

RANDOM ACCESS RENDERING OF ANIMATED VECTOR GRAPHICS USING GPU

IVAN LEBEN

Supervisor: GEOFF LEACH

Honours Thesis

School of Computer Science and Information Technology
RMIT University
Melbourne, AUSTRALIA

October, 2010

Abstract

Efficient random access to image data allows texture mapping of images onto curved surfaces to be accelerated by the GPU. This property has traditionally been associated with raster images, whereas a typical vector image representation cannot be efficiently interpreted at an arbitrary point by the established graphics pipeline. Recent research has allowed texture mapping of vector images by taking advantage of the programmable stages of the GPU. However, the new rendering approach introduces an expensive preprocessing step running on the CPU. In this thesis we investigate techniques to fully accelerate vector image rendering and texture mapping on the GPU, including the preprocessing steps. We present an alternative localised vector image representation and an efficient parallel encoding algorithm, achieving great performance improvements compared to a sequential CPU implementation and allowing random access rendering of animated vector images, where continuous preprocessing is required.

Contents

1	Introduction	3
2	Background	4
2.1	Vector Image Definition	4
2.2	Evaluation of Spline Segments	5
2.3	Winding Number Evaluation	6
2.4	Traditional Graphics Pipeline	6
2.5	Texture Lookups	7
2.6	GPU Programmability	8
2.7	Recent Advances in GPU Architecture	9
3	Previous Work	10
3.1	Forward Rendering Approaches	11
3.2	Random Access Rendering by Image Localisation	12
4	Our Approach	13
4.1	Pivot-Point Localised Image Representation	13
4.2	Image Processing Pipeline	14
4.3	Encoding Algorithm	14
4.3.1	Algorithm Overview	15
4.3.2	Segment List Construction	17
4.3.3	Pivot-Point Localisation	18
4.3.4	Auxiliary-Segment Localisation	19
4.4	Rendering Algorithm	19
4.4.1	Random Access Lookup	19
4.4.2	Antialiasing	20
5	Results and Analysis	20
5.1	Experiment Design	20
5.2	Experiment #1: Encoding Performance	22
5.3	Experiment #2: Rendering Performance	22
5.4	Experiment #3: Encoding + Rendering Performance	24
5.5	Experiment #4: Performance Scaling with Raster Size	24
5.6	Experiment #5: Performance Scaling with Geometric Complexity	26
5.7	Experiment #6: Memory Requirements	26
5.8	Limitations of Our Approach	27
5.9	Non-traditional Applications	27
6	Conclusion and Future Work	28



Figure 1: Vector image (top row) retains sharp colour discontinuities even at large zoom, while raster image (bottom row) suffers from noticeable artifacts.

1 Introduction

Random access rendering is a method whereby the colour of an image is retrieved at an arbitrary image space coordinate in an arbitrary order. Efficient random access to image data is a requirement for texture mapping and has traditionally been associated with raster images. The definition of a raster image as a grid of coloured pixels lends itself perfectly to the random access scheme. However, the main drawback is the artifacts such as jagged or blurred edges arising from missing detail in information when the image is scaled, as shown in Figure 1.

In contrast, vector graphics use mathematical equations and shapes to describe coloured regions. Such a definition allows scaling of a vector image while retaining sharp edges at colour discontinuities. As illustrated in Figure 2, in order to classify a point as inside or outside a vector shape, we need to consider all of its boundary segments. Moreover, when rendering a small portion of a shape under large zoom, the parts outside of the visible area still need to be processed, adding to unnecessary overhead as the zoom factor increases. This makes straightforward approaches to random access inefficient and has until recently prevented the use of vector images as textures.

Applications of vector graphics are many, ranging from high detail maps to general scalable user interfaces. An efficient random access algorithm would further broaden the usability of vector images. Regions of a map outside of the view could be efficiently discarded as the user zooms in on a particular detail. Graphical user interfaces could be mapped onto surfaces in 3D space allowing new ways of navigation and better screen estate management.

Historically, Graphics Processing Units (GPUs) have specialised into rendering of triangle primitives and have no built-in support for rasterization of curved shapes present in vector images. Nonetheless, dependency on the triangle-based rasterization has been alleviated by the increase in GPU programmability. Recently, [NH08] have developed a rendering technique which evaluates vector shapes directly in the programmable stages of the GPU, after localising vector data in an expensive preprocessing step on the CPU. This allows efficient random access, but limits the use to static non-animated graphics.

An important thing to consider is the paradigm shift taken by an approach such as [NH08]. In the traditional rendering technique the graphical primitives are processed once and displayed on the

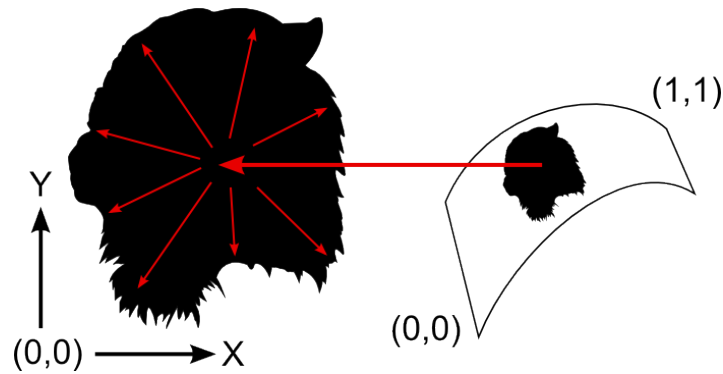


Figure 2: Straightforward approaches to random-access rendering of vector images are inefficient. To find the colour of one textured pixel on the screen, all the segments of the shape need to be processed.

screen directly after a single pass through the graphics pipeline. In contrast, the approach being discussed preprocesses the entire set of primitives in the scene to build a look-up data structure, which allows all the primitives affecting a given pixel to be looked up efficiently in the final rendering stage to determine the colour of a pixel. It is important to note that such an approach increases the per-pixel processing complexity, but is facilitated by the recent advance in graphics hardware architecture.

The focus of this research is to extend the above approach by investigating alternative vector image data structures and preprocessing techniques, suitable to parallelisation and acceleration via the GPU. Our goal is to implement the entire rendering process, including the preprocessing steps, on the GPU. By taking advantage of the parallel architecture, significant improvement in performance is achieved, allowing application to animated vector images where continuous preprocessing is required.

In general, this thesis investigates the following questions:

1. How to efficiently accelerate vector image localisation using the parallel architecture of the GPU?
2. What are the performance gains of a parallel preprocessing algorithm compared to a sequential CPU implementation?
3. Is there an alternative localised vector image representation, more suitable to parallel encoding and rendering?
4. What degree of vector data localisation provides an optimal balance between preprocessing and rendering time?
5. What is the benefit of vector data localisation with regard to animated vector graphics where continuous preprocessing is required?

2 Background

2.1 Vector Image Definition

A vector image is a set of vector shapes, each defined as an open or closed spline of straight line and curved segments. Closed shapes can be filled, meaning the spline defines a border between the exterior and the coloured interior of the shape. Shapes can also be stroked, meaning the boundary

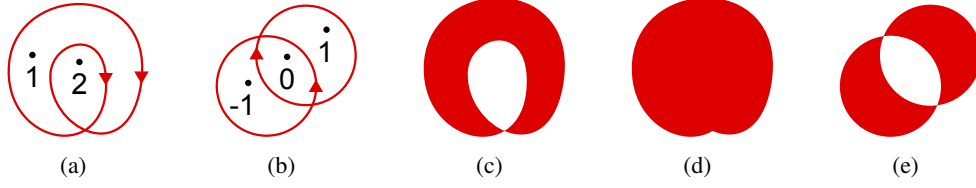


Figure 3: (a) a shape consisting of a single contour (b) a shape consisting of two contours with opposite orientation (c) shape A using even-odd fill rule; (d) shape A using non-zero fill rule; (e) shape B appears the same using both fill rules.

has a visible thickness. For practical purposes, stroked shapes can be adequately approximated by another more complex vector shape describing the resulting outline of the thick line, thus reducing the definition to closed, filled shapes.

The interior of the filled shapes is determined according to the value called *winding number*. For any point, the winding number represents the total number of times that the curve travels around the point. Clockwise movement counts as positive whereas anticlockwise movement counts as negative. Figure 3 shows examples of points in curves and their respective winding numbers. Note that one shape can consist of multiple contours in which case the winding number is the sum of the individual winding numbers of all the contours.

There are two common rules for determining the interior of a shape, namely *even-odd* and *non-zero*. The former classifies the areas with an odd winding number as interior while the latter considers any area with a winding number different than 0 to be inside. Depending on the number of contours and their orientation, a shape might have a different or equal appearance with different fill rules.

2.2 Evaluation of Spline Segments

Individual segments of a vector spline are most commonly defined using Bézier curves of first to third degree, as described in [Far02]. In general a Bézier curve of degree n is defined as follows:

$$\mathbf{B}(t) = \sum_{i=0}^n \mathbf{b}_{i,n}(t) \mathbf{P}_i, \quad t \in [0, 1]$$

where the points \mathbf{P}_i are called *control points* of the Bézier curve and the polynomials

$$\mathbf{b}_{i,n}(t) = \binom{n}{i} t^i (1-t)^{n-i}, \quad i = 0, \dots, n$$

are known as Bernstein basis polynomials of degree n with the following property:

$$\sum_{i=0}^n \mathbf{b}_{i,n}(t) = 1, \quad t \in [0, 1]$$

meaning that the sum of all the polynomials of the same degree is always 1. The Bézier curve equation therefore interpolates between the given control points according to the parameter t , beginning exactly at \mathbf{P}_0 and ending exactly at \mathbf{P}_n . A curve of degree n has $n + 1$ control points. A Bézier curve of degree 1 is essentially a straight line. Consecutive spline segments share a control point where one ends and the next one begins. The ordered set of control points of all the segments defines the *control polygon* of the spline as shown in Figure 4.

