On Viewpoint Invariance for Non-Planar Scenes

Andrea Vedaldi Stefano Soatto

UCLA CSD Technical Report #TR050012

Abstract

Current local feature detectors/descriptors implicitly assume that the scene is (locally) planar, an assumption that is violated at surface discontinuities. We show that this restriction is, at least in theory, un-necessary, as one can construct local features that are viewpoint-invariant for generic non-planar scenes. However, we show that any such feature necessarily sacrifices shape information, in the sense of being non shape-discriminative. Finally, we show that if viewpoint is factored out as part of the matching process, rather than explicitly in the representation, then shape is discriminative indeed. We illustrate our theoretical results empirically by showing that, even for simplistic scenes, current affine descriptors fail where even a naive 3-D viewpoint invariant succeeds in matching.

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: when do two or more images "belong to the same class"? A class reflects some commonality among scenes being portrayed [12, 15, 22, **?**]. 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? 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, as summarized in Sect. 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 the purpose of this introduction 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 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, occlusions, 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 using features." But what is a *feature*? and why do we need it? We will address these questions in Sect. 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, making this notion of correspondence meaningless in lack of additional assumptions. Probabilistic assumptions (e.g. priors) require endowing shape and reflectance with probability measures, not an easy feat. Therefore, we choose to make *physical assumptions* that allow us to give a meaningful answer to the correspondence problem. This problem naturally relates to wide-baseline matching [31, 13, 13, 10, 20].

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

1.2. Lambertian scenes in ambient light

While global correspondence can be computed for scenes with complex reflectance under suitable assumptions, 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 [8]. In particular, one can easily verify that if the illumination is assumed to be constant (ambient, or "diffuse") then local correspondence can be established. We therefore adopt such assumptions and relegate all non-Lambertian effects as "disturbances."

We can now make the formal notation above more precise: We represent an image as an array of positive numbers: $I : D \subset \mathbb{R}^2 \longrightarrow \mathbb{R}_+;\ x \mapsto I(x)$. 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 the case of ambient illumination, to first approximation[2] we have a global scaling α_t and an offset β_t. 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 "disturbance" 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), which we will address in Sect. 2.2. Eq. (2) is reminiscent of deformable templates [35, 9, 16], although here we do not know the template ρ.

1.3. State of the art and our contributions

The non-existence of general-case view invariants [6] has often been used to motivate local descriptors, for instance affine invariants. The results of [6], however, pertain to collections of points in 3-D space with no photometric signature associated to them. When one measures image intensity, on the other hand, we show that *viewpoint invariance can be achieved for scenes with arbitrary (continuous) shape, regardless of their albedo,* under suitable conditions which we outline in Sect. 2.1. While this result seems obvious in the aftermath, and by no means undermines the importance of affine descriptors, we believe it is important to state it precisely and prove it for the record, which we do in Theorem 1. The flip-side of general-case viewpoint invariants is that *they necessarily sacrifice shape information,* and therefore discrimination has to occur based solely on the photometric signature (Sect. 2.1). This result is straightforward to prove (Theorem 2), but since nobody has done so before, we believe it is important. It also explains the empirical success of "bags of features" in handling viewpoint variations [**?**]. Finally, we show that if viewpoint is factored out as part of the matching process, rather than in the representation, then *shape information is retained,* and can be used for discrimination. This may contribute to the discussion following [30] in the psycho-physical community. On illumination invariants, [8] showed that even for Lambertian scenes they do not exist. While they used a point light source model, diffuse illumination is perhaps a more germane assumption for cloudy days or indoor scenes, due to inter-reflections [21]. As we show, invariance to such a first-order model of Lambertian reflection in diffuse illumination can be easily achieved with our approach (Sect. 2.3). In deriving our results we lay out a general framework for designing detector/descriptor pairs that allows for *comparison of existing algorithms on analytical grounds,* in addition to experimental [28]. Our approach is reminiscent of [2], although more general. For the benefit of the reader that is unappreciative of theory alone, we illustrate our results with simple experiments that show that even a naive 3-D viewpoint invariant can support matching whereas current affine descriptors fail (Sect. 3). Of course, existing descriptors only fail at discontinuities, so our work serves to validate existing methods where appropriate, and to complement them where their applicability is limited. The point of this section is *not* to advocate use of our detector/descriptor as a replacement of existing ones. It only serves to illustrate the theory, and to point out that some of the restrictions imposed on existing methods may be un-necessary. 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. More discussion and a list of common objections to our theory can be found in Sect. 4.

