Clustering Appearance for Scene Analysis

Sanjeev J. Koppal and Srinivasa G. Narasimhan
Robotics Institute, Carnegie Mellon University, Pittsburgh, USA
Email: (koppal,srinivas)@ri.cmu.edu

Abstract

We propose a new approach called "appearance clustering" for scene analysis. The key idea in this approach is that the scene points can be clustered according to their surface normals, even when the geometry, material and lighting are all unknown. We achieve this by analyzing an image sequence of a scene as it is illuminated by a smoothly moving distant source. Each pixel thus gives rise to a "continuous appearance profile" that yields information about derivatives of the BRDF w.r.t source direction. This information is directly related to the surface normal of the scene point when the source path follows an unstructured trajectory (obtained, say, by "hand-waving"). Based on this observation, we transform the appearance profiles and propose a metric that can be used with any unsupervised clustering algorithm to obtain iso-normal clusters. We successfully demonstrate appearance clustering for complex indoor and outdoor scenes. In addition, iso-normal clusters serve as excellent priors for scene geometry and can strongly impact any vision algorithm that attempts to estimate material, geometry and/or lighting properties in a scene from images. We demonstrate this impact for applications such as diffuse and specular separation, both calibrated and uncalibrated photometric stereo of non-lambertian scenes, light source estimation and texture transfer.

1 Why Cluster Appearance?

Our world contains scenes of vastly varying appearances. These appearances depend on several different factors such as materials, 3D shapes of scenes and lighting and viewing geometry. Extracting these properties from images (or image sequences) for scene analysis is an important inverse problem in vision. Unfortunately, these properties usually interact non-linearly and estimating them becomes difficult.

In order to make this problem tractable, several works have assumed prior knowledge of either lighting or BRDFs or scene structure. Methods that assume known lighting include Woodham's classical photometric stereo ([31]) for lambertian scenes, as well as several extensions for non-lambertian low parameter BRDFs, such as the micro-facet model and the dichromatic model ([13],[23],[8],[30],[2],[28],[17],[29],[16]). Particularly, in the work of Goldman et al ([8]), clustering of material properties is shown to help with scene analysis. Complementary to the above methods is the class of 'inverse rendering' algorithms that estimate low parameter BRDFs and lighting ([27],[18]) using scanned 3D scene geometry. Recently, Ramamoorthi's thesis ([24]) provides a formal analysis of when exactly inverse rendering is possible for general BRDFs and lighting that are represented using Spherical Harmonics. Finally, Hertzmann and Seitz, ([11]) recover the geometry of objects by estimating combinations of "basis example spheres" that best describe scene BRDFs.

In this work, we present a novel approach for appearance analysis of static scenes containing a broad range of BRDFs, *without requiring any knowledge about scene geometry, material properties, lighting or example calibration objects*[1]. Our approach involves two key steps: (a) dividing a complex scene into geometrically consistent clusters (scene points with same or very similar surface normals) irrespective of the material properties and lighting, and (b) bootstrapping scene parameter estimations with this partial geometric information. The number of unknowns is reduced within each cluster, resulting in a reduced dimensional and more robust optimization for appearance parameters. The closest related work is by Healey ([10]), who segments a Lambertian scene into regions that have the same local geometry, using two images. Our approach uses many more images to extend these results to a much larger class of BRDFs. We note that our work is part of a recent trend ([8],[11]) to split the problem of appearance analysis into smaller, more manageable parts.

Our main contribution is a physically based technique to obtain iso-normal clusters of a static scene illuminated by a smoothly moving distant (directional) light source. The video camera observing the scene is assumed to be orthographic. As the source moves, observations at each scene point over time result in a *continuous appearance profile*. We believe that the smoothness (continuity) of the appearance profile is a powerful notion that has not yet been fully exploited in computer vision[2]. We present a comprehensive analysis of the information contained in these derivatives of this profile (specifically, extrema or inflection points) and how they relate to the surface normal of a scene point.

Unfortunately, direct unsupervised clustering of appearance profiles is not sufficient to obtain geometrically consistent clusters since profiles vary significantly with different material properties. Instead, we show that simply "hand-waving" a light source in an uncontrolled fashion (along an unstructured trajectory) minimizes the dependency of the appearance profile on the material properties. Based on the analysis of extrema locations, we then apply a transformation to the appearance profile that reduces this material dependency further. Finally, we present a metric that matches appearance profiles of same surface normals and can be used with *any supervised or unsupervised clustering technique* to obtain robust geometrically consistent scene clusters.

We mathematically analyze our clustering algorithm for a large class of linearly separable BRDFs introduced by Narasimhan et al., [19]. Within this broad group of models, we explore

[1] In previous work, this was achieved only for simple BRDF models such as lambertian or Torrance-Sparrow ([22],[21],[7],[1]).

[2] Exceptions include the space-time stereo [32] and the work of Hayakawa ([9]) that uses an arbitrary moving light source to alleviate the ambiguity in photometric stereo for Lambertian objects.

Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)
0-7695-2597-0/06 $20.00 © 2006 IEEE

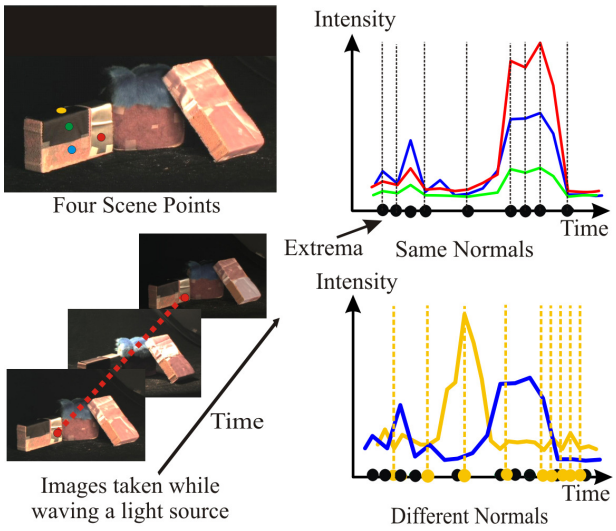


Figure 1. **Appearance Profile and Extrema:** Image sequence obtained by illuminating a static scene with a moving point source. The appearance profile of a scene point is the observed intensities at a single pixel over time. The appearance profiles show several extrema (peaks and valleys) as illustrated on the right. More often than not, scene points with same surface normal exhibit extrema at the same time instances. Similarly, most extrema locations do not match for scene points different normals. This makes extrema locations excellent features for our clustering algorithm.

more formally the notion of *orientation consistency*. This idea was first proposed by Hertzmann and Seitz ([11]) and used to compute the surface normal at a scene point by comparing it with an "example" object with known shape and BRDF. In constrast, we compute orientation consistencies *between* scene points of unknown normals and BRDFs, without requiring any example object. Of course, the trade-off here is that we require a longer sequence of images. Therefore, we also analyze the relationship between the length of the image sequence and the "complexity" of the BRDFs for a desired clustering error.