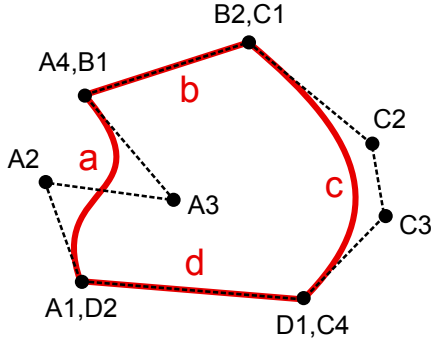


Figure 4: A closed spline (red) consisting of two line segments and two cubic Bézier curve segments with its control polygon (black).

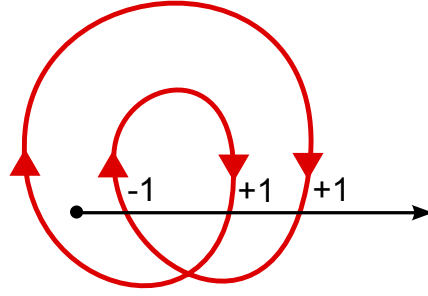


Figure 5: The direction of the intersections between a horizontal ray and the segments of a spline count towards the winding number of the point.

A simple way of calculating a point on the Bézier curve at an arbitrary parameter value is to substitute the value into every basis function, compute the products with the respective control points and add them together. While straightforward, this approach is not numerically stable. Instead, as described in [BM99] we can evaluate the points on the curve by recursive subdivision of the sides of its control polygon. This technique, known as *de Casteljau's algorithm* reduces numerical errors during the course of evaluation.

2.3 Winding Number Evaluation

The Bézier curve equations can also be used to evaluate the winding number of any point with regards to a vector spline. We start by casting a ray from the point in an arbitrary direction. Based on the *Jordan Curve Theorem*, the point is inside the polygon, if the ray crosses an edge of the polygon an odd number of times. Counting the number of intersections gives a value known as the *crossing number*. Now, by checking the direction of the intersecting edge, we instead add +1 for every left-to-right crossing and -1 for every right-to-left crossing. The resulting value is the winding number of the point.

Imagine the ray is being cast from point $P(x, y)$ horizontally in the positive direction of the x axis. We can find an intersection with any segment of the spline by finding the roots t_i of the equation $y = y(t)$ where $y(t)$ is the linear or higher order polynomial describing the spline segment. When $t_i \in [0, 1]$ and $x < x(t_i)$, the ray intersects the segment.

Obviously, the winding number at any point depends on the geometry of the entire shape outline. More precisely, to evaluate the winding number of a single point, all the segments of the shape to the right of this point need to be identified and checked for horizontal ray intersection. It is this requirement that poses the greatest obstacle when designing an efficient random access rendering algorithm.

2.4 Traditional Graphics Pipeline

Modern display devices require input in the form of a grid of colour values, called a *raster*. A colour value in a raster is called a *pixel*. For a graphical object or shape to be displayed, the pixels covered by the shape need to be determined in a process called *rasterisation* and the colour values in the display raster changed accordingly. The architecture of modern GPUs is optimised for rasterisation of three

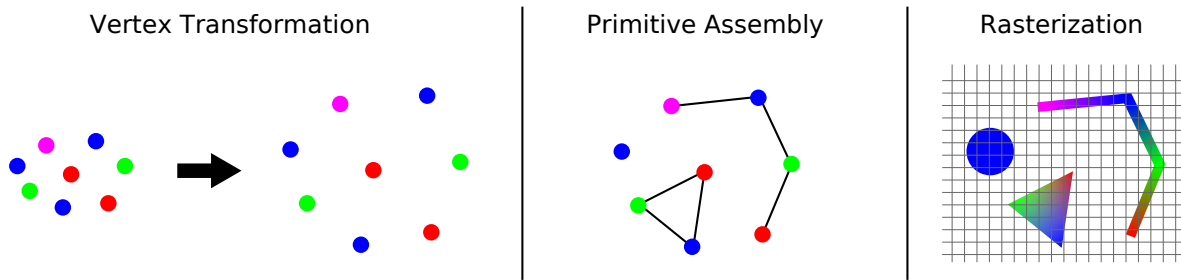


Figure 6: In traditional graphics pipeline, transformed vertices are assembled into one of the three primitive types and then rasterized into the framebuffer.

types of primitives: points, lines and triangles. The roots of this design reach into the early history of computer graphics when parallel computing architecture was not as powerful and programmable as today. Triangle rasterisation coupled with texture mapping of raster images was much simpler to implement in hardware compared to direct evaluation of higher order mathematical equations. Thus more complex shapes have to be converted into the basic primitives by *tessellation*, in order to accommodate and benefit from hardware-accelerated rendering.

Figure 6 shows how a traditional graphics pipeline processes the data. The input comes in the form of vertices which can carry multiple attributes, such as vertex coordinates, texture coordinates and colour. In a 3D application, the vertices carry specialised attributes required by the lighting equations, such as surface normals. The pipeline itself also has a set of attributes which represent its state. The state governs the processing of the input data and can be modified by the drawing commands. It includes important processing information, such as the vertex transformation matrix, the type of primitive for the current set of vertices, and the material and light parameters.

The pipeline consists of three stages. The vertex transformation stage applies the transformation matrix to the vertex coordinates. If lighting is enabled this stage performs lighting calculations and uses the result to determine the vertex colour. The primitive assembly stage combines the transformed vertices into one of the three basic primitive types, as defined in the pipeline state. The rasterization stage determines the pixels covered by each primitive and computes the interpolated values of the vertex attributes at every pixel. Finally, the resulting colour at each pixel is written into a temporary raster within the block of memory called *framebuffer*. When the entire image of one frame is constructed, the colour values are copied from the temporary raster into the actual display raster.

2.5 Texture Lookups

There are two common ways of determining the final pixel colour. As shown in Figure 6, the colour attribute of the vertex or the colour obtained as a result of the lighting equations can be directly interpolated across the primitive. Alternatively, every vertex can have a texture coordinate attribute assigned. The interpolated texture coordinate at every pixel is used to perform a lookup into a texture image, residing in video memory of the GPU. The lookup applies the colour of the texture at the point of texture coordinate to the pixel and might include some form of filtering to smooth out the colour values of adjacent pixels. This process is illustrated in Figure 7. A group of textured triangles small enough to approximate a curved surface can give an impression of the texture image being mapped onto an arbitrary curved surface.

Due to parallel architecture of the rasterization stage, every pixel is processed separately and the order of coordinates in the texture lookups is arbitrary. In order to be usable as a texture an image

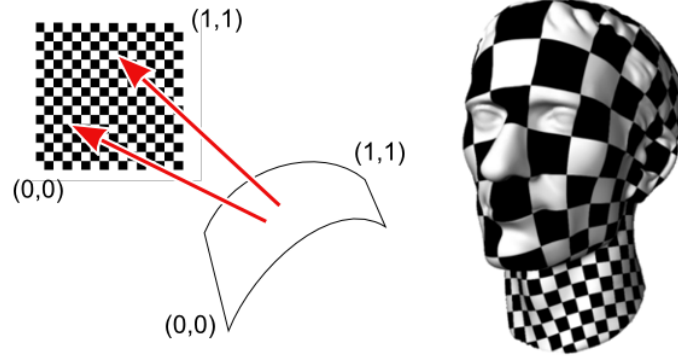


Figure 7: The process of texturing maps a texture image onto a curved surface by interpolating texture coordinates across the primitive and performing a lookup at every pixel.

must offer efficient random access to its data. To this end, raster images have long been used for texturing. If the colour values are laid out in row-major order, the following simple equation suffices to compute the memory location of a pixel within the raster data, corresponding to the given image space coordinate:

$$L = Y * W + X$$

where (X,Y) is the texture coordinate and W is the width of raster image in pixels. Once the memory location is computed, the colour value can be read directly from the raster grid and applied to the pixel being textured.

On the other hand, vector images do not offer such a simple and efficient colour lookup. As discussed in previous sections, the colour at an arbitrary coordinate in image space of a vector image depends on the winding number of that point with regards to every shape in the image. Furthermore, the winding number in turn depends on every segment of the respective shape. In order to use vector images as textures some preprocessing needs to be performed, to construct a more efficient lookup structure.

2.6 GPU Programmability

The GPU has evolved enormously in terms of speed and programmability since it first appeared. However the fundamental approach has not changed significantly, with the triangle rasterisation unit remaining a core built-in component. Nonetheless, the increase in programmability has enabled new ways of using the GPU.

Firstly, programmers can now write code that defines the behaviour of certain stages of the rendering pipeline as exposed by the graphics APIs such as OpenGL and DirectX. The pipeline is still based around points, lines and triangles being the basic rasterizable primitives, however, custom programs for processing of vertices and pixels can be uploaded onto the hardware. Figure 8 shows the programmable version of the traditional graphics pipeline as presented by the OpenGL API. The programmable stage processing the vertices is called a *vertex shader*, while the pixel processing stage is named a *fragment shader*. In addition, new programmable stages are being introduced, not present in the traditional fixed pipeline. The recently introduced *geometry shader* allows custom processing on the level of entire primitives, after the primitive assembly stage.

