Shiny Correspondence: Multiple-View Features for Non-Lambertian Scenes

Jason Meltzer Stefano Soatto
CSD-TR050019

May 20, 2005

Abstract

We present a method for developing viewpoint and illumination invariant feature descriptors for scenes that exhibit non-Lambertian reflection, for instance specularities or transparency. Such descriptors cannot be constructed from a single image; by observing many views from a moving vantage point and modeling image variation as a superposition of (diffuse and specular) radiance layers, we isolate diffuse albedo which acts as the invariant descriptor. In the process, we estimate the independent motion of such layers and their mixing coefficients. We illustrate our approach on shiny objects where existing feature descriptors fail to provide a stable signature for matching.

1. Introduction

Specularities are ubiquitous in natural and man-made environments. Most surfaces display some kind of specularity, be it highlighting, reflectivity, transparency, or a combination of many of these effects (consider floors, windows, people, picture frames, and computer parts, to name a few). Except under the most benign circumstances, such surfaces violate the assumption made by nearly all methods for recognition, namely that they obey the Lambertian reflectance model. This fact significantly complicates the tasks of object recognition, robot navigation, stereo, and any other systems which rely on finding correspondences amongst regions of multiple images.

A popular and powerful technique used throughout the literature on recognition, navigation, and reconstruction involves the computation of functions on regions of images. In the best case, these *feature descriptors* provide distinctive signatures of locations in space which can be matched reliably despite changes in the viewpoint of the camera. Research on invariant descriptors has progressed tremendously in recent years. A wide range of robust and accurate correspondence techniques for Lambertian scenes are available to facilitate higher-level vision-based tasks. A common procedure is to compute *affine invariants*, which are functions of image regions that are insensitive to affine warpings. By normalizing with respect to affine transformations, these descriptors can match planar Lambertian surface patches across wide viewpoint changes.

There are three common physical configurations for which affine invariant descriptors fail: occlusions, highly non-planar surfaces, and non-Lambertian reflectance. In the first two cases, the descriptors, which are developed from a single image, fail to accurately model the geometry of the scene; for the latter case, they fail to model the photometry. In this paper, we focus on non-Lambertian surfaces for the cases of reflection and transparency.

1.1. Single Views are Not Enough

Current approaches compute affine-invariant feature descriptors which approximate locally planar surfaces, and then match these across images. Since features are extracted from one image at a time, all such methods implicitly rely on the assumption of Lambertian reflection. The application of affine warping of an image neighborhood approximates the image of a viewpoint transformation of the corresponding plane in space. This is valid for a Lambertian surface under fixed illumination, but insufficient otherwise. The image of a surface displaying reflection or transparency is a composite of the radiance of many surfaces in the scene, each of which may have different depths and orientations. A change in viewpoint, therefore, will induce changes in the image which depend on the motions of more than one surface.

While Lambertian reflection is a necessary condition for single-view matching, it is not sufficient in the the presence of varying illumination, as one can have diffuse shading effects [15], or directed and self-shadowing [2] that impede the design of illumination invariants. Note that both diffuse and specular albedo are viewpoint-invariant, but the latter is difficult to estimate without a shape model [27], and reflectance and illumination can be factored out only in the presence of stringent conditions.

We seek to overcome these limitations by incorporating multiple views of an image patch. We assume that the scene exhibits a diffuse + specular reflectance model, and that global illumination is constant during the extraction of each feature descriptor (it can change arbitrarily between instances that are submitted to the system for matching). The feature learning process is an implicit multi-view