We experimentally demonstrate the accuracy of our clustering algorithm for several real scenarios with a broad range of materials. Our approach also shows good results with scenes not satisfying the assumptions of our method (anisotropic materials, outdoor scenes). Using iso-normal clusters makes estimating the lighting, geometry and material properties of a scene more robust. We demonstrate a range of applications such as separation of diffuse and specular components, calibrated and uncalibrated photometric stereo of non-lambertian scenes, estimation of light source directions, and lighting and geometrically consistent texture replacement. In addition, this approach can be also extended to methods like relighting and inverse rendering. Furthermore, because all our method requires is for a user to hand-wave a light source, image acquisition is very simple in contrast to methods that require complex illumination setups ([4]). Thus, we believe our approach to scene appearance analysis is broadly applicable in vision and graphics.

2 Appearance Profiles and their Extrema

Consider a static scene illuminated by a continuously moving distant point light source. An *appearance profile* is a vector of intensities measured at a pixel over time, as illustrated in Fig. 1. Direct clustering of these profiles fails, even after removing scale and offset, since the profile intensities are non-linear functions of geometric and material properties at a scene point. Hence, it is critical to obtain a feature from the appearance profile that is invariant (or insensitive) to material properties.

The continuity (smoothness) of an appearance profile yields information about the derivatives of the BRDF of the scene point w.r.t source direction. Our key observation is to exploit this smoothness by detecting *brightness extrema* (peaks and troughs), where the first order profile derivatives are zero. We will show analytical and empirical reasons why iso-normal appearance profiles exhibit the same extrema at the same time instance. (see Fig. 1). This makes extrema locations excellent features for clustering. In the remainder of this section, we will formalize this observation by mathematically analyzing extrema for a large class of BRDFs.

2.1 Extrema in Linearly Separable BRDF Models

The observed radiance of a scene point at a time instant t can be modeled as a dot product between "material terms" $\mathbf{M_i}$ (functions of material properties like diffuse and specular albedo, roughness) and "geometry terms" $\mathbf{G_i}$ (functions of surface normal $\mathbf{n}$, viewing $\mathbf{v}$, and illumination directions $\mathbf{s(t)}$) [19]:

$$E(t) = \sum_{i}^{k} \mathbf{M_i}(\rho, \sigma)\mathbf{G_i}(\mathbf{n}, \mathbf{v}, \mathbf{s(t)})\,. \qquad (1)$$

The above linearly separable BRDF model represents a broad class of BRDFs (since no specific expressions for $\mathbf{M_i}$'s or $\mathbf{G_i}$'s are assumed) and many well known BRDF models used in computer vision (lambertian, Oren-Nayar, dichromatic) are special instances of this model ([5],[6],[25],[12]).

The extrema of the appearance profile $E(t)$ are found by setting its first order derivative, $E^{'}(t)$ w.r.t time t to zero:

$$E^{'}(t) = \sum_{i}^{k} \mathbf{M_i}(\rho, \sigma)\mathbf{G_i^{'}}(\mathbf{n}, \mathbf{v}, \mathbf{s(t)}) = 0 \qquad (2)$$

One solution to this linear system occurs when $\forall \mathbf{i}\ \mathbf{G_i^{'}} = 0$, at a particular time instant t, irrespective of material terms $\mathbf{M_i}$ (which are assumed to be non-zero). These are precisely the extrema - which we call **Geometry-Extrema** - that we are interested in. All other solutions to Eq. 2 are **Material-Extrema** since their location depends on the material properties of the scene point $\mathbf{M_i}$'s. Two questions remain to be addressed: (a) How are the Geometry-Extrema related to surface normals? (b) How can we reduce the influence of Material-Extrema and increase the occurrences of Geometry-Extrema in appearance profiles? We will address these issues next.

2.2 Geometry-Extrema and Iso-Normal Clusters

Extrema are said to be *shared* between two appearance profiles if they occur at the same time instance in both profiles. Our key idea is to acquire images in a way that increases the number of shared Geometry-Extrema, while decreasing the number of shared Material-Extrema. Both goals are achieved by simply "hand-waving" the light source in an uncontrolled fashion (along an unstructured trajectory).

If the user does not plan a path, then hand-waving a light source is a random action. The light source changes its position smoothly, but randomly, at every time-step. Therefore the appearance profiles at every scene point are generated stochastically. We will now show, that as the number of shared Geometry-Extrema increases, the probability that two scene points have the same normal increases. As our results demonstrate, we require only a few Geometry-Extrema to achieve accurate clustering in practice.

