Two-and-a-Half Dimensional Acquisition and Rendering

Zhou Xue
LCAV, I&C, EPFL

***Abstract*—In this research proposal, we revisit the classic rendering problem for multiple views and illuminations. With the purpose of recreating interactive viewing experience of artwork (oil painting, wood cut, etc), we simplify the multi-view/light acquisition and rendering in 3D to a two-and-a-half dimensional problem, which assumes generally flat object and limited view range. We will start with the classic rendering method Polynomial Texture Maps (PTMs) [6] which creates realistic rendering results in full lighting space, then synthesize the rendering results for virtual views with depth information recovered from surface gradients by shapelet [4]. For better understanding and exploration of the correlation between texture and structure, we turn to 3D textons which gives powerful representations and effective schemes for constructing and using vocabulary of 3D textons [5]. Conclusively, the report discuss the sampling and rendering problem in two-and-a-half dimensions and proposes some potential research directions for the future work.**

***Index Terms*—Polynomial Texture Maps, shapelets,3D texton, Two-and-a-Half Dimensions**

I. Introduction

THERE is a world of difference between seeing an oil painting and viewing its image. Viewer miss the opportunity to reveal and experience the creation process and the subtle geometry details by simply changing the view angle. Besides, visual impression of an artistic painting is also strongly influenced by the intensity and spectral profile of the illumination. The absence of choices of illumination mode and view angle breaks an artwork into a set of pixel values. Therefore, we intend to recreate the interactive viewing experience, by reproducing the visual appearance under different viewing and lighting conditions.

Nayar [7] managed to achieve the relighting performance with specially designed display device, which is quite similar to our purpose. They proposed a 2D measurement method of the illumination field and produced a virtual image of a scene in response to it. Since they have a large set of images under controllable lighting conditions, PCA is applied to characterize the main components and represent the relighting results. However, it can only provide the rendering output for single view.

We are more ambitious for our rendering system. Our rendering should be fully aware of the environmental conditions and the viewers. The response have to be realtime to deliver immersive and realistic user experiences. Since we want to render the artwork through portable devices such as cell phones and tablets which have very limited computational resources, we need to find the balance between efficiency and performance. Therefore, the representation of artwork under multiple view and light conditions become the core problem to address. It not only directly decides the rendering method, but the acquisition strategy as well.

In the previous semester, we developed a prototype for the two-and-half dimensional rendering and it comprises three main parts: measure the environmental factors (view angle and illumination) contributing to the visual results; represent the image dataset under varying view/light conditions; implementing the realtime rendering algorithm on the Android platform. The first part can be easily achieved by illumination and face detection with the front camera on Android Tablet. The only issue is the balance between computational cost and performance. For the representation of image sequence, we only use images from one fixed camera under varying lights to simplify the acquisition. Polynomial texture map (PTM) [6] is used to model the variation under different light sources while shapelets [4] reconstruct the depth map to synthesize rendering results for virtual views. We have successfully delivered realistic rendering performances with only 9 parameters for each pixel. With the compact representation, it is very simple to implement the rendering in realtime with the help of GPU on the tablet device. Therefore, while we are trying to find

Proposal submitted to committee: August 27th, 2012; Candidacy exam date: September 3rd, 2012; Candidacy exam committee: Exam president, thesis director, co-examiner.

This research plan has been approved:

Date: _____

Doctoral candidate: _____
(name and signature)

Thesis director: _____
(name and signature)

Thesis co-director: _____
(if applicable) (name and signature)

Doct. prog. director: _____
(R. Urbanke) (signature)

new representations to better model the reflectance, we need to make sure they still keep the feature to be parallelized for realtime applications.

In this research proposal, we will give an overview of three research papers which are highly related to our work. They not only provide direct methodology to tackle our problem, but inspire us for the future work. The rest of the report is organized as follows: In section II we will review the polynomial texture maps [6] which is the starting point to tackle the rendering problem. Then we will introduce the tool [4] for synthetic views from the PTMs in section III. The third paper [5] discussed in section IV gives us a better understanding of how to model visual appearance across light/view. Finally, in section V, we will summarize our current work and discuss the directions of future research.

II. Polynomial Texture Maps [6]

In this section, we give a review on [6]. The paper proposes a texture mapping method which uses biquadratic polynomial presentation for each texel and provides plausible relighting results. The Polynomial Texture Maps (PTMs) uses an image-based acquisition strategy which successfully captures variations due to surface self-shadow and inter-reflections, therefore increases the realism. PTMs is also found to be useful for anisotropic and Fresnel shading models and multiple focus depth images.

The acquisition of PTMs is very practical. They only need a set of images under single directional light source from different angles. This image-based technology approximates each pixel independently without any complex geometry information. The biquadratic polynomial representation is compact and performs rendering in realtime. In the following sections, the implementation details will be introduced. We will also discuss its limitation for degrading high frequency components and incompetence for multiple view rendering.

