SIGGRAPH 98 Course Notes

Subdivision for Modeling and Animation

Organizers: Peter Schröder, California Institute of Technology
Denis Zorin, Stanford University—New York University

Lecturers

Peter Schröder
Caltech Multi-Res Modeling Group
Computer Science Department 256-80
California Institute of Technology
Pasadena, CA 91125
net: ps@cs.caltech.edu

Denis Zorin
Computer Graphics Laboratory
Gates Building, Rm 375
Stanford University
Stanford, CA 94305
net: dzorin@graphics.stanford.edu

Tony DeRose
Studio Tools Group
Pixar Animation Studios
1001 West Cutting Blvd.
Richmond, CA 94804
net: derose@pixar.com

David R. Forsey
Radical Entertainment Ltd.
1052 Homer St.
Vancouver B.C. Canada
V6B 2W9
net: dforsey@radical.ca

Leif Kobbelt
Computer Graphics Group
University of Erlangen
Am Weichselgarten 9
91058 Erlangen, GERMANY
net: kobbelt@informatik.uni-erlangen.de

Michael Lounsbery
Member of Technical Staff
Alias|wavefront
1218 Third Avenue, Suite 800
Seattle, WA 98101
net: louns@aw.sgi.com

Jörg Peters
Department of Computer Science
Purdue University
West Lafayette, IN 47907
net: jorg@cs.purdue.edu

Schedule

Morning Session: Introductory Material The morning section will focus on the foundations of subdivision, starting with subdivision curves and moving on to surfaces. We will review and compare a number of different schemes and discuss the relation between subdivision and splines. The emphasis will be on properties of subdivsion most relevant for applications.

Foundations I: Basic Ideas
Denis Zorin and Peter Schröder

Foundations II: Construction and Analysis
Denis Zorin and Jörg Peters

Afternoon Session: Applications and Algorithms The afternoon session will focus on applications of subdivision and the algorithmic issues practitioners need to address to build efficient, well behaving systems for modeling and animation with subdivision surfaces.

Interactive Multiresolution Mesh Editing
Denis Zorin

Subdivision Surfaces and Wavelets
Michael Lounsbery

A Variational Approach to Subdivision
Leif Kobbelt

Exploiting Subdivision in Modeling and Animation
David R. Forsey

Subdivision Surfaces in the Making of Geri's Game
Tony DeRose

Lecturers' Biographies

Peter Schröder is currently an assistant professor of computer science at the California Institute of Technology, Pasadena, where he directs the Caltech Multi-Res Modeling Group. He received a Master's degree from the MIT Media Lab and a PhD from Princeton University. During 1990/91 he worked for Thinking Machines. From 1992 to 1994 he held a visiting research fellow appointment at the German National Computer Science Research Center, and did postdoctoral work at the University of South Carolina. His research contributions range from massively parallel visualization, to physically based modeling, global illumination, wavelets and multiresolution geometry. For the past 5 years his work has concentrated on exploiting wavelets and multiresolution techniques to build efficient representations and algorithms for many fundamental computer graphics problems. He has taught in a number of Siggraph courses and most recently co-led the course on Wavelets in Computer Graphics (1996). His current research focuses on subdivision as a fundamental paradigm for geometric modeling and rapid manipulation of large, complex geometric models. The results of his work have been published in venues ranging from Siggraph to special journal issues on wavelets and WIRED magazine, and he is a frequent consultant to industry.

Denis Zorin is a research associate at the Computer Science Department of Stanford University. Starting in the fall of 1998, he will be an assistant professor at the Courant Institute of Mathematical Sciences, New York University. He received a BS degree from the Moscow Institute of Physics and Technology, a MS degree in Mathematics from Ohio State University and a PhD in Computer Science from the California Institute of Technology. His research interests include multiresolution modeling, the theory of subdivision, and applications of subdivision surfaces in Computer Graphics. He is also interested in perceptually-based computer graphics algorithms. He has published several papers in Siggraph proceedings.

