Silesian
University
of Technology

MASTER THESIS

Selected techniques of rendering shadow models in interactive computer graphics

Radosław Jacek RZECZKOWSKI
Student identification number: 290776

Programme: Informatyka
Specialisation: Interaktywna Grafika Trójwymiarowa

SUPERVISOR
Dr hab. inż. Agnieszka Szczęsna
DEPARTMENT Wydział Automatyki, Elektroniki i Informatyki
Faculty of Automatic Control, Electronics and Computer Science

Gliwice 2024

Thesis title

Selected techniques of rendering shadow models in interactive computer graphics

Abstract

The aim of this thesis is to survey different shadow rendering techniques, focusing on those that allow rendering in real-time, and test their performance and visual results. They are introduced in roughly chronological order, highlighting the improvements made to prior techniques over the years. They are briefly described along with references to their original sources and their use cases and downsides are discussed. Some of the techniques are implemented in a test application. For those, further implementation explanations are provided and tests are carried out.

Key words

computer graphics, rendering, shadow rendering

Tytuł pracy

Wybrane techniki renderowania modelu cieni w interaktywnej grafice komputerowej

Streszczenie

Celem tej pracy dyplomowej jest zbadanie różnych technik renderowania cieni, koncentrując się na tych, które umożliwiają renderowanie w czasie rzeczywistym, oraz przetestowanie ich wydajności i wyników wizualnych. Zostały one wprowadzone w przybliżeniu w porządku chronologicznym, podkreślając ulepszenia dokonywane na podstawie wcześniejszych technik na przestrzeni lat. Zostały one krótko opisane wraz z odniesieniami do ich oryginalnych źródeł, a także omówiono ich wady i zalety. Niektóre z technik zostały zaimplementowane w aplikacji testowej. W ich przypadku podano dalsze wyjaśnienia dotyczące implementacji i przeprowadzono testy.

Słowa kluczowe

grafika komputerowa, renderowanie, renderowanie cieni

Contents

Chapter 1

Introduction

1.1 The task of rendering

In this thesis, rendering is understood as the process of obtaining an image from a description of a three-dimensional scene. A scene can be rendered in a multitude of ways, with differences both in the specifics of the initial scene description and the rendered result. The rendered visuals can range from stylized to photorealistic. Rendering happens everywhere where a computer-generated imagery (CGI) is created for a viewer to see.

The wide spectrum of possible rendering results hints at the many ways in which rendering itself is performed. There are various techniques employed during the rendering process, which can be utilized together to create a desired look and fit within performance constraints. In more complex processes, there are many techniques used to render a scene, each responsible for modeling the visual aspects of different real-life phenomena or artificial effects. A renderer could be capable of adding stylized edge detection and cell-shading to an image, rendering glossy and rough surfaces, simulating shadows, reflections, caustics and refraction, dealing with hair and fur or volumetric participating media such as fog, smoke and clouds. Each of the mentioned effects can be rendered using one of many techniques, and many of them are being actively improved upon and researched.

The fact that there are many techniques currently in use to render a single type of effect creates and advantageous situation, where the techniques used can be chosen for each application depending on its characteristic. Two main factors can be discerned as defining the needs of an application: the performance requirements and the desired visual style.

The desired visual style is partially just a matter of preference defined by the style of the project. More importantly however it is a matter of clearly and efficiently conveying visual information, in a way that is consistent with the rest of the application and with what the end user expects. This means that realism, often touted the pinnacle and goal of computer graphics, is not necessarily always the best approach. A CAD (computer-aided

design) application or a 3D modelling tool would be made less useful by including realistic reflections, highly contrasting full shadows and motion blur in the rendered viewport. These programs need to clearly convey information about the shape and design of 3D objects, without distractions and obstructions. On the other hand, when a player starts a modern action-adventure game, they expect a level of realism in the game's graphics that allows for immersion in the presented world. The intricate visuals also make for a more engaging experience and can mean a better reception of a game. At the same time, design choices should be made to ensure that the realism or intricacy of the presented graphics do not get in the way of enjoying the gameplay, which in the end should be the main attribute of a video game.

The performance requirements, or performance constraints, are usually better defined and divide rendering into two general categories: real-time rendering and offline rendering. When designing an offline renderer the performance constraints would most likely concern memory usage and possibly general time, or energy, efficiency, as with offline rendering the time it takes to render an image is of small importance. Most important are image quality and fidelity, often also realism. Techniques used in this context can spend as much time performing calculations as is necessary for the desired output. Offline rendering work can also easily be distributed between many machines, so-called render farms. Real-time rendering on the other hand works with very strict and small time budgets. Because a real-time application needs to be responsive to user inputs and give the illusion of continuous motion it is expected to render at least 30 frames per second (FPS), giving the time budget for a single frame of at most 0.03 seconds. This time cannot all be spent on rendering, as user inputs need to be handled and application logic performed in the same time frame. Because of that, real-time applications need to balance visual complexity and performance. They are also often created for more casual consumers than offline rendering programs, so the hardware on which the application will run is expected to be moderately powerful.

1.2 Objective

It is apparent that the choice of rendering techniques is complex and requires in-depth knowledge about each of them as early as the design stages of an application. This thesis aims to provide a guide through the rendering techniques used to render cast shadows in scenes, focusing on those applicable in real-time contexts. Some of the chosen techniques are the most widely adopted techniques, but some, while possibly less popular, are still noteworthy due to their improved performance or visual quality. This thesis introduces the techniques, gives detailed explanations of their algorithms for a number of them, shows exemplary implementations and compares them in a series of tests. The comparison is made based on measures of performance (execution time), memory consumption and

visual fidelity.

To that end a test application is written, giving the ability to test different shadow rendering techniques and interact with the scene in real-time. This not only allows to test the techniques in practice, but also improves the understanding of each technique and allows to give more detailed explanations.

1.3 Chapter contents

In chapter 2, first the rationale behind the importance of shadows in computer generated images is described, followed by descriptions of specific shadow rendering techniques and their pros and cons. They are introduced in the context of the original publications that describe them and the techniques that came before them.

In chapter 3 the subject of the thesis is described, being the created test application and implementations of selected shadow rendering algorithms. Implementation details and author's observations from the implementation process are detailed.

Chapter 4 contains the experiments that were performed and their results. Tested are mainly the performance of each technique and the visual results obtained. The tests are carried out using few different scenes to observe how scene complexity impacts the performance. All experiments are accompanied by a discussion of the results and additional observations, for example based on profiling data.

Finally, chapter5 gives a summary of the work performed within the scope of this thesis, reviews the test results and conclusions from them and proposes possible future research directions.

1.4 Author's contribution

The author of this thesis is the sole author of its contents, including text, figures, drawings, renders and data, unless specifically stated otherwise. All materials not created by the author are appropriately referenced, including information and knowledge that is utilized thought the thesis. Within the scope of this thesis the author also created a test application with the use of open-source or third party software that is utilized in accordance with respective licenses.

Chapter 2

Problem analysis

2.1 The role of shadows in computer generated images

Computer graphics usually aim to reflect our surrounding reality, with techniques for rendering phenomena modeled from physical laws and after observations made in the real world. When not aiming for realism, the displayed graphics need to be at least rooted in the common human observation experience for the information to be understandable and interpretable. Shadows are a phenomenon that is very commonly present in our daily lives, deeply rooted in the common experience. An 1997 article [7] shows that even simple shadows provide the viewer essential information on the position and movement of objects, especially movement in depth. In some cases this information even overrides other cues such as change of size due to perspective. This effect is illustrated by Fig. 2.1.

The importance of shadows for understanding the 3D composition of a scene is especially great when an observer is devoid of stereo visual information, such as in the case of observing an image displayed on a regular monitor. Shadows, lighting and perspective then become the sole sources of information about depth and the relative positions of objects in the scene.

These observations are also supported by a more recent 2018 review on the perception of shadows [21]. It suggests more possible roles that shadows play in human perception of the world. Shadows help observers position objects in 3D space. Sometimes shadows are perceived as being a part of the object, so the lack of a shadow would make the object incomplete. Information about the shape of an object can also be derived from the shadow that it casts, as presented in Fig. 2.2.

The review also mentions that shadows in art, even if depicted in a way that is impossible in the real world, are an important element for an image to be perceived as realistic and of high quality. This also applies to computer graphics, making rendering shadows a necessary element for any application that requires elevated realism or sophist-

ication of the visual presentation. The difference between a scene with and without cast shadows is presented in Fig. 2.3 It is worth noting that since observers are not great at spotting physically incorrect shadows, as is stated in the studies, so graphics programmers can utilize this fact and take liberties with the realism of their shadows. This can lead to simplifying the algorithms producing them and improving their performance, with negligible impact on the apparent visual quality.

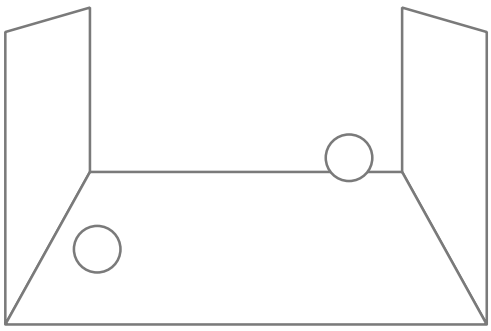
2.2 Shadow rendering techniques

2.2.1 Naming conventions

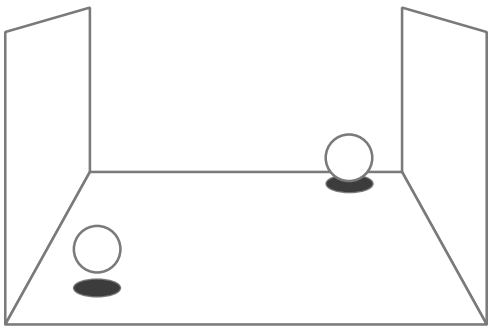
In this thesis the term 'shadows' is used interchangeably with term 'cast shadows', not to be confused with shading on the surface of objects, which stems from the orientation of the surface with regard to the light source.

As presented in the "Real-Time Rendering" book [1], a shadow is cast by a shadow caster, also called an occluder, onto a shadow receiver, when the caster occludes a line of sight from a point on the surface of the receiver to a light source.

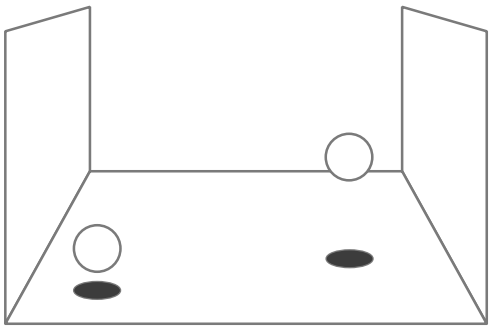
Shadows can be differentiated into hard and soft shadows. Perfectly hard shadows are not observed in reality, as such a shadow would require an infinitely small light source.



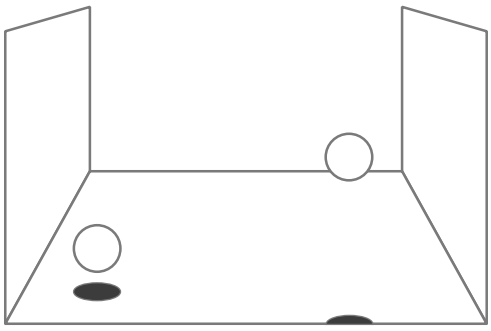
(a) Shadowless scene, the positions of the spheres are ambiguous.



(b) The upper ball seems placed on the surface.



(c) The upper ball seems to hover halfway.



(d) The upper ball seems high in the air and closer to the observer.

Figure 2.1: The position and scale of the spheres do not change, yet the perceived locations in 3D space are dictated by the shadow.

(a) The left shape could be a sphere, the right shape could be flat, like a coin.

(b) The left shape could be a cube, and the right a shape with top and bottom faces not being parallel.

Figure 2.2: The shadow can convey information on the shape of the object. It additionally describes whether an object is placed on a surface or above it.

Hard shadows however are often found in computer-generated images, as they are relatively simple and cost-efficient to render. Soft shadows require more complex solutions, but enhance the realism of the rendered scene. A soft shadow, as shown in Fig. 2.4 consists of an *umbra*, being the fully shadowed region, and *penumbra*, meaning the region that is partially in shadow. The appearance of a soft shadow, the sizes of the umbra and penumbra defining its softness, depend on the relative distances between the shadow caster, the shadow receiver, the light source and also the size of the light source.

2.2.2 Coarse distinction of shadow rendering techniques

In this thesis a coarse distinction of the shadow rendering techniques is proposed, dividing them into three groups. The division is decided based on the main mechanism utilized in a technique. Additionally, techniques in each group share similar issues that researchers have attempted and are still attempting to resolve with newer iterations of the original algorithms.

The first group consists of shadow map-based, or image-based, techniques. They render the scene from the viewpoint of the light and store how far away scene geometry is in a depth buffer. This is then used when rendering the actual view of the scene to determine which parts are in shadow and which are lit. A common issue for these techniques is aliasing, which occurs when signals are sampled at different rates, causing artifacts. In the case of computer graphics, spatial aliasing is most often observed, where the source signal, being geometry, a texture, or a depth buffer as in the case of shadow mapping, is sampled in discrete locations in space at varying sample densities. It is not trivial to counteract aliasing in shadow map techniques and there are many approaches to solving this problem. On the other hand, it is relatively simple to achieve approximate soft shadows, with more advanced techniques allowing for more realistic results.

The second group contains geometry-based techniques. These approaches modify the scene geometry to create shadow volumes, which contain areas where light from a source does not fall. The shadow volumes are then rendered with a clever use of a stencil buffer, producing the information whether a surface point is in light or not. Such techniques

(a) The scene rendered with no cast shadows, despite the high quality of models and materials, looks unrealistic and flat.

(b) The scene rendered with cast shadows is realistic and reads well.

Figure 2.3: The same scene rendered twice using Blender's Cycles renderer, with and without cast shadows. *Scene source: "Classroom" by Christophe Seux, Blender demo files.*

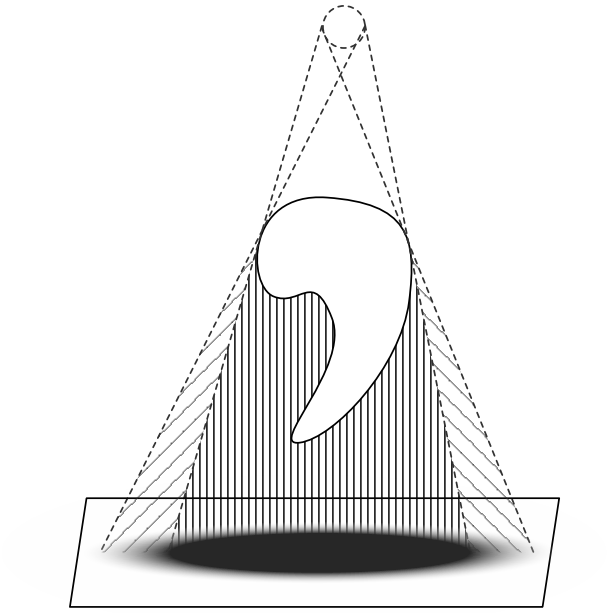


Figure 2.4: The umbra (under vertical lines) and penumbra (under slant lines) create a soft shadow.

suffer less aliasing problems than image-based ones, producing perfectly sharp shadow information on a per-pixel basis. They however introduce a lot of new geometry to the scene, the rendering time of which can be large and difficult to predict. It is also more difficult to achieve the effect of soft shadows with this approach.

The third group are ray-based techniques. They have only recently started being utilized in real-time scenarios, as they can be computationally costly and require powerful, modern hardware to run at interactive rates. This group of techniques is possibly the most intuitive, as it does not really rely on clever utilization of the rasterization hardware of a graphics card. To determine if a surface point is in shadow, a ray is cast in a straight line from the point to a light source. If any geometry is encountered along the ray, then the light source is occluded and the point is in shadow, otherwise it is lit. These techniques, apart from relying on powerful hardware, utilize specialized scene description data structures that accelerate these ray queries. They do not suffer from aliasing and can be extended to give photorealistic and physically-correct results.

The following sections introduce the techniques based on the publications which first described them. The sections attempt to mention them in logical order, highlighting the improvements and additions made to each over the years. More detailed descriptions of the algorithms used to implement these methods will be given in chapter 3. Additionally, the next section presents a now largely unused approach, that does not really fit into any of these categories, but is useful to build a geometric intuition of how shadows form.

2.2.3 Planar shadows

Planar shadows are one of the first techniques used for implementing shadow rendering and are not very robust, so they are mentioned in this thesis for historic reasons. They can be a help with understanding the geometry involved with shadows and their shape. That is because planar shadows are rendered using basic geometric projection.

Introduced by James Blinn in 1988 [3], planar shadows are created by projecting the shadow caster onto a flat surface of the shadow receiver. The caster vertices are projected along a vector between the vertex and a light source onto a planar surface of the receiver. This is achieved with a projection matrix, which multiplied with the vertex positions of the caster gives a new mesh, flattened onto the planar surface. This mesh can then be rendered on top of the shadow receiver with a black material, giving the illusion of a shadow. The technique is described in more detail in section 3.2.1.

While very simple and intuitive, this approach has many shortcomings that nowadays render it obsolete. The application needs to keep track of shadow casters and receivers separately, to know which objects should be projected onto which surfaces. Issues occur when a light source is between the caster and receiver. In such a situation an incorrect anti-shadow is created, as the operation of projection through a point is still valid. The receivers have to be planes, and special care needs to be taken to only draw the shadow on top of a receiver. Since the shadow is a separate mesh, it could be rendered beyond the mesh of a receiver. Additionally, objects cannot shadow themselves. On the other hand planar shadows give a lot of artistic control over the appearance of shadows and even their shape, as they can be manipulated just as any other mesh.

2.2.4 Basic shadow mapping

The general algorithm for shadow mapping was initially described by Lance Williams in 1978 in the article “Casting curved shadows on curved surfaces” [26]. The title immediately hints at the biggest strengths of this technique at the time: it is robust, allowing any scene objects to cast hard shadows onto others, regardless of their shapes and without the need for special treatment of casters and receivers as with planar shadows. In fact, similar shadow mapping techniques that utilized the z-buffer were already known before this article, but they could only be used for geometry consisting of planar polygons. This technique can be used to render shadows anywhere where any scene rendering can be obtained at all, which includes smooth, curved parametric surface patches.