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

[2] More precisely, the radiance at $p \in S$ is given by $R(p) \doteq \rho(p) \int_{V_p} \langle \nu_p, \lambda \rangle dA(\lambda)$ where ν_p is the normal and V_p the visibility cone at p and dA is the area form on the light source [21]. We coarsely approximate this model with a global affine transformation in (2).

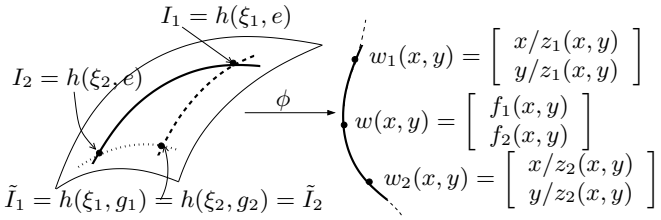

Figure 1: *Viewpoint invariant features sacrifice shape information by collapsing all homeomorphic closures of allowable warps onto a single equivalence class (see text for explanation).*

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* 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:[3] $\xi_1 \neq \xi_2 \Rightarrow \phi \circ h(\xi_1, \mu) \neq \phi \circ h(\xi_2, \nu)\ \forall\ \mu, \nu$. 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 this section we address viewpoint invariance, and therefore assume $\alpha_t = 1$ and $\beta_t = 0\ \forall\ t$ in eq. (2) until Sect. 2.3; we also assume no self-occlusions until Sect. 2.2, and therefore parametrize the surface S as $\Omega \subset \mathbb{R}^2 \to \mathbb{R}^3,\ x \mapsto S(x) = [x^T\ z(x)]^T$ for some choice of local coordinates, for instance $x = \pi(p)$. Since both ρ and S are unknown, and we only measure their composition through $I_t(x_t) = \rho \circ S(x)$, we rename the function $\rho \doteq \rho \circ S$. Similarly, we call $w_t \doteq \pi \circ g_t \circ S : \Omega \subset \mathbb{R}^2 \to \mathbb{R}^2$ 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 "disturbance" term n_t since that only affects the inference technique, not the general modeling paradigm and invariance considerations. Under these assumptions, w_t are homeomorphisms; as such, they induce a partition of the set of images $I(x)$ into 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. In particular, $\phi(I) = \{\rho(x),\ x \in \Omega\}$ is the maximal invariant (in the sense of inclusion, see Appendix A , uploaded as supplementary material, for details). This is a sketch of the argument that proves the following result.

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

The actual proof is *constructive,* and forms the basis for the design of general viewpoint invariant descriptors. Since it is somewhat technical we report it in Appendix A (uploaded). In Sect. 2.3 we show a simplified version of this construction. The restriction of this theorem to planar scenes is straightforward and forms the basis of the motivation behind affine invariant descriptors [27, 19, 32]. The claim is also latent in [4, 13, 26] for more general transformations, although to the best of our knowledge it has never been stated explicitly nor proven before.

Ideally one would like an invariant descriptor to be a "signature" of the scene, i.e. different scenes should result in different descriptors. Alas this cannot happen, as any viewpoint invariant necessarily sacrifices shape information. This is illustrated in Fig. 1 and proven below. The theorem is true for any viewpoint invariant, not just those that satisfy the sufficient conditions of Theorem 1, although we will adopt such assumptions for simplicity.

Theorem 2 (... but are not shape-discriminant)**.** *Under the hypotheses of Theorem 1, given a viewpoint invariant feature ϕ, for any scene with shape S_1 that yields an image I_1 there exists a scene with shape $S_2 \neq S_1$ that yields an image $I_2 \neq I_1$ such that $\phi(I_2) = \phi(I_1)$.*

[3] 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 [8].

