# SIGGRAPH2004

### **Precomputed Radiance Transfer**

Jan Kautz, Massachusetts Institute of Technology



Early interactive rendering techniques were limited to simple point lighting environments, and had almost no transport complexity – global effects like shadows were ignored.

Algorithms to generate interactive shadows exist – most notably crows shadow volume technique and the shadow zbuffer technique by Williams – there is a nice extension to this work at this years siggraph. However these techniques are restricted to point lights. Such algorithms can be combined with multipass rendering such as with an accumulation buffer to generate soft shadows, but the number of passes gets impractically large as the light sources get bigger.

Heidrich et al. modeled soft shadows and inter reflections with bump maps – but the technique is limited to point lights and performance isn't quite real-time. Polynomial texture maps can handle shadows and inter-reflections in real-time, but they too are limited to point lights.

Early work also extended lighting from point lights to general environment maps.

You have just seen two papers that generalize the early ideas to fairly general isotropic BRDFs. But environment map methods ignore transport complexity – they can't model shadows or inter-reflections.

As we have seen, shadow volumes can be extended to deal with single spherical light sources which may achieve real time rates on the next generation of hardware.

Ashikhmin et al. have a TOG paper that just came out that allows you to steer a small light source over a diffuse object.

Precomputed Radiance Transfer handles arbitrary, low-frequency lighting environments and provides transport complexity – including shadows, inter reflections and caustics.

Note that it is precisely soft shadows from low-frequency lighting that current interactive techniques have the greatest difficulty with.

Some earlier work had interesting transport complexity with frozen lighting environments. In particular the Irradiance volume paper allowed you to move a diffuse object through a precomputed lighting environment, dynamically lighting the object. However the object didn't shadow itself or the environment. Surface lightfields can model complex transport complexity and deal with interesting lighting environments, but the lighting environments are frozen. A recent example of that is Chens paper this year.

Matusik et al. have a nice paper here that captures both a surface light field and a reflectance field – allowing you to relight objects in novel lighting environments. The reconstruction is not interactive however.

#### Motivation – What we want

- What we want:
  - Illuminate objects with environment maps
  - Change lighting on-the-fly
  - Include self-shadowing and interreflections
  - In real-time







The basis functions we used are the spherical harmonics – they are equivalent to the fourier basis on the plane, but mapped to the sphere.

They have several nice properties:

Since the basis is orthonormal projection is simple, evaluation is also very simple.



The most important property for our application is that they are rotationally invariant. This means that given some lighting environment on the sphere, you can project that lighting environment into SH, rotate the basis functions and integrate them against themselves (ie: rotating the projection and re-projecting) you get identical results to rotating the original lighting environment and projecting it into the SH basis.

Integration of a product of two functions, which are represented in SH becomes a simple dot-product! Something we will exploit later on.

Rotation is simple (efficient evaluation formulae, just a linear operator on the SH coefs). We won't explain it in this talk.



Examples of a few basis functions. They are defined over the sphere. We show them here as sphere maps.

Note, that the higher basis function have higher frequencies too (like Fourier).



Here, we first convert a spherical function (environment map) into SH and then reconstruct the function (formula from before).

Again, the more basis functions are used, the higher frequencies can be represented...

For all examples in this talk, we've used 25 coefficients (assuming low-frequency lighting).



Alternatively, the Haar wavelet basis functions can be used (instead of SH). We will talk briefly about this at the end of the talk.

It should be noted, that wavelets are good at representing all-frequency detail (e.g. with 100 coeffs the bright windows are represented well, the not so important darker areas (floor) is represented with less accuracy).



To compute exit radiance from a point p, we need to integrate all incident lighting against the visibility function and the diffuse BRDF (dot-product between the normal and the light direction).



Same thing written down more accurately.



Visually, we integrate the product of three functions (light, visibility, and cosine).



The main trick we are going to use for precomputed radiance transfer (*PRT*) is to combine the visibility and the cosine into one function (*cosine-weighted visibility* or *transfer function*), which we integrate against the lighting.



This is not useful per se. We still need to encode the two spherical functions (lighting, cosine-weighted visibility/transfer function). Furthermore, we need to perform the integration of the product of the two functions quickly.



Using some more math again, we get the transfer function T(s).

Note, that this function is defined over the full sphere. It also implicitly encodes the normal at the point p! So, for rendering no explicit normal will be needed.



Now, when we plug the new T(s) into the rendering equation, we see that we have an integral of a product of two functions. We remember, that this special case boils down to a dot-product of coefficient vectors, when the two functions are represented in SH.

This is exactly, what we will do. We project the incident lighting and the transfer function into SH.



Then the expensive integral becomes a simple product between two coefficient vectors.



This has a number of implications:

Shadow computation/shading is independent of the number or the size of the light sources! All the lighting is encoded in the lighting vector, which is independent of that.