The article presents the general method of shadow mapping, along with its advantages and disadvantages. The rationale of the technique is based on the fact that all surfaces ‘seen’ by a light source from its point of view are lit by it, while all others are in shadow. To determine surface visibility for a light source, the scene is simply rendered from the viewpoint of the light source. This rendering is called the shadow map. Such an interme-

diate rendering, which will not be presented on screen but is only a step in the process of creating the final render of a scene, is often called a render pass. This rendering does not need to produce a color image, but only a depth- or z-buffer. This buffer is used in the standard rasterization pipeline to perform depth culling, so no special implementation is needed. Moreover, the existing highly optimized and hardware-supported rasterization architecture is utilized. The z-buffer stores the distances from the observer to the closest surface point in the scene per-texel. Once this view is obtained, the scene can be rendered again, in the usual manner from the point of view of the observer. When deciding whether a surface point is in shadow or not, the distance between the point and the light source can be calculated and compared to the value stored in the previously obtained shadow map. If the distance of the surface being rendered to the light source is greater than the depth read from the shadow map, then the surface is in shadow, and some other object that is closer to the light source occludes it. Otherwise, the surface is lit and further shading computations can take place, such as diffuse and specular lighting calculations.

Variants of shadow mapping have been described over the years, and techniques based on this approach are currently most widely utilized for real-time rendering. This is not surprising when considering the advantages of this method. The issues present in planar shadows such as the need to separate shadows casters and shadow receivers, difficulty casting shadows onto non-planar surfaces, impossibility of obtaining self-shadowing and possibility for shadows to appear in empty space beyond any receivers are all solved with shadow mapping. The technique is also relatively simple to implement and utilizes the rasterization pipeline in its full potential. The cost of the algorithm per light source is roughly twice the cost of rendering without shadow mapping, as scene objects need to be transformed and rendered into the z-buffer. The cost of rendering the shadow map will however be almost always less than rendering the scene, as there are no per-pixel lighting computations done to obtain color. In modern rendering pipelines the color render target buffer can be omitted altogether, making it also possible to not run the pixel shader at all.

The disadvantages of shadow mapping are outweighed by its advantages, but are not trivial to combat. One disadvantage is that basic shadow mapping generates only hard shadows, which are not realistic and are prone to visible aliasing in the final image. Multiple specializations of this technique exist that make it possible to render soft shadows with varying levels of realism and correctness. Another significant issue is the fact that this technique introduces a second level of aliasing to the shadows. This is because the shadow map itself is an image representation of the scene, making it a discrete view of continuous scene geometry. This creates a few problems that need to be addressed. For one, numerical representation resolution errors can be introduced into the shadow map. This can be easily minimized by rendering into a 16-bit or more z-buffer and setting the near and far clipping planes of the shadow map view to tightly encompass the scene. The

second problem being self-shadowing, also called surface-acne, is much more prominent and immediately observable in a naive implementation. As the samples of depth values stored in the shadow map are texels, they represent a non-zero area for a point location on a surface. When the exact depth values calculated when rendering the main view of the scene are compared with these spatially discretized samples, a surface can erroneously shadow itself, creating high-frequency, high-contrast patches of shadow, which often create intrusive moiré patterns on screen. This problem can be reduced by using a higher resolution shadow map, resulting in more dense depth sampling and decreasing the size of self-shadowing patches. This however is expensive and will not fully resolve the problem. Another method to deal with surface-acne is to use a depth bias when rendering the shadow map. A depth bias is a value that is added to the actual depth values before storing them in the z-buffer. By adding a positive depth bias, the depth of a surface can be pushed further away from the light source, under the actual surface. This way the actual distance between the light source and the surface being rendered in the final pass will always be less than the depth stored in the shadow map. This however can cause issues with light-leaking, or peter-panning, where contact shadows get detached from their casters. Methods to counteract this, as well as an example implementation will be shown in chapter 3.2.2. More issues with shadow mapping will be described in the following sections, along methods to circumvent them.

2.2.5 Fitting shadow maps

Fitting shadow maps is the simplest technique that can be used to reduce shadow map aliasing problems, by better utilization of the available shadow map resolution and precision. The main idea of fitting is to avoid rendering into the shadow map objects that are not necessary for the current scene and viewpoint. Initially introduced by Brabec [4], the method was developed over time and can be implemented in different levels of sophistication, more complex ones giving better results.

The simplest way to fit a shadow map is to tightly encompass the entire scene with the view frustum of the light source. This can work well enough for small scenes, but will be problematic for large or densely populated scenes.

One can set the light frustum width and height to encompass just the frustum of the viewer. Additional care needs to be taken to actually include all shadow casters that may cast shadows onto the scene objects present in the view frustum, as they may not necessarily be within it. This concerns only the near and far clipping planes of the light, and can be simply achieved by setting them to the extents of the scene.

There are more advanced techniques, including using the depth information of the scene observer to analyze the distribution of future shadow map samples in space and produce light frusta that tightly bind the rendered geometry [9]. This requires a depth

map of the scene from the observer's point of view to be available when rendering the shadow map, which can introduce additional performance cost but can also be hidden if the rendering pipeline already requires a depth pre-pass.

Fitting, since it depends on the view position and direction, can be temporally unstable. When the viewport changes there can be noticeable flickering on the edges of the shadows, caused by temporal aliasing. This can be minimized by quantizing the offset of the shadow map that is used when fitting to texel increments.

2.2.6 Warping shadow maps

Warping shadow maps was introduced by Stamminger [23] and aims to counteract perspective aliasing. Perspective aliasing means that more samples of the shadow map are taken closer to the viewer due to perspective foreshortening, meaning that a higher concentration of shadow map samples is required there than further away. A warped shadow map accomplishes exactly that with minimal performance impact by using a specific parametrized perspective transformation on the scene geometry before rendering into the shadow map. The mentioned technique is called perspective shadow maps (PSM). More approaches were build basing on this proposition, namely light space perspective shadow maps (LiSPSM) [27] and trapezoidal shadow maps (TSM) [15]. Both these techniques are very similar and give better results than PSM. It has been proven that LiSPSM actually gives best results out of the three as it leads to the best distribution of error over the shadow map area. [11]. These warping techniques can actually be used in tandem with shadow map partitioning approaches.

2.2.7 Filtering shadow maps

While the former techniques were mostly meant to counteract initial sampling aliasing by utilizing the resolution of the shadow map in the best possible way, shadow map filtering methods are mainly meant to help reduce aliasing during the resampling stage. Resampling happens when the shadow map is actually used to determine shadowed and lit areas in the final render, when the appearance of the shadows is determined. Usually to counteract such aliasing various filtering methods are employed. This is also the case in many areas of computer graphics, including sampling of different types of textures. Filtering can get rid of high frequency components from processed signals, such as pixelated edges in textures. Color textures can be filtered in simple and intuitive ways, for example by using linear filters or mipmaps.

A shadow map is special in the context of filtering, because of its specific content. It actually cannot be filtered directly. When shadow map filtering is considered more carefully, it becomes apparent that it is not the depth values that need to be filtered, but the actual shadow information obtained based on these depths from the shadow map and

depths of points in view. A shadow map filtered by, for example, linear interpolation would contain depths of non-existing surfaces, witch is not the desired outcome. Additionally, the shadows themselves would still be sharp and aliased, because the depth comparison function has binary output: a pixel is either in shadow or not.

Filtering shadow mapping has been approached in a few ways. One of them actually focuses on filtering the shadow information itself in the final render. The percentage-closer filtering (PCF) [20] algorithm does exactly that. The shadow computation is done in the regular way, but it is performed multiple times per pixel. A filter window and a kernel are used to sample the shadow map at offsets around the point that is being shaded, and the binary comparison results are averaged to obtain a filtered shadow. More robust techniques based on PCF improve the visual quality and performance, such as using specialized kernel weights and sampling patterns, for example a Poisson disk. The Poisson disk sampling pattern allows using fewer samples and covering a larger area. It can also be randomly rotated to avoid banding artifacts. Another one of such approaches was described by Yury Uralsky in the GPU Gems article "Chapter 17. Efficient Soft-Edged Shadows Using Pixel Shader Branching" [19]. It uses specifically generated arrays of sample offsets that are arranged in such a way that samples further away from the origin are first in the array. This way, when performing PCF, only the outer samples can be tested first. If their values agree, meaning the point is either fully in shadow or fully in light, the calculations can stop, allowing to omit most of the actual sample count. This early test-out provides significant performance gains. The technique also deals very well with banding, which is obvious in basic PCF. These techniques can be further improved to give smoother results by utilizing percentage-closer filtering that is built into the graphics hardware. By setting an appropriate sampler for the shadow map, bilinear PCF, with a $2x2$ kernel, can be achieved at virtually no additional performance cost. PCF is further described in section 3.2.4.

Another way to approach filtering shadow maps is to utilize mathematical and statistical principles to describe the shadow signal in a way that can be precomputed and filtered. One of the first such techniques were variance shadow maps (VSM) [5]. Mean and variance of the shadow map values are computed and later used to calculate the upper bound of the fraction describing how much light the receiver sees. This can lead to light leaks near object edges, which introduce high variance in the depth values, but significantly improves performance by allowing traditional filtering methods to be carried out on the depth map. VSM is further described in section 3.2.5. Many improved techniques have been developed basing on this achievement. One of them are layered variance shadow maps (LVSM) [8], which deal with light leaks by modifying the stored depth values to introduce less variance.

Variance shadow maps give better performance than methods like PCF and allow to filter the shadow map in traditional ways with multisampling and mipmapping, but are

prone to producing noisy artifacts in certain areas. A much more robust and recent approach is proposed called moment shadow mapping (MSM) [18]. It uses similar principles to VSM by computing and storing shadow map moments that statistically describe their contents, but instead of using two moments it uses four. This allows for much higher quality of the produced shadows and creates an overall more robust solution. The authors propose an algorithm that allows to store all four moments in a single 64-bit shadow map.

Revectorization-based shadow mapping is an interesting and an entirely different technique to combat shadow aliasing [13]. Instead of blurring the shadows to hide the aliasing, in RBSM the idea is to analyze the shadow map using discontinuity filters and compute discontinuity vectors for the aliased shadow boundaries. Based on this information, the shadow boundaries are revectorized, producing results similar to shadow volumes, with no visible shadow map texel boundaries. The article claims that the technique can provide real-time performance and can be used in applications such as video games. The authors also propose an approach to render soft shadows using this technique called revectorization-based soft shadow mapping [14]. They use a special revectorization-based filtering algorithm that can be used in screen space to create soft shadows.

All these techniques would struggle with rendering shadow effects with the inclusion of participating media such as smoke or fog, hair or fur. A technique called deep shadow maps aims to tackle this problem [12]. In this kind of shadow map, instead of storing the depth of the closest occluder, a function of z is stored per texel, whose value is that of the fractional visibility in the given direction. This approach creates prefiltered shadow maps which improve performance and visuals.

2.2.8 Rendering soft shadows with shadow maps

The techniques described so far can be used to create hard shadows and to smooth these hard shadows to help with aliasing. However, as discussed at the beginning of this section, to get realistic, or at least plausible looking shadows, the relative distances between the light source, receiver and occluder as well as the light size need to be taken into account, creating soft shadows. Soft shadows and their approximations are possible to obtain with shadow maps and are performant enough to allow for real-time implementations.

A realistic baseline can be set by algorithms that generate multiple shadow maps per light source, each time rendering from a different sampled point on the surface of the light. During rendering shadow mapping results for each of the maps are accumulated to create gradients in the shadowed areas, giving physically correct results. This approach however is very computationally intensive and not very practical in real-time scenarios. Attempts to make it possible to use in real-time have been made, such as the algorithm proposed by Scherzer [22]. Each frame more shadow maps for more light area samples are

created, and the quality is refined over time, with a relatively small performance impact. This however causes issues in motion, as objects and areas of the scene that have just appeared in the viewport need time to accumulate enough samples.

A very popular solution to the problem of rendering real-time soft shadows is percentage-closer soft shadows (PCSS) [6]. It basically utilizes the existing PCF technique, which can produce shadows of varying softness using its kernel size and offset scale. PCSS evaluates the PCF filter size based on an estimate of the size of the shadow penumbra, which is derived from depth information from the shadow map. It is therefore relatively simple to implement into a system that already uses PCF. The results, while not entirely physically correct, are good enough to fool the eye and certainly better than only hard shadows or uniformly blurry shadows. This approach is described in more detail in section 3.2.6.

PCSS can be accelerated by substituting PCF with techniques mentioned in the last section, such as CSM or VSM. With these the varying filter sizes can be accomplished by, once again, utilizing hardware supported algorithms like mipmapping. More advanced methods like summed area tables can be also used, allowing the computation of any sized rectangular filter using only four lookups. These methods are however significantly more involved implementation-wise.

Chapter 3

Subject of the thesis

3.1 Test application

For the purpose of getting familiar with the techniques and testing them, a test application was created. It allows a user to switch between implemented rendering techniques and observe their effects in an interactive first-person scene.

The application was written in the C++ programming language. It uses the Win32 platform to create windows in a Windows operating system. It uses DirectX 12 to utilize the GPU for rendering graphics. DirectX 12 was chosen because it is the newest of DirectX versions, allowing for granular control over the rendering process and boasting advanced features. It is also very widely adopted for all sorts of computer graphics tasks, making it worth specializing in. Additionally, the DirectX 12 API is much more verbose and involved as compared to, for example, OpenGL, which forces the programmer to gain more in-depth knowledge about graphics programming and all the processes involved.

Two notable open-source libraries were utilized in the project. Dear ImGui is a library simplifying the creation of graphical user interfaces. It is compatible with all the major graphics programming APIs, including DX12, making it fast and easy to integrate into a project, providing am intuitive and smooth user experience at the same time. The other library is the Tracy Profiler. The instrumentation library is just a part of the Tracy ecosystem, with separate applications like the profiler to gather and analyze performance data. It is simple to incorporate into an existing project and very performant, adding almost no overhead. Code zones and frames that will be captured when profiling are marked manually by the programmer, giving them full control over what data is captured. This also improves the readability of the outputs, which stands in contrast with other profilers that automatically gather data and can easily overwhelm the user and clutter the presentation with the amount of information provided.

The project uses CMake to fetch all the project dependencies and create a buildable MSVC project.

3.2 Implemented shadow rendering techniques

In the following sections, all the shadow rendering techniques that have been implemented during work on this thesis are described, along with explanations of their algorithms.

3.2.1 Planar shadows implementation

Planar shadows, as described in section 2.2.3, are a very simple technique to achieve hard shadows, introduced by Blinn [3]. The idea for the algorithm stems from the geometric understanding of how a shadow is formed. The shape of a shadow is basically the shadow caster projected onto the surface of a shadow receiver along rays from a light source, as shown in Fig. 3.1.

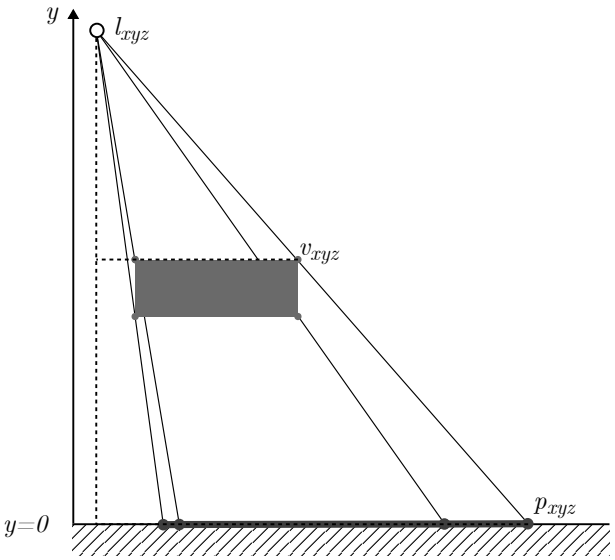


Figure 3.1: Vertices of a shadow caster projected onto $y = 0$ plane create a planar shadow. Dashed lines visualize the similar triangles.

Using this geometric understanding and similar triangles, marked in Fig. 3.1 in dashed lines, the following equations and a projection matrix can be derived. Assuming l to be the position of a light source, v to be the position of a vertex of the shadow caster and p the position of the projected shadow vertex, a projection onto the $y = 0$ plane is described by:

$$\frac{p_x - l_x}{v_x - l_x} = \frac{l_y}{l_y - v_y} \iff p_x = \frac{l_y v_x - l_x v_y}{l_y - v_y} \tag{3.1}$$

The z component is derived in analogous way:

$$\frac{p_z - l_z}{v_z - l_z} = \frac{l_y}{l_y - v_y} \iff p_z = \frac{l_y v_z - l_z v_y}{l_y - v_y} \tag{3.2}$$

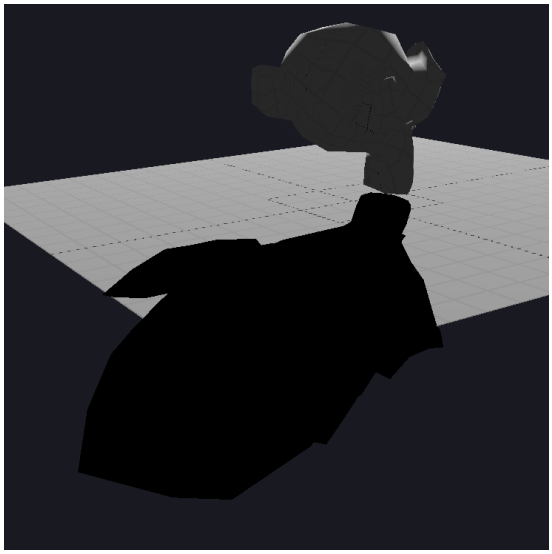
The y component will be $y = 0$ since the projection happens onto the xz plane. From that the projection matrix can be constructed:

$$\mathbf{M} = \begin{pmatrix} l_y & -l_x & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & -l_z & l_y & 0 \\ 0 & -1 & 0 & l_y \end{pmatrix} \tag{3.3}$$

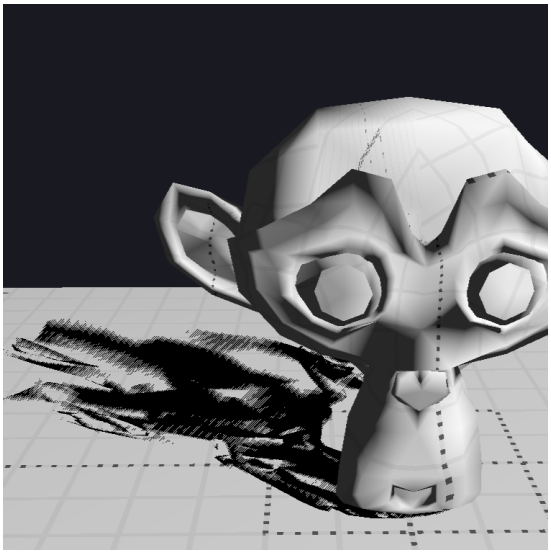
The $l_y - v_y$ factor in the denominator is obtained by utilizing the property of homogenous coordinates, where all components of a vector $\mathbf{v} = \begin{bmatrix} x & y & z & w \end{bmatrix}^T$ get divided by the w component, which gives $\mathbf{v} = \begin{bmatrix} x/w & y/w & z/w & 1 \end{bmatrix}^T$.