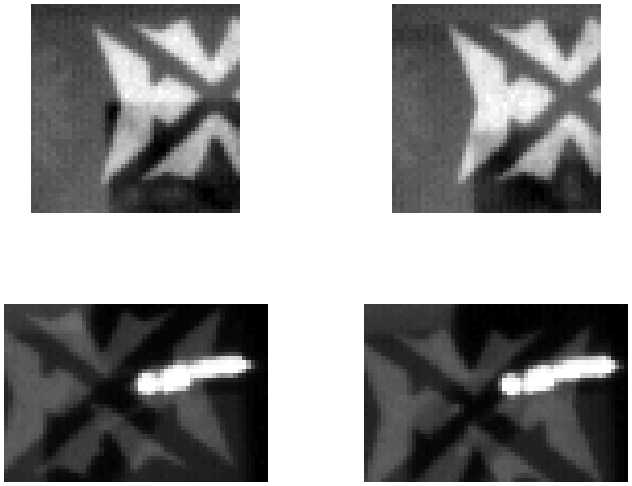

Figure 1: **Non-Lambertian Reflectance** These otherwise good features are corrupted by the motion of a specularity.

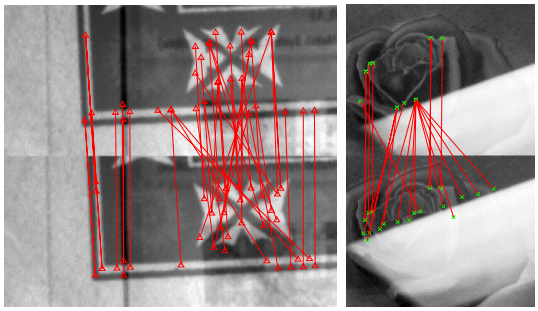

Figure 2: **SIFT on Non-Lambertian Scenes** These planar scenes display significant ambiguity in SIFT descriptor matching. Many points, despite being detected by the SIFT detector on both frames, are not matched or matched incorrectly due to the motion of the specularity in the scene.

reconstruction, whereby we assume that the surface is locally planar, and that image motion is affine. The image is therefore an additive composite of two layers, mixed linearly, each of which undergoes interdependent (though unknown) affine warping. We estimate the motion and albedo of each layer, corresponding to the affine motion of the diffuse albedo and the reflection of the light source relative to the moving viewpoint.

Once the albedos of the two layers have been recovered, an affine intensity-normalized version of each is warped to a canonical configuration around a feature point selected with standard techniques (refer to section 2). Since we cannot know which layer represents the diffuse albedo and which is specular, each constitutes a component of the invariant signature. During the matching phase, we compare all combinations of the descriptors of each layer and declare a successful match if at least one pair's distance is below a threshold.

2. Related Work

Feature descriptors for Lambertian scenes are well studied in the contexts of recognition, tracking, and navigation. Presently, there is a rich literature on *affine invariants*, which are functions of image patches which do not change with rotation, translation, scale, and skew transformations. The underlying assumptions made by all such methods are that the patches come from planar surfaces, the transformation of the image under a viewpoint change of the camera is approximately affine, and that the surfaces containing the patches are Lambertian. Typically, for small patches in many environments these assumptions are valid. [1, 7, 17, 18, 21, 25] are a selection of some recent works embodying this idea.

The most comprehensive affine invariant descriptor is presented by Mikolajczyk and Schmid in [18]. It searches an affine Gaussian scale space of three scale parameters (see [16] for the details of affine Gaussian scale space theory). An initial detection step uses a Harris corner detector ([10]) over multiple scales to find candidate points of interest. An iterative procedure finds a refined position and canonical scale and warping for each image neighborhood around each interest point. The descriptor is a 12 dimension vector built from normalized Gaussian derivatives (up to 4th order) of the rectified image neighborhood.

The scale invariant feature transform (SIFT) [17] is a similarity-invariant interest point detector and descriptor, which explicitly accounts for variability in translation, rotation, and changes of scale. The system builds an isotropic (one scale parameter) Gaussian scale-space pyramid and finds the maxima and minima of neighborhoods in the pyramid. These are assigned a canonical orientation and scale, and a descriptor is built based on Edelman *et al* [5]. The 128-element SIFT descriptor is produced by sampling the gradients around interest points at their assigned scales and binning them by orientation. [19] and [17] demonstrate the effectiveness of this descriptor for a wide range of orientations, scales, and some 3D perspective variation. Ke and Sukthankar modifiy this descriptor by generating a basis of principal components of the gradient vectors [14]. They increase the robustness of matching by comparing any two gradient vectors' projections onto this basis.