Tony DeRose is currently a member of the Tools Group at Pixar Animation Studios. He received a BS in Physics in 1981 from the University of California, Davis; in 1985 he received a Ph.D. in Computer Science from the University of California, Berkeley. He received a Presidential Young Investigator award from the National Science Foundation in 1989. In 1995 he was selected as a finalist in the software category of the Discover Awards for Technical Innovation.
From September 1986 to December 1995 Dr. DeRose was a Professor of Computer Science and Engineering at the University of Washington. From September 1991 to August 1992 he was on sabbatical

leave at the Xerox Palo Alto Research Center and at Apple Computer. He has served on various technical program committees including SIGGRAPH, and from 1988 through 1994 was an associate editor of ACM Transactions on Graphics.
His research has focused on mathematical methods for surface modeling, data fitting, and more recently, in the use of multiresolution techniques. Recent projects include object acquisition from laser range data and multiresolution/wavelet methods for high-performance computer graphics.

David Forsey is an assistant professor of Computer Science at the University of British Columbia in Vancouver, Canada, where he co-directs the Imager Computer Graphics Laboratory. He is currently on leave at Radical Entertainment Ltd. developing products based on Hierarchical B-splines. He obtained a BSc. in Zoology at the University of Guelph (1980) and a M.Math (1985) and PhD. (1990) from the University of Waterloo. He has taught in several SIGGRAPH courses, and has had two animations in the SIGGRAPH electronic theatre. His research interests include multiresolution modeling, facial animation as well as simulation and interactive tools as they relate to the artistic side of modeling and animation.

Leif Kobbelt currently holds a position as a post-doctoral research fellow at the University of Erlangen, Germany. His major research interest is sophisticated free-form modeling based on polygonal meshes. He received his master's (1992) and Ph.D. (1994) degrees from the University of Karlsruhe, Germany. He then spent one year at the University of Wisconsin, Madison as a visiting researcher in Carl de Boor's group. Since 1996 he has been working in the geometric modeling unit of the Computer Graphics Group at Erlangen. During the last 5 years he made significant contributions to the construction and analysis of subdivision schemes and pioneered the combination of the subdivision paradigm with variational methods from CAGD.

Michael Lounsbery is on the R&D staff for Alias|wavefront in Seattle. He has a B.S. in Computer Science (1984) from the University of Maryland. In 1994, he completed a Ph.D. in Computer Science from the University of Washington, where he studied the development of wavelets over subdivision surfaces. He has taught in previous SIGGRAPH courses on wavelets, and has also published research in surface interpolation. His other graphics interests include fairness and compression issues for 3D surface modeling.

Jörg Peters is an associate professor at the department of Computer Sciences at Purdue University. He received his PhD in 1990 from the University of Wisconsin-Madison advised by Carl de Boor. In 1991

and 1992 he held positions at the IBM T.J. Watson Research Center and Rensselaer Polytechnic Institute before joining the computer sciences department at Purdue in 1992.

In 1994, Dr. Peters was honored with a five year National Young Investigator Award for his work on Foundations and Tools for Surface Modeling. Currently Dr. Peters heads the SurfLab at CS Purdue and is a member of the Center for Visualization and Image Processing.

Peters' research focuses on representation and analysis of geometry on the computer. Notably, he has developed new tools for free-form modeling and design that combine the advantages of parametric spline representations and subdivision algorithms

Chapter 1

Introduction

Twenty years ago the publication of the papers by Catmull and Clark [2] and Doo and Sabin [3] marked the beginning of subdivision for surface modeling. This year, another milestone occurred when subdivision hit the big screen in Pixar's short "Geri's Game," for which Pixar received an Academy award for "Best Animated Short Film." The basic ideas behind subdivision are very old indeed and can be traced as far back as the late 40s and early 50s when G. de Rham used "corner cutting" to describe smooth curves. It was only recently though that subdivision surfaces have found their way into wide application in computer graphics and computer assisted geometric design (CAGD). One reason for this development is the importance of multiresolution techniques to address the challenges of ever larger and more complex geometry: subdivision is intricately linked to multiresolution and traditional mathematical tools such as wavelets.