This matrix can be generalized to project points onto any plane, but this was not implemented.

In the implementation the matrix is created in the vertex shader based on the position of the light source. The vertices of the caster, after being transformed into world space, are multiplied by it, projecting them onto the $y = 0$ plane. Then they are transformed with the view and perspective projection matrices as usual. The resulting mesh is rendered in black, with a pixel shader that just returns the black color. Special care should be taken when rendering the shadow mesh, since it perfectly overlaps the ground plane at $y = 0$. To avoid z-fighting, which is shown in Fig 3.2a, this mesh needs to be rendered after the receiver and before the caster, with z depth testing turned off in the rendering pipeline, or, as is the case in the implementation, a negative depth offset needs to be added to the mesh to bring it above the ground plane.



(a) The projected shadow mesh lies beyond the ground plane.



(b) Severe z-fighting caused by the fact that both the ground plane mesh and shadow mesh lie in the exact same plane.

Figure 3.2: Some of the issues that can arise when using projection shadows.

Multiple issues arise with this technique. The receiver must be a flat plane, objects cannot shadow themselves, casters and receivers need separate treatment. Additionally, care needs to be taken to only render the shadow meshes on top of the geometry of a receiver. Since the shadow is a separate mesh, it could be cast beyond the actual surface of the receiver, creating an impossible shadow in the air. This is illustrated in Fig 3.2b There is another possibility here for an impossible shadow. If the light source is between the caster and receiver, an anti-shadow will be cast as shown in Fig 3.3. This creates another special case that needs to be checked for by the application. Soft shadows are also difficult and costly to achieve, created by projecting the caster geometry multiple times using different discrete locations sampled over the area of the light source. The shadow meshes can then be blended, producing a darker color where more shadows are present. This gives very accurate results with a high number of samples, but is not a viable option for real-time rendering. All these problems make this technique virtually unused in modern rendering engines.

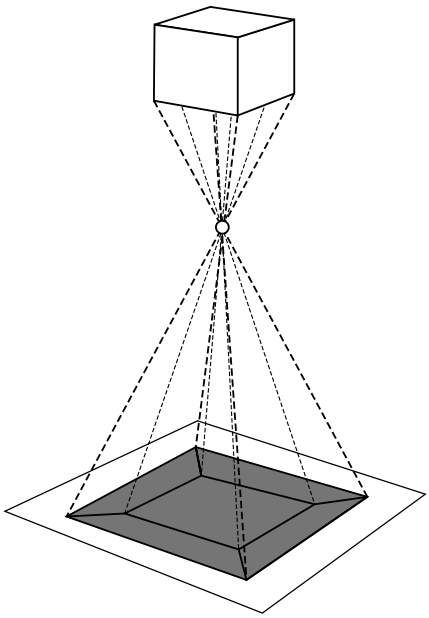


Figure 3.3: The shadow caster projected through a light source, creating an anti-shadow.

3.2.2 Implementation of basic shadow mapping

As mentioned in chapter 2.2.4, shadow mapping for a single light source is performed using two render passes. The first one generates the shadow map itself, the second and final pass renders the scene with shadows determined using the shadow map. The shadow map is a z-buffer, which means that, for each texel, it stores the values of the z coordinate of the scene surface point being rendered. These coordinates are stored after transforming the vertices of the scene into the NDC space, or normalized device coordinate space, of a given light source. This transformation is achieved in the vertex shader, by using the

regular world, view, projection (WVP) transformation matrices, with the distinction that the view and projection matrices are not of the scene observer but of the light source. In truth, multiplying the vertex positions by the WVP matrix will result in clip space homogenous coordinates, which then will be transformed automatically by the pipeline into NDC during the perspective division, when the homogenous coordinates are divided by their w component. The stored depth values will then be in the $0 - 1$ range, usually 0 being at the near clipping plane and 1 at the far clipping plane. This z-buffer can then be used as a texture shader resource in a pixel shader to render the final scene. To read the depth values from the shadow map, the NDC coordinates of each surface point need to be calculated again. The x any y components will then be used to access a texel in the shadow map, the z component will be compared with the contents of the shadow map. Do note that the actual Euclidean distance between the shaded point and the light source is not used, as these values would be in different coordinate spaces. When the shadow map value is accessed, it is compared with the z value of the point that is currently being shaded. If the shadow map value is greater, then the point being shaded is farther away from the light source than some other surface seen at this point from the point of view of the light. This means that it is not illuminated and is in shadow. Otherwise, the point is lit by the light source. The process is illustrated in Fig. 3.4. Fig. 3.5 shows a rendered view of a scene with shadow mapping and the shadow map itself.

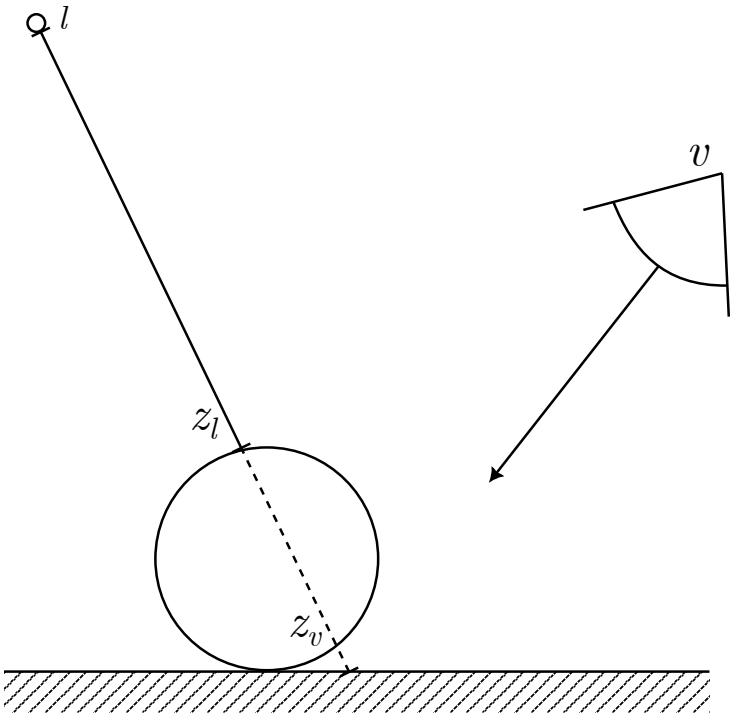
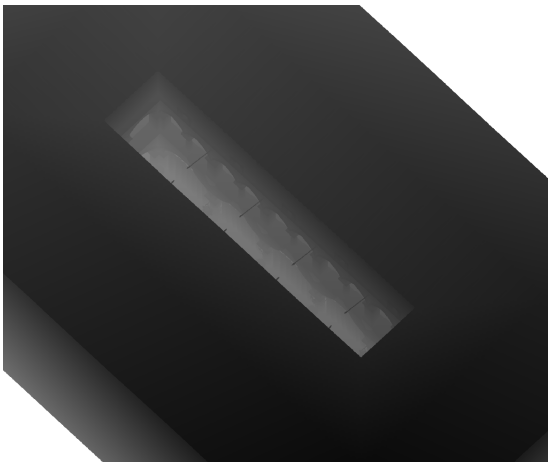


Figure 3.4: The light source l has stored the depth value z_l in its shadow map. When rendering the scene from the viewpoint of the observer v, the stored depth is compared to the depth z_v. This value will be larger than the one in the shadow map, so the point will be in shadow.

(a) A scene rendered with the use of shadow mapping.



(b) Section of the shadow map rendered for this scene.

Figure 3.5: An example scene rendered with a shadow map, which is shown in black and white.

To implement this technique a resource is needed that will be used both as an additional depth buffer and a texture. This can be easily achieved in an application programming interface (API) like DirectX 12 by creating one resource and two resource views, one being a Depth Stencil View (DSV) and another a Shader resource view (SRV). Once the rendering of the shadow map is finished using the DSV, the resource can be transitioned in preparation to be used as an SRV during rendering the main view using a resource barrier. Caution needs to be taken to transition the resource back to a DSV-compatible state before rendering the shadow map of the next frame. In this case, a resource barrier will also ensure that all write operations have finished before continuing with rendering, meaning that the rendering of the shadow map will be complete before it is used in the final pass. The shaders responsible for rendering the main scene view will evidently need access to the texture resource containing the shadow map, but also it will need to access the view and projection matrices of the light view pass. This is to calculate the corresponding NDC coordinates of the point being shaded in the shadow map. It is optimal to calculate light clip space coordinates in the vertex shader and let the pipeline interpolate them across the triangle for each fragment. Since the perspective division will not happen automatically here, it is needed to divide the coordinates by the w component in the pixel shader to obtain correct NDC values. Since the shadow map is sampled as a texture, the xy coordinates with range $[-1:1]$ need to be converted into uv texture coordinates with range $[0:1]$. Additionally, the y component needs to be flipped to address the texture correctly. This can be performed as a simple transformation in the pixel shader as $\begin{bmatrix} x & y \end{bmatrix}^T / \begin{bmatrix} 2 & -2 \end{bmatrix}^T + \begin{bmatrix} 0.5 & 0.5 \end{bmatrix}^T$ or incorporated into the MVP matrix of the light view. This can be achieved by multiplying the MVP matrix for the light source by the offset matrix as shown in equation 3.4.

$$\mathbf{L_{offset}} = \begin{pmatrix} 0.5 & 0 & 0 & 0.5 \\ 0 & -0.5 & 0 & 0.5 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \cdot \mathbf{MVP} \tag{3.4}$$

When sampling the shadow map as a texture, it is important to do so with a sampler that does not perform any filtering on the texture. Typically, texture filtering helps avoid different kinds of aliasing. Regular color textures however can be filtered and interpolated as they only store color data. The shadow map, which stores depth values, cannot be filtered in the same way as the produced values would be senseless in this process. They would create non-existing apparent surfaces especially noticeable between sharp depth changes, for example between a close object and the far clipping plane.

This observation brings forth one of the issues with shadow mapping: they cannot be filtered to avoid aliasing. One could also attempt to blur a shadow map to achieve soft shadows, but this is not a viable option for the same reason. Because of this basic shadow mapping can only be used to create hard shadows that suffer from aliasing issues, creating shadows that can appear blocky when up close, as in Fig. 3.6. Aliasing can be mitigated by rendering the shadow map into a higher resolution target, but only to a degree. Special techniques allowing for smoothing and filtering shadows generated with shadow mapping will be presented in the following sections.

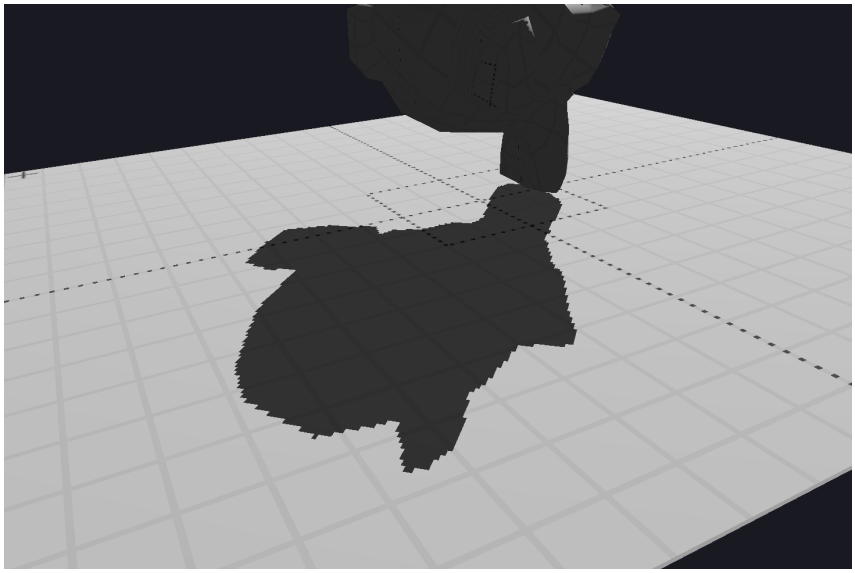


Figure 3.6: Aliased shadow appearance due to no filtering and low resolution of the shadow map.

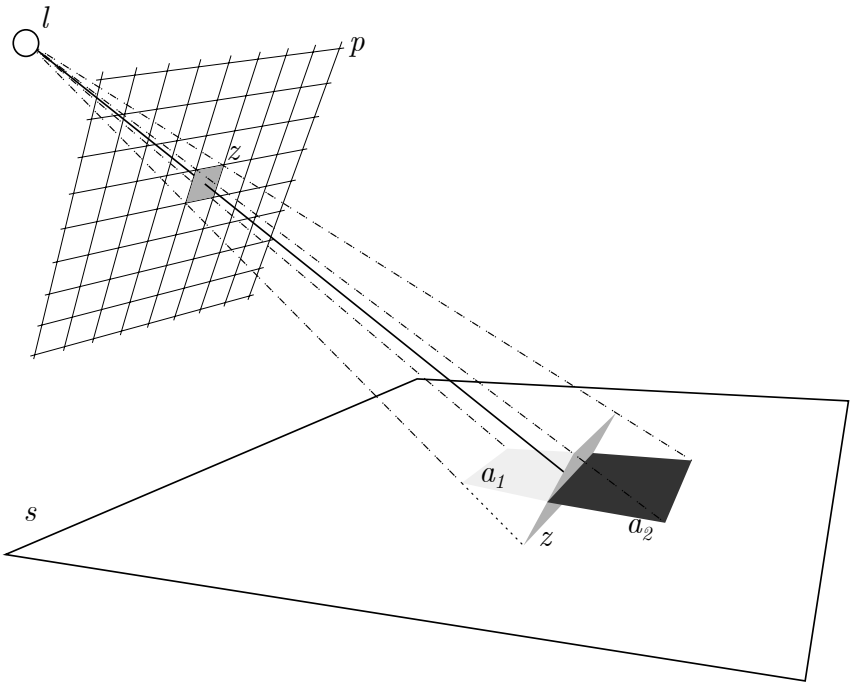
When rendering the shadow map an appropriate projection matrix has to be constructed. For a far away directional light, such as a sun light, an orthographic projection should be used. When rendering a spotlight a perspective projection can be used. Point lights can

also use a single shadow map with a perspective projection matrix, but only in situations when the light source is positioned outside the scene, or at least is not surrounded by shadow casters. In the opposite case, the most common technique is to use cubemaps and improved derivative techniques as described by Liang Wan [24]. A cubemap consists of six shadow map renders, each covering a face of a cube constructed around the light source. This way an omnidirectional shadow map is created.

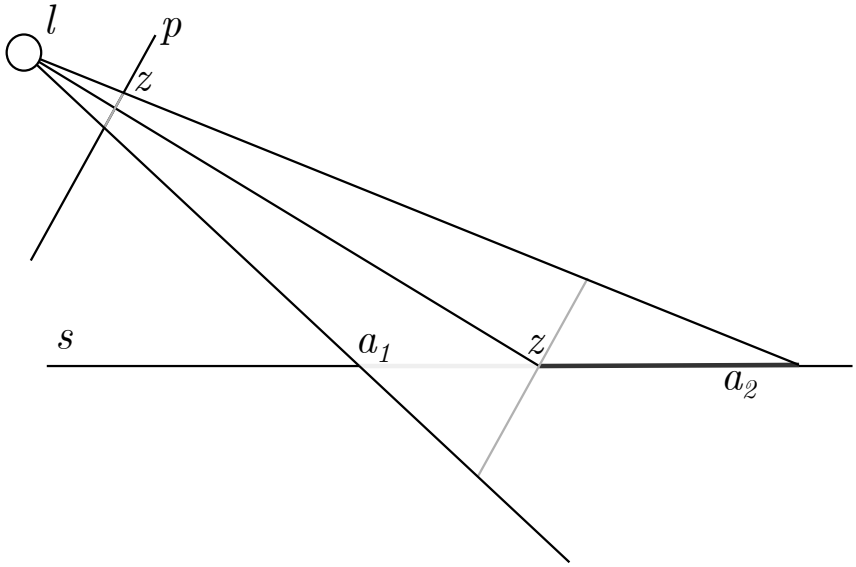
When creating a projection matrix as well as the view matrix for a shadow map, care has to be taken to capture all relevant shadow casters. This is especially true for large scale scenes and directional lights, as rendering the entirety of a scene into the shadow map would most of the time be unnecessary and suboptimal. Most of the rendered map would not be used in a given frame, and the distribution of shadow map resolution would be poor. It would be enough for the light view to tightly encompass the view frustum of the observer to obtain correct results. This is called fitting and will be presented in the following sections. This idea can be taken further as shown by Jiří Bittner [2] to allow for an approximate FPS gain of 1.5 times. Their method uses a custom advanced algorithm to cull shadow casters and render only those that contribute to the current main viewpoint.