Proof (sketch). Let a scene ξ_1 have surface $S_1(x)$ parametrized as the graph of a function, which is possible in the absence of self-occlusions, and similarly for a scene ξ_2. Let I_1 be the image generated by S_1 for some albedo ρ_1, and let $\tilde{I}_1$ be the image generated by the same scene under a different viewpoint, specified by g_1: $\tilde{I}_1 = h(\xi_1, g_1)$. Note that by assumption we have $\phi(I_1) = \phi(\tilde{I}_1)$. Now select any surface $S_2(x) \neq S_1(x)$ that is not occluded from both viewpoints that generated I_1 and $\tilde{I}_1$, and back-project the image $\tilde{I}_1$ onto S_2, to generate $\tilde{I}_2$. Then we have $I_1(x) = \tilde{I}_1(\pi g_1 S_1(x)) \doteq \tilde{I}_1(\tilde{x}) = \tilde{I}_2(\tilde{x}) = \tilde{I}_2(\pi S_2(\tilde{x}))$. Trivially, since $\tilde{I}_1 = \tilde{I}_2$, we have $\phi(\tilde{I}_1) = \phi(\tilde{I}_2)$. Now take an image of the scene ξ_2 from a different vantage point g_2 to get $I_2(x) \doteq \tilde{I}_2(\pi g_2 S_2(x))$. Unless albedo is constant, in general $I_2 \neq I_1$, while $\phi(I_2) = \phi(\tilde{I}_2) = \phi(\tilde{I}_1) = \phi(I_1)$. □

While the proof is simple, the claim is powerful because it shows that if we want to be viewpoint invariant, we have to "throw away" shape information. This does not mean that viewpoint invariant features are useless! In fact, scenes with different albedo yield different invariant descriptors, that are albedo-discriminative. The theorem suggests that approaches that do away with restrictive geometric variations in the configuration of feature descriptors in favor of loser topological requirements [**?**] or coarse quantization [3] of feature positions may be more robust to extreme viewpoint variations.

Also, the theorem does not imply that we cannot recognize objects that have different shape but the same albedo! Indeed, consider two scenes with different shape but identical albedo, e.g. $\rho = \text{const.}$, each generating an image I_1 and I_2, for instance the geometric structures of [30]. Now, given a new image $\tilde{I}$, we want to decide whether $\tilde{I}$ comes from ξ_1 or ξ_2. While this is not possible with a viewpoint invariant feature (as we show in App. A and briefly sketch below we can construct general homeomorphisms that make their feature coincide $\phi(I_1) = \phi(\tilde{I}) = \phi(I_2)$), it is still possible if the viewpoint is marginalized as part of the matching process ("recognition via reconstruction"): $\tilde{I} \leftrightarrow I_j \Leftrightarrow \exists\, \xi_j, g_j, \rho, \tilde{g} \mid \tilde{I} = h(\xi_j, \tilde{g}) = h(\xi_j, g_j)$. In this idealized case with no disturbance or uncertainty, the test will succeed for either $j = 1$ or $j = 2$. This argument can be used to prove the following:

Theorem 3 (Matching shapes)**.** *Discriminating scenes made of continuous surfaces with different shape but identical albedo from images that yield no self-occlusions can only be performed by estimation of the viewpoint as part of the matching process.*

Note that the estimate of the viewpoint may not be unique, as long as it yields a valid viewpoint-induced warp, as opposed to a general one. A pictorial illustration of this phenomenon is shown in Fig. 1. While the transformation from I_1 to $\tilde{I}_1$ can be performed by changes of viewpoint alone (but not shape), and the same for the transformation from I_2 to $\tilde{I}_2$, the composition of the two requires a change in shape. In other words, while the dotted orbit $w_1(x) = \pi(g[x^T, z_1(x)]^T)$ and the dashed one $w_2(x) = \pi(g[x^T, z_2(x)]^T)$ can be implemented by a change in viewpoint, their composition (solid line) cannot, and is instead a more general 2-D homeomorphism $w(x) = [f_1(x), f_2(x)]^T$. However, the computation of the feature collapses these three orbits onto the same equivalence class, making it impossible to distinguish warps that are due to changes in viewpoint (such as w_1, w_2) and those that are more general (such as w) (see App. A).

To avoid confusion, note that here "shape" means the 3-D geometry of the scene S. If we have, say, a planar contour, which we can view as a binary image ρ, we can build a viewpoint invariant descriptor (e.g. [3]) that can be legitimately used to recognize shape without searching for viewpoint during the matching procedure. Note, however, that the descriptor is *albedo-discriminative,* and it is only accidental that the albedo is used to represent (2-D) shape. Similarly note that the scene here includes everything visible, so the theorem does not apply to cases where the occluding boundary provides discriminative features, say to recognize a white sphere from a white cube on a black background. Finally, note that the notion of viewpoint can be generalized to an equivalence class under the action of a group, for instance the 3-D projective group, so that no *explicit* reconstruction is necessary during the matching phase.

2.2. Why features?

Before we marry to the notion of feature it is useful to recall Rao-Blackwell's theorem ([33], page 87) that, 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.[4] Unfortunately, Theorem 2 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.

[4] "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.

