Viewpoint Induced Deformation Statistics and the Design of Viewpoint Invariant Features: Singularities and Occlusions

Andrea Vedaldi and Stefano Soatto

University of California at Los Angeles, Los Angeles – CA 90095
{vedaldi, soatto}@cs.ucla.edu

Abstract. We study the set of domain deformations induced on images of three-dimensional scenes by changes of the vantage point. We parametrize such deformations and derive empirical statistics on the parameters, that show a kurtotic behavior similar to that of natural image and range statistics. Such a behavior would suggest that most deformations are locally smooth, and therefore could be captured by simple parametric maps, such as affine ones. However, we show that deformations induced by singularities and occluding boundaries, although rare, are highly salient, thus warranting the development of dedicated descriptors. We therefore illustrate the development of viewpoint invariant descriptors for singularities, as well as for occluding boundaries. We test their performance on scenes where the current state of the art based on affine-invariant region descriptors fail to establish correspondence, highlighting the features and shortcomings of our approach.

1 Introduction

This work is concerned with the design of viewpoint-invariant discriminative local features, *i.e.* local image statistics whose dependency on viewpoint can be made arbitrarily small while maintaining non-trivial dependency on other properties of the scene, such as its shape and reflectance. This problem has been largely solved under the assumption that the scene is locally planar, Lambertian, viewed under ambient illumination and moderate changes in viewpoint. Under these conditions, local deformations can be approximated by a similarity or affine transformation, and the resulting local invariant features (see [1, 2, 3, 4, 5, 6, 7, 8] and references therein) have proven to be a powerful tool in the recognition of individual objects as well as object categories [9, 10, 11, 12, 13, 14, 15, 4, 1, 16, 17]. But *what happens when such conditions are not satisfied?*

Changes of illumination can have minor or drastic effects depending on the reflectance of the scene [18] and we will not address them here; we will continue to assume that the scene is Lambertian and viewed in ambient light, leaving illumination out of the scope of this work. Drastic changes in viewpoint could be handled by concatenations of small changes if intermediate views are available [11, 14]. However, we will not make that assumption and allow large viewpoint changes which can induce visibility artifacts such as occlusions. The local planarity assumption is violated in two cases: At singularities (e.g. ridges or corners),

A. Leonardis, H. Bischof, and A. Pinz (Eds.): ECCV 2006, Part II, LNCS 3952, pp. 360–373, 2006.

and at occluding boundaries. Here the assumption of affine deformation is violated in a neighborhood of any size, and similarity/affine invariants often (but not always) fail.[1] But *how important are singularities and occlusions? How much weight do they carry in the recognition process?* We will show that singularities and occluding boundaries are few compared to interior regular points, but they carry significant weight in that they often correspond to photometrically salient regions (as also shown indirectly by [18], Sect. 5).

Now, assuming that we agree that singular regions and occlusions are important, *can we characterize the deformations they induce on the image under changes in viewpoint? Can we exploit this knowledge to design viewpoint-invariant features for such intrinsically non-planar portions of the scene?*

As we will show, in order to design a viewpoint invariant feature for singularities and occlusions we need to attach a curvilinear (or multi-linear) local frame to the image. This is still an open area of research, which we cannot address in the limited scope of this paper. We will therefore tap onto existing techniques that allow the extraction of some discrete representation (a graph) from local analysis of the image, such as [19, 10, 12] and their variants. We will discuss their role and their limitations in generating viewpoint invariants.

1.1 State of the Art

The literature on feature extraction is too extensive to review in the limited scope of a conference paper. The reader is encouraged to consult [20] and references therein. At one end of the spectrum of work on on feature-based recognition are simple parametric deformations, e.g. affine transformations yielding a procrustean density on feature constellations (see [21] and references therein). At the opposite end are "bags of features" that retain only feature labels regardless of their mutual position (see [22, 23, 24] and references therein). Viewpoint changes induce transformations more general than affine, but far less general than an arbitrary scrambling of feature positions. Our work concentrates on the case in between, following the steps of [25, 4, 26, 27, 28].[2] More specifically, [29, 30] have proposed region descriptors for salient regions detected at or near occluding boundaries. While feature selection is traditionally addressed as a representation issue, different from the final goal of recognition, the two processes are beginning to come together [31, 32, 33]. Since viewpoint variations (under the assumptions discussed) only induce changes in the domain of the image, this work generically relates to deformable templates [34] and deformable models [35]. Our attempt to characterize the "natural deformation statistics" follows the lead of [36, 37] and others that have characterized natural image and range statistics.[3] Specific relationships with other work will be pointed out throughout the manuscript.

[1] Part of the art of designing a descriptor is to give it slack to absorb violations of the underlying assumptions.

[2] Even work that allows arbitrary reordering of features relies on individual features being matched across views, and therefore the affine model restricts this approach beyond its ideal generality.

[3] Including [38] that has appeared while this manuscript was under review.