Consider two scene points with appearance profiles $E_1(t)$ and $E_2(t)$ and normals $\mathbf{n_1}$ and $\mathbf{n_2}$ respectively. Let the Geometry-Extrema of these profiles occur at the same time instances $\{t_j \,|\, j = 1..g\}$. Then, we must decide whether or not $\mathbf{n_1}$ is equal to $\mathbf{n_2}$. For this, let us compute the probability $P(\mathbf{n_1} \neq \mathbf{n_2})$. From Eq. 2,

$$\forall \mathbf{i} \;\; \mathbf{G_i}^{'}(\mathbf{n_1}, \mathbf{s}(t_j)) = 0 \,; \quad \mathbf{G_i}^{'}(\mathbf{n_2}, \mathbf{s}(t_j)) = 0 \; . \tag{3}$$

Let there be R (finitely many) roots of the function $G_i^{'}(\mathbf{n_2}, \mathbf{s}(t_j))$. Trivially, one of them is $\mathbf{n_2} = \mathbf{n_1}$. Let us assume all roots are equally likely (later we relax this requirement). Thus, $P(\mathbf{n_1} \neq \mathbf{n_2}) = (R-1)/R$ for a single extrema and term of the model. Given that there are k terms in Eq. 2, and g shared Geometry-Extrema, the probability is bounded by

$$P(\mathbf{n_1} \neq \mathbf{n_2}) = \left(\frac{R-1}{R}\right)^{k\,g} . \tag{4}$$

As the number of shared Geometry-Extrema, g, increases, the probability of the two appearance profiles being generated by different normals decreases. In the argument above we assume that all roots are equally likely. Even if this is not the case, the probability of any one root occurring in a particular video frame is *less than 1*. Thus, the product of such probabilities over the entire sequence of captured frames is bound to converge to zero if a long enough sequence is acquired. Note, that the appearance profiles of the scene points from the same normal may be vastly different due to material terms. In the next section we address how to ensure that material extrema do not break the clustering algorithm.

2.3 Acquiring Extrema by Hand-waving

How do we distinguish between Geometry-Extrema and Material-Extrema without knowing the material or geometric properties of the scene points? The key idea is that passing the source directly over a scene-point's normal gives rise to a Geometry-Extrema by creating a maxima in foreshortening (see Section 3 in [15]). Since real scenes contain several different surface normals, an unstructured, random path (as opposed to a structured one) will eventually cross many normals,

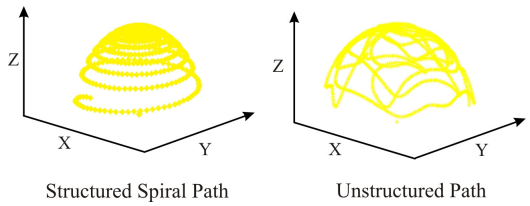


Figure 2. **Structured versus Unstructured Paths for Light Source.** Since scenes often contain many different surface normals, unstructured paths have a greater probability of inducing Geometry-Extrema, while keeping the acquisition process simple.

generating enough geometry-dependent extrema to create accurate clusters (see Fig. 2). This acquisition method is simple and does not require complicated illumination setups (such as in [4]).

While light source hand-waving produces shared Geometry-Extrema, it is critical that Material-Extrema are *not* shared since that would adversely influence the clustering algorithm described in Section 3. Consider two appearance profiles from two scene points with different local surface normals. We will show that the probability that these profiles share many Material-Extrema becomes significantly small over the course of a long hand-waving sequence.

From Eq. 2, if material extrema for two scene points with the same material terms $\mathbf{M}$ (written as a vector of $\mathbf{M_i}s$), but with different geometry derivative terms $\mathbf{K}^{'}$ and $\mathbf{L}^{'}$ (written as vectors of $\mathbf{G_i}^{'}s$), are coincident at time instance t_j then:

$$\mathbf{M}.\mathbf{K}^{'}(\mathbf{s(t_j)}) = 0; \quad \mathbf{M}.\mathbf{L}^{'}(\mathbf{s(t_j)}) = 0 \tag{5}$$

which means $\mathbf{K}^{'}(\mathbf{s}(t_j))$ and $\mathbf{L}^{'}(\mathbf{s}(t_j))$ both exist on the hyperplane in $\mathcal{R}_n$ defined by the normal $\mathbf{M}$. Now consider many time instances, $\{t_j | j = 1, 2...n\}$, where material extrema occur in both profiles. Since the $\mathbf{s}(t_j)$s are generated randomly and independently by waving a light source, the likelihood that all $(\mathbf{K}^{'}(\mathbf{s}(t_j)), \mathbf{L}^{'}(\mathbf{s}(t_j)))$ pairs occur on the same plane defined by $\mathbf{M}$ becomes small as n becomes large. We can apply a similar argument to scene points of different materials.

Therefore, for a long hand-waving sequence, material-extrema will not adversely effect our clustering algorithm. Note, that there may indeed be controlled and structured paths where our algorithm may work for particular scenes, but we believe these are rare and in any case, hard to prove and acquire. Thus, we advocate random paths over structured paths.

3 Algorithm to Create Iso-normal Clusters

The final clustering algorithm can be divided into four steps which are summarized in Table 1. These steps are **1)** collect images of a scene and detect brightness extrema, **2)** transform the appearance profiles (Figure 3) and **3)** use a common similarity metric to **4)** cluster the scene. Note that since we detect extrema occurrences using a moving window, we do not need to store the whole sequence of images in memory. Also our algorithm does not use the extrema locations by themselves, since different profiles may have different numbers of extrema. Instead, we transform the appearance profile by linearly interpolating between extrema locations, bounding the error due to Material-Extrema (see Sections 1 and 2 of our technical report [15]).

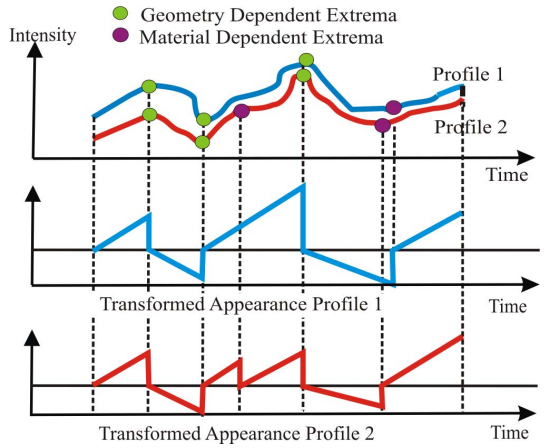


Figure 3. **Transformation Applied to Appearance Profiles:** This illustration shows the effect of transformation on two hypothetical appearance profiles. Consider the 'segments' between extrema. The slope of transformed profile is the sign of the first derivative of a segment. Therefore two segments that have positive first derivative (monotonically increasing), get the same positive slope of 1. Note that in segments where there are no material dependent extrema, the transformed values are identical.

