any deterministic function of it, *invariant descriptor*. The result indicates that the local structure of the image around a point can be used to determine a local "natural" frame. We discuss the consequences of this in Section 2.4.

Once detectors/descriptors have been obtained, matching can be based on just comparing the descriptors (since the domains have been normalized), or comparing the domains as well, for instance by quantifying the energy necessary to register them. A combination of the two can also be implemented [38, 16]. As we have pointed out, fixing w via equations (4) eliminates the dependency on g (the nuisance), but also eliminates the dependency on S, which is part of the description of the scene. Nevertheless, the resulting residual above depends on ρ, part of the descriptor. Also note that both w and Ω are unknown.

Now, suppose that the image I does *not* allow full inference of w via (4), for instance because it does not contain enough structure (e.g. local extrema) to provide a sufficient number of constraints. This means that, once the available constraints on w have been enforced via (4), the "residual" is already, by construction, invariant to w, and therefore g (and S). In the extreme case where I does not allow to infer any part of w, for instance when I or its statistics are constant, I is already a "descriptor" in the sense that it is invariant with respect to g.

Illumination invariance. Introducing illumination into the model does not modify the scheme just outlined for the simple case of ambient illumination and Lambertian scene. In fact, this case corresponds to an affine transformation of the range of the image, which simply enriches the equivalence class $[\phi]$. Normalization is trivial for the illumination parameters, since we can choose $\hat{\beta}_t = \int_{\hat{\Omega}} \hat{\rho}(x)dx$ and $\hat{\alpha}_t = \text{std}(\{\hat{\rho}(x)\ x \in \hat{\Omega}\})$. Naturally inference of the canonical elements (detection) has to be performed simultaneously with respect to all free parameters, which only increases the computational complexity, but not the conceptual derivation of the invariant.

2.4 Hierarchy of detectors/descriptors

The previous section established that the local structure of the image can be used to fix a "natural" frame via a detector. Any statistic of the image normalized relative to the natural frame is a descriptor. Depending on the choice of group structure w_t, that can range from simple planar translation (which we represent with a vector $T \in \mathbb{R}^2$) to infinite-dimensional groups of diffeomorphisms, one can adopt different normalization techniques.

In particular, corresponding to a geometric stratification of the group structure, one can build a hierarchy of detector/descriptor pairs. **Translation invariance** ($\mathbb{R}^2$) requires fixing a point $T \in \mathbb{R}^2$ and assigning its coordinates to $[0,\ 0]^T$ in the local frame. This is very common, and can be fixed using Harris' corner detector [20]. **Translation and scale invariance** ($\mathbb{R}^3$) requires a point with a scale associated with it, or a rotationally-symmetric intensity profile (e.g. a "blob"). The point is assigned to coordinates $[0,\ 0]^T$ in the local frame, and the unit is assigned isotropically to 1 in the two coordinate axes. This can be done by the Harris-Laplace detector [36], or in several other ways eloquently illustrated by Lindeberg [31]. **Similarity invariance** ($SE(2) \times \mathbb{R}$) requires a point, a scale and one direction. The point is assigned to the origin, the scale to the unit, and the direction to one of the coordinate axes. See for instance [33]. **Affine invariance** ($A(2)$) is well-known, and widely used. An affine invariant detector requires at least 3 points, or one point and two scaled directions. Schmid and coworkers have compared a number of affine detectors in their recent work [37]. Any affine-warped statistic, or "co-variant region," is an affine-invariant descriptor. **Viewpoint invariance, planar shape** ($H(2)$): planar projective transformations require 4 points. They are usually approximated by affine transformations because the added complexity only provides diminishing return. **Viewpoint invariance, generic shape** ($SE(3)$): here the map w is infinite-dimensional, due to its dependency on S. The homeomorphism w can be inferred only to the extent where the albedo ρ exhibits a sufficient degree of variability. We approximate w with a finite-dimensional map, and use a bank of local filters to determine a number of position, orientation and scale constraints. A variety of models can be used for the purpose, ranging from thin-plate splines [3, 2] (not a group, however) to polynomials (not recommended) for the representation of the warp w, and from curvelets [6] and local histogram (e.g. polar orientation histograms) to semi-global representations such as the sketch [11] for the analysis of the image. In Section 3 we illustrate this case with a piecewise affine deformation model.

All the detectors based on the invariance properties just outlined allow one to determine a *localized frame*, called a co-variant local frame,[6] that has a well-defined origin, hence the early nomenclature "feature point" even though a region Ω is used to determine the frame. However, often an image region Ω contains structure that is not localized or is repeated regularly. In other words, the frame associated to a certain point is only determined *up to a subgroup* which could be either continuous (e.g. the one-dimensional translational group for the case of an edge) or discrete (e.g. the repetition period for regular textures). In this case, one can associate the descriptor to any point along the equivalence class determined by the subgroup ambiguity. We distinguish the following frames: **edge in space** ($SE(3)$), fixed by an edge with the associated scale; **edge on the image** ($SE(2)/\mathbb{R}$), as a special case of the former when it is not possible to reliably associate a scale to the edge; **homogeneous periodic texture** ($SE(3)/\mathbb{Z}^2$) when the intensity profile is periodic with identical period along two spatial dimensions (possibly after warping or normalization). The construction can be extended for many special cases.

When we do not have a localized frame, the result of the detector is a warped image patch that contains an intensity profile with symmetries. Any statistic computed from such a profile is a valid descriptor . The descriptor does not contain any information on the geometry of the scene, like in the case of the localized frame, but, unlike that case, neither does the detector. Therefore, in such a case one can extend the region Ω to include all points that admit the same descriptor, i.e. the detector becomes a *segmentation procedure*.

Finally, one can integrate local descriptors in a global model by enforcing geometric information [30] (epipolar geometry, or shape statistics), or topological information [14, 17] (graphs of local descriptors).

3 Case study: 3-D corner

As an application of our theory, we develop a descriptor of a new kind of invariant features, corresponding to "3-D corners". We design a simple detector that select points where a suitable frame of reference can be

[6] Even though contra-variant would be a more appropriate name (Section 2.3).

easily attached. In particular, we focus on points x_0 that are projections of corners of the surface $S(x)$. A corner is a singular point of the object surface and cannot be approximated by a plane. Thus our descriptor works exactly under the conditions not supported by existing approaches [36, 15] (Figure 1).

3.1 Deformation under viewpoint change

We model a corner as a vertex with n planar faces. Barring occlusions, its image consists of n angular sectors, projections of the n faces, and a center x_0, projection of the vertex. These sectors are separated by edges, which we represent as vectors $v_i \in \mathbb{R}^2,\ i = 1, ..., n$. The length of the vectors will be used as a scale parameter.

When the viewpoint changes, the n faces of the corner are transformed by homographies, which we approximate by affine warps. This model locally captures the true transformation to an arbitrary degree of precision, which cannot be done by a single affine transformation as current approaches do. Since the corner surface is continuous, in the absence of occlusions so is the overall transformation. Thus, the n affine transformations are not independent and are fully specified by the mapping $x_0 \mapsto y_0$ of the center and the mappings $v_i \mapsto u_i,\ i = 1, ..., n$ of the edges (with their scales). Formally, let $\{\chi_i(x),\ i = 1, ..., n\}$ be a partition of $\mathbb{R}^2$ in n angular sectors, being $\chi_i(x)$ the indicator function of the i-th sector. We call *piecewise affine transformation* of degree n a function $w : \mathbb{R}^2 \to \mathbb{R}^2$ given by $w(x) = \sum_{i=1}^{n} \chi_i(x) A_i(x - x_0) + y_0,\ x \in \mathbb{R}^2$ where the matrices $A_i \in GL(2),\ i = 1, ..., n$ are chosen so that $w(x)$ is continuous.

Since the deformation of a corner under a viewpoint change is (locally) a PWA, PWAs are the minimal class of transformations with respect to which the feature has to be invariant, even though any more general class of transformations would fit. In particular, a PWA of degree m can be used for a corner that has $n < m$ physical edges, as long as the transformation is estimated consistently.

3.2 Feature detection

The detection process searches for corner structures in the image and attaches a reference frame to them. While there exist many possible procedures for detecting corners, including sketch primitives [11] or matched filters [19], our emphasis here is not in proposing yet another detector, but rather in how to arrive at a viewpoint invariant once a structure has been detected. Therefore, we choose a simple if not somewhat naive detector, designed to provide directly the structures that we need.

The procedure is articulated as follows. Initally, a set of Harris points [20] $X = \{x_1, ..., x_n\}$ is extracted. These points are used as candidate corners and as evidence for edge-like structures in the image (we use the fact that some Harris points are located along edges, particularly nearby the edge terminations). The algorithm checks for each pair $(x_i, x_j) \in X^2$ whether the image portrays an edge connecting x_i to x_j. Edges are modeled using the parametric template

$$T(x, y; w) = \text{sign}(y), \quad (x, y) \in [0, 1] \times [-w, w] \tag{5}$$

reminds what done in [1]. The template is matched[7] to the image by normalized cross correlation (NCC). Once the set E of edges has been extracted, the procedure attaches a reference frame to each point $x_0 \in X$. All edges connected to x_0 are considered: first the localization of each edge is refined using the model (5); then edges with the same orientation are clustered, because they relate to the same image structure; finally the edge that best covers the full extension of the underlying image structure is selected within each cluster. The selection uses an extension of the model (5) which represents explicitly the edge termination.

3.3 Feature canonization

Once a reference frame has been detected, we map it to a canonical configuration. In order to avoid singular configurations, we enforce the following conditions: (i) if all sectors are less than π radians wide, the normalized frame has n equally wide sectors; (ii) if one of the sectors is wider than π radians, we make this sector $3\pi/4$ radians wide and we fit evenly the others in the remaining $\pi/2$ radians[8]; (iii) if one sector is exactly

[7] There exists some simple yet effective heuristics that one can use to pre-prune the set of candidate edges and speed-up significantly the algorithm.