1.2 Notation and Formalization of the Problem

An image is a function $I : \Lambda \to \mathbb{R}^+;\ x \mapsto I(x)$ with local domain $\Lambda \subset \mathbb{R}^2$ and range in the positive reals. Under the assumptions of Sect. 1, the value of the image at a pixel x is approximately equal to the radiance ρ of the scene at a point p on a surface $S \subset \mathbb{R}^3$, $I(x) = \rho(p),\ p \in S$. In fixed coordinates, p projects onto $x = \pi(g_0 p)$ where $\pi : \mathbb{R}^3 \to \mathbb{P}^2$ is a canonical perspective projection and $g_0 \in SE(3)$ is the position and orientation of the camera. We say that x is the *image* of p, and p is the *pre-image* of x. These notions extend to sets; for instance, the pre-image of a ball of radius σ around x_0, $\mathcal{B}_\sigma(x_0)$, is the set $\{p \in S : \pi(g_0 p) \in \mathcal{B}_\sigma(x_0)\}$. If we consider multiple images of the same scene under changing viewpoint, we can choose one of the camera reference frames as the fixed frame, and parameterize the surface S relative to it. Then, with an abuse of notation, we can write $p = S(x)$ and we have that the generic image is given by

$$\begin{cases} I(\hat{x}) = \rho(S(x)) \\ \hat{x} = \pi(g_t S(x)) \doteq w(x), \qquad x \in \Omega. \end{cases} \tag{1}$$

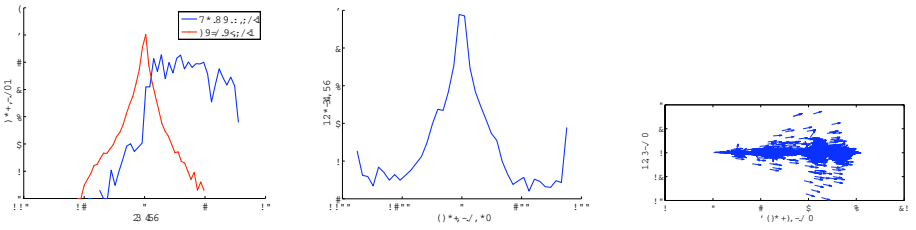
Can we characterize the structure and statistics of the function $w : \Omega \subset \mathbb{R}^2 \to \mathbb{R}^2$*? Can we use it to design viewpoint invariant features?*

2 Natural Warping Statistics

The structure of $w : \Omega \to \mathbb{R}^2$ obviously depends on the structure of S. We distinguish between three classes of points : x_0 is an *interior regular point* (**IR**) if there exists an σ and a neighborhood $\mathcal{B}_\sigma(x_0)$ whose pre-image $S(\mathcal{B}_\sigma(x))$ is simply connected and smooth. x_0 is an *interior singular point* (**IS**) if its pre-image is a $C^1(\mathcal{B}_\sigma(x_0))$ discontinuity, *i.e.* the scene is continuous around the pre-image of x_0 but not differentiable on it. An IS point can be the image of a *wedge* (the locus of singularities is a one-dimensional submanifold of S, locally approximated by a line), or an image of a *corner* (the locus of singularities is the tip of a generalized cone). Finally, x_0 is an *occluding boundary* (**OB**) if the pre-image of any neighborhood $\mathcal{B}_\sigma(x_0)$ is not simply connected, for any choice of σ. In this case, the occluding boundary could correspond to a singularity (OBS), as is the case for polyhedra, or it could correspond to regular points on S (OBR), as is the case for the silhouette of a smooth surface . Note that viewpoint variations can change the labeling of points. For instance, an IR point can become OB and vice-versa. However, if a point is IR there will always exist a neighborhood and a set of viewpoints (depending on σ) such that the point remains IR.

2.1 Deformation Statistics Around Interior Points

The goal here is to determine the distribution of the homeomorphism $w : \Omega \subset \mathbb{R}^2 \to \mathbb{R}^2;\ x \mapsto \pi(gS(x))$ defined in (1). In order to make the notation explicit we write $\Omega \doteq x_0 + \mathcal{B}_\sigma$, with x_0 a location on the image domain and $\mathcal{B}_\sigma$ a ball of radius σ centered at the origin, discretized into N points: $\mathcal{B}_\sigma = \{x_1, \ldots, x_N\}$. Therefore,



(a) (b) (c)

Fig. 1. *Camera motion statistics* depend heavily on the application. Ground vehicle navigation induces strongly non-Gaussian velocity distributions, as the statistics of Golem 2 driving in the DARPA Grand Challenge show (a) forward and lateral one-second displacements (b) one-second orientation variation (c) scatter plot of the vehicle relative displacement after one second (top view, restricted to fast parts of the track).

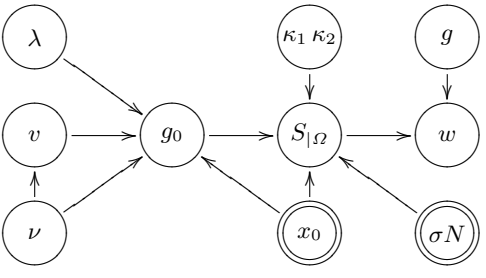


Fig. 2. *Statistical dependencies in a generative model of viewpoint warping.* Camera motion g and global shape S are rendered independent by conditioning on a local patch Ω. Local conditioning generates a dependency on x_0, σ and N, displayed as "observed" variables. The distributions $p(g_0|\lambda x_0, \nu, v)$, $p(S_{|\Omega}|g_0, \kappa_1, \kappa_2, x_0)$ and $p(w|S|_\Omega, g)$ encode deterministic functions resulting from simple geometrical and optical considerations (Sect. 1.2-2.1). The statistics of the other variables are determined empirically.