A powerful augmentation of the affine-invariant paradigm is presented in [7], whereby existing descriptors are used in conjunction with an *image exploration* technique to increase the proportion of correct matches. By propagating geometric constraints in areas surrounding candidate matches, the algorithm expands the regions of

correct correspondences and removes outliers. The affine invariant descriptor from [26] is used in the experimental system, along with the matching technique of [8] t generate initial correspondences.

There is a similarly broad literature on the topics of trans parent and reflective surfaces. Since our techniques involv motion rather than polarization or change of focus (such a [6], [20]), we concentrate here on prior research involvin the separation of transparent and reflective motions.

In [3], Bergen *et al* study the problem of recovering tw independent motions from three images. They assume a image generation model whereby two layers combine ac cording to some mixing (usually additive). The layers ca have independent velocities. Given a rough estimate of on component motion, they iteratively find both by alternatin coarse-to-fine motion estimation for each component.

Darrell and Simoncelli in [4] develop a bank of spatio temporal filters which give zero response when applied t points on images displaying corresponding motions. The generate many of these filters assuming that there are fe independent image velocities at any point, test combina tions of the filters, and select the set that showed the bes performance (maximum "nulling"). Their results demon strate the ability to separate foreground and backgroun motions in image sequences.

Szeliski *et al* in [23] assume an image is composed of a linear mixture of warped layers. Warping matrices operate on the pixels of the component layers to approximate their relative motions and the mixing coefficients. By observing that no pixel in a component layer can be less than zero, they develop a novel algorithm which alternatively tightens upper and lower bounds on the layers while simultaneously finding their motions.

Irani *et al* [11] use a model similar to [23] but process images in a temporal fashion instead of in batch. In [13], Ju *et al* seek to find optical flow in a robust manner by incorporating information from multiple adjacent regions. A potential motion at a point will have a set of surrounding inliers (surround points which move with the same motion) and outliers (those which move with differing motions). In their formulation, single motions are reinforced by inliers, and two motions (due to transparency) are estimated in a by analyzing the motion of outliers and separating these into two layers, plus outliers of both.

3. Problem Formulation and Modeling

Consider a semi-reflective surface S, for example a painting behind a glass frame or a glossy magazine. P is a generic point on S with a tangent plane T_P and albedo $\rho_d(P)$. A second surface, denoted L, contains points $P' \in L$ and albedo $\rho_s(P')$. S and L respectively represent the target object and the portion of the scene reflected from S toward the camera. Likewise, $\rho_d(P)$ and $\rho_s(P')$ constitute the diffuse and specular components of the radiance at the point

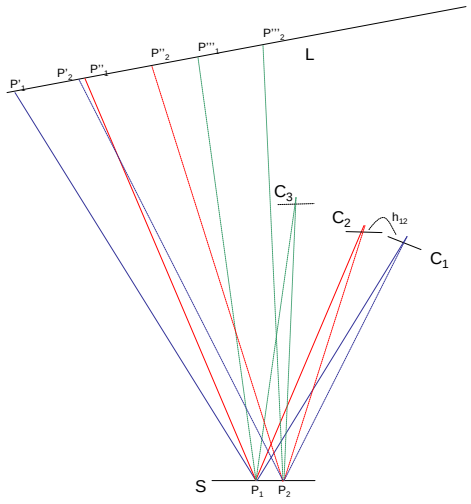

Figure 3: **Non-Lambertian Reflectance Model** The semi-reflective surface S contains point P, and the reflection points P' lie on the lighting surface L. C_1, C_2, C_3 are the three camera views

An image $I_1(x)$ is formed by a linear combination of two layers, $\rho_d(P)$ and $\rho_s(P')$, where P' is the point on L such that a ray cast from the image coordinate x reflects off T_P and intersects L. Under a change of viewpoint, a new image at coordinate y, where y and x correspond to the same point P on S, $I_2(y)$ is a linear combination of $\rho_d(P)$ and $\rho_s(P'')$, $P'' \in L$.