[8] We do this because no PWA (nor viewpoint) transformation can make the wide sector smaller than π radians

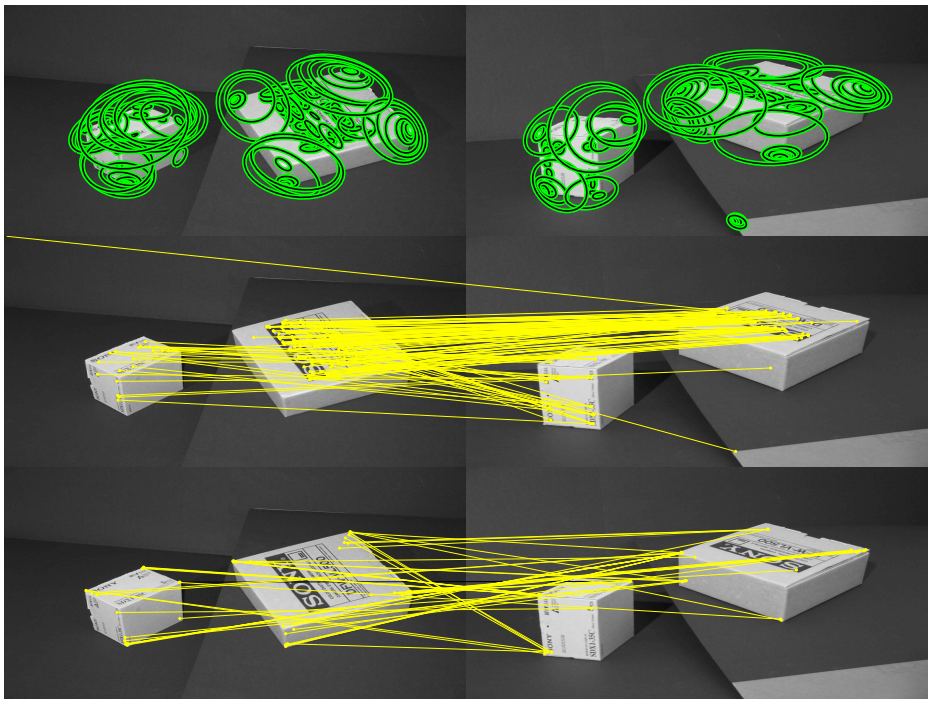

Figure 1: **Affine-invariant descriptors fail to capture non-planar structures:** *(top) two images of the same scene with detected regions; (middle and bottom) correspondence established using affine invariant signatures respectively for planar (middle) and non planar (bottom) regions. Several non planar regions are detected by the low-level detector, but are not matched because of the large discrepancy in the corresponding descriptor, caused by the non-planar structure of the scene.*

π radians wide (T-junction), we delete one edge and we reduce to the former case[9]. Note that at least two edges are required to compute the PWA transformation. If a point has less than two edges attached to it, it is discarded.

These rules fix the canonical reference frame up to a rotation. The rotation can be partially eliminated by requiring that one edge maps to $(1, 0)$. However, any edge could do, and we are left with a discrete subgroup of rotations to choose from. If the corner has a sector wider than π radians, we use this to uniquely identify an edge and eliminate the ambiguity. This is possible because there is at most one such sector and the property is preserved under viewpoint changes. If all sectors are narrower than π radians, we use the sector with maximal mean albedo as reference.

3.4 Feature description

Although the canonized features could be compared directly (e.g. by NCC), we compute a descriptor for each detected feature. This has two advantages: (1) makes the comparison much faster and (2) may absorb differences in the normalized features due to imprecise detections or unsatisfied assumptions (e.g. the surface is not Lambertian). Furthermore, most descriptors are insensitive to affine transformations of the albedo, so that we do not need to normalize explicitly the illumination. In the experiment we use the SIFT descriptor [33], one of the most widely used [33, 37]. We note however how this descriptor may not be as effective in our case as is for other kind of features. Indeed our canonized corners have strong oriented structures (the edges) in fixed position. This makes the SIFT descriptor (which is based on the gradient distribution) less discriminative.

Unilateral feature descriptors. The detector/descriptor works well under the assumptions we made. However, we wish to relax the hypotheses that the whole corner image is the projection of a single object.

[9]We do this because no PWA (nor viewpoint) transformation can change the π radians wide angle.

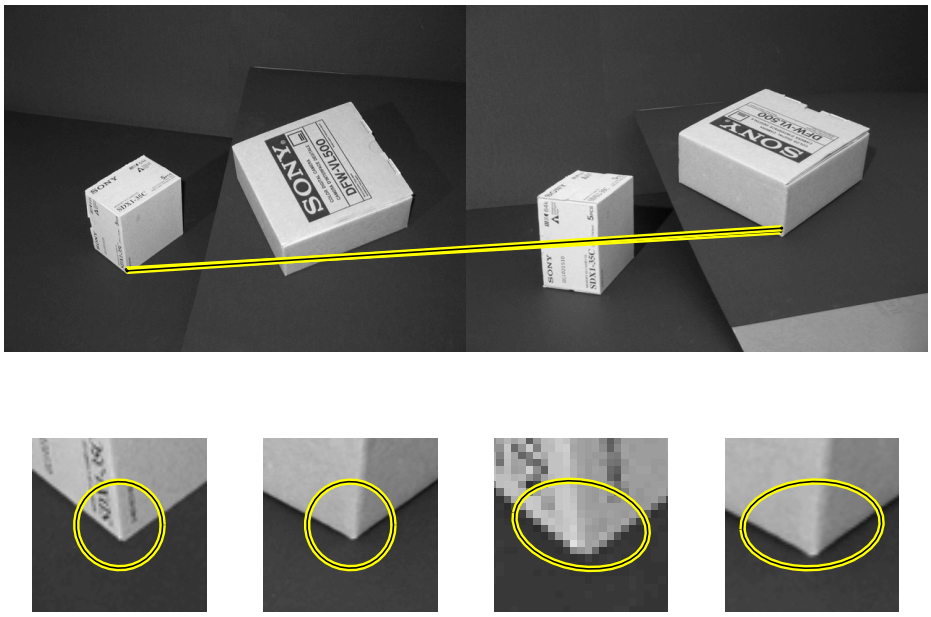

Figure 2: **Example of mismatched corner**. *By changing viewpoint we make the pose of a corner of one box similar to the pose of a corner of the other box. This confuses the affine descriptors.*

In fact, many corners are found on the boundaries of objects [48], and some sectors χ_i of the corner image may belong to the background. Clearly, we are not supposed to incorporate the background into the feature if we want to preserve invariance. We solve the problem by computing multiple descriptors for each possible assignment of the faces to the foreground or the background. In practice, the most common cases (objects with convex corners) are covered if we do so only for sectors larger than π radians, thereby obtaining no more than two descriptors for each detected feature.

3.5 Experiments

In the first experiment we explore the domain of applicability of our technique, by showing that it operates when established detector/descriptor techniques fail. To illustrate our point, we have purposefully chosen a simple scene (Figure 1 and 2): even on such scenes, most of the current affine-invariant methods fail to establish correspondence. Also, note that our goal is not to compare our method with existing affine-invariant schemes, since our method works on top of them. Since, to the best of our knowledge, nobody has presented viewpoint invariant schemes for non-planar scenes, we cannot do direct comparisons with any existing scheme.

As a typical representative [37] of the class of affine invariant detectors, we selected the Harris-Affine detector [36]. Figure 1 shows that most of the non-planar detected regions are incorrectly matched using the affine descriptor: of 186 features detected in the first image, 53 are successfully matched, 68 are mis-matched because the descriptor variability and 65 are not matched because the low-level detector fails to select the corresponding region. In contrast, Figure 3 and 4 show the performance of our method on the corners of the same images: almost all 3-D corners are matched correctly. There is just one mismatch, due to the almost identical appearance of the exchanged features (last two feature pair in Figure 3), and two missing corners, which are not extracted by the Harris detector in the very first stage of the algorithm. An exact comparison with the affine-invariant detector is difficult because the latter finds several times the same structures; roughly speaking, however, 70% of the mismatches (due to missing features or discrepancy of the descriptor) of the affine detector are fixed by the "3-D corner" model. As an additional advantage, our method extracts just one feature for each 3-D structure, while the Harris-Affine detector generates many duplicate detections of these structures.

In the second experiment we test our method on a more complex scene, made of various objects presenting a variety of 3-D corners. Figure 5 shows the detected reference frames and the matching pairs. One third

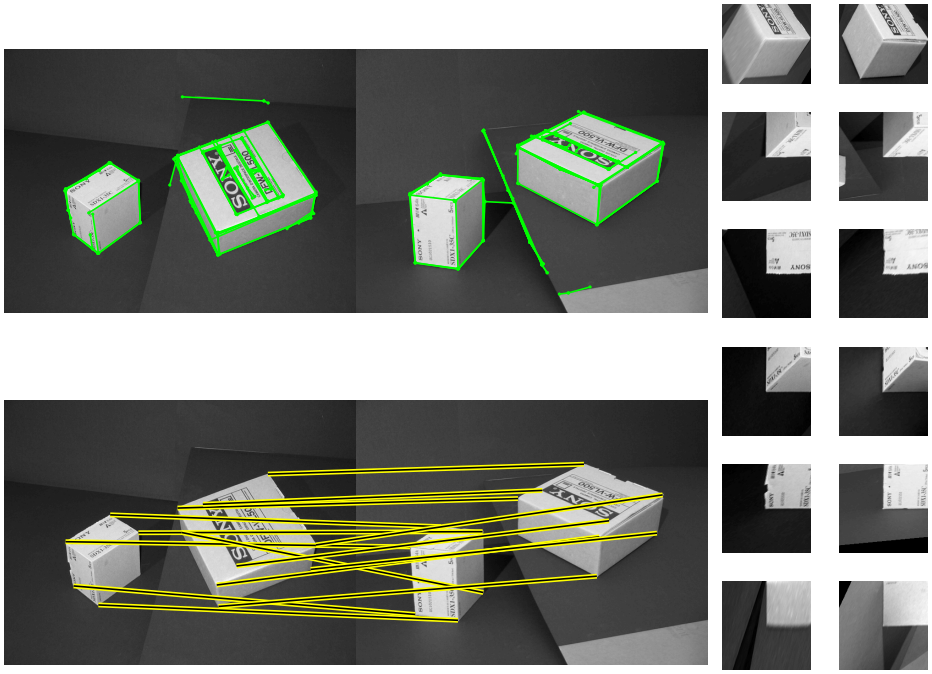

Figure 3: **General viewpoint invariants can match 3-D corners:** *(top) detected reference frames; (bottom) matched "3-D features"; (right) examples of canonized features. Most of the "3-D features" that are detected but mismatched using an affine-invariant descriptor are correctly matched using a more general viewpoint-invariant model, in this case a "3-D corner."*