$\Omega \doteq \Omega(x_0, \sigma, N)$. We then call $w_i \doteq w(x_0 + x_i) - x_0 - x_i$ the displacement of the pixel $i = 1, \ldots, N$, so that we can characterize the distribution of w via the vectors $w_1, \ldots, w_N$:

$$p(w|x_0, \sigma, N) \doteq p([w_1, \ldots, w_N]|x_0, \sigma, N). \tag{2}$$

Here x_0, σ and N are parameters of the distribution, the first indicating the position on the image plane, the second the *scale* at which the statistics are computed, the third the *sampling* of the discretization. We will now attempt to decompose the density above to elucidate its structure. The statistical dependencies are highlighted in Fig. 2.

The first step, following (1), would be to marginalize with respect to the scene S and the motion g. Done globally, this would be a tall order since g and S are not independent: One's motion within a scene depends on its shape. For instance, one typically walks on the floor while avoiding obstacles that are part

of the scene S. However, since we are considering regions away from occluding boundaries, w does not depend on the entire scene S, but only on its visible portion.[4] Therefore, we condition on the pre-image of the patch Ω and only consider the *local* dependency of w on S:

$$p(w|x_0, \sigma, N) = \int p(w|g, S_{|\Omega}) dP(g, S_{|\Omega}|x_0, \sigma, N). \tag{3}$$

The advantage is that g and the local pre-image $S_{|\Omega}$ are to first approximation independent, which yields

$$p(g, S_{|\Omega}|x_0, \sigma, N) = p(g)p(S_{|\Omega}|x_0, \sigma, N).$$

Note that local conditioning introduces the dependency of $S|_\Omega$ from x_0, σ and N, so empirical studies must take it into consideration.

The first factor $p(g)$ is the viewer *motion* density, which is crucially dependent on the application. For human motion (or hand-held cameras), the statistics have been computed in [38]. These are rather different than those for ground vehicle navigation: Fig. 1 shows statistics of displacement and rotation of the vehicle "Golem 2" during the DARPA Grand Challenge. Rotational and translational degrees of freedom are strongly correlated due to non-holonomic constraints. At the other end of the spectrum one can imagine a *tumbling robot* where the motion density is (improper and) close to uniform $p(g) \sim \mathcal{U}(SE(3))$.

The second factor can be further decomposed by locally approximating the scene $S_{|\Omega}$ using the Darboux frame g_0, and the two principal curvatures, κ_1 and κ_2, that encode *local shape*:

$$p(S_{|\Omega}|x_0, \sigma, N) = p(\kappa_1, \kappa_2, g_0|x_0, \sigma, N) \tag{4}$$

The Darboux frame g_0 is determined by the normal ν, the principal direction v, and the position of the point $\lambda x_0 \in \mathbb{R}^3$ where x_0 is written in homogeneous coordinates and $\lambda \in \mathbb{R}^+$ is the depth along the corresponding ray.

The first observation is that the dependency of this density on x_0 is non-trivial: On the top portion of an image we usually observe the ceiling (indoor) or the sky outdoor, on the bottom we usually have a flat ground; these significantly bias the pose statistics as shown in Fig. 4. There are also dependencies on the geometry of the sensor: The shape of the pre-image of Ω for a flat sensor, or for a cylindrical or conical mirror, depends on the location x_0 on the image. These, however, are second-order effects that can be easily compensated for. Having observed these effects, we then resort to computing aggregate statistics by marginalizing over x_0. So, we are left with having to estimate the density

$$p(\kappa_1, \kappa_2, \nu, v, \lambda|\sigma, N) = p(\kappa_1, \kappa_2|\sigma, N)p(v|\nu, \sigma, N)p(\nu|\sigma, N)p(\lambda|\sigma, N) \tag{5}$$

[4] Strictly speaking this assumption is incorrect, as a camera motion g can turn an interior point into an occluding boundary. However, here we assume that most interior points will remain so during motion.

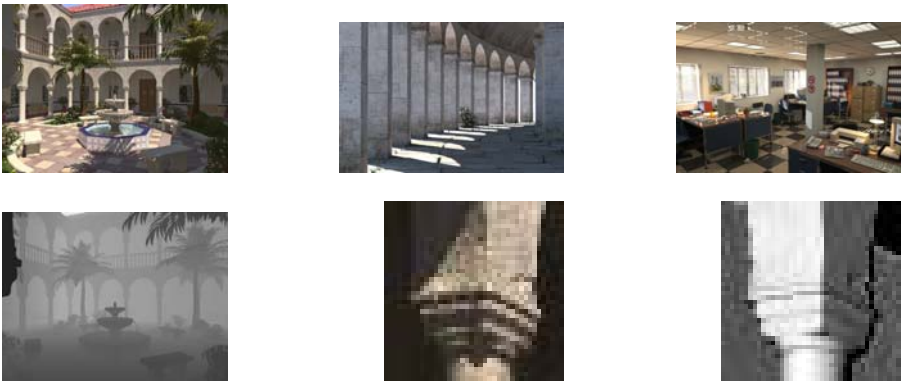
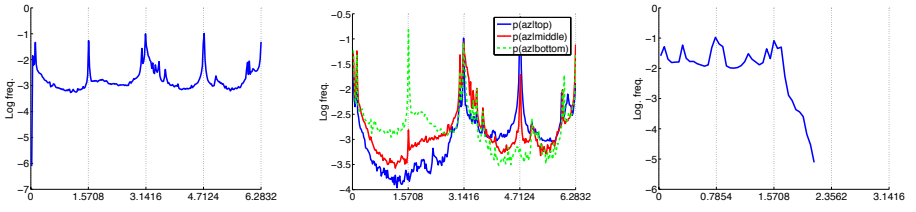