A. Background of Photo-realistic Rendering

The rendering performance heavily relies on the characterization of surface reflectance properties. In [8], Nicodemus describes the ratio of the reflected viewing illumination to the incident light per unit area with a fundamental Bidirectional Reflectance Distribution Function $BRDF(\Theta_i, \Phi_i, \Theta_e, \Phi_e, \lambda)$. It uses (Θ_i, Φ_i) to represent the incident light, (Θ_e, Φ_e) the exitant view and λ the wavelength. It is further extended to a spatially varying version with parameter (u, v) to represent the position on the surface [1]:$BTF_{r,g,b}(\Theta_i, \Phi_i, \Theta_e, \Phi_e, u, v)$. They manage to describe the reflectance with image-based technology. However it requires a large number of sampling images and sophisticated calibration which is a limitation in practice.

To keep the acquisition and rendering practical, PTMs sacrifices the freedom of view and the specular components. They assume a diffuse surface or reflectance excluded of specular parts. However, they still retain a good approximation for cast shadow and diffuse shading.

B. Biquadratic Polynomial Representation for Texture

A good representation for image sequence has to take full advantage of the redundancy mong the images. One apparent redundancy among multi-lighting images is the fairly constant chromaticity across the whole sequence. The change of the visual appearance largely results from the varying illumination. Therefore, the multi-lighting images can be modeled as a texel $(R_n(u,v), G_n(u,v), B_n(u,v))$ modulated by a luminance model $L(u,v)$:

$$\begin{aligned} R(u,v) &= L(u,v)R_n(u,v) \\ G(u,v) &= L(u,v)G_n(u,v) \\ B(u,v) &= L(u,v)B_n(u,v) \end{aligned} \tag{1}$$

where (u, v) stands for the texture coordinates. This model can be achieved with lower cost in color space like LUV and YC_bC_r, or higher cost for RGB channels respectively. Regardless of the color space, the illumination is represented with following biquadratic model for each texel:

$$\begin{aligned} L(u,v;l_u,l_v) = a_0(u,v)l_u^2 + a_1(u,v)l_v^2 + a_2(u,v)l_u l_v \\ + a_3(u,v)l_u + a_4(u,v)l_v + a_5(u,v) \end{aligned} \tag{2}$$

where (l_u, l_v) are light projections on the object surface, L is the resultant luminance at (u, v) and $\{a_i\}_{i=0}^5$ are coefficients of the polynomial fitting to the dataset for each pixel. The representation of equation (2) well characterize the smoothness widely existing in light space, even for the areas with high spatial frequencies. With $N+1$ images under different lighting conditions, the equation (3) can be used to calculate $\{a_i\}_{i=0}^5$.

$$\begin{bmatrix} l_{u0}^2 & l_{v0}^2 & l_{u0}l_{v0} & l_{u0} & l_{v0} & 1 \\ l_{u1}^2 & l_{v1}^2 & l_{u1}l_{v1} & l_{u1} & l_{v1} & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ l_{uN}^2 & l_{vN}^2 & l_{uN}l_{vN} & l_{uN} & l_{vN} & 1 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_5 \end{bmatrix} = \begin{bmatrix} L_0 \\ L_1 \\ \vdots \\ L_N \end{bmatrix} \tag{3}$$

Apparently, singular value decomposition (SVD) gives the best fit in the sense of ℓ_2 norm. The smoothing effect is only for light space, so no spatial blur will be introduced during the approximation procedure. Correspondingly, the hard shadow will become soften, the specularity will be neglected and the point light source will become area light. However, with uniformly chosen lighting positions, the approximation of PTMs retains the property of self shadowing, sub-surface scattering and inter-reflections to some extent. It gives particularly impressive rendering results when the rendering light is around original samples. This is also the reason why they sample the lighting space uniformly.

To better store and implement PTMs, the authors introduce two parameters: scale λ and bias ω for $\{a_i\}_{i=0}^5$ to eliminate the difference between different orders. All the coefficients are transformed to 8-bit integers which not only solves the storage issue, but speeds up the hardware efficiency as well. Fix point operations and parallel optimization make PTMs a very practical rendering method. Furthermore, since light dependency is linear, the sample rate can be very flexible with mip-maps. The authors also show the PTMs can be used to

render a lot of special effects such as anisotropic materials, Fresenel effects and Phong lighting equation.

Besides multi-lighting images, PTMs can also be generated from bump maps. The reflectance function for bump map can be represented as:

$$F(l_u, l_v) = l_u N_u + l_v N_v + \sqrt{1 - l_u^2 - l_v^2} N_w \tag{4}$$

where $\mathbf{N} = (N_u, N_v, N_w)$ stands for the surface normal. To approximate $F(l_u, l_v)$ with a biquadratic polynomial means to optimize the object function as follows:

$$\int_{-1}^{1}\int_{-1}^{1} (L(l_u, l_v) - F(l_u, l_v))^2 \, dl_u \, dl_v \tag{5}$$

To minimize equation (5), the PTMs basis $B = \{1, l_u, l_v, l_u l_v, l_u^2, l_v^2\}$ is transformed into the orthonormal basis:

$$B^{'} = \{\frac{1}{2}, \frac{\sqrt{3}}{2} l_u, \frac{\sqrt{3}}{2} l_v, \frac{3}{2} l_u l_v, \frac{\sqrt{45}}{4} l_u^2 - \frac{\sqrt{45}}{12}, \frac{\sqrt{45}}{4} l_v^2 - \frac{\sqrt{45}}{12}\} \tag{6}$$