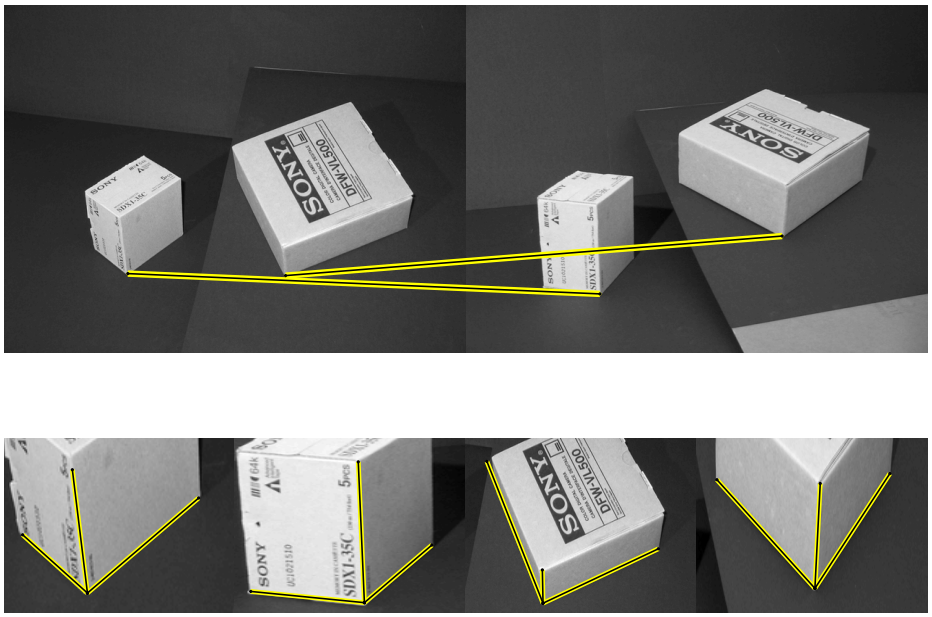

Figure 4: **A corner matched by the 3-D descriptor but not by the affine one**

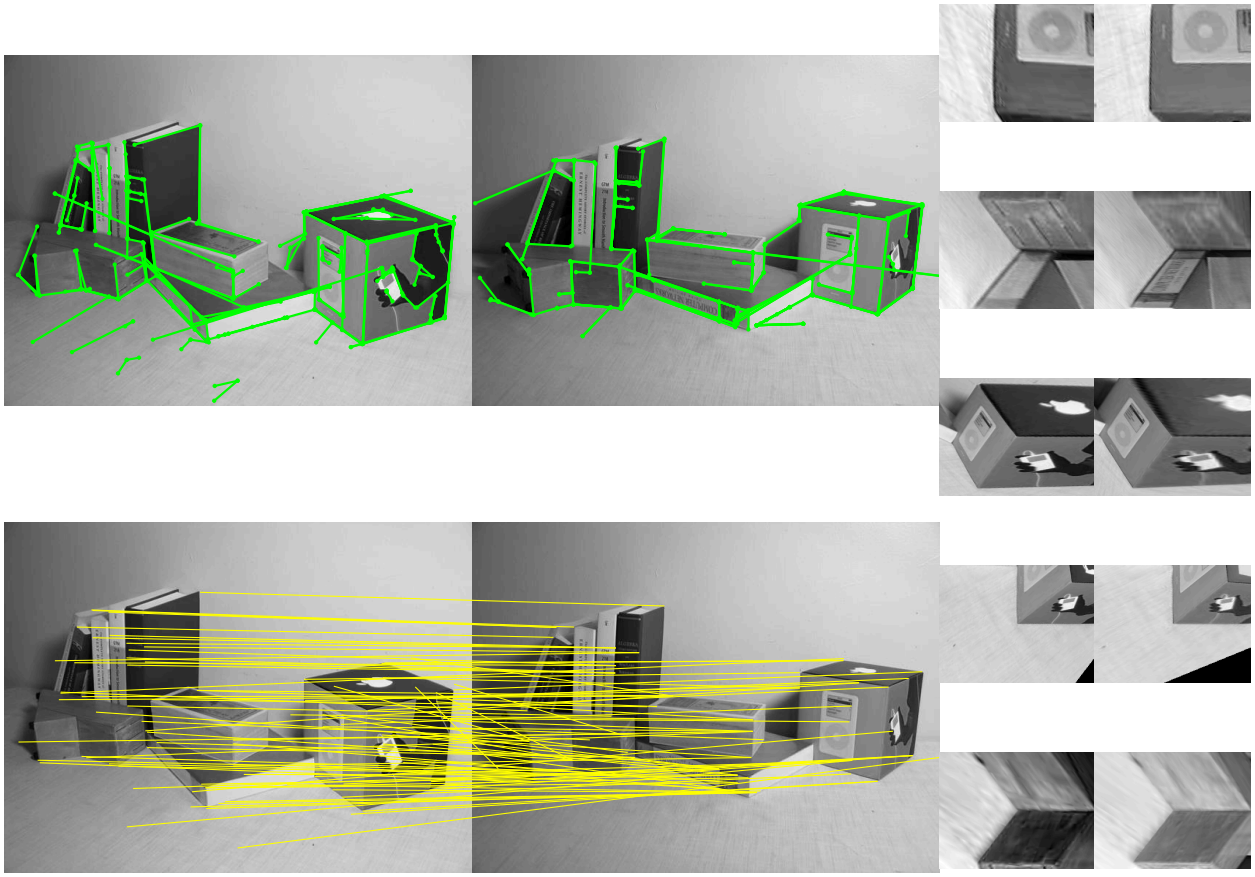

Figure 5: **Matching example:** *(top) all the features detected in the first image are connected to their nearest neighbors in the second image; (bottom) all the features detected on the first image are connected to their nearest neighbors in the second image and (right) a variety of normalized features. Of 93 detected features, 32 are present and correctly matched in the second image.*

of the detected features in the first image are correctly matched to the corresponding features in the second. Therefore, the performance is similar to that of the Harris-Affine detector on the planar structures of Figure 1, but in our case for non-planar structures. Some feature pairs are shown as well: they illustrate the most typical canonical configurations.

In the last experiment (Figure 6) we test our method on a more challenging scene, where several of our working hypotheses are not verified. We match two images of an highly non-planar, non-Lambertian scene. Not only the scene contains many 3-D corners, but these have non planar faces as well. Moreover, the two images are at two significantly different scales. The figure shows two corners that our method is able to match nevertheless, together with the corresponding canonized features. Note that the features are quite different, because of both the reflections and the non planarity of the corner faces. Still, these canonized features are similar enough to be matched using the SIFT descriptor, illustrating the importance of viewpoint canonization. As a further example of this fact and of the generality of our framework, we show the same two corners normalized using a thin-plate spline deformation, estimated by tracking and rectifying the edges. The matching distances are slightly smaller ($0.28 \mapsto 0.15$ and $0.4 \mapsto 0.36$ respectively) using this deformation as we compensate for the curvature of the edges.

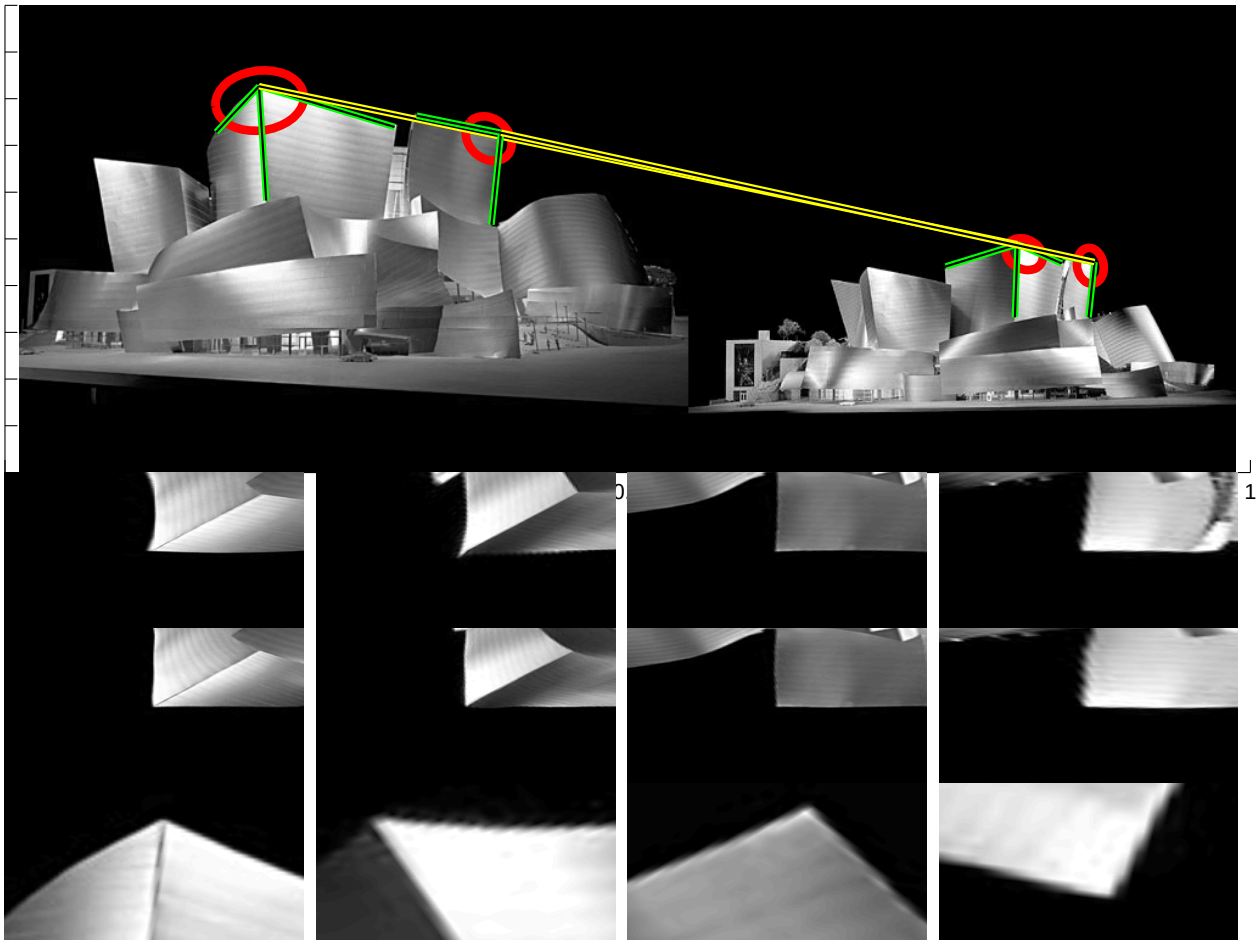

Figure 6: **Matching a challenging scene:** *two corners matched by our method; (top row) features canonized by a piecewise-affine transformation; (middle row) features canonized by a thin-plate spline transformation; (bottom row) features canonized by the Harris-Affine detector. Concert Hall, like all recent Gehry building, is challenging because severely non-Lambertian, and non-planar. Although the scene does not meet most of our working assmumptions, a few corners are still matched. The affine descriptor is shown to capture (up to a rotation) the deformation of the 2-D corner (right pair) but to fail the registration of the 3-D corner (left pair). On the contrary, our descriptor correctly normalize both cases.*

4 Discussion

Formalizing the simplest instance of the recognition problem makes it immediate to see that features cannot improve the quality of "recognition by reconstruction," if that was theoretically and computationally viable. However, features can provide a principled, albeit suboptimal, solution to the recognition problem: We have shown that under certain conditions viewpoint and illumination-invariant features can be constructed. Such features are local image statistics, and can be computed efficiently.