Fig. 3. A few samples from the *synthetic dataset* [39]. (Bottom-left) A range map computed from the model, with details (middle) showing fine-scale details (e.g. surface cracks) that are part of the geometry (shaded surface, right) and not just "painted" onto smooth surfaces.

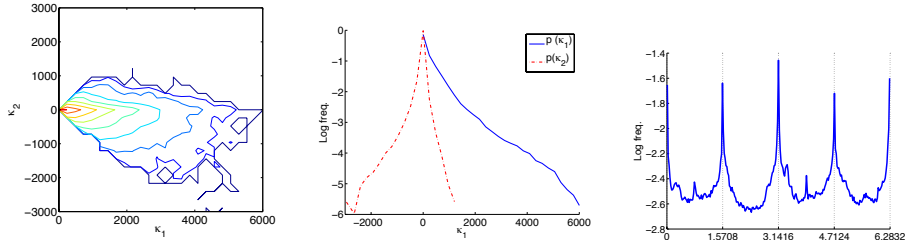


(a) Azimuth (b) Disaggregated (c) Elevation

Fig. 4. *Pose statistics.* (a) Histogram of the orientation of the normal vector relative to the optical axis. The peaks are due to horizontal and vertical surfaces. (b) The same statistics vary significantly if restricted to the top, middle and bottom third of the images. (c) Elevation of the normals relative to the optical axis.

which we do empirically. In order to have full control of sampling issues, we have decided to derive these statistics from simulated (ray-traced) images. While this choice presents potential dangers due to shortcuts often employed in ray-traced images, extensive sets of realistic images can be found, for which "ground truth" S is available. In our experiments we have used the datasets [39] that contains extremely detailed and realistic models (see Fig. 3).

We have observed that σ does not affect the nature of the statistics as long as $S(\Omega)$ can be approximated up to second order (recall that we are looking away from occluding boundaries). The choice of N is more delicate. Since the images are given to us at a fixed sampling rate, N and σ are naturally related. We have chosen σ corresponding to small windows of 5×5 pixels, and then implicitly chosen N by matching the scale of the mesh of S with the sampling of the image patch. We have done so by anisotropically smoothing the mesh proportionally to



(a) Joint curv. (b) Marginal curv. (c) Principal direction

Fig. 5. *Shape statistics.* (a) Joint histogram of the principal curvatures κ_1, κ_2. (b) Marginal histograms. (c) Histogram of the orientation of the principal direction (projected onto the image plane).

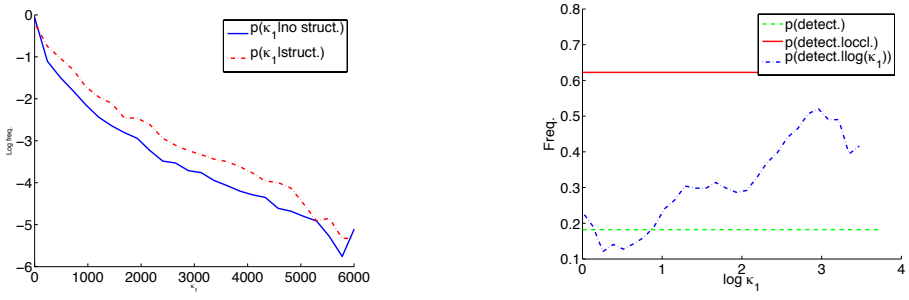
the area of the pre-image $S(\Omega(x_0, \sigma, N))$, while preserving occluding boundaries and sharp discontinuities. Curvatures and principal directions are computed using discrete differential operators [40] on the regularized meshes. The resulting marginal and joint histograms are shown in Fig. 5. As one can expect, these statistics exhibit high kurtosis, indicating that regions of high curvature are rare. Most non-planar structures are wedges ($\kappa_2 \approx 0$) and, interestingly, saddles ($\kappa_2 < 0$), consistent with the observations of [37].

2.2 Occlusion Statistics

Empirical distributions on the frequency of occluding boundaries can be obtained directly from range images. These have already been studied in [37], and show a kurtotic behavior similar to that of curvature, indicating that occlusions are a rare event.