Then the best coefficients $a_i^{'}$ can be computed directly with $L_i^{'}$, the i-th component of $B^{'}$:

$$a_i^{'} = \int_{-1}^{1}\int_{-1}^{1} L(l_u, l_v) F(l_u, l_v) \, dl_u \, dl_v \tag{7}$$

Then for each surface normal, the PTM coefficients are obtained as follows:

$$\begin{aligned} a_0 &= \frac{1}{2} a_0^{'} - \frac{\sqrt{45}}{12}(a_4^{'} + a_5^{'}), \; a_1 = \frac{\sqrt{3}}{2} a_1^{'} \\ a_2 &= \frac{\sqrt{3}}{2} a_2^{'}, \qquad\qquad a_3 = \frac{3}{2} a_3^{'} \\ a_4 &= \frac{\sqrt{45}}{4} a_4^{'}, \qquad\qquad a_5 = \frac{\sqrt{45}}{4} a_5^{'} \end{aligned} \tag{8}$$

Note that most of converting operations for bump map can be performed before the actual rendering. We need the previous calculations equation (7) and equation (8) to reconstruct a lookup table for sample normal $\mathbf{N}$ across the hemisphere. Afterwards, to render a bump map, we only need to compute the normal from bump map height and get the PTM coefficients from the lookup table.

C. Applications of PTMs

There are more information can be explored from the coefficients of PTMs. Firstly, inference of surface normal from PTMs can be solved by setting $\frac{\partial L}{u} = \frac{\partial L}{v} = 0$ and get solutions as follows:

$$l_{u0} = \frac{a_2 a_4 - 2 a_1 a_3}{4 a_0 a_1 - a_2^2} \tag{9}$$

$$l_{v0} = \frac{a_2 a_3 - 2 a_0 a_4}{4 a_0 a_1 - a_2^2} \tag{10}$$

With the normal map, they can enhance the geometry details by adding specular components to the rendering results. it can also be used for enhancement of diffuse objects as:

$$\begin{aligned} a_0^{'} &= g a_0, \; a_1^{'} = g a_1, \; a_2^{'} = g a_2 \\ a_3^{'} &= (1-g)(2 a_0 l_{u0} + a_2 l_{v0}) + a_3 \\ a_4^{'} &= (1-g)(2 a_1 l_{u0} + a_2 l_{v0}) + a_4 \\ a_5^{'} &= (1-g)(a_0 l_{u0}^2 + a_1 l_{v0}^2 + a_2 l_{u0} l_{v0}) + \\ &\quad (a_3 - a_3^{'}) l_{u0} + (a_4 - a_4^{'}) l_{v0} + a_5 \end{aligned} \tag{11}$$

where g is the factor for diffuse gain. The PTMs can also be used to expand the light space to render "impossible" images.

Moreover, applications of PTMs are not limited to the lighting related problems. The authors also use the biquadratic polynomial to represent multi-focus images and time-varying sequence. Although it is only an approximation method without theoretically proof, it still provides very impressive performance for these image sequences.

D. Acquisition Details for PTMs

Since we need to set up the PTM acquisition for our own rendering algorithm, we give a short review of the PTM systems. The essence of PTMs acquisition is to take photos under varying lighting conditions. The original lighting device is made out of hot-melt glue and wooden dowels in the shape of a subdivided icosahedron. They fix a camera facing down directly above the dome and use a hand held lamp for lighting at each triangular face to collect 40 images. They update the original dome to an automatic one which incorporates 50 custom-made flash boards under hardware control. A light-arm is used by the National Gallery in London to collect PTMs which has 12 commercial photographic flashes and the flexibility to rotate. A portable device is constructed for collecting PTMs of footprints, which uses flash bulbs as the illumination source and the arm is manually rotated to vary incident lighting direction in one axis. A very inexpensive PTM setup uses point-and-shoot digital camera to extend the flash unit. It is moved to various cutouts in a Styrofoam hemisphere to vary lighting direction.

Cultural Heritage Imaging (CHI) has built several successful acquisition systems for PTMs. They build a PTM dome that allows careful control of light direction and its color spectrum. The sophisticated design aims at avoiding the damage to the sensitive material on the artwork at certain light wavelengths. They use this dome to image fragile wax and lead seals attached with string to medieval documents, and oil paintings. Multi-spectral imaging has also been done using this equipment. They also have some simple photographic assembly and real time system which flashes 8 LEDs at 500 f/second while a synchronized high speed video camera captures frames.

When implementing the acquisition system, we choose a much simpler strategy. We fix the camera on the tripod facing to the painting. A sphere is placed next to the painting to calibrate the light direction. We also use a string to maintain a constant distance between the painting and hand-held flash. The flash is fired on the virtual dome uniformly facing the painting to have an unbiased data set. Based on our experimental results, 30 photos are definitely enough to provide a decent rendering result.

E. Conclusions

PTMs provides realistic rendering performances with a very low cost in storage and computations. However, it fails to capture the variation of geometric shadows and the specular components. The biquadratic polynomial is a decent approximation but only theoretically sound for some objects. There are potential models such as B-spline which can improve the approximation for reflectance function. The only issue lies in the acquisition part. More complicated acquisition strategy is a must for sophisticated models. Furthermore, the fixed view acquisition limits the rendering for multiple views. We use a strong assumption that the object is Lambertian and has a uniform luminance for all views. It holds for the most areas of our diffusive object except cast shadows and occlude areas. This dilemma inherits from the scarification of view freedom to simplify acquisition. Therefore, the main challenge is still the sampling strategy which has big influence on the representation and rendering.