Our effort allows one to compare existing methods for invariant feature detection on a common analytical footing. Also, our framework opens the grounds for a richer class of detectors/descriptors. As an illustrative example, we introduce a 3-D corner descriptor that can be employed to establish correspondence when the state of the art fails because of violation of the local-planarity assumption.

The major drawback of viewpoint invariant features is their inability to capture shape information, thereby limiting their discriminative power when shape is more important than albedo. We argue that, by making generic assumptions on the shape, it could be possible to design viewpoint invariant features that preserve part of the shape information as well.

Appendix

A An image formation model

In order to formalize any vision problem we need a model of the so called *image formation process*, that is the process for which images result from the interaction of an imaging device and the surrounding scene. In this section we introduce the most simple *image formation model* that is powerful enough to discuss the problems we are interested in.

A.1 What is the "image" ...

An "image" is just an array of positive numbers that measure the intensity (irradiance) of light (electromagnetic radiation) incident a number of small regions ("pixels") located on a surface. We will deal with gray-scale images on flat, regular arrays, but one can easily extend the reasoning to color or multi-spectral images on curved surface, for instance omni-directional mirrors. In formulas, a digital image is a function $I : [0, N_x - 1] \times [0, N_y - 1] \to [0, N_g - 1];\ (x, y) \mapsto I(x, y)$ for some number of horizontal and vertical pixels N_x, N_y and grey levels N_g. For simplicity, we will neglect quantization in both pixels and gray levels, and assume that the image is given on a continuum $\Omega \subset \mathbb{R}^2$, with values in the positive reals:

$$I : \Omega \subset \mathbb{R}^2 \to \mathbb{R}_+;\ x \mapsto I(x) \tag{6}$$

where $x \doteq [x,\ y]^T \in \mathbb{R}^2$. When we consider more than one image, we index them with t, which may or may not indicate time: $I(x, t)$. This abstraction in representing images is all we need for the purpose of these notes.

A.2 What is the "scene"...

A simple description of the "scene", or the "object", is less straightforward. This is a *modeling* task, for which there is no right or wrong choice, and finding a right model is as much of an art as it is a science; one has to exercise discretion to strike a compromise between simplicity and realism. We consider the scene as a collection of "objects" that are volumes bounded by closed, piecewise smooth surfaces embedded in $\mathbb{R}^3$. We call the generic surface S_i, with $i = 1, \ldots, N_o$, the number of objects. Each surface is described relative to a (Euclidean) reference frame, which we call $g_i \in SE(3)$. The two entities

$$S_i \subset \mathbb{R}^3;\ g_i \in SE(3)\ \forall\ i = 1, \ldots, N_o \tag{7}$$

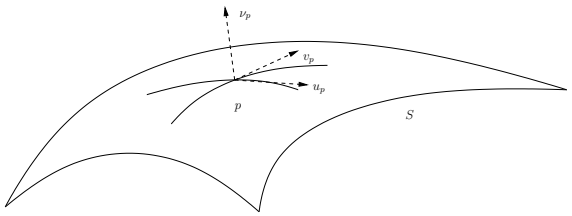

Figure 7: Local reference frame at the point p.

describe the **geometry** of the scene, and in particular we call g_i the *pose* relative to a fixed (or "inertial") reference frame[10] and S_i the *shape* of objects, although a more proper definition of shape would be the quotient S_i/g_i [29]. This is, however, inconsequential as far as our discussion is concerned.

In order to model the image formation process we need to complement our description of the objects with the elements necessary to describe their interaction with light. We explore this issue in the next section.

A.2.1 How objects interact with light: vanilla radiometry

Objects interact with light in ways that depend upon their *material* properties. Describing the interaction of light with matter is a nightmare if one seeks physical realism: one would have to start from Maxwell's equations and describe the scattering properties of the volume contained in each object. That is well beyond our scope. Besides, we do not seek physical realism, but only to capture the phenomenology of the material to the extent in which it affects the answer to our questions.

All objects emit, reflect and absorb lights and, in this regard, they are all equal. However, in order to simplify the discussion, we allow only some objects L to emit light and we restrict the other objects S to reflect/absorb light. We describe the radiation emitted by a surface L or S using its *radiance*, $R_L(p, l)$ or $R_S(q, l)$, which indicates the power density per unit area and unit solid angle emitted at a point $q \in L$ in a given direction $l \in \mathbb{H}^2$, and is measured in [W/sterad/m^2].

To make the notation more accurate, we define a Euclidean reference frame, called the *local frame*, centered at the point p with the third axis along the normal to the surface, $e_3 = \nu_p \in T_pS_i$ and first two axes parallel to the tangent plane. We call such a local reference frame g_p, which is described in homogeneous coordinates by

$$g_p = \begin{bmatrix} \begin{bmatrix} u_p & v_p & \nu_p \end{bmatrix} & p \\ 0 & 1 \end{bmatrix} \tag{8}$$

where u_p, v_p and ν_p are unit vectors. Therefore, a point q in the inertial reference frame will transform to $g_p q$ in the local frame at p. Similarly, a vector v in the inertial frame will transform to $g_{p*}v$ in the local frame where

$$g_{p*} = \begin{bmatrix} \begin{bmatrix} u_p & v_p & \nu_p \end{bmatrix} & 0 \\ 0 & 0 \end{bmatrix}. \tag{9}$$

When we consider the particular direction l from a point $q \in L$ on the light source towards a point $p \in S$ on the scene, this is given by $g_{q*}(p - q) = g_q p - 0 = g_q p$. Therefore, given a solid angle $d\Omega_L$ and an area element dL on the light source, the power per solid angle and unit foreshortened[11] area radiated from a point q towards p is given by

$$R_L(q, g_q p) d\Omega_L \langle \nu_q, g_q p \rangle dL \tag{10}$$

where $g_q p \in \mathbb{H}^2$ is intended as a unit vector. Now, how big a patch dL of the light we see standing at a point p on the scene depends on the solid angle $d\Omega_S$ we are looking through. Following Figure 8 we have that

$$dL = d\Omega_S \|p - q\|^2 / \langle \nu_q, l_{qp} \rangle \tag{11}$$

[10] If a point p is represented in coordinates via $\mathbf{X} \in \mathbb{R}^3$, then the transformed point gp is represented in coordinates via $R\mathbf{X}+T$, where $R \in SO(3)$ is a rotation matrix and $T \in \mathbb{R}^3$ is a translation vector. The action of $SE(3)$ on a vector is denoted by g_*v, so that if the vector v has coordinates $V \in \mathbb{R}^3$, then g_*v has coordinates RV. See [34], chapter 2 and appendix A, for more details.

[11] If the area element on the light source is dL, the portion of the area seen from p is given by $\langle \nu_q, g_q p \rangle dL$; this is called the *foreshortened area*.

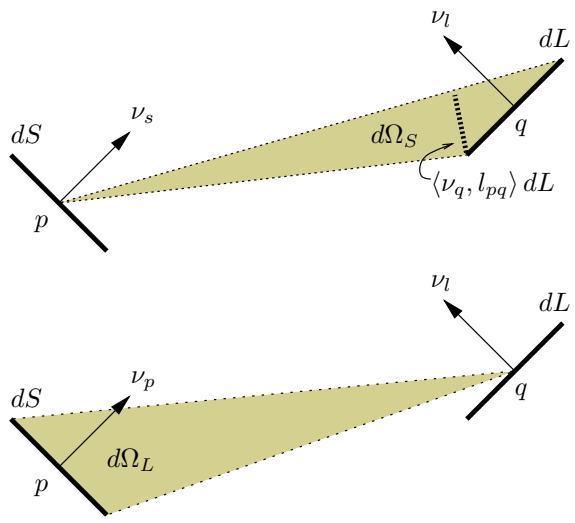

Figure 8: Energy balance: a light source patch dL radiates energy towards a surface patch dS. Therefore, the power injected in the solid angle $d\Omega_L$ by dL equals the power received by dS in the solid angle $d\Omega_S$. Equation (12) expresses this balance in symbols.

where we have defined $l_{qp} \doteq q - p/\|q - p\|$ and the inner product at the denominator is called *foreshortening*. Similarly, the solid angle $d\Omega_L$ shines a patch of the surface dS. The two are related by

$$d\Omega_L = \frac{dS}{\|p - q\|^2} \langle \nu_q, l_{pq} \rangle \tag{12}$$

where $l_{pq} = -l_{qp} = p - q/\|p - q\|$. Substituting the expressions of $d\Omega_L$ and dL in the previous two equations into (10), one obtains the infinitesimal power received at the point p.

Now, we want to write the portion of power exiting the surface at p in the direction of a pixel x through an area element dS. First, we need to write the direction of x in the local reference frame at p. We assume that x is a unit vector, obtained for instance via central perspective projection

$$\pi : \mathbb{R}^3 \longrightarrow \mathbb{S}^2; \; p \mapsto \pi(p) \doteq x. \tag{13}$$

However, the point p is written in the inertial frame, while x is written in the frame of the camera at time t. We need to first transform x to the inertial frame, via $g_*(t)^{-1}x$, and then express this in the local frame at p, which yields $g_{p*}^{-1} g_*(t)^{-1} x$. We call the normalized version of this vector $l_{px}(t)$. The total energy radiated by the point p in a direction v is obtained by integrating, of all the energy coming from the light source L weighted by the so called *bi-directional reflectance distribution function* (BRDF) $\beta_i : \mathbb{H}^2 \times \mathbb{H}^2 \to \mathbb{R}_+; (v, l) \mapsto \beta_i(v, l)$ at point p. The BRDF determines the portion of energy[12] coming from a direction l that is reflected in the direction v, each represented as a point on the half-sphere $\mathbb{H}^2$ centered at the point p and is therefore measured in [1/sterad]. This model neglects diffraction, absorption, subsurface scattering and other aberrations; the BRDF only describes the reflective properties of materials (*reflectance*). Note that β_i depends on the point p on the surface, and we are imposing no restrictions on such a dependency. For instance, we do *not* assume that β_i is constant with respect to p (homogeneous material). When emphasizing such a dependency we write $\beta(v, l; p)$.