Table 1

Step 1 (Input):

While acquiring frames by randomly waving a light,
Detect intensity extrema at each pixel and
store their occurrences in time.
(No need to store whole image sequence)
end

Step 2 (Transformation):

Construct a feature vector from each scene point's profile by piece wise linear interpolation of its extrema stored in Step 1 (Figure 3).

Step 3 (Metric):

Compute distance metric between (unit) feature vectors $\vec{A}$ and $\vec{B}$ using dot-product: Distance = 1 - $\vec{A}^t\vec{B}$.

Step 4 (Output):

Cluster the normals based on the metric in Step 3.

Table 1. **Algorithm:** Our method is simple to implement. The input to the algorithm is a sequence of scene images that are collected by hand-waving a light source. At each pixel, we only store the locations of all the brightness maxima and minima. We then linearly interpolate these extrema locations as shown in Figure 3. Therefore each pixel location is associated with a transformed profile. These profiles are then grouped using the dot-product dissimilarity metric with any clustering algorithm, such as k-means or hierarchical clustering.

Figure 4. Our acquisition setup with a Canon XL2 video camera, a 60 watt light attached to a wand. In real experiments the camera and light source are further away to satisfy orthographic assumptions.

Finally, we use the "dot-product" metric which is shown to be accurate for matching extrema locations of two profiles ([26]). Mathematically, if A and B are the transformed appearance profiles of two scene points, the "dot-product" metric is simply $1 - A^T\ B$.

Any number of sophisticated learning techniques (such as SVMs or spectral methods) can be used for the clustering part of our algorithm. However, the transformation and metric discussed above are powerful enough to allow the relatively simple k-means algorithm to produce accurate results.

4 Experiments: Simulated and Real Scenes

We will now demonstrate the accuracy of our algorithm using both simulations and a wide range of real indoor and outdoor scenes with complex scene structure and BRDFs.

Simulations: We performed extensive simulations with a scene containing 50 unique surface normals that were sampled from the hemisphere of directions. Our simulations spanned the *entire parameter* space of four models (Lambertian, Oren-Nayar, Torrance-Sparrow, Oren-Nayar + Torrance-Sparrow) with 20000 profiles per normal (we show only a few profiles in Fig. 5 for clarity). We recorded the extrema locations that were shared by over 95% of a normal's profiles. We compared these extrema locations across different normals using the dot-product measure and found that these shared extrema are unique to a particular normal. This supports our earlier idea that, compared to profiles from different normals, profiles from the same normal share more extrema. We also conducted experiments using our clustering algorithm, for simulated data of 8000 different profiles generated by the Torrance-Sparrow + Oren-Nayar model and with a scene of 60 different normals, getting 95% accuracy in iso-normal clustering.

Real Scenes: Our setup consists of a Canon XL2 digital video camera observing at a static scene as shown in Fig. 4. We first tested our algorithm on simple planar scenes consisting of real textures from the CURET project ([3]). Note the boxed regions at the top of Fig. 7. Our algorithm clustered all these textures together accurately, even though their materials properties were very different (specularities, roughness, 3D textures). In Fig. 8, our algorithm clusters anisotropic materials, implying that our method works even in some cases that do not satisfy our assumptions. We also show results for non-planar objects which contain an infinite number of normals. In these cases, our method degenerates gracefully by creating a piecewise approximation of the continuous curved surface.

In Fig. 9 we show more complex planar scenes, containing occlusion, cast shadowing and inter-reflection. In these regions, our method may over-cluster the scene, but note that the smaller clusters are still geometrically consistent. In Fig. 10, we show the clustering results obtained for outdoor images of a scene collected from the WILD database ([20]). We believe the diverse illumination due to weather (sunny, cloudy, fog, mist) creates appearance profiles with enough intensity variation to produce a good result.

Comparing two clustering algorithms: In Fig. 6 we show the results of experiments conducted to analyze and compare our clustering accuracy using two common unsupervised clustering methods, k-means and hierarchical clustering. We com-

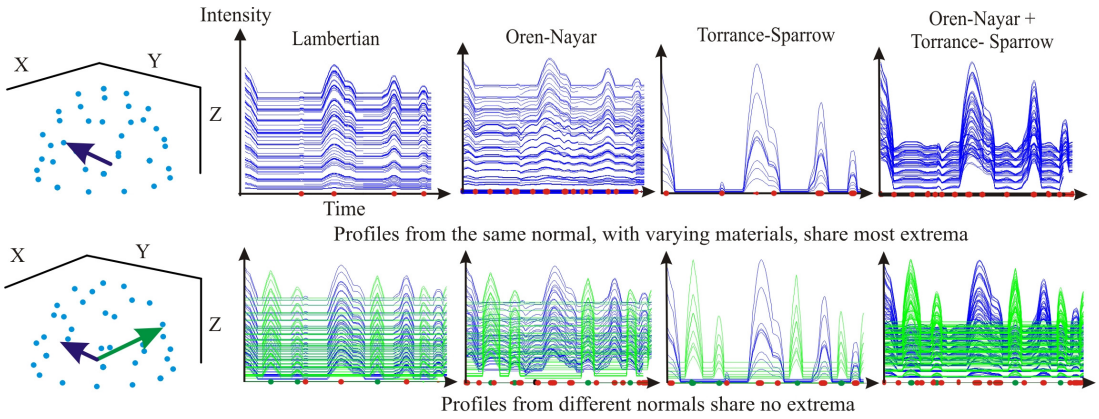


Figure 5. **Simulations showing the link between Extrema and Surface Normal:** Appearance profiles are simulated for four BRDFs over a range of 20000 material properties (only a few are shown for clarity). We only show two normals, although we simulated profiles for 50 (marked by blue dots on the hemisphere). The extrema location of a profile is marked on the x-axis by a colored dot. Note that profiles from the same local normal (top row) share most of the extrema locations, whereas profiles from different normals (bottom row) do not.

