Shadow Mapping Techniques

Pramook Khungurn

November 25, 2016

This document is written as I read a series of papers on recent shadow mapping techniques.

1 A Brief History of Shadow Mapping

- Shadow mapping was first introduced in 1978 [Williams, 1978]. Of course, it comes with aliasing problems, which needs to be solved in screen space.
- Reeves et al. introduced *percentage closer filtering* (PCF) in 1987 [Reeves et al., 1987]. The algorithm filters shadow testing results and so first achieves screen-space anti-aliasing. However, it does not supports *prefiltering* of shadow maps so is limitted to small kernels.
- Fernando introduced *percentage closer soft shadows* (PCSS) in 2005 [Fernando, 2005]. The algorithm renders plausible soft shadows caused by planar area lights by first approximating the average blocker depth and then use the blocker depth to determine the size of PCF kernel.
- By the way, these algorithms have been included as assignments in the 2016 version of CS 5625.

2 Variance Shadow Maps

- This is from the 2006 paper by Donnelly and Lauritzen [Donnelly and Lauritzen, 2006]. The paper attempts to make shadow map filtering faster.
- Main idea:
  - Instead of storing a single depth value in a shadow map pixel, the paper stores a representation of the distribution of depths at that pixel.
  - The paper represents a distribution by its first and second moments, i.e. the mean depth and the mean squared depth.
  - Two depth distributions can be average by averaging their moments. As such, prefiltering techniques such as mipmapping or anisotropic filtering can be applied.
  - From the first two moments, the bound on the fraction of depths that are more distant than the surface being shaded can be estimated. The paper uses this bound as the fraction of light that reaches the surface.
- Shadow map generation:
  - The shadow map has two channels. One channel stores the depth. The other stores the square of the depth.
  - After the shadow map is rendered, we can do filtering operations such as generating mipmaps or sum area table.

  - From the filtered depths and the filtered squared depths, we can compute the mean (μ) and the variance (σ^2) of each region on the shadow map.

- The paper makes use of the following inequality.

  Theorem 1 (Cantelli's inequality). *Let X be a real random variable whose distribution has mean μ and variance σ^2. We have that:*

  $$\Pr(X - \mu \geq \lambda) \leq \frac{\sigma^2}{\sigma^2 + \lambda^2}.$$

  for $\lambda > 0$.

  Consequently, let $t = \lambda + \mu$, we have that:

  $$\Pr(X \geq t) \leq \frac{\sigma^2}{\sigma^2 + (t - \sigma)^2}. \tag{1}$$

  Note though, that the inequality only works with $t > \mu$.

- When doing shadow map lookup, we are given a depth value t. The paper approximates the fraction of light reaching the shaded point with the RHS of (1). (For $t < \mu$, I think the paper simply returns 0.)

- The paper shows that (1) gives exact solution in the case where there is a single planar occluder casting shadow onto another planar occluder.

- Light bleeding artifacts can arise when σ^2 is large. This is a problem for scenes with high depth compleixity when viewed from the light.

3 Convolution Shadow Maps

- This is from the 2007 paper by Annen et al. [Annen et al., 2007].

- Let $\mathbf{x} \in \mathbb{R}^3$ be the world-space position of a pixel. Let $\mathbf{p} \in \mathbb{R}^2$ denote a position of a shadow map pixel. It is obtained by a surfective mapping $T : \mathbb{R}^3 \rightarrow \mathbb{R}^2$ between world-space and shadow map space: $T(\mathbf{x}) = \mathbf{p}$.

- The shadow map econdes a function $z(\mathbf{p})$ that represents the depth of the blocker that is closest to the light source for each $\mathbf{p}$.

- A pixel with world-space position $\mathbf{x}$ is considered in shadow when $d(\mathbf{x}) > z(\mathbf{p})$ where $d(\mathbf{x})$ is the depth of $\mathbf{x}$ with respect to the light source.

- The shadow function s is given by:

  $$s(\mathbf{x}) := f(d(\mathbf{x}), z(\mathbf{p})) = f(d(\mathbf{x}), z(T(\mathbf{x})))$$

  where $f(d, z) = 1$ if $d \leq z$ and $f(d, z) = 0$ otherwise.