Integrating the power received at point $p \in S$ through the BRDF, the power exiting from p in the direction of x through an area element dS results

$$R_S(p, x) dS(p) = \int_L \beta(l_{px}(t), g_p q) R_L(q, g_q p) d\Omega_L(q) \langle \nu_q, g_q p \rangle dL(q) \tag{14}$$

[12]The term "energy" is used colloquially here to indicate radiance, irradiance, radiant density, power etc.

where the arguments in the infinitesimal forms $dS, dL, d\Omega_L$ indicate their dependency. Now, we can substitute[13] the expression of $d\Omega_L$ from (12) and simplify the area element dS, to obtain the *radiance* of the surface at p

$$R_S(p,x) = \int_L \beta(l_{px}(t), g_p q) R_L(q, g_q p) \frac{\langle \nu_q, g_q p \rangle}{\|p-q\|^2} \langle \nu_p, l_{pq} \rangle dL(q) \tag{15}$$

Since the norm $\|p-q\|$ is invariant to Euclidean transformations, we can write it as $\|g_q p\|$. Now, if the size of the scene is small compared to its distance to the light, this term is almost constant, and therefore the measure

$$dE(q, g_q p) \doteq R_L(q, g_q p) \frac{\langle \nu_q, g_q p \rangle}{\|g_q p\|^2} dL(q) \tag{16}$$

can be thought of as a property of the light source. Since we cannot untangle the contribution of R_L from that of dL, we just choose dE to describe the power distribution radiated by the light source. Therefore, we have

$$R_S(p,x) = \int_L \beta(l_{px}(t), g_p q) \langle \nu_p, l_{pq} \rangle dE(q, g_q p). \tag{17}$$

This is the portion of power per unit area and unit solid angle radiated from a point p on a reflective surface towards a point x on the image at time t. The collection

$$\beta_i(\cdot,\cdot) : \mathbb{H}^2 \times \mathbb{H}^2 \to \mathbb{R}_+, \ i = 1, \dots, N_o; \ L \text{ and } dE : L \times \mathbb{H}^2 \to \mathbb{R}_+ \tag{18}$$

describes the **radiometry** of the scene (reflectance and illumination).

A.2.2 Dynamics

In addition, reflectance (BRDF) and geometry (shape and pose) are properties of each object that can change over time. So, in principle, we would want to allow $\beta_i, \ S_i, \ g_i$ to be functions of time. In practice, we will assume that the material of each object does not change, but only its shape, pose and of course illumination. Therefore, we will use

$$S_i = S_i(t); \ g_i = g_i(t), \quad t \in [0, T] \tag{19}$$

to describe the **dynamics** of the scene. The index t can be thought of as *time*, in case a sequence of measurements is taken at adjacent instants or continuously in time, or it can be thought of as an *index* if disparate measurements are taken under varying conditions (shape and pose). Note that, as we mentioned, the light source (L, dE) can also change over time. When emphasizing such a dependency we write $L(t)$ and $dE(q, l; t)$.

Example 1 *The simplest surface S_i one can conceive of is a plane: $S_i = \{p \in \mathbb{R}^3 \mid \langle \nu_i, p \rangle = d_i\}$ where ν_i is the unit normal to the plane, and d_i is its distance to the origin. For a plane not intersecting the origin, $1/d$ can be lumped into ν, and therefore three numbers are sufficient to completely describe the surface in the inertial reference frame. In that case we simply have S_i a constant, and $g_i = e$, the identity. A simple light source is an ideal point source, which can be modeled as $L \in \mathbb{R}^3$ with infinite power density $dE = E_l \delta(q - L)$. Another common model is a constant ambient illumination, which can be modeled as a sphere $L = \mathbb{S}^2$ with $dE = E_0 dL$. We will discuss examples of various models for the BRDF later.*

Remark 1 (Choosing a level of granularity in the representation) *Note that by assuming that the world is made of surfaces we are already imposing significant restrictions, and we are implicitly choosing a level of description for our representation. Consider for instance the fabric shown in Figure 9. There is no surface there. The fabric is made of thin one-dimensional threads, just woven tightly enough to give the impression of spatial continuity. Therefore, we choose to represent them as a smooth surface. Of course, the variation in the appearance due to the fine-scale structure of the threads has to be captured somehow, and we delegate this task to the reflectance model. Naturally, one could even describe each individual thread as a*

[13] Most often in radiometry one performs the integral above with respect to the solid angle $d\Omega_S$, rather than with respect to the light source. For those that want to compare the expression of the radiance R_S with that derived in radiometry, it is sufficient to substitute the expressions of dL and $d\Omega_L$ above, to obtain $R_S(p,x) = \int_{\mathbb{H}^2} \beta(l_{px}(t), g_p q) R_L(q, g_q p) \langle \nu_p, g_q p \rangle d\Omega_S(p)$. In our context, however, we are interested in separating the contribution of the light and the scene, and therefore performing the integral on L is more appropriate.

Figure 9: *A complex shape (woven thread) with simple reflectance (homogeneous material), or a simple shape (a smooth surface) with complex reflectance (texture)?*

cylindrical surface modeled as an object S_i, but this is well beyond the detail that we want to capture. This example illustrate the notion that defining objects entails a notion of scale. *Something (e.g. a thread) is an object at one scale, but is merely part of a texture at a coarser scale. Figure 9 highlights the modeling tradeoff between shape and reflectance: one could model the fabric as a very complex object (woven thread) made of homogeneous material (wool), or as a very simple object (a smooth surface) made of textured material. This is a modeling choice.*

Remark 2 (Tradeoff between shape and motion) *We note that, instead of allowing the surface S_i to deform arbitrarily in time via $S_i(t)$,* and *moving rigidly in space via $g_i(t) \in SE(3)$, we can lump the motion and deformation into $g_i(t)$ by allowing it to belong to a more general class of deformations G, for instance*

diffeomorphisms, and let S_i be constant. Alternatively, we can lump the deformation $g_i(t)$ into S_i and just describe the surface in the inertial reference frame via $S_i(t)$. This can be done with no loss of generality, and it reflects a fundamental tradeoff in modeling the interplay between shape and motion [52].

Now, if we agree that a scene can be described by its *geometry, photometry* and *dynamics*, we must decide how these relate to the measured images.

A.3 And how are the two related?

Given a description of the geometry, photometry and dynamics of a scene, a model of the image is obtained through a description of the *imaging device*. An imaging device is a series of elements designed to direct light propagation. This is typically modeled through diffraction, reflection, and refraction. We will ignore the first two propagation effects, and only consider the effects of refraction. For simplicity, we can also assume that the set of objects that act as light sources and those that act as light sinks are disjoint, so that $S_i \cap L = \emptyset$, i.e. we ignore inter-reflections. In that case, we can just lump all the objects into one, which we call the scene $S \doteq \cup_{i=1}^{N_o} S_i$ with its corresponding BRDF, $\beta = \cup_{i=1}^{N_o} \beta_i$. Note that S needs not be simply connected.

Equation (14) specifies how much power is radiated from the element dS at point p towards the pixel x. The next step consists of quantifying what portion of this energy gets absorbed by the pixel at location x. This follows a similar calculation, which we do not report here, and instead refer the reader to [21] (page 208). There, it is argued that the irradiance at the pixel x is equal to the radiance at the corresponding point p on the scene, up to an approximately constant factor, which we lump into R_S. The point p and its projection x onto the image plane at time t are related by the equations

$$x = \pi(g(t)p) \quad p = g(t)^{-1}\pi_S^{-1}(x) \tag{20}$$

where $\pi_S^{-1} : \mathbb{S}^2 \to \mathbb{R}^3$ denotes the inverse projection, which consists in scaling x by its depth $Z(x)$ in the current reference frame, which naturally depends on S. Therefore, the equation below, known as the *irradiance equation*, takes the form

$$I(x,t) = R_S(p, \pi(g(t)p)) = R_S(g(t)^{-1}\pi_S^{-1}(x), x). \tag{21}$$

After we substitute the expression of the radiance (17), we have the *imaging equation*

$$\boxed{\begin{cases} I(x,t) = \int_L \beta(l_{px}(t), g_p q)\langle \nu_p, l_{pq}\rangle dE(q, g_q p); \\ x = \pi(g(t)p);\ p \in S \end{cases}} \tag{22}$$

where the symbols above are defined as follows:

Notation: In the equation above, we have defined $l_{px} \doteq {g_{p_*}}^{-1} g_*(t)^{-1} x$, g_p and g_{p_*} are defined by Equation (8) and (9) respectively, $l_{pq} \doteq p - q/\|p-q\|$ and $g_p q$ indicates the (normalized) direction from p to q, and similarly for $g_q p$;

Light source: $L \subset \mathbb{R}^3$ is the (possibly time-varying) collection of light sources emitting energy with a distribution $dE : L \times \mathbb{H}^2 \to \mathbb{R}_+$ at every point $q \in L$ towards the direction of a point p on the

Scene: a collection of (possibly time-varying) piecewise smooth surfaces $S \subset \mathbb{R}^3$; $\beta : \mathbb{H}^2 \times \mathbb{H}^2 \times S \to \mathbb{R}$ is the bidirectional reflectance distribution function (BRDF) that depends on the incident direction, the reflected direction and the point $p \in S$ on the scene S and is a property of its material.

Motion: relative motion between the scene and the camera is described by the motion of the camera $g(t) \in SE(3)$ and possibly the action of a more complex group G, or simply by allowing the surface $S(t)$ to change over time.

Projection: $\pi : \mathbb{R}^3 \mapsto \mathbb{S}^2$ denotes ideal (pinhole) perspective projection, modeled here as projection onto the unit sphere, although the same model applies if $\pi : \mathbb{R}^3 \to \mathbb{P}^2$, in which case l_{px} has to be normalized accordingly.

Visibility and cast shadows: One should also add to the equation two characteristic function terms: $\chi_v(x,t)$ outside the integral, which models the visibility of the scene from the pixel x, and $\chi_s(p,q)$ inside the integral to model the visibility of the light source from a scene point (cast shadows). We are omitting these terms here for simplicity. However, in some cases that we discuss in the next section, discontinuities due to visibility or cast shadows can be the only source of visual information.