The assumptions to prove Theorem 1 require no visibility artifacts, such as self-occlusions 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. [5] This motivates relaxing 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: $I_1 \leftrightarrow I_2 \Leftrightarrow \exists\, \Omega \subset D, \xi \mid \forall\, x \in \Omega : I_1(x) = h(\xi(x), \nu_1),\ I_2(x) = h(\xi(x), \nu_2)$.

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 Sect. 2.3.

2.3. Invariance by canonization

From (3) we can easily infer that $\{\rho(x),\ x \in \Omega\}$ is the "ideal" invariant feature, in the sense that any other invariant feature is a function of it. Of course, we do not know ρ nor Ω. In App. A we will construct the maximal feature explicitly; here we derive a simplified version of the argument that is more intuitive. We start by expressing what we have in terms of what we want: $I_t(x_t) = \rho(w_t^{-1}(x_t)),\ x_t \in w_t(\Omega)$. It is obvious that if we take any homeomorphism $v : \mathbb{R}^2 \to \mathbb{R}^2$ and we 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: $[\phi] \doteq \{\rho(v(x)),\ x \in v^{-1}(\Omega),\ v : \mathbb{R}^2 \to \mathbb{R}^2 \text{ a homeo}\}$. Now we have two options to proceed. One is to define a distance between equivalence classes, $d([\phi_1], [\phi_2])$, that requires marginalizing the nuisance as part of the correspondence process, what we called "recognizion by reconstruction" earlier. The alternative is to identify, for each equivalence class, a *canonical representative*, that is a unique element of the class, $\hat{\phi}$, corresponding to a choice of v, and then define a distance between feature elements, $d(\hat{\phi}_1, \hat{\phi}_2)$. A choice of canonical element $\hat{\phi}$ in the equivalence class $[\phi]$ must be determined uniquely from the available data, that is $I_t(x_t),\ x_t \in D$.

Feature detectors. Based on the discussion above, a detector is a contra-variant functional $F_i,\ i = 1, 2, \ldots$, such that $F_i([\phi]) = F_i(x, v, \Omega) = e_i$ uniquely determines $\hat{v}$, and therefore $\hat{\phi}$. Without loss of generality[6], 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. One can derive most existing detectors by changing the functional or the norm, second order moments [24, 27], edge/intesity [13], saliency [19], level set-based regions [26, 1], affine homogeneous-texture regions [32].

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*. This indicates that the local structure of the image around a point can be used to determine a local "natural" frame.[7]

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 [29, 14].

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.

Introducing illumination into the model does not modify the scheme just outlined for the simple case of ambient illumination and Lambertian reflection. In fact, to first-order, 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, e. g. $\hat{\beta}_t = \int_{\hat{\Omega}} \hat{\rho}(x)dx$ 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.

[5] Of course one can attempt to characterize clutter probabilistically, but this is well beyond our scope here.

[6] Note that all existing detectors assume Ω is given (e.g. a unit circle), and estimate the adapted region $w(\Omega)$ (e.g. an ellipse) from the transformation.

[7] In particular, *translation invariance* $g \in \mathbb{R}^2$ [18], *scale invariance* $g \in \mathbb{R}^3$ [23]; *Euclidean invariance* ($g \in SE(2)$) and *Similarity invariance* ($g \in SE(2) \times \mathbb{R}$) [25]; *Affine invariance* $g \in \mathbb{A}(2)$ [28] are all well-known. *Viewpoint invariance for generic shape* ($g \in SE(3)$) requires fixing a homeomorphism tailored to the local structure of the image, e.g. a thin-plate splines [5, 4] (not a group, however) polynomials (tricky numerics), [7] or local histograms (e.g. polar orientation histograms) to semi-global representations such as the sketch [11]. In Sect. 3 we illustrate this case with a piecewise affine deformation model.

Remark 1 (Non-localized frames)**.** *All the detectors based on the invariance properties just outlined allow one to determine a* localized frame*, called a co-variant local frame,*[8] *that has a well-defined origin, hence the early nomenclature "feature point" although 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 symmetry subgroup *which could be either continuous or discrete. In this case, one can associate the descriptor to any point on the equivalence class determined by the subgroup ambiguity: for instance, for an* edge in space *($g \in SE(3)/SE(2)$) on can fix the gradient direction and scale, and similarly for an* edge on the image *($g \in SE(2)/\mathbb{R}$), a special case of the former when it is not possible to reliably associate a scale to the edge; in* homogeneous periodic textures *($SE(3)/\mathbb{Z}^2$) the intensity profile is isotropically periodic (possibly after warping or normalization), and so on.*