- A convolution on a function g with kernel w, supported over a neighborhood $\mathcal{N}$, is defined as:

  $$[w * g](\mathbf{p}) := \sum_{\mathbf{q} \in \mathcal{N}} w(\mathbf{q}) g(\mathbf{p} - \mathbf{q}).$$

- The convolution of the shadow function s with w is more complicated. It is formulated as follows:

  $$s_{w*f}(\mathbf{x}) = \sum_{\mathbf{q} \in \mathcal{N}} w(\mathbf{q}) f(d(\mathbf{y}), z(\mathbf{p} - \mathbf{q})).$$

  Notice that the actual convolution happens in the shadow map space because it doesn't make sense to talk about 2D convolution in 3D. The definition also contain a new variable $\mathbf{y}$, informally defined as the point that lies near $\mathbf{x}$ such that $T(\mathbf{y}) = T^{-1}(\mathbf{p} - \mathbf{q})$. However, there is no unique $\mathbf{y}$ because T is not invertible. Hence, the above definition does not quite work.

  In order to get a sound mathematical definition, we make the assumption that $d(\mathbf{y}) = d(\mathbf{x})$. This says that $d(\mathbf{x})$ is the representative depth for the neighborhood $\mathcal{N}$. This is only correct for a planar receiver parallel to the shadow map plane, but it is quite reasonable and has been used in PCF [Reeves et al., 1987]. So, now the definition of shadow map convolution becomes:

  $$s_{w*f}(\mathbf{x}) = \sum_{\mathbf{q} \in \mathcal{N}} w(\mathbf{q}) f(d(\mathbf{x}), z(\mathbf{p} - \mathbf{q})) = [w * f(d(\mathbf{x}), z)](\mathbf{p}).$$

- Note that filtering cannot be applied directly to the shadow map values:

  $$[w * f(d(\mathbf{x}), z)](\mathbf{p}) \neq f(d(\mathbf{x}), [w * z](\mathbf{p})).$$

  The goal of the paper is to circumvent this limitation.

- The paper achieves the goal by expanding $f(d, z)$ as follows:

  $$f(d, z) = \sum_{i=1}^{\infty} a_i(d) B_i(z).$$

  The expansion has to be truncated to some truncation order N. So, the shadow map function can be written as:

  $$s(\mathbf{x}) = \sum_{i=1}^{N} a_i(d(\mathbf{x})) B_i(z(\mathbf{p})).$$

  We will choose the functions $B_1, B_2, \ldots, B_N$ and $a_1, a_2, \ldots, a_N$ later.

- The expansion is useful because we can now apply filtering on the basis function values of the shadow map pixels:

  $$s_{w*f}(\mathbf{x}) = [w * f(d(\mathbf{x}), z)](\mathbf{p}) = \left[w * \sum_{i=1}^{N} a_i(d(\mathbf{x})) B_i \right](\mathbf{p}) = \sum_{i=1}^{N} a_i(d(\mathbf{x})) [w * B_i](\mathbf{p}).$$

  In other words, *any convolution on the shadow function is equivalent to convolving the individual basis images $B_i(z(\mathbf{p}))$.*

- It is time to choose the basis functions. The paper uses *Fourier expansion.*

- Any periodic function $g(t)$ can be represented as an infinite sum of sinusoids:

  $$g(t) = \frac{1}{2} a_0 + \sum_{n=1}^{\infty} \left[a_n \cos\left(\frac{2\pi n}{T} t \right) + b_n \sin\left(\frac{2\pi n}{T} t \right) \right]$$

  where

  $$a_0 = \int_0^T g(t)\ \mathrm{d}t, \qquad a_n = \int_0^T \cos\left(\frac{2\pi n}{T} t \right) g(t)\ \mathrm{d}t, \qquad b_n = \int_0^T \sin\left(\frac{2\pi n}{T} t \right) g(t)\ \mathrm{d}t.$$