The imaging equation is relevant because most of computer vision is about inverting it; that is, inferring properties of the scene (shape, material, motion) regardless of pose, illumination and other nuisances (the visual reconstruction problem). However, it the general formulation above, one cannot infer photometry, geometry and dynamics from images alone. Therefore, we are interested in deriving a model that strikes a balance between invertibility (i.e. it should contain only parameters that can be identified) and realism (i.e. it should capture the phenomenology of image formation). In the next section we illustrate simple models that are widely used in computer graphics to generate realistic, albeit non-perfect, looking images: Phong (corrected) [42], Ward [51] and Torrance-Sparrow (simplified) [49]. All these models include a function $\rho_d(p)$ called *(diffuse) albedo*, and a function $\rho_s(p)$ called *specular albedo*. Diffuse albedo is often called just albedo, or, improperly, *texture*. We will discuss various special cases of the imaging equation and the role they play in visual reconstruction. Here we limit ourselves to deriving the model under a generic[14] illumination consisting of an ambient term and a number of concentrated point light sources at infinity: $L = \mathbb{S}^2 \cup \{L_1, L_2, \ldots, L_k\},\ L_i \in \mathbb{R}^3$, $dE(q) = E_0 dL(q) + \sum_{i=1}^k E_i \delta(q - L_i)$. In this case the imaging equation reduces to

$$\boxed{\begin{cases} I(x,t) = \rho_d(p) \left(E_0 + \sum_{i=1}^k E_i \langle \nu_p, L_i \rangle \right) + \rho_s(p) \sum_i E_i \frac{\left\langle g^{-1}(t)x + L_i/\|L_i\|, \nu_p \right\rangle^c}{\langle g^{-1}(t)x, \nu_p \rangle} \\ x = \pi(g(t)p); \quad p = S(x_0). \end{cases}} \tag{23}$$

Remark 3 *Note that this model does not explicitly include occlusions and shadows. Also, note that the first (diffuse) term does not depend on the viewpoint g, whereas the second term (specular) does. However, note that, depending on the coefficient c, the second term is only relevant when x is close to the specular direction, and therefore if one assumes that the light sources are concentrated, the second term is relevant in a small subset of the scene. If we threshold the effects of the second term based on the angle between the viewing and the specular direction, then we can write the above model as*

$$I(x,t) = \begin{cases} \rho_d(p) \left(E_0 + \sum_{i=1}^k E_i \langle \nu_p, L_i \rangle \right) & \text{if } \ \left\langle g^{-1}(t)x + L_i/\|L_i\|, \nu_p \right\rangle < \gamma(c) \ \forall\ i \\ \rho_s(p) E_{\hat{i}} & \text{otherwise} \end{cases} \tag{24}$$

where $\hat{i} = \arg\min_i \frac{\left\langle g^{-1}(t)x + L_i/\|L_i\|, \nu_p \right\rangle^c}{\langle g^{-1}(t)x, \nu_p \rangle}$ *which justifies the rank-based model of [26]. Empirical evaluation of the validity of this model, and the resulting "brightness constancy constraint" discussed in the next subsection, has not been addressed thoroughly in the literature.*

The "identity" of a scene or an object is specified by its shape S and its reflectance properties β. The illumination $L(t), dE(\cdot,t)$, visibility $\chi(x,t)$ and pose/deformation $g(t)$ are "nuisance factors" that affect the measurements but *not* the identity of the scene/object[15]. They change with the view, whereas the identity of the object does not (see Figure 10). In the imaging equation (23) we measure $I(x,t)$ for all $x \in \Omega$ and $t = t_1, t_2, \ldots, t_m$, and the unknowns are $L(t_j), dE(\cdot, t_j), g(t_j)$, which for simplicity we indicate as $L_j, dE_j(\cdot), g_j$ respectively, for all $j = 1, \ldots, m$. For simplicity, we indicate all the unknowns of interest with the symbol ξ (note that some unknowns are infinite-dimensional), and all the nuisance variables with ν. Equation (23), once we write the coordinates of the point p relative to the pixel in the moving frame, $p = g(t)^{-1}\pi^{-1}(x,t)$, can then be written as a functional h, formally, as follows:[16]

$$\boxed{I = h(\xi, \nu) + n.} \tag{25}$$

[14] See Problem 5.

[15] Depending on the problem at hand, some unknowns may play either role: motion, for instance, could be a quantity of interest in tracking, but it is a nuisance in recognition. Illumination will almost always be a nuisance.

[16] Note that the symbol ν for "nuisance" in the symbolic equation may be confused with ν_p, the normal to the surface in the physical model. Since the two symbols will be used exclusively in different contexts, it should be clear which one we are referring to.

Figure 10: *Examples of variability among different images of the same scene (top-left): illumination (top-center), viewpoint (top-right, bottom-left), removal/replacement of parts (bottom-center), partial occlusion (bottom-right).*

where, to summarize the equivalence of (25) with (23), we have

$$\begin{cases} I : \Omega \subset \mathbb{R}^2 \to \mathbb{R}^3 \\ \xi \in C(\mathbb{R}^2 \backslash \mathcal{D} \to \mathbb{R}^3) \times BV(\mathbb{S}^2 \times \mathbb{S}^2 \to \mathbb{R}^+) \doteq \mathcal{S} \\ \nu \in SE(3) \times BV(\mathbb{R}^3 \to \mathbb{R}^+) \times \mathbb{P}(\mathbb{R}^3 \to \mathbb{R}^2) \doteq \nu \\ h : \mathbb{R}^3 \times BV(\mathbb{R}^3 \to \mathbb{R}^+) \times SE(3) \times \mathbb{R}^+ \times \mathbb{R}^{3k} \times \mathbb{R}^k \to \mathbb{R}^+ \\ \qquad (p, \beta, g, E_0, \{L_1, \dots, L_k\}, \{E_1, \dots, E_k\}) \mapsto I \\ n \sim \mathcal{N}(0, Q) \end{cases} \tag{26}$$

where $\mathcal{D}$ is a subset of measure zero (the set of discontinuities), BV denotes functions of bounded variation. We will use the symbolic notation of (25) and the explicit notation of (23) interchangeably, depending on convenience. In some cases we may indicate the arguments of the functions I, ξ, ν, n explicitly.

Remark 4 (Occlusions and cast shadows) *Occlusions are an accident of image formation that significantly complicates our modeling efforts. In fact, while they are "nuisances" in the sense that they do not depend solely on the scene, the* do *depend on both the scene and the viewpoint (for occlusions) and illumination (for cast shadows). That is why, despite depending on the nuisance, under suitable conditions they can exploit to infer the shape of the scene (see [53] for occlusions and [4] for cast shadows).*

B Special cases of the imaging equation and their role in visual reconstruction (taxonomy)

All vision algorithms make implicit or explicit assumptions on the imaging equation (23). In this section we discuss the most typical of such assumptions, particularly for what concerns the reflectance of the objects and the illumination.

B.1 Empirical reflectance models

Most common materials can be described by a BRDF. Exceptions include translucent materials (e.g. skin), anisotropic material (e.g. brushed aluminum), micro-structured material (e.g. hair) etc. However, since our goal is not realism in a physical simulation, we are content with some common BRDF that are well established in computer graphics: Phong (corrected) [42], Ward [51] and Torrance-Sparrow (simplified) [49].

Phong (corrected) $\beta(v,l) = \rho_d(p) + \rho_s(p)\cos^c\delta/\cos\theta_i\cos\theta_o$.
Here $\cos\delta = \langle g(t)^{-1}x + q/\|q\|, \nu_p\rangle$ where each term in the inner product is normalized, and $\theta_i \doteq \arccos\langle l, \nu_p\rangle$, and $\arccos(\theta_o) \doteq \langle v, \nu_p\rangle$; $c \in \mathbb{R}$ is a coefficient that depends on the material.

Ward $\beta(v,l) = \rho_d(p) + \rho_s(p)\frac{\exp(-\tan^2(\delta)/\alpha^2)}{\sqrt{\cos\theta_i\cos\theta_o}}$.
Here $\alpha \in \mathbb{R}$ is a coefficient that depends on the material and is determined empirically.

Torrance-Sparrow (simplified) $\beta(v,l) = \rho_d(p) + \rho_s(p)\frac{\exp(-\delta^2/\alpha^2)}{\cos\theta_i\cos\theta_o}$.

Separable radiance As Nayar and coworkers point out [41], the radiance for the latter model can be written as the sum of products, where the first factor depends solely on material (diffuse and specular albedo), whereas the second factor compounds shape, pose and illumination.

In all these cases, $\rho_d(p)$ is an unknown function called *(diffuse) albedo*, and $\rho_s(p)$ is an unknown function called *specular albedo*. Diffuse albedo is often called just albedo, or, improperly, *texture*.

Note that the first term (diffuse reflectance) is the same in all three models. The second term (specular reflectance) is different. Surfaces whose reflectance is captured by the first term are called Lambertian, and are by far the most studied in computer vision. The rest of this appendix discusses various models of the illumination for the Lambertian and non-Lambertian reflection cases.

B.2 Lambertian reflection

Lambertian surfaces essentially look the same regardless of the viewpoint: $\beta(v,l) = \beta(w,l)\ \forall w \in \mathbb{H}^2$. This yields to major simplifications of the image formation model. Moreover, in the case of constant illumination, it allows relating different views of the same scene to one another directly, bypassing the image formation model. This is known as the *correspondence problem*, which relies crucially on the Lambertian assumption and the resulting brightness constancy constraint.[17] We address this case first.

B.2.1 Constant illumination

In this case we have $L(t) = L$ and $dE(q,l;t) = dE(q,l)$. We consider two simple light source models first.

Ambient light

Ambient light is due to inter-reflection between different surfaces in the scene. Since modeling such inter-reflections is quite complicated,[18] we will approximate it by assuming that there is a constant amount of energy that "floods" the ambient space. This can be approximated by a sphere radiating constant energy: $L = \mathbb{S}^2$ and $dE = E_0 dL$. In this case, the imaging equation reduces to

$$I(x,t) = \rho_d(p)E_0 \int_{\mathbb{S}^2} \langle \nu_p, l\rangle d\Omega(l) \tag{27}$$

[17] Although the constraint is often used *locally* to approximate surfaces that are *not* Lambertian.
[18] There is some admittedly sketchy evidence that inter-reflections are not perceptually salient [12].

Due to the symmetry of the light source, assuming there are no shadows, we can always change the global reference frame so that $\nu_p = e_3$; therefore, the integral does not depend on p, and is a constant that, together with E_0, can be lumped into ρ_d, yielding the simplest possible model that, when written with respect to a moving camera, gives

$$\boxed{\begin{cases} I(x,t) = \rho(p) \\ x = \pi(g(t)p); \quad p = S(x_0). \end{cases}} \tag{28}$$

Note that this model effectively neglects illumination, for one can think of a scene S that is self-luminous, and radiates an equal amount of energy $\rho(p)$ in all directions. Even for such a simple model, however, performing visual inference is non-trivial. It has been done for a number of special cases:

Constant albedo: silhouettes When $\rho(p)$ is constant, the only information in Equation (28) is at the discontinuities between $x = \pi(g(t)p), p \in S$ and $p \notin S$, i.e. at the occluding boundaries. Given suitable conditions, that have been first studied by Aström et al. [8], motion $g(t)$ and shape S can be recovered. The reconstruction of shape S and albedo ρ has been addressed in an infinite-dimensional optimization framework by Yezzi and Soatto [53] in their work on stereoscopic segmentation.

Smooth albedo The stereoscopic segmentation framework has been extended to allow the albedo to be smooth, rather than constant. The algorithm in [25] provides an estimate of the shape of the scene S as well as its albedo $\rho(p)$ given its motion relative to the viewer, $g(t)$.

Piecewise constant/piecewise smooth albedo The same framework has been recently extended to allow the albedo to be piecewise constant in [26]. This amount to performing region-based segmentation a' la Mumford-Shah [40] on the scene surface S. Although it has not been done yet, the same ideas could be extended to piecewise smooth albedo.