Remark 2 (Segmentation as a detector)**.** *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 and any statistic computed from such a profile is a valid descriptor. Unlike the localized frame, the detector here does not contain any shape information (neither does the descriptor, in both cases), and is realized by a* segmentation procedure *that extends the domain of the descriptor Ω to include all points that have common statistics. Therefore, our approach gives theoretical grounds to segmentation beyond simple computational considerations.*

3. A case study: 3-D corners

To illustrate the analytical results we explicitly construct a viewpoint invariant descriptor for 3-D scenes. Our goal here is not to propose yet another detector/descriptor to replace existing ones. Rather, we illustrate how their limitations can be overcome. We focus on singular points of the surface $S(x)$ that cannot be locally approximated by a plane (Figure 2), hence defying current affine descriptors. We model a corner as a vertex with n planar faces that, barring occlusions, produces an image with n angular sectors and a center x_0, projection of the vertex. These are separated by edges, which we represent as vectors $v_i \in \mathbb{R}^2,\ i = 1, ..., n$, their lengths being scale parameters. When the viewpoint changes the n sectors are transformed by homographies, which we approximate with affine warps. This model locally captures the true transformation to an arbitrary degree, unlike a single affine transformation that current descriptors are based on. 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 mappings $x_0 \mapsto y_0\ v_i \mapsto u_i,\ i = 1, ..., n$.[9]

Detection: While there exist many possible procedures for detecting corners in images [**?**], including sketch primitives [11] or matched filters [17], our emphasis here is on how to arrive at a viewpoint invariant once a structure has been detected. Therefore, we choose a simple if not somewhat naive detector that yields directly the corner structure.[10]

Canonization: Once a frame is detected we map it to a canonical configuration that avoids singular transformations[11]. This step fixes the canonical frame up to a rotation, which can be partially eliminated by requiring that one edge maps to $(1, 0)$, which leaves us with a discrete subgroup that can be further resolved with radiance information.[12]

Descriptor: Although the canonized features could be compared directly (e.g. by NCC), we compute a descriptor for each detected feature. This has the advantage of makeing the comparison faster, absorbing differences in the normalized features due to imprecise detections or unsatisfied assumptions (e.g. the surface is not Lambertian), and illustrates how our approach complements, rather than replaces, existing descriptors. 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 [25], one of the most widely used [25, 28]. 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: Many corners are found on the occluding boundaries [34], and some sectors χ_i may belong to the background. We therefore compute multiple descriptors, one per possible assignment of the faces to the foreground or

[8] Although contra-variant would be a more appropriate name (Sect. 2.3).

[9] Formally, let $\{\chi_i(x),\ i = 1, ..., n\}$ be a partition of $\mathbb{R}^2$ in n angular sectors, $\chi_i(x)$ the indicator function of the i-th sector. We call *piecewise affine transformation* (PWA) 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 $A_i \in GL(2),\ i = 1, ..., n$ are chosen so that $w(x)$ is continuous.

[10] A set of putative corners $X = \{x_1, ..., x_n\}$ is extracted [18] and edges checked for each pair $(x_i, x_j) \in X^2$ using a parametric template $T(x, y; w) = \text{sign}(y), \quad (x, y) \in [0, 1] \times [-w, w]$ reminiscent of [2], via normalized cross correlation (NCC). A reference frame is then attache to each point $x_0 \in X$ and all edges connected to x_0 are detected, localized (using an extension of the edge model with explicit terminations), refined and clustered via vector quantization.

[11] This can be achieved by enforcing the following conditions: (i) if all sectors are less than π, the normalized frame has n equall sectors; (ii) if one of the sectors is wider than π we make this sector $3\pi/4$ and fit the others in the remaining $\pi/2$ radians; (iii) if one sector is exactly π (e.g. a T-junction), we delete one edge and reduce to the former case.

[12] If the corner has a sector wider than π, we use this to uniquely identify an edge and eliminate the ambiguity, since 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.