Another issue of shadow maps is self-shadowing, or surface acne. This is also a result of aliasing, since the continuous scene geometry and its depth data is sampled in a discrete manner into an image when rendering the shadow map. If the shadow map texel does not correlate one-to-one with the final view samples, which happens very often and is highly likely, the shadow map samples will cover a non-zero area on the geometry surface in the final view of the scene, which is illustrated in Fig. 3.7. This can cause the surface to erroneously shadow itself, as shown in Fig. 3.8a. This issue gets more pronounced the higher the angle between the light source view ray and the surface normal. It can be resolved by adding a depth bias when rendering the shadow map, which will ensure the samples are below the actual surface. In modern rendering APIs the depth bias can be configured for a pipeline and applied automatically when rendering. Especially useful is the combination of a constant depth bias and a slope-scaled depth bias. The second bias helps eliminate the problem when surfaces appear more edge-on in the light view by scaling the depth bias with the angle between the light direction and surface. Too high depth bias can cause peter-panning, a situation shown in Fig. 3.8b where a shadow caster gets detached from its shadow. Because of this, depth bias values often need to be hand-tuned on a per-scene basis. Other approaches have also been proposed, such as using only the back faces of objects to render the shadow map [25] or utilizing depth peeling to store average depths between the back and front faces of an object [28], but they both still require some bias to avoid shadow-acne in areas such as silhouette edges or concavities.

The z-buffer of a rendered image can be actually non-linear in its $[0 - 1]$ range. Perspective projection matrices are sometimes built in such a way that the precision closer to 1 is lower than closer to 0. This is purposefully done to gain more precision closer to the

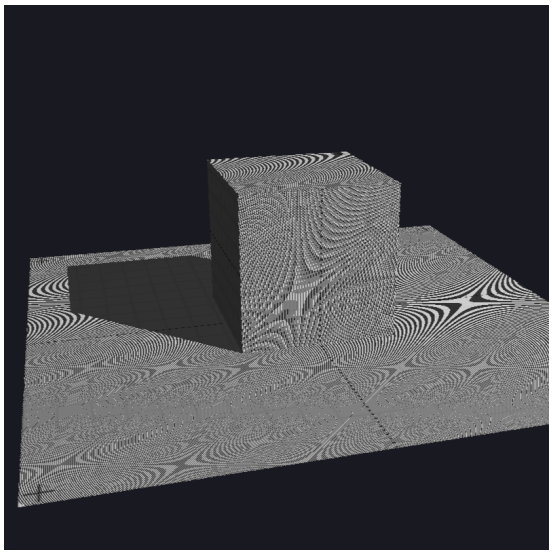


(a) Depth samples being compared with depth values over an area of the actual mesh lead to aliasing and shadow acne.

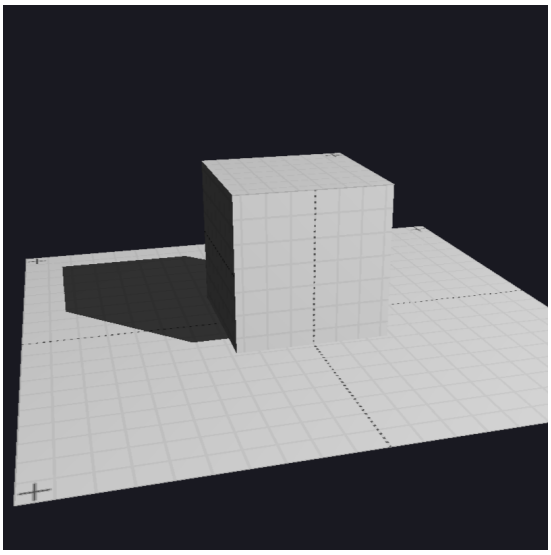


(b) The aliasing issue present in shadow maps illustrated in two dimensions from the side.

Figure 3.7: The cause of shadow acne. Light source l renders the shadow map onto its projection plane p, having some finite resolution. The sampled depth value z is stored in a texel. This creates a virtual surface orthogonal to the light direction on top of the actual geometry surface s. When rendering the scene, depth values in the area a_1 will have a lower depth value stored in the shadow map texel than their actual depth, causing them to be in light. Samples in the a_2 area will have greater depth, causing them to be evaluated as being in shadow.



(a) Shadow acne due to using not enough bias.



(b) The cube is disconnected from its shadow due to too much depth bias.

Figure 3.8: Some of the issues that can arise when using projection shadows.

viewer, where it is more needed when performing hidden-surface removal than far away, where objects will become very small due to perspective foreshortening. This can be an issue with shadow mapping, since objects close to the light source are not necessarily close to the viewer. In fact the opposite can often be true. As shown by Stefan Brabec [4], the z-buffer for the use in shadow mapping can be linearized using a linear transformation dependent on the projection matrix used. The following perspective projection matrix present in the DX Math library can be linearized with equation 3.7.

$$\begin{pmatrix} W & 0 & 0 & 0 \\ 0 & H & 0 & 0 \\ 0 & 0 & \frac{f}{f-n} & \frac{-fn}{f-n} \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \begin{pmatrix} Wx \\ Hy \\ \frac{zf}{f-n} - \frac{wfn}{f-n} \\ z \end{pmatrix} \Rightarrow \begin{pmatrix} Wx \\ Hy \\ \frac{zf}{f-n} - \frac{fn}{f-n} \\ z \end{pmatrix} \quad \text{for} \quad w = 1 \tag{3.5}$$

$$\begin{pmatrix} Wx \\ Hy \\ \frac{zf}{f-n} - \frac{fn}{f-n} \\ z \end{pmatrix} /z = \begin{pmatrix} \frac{Wx}{z} \\ \frac{Hy}{z} \\ \frac{f}{f-n} - \frac{fn}{z(f-n)} \\ 1 \end{pmatrix} \tag{3.6}$$

After the perspective division the final form of the z-value equation is obtained. Below, in Fig. 3.9, the z-values are plotted for exemplary near and far clipping planes equal $n = 2$, $f = 15$.

The values can be linearized back again simply by multiplying the clip-space z-value by $\frac{z}{f}$, which gives the following equation.

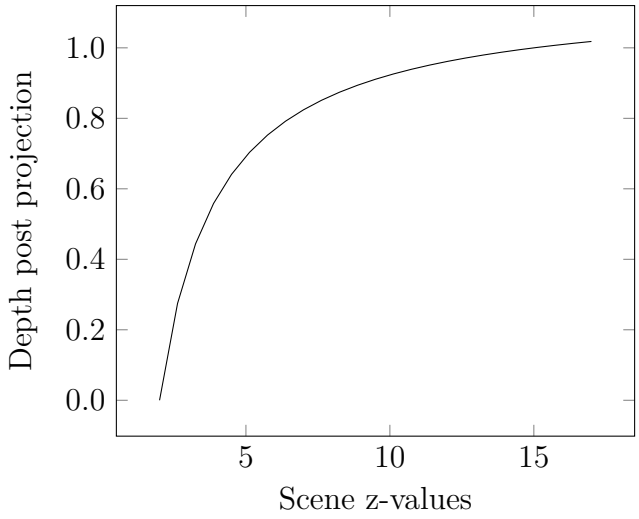


Figure 3.9: The z-values after a perspective projection become non-linear.

$$\left(\frac{f}{f-n} - \frac{fn}{z(f-n)}\right) * \frac{z}{f} = \frac{z-n}{f-n} \tag{3.7}$$

In the application z-buffer linearization was not implemented, as after investigation it turned out that an orthographic projection, which is used for the directional light, does not introduce nonlinearity to the z-buffer range. This can be seen by inspecting how the matrix will affect the w component. The orthographic projection matrix as implemented in the DX Math library is presented below.

$$\mathbf{M_o} = \begin{pmatrix} W & 0 & 0 & 0 \\ 0 & H & 0 & 0 \\ 0 & 0 & \frac{1}{f-n} & \frac{-n}{f-n} \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{3.8}$$

This matrix does not modify the w component and the z values will remain linear. In fact the formula for depth values after using this matrix is exactly the same as the perspective one after linearization.

3.2.3 Fitting shadow maps

An attempt to implement LiSPSM was made, but was not successful due to poor coverage of the implementation details both in the original paper and on the internet.

Only a part of the algorithm was implemented, basically providing simplified fitting of the shadow map view. The shadow map camera has certain extents set up by hand for each scene, such that the whole scene fits within view. Since the entire scene is not

necessarily viewed at once by the observer, only the part that is within the observer view frustum can be rendered to the shadow map. Therefore, the shadow map post-projection space can be adjusted to include the area viewed by the observer.

This can be achieved by building a fitting transformation matrix $\mathbf{F}$ that includes translation and scaling, that is applied after the light view and projection matrices, resulting in $\mathbf{F}\mathbf{P}_L\mathbf{V}_L$. To obtain $\mathbf{F}$, first the observer view frustum extents need to be found in world space, which can be obtained by applying the inverse of observer view and projection matrices to an NDC cuboid. This can be a cube with ranges $[-1:1]$ in x, y, z in APIs like OpenGL, or a cuboid with ranges $[-1:1]$ in x and y, and $[0:1]$ in z in DirectX. Once the vertices of this cuboid are moved into world space, they can be multiplied by the view and projection matrices of the light source, $\mathbf{P}_L\mathbf{V}_L$, and their components divided by w to bring them into the post-projection space of the light source. Once there, a bounding box can be found enclosing these points. To not move the contents of the shadow map beyond the original extent specified for the light source, the components of the vertices of this bounding box can be clamped to the NDC range. The whole process is shown in Fig. 3.10. Finally, the minimum and maximum x and y values of these bounding vertices

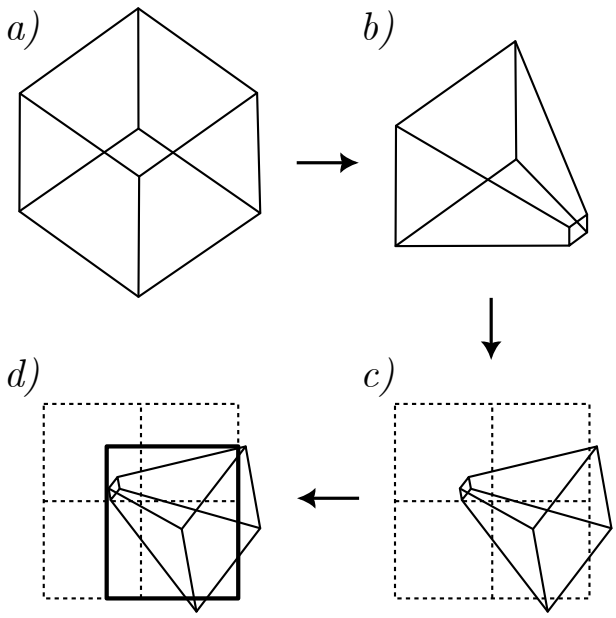


Figure 3.10: The steps involved in fitting the shadow map. The NDC cuboid in a) is in observer post-projection space, or clip space. It is moved to the state presented in b) using inverse matrices, resulting in a world space representation of the viewing frustum. It is them transformed into c), the light view clipping space. Finally, a clamped bounding box is found, shown in d).

can be found and used in the following fitting matrix.

$$\mathbf{F} = \begin{pmatrix} s_x & 0 & 0 & o_x \\ 0 & s_y & 0 & o_y \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \text{where} \tag{3.9}$$

$$s_x = \frac{2}{x_{max} - x_{min}} \quad \text{and} \quad o_x = \frac{-s_x(x_{max} + x_{min})}{2} \tag{3.10}$$

The calculation of s_y and o_y is analogous.

As was implemented, the fitting focuses on an approximate convex body of $\mathbf{L} \cap \mathbf{V}$, where $\mathbf{L}$ stands for the light view frustum and $\mathbf{V}$ for the viewer frustum. The shadow map is then fitted only in the clip space x and y dimensions, ignoring depth z. A more robust solution would be to find the convex hull of $\mathbf{L} \cap \mathbf{V} \cap \mathbf{S}$, with $\mathbf{S}$ meaning the scene bounding box. Then the shadow map would only focus on areas within the viewer frustum where there exists geometry to render. The near plane of the light frustum can then be set to the upper scene extent and the far plane to the farthest viewer frustum point. The shadow map would then require no hand adjustments to its frustum and would adapt to different scenes and viewer orientations.

3.2.4 Filtering shadow maps with PCF

As established in section 2.2.7, percentage-closer filtering performs filtering of the shadow information to avoid aliasing artifacts. Basic shadow mapping performs a depth comparison giving a binary result on whether a scene point is in shadow of some other occluder or is directly in view of the light source. This comparison can be described with a formula $s(t, z) = H(d(t) - z)$, where function $s(t, z)$ is the shadow function that takes texture coordinates t and the depth value z of a scene sample in light space. Function $d(t)$ is the depth lookup into the shadow map and H is the Heaviside function. This binary result can be obtained for positions near the initial scene sample and averaged, to get a value in the $[0 : 1]$ range describing how much light the point receives. This makes the technique relatively simple to implement in a renderer that already uses shadow mapping.

Possibly the simplest way to get PCF results is to use the hardware implementation made available by modern graphics APIs. This is accessible via the settings of the sampler that will be used with the shadow map. In DirectX 12 one can specify a linear comparison filter operation and a comparison function when creating a sampler. When later used in an HLSL shader, the SampleCmp function should be used. It takes the value to compare against as one of its arguments. It will perform comparison according to the comparison function with four nearest texels and bilinearly interpolate the results. The effect is shown in Fig. 3.11, in subfigure 3.11a. This is done at practically no additional performance cost.

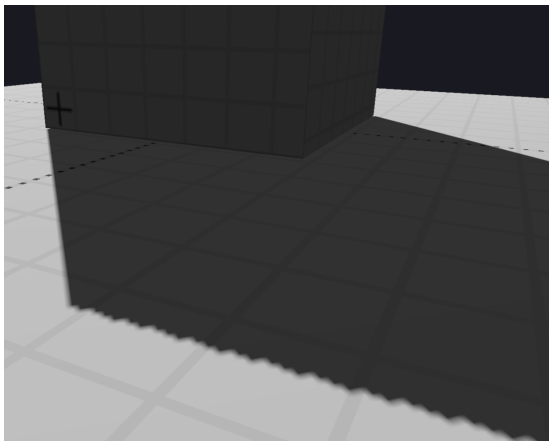
Unfortunately, hardware bilinear PCF is not going to significantly improve the shadow mapping visuals. In most cases there will still be visible aliasing, which is due to the fact that the kernel used is very small, being only 2×2. A larger kernel can be built manually and used in the same manner to get smoother shadows. A simple box filter with an $n \times n$ kernel can cover a larger area in the shadow map. The larger the kernel, the smoother the shadow and the larger the penumbra. This can be implemented by generating an array of offsets either in the shader or, to preserve computing power, on the CPU and later sent to the GPU via a uniform buffer. The offsets should be in texel-aligned increments, which is easily achieved by expressing them in multiples of $1/s$, where s is the shadow map size. These offsets are then added to the scene sample position expressed in light space. It is important to note, that even though different texels are sampled, the initial scene sample z value remains the same. Unfortunately, the visual quality of such filtering is visibly sub-par, as pixelated banding artifacts occur as shown in 3.11b. This is due to the fact that only $n^2 + 1$ shades can be obtained in the penumbra. The kernel size could obviously be raised, but sizes even as relatively low as 11 have a significant impact on performance. Hardware PCF can be used with this method to smooth out the banding as shown in Fig. 3.11c, and an offset scale can be utilized to control the size of the penumbra regardless of the size of the kernel.

A more robust and efficient approach allows using fewer samples while giving good visual results, sampling the shadow map adaptively. Instead of using a box filter it creates sets of randomly jittered offsets that are used when performing PCF. Most importantly, these offsets are stored in memory in such a way that allows to decide whether a point is fully lit or fully in shadow without having to test all samples. The offsets are first generated on the CPU in a jittered grid pattern of $n \times n$ size. Their coordinates are normalized, so all samples lie within the $[0 : 1]$ range in the x and y axes. Then, all the samples are moved from the grid into a disk domain using the following equations.

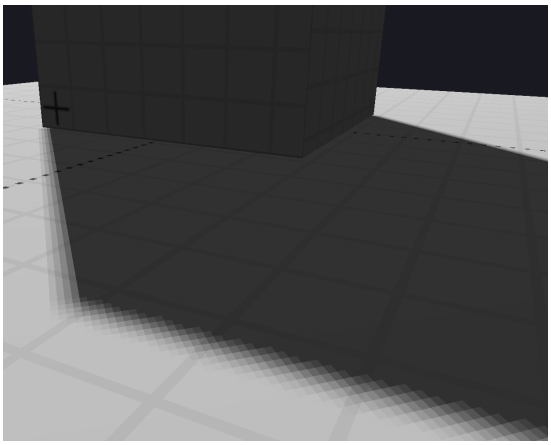
$$u = \sqrt{y} \times \cos(2\pi x) \tag{3.11}$$

$$v = \sqrt{y} \times \sin(2\pi x) \tag{3.12}$$

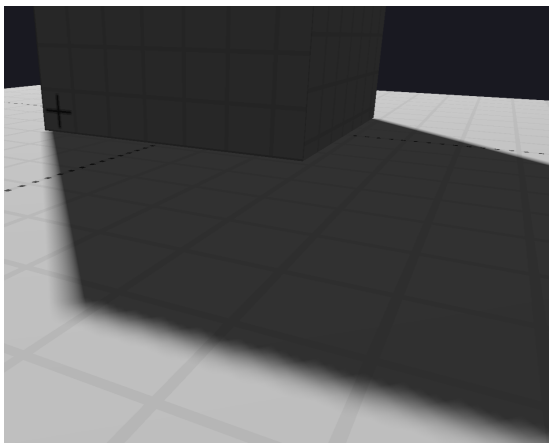
The angular position of the sample on the disk is then determined by x and the y defines the radial distance from the center. If the loop that generates these points and stores them into an array is constructed in a smart way, the highest y coordinates are generated first, so the offsets the furthest away from the center are first in the array. This is then utilized in shader code when looping over the sample offsets and performing PCF: only the first n samples are tested at first, and if they all result in either 0 or 1, meaning the point is entirely in shadow or in light, further PCF processing for this point can stop. This delivers a significant performance boost, saving processing power where computations do not contribute to the visual outcome. The full set of samples will only be used in penumbras.