2.3 Saliency of Singularities and Occlusions

Although occlusions and singularities are rare events, in the sense that they represent a zero-measure subset of the scene and project onto a small subset of the image (by area), they are salient in that such geometric discontinuities often correspond to photometric discontinuities that are selected by feature detectors. For the case of occlusions, this is obvious since at occluding boundaries an arbitrarily small neighborhood contains the image of different objects. For the case of singularities, Chen et al. [18] have argued that in homogeneous materials they yield photometrically salient profiles, that they have measured empirically. To validate these results, we have tested a standard edge detector (Canny) on ray-traced images and we have examined the co-location of their responses to the curvature of the local pre-images (Fig. 6). This experiment illustrates that occlusions and singularities, although rare, are photometrically salient, and therefore *there remains the need to study feature descriptors for regions that include discontinuities.* We now move on to that problem.



(a) Curvature on edges (b) Edges on bent patches

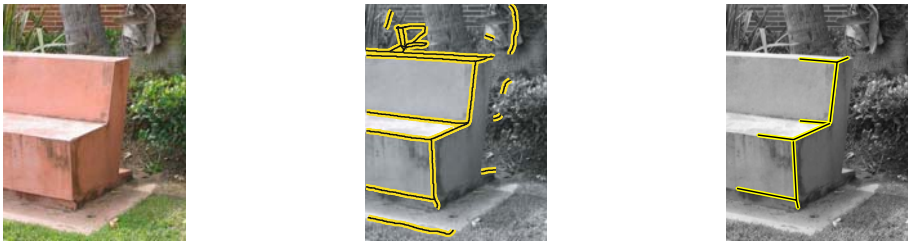
Fig. 6. *Saliency of singularities.* A Canny edge detector is implemented at a scale comparable to the scale of the patches used for the statistics (5×5 pixels). In (a) we show how the histogram of the principal curvature κ_1 varies if restricted to those patches that contain an edge (we discarded patches that contain an occluding boundary): On average, "Canny patches" have higher curvature. In (b) we sorted the patches in increasing curvature (log scale) and computed the fraction that contains an edge. We repeated this computation for all patches and the patches that contain an occluding boundary. The fraction increases significantly with the curvature and is even higher for occluding boundaries.

3 Designing Viewpoint Invariant Descriptors

The empirical evidence in the previous section suggests that image regions with discontinuities or occluding boundaries are photometrically salient, which in turn suggests that they may be distinctive and therefore useful for recognition. In this section we illustrate how to construct viewpoint invariant features for such regions. We first show how a general methodology has been used before for the case of interior-regular points and singularities, and extend it to occluding boundaries.

We will assume that we have a mechanism available to establish the origin of a local reference frame. This is the role of a *feature detector* that can pool statistics from regions of various shape and size. Detectors may localize a point on the image, or select entire regions (in case the pooled statistics are constant), which in turn can be used to establish a local frame. Around the origin we will construct a local viewpoint invariant region statistic, or *feature descriptor*.

From the image formation model (1) it is immediate to see that the equivalence class of image deformations, *i.e.* the set $\phi(I, \Omega) \doteq [I(w(x)),\ x \in \Omega\ \forall\ w]$ is a viewpoint invariant. Indeed, it is the *maximal* viewpoint invariant, in the sense that any other invariant is a function of it. Unfortunately, comparing such invariants could be difficult because it entails a search over w. Since $\phi(I, \Omega)$ is an equivalence class, any element can represent it. Therefore, one can seek a mechanism to associate a *canonical warping* $\hat{w}$ to the local image structure, as



(a) Image (b) Sketch (c) Frames

Fig. 7. *Local image structure as extracted in the pre-processing step.* The image in the middle shows the structures extracted by the sketch. On the right we show some of the star-like subgraphs that we manually select as candidate feature frames. The subgraphs are centered on junctions and have linear branches. Whenever no natural termination of a branch is found, a nominal value (established by looking at the maximum of a Laplacian operator centered at the junction) is used.

well as a canonical domain $\hat{\Omega}$, and use $[I(\hat{w}(x)),\ x \in \hat{\Omega}]$, or any function of it, as the invariant descriptor.[5]

3.1 Interior Regular and Singular Points

The program sketched above has been carried out successfully by many researchers for the case of affine warps: $w(x) = Ax + b$. Note that the linear terms in the local approximation can be written out, spelling explicitly the rotational and translational components of g, as $w(x) = (R + T\nu^T)x$**,** and the homography $(R + T\nu^T)$ can be approximated with an affine transformation $[A\ b]$. Therefore, the transformation induced by any IR point can be annihilated by an appropriate affine transformation: The scene is a plane, $S = \mathbb{R}^2$, the translational term b is fixed by a feature detector (*e.g.* Harris [42], so without loss of generality we can assume $b = 0$), and the second moment matrix, or other local intensity statistic [43], can be used to determine A. The transformation that inverts A can be interpreted as a warping of a canonical circular neighborhood $[I(\hat{w}(x)),\ x \in \mathbb{S}^1]$, or $\hat{w}$ can be though of as the transformation of a detected elliptical region $\hat{\Omega}$ into a circle $\mathbb{S}^2 = \hat{w}^{-1}(\hat{\Omega})$, as in [2].