The programmable stages can be used to improve traditional rendering approaches. For example,

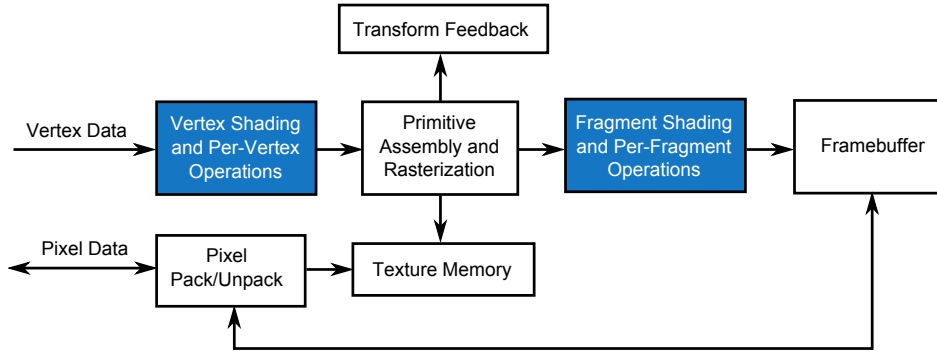


Figure 8: OpenGL rendering pipeline. Programmable parts are coloured in blue.

a user-defined vertex displacement program can modify the input coordinates on a per-pixel basis, rather than by a unique transformation matrix. A specialised fragment program can perform per-pixel lighting, thus achieving higher realism of the rendered image. A custom geometry shader can discard the input primitives and generate new ones, therefore dynamically adjusting the object geometry.

More generally, the programmability of the GPU is allowing it to be used as a general processing unit, an approach known as *General Purpose GPU processing* (GPGPU). Arbitrary input data can be encoded into vertices and then decoded and processed by a shader program. The ability to perform texture lookups at any point in a shader program allows arbitrary data encoded into textures to be read directly from the GPU memory. The option to render into the raster of a texture image, rather than the display raster, allows generic output to be encoded and written out as colour values.

Furthermore, new programming interfaces such as CUDA [Cud] and OpenCL [Ope] have emerged that expose the GPU hardware as a generic multi-core computing device for computationally intensive tasks. This approach is known as *GPU Computing*. These APIs are completely devoid of any graphics-related paradigms, so there is no need to encode the data into graphical entities such as vertices and textures. Nonetheless, the results of the computation are often used for graphical purposes or end up being visualised via traditional rendering pipeline.

2.7 Recent Advances in GPU Architecture

With the increasing use of GPUs for general-purpose computation and the shift of the CPU towards multi-core computing, the architectures of the two processing units are in some ways starting to converge. On the one hand, the number of cores on the CPU increases with every generation. On the other hand, recent GPU architectures include features that were previously only found on a CPU.

NVIDIA has recently released a new generation of GPUs named *Fermi* [Fer]. Among other improvements, the new architecture introduced two very important features for our research: global memory space and a unified cache architecture. A new set of 64-bit addressing instructions allows C-style memory pointers to be used in shader code. The cache architecture ensures coherence in memory accesses between multiple shading threads, allowing data structures in the global memory to be shared between parallel executions of shader code. Concurrent access can be synchronised via atomic operations, which received a significant performance improvement compared to previous GPUs.

Although these changes are mostly targeted towards GPU Computing, some of them are exposed to the graphics pipeline as well. The OpenGL API allows new features to be accessed through third party *extensions*, which accelerate the evolution of the API even before the new features are approved as part of the core function set. With the new architecture, NVIDIA also released extensions to OpenGL

which take advantage of the new GPU capabilities in order to increase the versatility of the graphics pipeline.

Of particular relevance to our approach are the extensions `EXT_shader_image_load_store` and `NV_shader_buffer_store` which add new features to the shading language, allowing writes to arbitrary locations in the GPU memory. The former achieves this by allowing a texture image to be bound as a new type of rendering target, thus limiting the valid write locations to a subset of memory. The latter enhances the shading language with C-style pointers, through which the memory can be read or written at any location within the global memory space. Both extensions add new built-in functions to the language to perform atomic read and write operations.

The reason these two extensions are so important is they allow complex data structures to be constructed in the GPU memory. Previously, the output of a fragment shader had a fixed location in the target framebuffer at the coordinate of the corresponding pixel. With the new memory access functions, a shading program can lay out the data structures in memory much like a typical CPU program would. By using these extensions, our preprocessing algorithm can build an efficient vector image lookup structure in parallel from multiple executions of the shader program.

3 Previous Work

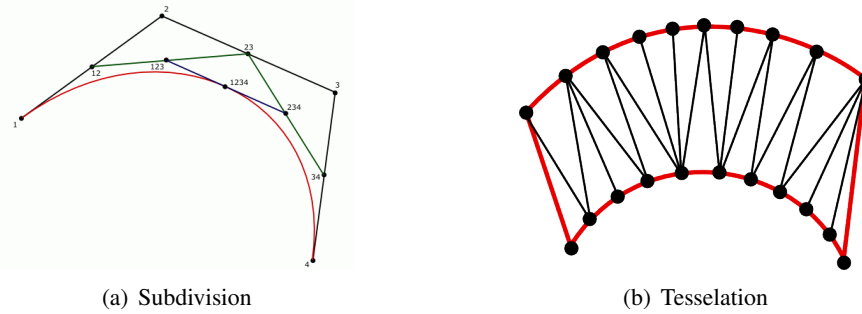


Figure 9: Traditional CPU renderer subdivides curves into straight line segments and then tessellates the shape interior into triangles.

In a typical CPU implementation of a vector graphics renderer, every spline is first approximated by linear segments. Intermediate points on the curved segments can be obtained in a straightforward manner by curve evaluation at a fixed interval. Alternatively, the process can be optimised via adaptive algorithms which generate less points in the straight regions and more points in the regions of higher curvature. Common approaches include forward difference evaluation and recursive subdivision based on de Casteljau algorithm. The interior of the resulting piecewise linear polygon is triangulated according to the fill-rule applied to the shape. Finally, the triangles are rasterized into the framebuffer. Figure 9 illustrates the process of subdivision and tessellation.

Unfortunately, the curve subdivision and polygon triangulation algorithms are expensive in terms of required processing time compared to the final rendering of the image. Moreover, they are inherently sequential, which means they cannot be efficiently parallelised and therefore cannot take advantage of the GPU.

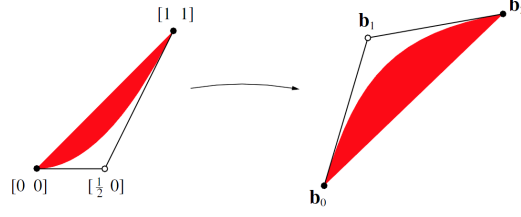


Figure 10: A quadratic curve is projected from texture space into screen space via coordinate interpolation. (Image source: [LB05])

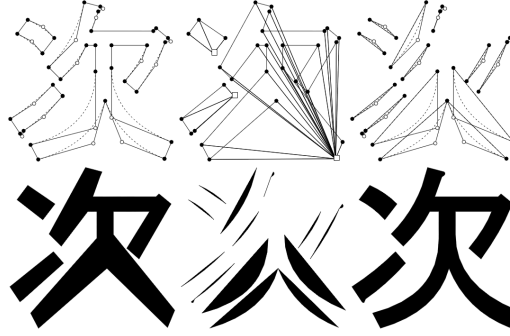


Figure 11: Stencil-buffer rendering technique based on Jordan Curve theorem avoids the need for triangulation of shape interior. (Image source: [KSST06])

3.1 Forward Rendering Approaches

[LB05] present an approach where the implicit equation of quadratic curves is evaluated directly in the fragment shader to perform inside/outside classification. They start by observing that “the implicit form of any rational parametric quadratic curve is a conic section; and that any conic section is the projected image of a single canonical parabola.” They show that a quadratic Bézier curve with control points $(0, 0)$, $(0.5, 0)$, and $(1, 1)$ plots the quadratic equation $v = u^2$ exactly. They assign these points as the (u, v) texture coordinates to the vertices of a triangle. Suppose the vertices are the control points of another arbitrary quadratic curve. As the hardware interpolates the texture coordinates across the triangle, it essentially projects the quadratic curve from texture space into screen space. They determine if the pixel is inside or outside the curve by evaluating $f(u, v) = u^2 - v$ in a pixel shader program. If $f(u, v) < 0$ then the pixel is inside the curve, otherwise it is outside.

This approach removes the need for curve subdivision, but the interior of the shape still needs to be triangulated. [KSST06] combine the projective rendering of the quadratic curves with a stencil polygon rendering technique. First, an arbitrary point P is selected among the segment control points. For each linear segment a triangle is drawn with one of its vertices at P and the other two at the vertices of the line segment. For each quadratic segment a triangle is drawn between P and the first and the last control point, skipping the middle control point. Additionally, a triangle is drawn for each quadratic curve, discarding the pixels based on the projective technique of [LB05]. For every pixel drawn, the binary value in the stencil buffer is inverted. After all the triangles have been drawn, the values in the stencil buffer essentially represent the crossing number for every point as if the intersection ray was cast towards P . This allows the pixels in the framebuffer to be coloured according to even-odd fill rule.