III. Shapelets Correlated with Surface Normals Produce Surfaces [4]

In this section, we review an approach which reconstruct the 3D depth from the surface gradient with basis functions, referred here as shapelets. Unlike the typical reconstruction from the integration of surface gradients, this method is robust to noise, independent from the integration path and able to accommodate ambiguity in tilt. Most importantly, the surface reconstruction is simply a summation of correlations between the gradients of shapelets and the gradients of the object surface.

These features meet our needs for reconstructing depth map from PTMs perfectly. The basis functions used in reconstruction are linear and finite support. Therefore it provides the possibility for us to embed the reconstruction in the rendering operations. Furthermore, the shapelets provide decent results for our specific objects and lay no extra computational burdens.

A. Background for Shape from Surface Gradients

Kovesi [4] addresses the problem of deducing depth map from surface normals, which are normally generated from shape from shading, photometric stereo or shape from texture algorithms. These surface gradients suffer from noise and usually ill-conditioned. Although the intuitive solution would be direct integration, the discontinuity of surface will lead to different results because of the choice of different integration path, let alone the noise wide existing in these gradients.

Terzopoulos [9] solves the problem with a variational formulation and optimize the depth map subject to the minimization of deviation from depths, orientations and surface discontinuities. Therefore, there will be issues for parameter choices and convergence concerns. The optimization involves a lot possible variables: different energy term, such as l_0, l_1 norms, carefully designed smooth term or even explicit operations for occluded and discontinuous areas.

Frankot anc Chellappa [2] recovery the shape from gradients in a much simpler way while enforcing the integrability of the surface. To avoid the ambiguity of integration or presence of noise, they simply project the gradients of the surface to a set of integrable basis, here as Fourier basis functions. The natural idea to improve this method is to modify the basis functions which is crucial for the performance. In [3] wavelet are used to improve the reconstruction performance.

This paper follows the idea of projecting normals onto basis functions. They operate with slant and tilt instead of gradients (x, y, z) for each position. Therefore the approach has advantage of tolerating ambiguities in tilt of π. A redundant set of non-orthogonal basis functions of finite support are used for the reconstruction referred as shapelets.

B. The Shapelets for Surface Reconstruction

The surface reconstruction is the computation to recovery of the depth range for input images. Apparently the depth map can be decomposed into a set of basis functions. Since the differentiation is a linear operator, the correlation between a signal and the basis function can be easily deduced from the correlation between the gradients of the signal and the basis function. It means by correlating the normal map and the gradients of a set of carefully designed shapelets we can recovery the depth range up to an offset. The continuity of the surface, on the other hand, can be easily imposed by the choice of basis functions. The constrains of the shapelets can be concludes from the following aspects:

The shapelets must have minimal ambiguity of shape with respect to its gradient and its gradient must satisfy the admissibility condition of zero mean. In 1D scenario, this constraint can be easily achieved by imposing one symmetric peak for the function. Then its gradient function will have one positive and one negative peak respectively.

Another important constraint is the feature to preserve the phase information in signal which means the correlation between the gradients of the shapelets and the normal map will preserve the phase information either. For practical purpose, the shapelets should only have finite supports which imposes the function to be a Gaussian or a near Gaussian.

Finally, since the reconstruction performance largely depends on how well the shapelets represent the signal, the spectrum coverage of the bank of shapelets is the most important constraint. The spectrum of the surface $F(\omega)$ and its gradient $j\omega F(\omega)$ have a $\pi/2$ difference in phase and scaled amplitude by frequency. Therefore, an uniform coverage need the spectrum of the shapelet gradients to be inversely proportional to frequency.

C. Surface Reconstruction with Shapelets

For a 1D signal, the reconstruction from its gradients by shapelets can be represented as follows:

$$R = \sum_i \nabla_f \star \nabla_{si} \tag{12}$$

where ∇_f stands for the gradients of the 1D signal, ∇_{si} the shapelets at scale i and R the reconstruction of signal.

The reconstruction in 2D is not that straightforward. They specifically formulated the gradients with slant σ and tilt τ

other than gradients (x, y). This formulation can be useful for the ambiguities in tilt of π. Then the correlation between the gradients of the shapelets and surface is performed at each level i: $C_i = C_{\nabla i} C_{\tau i}$ where $C_{\nabla i}$ is the gradient correlation with slant and formed as:

$$C_{\nabla i} = |\nabla_f| \star |\nabla_{si}| \tag{13}$$

where $|\nabla|$ denotes the magnitude of gradient given by $\tan(\sigma)$ and . Without the tilt information $C_{\tau i}$, $C_{\nabla i}$ can only provide the correlation information with an ambiguity of π. It means the sign of the result need to be decided by tilt correlation. The same directions between shapelet and surface lead to positive values while opposite ones negative values. The orthogonal directions mean no correlations between the two inputs. Therefore, $C_{\tau i}$ is defined by the tilt difference as follows:

$$C_{\tau i} = \cos(\tau_f) \star \cos(\tau_{si}) + \sin(\tau_f) \star \sin(\tau_{si}) \tag{14}$$