The same ideas can be easily extended to non-planar scenes [41]. In this case, the reference frame we seek to normalize using intensity statistics is not affine, but curvilinear and possibly known only up to symmetries, when the image presents regular textures or homogeneous regions [41]. The deformation induced by changes in viewpoint can be represented by a *piecewise affine transformation*, with as many components as connected elements of the singularity. For instance, an edge has 2

[5] Invariance is achieved through a local homeomorphic deformation of the image domain into a canonical configuration tailored to the local image structure. While fixing a homeomorphism of the image domain forces viewpoint invariance, the converse is not necessarily true; i.e. image domain deformations induced by changes of viewpoint do not cover the set of all possible homeomorphisms [41], unless the scene is planar.

affine components, a 3-D corner has 3, etc. with the tip of a cone with smooth section as a limiting case. Naturally these affine transformations are *not independent* because they have to satisfy compatibility constraints (see [41] for details).

3.2 Occluding Boundaries and Unilateral Descriptors

It is easy to show that the deformation induced by the motion of an arbitrary shape does not preserve any geometric or topological property of the silhouette [44, 45]. Indeed, given two curves, one can construct objects that, under suitable viewpoints, have the curves as silhouettes. This is not true when the object has symmetries, or when it has a particular structure, for instance a polyhedron. In the former case one can derive case-by-case invariants, which is beyond our scope here. In the latter case, occluding boundaries correspond to singularities, and we can build a *unilateral descriptor* following the lines of the previous section. We proceed with a detector in the exact same way as we did in Sect. 3.1, since a-priori we do not know whether edges in the image are due to albedo or shape. Then for each local neighborhood we construct not one, but several descriptors based on masking different sectors of the local graph, followed by rectification. Whether a given region is a singularity or an occluding boundary will only be clear at matching: If matching all N regions independently produces similarly small residuals, the singularity hypothesis is accepted, and the entire region is normalized and matched. If at least one of the N matches yields a low residual, the occlusion hypothesis is accepted, and matching is based on one sector only Figs. 8-9 illustrates few representative examples.

Once the local structure in a neighborhood of the image is extracted by a low-level feature detector, one could build a discrete representation (local graph) and compare regions by comparing their graphs. Unfortunately, such graphs are highly unstable with respect to changes of viewpoint, as failure to detect local

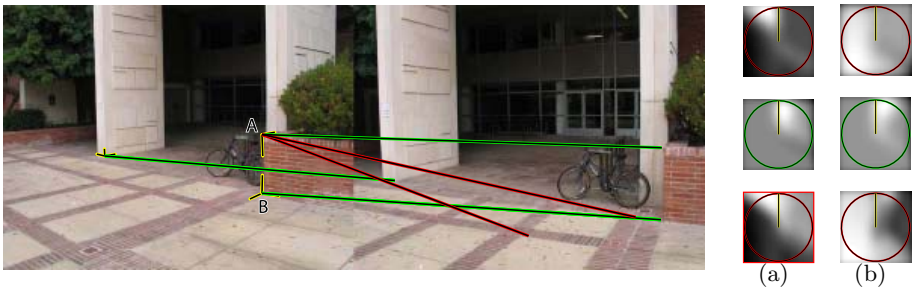


Fig. 8. *Background resilience.* Feature A: To achieve insensitivity to the background, we generate two unilateral SIFT descriptors, one for each side of the local frame. (a) and (b) show three patches on which the three descriptors are computed (scale is 1/6 of the radius). These features have been added to the pool of features detected by SIFT to see whether they enable correct discrimination. Of these, two do not match correctly (red lines) because they cover the background, while the other (green line) does as it covers only the foreground. Feature B: Both the bilateral and unilateral descriptors match because the background does not change substantially.

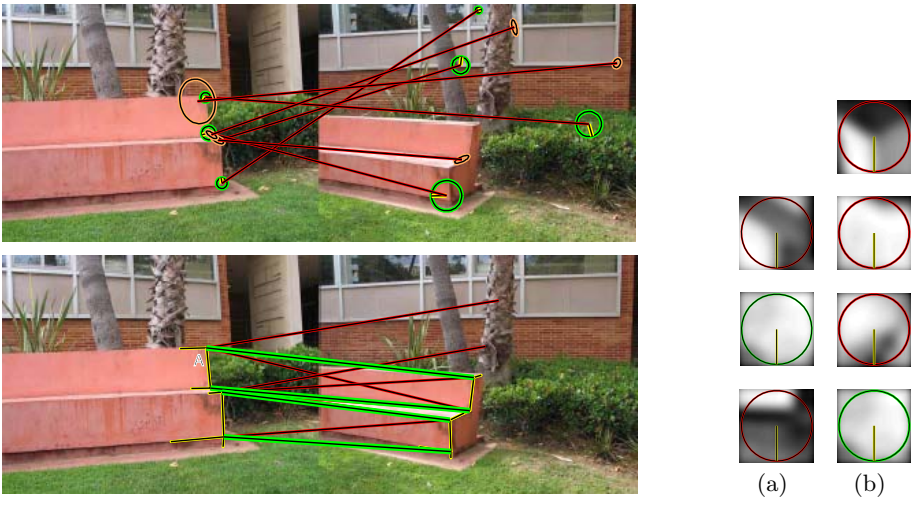


Fig. 9. *Occlusion resilience.* We generate several descriptors for each selected corner structure of Fig. 7, then add them to the pool detected by SIFT/Harris-Affine. (top) Due to visibility effects, SIFT (green) and Harris-Affine (orange) fail to match all four corners (red lines). (bottom) The unilateral descriptors that cover the foreground portion of the object are matched correctly, while the others fail. Eventually each feature is associated to its best matching descriptor (green lines). Columns (a) and (b) show the three and four descriptors extracted in the two images for the feature denoted as A (in green the matching descriptors).