Rendering this kind of shadows is extremely cheap. It is in fact cheaper than rendering hard shadows!

The transfer vectors can be computed off-line, thus incurring no performance penalty at runtime.

The lighting vector for the incident light can be computed at run-time (fast enough, takes a few milliseconds).



The precomputation of transfer coefficients means that the models have to be static!

Also, there is an implicit assumption, that all points on the surface receive the same incident illumination (environment map assumption). This implies that no half-shadow can be cast over the object (unless, it's part of the object preprocess).



This shows the rendering process.

We project the lighting into SH (integral against basis functions). If the object is rotated wrt. to the lighting, we need to apply the inverse rotation to the lighting vector (using the SH rotation matrix).

At run-time, we need to lookup the transfer vector at every pixel (or vertex, depending on implementation). A (vertex/pixel)-shader then computes the dot-product between the coefficient vectors. The result of this computation is the exitant radiance at that point.

## **PRT Results**





Unshadowed



Shadowed









The same technique can be applied to diffuse volumes. Here we have a transfer vector at each voxel, instead of each pixel/vertex.



Since, the normal is implicitly encoded in the transfer function, we can easily do bump mapping as well...



Some detail has been lacking so far, which we will explain in the following.



As we've seen before, both the lighting and the transfer function need to be projected into SH. Here we will talk about the projection of the transfer function.

As introduced in the Background-Section, projecting a function into SH boils down to integrating that function against the SH basis functions. This results in a vector of coefficients.

As a reminder: the transfer function is the visibility times the dot-product between the normal *n* and the sample direction *s* (multiplied by the albedo, which says how reflective the surface is). $\langle$ 



The main question is how to evaluate the integral. We will evaluate it numerically using Monte-Carlo integration. This basically means, that we generate a random (and uniform) set of directions  $s_j$ , which we use to sample the integrand. All the contributions are then summed up and weighted by 4\*pi/(#samples).

The visibility  $V_p()$  needs to be computed at every point. The easiest way to do this, is to use ray-tracing.

-----

Aisde: uniform random directions can be generated the following way.

- 1) Generate random points in the 2D unit square (x,y)
- 2) These are mapped onto the sphere with:

theta =  $2 \arccos(\operatorname{sqrt}(1-x))$ 

phi = 2y\*pi



Visual explanation 1).



Visual explanation 2):

This slide illustrates the precomputation for direct lighting. Each image on the right is generated by placing the head model into a lighting environment that simply consists of the corresponding basis function (SH basis in this case illustrated on the left.) This just requires rendering software that can deal with negative lights.

The result is a spatially varying set of transfer coefficients shown on the right.

To reconstruct reflected radiance just compute a linear combination of the transfer coefficient images scaled by the corresponding coefficient for the lighting environment.



Pseudo-code for the precomputation.

The function selfShadow( p, sample[j].dir ) traces a ray from position p in direction sample[j].dir. It returns true if there it hits the object, and false otherwise.



Bump mapping is really easy to incorporate (if transfer vectors are stored in a texture).

The precomputation algorithm from the previous slide remains the same. Only the normal needs to be looked up from a bump map in the precomputation phase!

If the bump map is too incorporate shadows from the bumps, then the selfShadow() function needs to be augmented to test for intersections with the bump map (e.g. convert to a heightfield for intersection tests).



Not only shadows can be included into PRT, but also interreflections.

Light arriving at a point q can be subsequently scattered onto a point p. I.e. light arriving from  $s_q$  can arrive at p, although there is may be no direct path (along  $s_q$ ) to p (as in this example).

Note, that light is arriving from infinity, so both shown direction  $s_q$  originate from the same point in infinity.



More formally, we do not only have direct illumination  $L^DS$  (Direct Shadowed), but also light arriving from directions *s*, where there is self-shadowing (i.e.  $l-V_p(s)$ ). The light arrives from positions *q*, which are the first hit along *s*.



To account for interreflections, the precomputation has to be changed again.

Each additional bounce *b* generates a vector  $T^b_{\{p,i\}}$ , which is computed as shown on the slide. Each of these additional transfer vectors is for a certain bounce.

To get the final transfer vector, they have to be added. Again, the run-time remains the same!



This set of images shows the buddha model lit in the same lighting environment, without shadows, with shadows and with shadows and inter reflections.

#### Rendering

• Reminder:

$$L_p^{out}(\vec{v}) = \sum_i^n L_i^{in} T_i$$

SIGGRAPH2004

• Need lighting coefficient vector:

$$L_i = \int L^{in}(\vec{s}) y_i(\vec{s}) d\vec{s}$$

- Compute every frame (if lighting changes)
- Projection can e.g. be done using Monte-Carlo integration (see before)

Rendering is just the dot-product between the coefficient vectors of the light and the transfer.

The lighting coefficient vector is computed as the integral of the lighting against the basis functions (see slides about transfer coefficient computation).