Our imaging model is therefore

$$\begin{aligned} I_1(x_i) &= \alpha_1 \rho_d(P_i) + \beta_1 \rho_s(P_i^{'}) \\ I_2(y_i) &= \alpha_2 \rho_d(P_i) + \beta_2 \rho_s(P_i^{''}) \end{aligned} \quad (1)$$

where $\alpha_i + \beta_i = 1 \; \forall i$. Note that in the above model (1) if the surface is Lambertian, then $\beta_i = 0 \; \forall i$. It is easy to see why any single-view descriptor of a non-Lambertian scene will fail: under a change in viewpoint, the second term of the right-hand side changes. Thus, descriptors developed from single views cannot be invariant even for simple non-Lambertian scenes.

We assume that at each feature point in the scene, the image of the neighborhood of P consists only of a superposition of the images of two planes, T_P and $T_{P'}$, which are the tangent plane of S at P and the tangent plane of L at P'. Under these conditions, a viewpoint transformation of the scene can be represented with two affine transformations, one for each layer. In principle, these affine transformations

are interdependent if the scene is fixed and only the camera moves. In practice, however, since we do not know the motion of the camera or the geometry of the scene, we assume that the camera observes independent motions of each layer.

Thus, our model simplifies to

$$\begin{aligned} I_1(x_i) &= \alpha_1 \rho_d(h_{d1}(x_i)) + \beta_1 \rho_s(h_{s1}(x_i)) \\ I_2(y_i) &= \alpha_2 \rho_d(h_{d2}(y_i)) + \beta_2 \rho_s(h_{s2}(y_i)) \end{aligned} \quad (2)$$

or, if written in terms of warped images J_d and J_s,

$$\begin{aligned} I_1(x_i) &= T_1(\alpha_1 J_d(x_i)) + T_2(\beta_1 J_s(x_i)) \\ I_2(y_i) &= T_3(\alpha_2 J_d(y_i)) + T_4(\beta_2 J_s(y_i)). \end{aligned} \quad (3)$$

For a quantized images of size $m \times n$, we have mn equations and $2mn + 13$ unknowns: two $m \times n$ sized images, two 6-parameter affine transformations, and one scalar mixing factor α_i ($\beta_i = 1 - \alpha_i$). We therefore need at least 3 images in order to recover all of the parameters.

On the discrete pixel grid, we can write this system of equations as

$$\begin{aligned} I &= TP \\ I &= \begin{bmatrix} I_1 \\ I_2 \\ I_3 \end{bmatrix}, \; T = \begin{bmatrix} T_1 & T_2 \\ T_3 & T_4 \\ T_5 & T_6 \end{bmatrix}, P = \begin{bmatrix} J_1 \\ J_2 \end{bmatrix} \end{aligned} \quad (4)$$

where the mixing coefficients are rolled into the transformation matrices. As the motion of the layers is relative to the first image, we take $T_1 = T_2 = I_{n \times n}$. In order to find the transformations T, observe that

$$T^{\perp} I = 0.$$

Since the form of T is known given the parameters of the affine transformations, we can compute $T^{\perp}$ via non-linear minimization.

$$\hat{T} = \arg\min_{T} \|T^{\perp} I\|^2$$

following which the layers can be computed as

$$\hat{P} = \hat{T}^{\dagger} I$$

where $A^{\dagger}$ denotes the pseudoinverse of the matrix A.

The separation of the layers allows us to calculate affine-invariant descriptors for each using standard techniques (refer to section 2). Since we do not know *a priori* which layer represents the diffuse surface (which is typically the object of interest), we store both descriptors. During a matching phase, the algorithm is run again on a new set of images, new descriptors are extracted for the two layers, and all pairwise distances are computed. If one pair matches, a successful correspondence is declared.

4. Experiments

Our experiments are in two parts. First, we test the layer separation using real images but synthetic layer motions and mixings. This allows us to measure the results of the separation relative to ground truth. Following, we test our algorithms on real data taken of a fixed scene from a moving camera. In both cases, the first step of our experimental algorithm does a dominant-motion alignment using Lucas-Kanade ([22, 24]) on a small patch. This allows us to constrain the search directions since one layer will have a small motion after alignment.