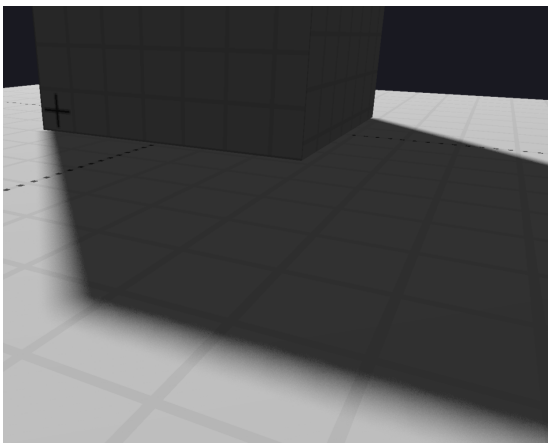
(a) Shadow mapping with hardware bilinear filtering.



(b) Shadow mapping with a manually applied 5×5 kernel and box filter.



(c) Shadow mapping with a manually applied 5×5 kernel and hardware bilinear filtering.



(d) Shadow mapping with 16×16 sets of 8×8 jittered offsets.

Figure 3.11: Different results obtained with different PCF algorithms used to filter a 512×512 shadow map.

The fact that all offsets have been normalized means that they can be multiplied by a scalar scale factor to control the amount of smoothing, or the size of the penumbra.

Unfortunately, the problem of banding in penumbra regions observed in simple box filtered PCF still persists. Hardware PCF can again be used to smooth the results further, but there is another solution to this problem. Instead of using a single set of jittered sample offsets, $m \times m$ such sets can be generated and stored in a three-dimensional texture. In the shader this texture is accessed using the screen pixel coordinates, to choose a set of offsets, and then the set is indexed to perform PCF. This way banding is stochastically smoothed out in the final result, as shown in Fig. 3.11d. Creating an offset set for each pixel of the final image would consume a lot of memory and not provide significant gains, so small sets, such as 16×16 are used and tiled over the whole render area. This is enough to generate sufficient noise in the penumbra areas to create a smooth appearance.

As an argument against this technique one could point out the fact that GPU hardware

does not handle dynamic branching well, and that it should be avoided to maintain the best performance. While this statement is generally true, it is not fully accurate. On modern GPUs the parallel threads that perform the computations run in groups, partially determined by their screen space positions, called warps or wavefronts. This is well explained and visualized by Nvidia in their developer article on the logical structure of the rendering hardware [10]. Since the branching in this algorithm depends on whether a pixel is in the penumbra region, it is dependent on screen space position and penumbra points will likely be in warps together, making the optimization effective.

A substantial problem of PCF algorithms is surface acne, which gets worse the larger the kernels and offset scales are used. This is because the sampled area of the shadow map increases and read depth values may vary more, while the depth of the scene point being shaded does not change. This can be again combated with a depth bias. An offset cone can be also used by applying higher bias the further from the original sample a point is. These however still need to be hand-tuned on a scene by scene basis. Additionally, it is important to note that, while PCF approaches do generate smooth shadows, they are not physically correct soft shadows. This is both due to the fact that the amount of smoothing is constant and that scene surface points are sampled. For an accurate shadow the area of the light source would need to be sampled for visibility from a given scene point, not the other way around.

3.2.5 Variance shadow maps

As introduced in section 2.2.7, variance shadow maps take a different, more robust approach to shadow mapping. Two moments of depth values, $M_1 = E(x)$ and $M_2 = E(x^2)$, being the depth and depth squared, possibly averaged over an area, are used to compute the mean and variance squared. Those are in turn used to calculate the upper bound that a given point is occluded using the Chebyshev inequality. The equations below show the calculation of the Chebyshev inequality p_{max}.

$$\mu = E(x) = M_1 \tag{3.13}$$

$$\sigma^2 = E(x^2) - E(x)^2 = M_2 - M_1^2 \tag{3.14}$$

$$P(x \geq z) \leq p_{max}(z) = \frac{\sigma^2}{\sigma^2 + (z - \mu)^2} \tag{3.15}$$

The value of p_{max} can be directly used in a shader as determining the percentage by which a point is not in shadow.

To actually render a VSM the regular shadow mapping algorithm requires a few modifications. The shadow pass will now have to use a color render target, as each moment will be stored in a single channel of the shadow map. It is best to use 32-bit per channel, two-channel format to avoid numeric stability problems. This shadow map texture can

then be used in a modified main-pass fragment shader to perform the calculations from 3.13. Additionally, it is good practice to clamp the minimum variance to some very small floating-point number, for example 0.00002. This again serves to avoid numeric stability issues and can also help with shadow acne and self-shadowing.

Another method to avoid erroneous self shadowing is to use the second moment of the VSM to introduce dynamic biasing using the following formula.

$$M_2 = \mu^2 + \frac{1}{4}\left(\left(\frac{\partial f}{\partial x}\right)^2 + \left(\frac{\partial f}{\partial y}\right)^2\right) \tag{3.16}$$

The partial derivative values can be easily obtained in-shader using built in partial derivative keywords. This addition basically removes the need for additional biasing, the hand tuning of which was a major issue for techniques utilizing regular shadow maps.

Variance shadow maps like this would be usable, but could suffer from light leaking near object edges, where the variance is too large and causes pixels that should be in shadow to receive light. This situation can be improved with another modification, by simply remapping the obtained p_{max} to a range $[a : 1]$, where a is some hand-picked value that suits the needs of a particular scene.

The most interesting difference of VSM for a graphics programmer as compared to shadow mapping is the fact that the depth moments can be filtered like regular color data. This fact allows to utilize the entire arsenal of techniques available for dealing with aliasing problems in regular color textures like mipmapping, anisotropic filtering and even ordinary blurring. When a separable blur is chosen, like a Gaussian blur or a simple box filter, the operation can be performed very efficiently by blurring the VSM first horizontally and then vertically. For this another temporary texture has to be used, of same size and format as the shadow map, thus increasing the memory footprint of the algorithm. This cost however is comparatively not great on modern hardware and in most cases is worth paying for the performance gain. Filtering with a PCF filter can take $M^2 \times N^2$ operations, where M is the texture size and N the filter size. A separable filter brings this cost down to $2 \times N \times M^2$. To improve the shadow appearance even further, a VSM can be rendered using multisample anti-aliasing.

To further improve the filtering of VSM, the usage of summed-area tables is proposed [17]. They allow the sum of values in any rectangular area in the shadow map to be obtained in constant time with only four texture lookups. This means that filtering with kernels of arbitrary sizes can be carried out in a time less than even a 3×3 PCF operation. This is a very significant speedup and can be used to generate soft shadows by dynamically querying information filtered by different kernel sizes that depend on the contents of the scene.

3.2.6 Implementing PCSS

Percentage-closer soft shadows, as described in section 2.2.8, utilize the PCF technique and the fact that the shadows created with it can have varying levels of smoothness. The smoothness, or the size of the penumbra can be controlled by the kernel size and the scale of the sample offsets. When correctly driven, varying the size of the penumbra on a per-pixel basis can create plausible soft shadows that harden when close to the occluder. To achieve that, the estimated size of the penumbra has to be computed based on information about occluders and the size of the light source.

The algorithm first gathers information on the occluders that block the light source from the point of view of a point in the scene that is being shaded. For this a search radius is calculated from the following formula by projecting the base of a vision cone of the shaded point of the light source onto the shadow map.

$$s_{lv} = \frac{s_{lw}}{s_{lf}} \tag{3.17}$$

$$r = s_{lv} \cdot \frac{p_{vz} - n}{p_{vz}} \tag{3.18}$$

In equation 3.17 s_{lv} is the size of the light source in light view space, s_{lw} is the size in world space and s_{lf} is the size of the light frustum. A square frustum is assumed, of size $s_{lf} \times s_{lf}$. The sampling radius r is then calculated using the n - near clipping plane and p_{vz} being the z coordinate of the scene point in light view space, as illustrated in Fig. 3.12. This search radius is then used to sample the shadow map in the area defined by r around the coordinates of the initial sample position in light space. Sample offsets can be

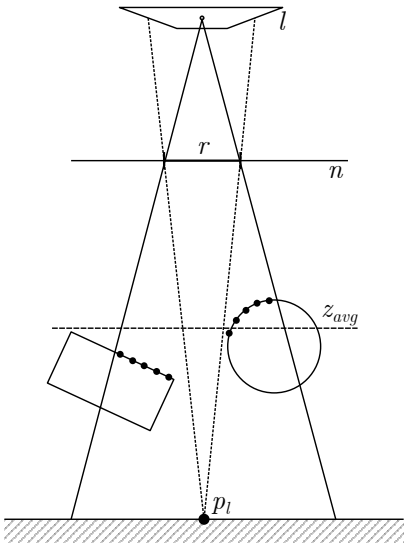


Figure 3.12: The scene point in light view space p_l and its view cone marked in dashed lines project the area of the light source l onto the light frustum near plane n. This creates the search radius r where depth samples are taken from the depth map, marked with black dots. These are averaged to obtain the average occluder depth z_{avg}.

obtained in any way, as long as they are in the range $[0:1]$ and can be scaled with r, for example using the sample offsets generated for adaptive PCF. The obtained depth values are averaged together and produce the average blocker depth z_{avg}. Once this is known, the following formulas can be used to calculate the penumbra width, and from that the filter size.

$$s_p = \frac{p_{vz} - z_{avg}}{z_{avg}} s_{lv} \tag{3.19}$$

$$s_f = \frac{n}{p_{vz}} s_p \tag{3.20}$$

Equation 3.19 calculates the penumbra size s_p and is based on similar triangles. Then the filter size s_f can be derived. This can be directly used for performing PCF.

The size of the penumbras is not dependent on the resolution of the shadow map thanks to equation 3.17. PCSS however does have its problems. It is not practical to perform the occluder search exhaustively, so the number of samples needs to be carefully chosen to avoid undersampling and give meaningful estimates. The amount of depth map texture lookups is even higher than with PCF, which puts a strain on texture memory bandwidth. The contact-hardening shadows that it produces are not physically correct. This is in part due to the fact that the algorithm uses the average depth of occluders and assumes that the average occluder, light and receiver are planar and parallel. Additionally, shadow map samples taken for averaging often stretch beyond the shadow map area containing the actual occluders.

Chapter 4

Experiments

4.1 Methodology

One of the objectives of the thesis was to compare multiple shadow rendering methods both in a qualitative and quantitative ways. To achieve this a test renderer application was created, which implements the rendering techniques and allows a user to observe their results as well as profile the performance the program.

The tests were performed in a semi-controlled environment. Care was taken to avoid other processes interrupting and affecting the test results, but their impact could not be entirely eliminated in. Results of repeated tests at different moments in time or on different machines with the same hardware will differ, but that difference should not be large enough to impact the overall comparison results. The machine used for performing measurements had the following specification:

- OS: Windows 10, 22H2
- CPU: AMD Ryzen 5 1600
- GPU: Nvidia GTX 1080, 8 GB VRAM
- RAM: 16 GB

4.1.1 Created test application

The test application allows the user to choose a rendering mode and observe the results in real-time. It also allows to change the viewpoint and observe how the shadows behave in motion. Most rendering modes implement different shadow rendering techniques and some include debug views like wireframe mode or mesh colored by normals. The application is also instrumented with profiling commands which make it possible to connect to an external profiler and collect data in real-time, as well as save and load existing profiling data.

As mentioned in section 3.1, the Tracy profiler open source library and tool was used to gather and visualize profiling data. For additional profiling and debugging RenderDoc, PIX and Nsight Graphics were used.

4.2 Data sets

The data sets consist of multiple scenes used for testing. Three scenes, Crytek Sponza, Power Plant and Chinese Dragon are from the Computer Graphics Archive [16]. The Crytek Sponza scene contains 262267 triangles and 184330 vertices. The Power Plant scene contains 12759246 triangles and 10614919 vertices. The Chinese Dragon contains 871306 triangles and 438929 vertices. They were chosen because they are standard test scenes in the computer graphics community, which are freely available on the internet. They also contain highly varying amounts of vertices, with the Power Plant scene having almost sixty times more vertices than Cytek Sponza, which allows to test the influence of geometric complexity on the performance of the algorithms. The Chinese Dragon scene is much more contained and showcases small-scale shadows and self-shadowing. Additionally, a simple scene containing a unit cube positioned on a ground plane was created in Blender. This scene contains only twenty-eight vertices. Its simplicity allows to easily spot differences in the appearance of shadows in the render and highlights issues such as peter-panning.

4.3 Results

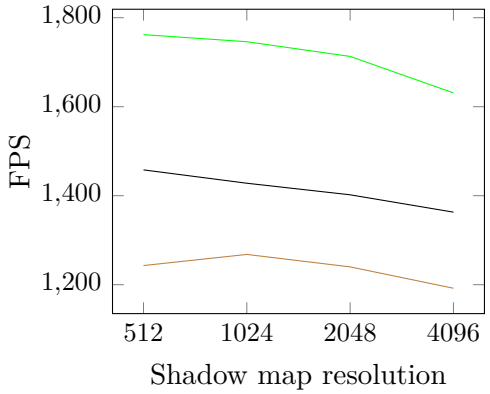
The results of testing each implemented shadow rendering method are presented in the following sections. Each set of results is complemented by a description and discussion.

4.3.1 Basic shadow maps

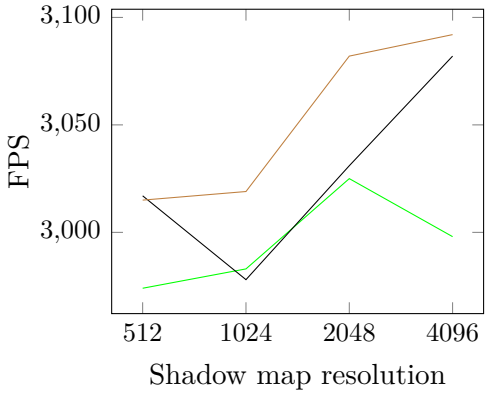
The basic implementation of shadow mapping from section 3.2.2 was tested with all four scenes at different shadow map and output resolutions. The results are shown in Fig. 4.1.

The test results mostly match the expectation that fewer FPS will be produced with increasing shadow map resolution and output resolution. The only exception is plot 4.1b of the Cube scene results. This is however probably due to the fact that this scene is so simple, and in effect is rendered with such high FPS (highest in this test set, over 3000), that any interruptions from other programs on the machine will be visible. Inconsistencies might also be produced by hardware cores' dynamic clock rates, which can vary over the duration of the test.

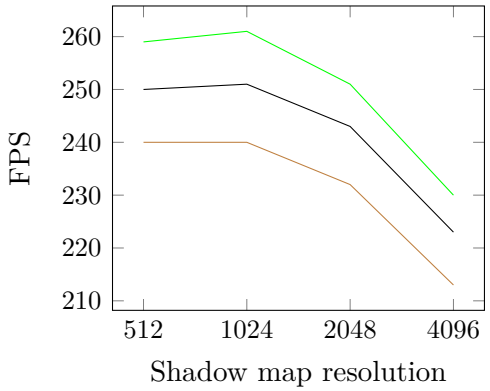
The basic shadow mapping algorithm is simple enough to produce high FPS. Even in the most complex scene, the Power Plant, over 210 frames are rendered at 2k output and



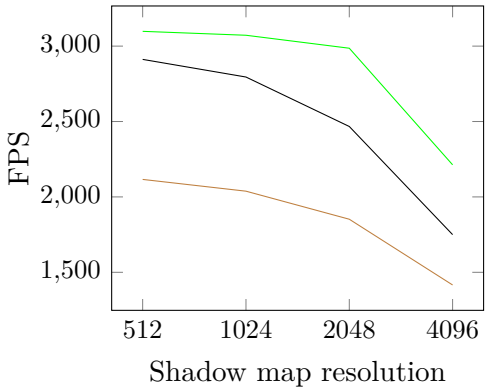
(a) Results for Chinese Dragon scene.



(b) Results for the Cube scene.



(c) Results for the Power Plant scene.



(d) Results for the Crytek Sponza scene.

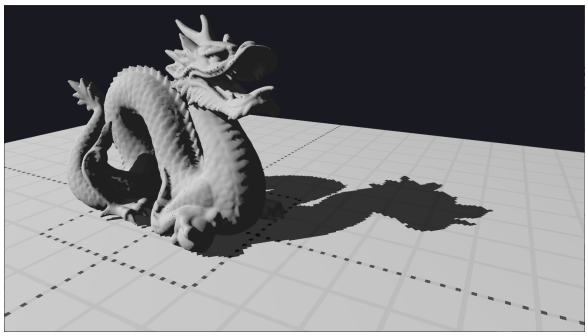
Figure 4.1: Frames per second for all test scenes, for different sizes of the shadow map and output resolutions. In green 1280×720, in black 1920×1080 and in brown 2560×1440. Rendering with the basic shadow map implementation.

4k shadow map resolution.

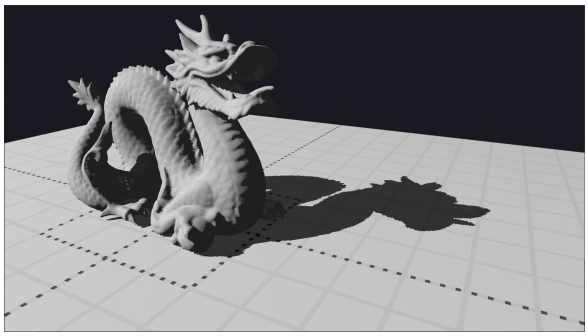
Looking at the shapes of the plots, it can be said that the basic shadow mapping algorithm is not dependent on output resolution. The lower FPS for higher output resolutions stem mostly from the sole fact that more pixel need to be rendered, because the FPS falloffs along the x-axis follow the same shape. In most cases the drop in FPS between the lowest and highest shadow map resolution is the same, regardless of the output resolution.