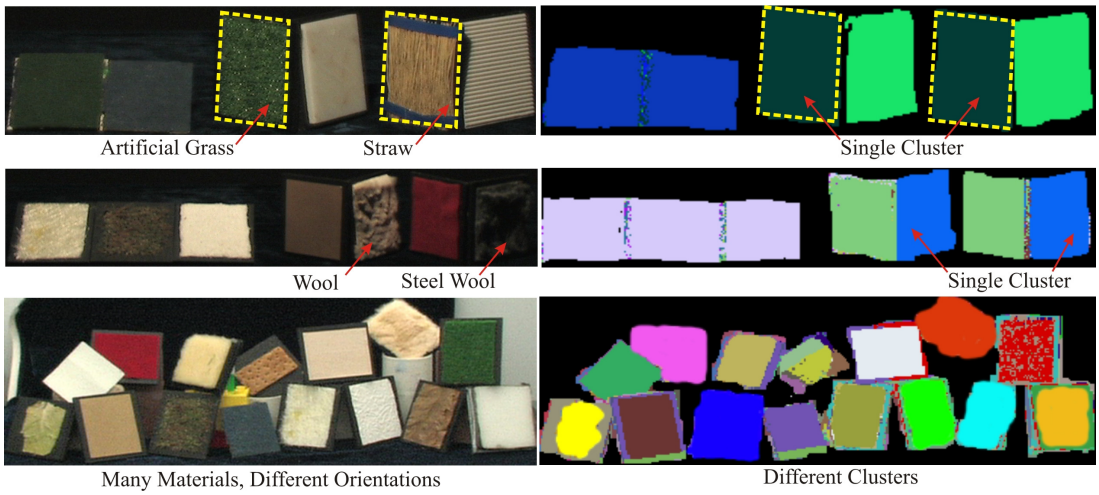


Figure 7. **Results obtained when our algorithm is used to cluster materials in the CURET Database.** We acquired image sequences of the real sample materials by waving a light source (and did not use the still images distributed by Columbia University). Notice the top row containing materials such as artificial grass and straw and the middle row with examples of real wool and steel wool. Despite significant appearance differences, these samples cluster together accurately because they share the same surface normal. **Please see video at [14] for better visualization.**

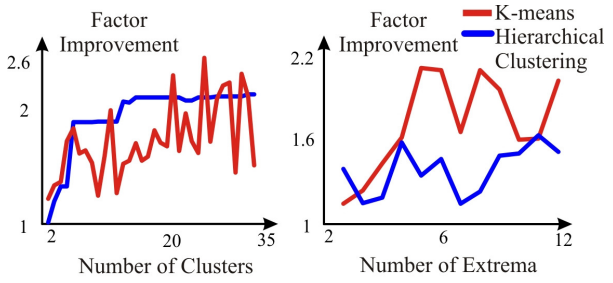


Figure 6. **K-means versus Hierarchical Clustering.** The factor of improvement is the ratio of the clustering accuracy using our metric to the clustering accuracy using the Euclidean metric. The left hand side plot shows the factor of improvement variation with increase in number of clusters. As expected, this graph plateaus due to over clustering. The right hand side plot shows the factor of improvement obtained for different number of extrema used in the appearance profiles. Note that the k-means graph is jagged because initialization is non-deterministic. In both cases, our recommended metric performs significantly better than the Euclidean metric.

pared the results obtained using both our dot-product metric and the commonly used Euclidean distance metric. We plot the factor of improvement achieved with varying numbers of clusters and numbers of extrema used in clustering. In all cases, our recommended metric shows significant improvement. We have used k-means in all our real experiments and obtained accurate results. We note that, in addition to k-means, our method can be used with any sophisticated clustering techniques.

5 Scene Analysis using Geometry Clusters

Thus far, we described our method to cluster a scene into regions of same (or similar) surface normals. We will now demonstrate how this clustering can be used to estimate important scene properties such as shape and material properties, and light source directions.

The key benefits of clustering geometry of the scene *before* estimating the appearance properties are two-fold. First, the total number of unknowns is reduced, making the inverse estimation easier; only one surface normal per cluster needs to be

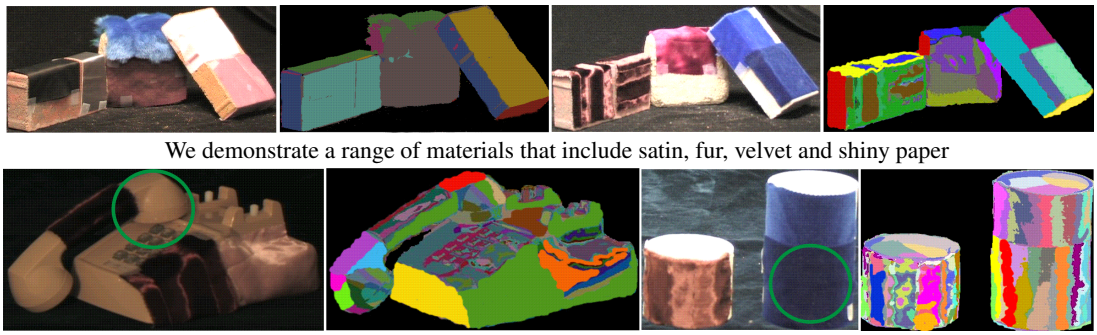


Figure 8. **Clustering curved surfaces with complex (possibly anisotropic) materials** When anisotropic BRDFs are present in the scenes, our method still produces meaningful clusters. Furthermore, for curved surfaces, our method produces a piecewise planar approximation.