4.1. Synthetic Motion

To demonstrate the ability of our system to accurately separate the component layers of composite images, we performed experiments on real images mixed synthetically so that ground truth is available for the layers and the motions.

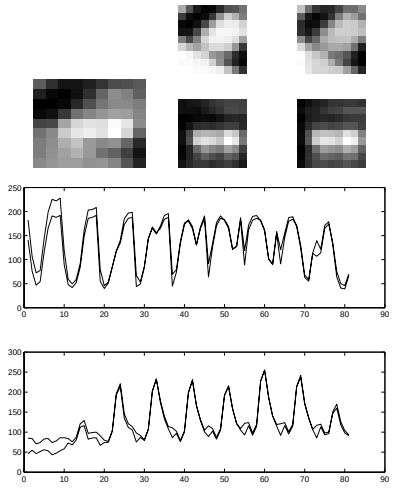

Figure 4: **Separation of layers for synthetic motion.** The top row left shows one of the input images, which was mixed synthetically. The center column of the top row are the original layers, and the right column the corresponding inferred layers. The plots on the second and third compare the original layers to the recovered layers, one plot for each.

4.2. Real Sequences

We demonstrate our algorithm on real image sequences, showing both the extraction of the layers and the ability to match between viewpoints. In the following experiment, the target object is the painting "Meditative Rose" by Salvador Dali, and the specular reflection is the box for Microsoft Office. Our algorithm separates the layers for pre-selected regions and matches based on SSD residuals.

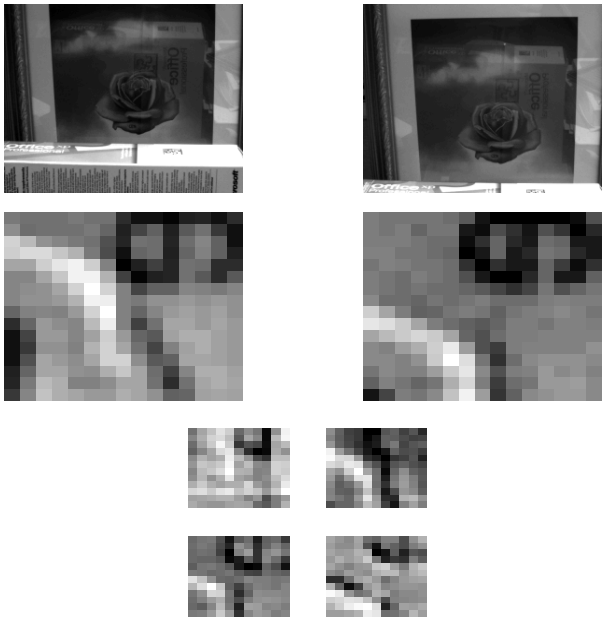

Figure 5: **Separation of layers for real motion.** The top row shows two of the original input images, taken from a moving camera. The row shows the clipped input images after tracking a feature (the 'e'), and the bottom four images the corresponding inferred layers, each column representing a layer, and each row derive from a separate sequence of input images. The separated layers in first column on the bottom can be matched using SSD to find correspondence between the 'e' on various image sequences, despite the specular motion.

5. Conclusion

We have presented an algorithm which can be used to find correspondence amongst image sequences of non-Lambertian scenes displaying reflection or transparency. Multiple views of any image feature are required to perform matching, since no single-view statistic is invariant under non-Lambertian reflectivity. Our algorithm first separates motion layers, reconstructs the albedos, and matches based on these component layers.

References

[1] A. Baumberg. "Reliable feature matching across widely separated views," *Proc. CVPR,* 2000.

[2] P. Belhumeur, D. Kriegman, A. L. Yuille. "The Generalized Bas Relief Ambiguity." *Int. J. on Computer Vision,* 35(1):33-44, 1999.

[3] J. Bergen, *et al*. "Computing Two Motions from Three Frames." *Proc. ICCV,* 1990.