The plots have their x-axis in logarithmic scale. Taking that into account, it can be observed that the FPS counts fall linearly with growing shadow map resolutions.

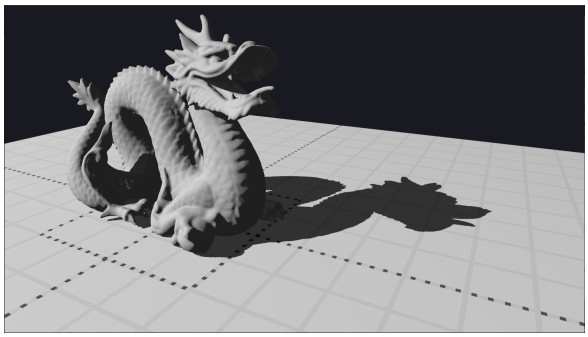
The algorithm's performance is not view-dependent. When traversing the scene the frame rates are stable. This is backed by the fact that in basic shadow mapping the same work is performed for each rendered pixel, consisting of a comparison operation with the depth value stored in the depth map. This however would not be true if shadow map fitting was used. In such case, the performance will become view dependent as more or less geometry will be rendered into the shadow map. This can provide a useful performance boost in large scenes.



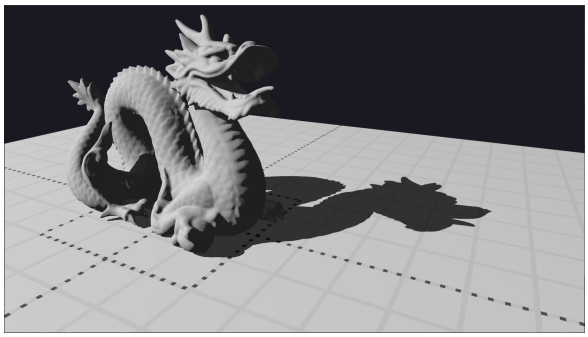
(a) The Chinese Dragon rendered with 512 × 512 shadow map.



(b) The Chinese Dragon rendered with 1024 × 1024 shadow map.



(c) The Chinese Dragon rendered with 2048 × 2048 shadow map.



(d) The Chinese Dragon rendered with 4096 × 4096 shadow map.

Figure 4.2: The Chinese Dragon scene rendered with different shadow map resolutions, using the basic shadow mapping algorithm.

Using the Tracy profiler it can be confirmed that the renderer's performance is limited by the computations performed on the GPU, as the CPU thread spent approximately 30% of time waiting for the GPU resources to be available. Tracy also reports that approximately 40% of the frame time was spent on the shadow map render pass and 52% on the main render pass. The remaining time was spent rendering the GUI and synchronizing. This can be expected as with this simple technique the shadow map pass is almost as expensive as the main pass. The main pass is slightly more involved as it outputs pixels and performs simple shading, as well as performs the shadow mapping itself, which requires texture lookups and comparisons that add time.

The rendering results for the Chinese Dragon are presented in Fig. 4.2 and in Fig. 4.3 the Crytek Sponza is presented.

The Chinese Dragon results are quite good for higher resolution shadow maps, since the shadow map texels map many to one with regard to screen pixels. Aliasing is however very visible for lower resolutions. Crytek Sponza suffers from shadow acne on vertical surfaces. This is a sign of badly chosen shadow bias. It gets less visible as resolution rises, but is always present. Even at the highest resolution, shadow map texel borders are still visible, where they were not visible in the Chinese Dragon scene. This is due to the fact that the shadow map covers a larger area and each shadow map texel covers more space

(a) The Crytek Sponza rendered with 512 × 512 shadow map.

(b) The Crytek Sponza rendered with 1024 × 1024 shadow map.

(c) The Crytek Sponza rendered with 2048 × 2048 shadow map.

(d) The Crytek Sponza rendered with 4096 × 4096 shadow map.

Figure 4.3: The Crytek Sponza scene rendered with different shadow map resolutions, using the basic shadow mapping algorithm.

on the screen.

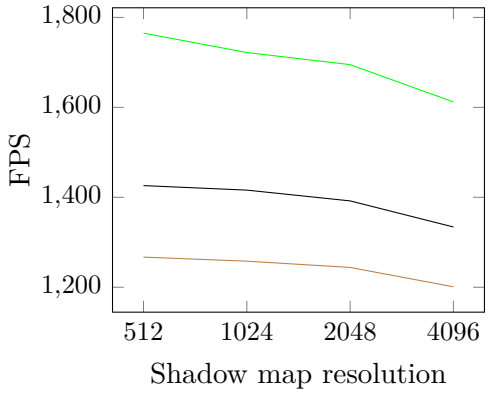
The memory usage is wholly dependent on the resolution of the shadow map. If memory is very scarce a depth buffer of lower precision can be used, but that is unlikely on modern graphics hardware to be the case. The depth map in this case was created to store the depth in 24 bits, plus additional 8 bits for the stencil buffer, totaling 32 bits per shadow map texel.

4.3.2 Filtered shadow maps

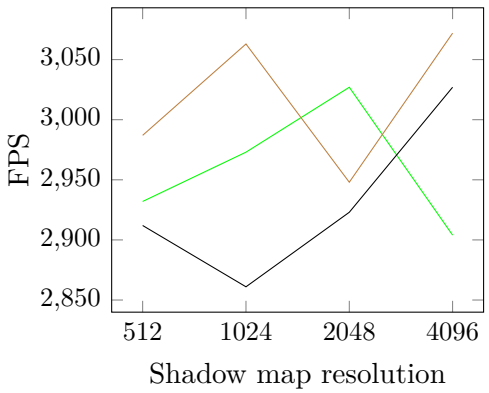
The following sections present experiments and their results performed using the implementations of different filtering techniques for shadow maps, which were described in sections 2.2.7 and 3.2.4.

Hardware bilinear filtering

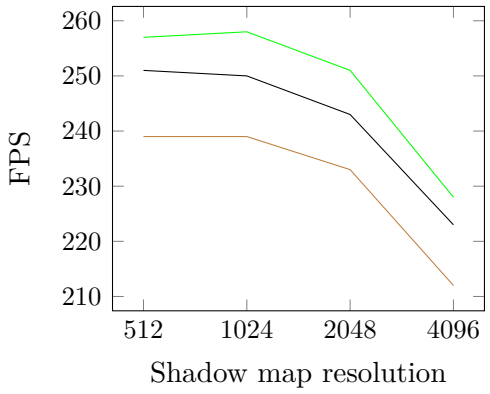
Hardware bilinear filtering is considered a simple filtering method that can improve appearance of shadows generated with shadow maps with practically no performance cost. This is tested for different output and shadow map resolutions in different test scenes and the results are presented in Fig. 4.4.



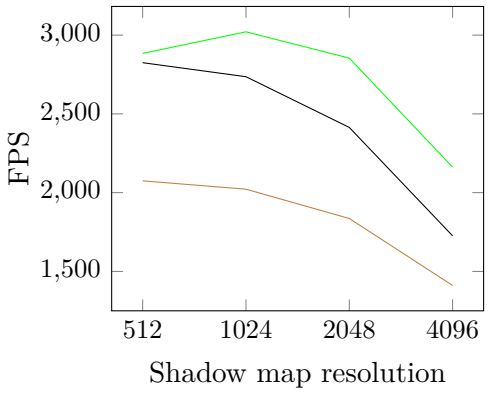
(a) Results for Chinese Dragon scene.



(b) Results for the Cube scene.



(c) Results for the Power Plant scene.



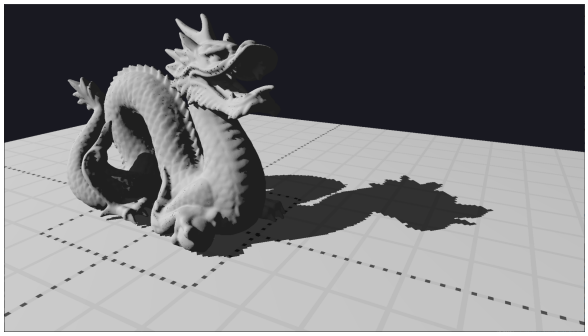
(d) Results for the Crytek Sponza scene.

Figure 4.4: Frames per second for all test scenes, for different sizes of the shadow map and output resolutions. In green 1280×720, in black 1920×1080 and in brown 2560×1440. Rendering with the basic shadow mapping technique with bilinear filtering enabled.

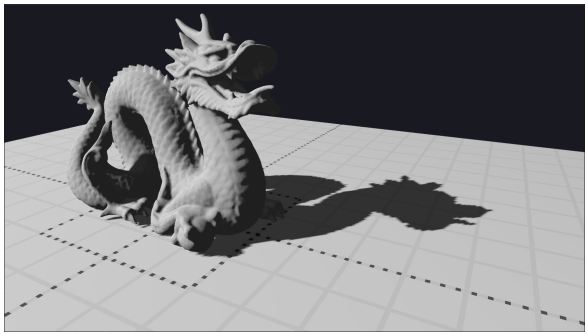
The results are very similar, with no immediately visible drop in performance. Taking the Power Plant scene as an example, the FPS range measured at Full HD output resolution is [251 : 223]. In the shadow mapping without filtering test the FPS range was [250 : 223]. The results are the same when accounting for instabilities caused by other software and dynamic clock rates. This is interesting when compared to results of Crytek Sponza tests, where the Full HD range was [2825 : 1726] with bilinear filtering and [2912 : 1750] without. Here, a significant drop in performance can be seen when using bilinear filtering. This might be caused by the fact FPS measurements are more sensitive in higher rages, as 1 FPS difference at 250 FPS is equal to 0.016ms, while 1 FPS difference at 2800 FPS is equal to only 0,000127ms. This is somewhat visible in the Crytek Sponza FPS ranges, as the upper ends have a higher difference between them than lower ends. Additionally, since the Power Plant is a much heavier scene performance-wise, the GPU has much more work to perform and the slowdown from using bilinear filtering can get masked better. When toggling bilinear filtering on and off during testing, the FPS readout would most of the time stay stable within reasonable FPS ranges. It seems that the cost

of hardware bilinear filtering is indeed insignificant in realistic scenarios, where the GPU is fully utilized.

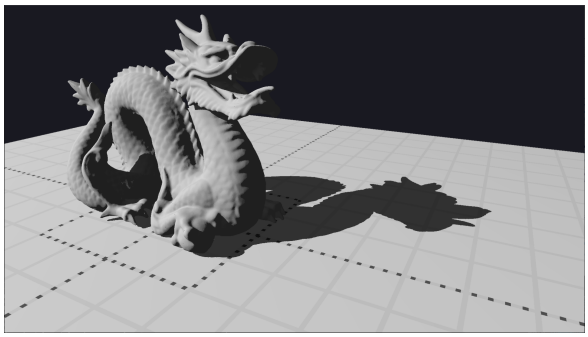
The effect on the final rendered result is on the other hand very significant, especially at lower shadow map resolutions, as presented in figure 4.5. The shadow boundaries get smoother, which improves the visuals even at low shadow map resolutions. The texels are less noticeable and the shape is more easily readable as a shadow.



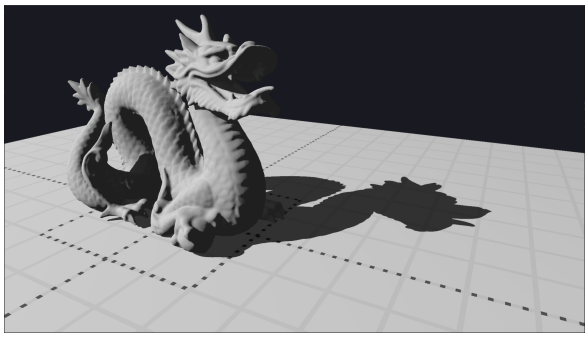
(a) The Chinese Dragon rendered with a 512× 512 shadow map with no filtering.



(b) The Chinese Dragon rendered with a 512× 512 shadow map with bilinear filtering.



(c) The Chinese Dragon rendered with a 4096 × 4096 shadow map with no filtering.



(d) The Chinese Dragon rendered with a 4096 × 4096 shadow map with bilinear filtering.

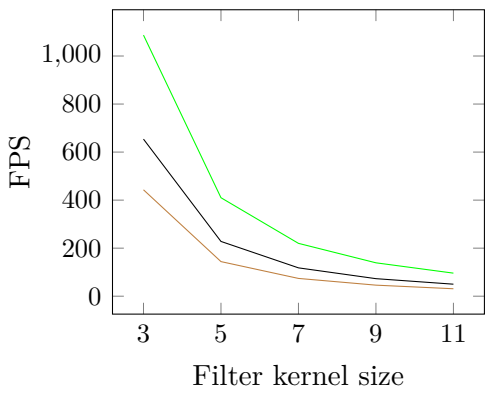
Figure 4.5: The Chinese Dragon rendered with and without bilinear comparison filtering.

Unfortunately, this filtering means that biasing has to be adjusted, as samples will be taken around the original position. This results in visible shadow acne, especially at lower shadow map resolutions, where shadow map texels cover a larger area on screen. It's noteworthy that, while there will be more pixels that erroneously self-shadow, the contrast of such artifacts will be smaller. This can help hide them at higher shadow map resolutions.

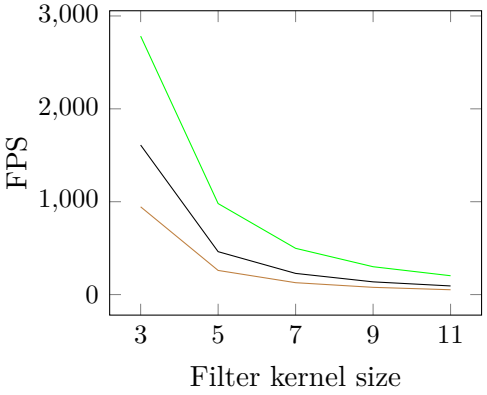
Percentage-closer filtering

PCF is another filtering technique that allows for the creation of arbitrarily sized filter kernels used to filter the results of shadow computations. For the performance tests, the shadow map resolution will be kept constant to measure how kernel size and output resolution affect the FPS counts. Shadow map resolution does not impact the results in

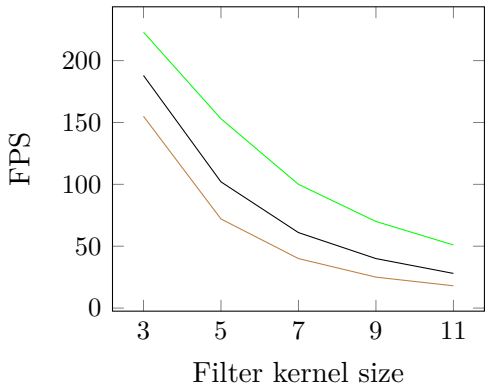
the context of this technique, apart from the performance degradation with higher shadow map resolutions observed in previous tests. The results for all scenes, for different square filter kernel sizes and output resolutions are presented in Fig. 4.6.



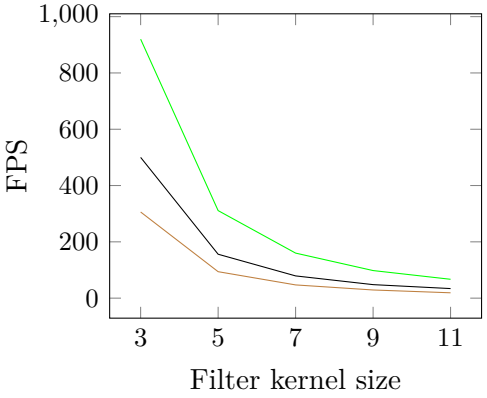
(a) Results for Chinese Dragon scene.



(b) Results for the Cube scene.



(c) Results for the Power Plant scene.

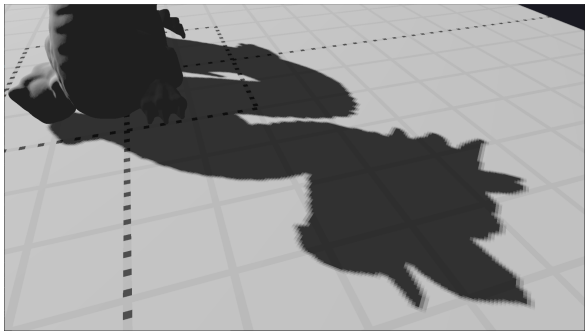


(d) Results for the Crytek Sponza scene.

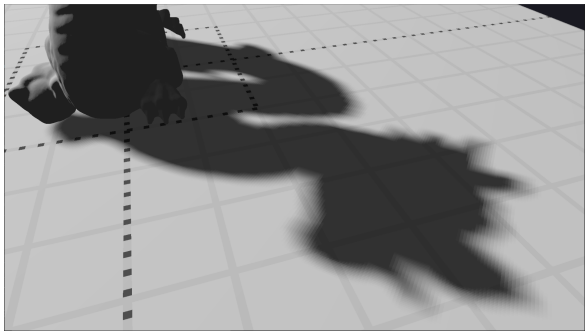
Figure 4.6: Frames per second for all test scenes, for different sizes of the filter kernel and output resolutions. In green 1280×720, in black 1920×1080 and in brown 2560×1440. Rendering with the basic shadow mapping technique with PCF at constant shadow map size 1024×1024.

The results show performance dropping exponentially with regard to increasing filter kernel size. This is expected, as for a kernel of size n there are n^2 depth comparisons performed. This fact makes high kernel sizes impractical and limits the usability of the approach. A larger smoothing effect is better achieved by using a scale factor to scale the sample offsets within the filter kernel than by increasing the kernel size.