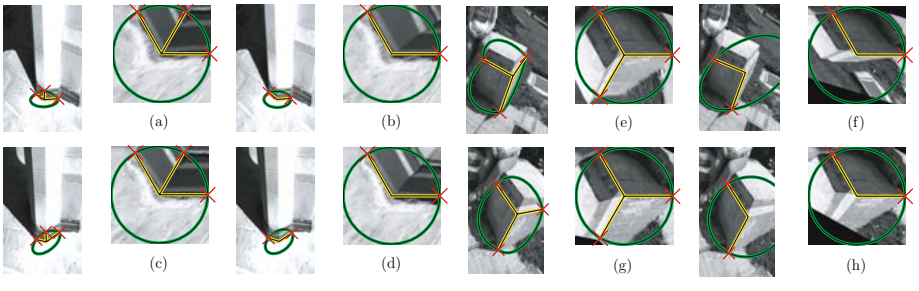


Fig. 10. *Comparing discrete structures through a generative process.* Pictures (a) and (c) show two images of the same structure re-projected by means of the normalized graph: They look similar as expected. Pictures (b) and (d) show the patches obtained when the weak structure (the central edge) is removed from the graph. Despite the graph topology changing drastically (a 3-junction becomes a corner), the re-projected patches look quite similar. Unfortunately this does not work in all cases, as depicted in pictures (e-h). Here normalization is inconsistent because the detector only considers edge-like structures.

structures results in changes of topology of the corresponding graph. Since we compare intensity statistics in a normalized frame, one could argue that if a local structure is not stable enough with respect to changes of viewpoint, that structure should not matter for matching. We illustrate this in Fig. 10 (left) where the instability in inferring local image structure is anihilated by the synthesis of the normalized patch. Indeed, since the canonical configuration is arbitrary, one can choose it to compensate for failures of the low-level feature detector.

This, however, does not always work, since missed detection changes the canonization procedure, as we illustrate in Fig. 10 (right). This is the weakest point of our method, which can be improved with mid-level processing and grouping procedures that are beyond the scope of this paper.

4 Discussion

We have derived a statistical characterization of the deformations of the image domain induced by changes of viewpoint. This shows that, while occlusions and surface singularities are rare, they are photometrically salient, which motivates their use for recognition. This prompts us to develop dedicated viewpoint invariant descriptors.

For singularities, we rely on existing methods to extract local image structure, and construct an invariant descriptor by normalizing such structure and generating a canonical radiance from it. Although the technique is general, it relies on pre-processing steps that, with the current state of the art, are problematic. Alternatively, one could use region-based segmentation approaches as a means to extract local structure ahead of computing invariant statistics. For the case of occlusions, we have developed unilateral descriptors based on masking portions of the detected regions. We have shown a few representative examples of the behavior of such descriptors for cases where existing affine invariants fail to establish correspondence. Note that we do not advocate the descriptors in Sect. 3.1-3.2 as an *alternative* to existing descriptors. They are designed to cover conditions that current descriptors are not designed for, hence be complementary. Note also that the best affine descriptors can tolerate a great deal of violation of the assumptions they are designed for, therefore many of the cases where our approach would be best suited is already covered by, say, SIFT or Harris-affine.

Considerable work remains to be done to design robust and stable low and mid-level detection schemes, but we hope that this study illustrates a general methodology that can be used to design viewpoint invariant descriptors for non-planar portions of the scene.

References

1. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV **2** (2004) 91–110
2. Mikolajczyk, K., Schmid, C.: An affine invariant interest point detector. In: Proc. ECCV, Springer-Verlag (2002) 128–142