Nowhere constant albedo When $\nabla\rho(p) \neq 0$ everywhere in p, the image formation model can be bypassed altogether, leading to the so-called correspondence problem which we will see shortly. This is at the base of most traditional stereo reconstruction algorithms and structure from motion. Since these techniques apply without regard to the illumination, we will address this after having relaxed our assumptions on illumination.

Point light(s)

A countable number of stationary point light sources can be modeled as $L = \{L_1, L_2, \ldots, L_k\},\ L_i \in \mathbb{R}^3$, $dE = \sum_{i=1}^{k} E_i \delta(q - L_i)$. In this case the imaging equation reduces to

$$I(x,t) = \sum_{i=1}^{k} E_i \rho_d(p) \langle \nu_p, p - L_i / \|p - L_i\| \rangle. \tag{29}$$

Note that, if we neglect occlusions and cast shadows, the sum can be taken inside the inner product and therefore there is no loss of generality in assuming that there is only one light source. If the light sources are at infinity, p can be dropped from the inner product; furthermore, the intensity of the source E multiplies the light direction, so the two can be lumped into the vector L. We can therefore further simplify the above model to yield, taking into account camera motion,

$$\boxed{\begin{cases} I(x,t) = \rho(p) \langle \nu_p, L \rangle \\ x = \pi(g(t)p); \quad p = S(x_0) \end{cases}} \tag{30}$$

Inference from this model has been addressed for the following cases.

Constant albedo Yuille et al. [55] have shown that given enough viewpoints and lighting positions one can reconstruct the shape of the scene. Jin et al. [24] have proposed an algorithm for doing so, which estimates shape, albedo and position of the light source in a variational optimization framework. If the position of the light source is known and there is no camera motion, this problem reduces to classical shape from shading [22].

Smooth/piecewise smooth albedo In this case, one can easily show that albedo and light source cannot be recovered since there are always combinations of the two that generate the same images. However, under suitable conditions shape can still be estimated, as we discuss next.

Nowhere constant radiance If the combination of albedo and the cosine term (the inner product in (30)) result in a radiance function that has non-zero gradient, we can think of the radiance as an albedo under ambient illumination, and therefore this case reduces to multi-view stereo, which we will discuss shortly. Naturally, in this case we cannot disentangle reflectance from illumination, but under suitable conditions we can still reconstruct the shape of the scene, as we discuss shortly in the context of the correspondence problem.

Cast shadows If the visibility terms are included, under suitable conditions about the shape of the object and the number and nature of light sources, one can reconstruct an approximation of the shape of the scene.

General light distribution

An arbitrary distant light distribution can be modeled as a positive density on the sphere at infinity: $L = \mathbb{S}^2$. Any positive density on the sphere can be approximated arbitrarily well by a sum of Gaussians, a result known to Wiener, slightly modified to take into account the spherical ambient space. However, each Gaussian can be represented as a convolution of a delta measure with a canonical Gaussian (zero-mean). When inserted into the imaging equation, the effect of the Gaussian kernel and the BRDF compound in a way that cannot be discerned from the data alone. Therefore, we can lump the Gaussian kernel into the BRDF and be left with point light sources with no loss of generality. Naturally, in practice each each light may have a different dispersion matrix, which in general results in an empirical coefficient (α in the Torrance-Sparrow model) that is direction-dependent. If we allow the BRDF to be anisotropic, we can never distinguish reflectance from illumination. Consider for instance a polished sphere illuminated by a Gaussian light sources, compared to a rougher sphere illuminated by a point.

Note that most current work on general representation of illumination uses a series expansion of the distribution dE on $L = \mathbb{S}^2$ into spherical harmonics [44]. This is problematic for two reasons: first, spherical harmonics are *global*, so the introduction of another term in the series affects the entire image. Second, while any function on the sphere can be approximated with spherical harmonics, there is no guarantee that such a function be *positive*. Indeed, the harmonic terms in the series are themselves not positive, and therefore each individual component does not lend itself to be interpreted as a valid illumination, and there is no guarantee except in the limit where the number of terms goes to infinity that the truncated series will be a vaild illumination. The advantage of a sum of Gaussian approximation is that one can approximate any positive function, and given any truncation of the series one is guaranteed to have a positive distribution dE.

Given these considerations, we restrict our attentions to illumination models that consist of the sum of a constant ambient term and a countable number of point light sources. The general case, therefore, reduces to the special cases seen above:

$$L = \mathbb{S}^2; \quad dE(q) = E_0 dL(q) + \sum_{i=1}^{k} E_i \delta(q - L_i). \tag{31}$$

Note that the energy does not depend on the direction, since for distant lights (sphere of infinite radius) all directions pointing towards the scene are normal to L.

Multi-view stereo and the correspondence problem

If the radiance of the scene $R_S(p)$ is not constant, under suitable conditions one can do away with the image formation model altogether. Consider in fact the irradiance equation (21). Under the Lambertian assumption, given (at least) two viewpoints, indexed by t_1 and t_2, we have that

$$I(x_1, t_1) = R_S(p, \pi(g(t_1)p)) = I(x_2, t_2) \tag{32}$$

without regard to how the radiance R_S comes to existence. The relationship between x_1 and x_2 depends solely on the shape of the scene S and the relative motion of the camera between the two time instants, $g_{12} \doteq g(t_1)g(t_2)^{-1}$:

$$x_1 = \pi(g_{12}\pi_S^{-1}(x_2)) \doteq w(x_2; S, g_{12}). \tag{33}$$

Therefore, one can forget about how the images are generated, and simply look for the function w that satisfies (substitute the last equation into the previous one)

$$\boxed{I(w(x_2; S, g_{12}), t_1) = I(x_2, t_2)}. \tag{34}$$

Finding the function w from the above equation is known as the *correspondence problem*, and the equation above is the *brightness constancy constraint*.

More recently, Faugeras and Keriven have cast the problem of stereo reconstruction in an infinite-dimensional optimization framework, where the equation above is integrated over the entire image, rather than just in a neighborhood of feature points, and the correspondence function w is estimated implicitly by estimating the shape of the scene S, with a given motion g. This works even if ρ is constant, but due to a non-uniform light and the presence of the Lambertian cosine term (the inner product in equation (30)) the radiance of the surface is nowhere constant (shading effect, or attached shadow) and even in the case of cast shadows, if the light does not move. In the presence of regions of constant radiance, the algorithm interpolates in ways that depend upon the regularization term used in the infinite-dimensional optimization (see [13] for more details).

B.2.2 Constant viewpoint: photometric stereo

When the viewpoint is fixed, but the light changes, inverting the model above is known as photometric stereo [21]. If the light configuration is not known and is allowed to change between views, Belhumeur and coworkers have shown that this problem cannot be solved [7]. In particular, given two images one can pick a surface S at will, and construct two light distributions that generate the given images, even if the scene is known to be Lambertian. However, this result relies on the presence of a single point light source. We conjecture that if the illumination is allowed to contain an ambient term, these results do not apply, and therefore reconstruction could be achieved. Note that psychophysical experiments suggest that face recognition is extremely hard for humans under a point light source, whereas a more complex illumination term greatly facilitates the task.

B.3 Non-Lambertian reflection

In this subsection we relax the assumption on reflectance. While, contrary to intuition, a more complex reflectance model can in some cases facilitate recognition, in general it is not possible to disentangle the effects of shape, reflectance and illumination. We start by making assumptions that follow the taxonomy used for the Lambertian case in the previous subsection.

B.3.1 Constant illumination

Ambient light

In the presence of ambient illumination, the specular term of an empirical reflection model, for instance Phong's, takes the form

$$\rho_s(p) \int_{-\pi}^{\pi} \int_{0}^{\pi/2} \frac{\cos^k \delta}{\cos \theta_o} \sin \theta_i d\theta_i d\phi_i \tag{35}$$

If the exponent $c \to \infty$, only one point on the light surface $\mathbb{S}^2$ contributes to the radiance emitted from the point p. Since the distribution dE is uniform on L, we conclude that, if we exclude occlusions and cast shadows, this term is a constant. This can be considered as a limit argument to the conjecture that, in the presence of ambient illumination, the specular term is negligible compared to the diffuse albedo. Naturally, if an object is perfectly specular, it renders the viewer an image of the light source, so in this case inter-reflection is the dominant contribution, and the ambient illumination approximation is no longer justified. See for instance Figure 11.

Figure 11: In the presence of strongly specular materials, the image is essentially a distorted version of the light source. In this case, modeling inter-reflections with an ambient illumination term is inadequate.

Point light(s)

In the presence of point light sources, the specular component of the Phong models becomes

$$\sum_i E_i \rho_s(p) \frac{\langle g^{-1}(t)x + L_i/\|L_i\|, \nu_p \rangle^c}{\langle g^{-1}(t)x, \nu_p \rangle} \tag{36}$$

where the arguments of the inner products are normalized. In this case, assuming that a portion of the scene is Lambertian and therefore motion and shape can be recovered, one can invert the equation above to estimate the position and intensity of the light sources. This is called "inverse global illumination" and was addressed by Yu and Malik [54]. If the scene is dominantly specular, so no correspondence can be established from image to image, we are not aware of any general result that describes under what condition shape, motion and illumination can be recovered. Savarese and Perona [45] study the case when assumptions on the position and density of the light, such as the presence of straight edges at known position, can be exploited to recover shape.

General light

In general, one cannot separate reflectance properties of the scene with distribution properties of the light sources. Jin et al. [26] showed that one can recover shape S as well as the radiance of the scene, which mixes the effects of reflectance and illumination.

B.3.2 Constant viewpoint

In the presence of multiple point light sources, Many have studied the conditions under which one can recover the position and intensity of the light sources, see for instance [41] and references therein. Variations of photometric stereo have also been developed for this case, starting from [23].

B.3.3 Reciprocal viewpoint and light source

Zickler et al. [56] have developed techniques to exploit a very peculiar imaging setup where a point light source and the camera are switched in pairs of images, which allows us to eliminate the BRDF from the

imaging equation.