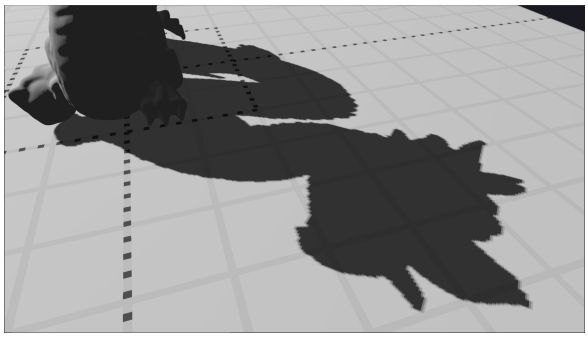
During testing interesting behavior was observed. In all tests, when the application was running with its initial kernel parameter unchanged, the frame rate would be much higher than at the same kernel size after any change to it was made. This difference in some cases was as large as the entire FPS value after a change, meaning the loss of half the frame rate. The reason for this is not known, because default initialization and optimization of the code path is unlikely, as neither the shader code nor the compiler



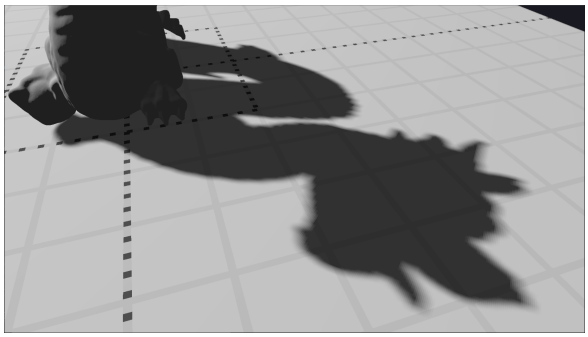
(a) The shadow of the Chinese Dragon filtered with a 3×3 kernel.



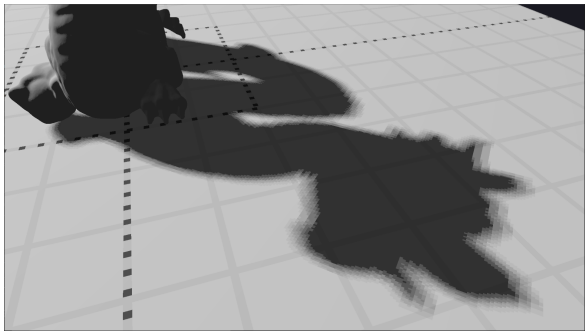
(b) The shadow of the Chinese Dragon filtered with an 11×11 kernel.



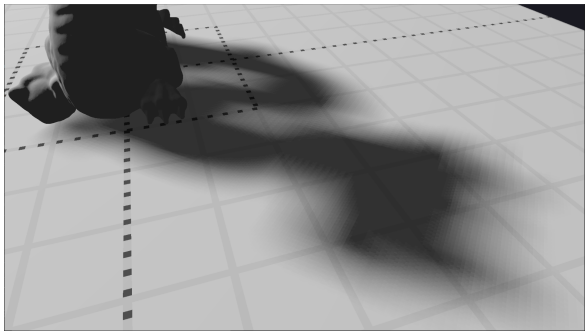
(c) The shadow of the Chinese Dragon filtered with a 3×3 kernel and a scale of 0.5 applied to each kernel offset.



(d) The shadow of the Chinese Dragon filtered with an 11×11 kernel and a scale of 0.5 applied to each kernel offset.



(e) The shadow of the Chinese Dragon filtered with a 3×3 kernel and a scale of 3 applied to each kernel offset.



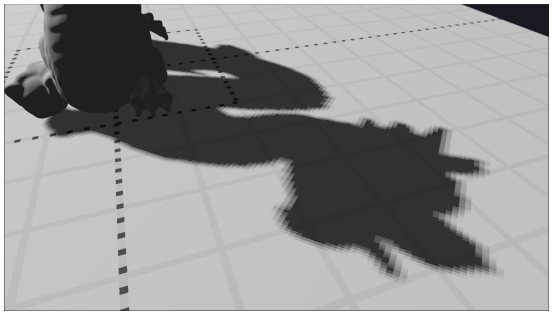
(f) The shadow of the Chinese Dragon filtered with an 11×11 kernel and a scale of 3 applied to each kernel offset.

Figure 4.7: The Chinese Dragon rendered with 3×3 and 11×11 kernels with different offset scales.

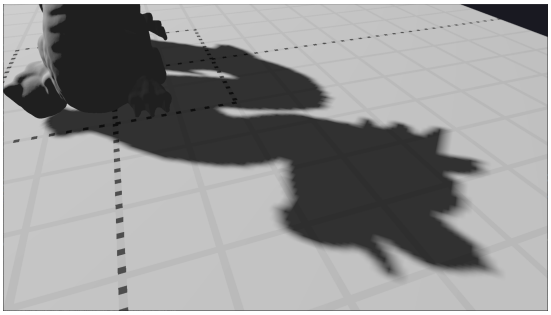
know of the initial settings that will be used for the kernel size. All results shown in Fig. 4.6 present FPS values after the initial change was made.

The appearance of shadows rendered with different sizes of kernels and offsets is presented in Fig. 4.7. Banding is very much visible in the smoothed areas of the shadows, even with the largest kernel. This is because of the limited range of shades that can be produced in these transitional areas. A transition that appears smoother can be achieved by setting a smaller than 1 offset scale, but that limits the size of the smoothed shadow areas. Additionally, texel boundaries are always visible with this method.

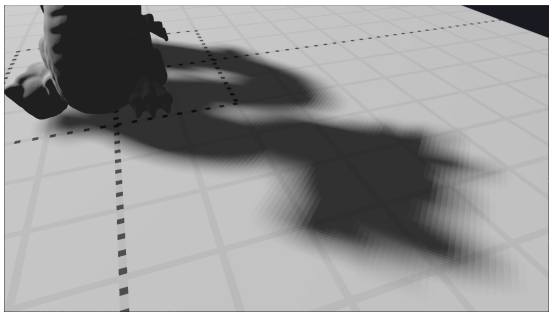
The usage of even the smallest 3×3 kernel requires adjusting the depth bias to get clean results. The larger the kernel, the more bias is needed. This could be partially alleviated by using the distance from the kernel center to scale a bias that is added before the comparison happens, but overall bias would still depend on receiver orientations and would need to be adjusted per-scene.



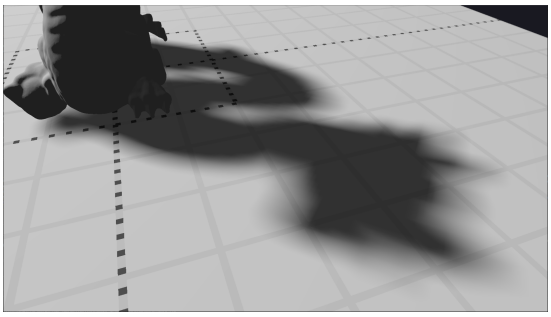
(a) The Chinese Dragon rendered with basic 3×3 PCF.



(b) The Chinese Dragon rendered with 3×3 PCF enhanced by bilinear filtering.



(c) The Chinese Dragon rendered with basic 11×11 PCF.



(d) The Chinese Dragon rendered with 11×11 PCF enhanced by bilinear filtering.

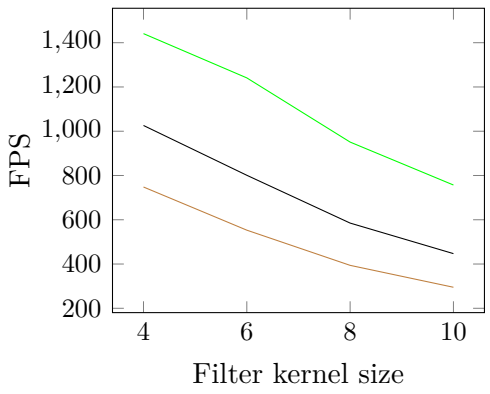
Figure 4.8: The Chinese Dragon rendered with PCF and with and without additional bilinear comparison filtering, using a 512×512 shadow map resolution.

The performance cost and underwhelming results make this technique mainly a stepping stone for more robust approaches. It is not practical to use it in real applications that need to run in real-time. There is however one adjustment that can be made to significantly improve the shadow's appearance produced by PCF. As established in section 4.3.2, hardware bilinear comparison filtering comes at virtually no performance cost. It can be applied to PCF, the results of which are shown in Fig. 4.8

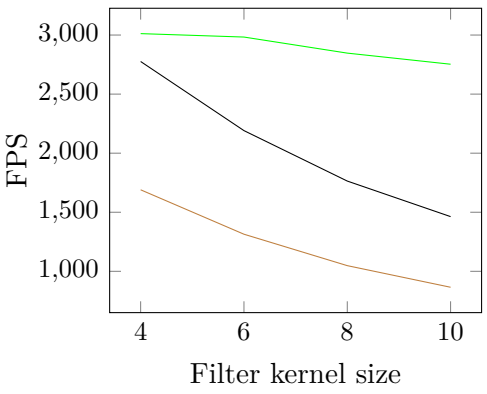
These results show that bilinear comparison filtering combined with PCF can lead to a great improvement of the visuals, even at kernel sizes as small as 3×3. Additionally, the shadow map used there has only 512×512 resolution, making it a very performant option. The impractical kernel size 11×11, when combined with bilinear filtering gives very smooth results, without any banding, hiding the texel boundaries very well, especially considering the low resolution of the shadow map.

Adaptive percentage-closer filtering

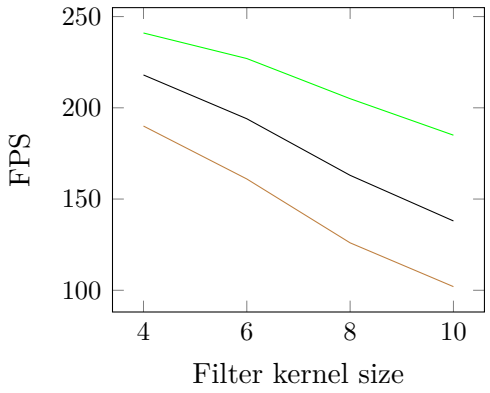
Adaptive PCF aims to improve the performance of regular PCF by allowing only some of the kernel samples to be processed when the shaded pixel in not in a penumbra region. It also improves the control over the penumbra size while maintaining visual quality. This is tested in this section, beginning with the results of performance testing presented in Fig. 4.9.



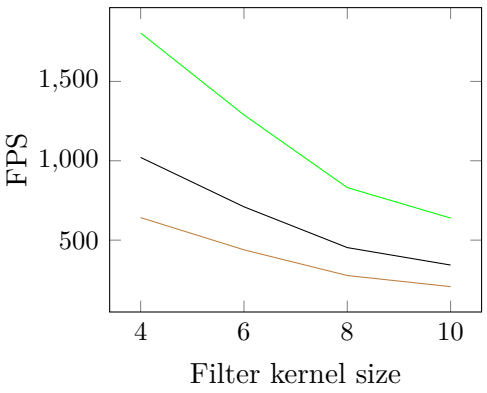
(a) Results for Chinese Dragon scene.



(b) Results for the Cube scene.



(c) Results for the Power Plant scene.



(d) Results for the Crytek Sponza scene.

Figure 4.9: Frames per second for all test scenes, for different sizes of the filter kernel and output resolutions. In green 1280×720, in black 1920×1080 and in brown 2560×1440. Rendering with the basic shadow mapping technique with adaptive PCF at constant shadow map size 1024×1024.

The produced graphs show that the technique does indeed improve performance in a

significant way. Not only are the FPS higher, giving for the smallest kernel sizes 15% up to 104% more FPS depending on the scene, but the frame rate also falls off less rapidly with increasing kernel size, giving between 392% up to 1473% more FPS for the largest kernel sizes. This is a very meaningful performance boost, making this technique capable of real-time performance. This is owed to the fact that small kernels are used to determine if a position is in the penumbra region which needs to be smoothed, and filtering stops if it is not.

Because this optimization technique is adaptive, its performance is view dependent. More kernel samples will be used if the view contains large penumbra regions than when it only contains fully lit or fully shadowed surfaces. The FPS difference can be even that of 50%. It can also vary depending on the scene for the same reason, so actual performance gains will vary. Performance will also be directly dependent on the sample offset scale, since with higher offsets larger penumbras are produced. This technique however will always give better or equal performance as basic PCF of the same kernel size.

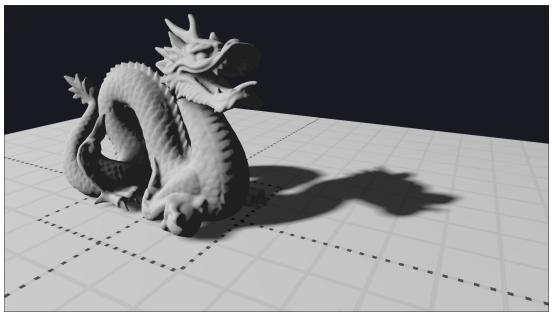
As observed in previous tests, the bias has to be adjusted carefully when using filtering. In the case of adaptive PCF, a bad bias will not only result in degradation of the produced visuals but also degradation of performance, as more surface points will utilize the full kernel size when in the penumbra of an erroneous shadow.

Adaptive PCF in this implementation uses $m \times m$ sets of randomized kernels. A specific kernel is chosen per-pixel based on screen coordinates. This is to minimize banding by introducing noise into penumbra areas, which is treated by the eye as a smoothed shadow boundary. Figure 4.10 presents results rendered with different amounts of kernels for a large offset scale.

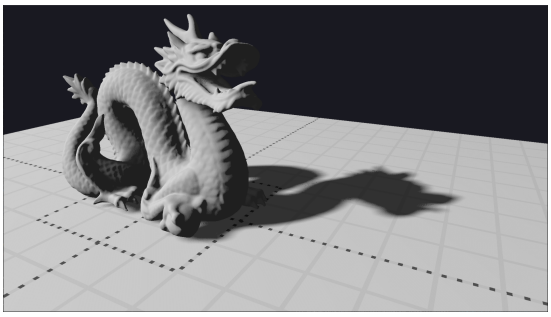
Testing shows no real performance cost of increasing the filter sets count in this range. The kernels are stored in a 3D texture of dimensions $m \times m \times (n^2/2)$, which is itself generated once on application startup on the CPU, so the cost of generating a large set does not affect FPS. Additionally, accessing a set is simple and efficient, as each output fragment coordinate is simply divided by m to sample the 3D texture. The sampler is set to wrap the UV addresses, so the filter sets repeat in each direction on screen every m pixels.

The effect of kernel set count on the visuals is small. For the smallest 2×2 set there is visible repetition in the structure of the penumbra in the mouth of the dragon or in the shadow near its head. Higher counts provide improvements to the apparent smoothness of the shadow. The difference is a bit more noticeable in movement, where too few sets cause visible repeating patterns in the penumbra regions, that are overlaid on the shadow effect. These artifacts tend to disappear at sizes higher than 14×14.

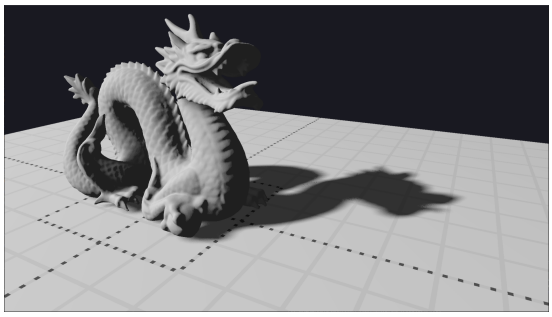
As it was mentioned, the 3D texture used to store the kernel offsets has dimensions $m \times m \times n^2/2$. This means that memory requirements of this technique grow exponentially both with the dimension m and n. The size is slightly reduced by using a texture with



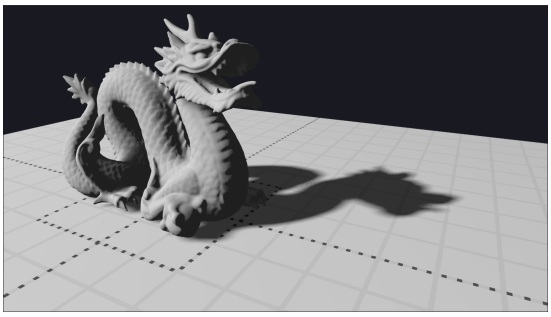
(a) The Chinese Dragon rendered with 2×2 set of kernels with adaptive PCF.



(b) The Chinese Dragon rendered with 6×6 set of kernels with adaptive PCF.



(c) The Chinese Dragon rendered with 14× 14 set of kernels with adaptive PCF.



(d) The Chinese Dragon rendered with 24× 24 set of kernels with adaptive PCF.

Figure 4.10: The Chinese Dragon rendered with adaptive PCF, using a 1024×1024 shadow map resolution, kernel size of 8 × 8, kernel offset scale of 5 and varying counts of kernel sets.

four, 32-bit channels, and storing each xy offset into two of the channels. This means that two kernel offsets can be stored in a single texel of the texture.

It is worth noting a caveat of this technique that is visible in the results in Fig. 4.10. High contrast noise appears in lit areas that are surrounded by shadow, such as in between the neck and back of the dragon's shadow. These erroneous areas stem from the optimization technique. Since the samples farthest away from the kernel center determine if the point is in shadow or not, such areas can get incorrectly classified as shadowed. This issue could be alleviated by using weights when averaging the filter samples, that decrease with the distance from the kernel center.

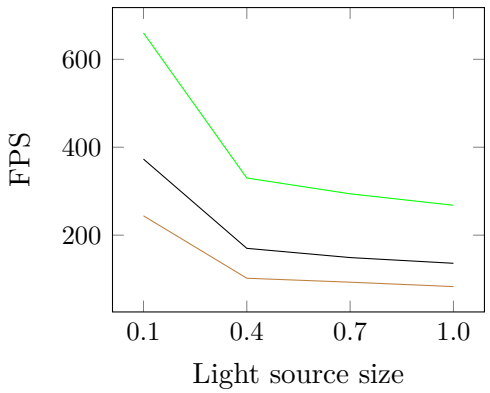
Overall, the performance and visual results of adaptive PCF make it a very practical solution to filtering shadow maps. The shadows look smooth, and their penumbras are easily controlled with the offset scale.