- The shadow test function f is a 2D function, but it can be represented as the Heaviside step function as follows:

  $$f(d,z) = H(z-d)$$

  where

  $$H(t) = \begin{cases} 1, & t > 0 \\ 1/2, & t = 0 \\ 0, & t < 0 \end{cases}.$$

- Let $S(t)$ be the square wave function with period 2 and amplite 1. We have that, when $t \in (-1,1)$, we have that $H(t) = 1/2 + S(t)/2$. The Fourier series for $S(t)$ is given by:

  $$S(t) = \frac{4}{\pi} \sum_{n=1}^{\infty} \frac{\sin((2k-1)\pi t)}{2k-1}.$$

  As a result,

  $$f(d,z) = H(z-d) = \frac{1}{2} + 2 \sum_{k=1}^{\infty} \frac{\sin(c_k(z-d))}{c_k}.$$

  where $c_k = (2k-1)\pi$. Now

  $$\sin(c_k(z-d)) = \sin(c_k z - c_k d) = \sin(c_k z)\cos(c_k d) - \cos(c_k z)\sin(c_k d).$$

  We have that:

  $$f(d,z) = \frac{1}{2} + 2 \sum_{k=1}^{\infty} \frac{\sin(c_k z)\cos(c_k d)}{c_k} - 2 \sum_{k=1}^{\infty} \frac{\cos(c_k z)\sin(c_k d)}{c_k}$$

  As a result, we may pick the basis functions a_i, B_i as follows:

  $$\begin{aligned} a_{2k-1}(d) &= \frac{2\cos(c_k d)}{c_k}, & B_{2k-1}(z) &= \sin(c_k z), \\ a_{2k}(d) &= -\frac{2\sin(c_k d)}{c_k}, & B_{2k}(z) &= \cos(c_k z) \end{aligned}$$

  where $k = 1, 2, \ldots, M = N/2$. DON'T FORGET TO ADD 1/2.

- Pros of Fourier expansion:
  - It is shift-invariant with respect to d and z. This allows the paper to move the Heaviside function without affecting the truncation error. (More on this later.)
  - The basis functions are bounded to $[-1,1]$, which allows the paper to used fixed point representation.

- Cons of Fourier expansion:
  - Rigging.
  - The Heaviside step function is smoothed, which can cause incorrect shadowing if not handled.

- The paper reduces ringing by attentuating the $(2k-1)$th and $(2k)$th terms by $\exp(-\alpha(k/M)^2)$ where the parameter α controls the attenuation strength. This introduces the trade-off where, the more ringing is reduced, the less steep the step function becomes.

- Approximating the shadow test with a low order Fourier expansion results in a smooth transition between shadowed and unshadowed region. This means that, when $z - d \approx 0$, the shadow function evaluates fo 0.5 which is not desirable. The paper solves this problem by:
  1. Subtracting an offset from d to shift the shadow boundary. This does not affect the approximation error incurred by the Fourier expansion because it is shift-invariant.
  2. Scaling the output of the shadow up to make the function steeper and the transition area narrower. This might introduce aliasing though.

4 Soft Shadows with Convolution Shadow Maps

- This is from the paper [Annen et al., 2008]. It extends convolution shadow maps (CSM) by proposing how to efficiently compute the average blocker depth, which is then used in the same manner as PCSS.
- The average blocker depth $z_{\mathrm{avg}}(\mathbf{x})$ of a world-space point $\mathbf{x}$ is the average of depths values of points *above* $\mathbf{x}$ around $T(\mathbf{x})$. To formulate this function mathematically, let us introduce the "complementary" shadow test $\bar{f}$:

  $$\bar{f}(d(\mathbf{x}), z(\mathbf{p})) = \begin{cases} 1, & d(\mathbf{x}) > z(\mathbf{p}) \\ 0, & d(\mathbf{x}) \leq z(\mathbf{p}) \end{cases}.$$

  We can use the shadow test to define the average blocker depth as follows:

  $$z_{\mathrm{avg}}(\mathbf{x}) = \frac{[w_{\mathrm{avg}} * (\bar{f}(d(\mathbf{x}), z) \times z)](\mathbf{p})}{[w_{\mathrm{avg}} * \bar{f}(d(\mathbf{x}), z)](\mathbf{p})}$$

  where w_{avg} is an averaging kernel. The denominator is just $1 - s_{w_{\mathrm{avg}} * f}(\mathbf{x})$. For the nominator, we can approximate it with the same trick in CSM. That is, we expand $\bar{f}$ as a sum of products of functions of d and z so that we have:

  $$\bar{f}(d(\mathbf{x}), z) = \sum_{i=1}^{N} \bar{a}_i(d(\mathbf{x}))\bar{B}_i(z(\mathbf{p}))z(\mathbf{p}).$$