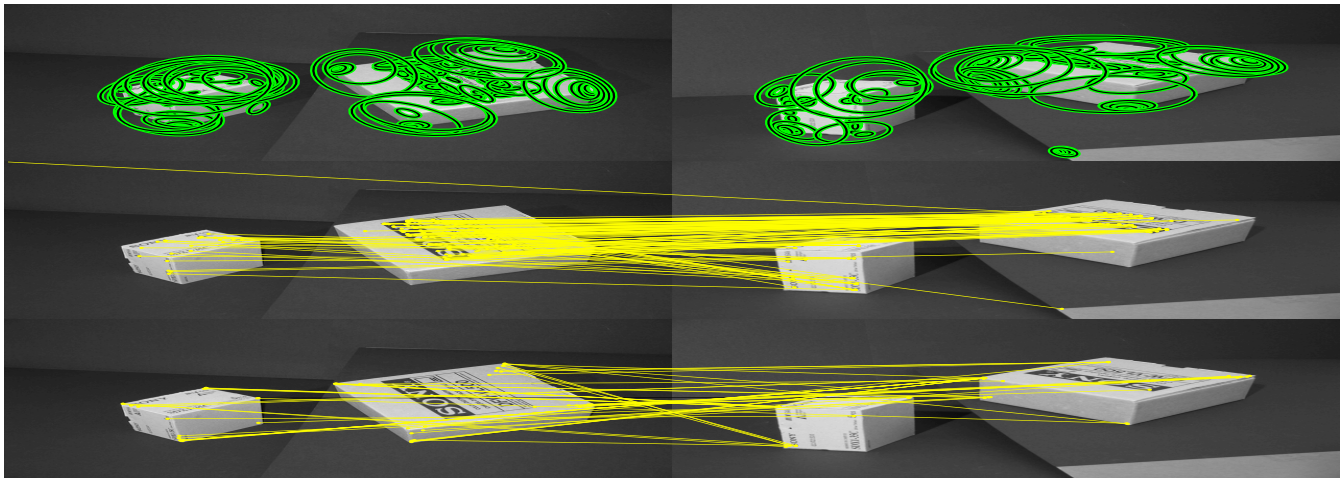

Figure 2: **Affine-invariant descriptors fail to capture non-planar structures:** *(top) two images of the same scene with detected regions; (middle and bottom) correspondence established using affine invariant signatures on the planar (middle) and non-planar (bottom) regions of the scene. While several non-planar regions are detected, they are mismatched because of the large discrepancy in the corresponding descriptor, caused by the non-planar structure of the scene.*

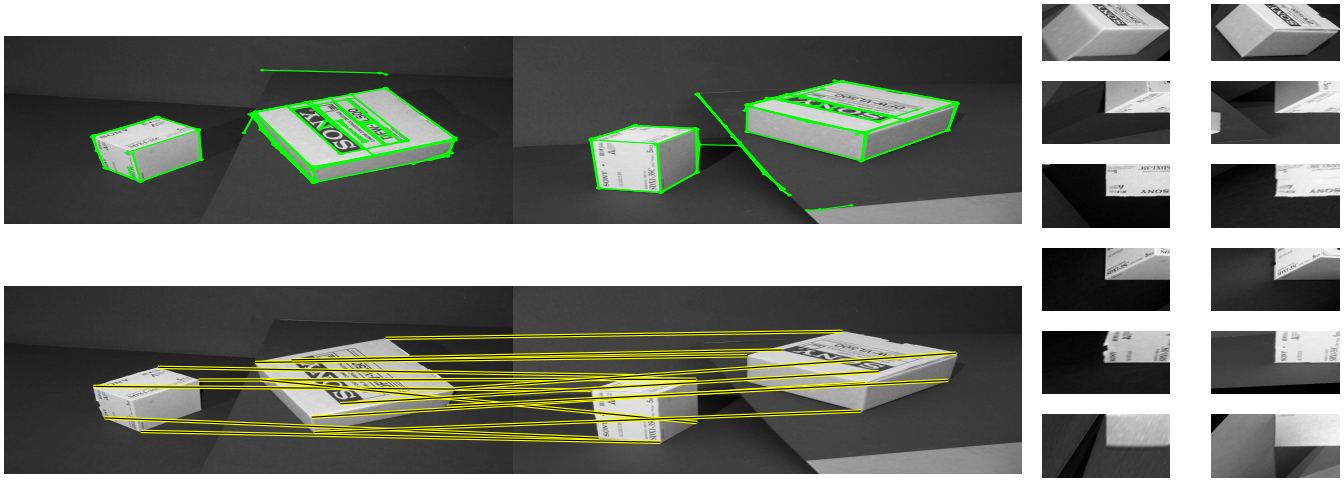

Figure 3: **General viewpoint invariants can match 3-D corners:** *(top) detected reference frames; (bottom) matched "3-D features"; (right) examples of canonized features. Most of the "3-D features" that are detected but mismatched by affine-invariant descriptors are correctly matched by a more general viewpoint-invariant.*

the background. [13]