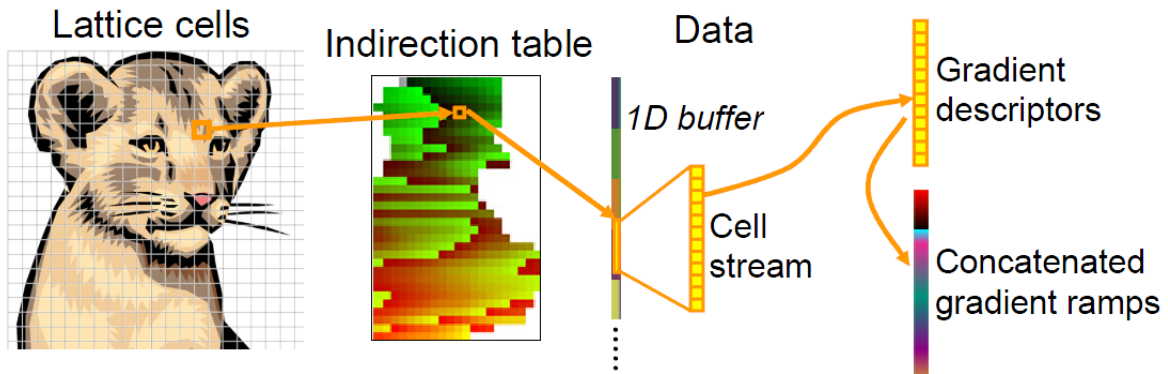


Figure 12: Random access rendering introduces a new paradigm in vector image rendering. (Image source: [NH08])

3.2 Random Access Rendering by Image Localisation

The algorithms described so far enable rendering of vector graphics to be accelerated by the GPU without the need for subdivision or tessellation. However, they are strongly based on the traditional primitive rasterization pipeline. As such, the rendered image can only be transformed by transforming the vertices of the triangle primitives before they are rasterized. Instead, we are aiming to use vector images as textures and map them onto other surfaces. As discussed previously, texture lookups require a random access scheme, where the pixel colours are queried in an arbitrary order at arbitrary locations in the image space. To achieve this capability, we need to construct an efficient lookup structure, that will allow evaluation of the winding number of any point with respect to any vector shape from within the fragment shader.

[NH08] presented an approach to this problem which localises vector image representation to coarse lattice cells in a preprocessing step implemented as a sequential CPU algorithm. Each lattice cell contains a localised description of the graphics primitives it overlaps. This information is encoded into a GPU texture and interpreted by the fragment shader. For every fragment, the vector data is read from the lattice cell that the texture coordinate falls in. The winding number is calculated using the ray-casting intersection technique, but thanks to the localisation, only the segments overlapping the cell need to be considered. They report that increasing the lattice resolution decreases the number of spline segments to intersect per pixel, as long as the detail in the image is fairly uniform. This in turn reduces the rendering time and increases the rendering framerate. However, they find that the preprocessing takes too long for real-time applications which prevents their approach to be used with animated images.

The above approach represents an important paradigm shift in vector image processing. Previous rendering techniques used vector image data as direct input to the traditional pipeline. Control vertices of vector curves were represented as vertices of basic graphics primitives and the primitive rasterization stage was cleverly enhanced with the specialized shading programs to perform the inside/outside classification of every pixel. In contrast, the image localisation approach does not represent any part of the vector image data using traditional graphics pipeline entities. A GPU texture is used merely as a memory buffer to contain encoded vector data and the fragment shader is used as a generic unit of computation that performs a look up into the encoded data and emits the resulting pixel color.

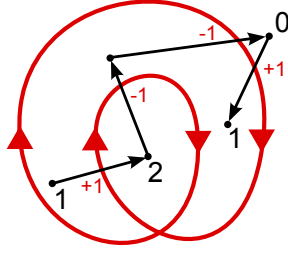


Figure 13: Starting from a known winding number, the winding number of adjacent regions can be determined by a single crossing of an edge.

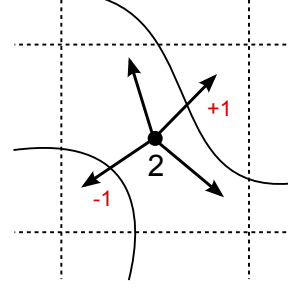


Figure 14: Casting a ray from the pivot point and intersecting the segments in the cell gives the winding number of any other point within the same cell.

4 Our Approach

In order to extend the random access rendering technique by [NH08] to animated vector images, the preprocessing step that constructs the lookup data structure needs to be accelerated. Their localised image representation is organized into grid cells, so it lends itself directly to a parallel encoding algorithm, meaning it can benefit from the parallel architecture of the GPU.

However, their “fast lattice-clipping algorithm” involves certain steps which cannot be efficiently parallelised. To preserve the winding number within a cell, their algorithm creates additional linear segments on the right boundary of each cell. We thus refer to the resulting lookup structure as *auxiliary-segment* (AS) image representation. The presence and direction of the auxiliary segments affects the winding number of the points in the cell and depends on the segments in other cells, to the right of the cell in question. Therefore, for every segment crossing a cell, the winding number needs to be integrated into the cells to its left. The dependency of each cell on the contents of other cells makes it challenging to devise a parallel localisation algorithm.

To avoid bottlenecks in our approach, we develop an alternative *pivot-point* (PP) localised image representation in an effort to minimise inter-cell dependencies and allow our encoding algorithm to take maximum advantage of the parallel GPU architecture. We also present an adjusted rendering algorithm that utilizes our image representation for texture mapping.

4.1 Pivot-Point Localised Image Representation

There are two main requirements for our localised representation: (i) it must exhibit a high degree of localisation to allow efficient random-access and (ii) it must be obtainable by an algorithm which can achieve a high degree of parallelism. Like [NH08], we create a coarse grid across the original vector image. However, our approach differs in the contents of the grid cells.

We start by observing that the winding number of any region in a vector shape can be determined by a single crossing of an edge, as long as the winding number of another adjacent region is known. As shown in Figure 13 it is possible to move from one region to another, adjusting the winding number according to the direction of the crossing edge and thus finding the winding number of every region.

We take advantage of this observation by assigning a pivot point to every cell, with its location implicitly defined as the center of the cell. We find the winding number of the pivot point in the preprocessing step and then use it as the source for ray-casting, similar to how the forward rendering technique of [KSST06] uses one of the control points as the pivot point for construction of auxiliary

triangles.

We adjust our rendering algorithm accordingly. While [NH08] cast a horizontal ray in the positive x direction, we cast a ray from the pivot point P of the cell in an arbitrary direction towards any other point W in the cell for which the winding number is needed. The intersections of the ray with vector segments of the shape give the change in winding number of W from the original winding number of P . Note that only the segments within the same cell as P and W need to be checked for intersection, as those are the only segments that can potentially stand in between the two points. Therefore, the number of segments that need to be intersected effectively decreases with the increase in grid resolution, thus satisfying requirement (i) for our localised representation.

In contrast to cell contents of an AS representation, the pivot winding number in PP representation only depends on the geometry of the shape and not the resulting contents of other cells in the grid. This makes our representation more suitable to a parallel construction algorithm. As explained in Section 2.3, finding the winding number at an arbitrary location on a point-by-point basis is not efficient. However, the subsequent sections show that it is possible to find the winding numbers for a set of points efficiently with a parallel algorithm, as long as the coordinates of the points are known in advance. Since locations of cell pivot points become known as soon as the grid resolution is chosen, our localised representation satisfies requirement (ii) set above.

Our PP representation of a vector image thus consists of a set of grid cells containing a pivot winding number and the list of intersecting spline segments. In contrast to AS representation, no additional auxiliary segments are required, resulting in a smaller number of segments per grid cell.

4.2 Image Processing Pipeline

We split our image processing pipeline into two stages: encoding and rendering. The purpose of the encoding stage is to convert the original vector image data into a localised representation. This stage performs all the preprocessing required so that the resulting lookup structure can be directly used for rendering. As long as the image data does not change, only the rendering stage needs to be invoked in order to display the image. When the image is animated, the encoding stage is invoked on every frame to rebuild the lookup structure before the rendering stage.

The following sections describe our encoding and rendering algorithms in detail. In order to be able to compare the performance of the two different localised representations, we implemented the encoding pass and the rendering pass for both our PP representation as well as the AS representation by [NH08]. The algorithms are mostly similar for both image representations, except where we explicitly mention the differences.

4.3 Encoding Algorithm

Our encoding algorithm requires minimal setup by the CPU which includes the following steps:

1. Approximate cubic curve segments with multiple quadratic segments.
2. Compute the coarse bounding box of the entire image and every separate object based on the segment control points.
3. Compute the memory size of the main grid and object sub-grids to format the grid memory pool.
4. Upload grid info, object info, and segment control points into the GPU memory.