Sofar, the transfer coefficient could be single-channel only (given that the 3-channel albedo is multiplied onto the result later on). If there are interreflections, color bleeding will happen and the albedo cannot be factored outside the precomputation. This makes 3-channel transfer vectors necessary, see next slide.

### Rendering



• In case of interreflections (and color bleeding):





In the following we talk about 4 extensions.

- 1) How to reduce the storage cost
- 2) How to incorporate animated objects
- 3) Other materials (other than diffuse)
- 4) Use of other basis functions



The original storage cost is fairly high (25 coefficients per texel).



Standard texture compression can work, but there are better techniques.



PCA (Principle Component Analysis):

Plug all the vectors  $T_p$  (for all p) into PCA, and you will get above result/approximation.

The quality of the approximation depends on the number of basis vectors. If the variation in the  $T_p$  is high, a high number of basis vectors is needed.

The beauty of this is, that now we only need to store weights per pixel (hopefully far less than there are original coefficients). Unfortunately, this might not be the case, rendering pure PCA not very usuful (or only limited to certain objects, where  $T_p$  is well-approximated with less than 25 basis vectors).



#### VQ:

We plug all  $T_p$  into VQ and get a set of representative  $T^k$  back. Each  $T_p$  is uniquely associated with a  $T^k$ , but different  $T_p$  may map to the same  $T^k$  (quantization).

Now we only need to store the index at teach pixel, instead of the actual vector.

Storage is hereby reduced, as the codebook (containing the  $T^k$ ) is usually much smaller than the number of original vectors  $T^p$ .

Unfortunately, quantization artifacts may be visible and mip-mapping is not possible anymore.



The best compression technique combines both algorithms.

First run VQ. All the  $T_p$  that map to the same  $T^k$  are considered a cluster. Within that cluster PCA is used, which now works much better, because all  $T_p$  of a cluster are similar (due to VQ).

Rendering is slightly different now. The dot-products between  $T_p$  and lighting vector L is now:  $T_p dot L = SUM(w_k T_k) dot L = SUM(w_k (T^k dot L)) = SUM(w_k R^k)$  $R^k$  can be computed on the CPU, since the  $T^k$  don't change per-pixel, but only the weights. Only the SUM( $w_k ^k R^k$ ) is then done on the GPU.

Attention has to be paid to the different areas of an object, that belong to different clusters (they have different basis vectors).



Left: VQ producing 64 clusters. 2 PCA Vectors (mean + first) - Quantization artifacts are still visible (varies only linearly within a cluster)

Middle: VQ producing 64 clusters. 3 PCA Vectors (mean + first, second) - Artifacts almost gone (variation is now bilineaer)

Right: VQ producing 64 clusters. 4 PCA Vectors (mean + first, second, third) - Visually as good as original (variation is trilinear within cluster)



Assumption so far was: static models (because of precomputation).



Key-framed objects can be easily used with PRT. Just precompute transfer coefficient for every key-frame and interpolate coefficients within key-frames (as the rest is interpolated as well).

Storage cost can be decrease by using the same compression technique as before.

The precomputation becomes a lot more expensive... (Nothing really to prevent this).



Fully dynamic models:

No solution in sight. Visibility changes radically (and in unknown ways). Computation of transfer vectors would need to be done on-the-fly.



Example of a key-framed animation.



Other effects that can be incorporated with an enhance pre-process: caustics and subsurface-scattering.

View-dependent effects require transfer matrices and not vectors (see original PRT paper).

## PRT with glossy BRDFs





Measured Vinyl



Quality of SH solution.

0 degree (point light) source, 20 degree light source, 40 degree light-source.

Light is blocked by a blocker casting a shadow onto the receiver plane. Different order of SH is shown (order  $^2$  = number of basis functions). Very right: exact solution.

As stated before, lighting is assumed low-frequency, i.e. point light doesn't work well, but large area lights do!



As we've noticed SH make shadows smooth (few coefficients are used, hence the low-frequency lighting assumption). Even if more are used (say 100), still low-frequency.

Alternative: Use Haar Wavelets. Quality improves quite a bit. Basic algorithm remains the same.



As shown in the comparison on the right, with more coefficients, wavelets do much better represent the lighting than the SH (which show a lot of ringing artifacts).

There are a few differences when using Haar instead of SH:

- 1) All transfer coefficients need to be computed!
- 2) Because the actual N coefficients used, is decided at run-time based on the lighting's most important N coefficients (N=100 seems sufficient).
- 3) This requires all transfer coefficients to be stored as well (can be compressed well, like lossy wavelet compressed images).
- 4) Since the coefficients to be used change at run-time, this is not well-suited to a GPU implementation (but works fine on CPU)

#### Conclusions



#### Pros:

- Fast, arbitrary dynamic lighting
- PRT: includes shadows and interreflections

Cons:

- Works only well for low-frequency lighting
- Animated models are difficult to handle

I