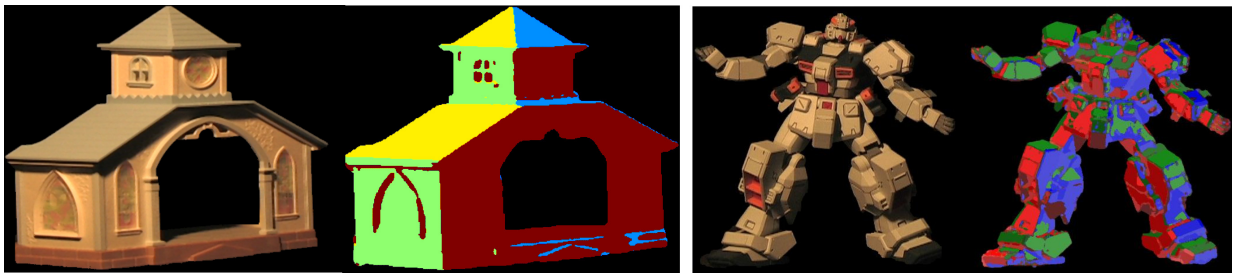


Figure 9. **Clustering surfaces with cast shadows.** When complex effects such as cast shadows and inter-reflections are present in the scenes, our method works for simpler scenes such as on the left. For more complex scenes, such as on the right, our method may fail to group all pixels in the scene that have the same normal. Instead, the algorithm simply over clusters the scene into smaller iso-normal clusters.

estimated instead of one normal per scene point. Second, clustering provides important spatial coherency in scene geometry. It is common practice in computer vision to use smoothness priors to ensure stability of any estimation algorithm. *Analogously, iso-normal clusters are excellent priors to estimate scene geometry*.

Consider the $2k$ terms of $\mathbf{G_i}$'s and $\mathbf{M_i}'s$ in the linearly separable BRDF model of Eq. 1 that describe the intensity at any pixel. Let these terms be represented by analytic forms of specific models (say, Lambertian, Oren-Nayar, Torrance-Sparrow, etc see Table 1 in [19]). Estimating the $\mathbf{G_i}$'s and $\mathbf{M_i}'s$ allows us to explicitly estimate scene properties such as surface normals, albedos, and source directions.

Note that all pixels within a cluster share the same normal and hence, the same values for $\mathbf{G_i}$'s. Now, consider the observed intensities of the pixels within a cluster over k frames, where k is the number of terms in Eq. 1. These measurements can be written in matrix form as:

$$\begin{pmatrix} M_{11} M_{12} ... M_{1k} \\ . \\ . \\ . \\ M_{p1} M_{p2} ... M_{pk} \end{pmatrix} \begin{pmatrix} G_{11} ... G_{k1} \\ . \\ . \\ . \\ G_{1k} ... G_{kk} \end{pmatrix} = \begin{pmatrix} E_{k1} \\ . \\ . \\ . \\ E_{kp} \end{pmatrix} \quad (6)$$

where M_{ij} is the j^{th} material term at pixel i, G_{pq} is the p^{th} geometry term at frame q, and E_{kl} is the appearance profile of length k at pixel l.

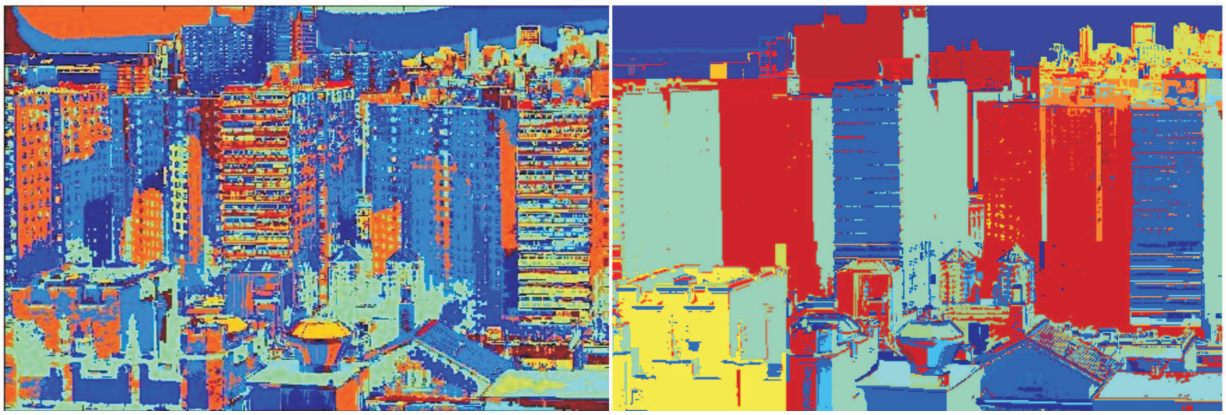
The estimation of the model parameters consists of two steps, **(a)** Start with an initial guess for the k^2 geometry terms G_{ij} and then use non-negative least squares to solve the above linear system to obtain the kp material terms M_{ij} within each cluster, and **(b)** Use the computed material terms to estimate the geometry terms for all the frames of the sequence (not just k frames).

We iterate this alternate estimation material and geometry terms until convergence. In many vision algorithms, $k = 2$ suffices and the above problem requires a 4-dimensional optimization. We now demonstrate the accuracy of the algorithm for the case $k = 2$ using several applications. Note that there exists an ambiguity in the estimation of $\mathbf{M}$ and $\mathbf{G}$. In our example application, this ambiguity is resolved by constraining the form of the material ($\mathbf{M}$) and geometry ($\mathbf{G}$) terms. For example, in the 2-dimensional case, we constrained the first term to be Lambertian ($\mathbf{M_1} = \rho$, the albedo, and $\mathbf{G_1} = \vec{n}.\vec{s}$, the foreshortening term). We will leave a more formal treatment of the optimization for $k > 2$ for future work.

Separating Diffuse and Specular Components: On the left of Figure 11 we show a frame from a real video sequence. We use the 4-dimensional estimation described above to separate this sequence into diffuse and specular components. The center and right images are frames from the two extracted sequences. Note that the large highlight present in the leftmost book has been completely removed in the diffuse image.