4.3.1 Algorithm Overview

A high level description of our encoding algorithm is given in pseudocode below, excluding the details that differ between the auxiliary-segment and pivot-point representations:

```
//Initialisation pass
init stream length counter
init main grid counters
init object grid counters

[Sync memory]

//Segment encoding pass
for (every object)
    for (every line segment)
        find line bounding box
        for (every object grid cell in bbox)
            if (intersects cell)
                insert line segment node into cell stream
                link line node to previous node in the same cell
            endif
        loop
    loop
    for (every quadratic segment)
        find quad bounding box
        for (every object grid cell in bbox)
            if (intersects cell)
                insert quad segment node into cell stream
                link quad node to previous node in the same cell
            endif
        loop
    loop
loop

[Sync memory]

//Object header encoding pass
for (every object)
    for (every object grid cell in bbox)
        insert object header node into cell stream
        link object node to previous object node in the same cell
        link object node to its first segment node in the same cell
        link main grid cell to its first object node
    loop
loop

[Sync memory]

//Object sorting pass
for (every main grid cell)
    sort object headers in the stream back-to-front
loop

[Sync memory]
```

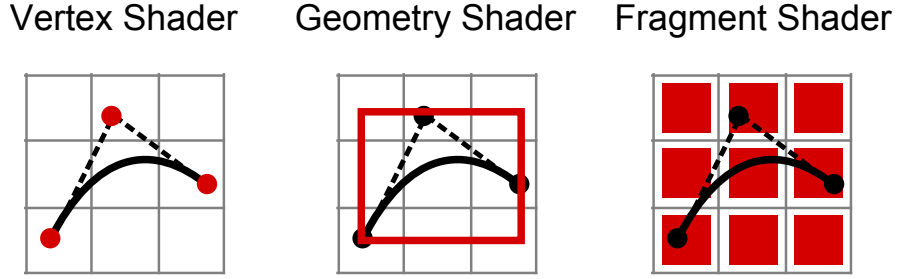


Figure 15: Various programmable stages of the graphics pipeline collaborate in encoding of a quadratic curve segment.

Despite the output of the encoding stage being used for rendering purposes, the encoding algorithm itself cannot be modeled directly by the traditional graphics pipeline. Like [NH08], we use a GPGPU approach rather than a fully generic GPU Computing approach, meaning the steps of the algorithm are expressed in terms of graphical constructs. The main reason for this choice is that the GPU Computing is still rapidly evolving, with tools maturing, and that to our knowledge the GPGPU approaches still perform better.

In the GPU implementation of the algorithm the steps of every loop actually run in parallel. The input data is encoded into graphical primitives and the drawing calls are issued so as to trigger multiple invocations of the shader programs for every element of the input. Parallelism is achieved by distributing the computation between multiple shader processors on the GPU with regard to segments, control points or grid cells. Figure 15 illustrates the distribution of computation across the various programmable stages of the graphics pipeline. Vertex shaders process multiple control points, geometry shaders find the bounding box of multiple segments, and fragment shaders encode the segments into multiple cells covered by their bounding box in parallel.

Every top-level loop in our encoding algorithm represents one graphics pipeline pass. The input required by every pass is represented with the smallest number of vertices that define the respective geometry, in order to minimize the data transfer between the main and the GPU memory. Objects are represented by a line primitive extending from the minimum point to the maximum point of the object's bounding box. Straight line segments are represented directly by a line primitive. Quadratic curve segments are represented by a triangle which encodes the three control points of the curve. Note that these input primitives are not processed by the graphics pipeline in the traditional way. Instead, we are taking advantage of the programmable stages in the pipeline to decode the input data, process it as needed, and then encode the results so they can be passed to the next stage of the pipeline.

Thanks to recent advances in graphic hardware, described in Section 2.7, the algorithm can interact with the intermediate and output memory buffers directly. However, due to multiple levels of cache the concurrent memory access might result in memory discrepancies as seen by the processors from different processing groups. To avoid this, an explicit function call needs to be issued to synchronise the memory across all the processors. It is important to note that none of the points of memory synchronisation are inside a loop, which means the number of required synchronisations is fixed and does not increase with the complexity of the input data.

This was especially challenging to achieve on the level of objects when an image consists of multiple overlapping shapes of different colors. The encoded segment lists for each object need to be kept separate within each cell, so that the inside/outside classification in the rendering stage can be done on per-object basis. The following sections describe the segment list construction in detail, including our solution to the problem of segment list separation.

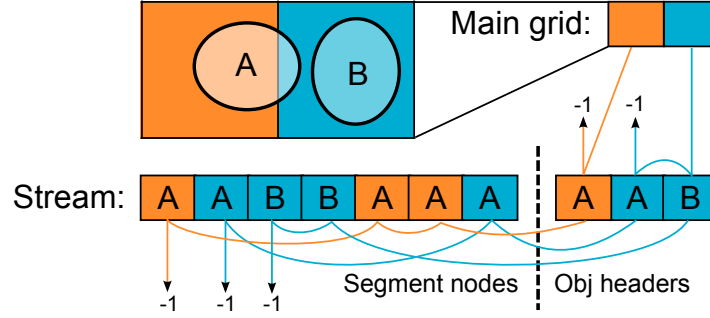


Figure 16: The stream buffer contains interleaved nodes of multiple linked lists. Each linked list contains segments of one particular object intersecting one particular grid cell.

4.3.2 Segment List Construction

The output of our encoding algorithm is written into the stream buffer. The data in the stream buffer is organized into nodes of different types, inter-linked via the stream index of the first data in the node. Different nodes have different number of data values and therefore different sizes. The first value in every node defines the type of the node. The second value in every node is the stream index of the previous node in the same list. The following values in the node encode additional parameters, depending on the node type. A stream index of value -1 indicates the end of a linked list. Such organisation allows us to interleave multiple lists with nodes of different types and sizes within the same memory space.

We use three types of nodes in our localised representation. A line segment node encodes the two control points of a linear segment. A quadratic segment node encodes the three control points of a quadratic curve segment. An object header node encodes the ID of the object to which a list of segments belongs. This ID represents the back-to-front order of the object within the vector image and is also used in the rendering stage to look up object properties (such as colour) in the object info buffer.

There might be multiple objects intersecting the same cell and we need to be able to distinguish which segments belong to which object. To this end, we create a separate segment list for every object intersecting a cell and connect all the segment lists of a cell via another linked list containing object header nodes. We also create an indirection grid which contains pointers to the beginning of the respective object header list for every cell. Figure 16 shows an example of one possible final structure of the stream buffer.

The stream node insertion algorithm is given in the following pseudocode:

```

I = atomicAdd( L, S )
P = atomicExchange( H, I )
stream[ I ] = node type
stream[ I+1 ] = P
stream[ I+...] = additional node data...

```

where L is the atomic counter holding the total stream length, S is the size of the new node as number of stream words, I is the stream index of the new node in the list, H is the atomic counter holding the stream index of the last node in the list, and P is the stream index of the previous node in the list. Function *atomicAdd* adds the given value to an atomic counter and function *atomicExchange* swaps

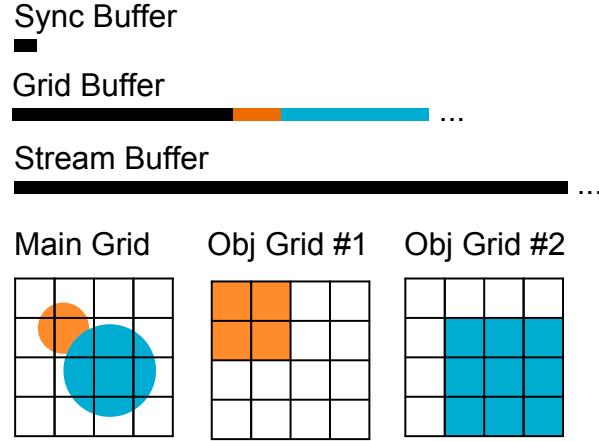


Figure 17: Three buffers used by the encoding algorithm. The main indirection grid and the stream buffer are the input to the rendering pass.

the value of the atomic counter with the new value. Both atomic functions return the old value read prior to the modification.

Atomic counters L and H are the key to synchronisation of the encoding process between multiple shading processors. These counters are stored in the intermediate buffers shown in Figure 17. The synchronisation buffer holds a single global stream length counter L , while every cell of the grid buffer holds a list head counter H . The grid buffer is formatted so as to contain the cells of the main grid, followed by the cells of every object grid in row major order. The size of an object grid is determined by the bounding box of the respective object, aligned to the cells of the main grid.

Since all the data is interleaved in a single stream buffer, only one L counter is used. On the contrary, the use of multiple H counters allows us to construct multiple linked lists simultaneously. The segment encoding pass uses atomic counter H in the grid of the object to which the current segment belongs. The object encoding pass uses the atomic counter H in the main grid, but reads the H counter from the respective object grid cell and stores it into the object header node.

After the object header encoding pass is done, all the segments within a grid cell can be accessed by reading the indirection index H from the main grid cell and then iterating through the object header list and every segment sub-list as shown in Figure 16. The final back-to-front sorting pass iterates through the object header list for every cell and sorts the contents of the object header nodes based on the object ID.

4.3.3 Pivot-Point Localisation

In the PP version of our encoding algorithm, the segment encoding pass performs an additional task of computing the exact winding number for the pivot points of every object grid cell. To achieve this, we need to compute additional intersections between every segment and the horizontal rays cast from the pivot points to its left.

The computation of these intersections is parallelised by triggering additional shader invocations, not only for the cells covered by the segment, but also for every cell to its left. Each shader invocation computes the ray-segment intersection for the respective pivot point and updates the winding number counter in the object grid cell with an atomic operation. When the object header encoding pass begins, the grid cell counters contain the final winding number of their pivot point. This winding number is

stored into the object header and used in the rendering pass to offset the winding number obtained with segment intersections between the pivot point and an arbitrary point inside the cell.

4.3.4 Auxiliary-Segment Localisation

The AS version of our encoding algorithm is essentially a parallel version of the “fast lattice-clipping algorithm” as described in [NH08] with minor adjustments to improve the parallelism. The segment encoding pass of our algorithm performs the cell boundary intersection tests and computes the change Δh in winding number. The object header encoding pass performs insertion of auxiliary vertical segments based on the final winding number sum h .