In conclusion, the 2D surface reconstruction can with shapelets in form of slant and tilt can be represented as:

$$\begin{aligned} R &= \sum_i C_i \\ &= \sum_i C_{\nabla i} C_{\tau i} \\ &= \sum_i [|\nabla_f| \cos(\tau_f)] \star [|\nabla_{si}| \cos(\tau_{si})] + \\ &\qquad [|\nabla_f| \sin(\tau_f)] \star [|\nabla_{si}| \sin(\tau_{si})] \end{aligned} \tag{15}$$

D. Performance Evaluation

The author tests the algorithm on some basic shapes (both synthetic and real data) and achieve very impressive performance without further optimization. They also test shapelets under the presence of noise (standard deviations of 0.3). The shapelets show strong robustness to noise which is consistent with [2]. With ambiguity of π in tilt information, the reconstruction can still describe the general structure but it appears to be "inflated" since both positive and negative correlations are treated as positive ones. The author also generate some interesting "2.5D" reconstruction with only occluding contours. The contours are treated as normals orthogonal to the view direction and other areas are assumed to be flat.

The implementation of shapelets is very straightforward. We use a bank of Gaussian functions at different scale and realize the convolutions in frequency domain. As mentioned in the introduction section, we need the depth information to render for synthetic views. We use normal map from PTMs and reconstruct the image range with shapelets. It works well for some samples as shown in Fig.1. The *shell* and *paper* have smooth surface variations and their geometry details are repeated or sparse. On the other hand, the reconstruction for the *ancient piece* is not very good. Due to the low pass nature of Gaussian functions, those subtle details are missing in the depth map. Furthermore, plenty of high frequency geometry structures are also degraded.

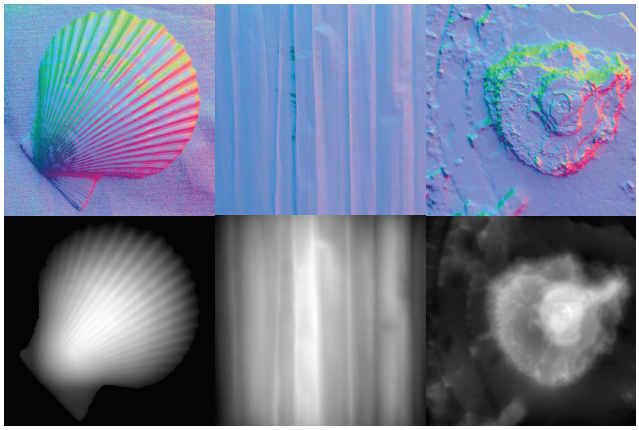

Fig. 1. Reconstruction Comparison: Top row is the normal maps and the bottom row the reconstructed depth range; From left to right is data sample *shell*,*paper* and *ancient piece*.

E. Implementation Details and Conclusions

In this section, we will discuss how we use the geometry information from surface gradients to create the depth sensitivity for synthetic view rendering. The straightforward way is to assign the depth range for each vertices. Then the projecting operations are performed for each vertex. However, heavy operations for each pixel not only largely degrade the realtime performance, bus also differ from PTMs rendering scheme based on texture operations.

To fully use the reconstruction results and limited resources, we address this problem by studying our specific rendering scenarios. Unlike the usual 3D rendering, our reconstruction objects, oil painting, are generally flat. The sense of depth variation and consistency between synthetic views are much more important than the exact accuracy of the 3D structures. Therefore, we calculate the offsets from depth map for the 3D projection. Since the surface is continuous, we can just use the offsets to achieve projecting effect and add convincing depth sensitivity.

Shapelets deliver plausible performance for most cases, but for object full of high frequency details, both reconstruction and rendering show distortion and inaccuracy. Compared with traditional shape-from-gradient problems, we have the extra texture information from PTMs. The reconstructed depth map organizes the texture for synthetic views. We argue that the texture being projected should also contribute to the shapelets and surface recovery. It actually makes sense since our main purpose is texture projection while preserving the details for all the views. We should find a way to put texture constraints to the surface reconstruction to ensure the all the important details for a new view.

IV. Representing and Recognizing the Visual Appearance of Materials using Three-dimensional Texton [5]

In this section, we give a review on [5] in which Leung and Malik propose a unified model for both reflectance and surface normal from texture. This is quite a challenge. The normal of surface creates several visual effects, such