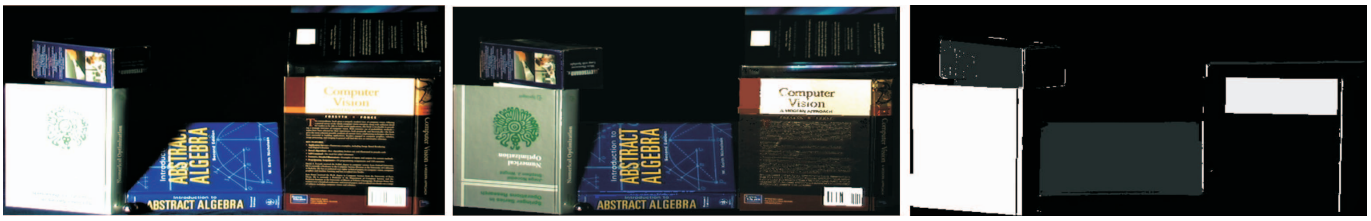
Extracting 3D Shape: Once we extract the diffuse component from a video sequence, we can apply algorithms that assume a Lambertian model. One example of such an algorithm is calibrated photometric stereo. In Figure 12 we show a few frames of a video sequence of a cup. Note that there is a sharp specularity and the cup is not lambertian. After clustering and

Images of an Outdoor Scene



K-Means Our Method

Figure 10. **Clustering WILD Database:** Note the complex appearance effects that occur in this data set. Our transformation of the appearance profile and the metric does significantly better than using Euclidean distance metric. In both cases, k-means was used to cluster appearance profiles. Note: some sub-clusters were merged by user for better viewing only. **Please see video at [14] for the variation in appearances in the input image sequence.**



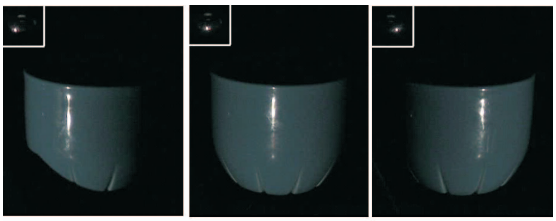
Real Image Diffuse Component Specular Component

Figure 11. **Separation of Specular and Lambertian Components:** When the number of components that make up the input video sequence is 2, then we can set up a 4-dimensional optimization to extract diffuse and specular terms as shown.

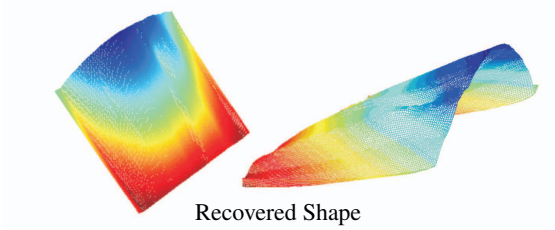
extracting the diffuse component, 3D shape is estimated with photometric stereo using measured light directions.

We also used Hayakawa's method for uncalibrated photometric stereo ([9]) with extracted diffuse components. In Figure 13, we were able to estimate the structure from the normals with no input light directions. This method also gives us lighting directions up to a rotation, and we display them at the bottom of the figure. Please view [14] for many more views of these estimated 3D shapes.

Appearance Consistent Texture Transfer: Profiles in appearance clusters share the same intensity extrema at the same time. Copying profiles from one part of an appearance cluster to another creates new pixels that vary *consistently* within their cluster. In Figure 14 is a screen shot is shown where the replaced profiles come from pixels in the same cluster with different texture. The complex appearance effects of the materials are preserved through the length of the video sequence.

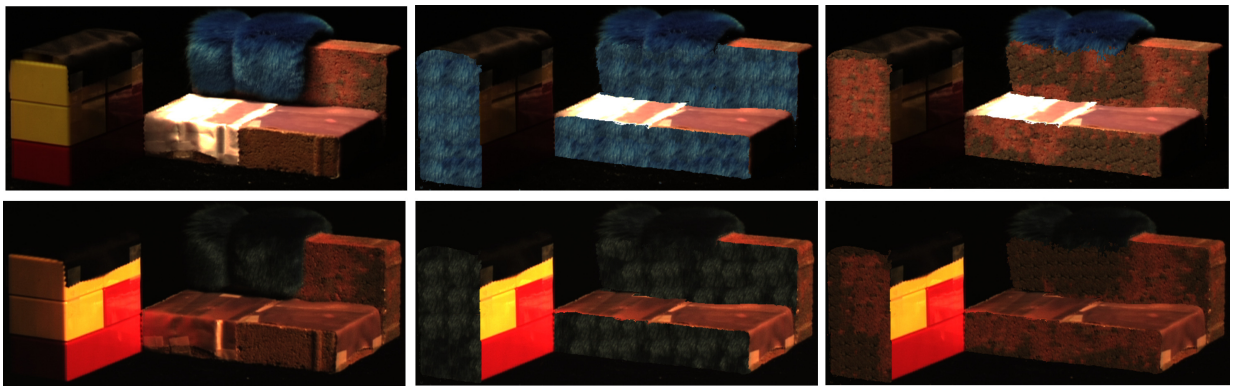


Non-Lambertian Scene



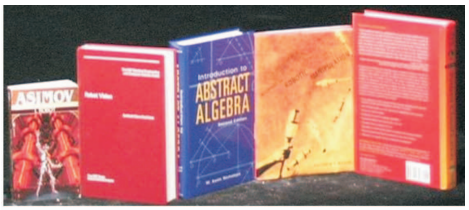
Recovered Shape

Figure 12. **Extracting Scene Structure:** We extract the Lambertian terms from a scene and apply Photometric Stereo. Integrating the normals gives us 3D shape. We show two views of the structure, and **for more views on this and other results please see [14]**.

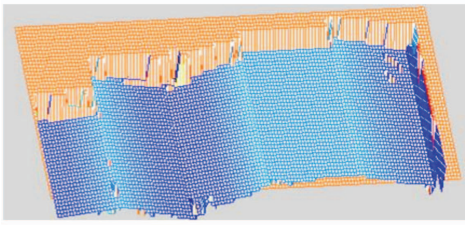


Two Lighting Directions Fur Transfer Brick Transfer

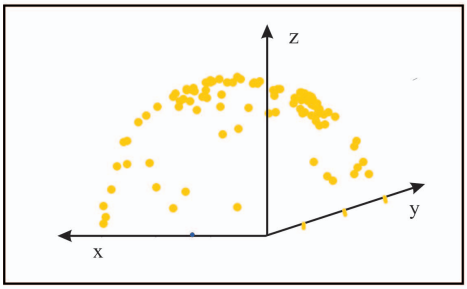
Figure 14. **Texture transfer** of complex materials (such as fur, and brick) between similar surface normals in a scene. A patch of the original scene is chosen by the user and a simple repetitive texture synthesis method is used to transfer this patch onto other areas of the scene with the same surface normal. Note the consistency in geometry and lighting in the transferred regions. **Please see video at [14] for many more lighting variations.**



One Image From Input Sequence



Computed Structure



Estimated Light Sources

Figure 13. **Uncalibrated Photometric Stereo:** Our clustering and optimization allow algorithms that assume diffuse model to work with non-Lambertian objects. Here we use Hayakawa's method ([9]) to get the 3D structure of the books and the corresponding lighting. Note that we obtain only the normals from photometric stereo, and we have to compute the book planes in an extra step.