3. Schaffalitzky, F., Zisserman, A.: Viewpoint invariant texture matching and wide baseline stereo. In: Proc. ICCV. (2001)
4. Triggs, B.: Detecting keypoints with stable position, orientation, and scale under illumination changes. In: Proc. ECCV. (2004)
5. Ullman, S., Vidal-Naquet, M., Sali, E.: Visual features of intermediate complexity and their use in classification. Nature **5** (2002)
6. Kadir, T., Brady, M.: Scale saliency: A novel approach to salient feature and scale selection. In: International Conference Visual Information Engineering. (2003)
7. Lindeberg, T.: Feature detection with automatic scale selection. IJCV **30** (1998) 77–116
8. Matas, J., Chum, O., Urban, M., Pajdla, T.: Robust wide baseline stereo from maximally stable extremal regions. In: Proc. BMVC. (2002)
9. Fei-Fei, L., Fergus, R., Perona, P.: A Bayesian approach to unsupervised one-shot learning of object categories. In: Proc. ICCV. (2003)
10. Weber, M., Welling, M., Perona, P.: Unsupervised learning of models for recognition. In: Proc. ECCV. (2000)
11. Rothganger, F., Lazebnik, S., Schmid, C., Ponce, J.: 3D object modeling and recognition using affine-invariant patches and multi-view spatial constraints. In: Proc. CVPR. (2003)
12. Fergus, R., Perona, P., Zisserman, A.: Object class recognition by unsupervised scale-invariant learning. In: Proc. CVPR. (2003)
13. Sivic, J., Zisserman, A.: Video Google: A text retrieval approach to object matching in videos. In: Proc. ICCV. (2003)
14. Ferrari, V., Tuytelaars, T., Van Gool, L.: Integrating multiple model views for object recognition. In: Proc. CVPR. (2004)
15. Tuytelaars, T., Van Gool, L.: Wide baseline stereo matching based on local, affinely invariant regions. In: Proc. BMVC. (2000) 412–425
16. Dorkó, G., Schmid, C.: Object class recognition using discriminative local features. PAMI (submitted) (2004)
17. Fritz, G., Seifert, C., Paletta, L., Bischof, H.: Rapid object recognition from discriminative regions of interest. In: Proc. 19th National Conference on Artificial Intelligence. (2004) 444–449
18. Chen, H.F., Belhumeur, P.N., Jacobs, D.W.: In search of illumination invariants. In: Proc. CVPR. (2000)
19. Guo, C., Zhu, S.C., Wu, Y.N.: Towards a mathematical theory of primal sketch and sketchability. In: Proc. ICCV. (2003) 1228
20. Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., Van Gool, L.: A comparison of affine region detectors. IJCV **1** (2004) 63–86
21. Fergus, R., Perona, P., Zisserman, A.: A sparse object category model for efficient learning and exhaustive recognition. In: Proc. CVPR. (2005)
22. Csurka, G., Dance, C.R., Dan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: Proc. ECCV. (2004)
23. Grauman, K., Darrell, T.: Efficient image matching with distributions of local invariant features. In: Proc. CVPR. (2005)
24. Malik, J., Belongie, S., Shi, J., Leung, T.: Textons, contours and regions: Cue integration in image segmentation. In: Proc. CVPR. (1999)
25. Thureson, J., Carlsson, S.: Appearance based qualitative image description for object class recognition. In: Proc. ECCV. Volume 2. (2004)
26. Ferrari, V., Tuytelaars, T., Van Gool, L.: Simultaneous object recognition and segmentation by image exploration. In: Proc. ECCV. (2004)

27. Brown, M., Lowe, D.G.: Invariant features from interest point groups. In: Proc. BMVC. (2002)
28. Fraundorfer, F., Bischof, H.: A novel performance evaluation method of local detectors on non-planar scenes. In: Proc. CVPR. (2005)
29. Mikolajczyk, K., Zisserman, A., Schmid, C.: Shape recognition with edge-based features. In: Proc. BMVC. (2003)
30. Stein, A., Hebert, M.: Incoporating background invariance into feature-based object recognition. In: Seventh IEEE Workshop on Applications of Computer Vision (WACV). (2005)
31. Vasconcelos, N.: Feature selection by maximum marginal diversity: optimality and implications for visual recognition. In: Proc. CVPR. (2003)
32. Wold, L., Shashua, A.: Feature selection for unsupervised and supervised inference: the emergence of sparsity in a weighted-based approach. In: Proc. ICCV. (2003)
33. Levi, K., Weiss, Y.: Learning object detection from a small number of examples: the importance of good features. In: Proc. CVPR. Volume 2. (2004)
34. Grenander, U.: General Pattern Theory. Oxford University Press (1993)
35. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active shape models – their training and application. Comput. Vis. Image Underst. **61** (1995) 38–59
36. Huang, J., Lee, A.B., Mumford, D.: Statistics of range images. In: Proc. CVPR. (2000)
37. Yang, Z., Purves, D.: Image/source statistics of surfaces in natural scenes. Network: Comput. in Neural Syst. **14** (2003)
38. Roth, S., Black, M.J.: On the spatial statistics of optical flow. In: Proc. ICCV. (2005)
39. Piqueres, J.V.: The persistence of ignorance. http://www.ignorancia.org/ (2006)
40. Meyer, M., Desbrun, M., Schröder, P., Barr, A.H.: Discrete differential geometry for triangulated 2-manifolds. In: Proc. of VisMath. (2002)
41. Vedaldi, A., Soatto, S.: Features for recognition: Viewpoint invariance for non-planar scenes. In: Proc. ICCV. (2005)
42. Harris, C., Stephens, M.: A combined corner and edge detector. In: Proceedings of The Fourth Alvey Vision Conference. (1988) 147–151
43. Bauer, J., Bischof, H., Klaus, A., Karner, K.: Robust and fully automated image registration using invariant features. In: Proc. of Intl. Arch. of Photogram., Remote Sensing and Sptl. Inf. Sci. (2004)
44. Lazebnik, S., Sethi, A., Schmid, C., Kriegman, D.J., Ponce, J., Hebert, M.: On pencils of tangent planes and the recognition of smooth 3D shapes from silhouettes. ICJV (2002)
45. Schmid, C., Zisserman, A.: The geometry and matching of curves in multiple views. In: ECCV. (1998)