Experiments: We choose H-A [27] as representative of affine invariant detectors/descriptors. Figure 2 shows that of 186 features detected in the first image, 53 are successfully matched in the second, 68 are mismatched because of the descriptor variability and 65 are not matched because the detector fails to select them in the second image. Figure 3 shows that even the naive 3-D descriptor introduced can match most 3-D corners. There is just one mismatch, due to the almost identical appearance of the last two feature pairs in Figure 3, and two missing corners, which are not extracted by the Harris detector in the first stage. An exact comparison with affine-invariant detectors is difficult because the latter find several times the same structures; roughly speaking, however, 70% of the mismatches of the affine detector are fixed by the "3-D corner" model.[14] Note that no direct comparison with 3-D viewpoint invariants is possible since, to the best of our knowledge, there are none in the literature. In the second experiment we test a scene presenting a variety of 3-D corners. Figure 4 shows the detected frames and the matching pairs: One third of the features in the first image are correctly matched in the second. In this, the performance is similar to that of the H-A detector on the planar structures of Figure 2, but in our case for non-planar structures. In the

[13] In practice, the most common cases (objects with convex corners) are covered if we do so only for sectors larger than π, thereby obtaining no more than two descriptors for each detected feature.

[14] 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.

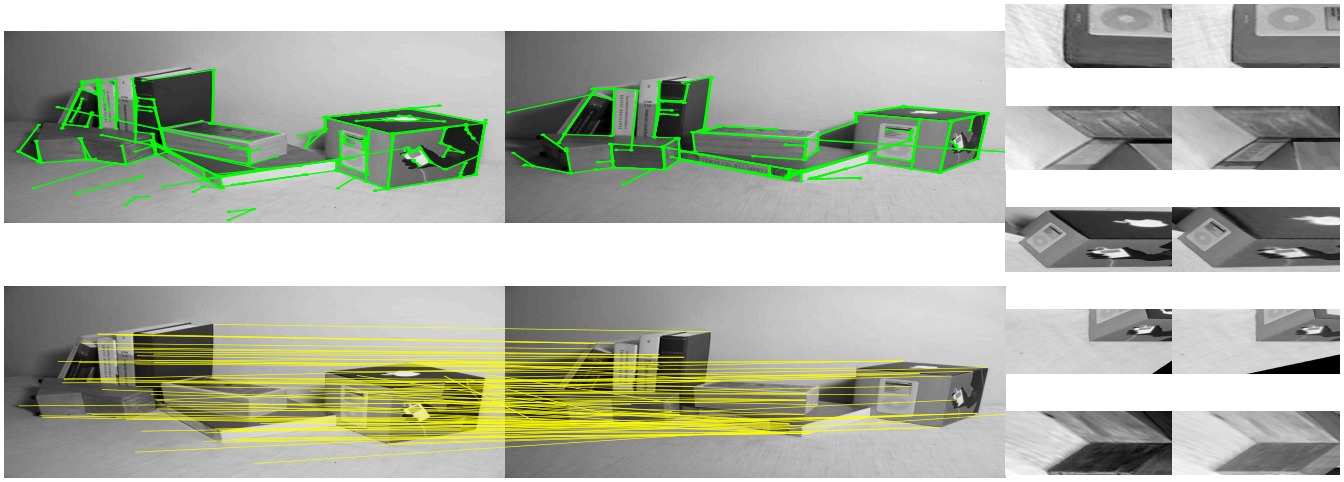

Figure 4: **Matching example:** *(top) all the features detected in the first image are connected to their nearest neighbor descriptor in the second image (bottom); (right) a variety of normalized features. Of 93 detected features, 32 are present and correctly matched in the second image.*