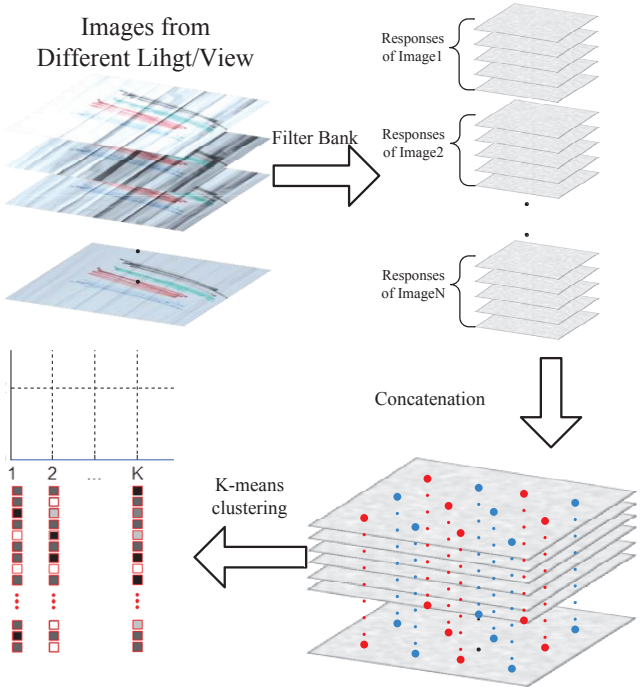

Fig. 2. Constructing vocabulary of 3D textons. The multiple images under varying view/light are filtered with predesigned filter banks. Then concatenation is performed for each pixel to form $N_{vl}N_{fil}$ long vectors. Finally, K-means is used to find most representative 3D textons. The histogram of these K textons is used for characterization.

as specularities, shadows and occlusions which can be easily confused. Meanwhile, the reflectance appearance of the same position sometimes drastically changes under different light/view which makes modeling much more difficult.

This is a very first paper to model the texture variation across view and light space. It provides good guidance on how to combine features under different conditions and recognize the consistency with quite different appearances. It inspires us on how to model and optimize PTMs and shapelets in the future work.

A. Building a 3D Texton Vocabulary

Texture with apparent geometry details has quite different visual appearance under different light/view conditions. Those variation can not be modeled simply by brightness normalization or intensity transforms. Those high frequency impulses, such as specularities and cast shadows, are even more difficult to deal with. These challenges in modeling the relation between the image intensity values under varying view/light settings and the properties of 3D texture lead to an interest in building explicit models for 3D textures. Previous works lack the expressiveness to solve the general problem of natural material representation, recognition, and synthesis under varying lighting and viewing conditions.

To tackle the problem, Leung and Malik focus on properties at the local scale where only a small number of perceptually distinguishable micro-structures exist on the surface. The textures have various appearance, but geometrically the local surface relief may fall into only several classes such as ridges, grooves, bumps, hollows, etc. Similarly, reflectance variations have even more repeatability which makes the classification easier. Therefore building a small, finite vocabulary of micro-structures referred as 3D textons should be an effective way characterize the texture locally based on the observations.

They use multiple filter response to characterize the texture, which has spatially repeating properties by definition. With the redundant representations for each pixel, they are able to identify several distinct filter response vectors that can describe most of the textures. They use 48 filters which include 36 elongated filters at 6 orientations, 3 scales, and 2 phases, 8 center surround difference of Gaussian filters, and 4 low-pass Gaussian filters. The filter responses for each pixel form a 48 dimensional vector and those vectors are clustered with K-means to get actual textons that describe the appearance effectively.

For natural 3D textures, the effects of shadows, specularity and multiple lighting under varying lighting and viewing conditions will lead to drastically change of visual appearance. When performing texture clustering in 2D, the shadows only stands for the low intensity of pixel values. However, they can have several quite different meanings under the 3D textons descriptions: The shadows could be caused by the lighting angle on a Lambertian surface which changes according to the cosine of incident light. These changes usually have smooth variations. It can also be cast shadow which results from changes of light direction and surface heights of neighboring areas. These changes are usually recognized as sharp impulses in the reflectance function. A more interesting case is the deep groove which will have the dark appearance for a wide range of lighting and viewing change.

Of cause, there are some assumptions to simplify the problem: Lambertian surface model and absence of occlusion, shadows, mutual illumination and specularity. However, these assumption will largely narrow the usage of the representations. To overcome these difficulties, a large number of image samples across lighting and viewing space are used to characterize the 3D texture. Let the number of images be N_{vl} with $N_{vl} \gg 1$. With those texture registered to the same view/light conditions, the equivalence under N_{vl} different lighting and viewing conditions should guarantee the similarity under all lighting and viewing conditions. Furthermore, the co-occurrence of filter responses across different lighting and viewing conditions specifies the local geometric and photometric properties of the surface. Based on these observations, they propose the constructing scheme for 3D textons vocabulary as shown in Fig.2. They use $N_{mat} = 20$ materials and $N_{vl} = 20$ different lighting and viewing conditions, which are registered by standard area-based sum-of-square-differences(SSD). The details are as follows:

1.Apply N_{fil} filter banks to all the materials under N_{vl} lighting and viewing conditions which form data vectors with $N_{vl}N_{fil}$ long for each pixel.

2.Apply K-means clustering to N_{mat} materials separately. The center number K for clustering is set to 400. The clustering optimize the error distance with SSD:

$$Err = \sum_{i=1}^{N}\sum_{k=1}^{K} q_{ik}\|x_i - c_k\|^2 \tag{16}$$

where

$$q_{ik} = 1if\|x_i - c_k\|^2 < \|x_i - c_j\|^2 \quad (17)$$

$$\forall j = 1, \ldots, K \; and \; j \neq k \quad (18)$$

$$q_{ik} = 1otherwise \quad (19)$$

where N represents the pixel number, x_i the concatenated filter response for pixel i and c_k the kth center of the appearance vector.