We use the same boundary conditions to insert auxiliary vertical segments on the right edges of the cells. As discussed at the start of Section 4, computation of the running sum of Δh itself cannot be efficiently parallelised. However, within our parallel algorithm we can still gain some advantage by not calculating it at once in the object encoding pass. Instead, we update the sum across all the cells in the same row to the left of the current cell in the segment encoding pass as soon as an intersection with a segment is found. Although more work is done in total this way, it proved to work better in parallel, achieving a noticeable speed-up.

The reason for this is that the complexity of the segment encoding pass is much higher than the complexity of the final object header encoding pass. Since the Δh is read and written concurrently from multiple shaders, it needs to be accessed through atomic operations. The result of every atomic operation needs to be synchronised between the local caches of multiple shading processor groups. Due to the higher complexity, there is more time between invocations of the segment encoding shader to perform the synchronisation, before the result needs to be read by another invocation, so the impact on the performance of the entire rendering pass is much lower.

When the object header encoding pass begins, the grid cell counters already contain the final winding number sum h , which can then be used directly to insert additional auxiliary segments, before the linked list is closed with an object header.

4.4 Rendering Algorithm

Cell stream data generated by the encoding algorithm can be used at any time during rendering to perform an efficient colour lookup at any coordinate in image space. An encoded vector image can thus be used in the same way as a traditional raster texture. To display the image in its original appearance, a rectangle of matching size with appropriate vertex texture coordinates can be rendered. Alternatively, the image can be mapped onto an arbitrary surface, even under perspective transformation, as shown in Figure 18.

4.4.1 Random Access Lookup

The lookup process begins by transforming the image space texture coordinate into grid space. When the target grid cell is determined, the pointer to the beginning of the stream is read. In this sense, the grid acts as an indirection table. After obtaining the stream pointer, the nodes in the linked list are processed one after another.

Every line and quadratic curve node is tested for intersection and the winding number updated when a valid intersection is found. If the final winding number classifies the point as inside the shape according to the fill rule in use, the colour of the shape is returned, otherwise the current fragment is discarded.

Note that the segments in the cell stream might belong to multiple shapes. The encoding algorithm only discards segments from a cell when entirely occluded by another shape, so a particular point might still be determined to be inside two different shapes. Since the shapes in the stream are sorted back to front, the color of the last occluding shape overwrites the previous one, but only if it is fully opaque. When a semi-transparent shape is encountered, the colour is blended with the current value.

4.4.2 Antialiasing

As discussed in [NH08] a localised image representation offers additional benefit of being able to perform adaptive supersampling based on the amount of detail in a cell. For instance, when a cell is entirely filled with one solid colour the lookup can revert to a single sample. Additionally, the cell stream data can be read from the GPU memory only once and shared by all the samples.

However, we found that this approach only works correctly when the transformation from the screen space into texture space is known within the fragment shader performing the lookup and the samples of a pixel actually fall into the same localisation cell. This is only the case when the image is being rendered onto a flat surface or under known texture coordinate perturbation and not when the image is being mapped onto an arbitrary surface under perspective transformation. In the latter case, adjacent samples of the same pixel might fall onto a different primitive of the textured object and require a lookup from a different grid cell.

To overcome these issues, we resorted to using a multi-sampling approach, which has good support in modern hardware, and evaluate the texture lookup at every sample. This approach is completely generic and offers correct antialiasing in every scenario, but cannot take advantage of the potential optimisations described above.

5 Results and Analysis

The goal of our experiments was to test the performance of every stage in our image pipeline and compare it between the various approaches. We performed a number of experiments stressing the encoding and the rendering stage in combination with the pivot-point and the auxiliary-segment localised image representations. All results were obtained using Intel Core i7 2.80 GHz CPU with an NVIDIA GTX460 1GB GPU.

5.1 Experiment Design

Two images were used in our experiments: the tiger image, which has long been the traditional subject of benchmarking for vector graphics, as well as a block of text containing variable numbers of glyphs. Both images are shown in Figure 18.

The tiger image features many larger and smaller shapes of high complexity. For this reason, it is particularly good for testing the effect of grid resolution on the encoding and rendering performance. Moreover, the shapes occluding each other present a good test for the part of our encoding algorithm which discards segments from a cell when they are entirely occluded by another shape in front.

The text image served a different purpose. It was mainly used to see how the encoding and rendering performance scales with the amount of input data. The glyphs of the text are fairly spread out and do not occlude each other. The details are thus regularly spread across the grid, eliminating the unfavorable scenario of unbalanced stream lengths between the grid cells. It is also easy to control the amount of data by adding or removing characters from the text.

Firstly, we wanted to compare the performance of our parallel GPU encoding algorithm against a sequential CPU approach. Even though modern CPUs include multiple processing cores, our CPU implementation only uses one thread of execution to execute the loops of our encoding algorithm iteratively. Of course, our parallel algorithm could probably be implemented with multiple threads on a CPU, but the purpose of this test was to compare a parallel processing approach against a sequential approach. As a matter of fact, it might be interesting to compare the performance of two parallel implementations, one on the CPU and one on the GPU, but this was left for future work.

We also wanted to see which of the two localised image representations can be encoded faster with the CPU and the GPU approach and how the grid resolution affects the encoding performance. Our experiment #1 includes measurement of achievable encoding rate (i.e. number of executions per second) for the tiger image, at various levels of grid resolution.

Secondly, we wanted to test the random-access rendering performance using the two different image representations (PP and AS) and compare it to the performance of a traditional forward rendering approach. Experiment #2 involves measurement of achievable rendering rate at various levels of grid resolution, for the same tiger image.

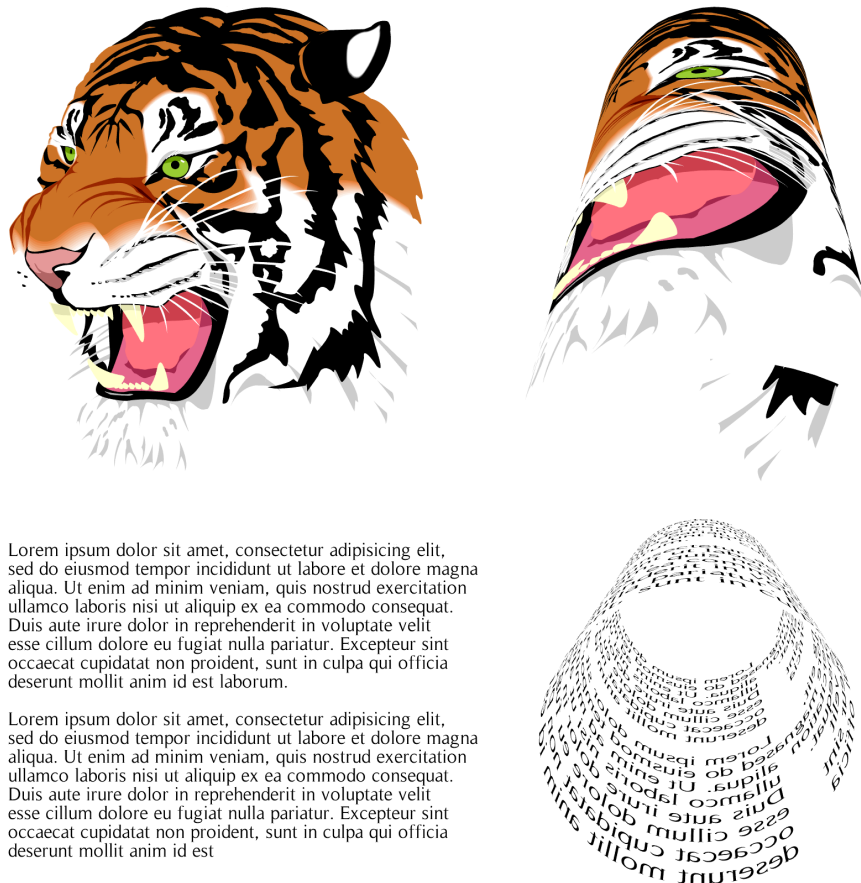


Figure 18: Tiger and text vector images used in our tests. On the left, the images rendered onto a flat surface. On the right, the images mapped onto a curved surface of a cylinder.

Next, we wanted to simulate a scenario where the vector image is animated, in which case the encoding needs to be repeated on every frame. Unfortunately, our time limits did not permit us to implement the actual process of the animation of a vector image. Nonetheless, we could still evaluate the effective performance by measuring the combined encoding and rendering framerate. Again, we used the tiger image and repeated experiment #3 at various grid resolutions, in order to find the optimal balance between the encoding and rendering speed.

The first three experiments test the performance as a function of grid resolution. In order to see how other factors affect the performance, we designed two more experiments. Experiment #4 measures the rendering speed of the tiger image at various levels of zoom, to show how the performance scales with the number of coloured pixels in the resulting raster. For the random-access rendering approach this means more texture lookups. For the forward rendering approach this means larger area of rasterization.

We used the text image in experiment #5 to show how the rendering and encoding performance scales with the increasing geometric complexity. We measured the framerate of the random-access rendering and the forward-rendering approach while adding and removing characters from the text. In order to expose the geometric adaptability of the random-access approach, we adjusted the grid resolution every time the number of characters changed, so as to maintain a steady number of segments per cell. We also tested performance scaling with geometry in an animated image scenario, by measuring the combined performance of encoding and rendering with the random-access approach.