  As a result,

  $$z_{\mathrm{avg}}(\mathbf{x}) \sum_{i=1}^{N} \bar{a}_i(d(\mathbf{x}))[w_{\mathrm{avg}} * \bar{B}_i(z)z](\mathbf{p}),$$

  so we will need to compute new basis images $\bar{B}_i(z(\mathbf{p}))z(\mathbf{p})$ along with the regular CSM basis images. This approach of computing the average blocker depth is called CSM-Z.
- The paper also proposes improvements on the Heaviside function's expansion.
  - With appropriate scaling, shifting, and subsequent clamping, ringing can be avoid completely. This can be done by shifting and scaling the response so that ringing occurs when the response is above 1 or below 0. If we clamp the response to $[0, 1]$, then ringing is avoided completely.
  - There's also the problem that the slope of the shadow test function is not sharp enough around $z - d \approx 0$. The paper applies a non-linear transformation $G(v) = v^p$ to the filtered shadow value $s_{w*f}(\mathbf{x})$ with $p \geq 1$. (If $p = 1$, then nothing changes.)

    However, this might remove smooth transitions from penumbra regions, so it selectively applies the transformation. When $d(\mathbf{x}) - z_{\mathrm{avg}}(\mathbf{p})$ is small, we know that $\mathbf{x}$ is near a contact point where light leaking will likely occur and penumbra is likely to be sharp. So, the paper chooses the exponent p as follows:

$$p = 1 + A \exp(-B(d(\mathbf{x}) - z_{\text{avg}}(\mathbf{p}))).$$

Here, the parameter A contains the strength of darkening, and B determines how fast darkening effect fades away as $d(\mathbf{x}) - z_{\text{avg}}(\mathbf{p})$ increases.

- The rendering algorithm:
  - The scene is rendered from the center of the area light source, and the z-value is written to the shadow map.
  - From the depth map, the following images are produced: the Fourier series basis and its complementary basis images multiplied by the shadow map z-values.
  - Mipmaps or sum-area-table of the above generate images are computed.
  - For each camera pixel, the following process is carried out:
    - The initial filter size is determined according to the cone defined by the intersection of the cone defined by the intersection of the area light source and the shadow map plane.
    - The average block depth is determined on the window defined by the initial filter size.
    - The final filter width is determined using the same algorithm employed by PCSS.
    - Then, the filtered shadow value is determined with CSM using the final filter width.
- The paper also proposes a greedy algorithm to decompose an environment map into a number of area light sources. Since this is not directly related to shadow mapping, we skip its discussion here.

References

[Annen et al., 2008] Annen, T., Dong, Z., Mertens, T., Bekaert, P., Seidel, H.-P., and Kautz, J. (2008). Real-time, all-frequency shadows in dynamic scenes. In *ACM SIGGRAPH 2008 Papers*, SIGGRAPH '08, pages 34:1–34:8, New York, NY, USA. ACM.

[Annen et al., 2007] Annen, T., Mertens, T., Bekaert, P., Seidel, H.-P., and Kautz, J. (2007). Convolution shadow maps. In *Proceedings of the 18th Eurographics Conference on Rendering Techniques*, EGSR'07, pages 51–60, Aire-la-Ville, Switzerland, Switzerland. Eurographics Association.

[Donnelly and Lauritzen, 2006] Donnelly, W. and Lauritzen, A. (2006). Variance shadow maps. In *Proceedings of the 2006 Symposium on Interactive 3D Graphics and Games*, I3D '06, pages 161–165, New York, NY, USA. ACM.

[Fernando, 2005] Fernando, R. (2005). Percentage-closer soft shadows. In *ACM SIGGRAPH 2005 Sketches*, SIGGRAPH '05, New York, NY, USA. ACM.

[Reeves et al., 1987] Reeves, W. T., Salesin, D. H., and Cook, R. L. (1987). Rendering antialiased shadows with depth maps. In *Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '87, pages 283–291, New York, NY, USA. ACM.

[Williams, 1978] Williams, L. (1978). Casting curved shadows on curved surfaces. In *Proceedings of the 5th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '78, pages 270–274, New York, NY, USA. ACM.