3.Merge the the centres for N_{mat} material and prune the numbers, from $400N_{mat}$ to 100, by merge the close points and discard those with too few neighbours.

4.Apply the K-means again with new centers to achieve minimum distance for all images.

There are two important properties of their texture vocabulary. Firstly, expressiveness: the texture from different materials should be discriminated by the vocabulary. Secondly, generalization: the novel material which are not from the training data should also be expressive by the vocabulary.

B. Texture Recognition with 3D Textons

The purpose of building such a vocabulary of 3D textons is mainly for recognition and classification of different textures and materials. Given the materials under all viewing and lighting conditions, we could get the filter responses for each position on the surface and identify them with the 100 texton labels from the vocabulary. Then each material has a symbol map which consist 100 different values. Since the spatial arrangement for the symbol map is not essential for the recognition and classification, they focus on the statistic feature of each material. The histogram of the labels is used to distinguish different materials and the chi-square significance test to measure the distance between two histograms as follows:

$$\chi^2(h_1, h_2) = \frac{1}{2}\sum_{n=1}^{bins}\frac{(h_1(n) - h_2(n))^2}{h_1(n) + h_2(n)} \quad (20)$$

They used the chi-square probability function $P(\chi^2|\upsilon)$ to measure the significance of chi-square distance. Therefore,

$$P(\chi^2|\upsilon) = Q(\upsilon/2, \chi^2/2) \; and \quad (21)$$

$$Q(a, x) = \frac{1}{\Gamma(a)}\int_0^x e^{-t}t^{a-1}\,dt \quad (22)$$

The strategy to represent material under varying lighting and viewing conditions with histogram can be considered as a compressing operation for the data. It largely reduces the dimensions of the data and removes the spatially influence to focus on the statical features of materials.

To verify the effectiveness of texton representation, the 3D textons are compared with principle component analysis (PCA) which is widely used in this area. The PCA representation generates a linear subspace with K sample images R^N where $K \ll N$. The classification and recognition are performed in R^K space which are largely simplified. But PCA is intrinsically linear which cannot fit some nonlinear effects of the surface, such as shadows, occlusion, specularities, mutual illumination and so on. Moreover, PCA cannot be applied to non-Lambertian materials. Therefore, the 3D texton is a better representation since it doesn't make Lambertian assumption or rule out any special surface effects.

The true challenge for texture recognition is the case only one view/lighting texture is available with a vocabulary built from multi-light/view conditions. This problem is rarely studied until this paper. Under the given lighting and viewing conditions, the filter responses of the input image correspond to the known positions of the appearance vector. However, as mentioned in the previous parts, different geometry feature or materials may have the same reflectance under different view/light conditions. Therefore it is impossible to label each pixel accurately let alone acquire statistical feature from histogram of the big picture.

They solve this problem with a Markov chain Monte Carlo (MCMC) algorithm. Firstly, they label pixel i randomly from N_i possible alternatives and denote the initial state as $x^{(t)}$ with $t = 0$. Then with texton labeling of the input image, they can decide the material type and further measure the distance $P(x^{(t)})$ between incoming image and the training materials as shown in equation (21). To update the state, they randomly change the label of M pixels and compute the new probability $P(x^{'})$. The parameter $\alpha = \frac{P(x^{'})}{P(x^{(t)})}$ is the accepting probability for the new state($\alpha = 1$ when $\alpha > 1$). Finally, repeat all these steps until the state converges to a stable output.

The essence of the algorithm is estimating the distribution $P(Label|Material)$ by randomly changing pixel labels to draw samples. The performance is very impressive with 87% detection rate for 40 material under 4 different view/light. The possible label number is 5 and the random change of labels is 5%.

C. Novel View/Light Prediction with 3D Textons

The prediction problem is very similar to the texture recognition with only partial information. Given image set of a novel texture taken from n different known light/view conditions, a nN_{fil} data vector can be generated from the filter response. As the photos are taken under known conditions, we can compare the data vector to the corresponding parts of the 3D textons. Even for incomplete and ambiguous scenarios, they are able to use MCMC to determine the material and labels for each pixel.

As the K labels are $N_{vl}N_{fill}$ long data vectors, the other $N_{vl} - n$ conditions can be synthesized with these K vectors. Therefore the texture appearance can be changed according to the novel lighting or viewing directions. The visual performance is plausible. Especially compared with traditional texture mapping, the synthesis has more sense of depth and sharp lighting response, such as surfacełhighlights, shadows, and occlusions.

D. Conclusions

The 3D textons are very effective representations for textures with both reflectance and surface relief variations. It demonstrates impressive performance for recognizing materials with only one image under known light/view condition.

Furthermore, it can also be used to predict and synthesize the appearance under training lighting/viewing conditions. The paper shows an empirical way to deal with texture information across different spaces. The similar idea can be very helpful for optimization of shapelets with the constraints of texture information.

However, both recognition for one view image and light/view prediction are limited to the training conditions. To expand the applicable range, the image acquisition will be exhausting and the 3D textons end up being a dimension disaster. The expressiveness and compactness are both potential directions for further improvement.