[4] T. Darrell, E. Simoncelli. ""Nulling" Filters and the Separation of Transparent Motions." *Proc. ICCV,* 1993.

[5] S. Edelman, N. Intrator, and T. Poggio. "Complex cells and object recognition," unpublished manuscript, 1997.

[6] H. Farid, E. Adelson. "Separating Reflections from Images Using Independent Components Analysis." *J. of the Opt. Soc. of Am.* 16(9):2136-2145, 1999.

[7] V. Ferrari, T. Tuytelaars and L. Van Gool. "Simultaneous Object Recognition and Segmentation by Image Exploration." *Proc. ECCV,* 2004.

[8] V. Ferrari, T. Tuytelaars and L. Van Gool. "Wide-baseline muliple-view correspondences." *Proc. CVPR,* 2003.

[9] B. Frey and N. Jojic. "Transformed Component Analysis: Joint Estimation of Spatial Transformations and Image Components." *Proc. ICCV*, 1999.

[10] C. Harris and M. Stephens. "A combined corner and edge detector." *Alvey Vision Conference,* 1988.

[11] M. Irani, B. Rousso, S. Peleg. "Computing Occluding and Transparent Motions." *Int. J. on Computer Vision.* 12(1):15-16, 1994.

[12] H. Jin, S. Soatto, A. Yezzi. "Multi-view stereo beyond Lambert." *Proc. CVPR,* 2003.

[13] S. Ju, M. Black, A. Jepson. "Skin and Bones: Multi-layer, Locally Affine, Optical Flow and Regularization with Transparency." *Proc. CVPR,* 1996.

[14] Y. Ke and R. Sukthankar. "PCA-SIFT: A More Distinctive Representation for Local Image Descriptors." *Proc. CVPR,* 2004.

[15] M. Langer, S. Zucker. "Shape From Shading on a Cloudy Day." *J. Opt. Soc. Am. A,* 11(2):467-478, 1994.

[16] T. Lindeberg. *Scale-Space Theory in Computer Vision.* Boston: Kluwer Academic Publishers, 1994.

[17] D. Lowe. "Distinctive image features from scale-invariant keypoints," *Int. J. on Computer Vision.* 60(2):91-110, 2004.

[18] K. Mikolajczyk, C. Schmid. "An affine invariant interest point detector." *Proc. ECCV,* 2002.

[19] K. Mikolajczyk and C. Schmid. “A performance evaluation of local descriptors.” *Proc. CVPR,* 2003.

[20] B. Sarel, M. Irani. “Separating Transparent Layers through Layer Information Exchange.” *Proc. ECCV,* 2004.

[21] F. Schaffalitzky and A. Zisserman. “Viewpoint invariant texture matching and wide baseline stereo,” *Proc. ICCV,* 2001.

[22] J. Shi and C. Tomasi. “Good Features to Track,” *Proc. CVPR,* 1994.

[23] R. Szeliski, S. Avidan, P. Anandan. “Layer Extraction from Multiple Images Containing Reflections and Transparency.” *Proc. CVPR,* 2000.

[24] C. Tomasi, T. Kanade. “Detection and tracking of point features.” Tech. Rept. CMU-CS-91132. Pittsburgh: Carnegie Mellon U. School of Computer Science, 1991.

[25] T. Tuytelaars, and L. Van Gool. “Matching Widely Separated Views based on Affine Invariant Regions,” *Int. J. on Computer Vision,* 59(1):61-85, 2004.

[26] T. Tuytelaars and L. Van Gool. “Wide Baseline Stereo based on Local, Affinely invariant Regions,” *British Machine Vision Conference,* 2000.

[27] Y. Yu, P. Debevec, J. Malik, T. Hawkins. “Inverse Global Illumination: Recovering Reflectance Models of Real Scenes from Photographs.” *Proc. of the AMS SIGGRAPH*, 1999.

[28] R. Zabih and J. Woodfill. “Non-Parametric Local Transforms for Computing Visual Correspondence.” *Proc. ECCV,* 1994.