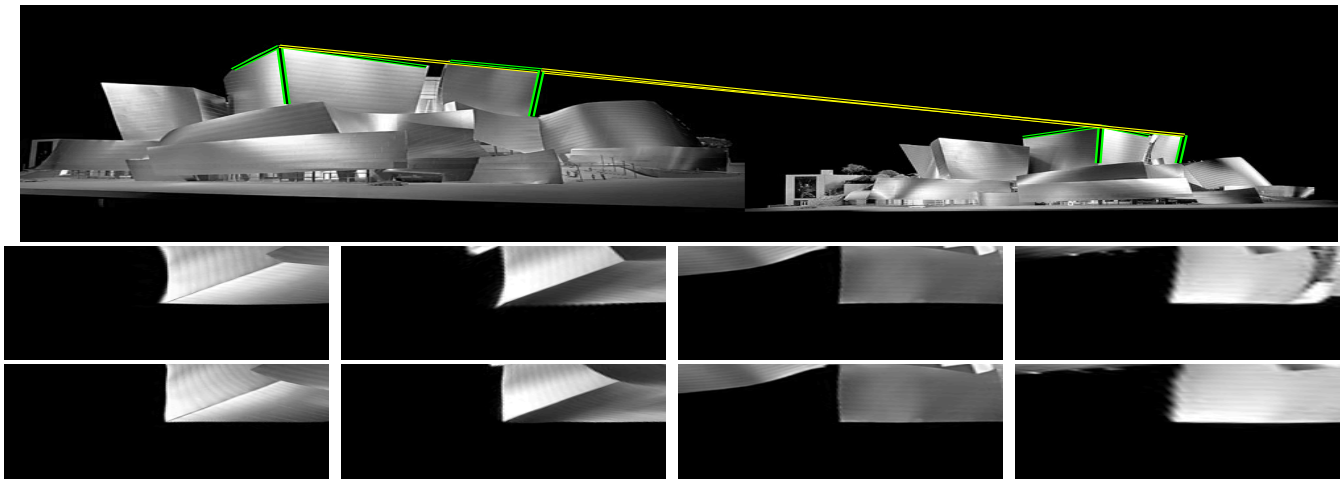

Figure 5: **Matching Gehry:** *(top) two corners matched by our method; (middle) features canonized by a piecewise-affine transformation; (bottom) features canonized by a thin-plate spline transformation. Although the scene does not meet most of our working assumptions, a few corners are still matched (see also supplementary material).*

last experiment (Figure 5) we test our method on a scene where several of our working hypotheses are not verified because of highly non-planar, non-Lambertian surfaces. The figure shows two corners that our method is able to match, together with the corresponding canonized features. The canonized features are similar enough to be matched using SIFT, illustrating the importance of viewpoint canonization. We also show the same two corners canonized using a thin-plate spline, estimated by 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.

4. Discussion

Our contributions in this manuscript are mainly theoretical: We clarify some misunderstandings that are lingering in the literature, where affine-invariant detectors/descriptors are often motivated by the non-existence of general-case viewpoint invariants following [6]. Our results do not imply that affine-invariant descriptors are not useful. On the contrary, they may very well be the way to go, but we believe it is important that their motivations be clear and that overly restrictive assumptions are not imposed. Furthermore, by showing that viewpoint invariants are not shape-discriminative we validate "bags of features" approaches to recognition (see [**?**] and references therein), where spatial relations among features (i.e. shape) are either discarded or severely quantized or "blurred" [3, 4]. Finally, we show that if instead of using a feature-based approach one factors out viewpoint as part of the matching process, then shape is discriminative indeed. This, however, requires (explicit or implicit)

optimization with respect to the viewpoint, which may help explain some of the psycho-physical results following [30], where albedo is non-discriminative and therefore shape is the only "feature." 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, representation for recognition: We have shown that under certain conditions viewpoint and illumination-invariant features can be constructed explicitly. Our framework allows comparison of existing methods and opens the way to design richer classes 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.

References

[1] L. Alvarez and J. M. Morel. Morphological approach to multiscale analysis: From principles to equations. In *In B. M. ter Haar Romeny (ed.), Geometric–Driven Diffusion in Computer Vision*, 1994.

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

[3] S. Belongie, J. Malik, and J. Puzicha. Matching shapes. In *Proc. of the IEEE Intl. Conf. on Computer Vision*, 2001.

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

[5] 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.

[6] J. B. Burns, R. S. Weiss, and E. M. Riseman. The non-existence of general-case view-invariants. In *Geometric Invariance in Computer Vision*, pages 120–131, 1992.

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

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

[9] T. F. Cootes, G. J. Edwards, and C. J. Tayolr. Active 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] 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.

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

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

[15] 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.

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

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

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

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

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

[21] M. Langer and S. Zucker. Shape from shading on a cloudy day. *J. Opt. Soc. Am. A*, 11(2):467–478, 1994.

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

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

[24] T. Lindeberg and J. Garding. 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.

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

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

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

[28] 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.

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

[30] R. N.Shepard and J. Metzler. Mental rotation of three-dimensional objects. *Science*, 171:701–703, 1971.

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

[32] 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.

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

[34] 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.

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