With the concept of 3D textons, we revisit our work on PTMs and shapelets. They both share the same challenge to model input from different conditions. The PTMs use a parametric model but the coefficients themselves are standing for different reflectance features. It would be more valuable if we can construct the 3D texton for full lighting space from PTMs. The shapelets are facing only two different conditions: 3D gradients and textures. The dimension is low but the difference is significant. The simple concatenation may not be the answer. Moreover, we also need to take the texture preservation into consideration while projecting them for synthetic views.

V. Research Proposal

The work of previous semester provided us with good insights and exciting challenges for the acquisition and rendering of two-and-a-half dimensional objects. Meanwhile, reviewing [6] [4] [5] improves us with better understanding of the problems and shows potential directions to continue our work. We will first summarize ongoing work with short term plan and then discuss the possible directions of long term research.

A. Current Work

We have set up our own PTM acquisition system with tripod, wireless flash and calibration sphere. Although there are some geometry distortions for cast shadows, the approximation overall is still plausible. The sphere we use for calibration can provide both light direction and camera pose. We are working on the acquisition for stereo PTMs which uses two fixed DSLR on tripod and for multiple view PTMs which uses a main DSLR and auxiliary mobile cameras registered to front PTMs.

Our rendering algorithm on Android tablet has a frame rate around 25 for images with resolution 512×512 and 15 for images 2000×2000. The rendering results appear to be naturally responding to the viewer and environmental lighting conditions. The texture wrapping with normal map performs well for regular view angles and delivers appealing visual experience of depth sensitivity.

We are working with texton training scheme for both shapelets and PTMs. Combining the texture and gradients response for shapelets optimization is not trivial. Using the texture preservation while projecting is still a big challenge. On the other hand, constructing a 3D texton vocabulary from the PTMs appears to be a better idea compared with fixed view/light conditions. PTMs can expand the whole lighting space and preserve the geometry information with normal maps. The redundancy in the coefficients has never been fully discovered. We are constructing 3D textons for oil paintings and characterizing these paintings with representative textons $\{\{a_i\}_{i=0}^{5}, RGB\}$ from PTMs. We may be able to rendering full lighting space with limited acquisition data by using textons of PTMs.

B. Future Work

There are two main directions for the future work. Firstly, representing the reflectance function with better parametric model other than biquadratic polynomial will be fundamental improvements. Apparently, the reflectance function mainly contains three components: diffuse components, specularity and cast shadows. A spline models fits the description perfectly. The only issue is how to sample the reflectance. [10] shows us a potential direction for the spline modeling. For 1D signal, only a few samples are needed to form the perfect reconstruction. Sampling the reflectance function in 2D lighting space will be both challenging and useful.

Secondly, designing filter banks for shape-from-gradient is not a new topic. However, for our specific rendering scenarios, we need to avoid distorting the texture information while synthesize new view with depth maps. Therefore, it is possible to combine the reconstruction and rendering since they can all be linear operations. Afterwards we can train the filter banks with the consideration from both gradients and textures. A redundant basis is preferable since we can choose the rendering filter banks for specific view angles. This will not only contribute to the theoretical side, but very practical for rendering since a lot of textures are represented by normal maps. The sense of depth for texture will definitely improve the visual experience.

References

[1] Kristin J. Dana, Bram van Ginneken, Shree K. Nayar, and Jan J. Koenderink. Reflectance and texture of real-world surfaces. *ACM Trans. Graph.*, 18(1):1–34, January 1999.

[2] Robert T. Frankot and Rama Chellappa. A method for enforcing integrability in shape from shading algorithms. *IEEE Trans. Pattern Anal. Mach. Intell.*, 10(4):439–451, July 1988.

[3] Jun-Wei Hsieh, Hong-Yuan Mark Liao, Ming-Tat Ko, and Kuo-Chin Fan. Wavelet-based shape from shading. *CVGIP: Graphical Model and Image Processing*, 57(4):343–362, 1995.

[4] Peter Kovesi. Shapelets correlated with surface normals produce surfaces. In *In ICCV05*, pages 994–1001, 2005.

[5] Thomas Leung and Jitendra Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *Int. J. Comput. Vision*, 43(1):29–44, June 2001.

[6] Tom Malzbender, Tom Malzbender, Dan Gelb, Dan Gelb, Hans Wolters, and Hans Wolters. Polynomial texture maps. In *In Computer Graphics, SIGGRAPH 2001 Proceedings*, pages 519–528, 2001.

[7] Shree K. Nayar, Peter N. Belhumeur, and Terry E. Boult. Lighting sensitive display. *ACM Trans. Graph.*, 23(4):963–979, October 2004.

[8] F. E. Nicodemus, J. C. Richmond, J. J. Hsia, I. W. Ginsberg, and T. Limperis. Radiometry. chapter Geometrical considerations and nomenclature for reflectance, pages 94–145. Jones and Bartlett Publishers, Inc., USA, 1992.

[9] Demetri Terzopoulos. The computation of visible-surface representations. *IEEE Trans. Pattern Anal. Mach. Intell.*, 10(4):417–438, July 1988.

[10] Martin Vetterli, Pina Marziliano, and Thierry Blu. Sampling signals with finite rate of innovation. *IEEE Transactions on Signal Processing*, 50(6):1417–1428, 2002. IEEE Signal Processing Society's 2006 Best Paper Award.