6 Conclusions

In this paper we described how the derivatives of BRDF (encoded as extrema locations) are related to scene geometry. We demonstrated an algorithm to exploit these extrema to create iso-normal clusters of a scene and to use these clusters for effective scene analysis. Our algorithm has no prior information about geometry, material or light sources. We believe that our method is simple and has several applications for vision and graphics.

7 Acknowledgements

This research was supported by NSF Awards #CCF-0541230 and #CCF-0541307, and an ONR Award #N00014-05-1-0188. The authors thank Alexei Efros for early technical discussions and Shree Nayar for providing the CURET textures ([3]).

References

[1] R. Basri and D. W. Jacobs. Photometric stereo with general, unknown lighting. *CVPR*, 2001.

[2] E. Coleman and R. Jain. Obtaining 3-dimensional shape of textured and specular surfaces using four-source photometry. *Intl Conf. Color in Graphics and Image Processing*, 1982.

[3] K. J. Dana, B. V. Ginneken, S. K. Nayar, and J. J. Koenderink. Reflectance and texture of real world surfaces. *CVPR*, 1997.

[4] P. Debevec, T. Hawkins, C. Tchou, H. Duiker, W. Sarokin, and M. Sagar. Acquiring the reflectance field of a human face. *In Proc. SIGGRAPH*, 2000.

[5] J. DeYoung and A. Fournier. Properties of tabulated bidirectional reflectance distribution. *Graphics Interface*, 1997.

[6] A. Fournier. Separating reflection functions for linear radiosity. *Eurographics Workshop on Rendering*, 1995.

[7] A. S. Georghiades. Recovering 3-d shape and reflectance from a small number of photographs. *Eurographics Workshop on Rendering*, 2003.

[8] D. Goldman, B. Curless, A. Hertzmann, and S. Seitz. Shape and spatially-varying brdfs from photometric stereo. *ICCV*, 2005.

[9] H. Hayakawa. Photometric stereo under a light-source with arbitrary motion. *JOSA*, 1994.

[10] G. Healey and L. Z. Wang. Segmenting surface shape using colored illumination. *SCIA*, 1997.

[11] A. Hertzmann and S. Seitz. Shape and materials by example: a photometric stereo approach. *CVPR*, 2003.

[12] J. Kautz and M. D. McCool. Interactive rendering with arbitrary brdfs using separable approximations. *Eurographics Workshop on Rendering*, 1999.

[13] G. J. Klinker, S. A. Shafer, and T. Kanade. A physical approach to color image understanding. *IJCV*, 1990.

[14] S. J. Koppal and S. Narasimhan. Appearance clustering video. *http://www.cs.cmu.edu/∼koppal/cvpr.mp4*, 2006.

[15] S. J. Koppal and S. G. Narasimhan. Appearance clustering: A novel approach to scene analysis, *Carnegie Mellon University Technical Report TR-06-14*. 2006.

[16] S. Mallick, T. Zickler, D. Kriegman, and P. Belhumeur. Beyond lambert: Reconstructing surfaces with arbitrary brdfs. *ICCV*, 2001.

[17] S. Mallick, T. Zickler, D. Kriegman, and P. Belhumeur. Beyond lambert: Reconstructing specular surfaces using color. *ICCV*, 2005.

[18] S. Marschner, S. Westin, E. Lafortune, K. Torrance, and D. Greenberg. Image-based brdf measurement including human skin. *Eurographics Workshop on Rendering*, 1999.

[19] S. G. Narasimhan, V. Ramesh, and S. K. Nayar. A class of photometric invariants: Separating material from shape and illumination. *ICCV*, 2003.

[20] S. G. Narasimhan, C. Wang, and S. K. Nayar. All the images of an outdoor scene. *ECCV*, 2002.

[21] S. K. Nayar, K. Ikeuchi, and T. Kanade. Determining shape and reflectance of hybrid surfaces by photometric sampling. *IEEE Transactions on Robotics and Automation*, 1990.

[22] S. K. Nayar, K. Ikeuchi, and T. Kanade. Surface reflection: Physical and geometrical perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1991.

[23] M. Oren and S. K. Nayar. Generalization of the lambertian model and implications for machine vision. *IJCV*, 1995.

[24] R. Ramamoorthi and P. Hanrahan. A signal-processing framework for inverse rendering. *SIGGRAPH*, 2001.

[25] S. Rusinkiewicz. A new change of variables for efficient brdf representation. *Eurographics Workshop on Rendering*, 1998.

[26] G. Salton. Automatic text processing: The transformation, analysis, and retrieval of information by computer. 1989.

[27] Y. Sato, M. D. Wheeler, and K. Ikeuchi. Object shape and reflectance modeling from observation. *SIGGRAPH*, 1997.

[28] A. Shashua. On photometric issues in 3d visual recognition from a single 2d image. *IJCV*, 1997.

[29] H. D. Tagare and R. J. P. deFigueiredo. A theory of photometric stereo for a class of diffuse non-lambertian surfaces. *IEEE Transactions on PAMI*, 1991.

[30] K. E. Torrance and E. M. Sparrow. Theory for off-specular reflection from roughened surfacesw. *JOSA*, 1967.

[31] R. J. Woodham. Photometric stereo. *MIT AI Memo*, 1978.

[32] L. Zhang, B. Curless, and S. Seitz. Spacetime stereo: Shape recovery for dynamic scenes. *In Proc. CVPR*, 2003.