4.3.3 Soft shadows with shadow maps

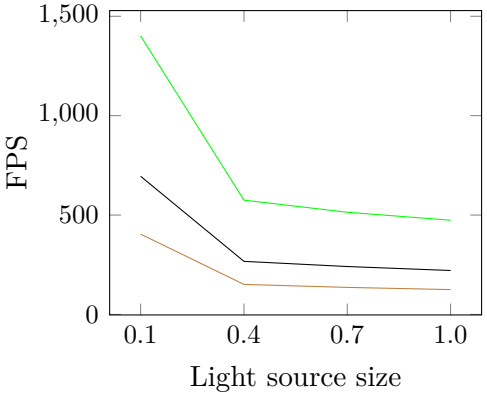
PCSS

Implementation of percentage-closer soft shadows was presented in section 3.2.6. This section presents the results of testing this technique. Figure 4.11 shows FPS measurements

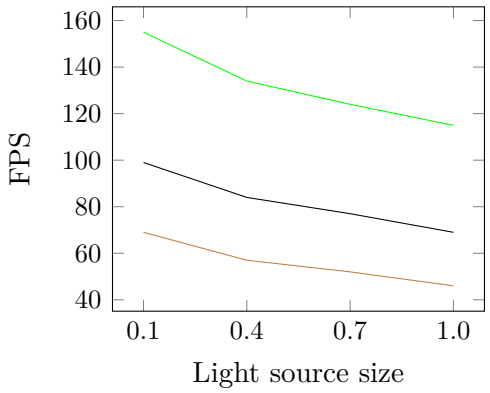
for all test scenes with different light source sizes, which define the size of the penumbra and impact overall performance. PCSS is implemented here using adaptive techniques for both finding the average occluder depth and filtering.



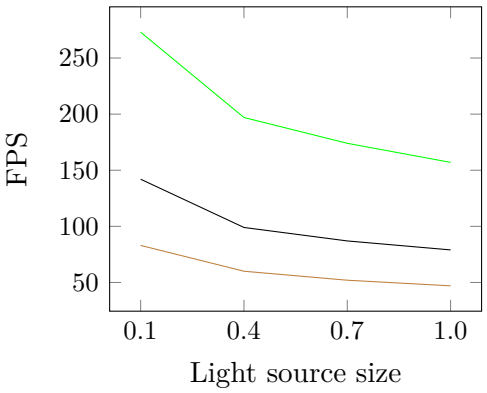
(a) Results for Chinese Dragon scene.



(b) Results for the Cube scene.



(c) Results for the Power Plant scene.



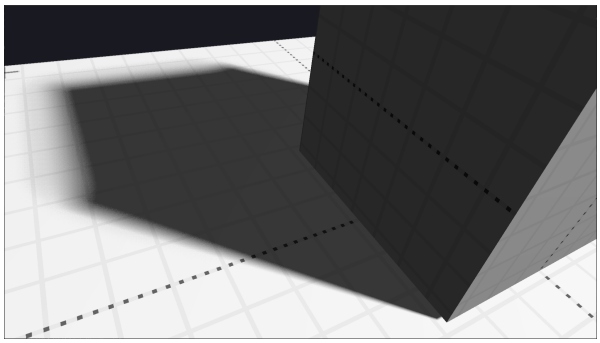
(d) Results for the Crytek Sponza scene.

Figure 4.11: Frames per second for all test scenes, for different sizes of the light source and output resolutions. In green 1280×720, in black 1920×1080 and in brown 2560×1440. Rendering with shadow mapping with PCSS at constant shadow map size 1024×1024 and 16×16 8×8 kernels.

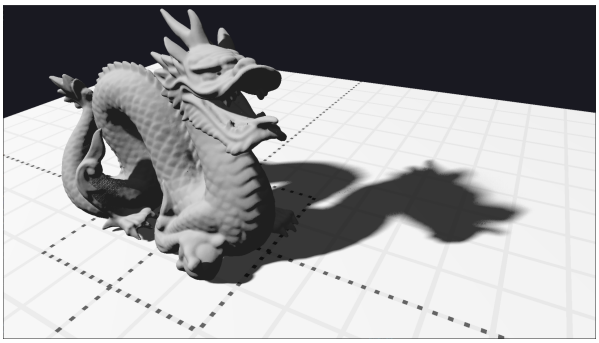
Performance is affected more negatively than with adaptive PCF. This decrease happens because PCSS basically performs two adaptive filtering-like steps per pixel. It first has to sample the shadow map to find the average occluder depth and then again for PCF with a calculated size. For simpler scenes however, as the size of the light source increases, there seems to be a threshold value beyond which the performance degradation slows down. This is probably because in the case of a simple scene like the Chinese Dragon or the Cube, once the light source is large enough almost all surfaces will be evaluated as if in a penumbra. Additionally, very high light source sizes produce unpleasant results and should not be used in a real application anyway.

The light size needs to be carefully tuned on a per-scene basis to get plausible results. Additionally, bias has to be adjusted to account for wider kernels. Images rendered with

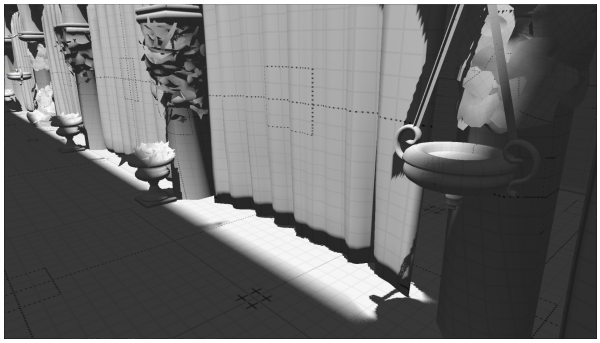
PCSS are presented in Fig 4.12.



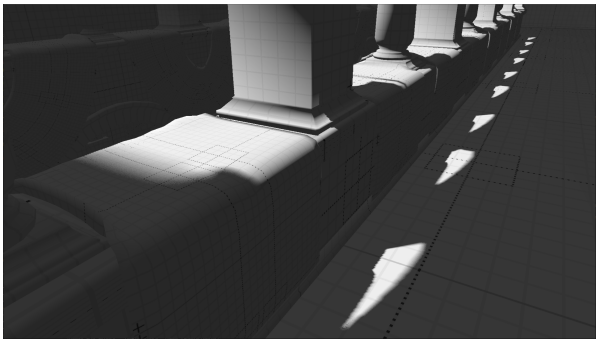
(a) The Cube scene showing smooth shadow penumbra falloff with changing distance between shadow caster and receiver.



(b) The Chinese Dragon showcasing contact hardening shadows.



(c) PCSS shadows in Crytek Sponza, creating soft shadows cast by the rooftop above and harder shadows cast by the hanging bowl.



(d) Shadows hardening with changing distance to occluder in Crytek Sponza.

Figure 4.12: The results of PCSS in different scenes, rendered with 4096×4096 shadow map resolution.

PCSS has its performance cost, but the results it produces can certainly elevate the graphical presentation of a real-time application.

4.4 Summary and conclusions

Basic shadow mapping lays the foundation on which more advanced techniques can be built. It has a performance cost that is slightly smaller than double the main render pass. On its own it requires very high resolution to get pleasant results, so that aliasing it eliminated.

It can be enhanced with various filtering methods. Bilinear comparison filtering is implemented within GPU hardware, so it can be very fast. It can be used to blur the edges of shadows for almost no performance cost. This helps separate hard shadows from actual geometry. Percentage-closer filtering allows control over the strength and quality of the shadow smoothing effect. It uses a square filter kernel to average shadow function results. The performance drops very rapidly with increase in the filer size and the results

feature quite severe banding and visible shadow map texel borders. This can be greatly improved by combining PCF with hardware bilinear filtering. Then, even a very small shadow map and a modestly sized kernel can produce very pleasant shadows. The last filtering method, adaptive PCF, improves both performance and visual results. By using sets of filter kernels stored in a 3D texture, the algorithm can quit filtering early where further samples are not required. The blurring gives very smooth results, hiding any banding or texel boundaries and is easily controllable. Adaptive PCF increases the memory consumption by $m \times m \times (n^2/2)$. It can also unfortunately introdce high-contrast noise in specific areas of the shadows.

Finally, percentage-closer soft shadows can be used to render contact-hardening shadows. This technique gives the most robust visuals, but requires careful hand tuning of the parameters such as bias and light size. It gives realistic looking results, which are however not physically correct, and are only an approximation. The performance is approximately half of adaptive PCF on its own.

If the visuals are of the highest importance and the application has a lot of frame time budget available, PCSS is the best choice. Otherwise, a balance between shadow map resolution, kernel size and kernel offset scale should be found within adaptive PCF. Basic PCF should probably be avoided, unless combining a small kernel size with hardware bilinear filtering and high shadow map resolution for basic antialiasing of shadows.

Chapter 5

Summary

Within the realm of computer graphics rendering, many approaches to render various physical phenomena exist. Shadow rendering is no exception, offering a plethora of choices varying in visual robustness, achievable performance and realism. This thesis analyzed some of the techniques meant for real-time shadow rendering, providing their background and explaining their algorithms. Within the scope of the thesis, a test application was built in c++ using the DirectX 12 graphics API. The application implemented chosen techniques and allowed them to be tested and compared against each other on the basis of performance and visual quality.

The tests demonstrated that each technique has its pros and cons. The beast quality of the rendered results in percentage-closet soft shadows is reflected in the heavy performance requirements. Fast basic shadow mapping would look out of place within a modern renderer, giving unrealistic hard shadows and heavy aliasing. A solid middle ground seems to be adaptive percentage-closer filtering, which produces pleasant results and attempts to improve the performance of basic PCF with a clever branching algorithm.

Within the scope of this thesis five approaches to shadow rendering were implemented and tested. In the future it would be worthwhile to investigate more advanced and modern rendering techniques to see how some problems present in the ones already covered are getting solved and in what directions the field is heading.

Bibliography

[1] Thomas Akenine-Moller, Eric Haines and N. Hoffman. *Real-Time Rendering, Fourth Edition*. Boca Raton: Taylor & Francis Ltd, 2018. ISBN: 978-1-1386-2700-0.

[2] Jiří Bittner, Oliver Mattausch and A. Silvennoinen. 'Shadow caster culling for efficient shadow mapping'. In: *Symposium on Interactive 3D Graphics and Games*. 2011, 81—88.

[3] J. Blinn. 'Me and My (Fake) Shadow'. In: *IEEE Comput. Graph. Appl.* 8.1 (1988), 82—86.

[4] Stefan Brabec, Thomas Annen and H.-P. Seidel. 'Practical Shadow Mapping'. In: *Journal of Graphics Tools* 7.4 (2002), pp. 9–18.

[5] William Donnelly and Andrew Lauritzen. 'Variance shadow maps'. In: *Proceedings of the 2006 Symposium on Interactive 3D Graphics and Games*. I3D '06. Redwood City, California: Association for Computing Machinery, 2006, 161–165. ISBN: 159593295X. DOI: 10.1145/1111411.1111440. URL: https://doi.org/10.1145/1111411.1111440.

[6] Randima Fernando. 'Percentage-closer soft shadows'. In: *ACM SIGGRAPH 2005 Sketches*. SIGGRAPH '05. Los Angeles, California: Association for Computing Machinery, 2005, 35–es. ISBN: 9781450378277. DOI: 10.1145/1187112.1187153. URL: https://doi.org/10.1145/1187112.1187153.

[7] Daniel Kersten, Pascal Mamassian and D. Knill. 'Moving Cast Shadows Induce Apparent Motion in Depth'. In: *Perception* 26.2 (1997). PMID: 9274752, pp. 171–192. DOI: 10.1068/p260171. eprint: https://doi.org/10.1068/p260171. URL: https://doi.org/10.1068/p260171.

[8] Andrew Lauritzen and Michael McCool. 'Layered variance shadow maps'. In: *Proceedings of Graphics Interface 2008*. GI '08. Windsor, Ontario, Canada: Canadian Information Processing Society, 2008, 139–146. ISBN: 9781568814230.

[9] Andrew Lauritzen, Marco Salvi and Aaron Lefohn. 'Sample distribution Shadow Maps'. In: Feb. 2011, pp. 97–102. DOI: 10.1145/1944745.1944761.

[10] *Life of a triangle - NVIDIA's logical pipeline*. URL: https://developer.nvidia.com/content/life-triangle-nvidias-logical-pipeline (visited on 14/10/2024).

[11] Brandon Lloyd, David Tuft, Sung-eui Yoon and Dinesh Manocha. 'Warping and Partitioning for Low Error Shadow Maps'. In: Jan. 2006, pp. 215–226. DOI: 10.2312/EGWR/EGSR06/215-226.

[12] Tom Lokovic and Eric Veach. 'Deep Shadow Maps'. In: *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*. 1st ed. New York, NY, USA: Association for Computing Machinery, 2023. ISBN: 9798400708978. URL: https://doi.org/10.1145/3596711.3596746.

[13] Márcio C. F. Macedo and Antônio Lopes Apolinario. 'Revectorization-Based Shadow Mapping'. In: *Graphics Interface*. 2016. URL: https://api.semanticscholar.org/CorpusID:33053694.

[14] Márcio C. F. Macedo, A. L. Apolinário Jr. and K. A. Agüero. 'Revectorization-Based Soft Shadow Mapping'. In: *Computer Graphics Forum* 39.1 (2020), pp. 389–404. DOI: https://doi.org/10.1111/cgf.13810. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.13810.

[15] Tobias Martin and Tiow-Seng Tan. 'Anti-aliasing and Continuity with Trapezoidal Shadow Maps.' In: Jan. 2004, pp. 153–160. DOI: 10.2312/EGWR/EGSR04/153-160.

[16] Morgan McGuire. *Computer Graphics Archive*. 2017. URL: https://casual-effects.com/data (visited on 23/10/2024).

[17] Hubert Nguyen. *GPU Gems 3*. Addison-Wesley Professional, 2007. ISBN: 9780321515261.

[18] Christoph Peters and Reinhard Klein. 'Moment shadow mapping'. In: *Proceedings of the 19th Symposium on Interactive 3D Graphics and Games*. i3D '15. San Francisco, California: Association for Computing Machinery, 2015, 7–14. ISBN: 9781450333924. DOI: 10.1145/2699276.2699277. URL: https://doi.org/10.1145/2699276.2699277.

[19] Matt Pharr and Randima Fernando. *GPU Gems 2: Programming Techniques For High-Performance Graphics And General-Purpose Computation*. Addison-Wesley Professional, 2005.

[20] William T. Reeves, David H. Salesin and R. L. Cook. 'Rendering antialiased shadows with depth maps'. In: *Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH '87. New York, NY, USA: Association for Computing Machinery, 1987, 283–291. ISBN: 0897912276. DOI: 10.1145/37401.37435. URL: https://doi.org/10.1145/37401.37435.

[21] Paulo E. Santos, Roberto Casati and P. Cavanagh. 'Perception, cognition and reasoning about shadows'. In: *Spatial Cognition & Computation* 18.2 (2018), pp. 78–85. DOI: 10.1080/13875868.2017.1377204. eprint: https://doi.org/10.1080/13875868.2017.1377204. URL: https://doi.org/10.1080/13875868.2017.1377204.

[22] Daniel Scherzer, Michael Schwärzler, Oliver Mattausch and M. Wimmer. 'Real-Time Soft Shadows Using Temporal Coherence'. In: *Advances in Visual Computing*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 13–24. ISBN: 978-3-642-10520-3.

[23] Marc Stamminger and George Drettakis. 'Perspective shadow maps'. In: *ACM Trans. Graph.* 21.3 (2002), 557–562.

[24] Liang Wan, Tien-Tsin Wong and C. Leung. 'Isocube: Exploiting the Cubemap Hardware'. In: *IEEE Transactions on Visualization and Computer Graphics* 13.4 (2007), pp. 720–731.

[25] Yulan Wang and Steven Molnar. *Second-Depth Shadow Mapping*. Tech. rep. USA, 1994.

[26] Lance Williams. 'Casting curved shadows on curved surfaces'. In: *SIGGRAPH Comput. Graph.* 12.3 (1978), 270—274.

[27] Michael Wimmer, Daniel Scherzer and W. Purgathofer. 'Light space perspective shadow maps'. In: *Proceedings of the Fifteenth Eurographics Conference on Rendering Techniques*. EGSR'04. Norrköping, Sweden: Eurographics Association, 2004, 143–151. ISBN: 3905673126.

[28] Andrew Woo. 'The shadow depth map revisited'. In: *Graphics Gems III*. USA: Academic Press Professional, Inc., 1992, 338—342. ISBN: 0124096719.

Appendices

Technical documentation

.1 Building the test application

The test application uses CMake to get all the dependencies and create a buildable MSVC project. The application was built using Visual Studio Code and CMake Tools extension. In the CMake Tools extension tab the configuration, for example 'msvc Release', should be selected. After that the build target 'ShadowRendering' and a preset have to be set. The extension will then begin fetching the dependencies and configuring the project. Once ready the program can be built. The only dependency that is not handled by CMake is the Windows SDK. The user has to install it manually in the default installation location. The executable is built into an appropriate location in the /build directory, depending on the configuration.

.2 Using the test application

Once launched, the application will load the Cube scene and present a window containing the rendered output. ImGUI can be toggled with the H key. It contains information on controls, allows selection of rendering modes and provides brief explanations. It also displays the FPS and frame time of the application if the widget is enabled.

.3 Profiling the test application

To profile the test application, an executable of TracyProfiler can be downloaded from its GitHub repository. First the tracy-profiler.exe should be run, then the test application. The test application should be listed in Tracy as ready for profiling. Upon connecting, data gathering will begin immediately and frames will start appearing in the Tracy GUI. They can be later analyzed and inspected, for example by checking how the GPU is utilized, inspecting the time the CPU is waiting for the GPU or analyzing the timing statistics that Tracy provides.

List of abbreviations and symbols

CGI	computer-generated imagery
CAD	computer-aided design
FPS	frames per second
PSM	perspective shadow maps
LiSPSM	light space perspective shadow maps
PCF	percentage-closer filtering
VSM	variance shadow maps
LVSM	layered variance shadow maps
PCSS	percentage-closer soft shadows
MSM	moment shadow mapping
RBSM	revectorization-based shadow mapping
NDC	normalized device coordinates
WVP	world–view–projection
API	application programming interface
DSV	Depth Stencil View
SRV	Shader resource view

List of additional files in electronic submission (if applicable)

Additional files uploaded to the system include:

- source code of the test application,
- the Cube test scene and sample texture,
- a video file showing the running test application

List of Figures

List of Tables