Finally, in experiment #6 we measured the resulting memory size of the encoded PP and AS image representations at various grid resolutions. Since a vector image allows rendering at any level of scale, it is hard to perform any meaningful comparison in memory size with a raster image. Still, to put the memory size in perspective, we included a comparison with the memory requirement of a raster image of the same pixel size as the vector image at 1:1 scale.

In the following sections we analyse the results obtained in every experiment.

5.2 Experiment #1: Encoding Performance

As seen in Figure 19 both parallel GPU encoding approaches performed significantly better than the sequential CPU approaches, offering 2 to 3 times increase in encoding speed. Even at higher grid resolutions, when the CPU performance starts dropping below rate of 30 frames per second, the GPU encoding remains well above the interactive threshold.

According to our expectations, the CPU was faster at encoding the AS representation, while the GPU encoding performed better with the PP representation. Maintaining the auxiliary segment counter is computationally simpler compared to the calculation of ray-curve intersections at pivot points, thus resulting in faster sequential execution. However, as previously discussed, this same step reduces the efficiency of our parallel approach when the AS representation is used.

Surprisingly, though, the difference in GPU performance between the PP and AS representation was not very large. This hints that the atomic operations might still be the bottleneck of our parallel approach, even with the PP representation. Hopefully, with further improvements in hardware, the effect of the atomic operations on the overall performance will be reduced.

5.3 Experiment #2: Rendering Performance

Figure 20 plots the rendering speed as a function of grid resolution. At higher grid resolutions the random-access rendering approach performed more than twice as fast as the traditional forward rendering approach. Due to lower number of segments per cell, our PP localised representation achieved

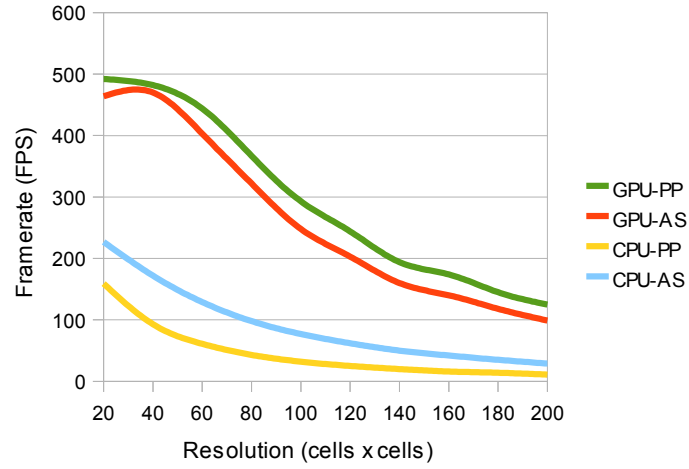


Figure 19: Encoding speed as a function of grid resolution for the tiger image in Figure 18.

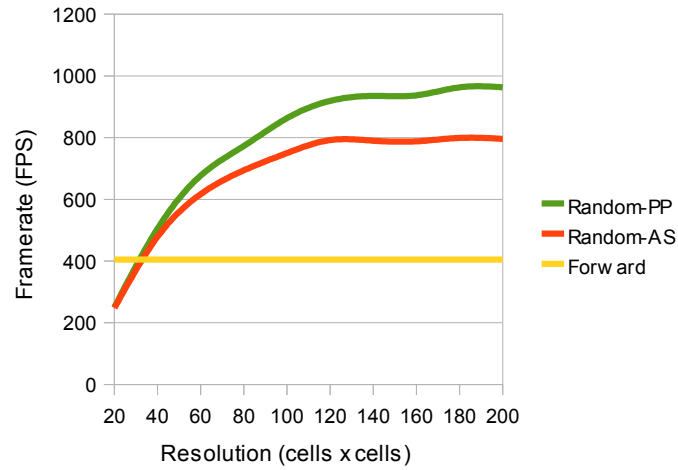


Figure 20: Rendering speed as a function of grid resolution for the tiger image in Figure 18.

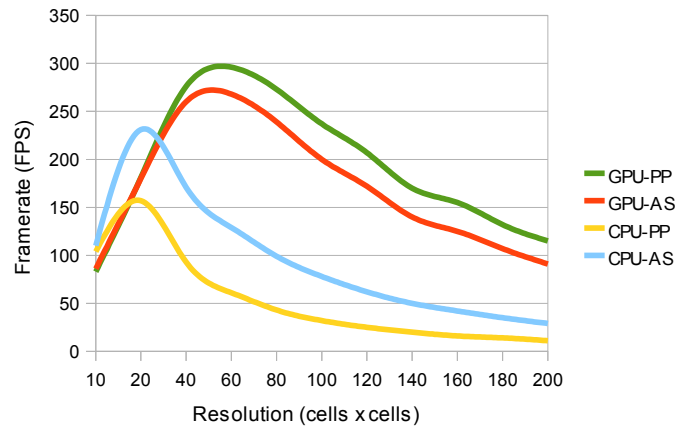


Figure 21: Encoding + rendering speed as a function of grid resolution for the tiger image in Figure 18.

additional 25% speedup compared to the AS representation.

Obviously, the forward rendering approach is not affected by the grid resolution, so it shows as a constant value. With the random-access rendering approach, the framerate sharply increases at lower resolutions but eventually reaches a plateau. At that point the resolution of the grid becomes high enough that further increase in resolution does not result in any significant decrease in the number of segments per cell. Obviously, the resolution at which this occurs varies with the level of detail in the image, so the optimal resolution depends on the image being rendered.

It could be worth investigating techniques to estimate the optimal grid resolution for a given image in advance, in order to achieve the optimal rendering performance. Unfortunately, our time limits did not permit us to investigate such techniques.

5.4 Experiment #3: Encoding + Rendering Performance

The differences in encoding and rendering performance observed in experiments #1 and #2 clearly affect the combined performance of the two stages, as shown in Figure 21. Our parallel GPU encoding approach which performed significantly better than the sequential CPU approach in the encoding stage, in combination with our PP localised representation which performed best in the rendering stage, achieved the highest combined performance of the two stages.

Similar to the rendering framerate, the combined framerate of the two stages sharply increases at lower resolutions. However, at a certain point the decrease in encoding performance outweighs the gains in the rendering performance and the combined framerate starts to decrease. Both the CPU and the GPU data lines show a clear point at which the balance between encoding and rendering speed is optimal.

For the parallel encoding approach, the peak in performance occurred at the resolution of 60×60 cells. With the same vector image, the rendering performance in experiment #2 reached a plateau at the resolution of 120×120 cells. Optimally, these two resolutions would coincide, allowing the combined performance to scale with the grid resolution to the point where the rendering performance cannot increase anymore.

This could be achieved by further improving the encoding speed to perform better at higher resolutions. Alternatively, the same could be achieved by improving the gains in rendering speed with the increase in grid resolution. As future work, it might worth investigating the use of irregular spatial data structures, to achieve better distribution of image detail across the localised lookup structure and improve the rendering performance gains.

5.5 Experiment #4: Performance Scaling with Raster Size

Figure 22 plots the framerate as a function of image size on the screen after scaling. The higher the scaling factor, the more pixels in the target raster are covered by the image. This affects the two rendering approaches in a different way.

With the forward rendering approach, a larger scale results in larger triangle primitives, and thus more work for the rasterization stage of the graphics pipeline. Regardless of the scale, the entire geometry of the image always needs to be processed and therefore the load on the vertex transformation and primitive assembly stages remains constant. In the case of the tiger image, the geometric complexity is so high that it significantly outweighed rasterization and there was no noticeable slowdown at larger levels of scale.

On the other hand, the random-access rendering approach behaves more similar to texturing using a raster image; at large enough levels of grid resolution, the geometric complexity is removed from

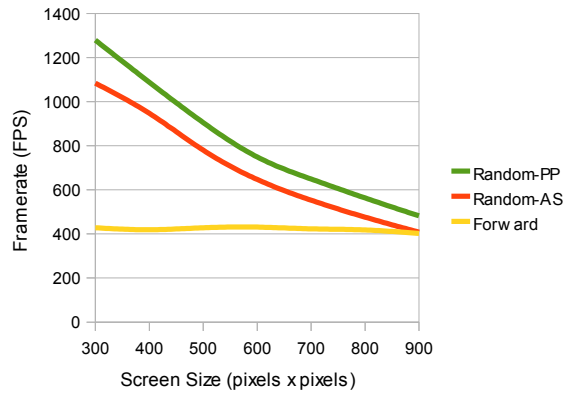


Figure 22: Rendering speed as a function of actual image screen size for the tiger image in Figure 18.

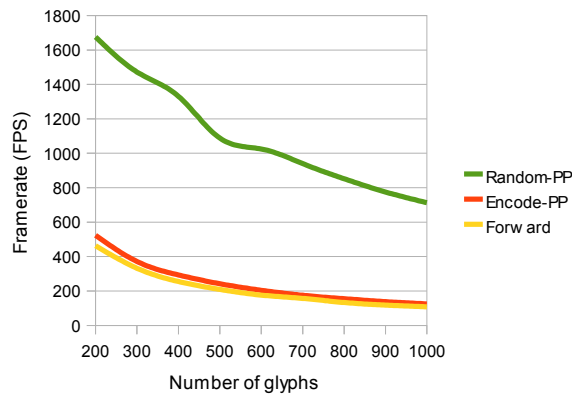


Figure 23: Rendering speed as a function of number of glyphs for the text image in Figure 18.