Remark 5 *It is possible to approximate to an arbitrary degree any positive real-valued function by a sum a (spherical) Gaussians plus a constant. Therefore, it is possible that, given an illumination* $dE(q) = (E_0 + \sum_{i=1}^{n} G(q - L_i; \sigma_i))dL$ *and a BRDF* β *belonging to one of the phenomenological models here described that generate a given set of images* $\{I_j\}$*, there exists another BRDF* $\tilde{\beta}$ *belonging to the same class such that the illumination* $d\tilde{E}(q) = E_0 dL + \sum_{i=1}^{n} \delta(q - L_i)$ *generates the same collection of images* $\{I_j\}$*.*

Remark 6 (Illumination variability of a Lambertian plane) *Consider an image generated by a model (24). We are interested in modeling the variability induced in two images of the same scene under different illumination. We will assume that illumination can be approximated by an ambient term* E_0 *and a concentrated point source with intensity* E_1 *located at* L*, so that each image* $I(x_i, t_i)$ *can be approximated by* $\rho_d(p)(E_0(t_i i) + E_1(t_i)\langle \nu_p, L(t_i i)\rangle + \beta(i)$ *where the latter term lumps together the effects of non-Lambertian reflection. The relationship between two images, the, can be obtained by eliminating the diffuse reflection* ρ_d*, so as to obtain*

$$I(x_1, t_1) = I(x_2, t_2)\frac{E_1(t_1) + E_1(t_1)\langle \nu_p, L(t_1)\rangle}{E_0(t_2) + E_1(t_2)\langle \nu_p, L(t_2)\rangle} - \frac{-\beta(t_1)}{E_0(t_2) + E_1(t_2)\langle \nu_p, L(t_2)\rangle} + \beta(t_2). \tag{37}$$

Now, if the scene is a plane, then the first fraction on the right hand side does not depend on p*, i.e. it is a constant, say* α*. The second and third term depend on* p *if the scene is non-Lambertian. However, if non-Lambertian effects are negligible, or absent like in our assumptions, then the second term can also be approximated by a constant, say* β*. Furthermore, for the case of a plane* x_1 *and* x_2 *are related by a homography,* $x_1 = Hx_2$ *where* x_1 *and* x_2 *are intended in homogeneous coordinates. Therefore, the relationship between the two images can be expressed as*

$$I(x_2, t_2) = \alpha I(Hx_2, t_2) + \beta. \tag{38}$$

One can therefore think of one of the images (e.g. $I(\cdot, t_0) \doteq \rho$*) as the scene, and the images are obtained by a warping* H *of the domain and a scaling* α *and offset* β *of the range. All the nuisances,* H, α, β *are* invertible*, and therefore a planar Lambertian scene one can construct an invariant descriptor.* H *can be moded-out by fixing 4 points, and* α *and* β *by normalizing the image intensity histogram. Images for which* H, α, β *cannot be computed are already nuisance-invariant.*

References

[1] S. Baker, S. K. Nayar, and H. Murase. Parametric feature detection. *IJCV*, 27(1):27–50, 1998.

[2] A. Berg and J. Malik. Geometric blur for template matching. In *Proc. CVPR*, 2001.

[3] F. L. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *Transactions on Pattern Analysis and Machine Intelligence*, 11(6):567–585, June 1989.

[4] J.-Y. Boubuet, M. Weber, and P. Perona. What do planar shadows tell about scene geometry? In *Proc. CVPR*, 1999.

[5] J. B. Burns, R. S. Weiss, and E. M. Riseman. View variation of point-set and line-segment features. *IEEE Trans. Pattern Anal. Mach. Intell.*, 15(1):51–68, 1993.

[6] E. Candès and D. L. Donoho. New tight frames of curvelets and optimal representations of objects with smooth singularities. Technical report, Stanford University, 2002.

[7] H. F. Chen, P. N. Belhumeur, and D. W. Jacobs. In search of illumination invariants. In *Proc. CVPR*, 2000.

[8] R. Cipolla, K. Åström, and P. J. Giblin. Motion from the frontier of curved surfaces. In *Proc. ICCV*, 1995.

[9] T. F. Cootes, G. J. Edwards, and C. J. Tayolr. Acrive appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6), June 2001.

[10] D.Chetverikov, Z.Megyesi, and Z.Jankó. Finding region correspondences for wide baseline stereo. In *Proc. ICPR 2004*, volume 4, pages 276–279, 2004.

[11] C. en Guo, S.-C. Zhu, and Y. N. Wu. Towards a mathematical theory of primal sketch and sketchability. In *Proc. ICCV*, page 1228, 2003.

[12] J. Enns and R. Rensink. Influence of scene-based properties on visual search. *Science*, 247:721–723, 1990.

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

[14] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *IEEE CVPR Workshop of Generative Model Based Vision*, 2004.

[15] V. Ferrari, T. Tuytelaars, and L. V. Gool. Wide-baseline multiple-view correspondences. In *Proc. CVPR*, volume 1, pages 718–725, June 2003.

[16] A. Fitzgibbon and A. Zisserman. Joint manifold distance: a new approach to appearance based clustering. In *Proc. CVPR*, page to appear, 2003.

[17] D. A. Forsyth, J. Haddon, and S. Ioffe. Finding objects by grouping primitives. In D. A. Forsyth, J. L. Mundy, V. D. Gesù, and R. Cipolla, editors, *Shape, contour and grouping in computer vision*. Springer-Verlag, 2000.

[18] U. Grenander and M. I. Miller. Representation of knowledge in complex systems. *J. Roy. Statist. Soc. Ser. B*, 56:549–603, 1994.

[19] L. Haglund and D. J. Fleet. Stable Estimation of Image Orientation. In *Proc. ICIP*, pages 68–72. IEEE, 1994.

[20] C. Harris and M. Stephens. A combined corner and edge detector. In *Proceedings of The Fourth Alvey Vision Conference*, pages 147–151, 1988.

[21] B. K. P. Horn. *Robot Vision*. MIT Press, Cambridge, Massachusetts, 1986.

[22] B. K. P. Horn and M. J. Brooks. *Shape from Shading*. MIT Press, Cambridge Massachusetts, 1989.

[23] K. Ikeuchi and B. K. P. Horn. Numerical shape from shading and occluding boundaries. *Journal of Artificial Intelleigence*, 1980.

[24] H. Jin, A. J. Yezzi, and S. Soatto. Region-based segmentation on evolving surfaces with application to 3d reconstruction of shape and piecewise constant radiance. In *European Conference on Computer Vision*, volume 2, pages 114–125, 2004.

[25] H. L. Jin, A. J. Yezzi, Y. H. Tsai, L. T. Cheng, and S. Soatto. Estimation of 3D surface shape and smooth radiance from 2D images: A level set approach. *J. Scientific Computing*, 19:267–292, 2003.

[26] W. H. Jin, S. Soatto, and A. J. Yezzi. Multi-view stereo beyond lambert. *Proc. CVPR*, 1:171–178, 2003.

[27] T. Kadir, A. Zisserman, and M. Brady. An affine invariant salient region detector. In *Proc. ECCV*, 2004.

[28] A. Kaplan, E. Rivlin, and I. Shimshoni. Robust feature matching across widely separated color images. In *Proc. CVPR*, 2004.

[29] D. G. Kendall. Shape manifolds, procrustean metrics and complex projective spaces. *Bullettin of the London Mathematical Society*, 16:81–121, 1984.

[30] S. Lazebnik, C. Schmid, and J. Ponce. Semi-local affine parts for object recognition. In *Proc. BMVC*, pages 959–968, 2004.

[31] T. Lindeberg. Feature detection with automatic scale selection. *IJCV*, 30(2):77–116, 1998.

[32] T. Lindeberg and J. Gading. Shape-adapted smoothing in estimation of 3-d depth cues from affine distortions of local 2-d brightness structure. *Springer-Verlag Lecture Notes in Computer Science*, 80o:389–400, 1996.

[33] D. G. Lowe. Distincive image features from scale-invariant keypoints. *IJCV*, 2(60):91–110, 2004.

[34] Y. Ma, S. Soatto, J. Kosecka, and S. S. Sastry. *An Invitation to 3-D Vision*. Springer Verlag, 2003.

[35] J. Matas, O. Chum, M.Urban, and T.Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *BMVC 2002*, 2002.

[36] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *IJCV*, 11(60):63–86, October 2004.

[37] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. V. Gool. A comparison of affine region detectors. *IJCV*, 1(60):63–86, 2004.

[38] M. I. Miller and L. Younes. Group actions, homeomorphisms, and matching: A general framework. *IJCV*, 1/2(41):61–84, December 2002.

[39] G. Mori, S. Belongie, and H. Malik. Shape contexts enable efficient retrieval of similar shapes. In *Proc. CVPR*, volume 1, 2001.

[40] D. Mumford and J. Shah. Optimal approximations by piecewiese smooth functions and associated variational problems. *Comm. Pure and Applied Mathematics*, 1989.

[41] S. K. Nayar, K. Ikeuchi, and T. Kanade. Surface reflection: Physical and geometrical perspectives. *PAMI*, 13(17):611–634, 1991.

[42] B.-T. Phong. Illumination for computer generated images. *Comm. ACM*, 6(18):311–317, June 1975.

[43] P. Pritchett and A. Zisserman. Wide baseline stereo matching. In *Proc. ICCV*, pages 754–760, 1998.

[44] R. Ramamoorthi and P. Hanrahan. On the relationship between radiance and irradiance: Determining the illumination from images of a convex lambertian object. *Journal of The Optical Society of America*, 2001.

[45] S. Savarese and P. Perona. Local analysis for 3D reconstructiuon of specular surfaces. In *Proc. CVPR*, 2001.

[46] F. Schaffalitzky and A. Zisserman. Viewpoint invariant texture matching and wide baseline stereo. In *International Conference on Computer Vision (ICCV'01)*, volume 2, July 2001.

[47] J. Shao. *Mathematical Statistics*. Springer Verlag, 1998.

[48] A. Stein and M. Hebert. Incoporating background invariance into feature-based object recognition. In *Seventh IEEE Workshop on Applications of Computer Vision (WACV)*, January 2005.

[49] K. E. Torrance and E. M. Sparrow. Theory for off-specular reflection from roughened surfaces. *Journal of Optical Society of America*, pages 1105–1114, 1967.

[50] K. Toyama and A. Blake. Probabilistic tracking with exemplars in a metric space. *IJCV*, 48(1):9–19, 2002.

[51] G. Ward. Measuring and modeling anisotropic reflection. *Computer Graphics*, 1992.

[52] A. J. Yezzi and S. Soatto. Deformotion: Deforming motion, shape average and the joint registration and approximation of structures in images. *Int. J. Comput. Vision*, 53(2):153–167, 2003.

[53] A. J. Yezzi and S. Soatto. Stereoscopic segmentation. *IJCV*, 2003.

[54] Y. Yu, P. Debevec, J. Malik, and T. Hawkins. Inverse global illumination: recovering reflectance models of real scenes from photographs. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 215–224. ACM Press/Addison-Wesley Publishing Co., 1999.

[55] A. L. Yuille, J. M. Coughlan, and S. Konishi. The kgbr viewpoint-lighting ambiguity. *Journal of The Optical Society of America*, 2002.

[56] T. Zickler, P. N. Belhumeur, and D. J. Kriegman. Helmholtz stereopsis: Exploiting reciprocity for surface reconstruction. In *Proc. ECCV*, 2002.