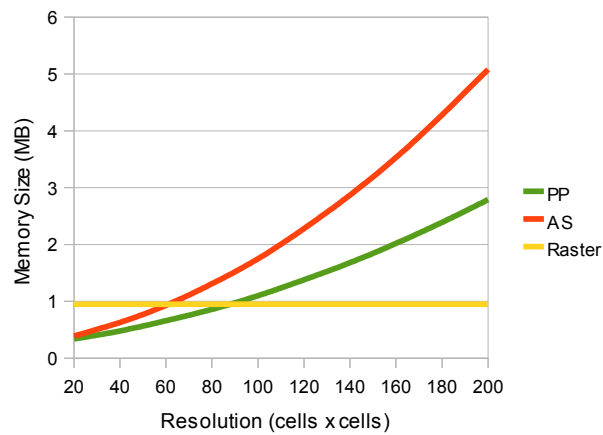


Figure 24: Total memory size of the stream data as a function of grid resolution for the tiger image in Figure 18.

the rendering stage and the number of textured pixels becomes the most significant performance factor. The results in Figure 22 were achieved with a grid resolution of 120×120 cells. Despite the high resolution, the per-pixel cost of performing the lookup into the localised image representation turned out to be higher than that of the traditional forward rendering rasterization, and the framerate noticeably decreased as the image was scaled. At the screen size of 300×300 pixels, the random-access rendering performed at around 1200 frames per second, essentially tripling the forward rendering performance. As the screen size increased towards 900×900 pixels, the random-access rendering speed linearly decreased, eventually matching the forward rendering speed of 400 frames per second.

Unfortunately, we couldn't measure performance reliably at larger screen sizes, as we were limited with the size of our display device. By extrapolating the obtained information, we expect that the forward rendering approach might be more appropriate to use with very large display resolutions, as long as the type of application does not require additional capabilities of random-access rendering.

5.6 Experiment #5: Performance Scaling with Geometric Complexity

In this experiment, we rendered the text image, while changing the number of characters in the text, and therefore the number of vector glyph objects in the image. As opposed to image screen size in experiment #4, the increasing geometric complexity of the image noticeably affected the forward rendering technique. Figure 23 shows an exponential decrease in forward rendering framerate as the number of glyphs increases.

When testing the random-access rendering technique, we took advantage of the adaptable capabilities of the localised image representation. We adjusted the grid resolution dynamically according to the number of characters in the text, so that each glyph object covered approximately a 2×2 block of cells. The random-access approach significantly outperformed the forward rendering approach at any number of characters, achieving a 3 to 4 times factor of improvement in rendering speed.

It is also interesting to note that the performance of random-access encoding and rendering combined almost perfectly matched traditional rendering of the text image, which means animated text can be rendered with equal efficiency using both techniques.

5.7 Experiment #6: Memory Requirements

Figure 24 shows the total memory size of the cell stream data for the tiger image, at various grid resolutions. This is the memory space taken by the localised representation of the image which needs to be kept in the GPU memory for continuous rendering.

Despite the similarity in approach, the memory sizes measured in our experiment are generally higher than those of [NH08]. This is mainly due to two important design choices. Firstly, the straight-forward approximation of cubic curves with a fixed number of four quadratic segments greatly increases the total number of segments in the encoded image. This could be much improved by implementing a more flexible adaptive approximation. Secondly, due to the parallelism in our cell stream construction, the segments might be encoded into the stream out of order, so the end and start point of two consecutive segments cannot be shared and needs to be encoded twice in the node of each segment.

Nonetheless, considering the diminished gains in rendering performance at large grid resolutions, the memory requirements of our encoded vector image do not gravely exceed those of a traditional raster image. The pixel size of the tiger vector image used in this experiment at 1:1 scale is approximately 500×500 pixels. At this size a 32-bit raster image requires roughly 1MB of memory space. At the grid size of 120×120 cells, when the rendering performance with the same image reaches a

plateau, our pivot-point representation barely exceeds that memory size. Of course, this comparison is not necessarily meaningful, as the two image representations are qualitatively different, but at least puts the results of this experiment in perspective.

It is also worth noting, that the memory size of our pivot-point representation grows significantly slower than the memory size of the auxiliary-segment representation, since the fully opaque cells do not require an auxiliary segment to mark them as interior.

5.8 Limitations of Our Approach

The localisation of a vector image to cells of a regular grid reduces the number of segments to be intersected in one texture lookup as long as the level of detail is fairly uniform across the image. However, if the image has large areas of low detail and a few clusters of very high detail, a regular grid becomes less efficient at dividing the detail among the cells. In the worst case, increasing the resolution of the grid will not improve the rendering performance at all, if the details remain within one single cell.

This issue could be solved by using an adaptive spatial data structure, such as a kd-tree. In that regard, our pivot point representation has an additional advantage over auxiliary-point representation of being more generically defined - there are no dependencies between adjacent cells, only a cell and its pivot point. Our localised representation can be easily generalised from rectangular cells to any type of irregular convex regions. The only requirement is that a ray cast from the pivot point towards an arbitrary point within the region cannot intersect its boundary, which would make the winding number depend on the intersections with the segments in the adjacent region. This condition is satisfied in any convex region.

Due to our time constraints, the investigation of such an approach was left for future work.

5.9 Non-traditional Applications

Besides the traditional advantage of scalability, a localised vector image representation offers additional advantage in the form of efficient random-access which broadens the range of applications.

Figure 18 shows the tiger and text images used as textures and mapped onto the surface of a cylinder under perspective transformation. Such applications were previously limited to raster images with the artifacts introduced at larger levels of zoom limiting the flexibility of use. Instead, vector textures could be used in many ways to enhance the user experience in desktop applications as well as virtual worlds.

Vector-based glyphs and other user interface elements can retain sharpness when rendered on a slanted or curved surface at any scale. This allows for much greater flexibility and creativity in user interface design and facilitates development of 3-dimensional desktop interfaces, which can lead to general improvement in screen estate management.

In virtual worlds, raster textures suffer from artifacts when the viewer approaches the textured surface. In order to retain the image clarity, a sufficiently large image has to be used to still display well at the closest distance it is expect to be observed from. This leads to suboptimal memory usage when a textured object is observed from afar. Traditionally, to work around this issue, multiple versions of the same image of various sizes would be used and loaded depending on the viewer's position. In contrast, when a vector texture is used, a single version of the image needs to be loaded in memory to be able to render the textured object with maximum clarity regardless of the viewer's position.

6 Conclusion and Future Work

We have presented a preprocessing and rendering technique for vector images which takes advantage of the parallel architecture of the GPU and allows fast random access rendering of animated vector images. We have shown that our parallel GPU approach outperforms the sequential CPU approach in the preprocessing step, and that our random-access rendering technique in most cases significantly outperforms the traditional forward rendering technique.

We have also presented an alternative pivot-point vector image representation, which allows slightly faster encoding and rendering, but grows significantly slower in memory size compared to auxiliary-segment representation, while approximately matching the memory size of a traditional raster image. It retains the traditional vector image advantage of scalability over raster images, while offering additional advantage in the form of efficient texture mapping, allowing alternative non-traditional applications.

Our experiments showed that the optimal level of image localisation depends on the geometric complexity of the image, and therefore an estimation algorithm based on the input data is required in order to choose the optimal grid resolution for our localised image representation.

In an animated image scenario, our approach achieved a combined performance of the preprocessing and rendering stage well above the interactive rate of 30 frames per second with high-detail images, as well as reasonably large sections of text used in typical user interfaces.

Due to time constraints, our implementation does not support all the features of a typical vector image. Moreover, there are areas where alternative approaches could provide further performance gains. Avenues for future work include:

1. Rendering of wide strokes by approximation with filled paths
2. Colouring shape interior based on a linear or radial gradient
3. Use of adaptive rather than fixed approximation of cubic segments with quadratic segments
4. Implementation of efficient ray intersection with cubic segments to avoid approximation entirely and reduce stream memory size
5. Investigation of optimal grid resolution estimation techniques
6. Generalisation of pivot-point representation to irregular spatial data structures

References

- [BM99] Wolfgang Boehm and Andreas Miller. On de casteljau's algorithm. *Computer Aided Geometric Design*, 16(7):587 – 605, 1999.
- [Cud] Cuda. http://www.nvidia.com/object/what_is_cuda_new.html. Online; accessed 31 October 2010.
- [Far02] Gerald Farin. *Curves and surfaces for CAGD: a practical guide*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 5th edition, 2002.
- [Fer] Next generation cuda architecture, code named fermi. http://www.nvidia.com/object/fermi_architecture.html. Online; accessed 31 October 2010.
- [KSST06] Yoshiyuki Kokojima, Kaoru Sugita, Takahiro Saito, and Takashi Takemoto. Resolution independent rendering of deformable vector objects using graphics hardware. In *SIGGRAPH '06: ACM SIGGRAPH 2006 Sketches*, page 118, New York, NY, USA, 2006. ACM.
- [LB05] Charles Loop and Jim Blinn. Resolution independent curve rendering using programmable graphics hardware. In *SIGGRAPH '05: ACM SIGGRAPH 2005 Papers*, pages 1000–1009, New York, NY, USA, 2005. ACM.
- [NH08] Diego Nehab and Hugues Hoppe. Random-access rendering of general vector graphics. In *SIGGRAPH Asia '08: ACM SIGGRAPH Asia 2008 papers*, pages 1–10, New York, NY, USA, 2008. ACM.
- [Ope] Opencl. <http://www.khronos.org/opencl/>. Online; accessed 31 October 2010.