Constructing surfaces through subdivision elegantly addresses many issues that computer graphics practitioners are confronted with

- **Arbitrary Topology:** Subdivision generalizes classical spline patch approaches to arbitrary topology. This implies that there is no need for trim curves or awkward constraint management between patches.

- **Scalability:** Because of its recursive structure, subdivision naturally accommodates level-of-detail rendering and adaptive approximation with error bounds. The result are algorithms which can make the best of limited hardware resources, such as those found on low end PCs.

- **Uniformity of Representation:** Much of traditional modeling uses either polygonal meshes or spline patches. Subdivision spans the spectrum between these two extremes. Surfaces can behave

as if they are made of patches, or they can be treated as if consisting of many small polygons.

- **Numerical Stability:** The meshes produced by subdivision have many of the nice properties finite element solvers require. As a result subdivision representations are also highly suitable for many numerical simulation tasks which are of importance in engineering and computer animation settings.

- **Code Simplicity:** Last but not least the basic ideas behind subdivision are simple to implement and execute very efficiently. While some of the deeper mathematical analyses can get quite involved this is of little concern for the final implementation and runtime performance.

In this course and its accompanying notes we hope to convince you, the reader, that in fact the above claims are true!

The main focus or our notes will be on covering the basic principles behind subdivision; how subdivision rules are constructed; to indicate how their analysis is approached; and, most importantly, to address some of the practical issues in turning these ideas and techniques into real applications.

The following 2 chapters will be devoted to understanding the basic principles. We begin with some examples in the curve, i.e., 1D setting. This simplifies the exposition considerably, but still allows us to introduce all the basic ideas which are equally applicable in the surface setting. Proceeding to the surface setting we cover a variety of different subdivision schemes and their properties.

With these basics in place we proceed to the second, applications oriented part, covering algorithms and implementations addressing

- **Interactive Multiresolution Mesh Editing:** This section discusses many of the data structure and algorithmic issues which need to be addressed to realize high performance. The result is a system which allows for interactive, multiresolution editing of fairly complex geometry on PC class machines with little hardware graphics support.

- **Subdivision Surfaces and Wavelets:** This section shows how subdivision is the key element in generalizing the traditional wavelet machinery to arbitrary topology surfaces. The result are a class of algorithms which open up applications such as compression for subdivision surfaces, for example.

- **A Variational Approach to Subdivision:** Most subdivision methods are stationary, i.e., they use a fixed set of rules. They are generally designed to exhibit some order of differentiability. In practice it is often much more important to consider the fairness of the resulting surfaces. Variational subdivision incorporates fairness measures into the subdivision process.

- **Exploiting Subdivision in Modeling and Animation:** One reason subdivision is becoming very popular is that it supports hierarchical editing and animation semantics. This was discovered originally in the traditional spline setup and lead to the development of hierarchical splines. From that technique it is only a small step to multiresolution modeling using subdivision. This section discusses some of the issues in controlling animation hierarchically.

- **Subdivision Surfaces in the Making of Geri's Game:** This section discusses how subdivision surfaces successfully address the needs of very high end production environments. in the process new techniques had to be developed which are detailed in this part of the notes.

Beyond these Notes

One of the reasons that subdivision is enjoying so much interest right now is that it is very easy to implement and very efficient. In fact it is used in many computer graphics courses at universities as a homework exercise. The mathematical theory behind it is very beautiful, but also very subtle and at times technical. We are not treating the mathematical details in these notes, which are primarily intended for the computer graphics practitioners. However, for those interested in the theory there are many pointers to the literature.

These notes as well as other materials such as presentation slides, applets and snippets of code are available on the web at `http://www.multires.caltech.edu/teaching/courses/subdivision/` and all readers are encouraged to explore the online resources. A repository of additional information beyond this course is maintained at `http://www.mrl.nyu.edu/dzorin/subdivision`.

Chapter 2

Foundations I: Basic Ideas

In this chapter we focus on the 1D case to introduce all the basic ideas and concepts before going on to the 2D setting. Examples will be used throughout to motivate these ideas and concepts. We begin initially with an example from interpolating subdivision, before talking about splines and their subdivision generalizations.

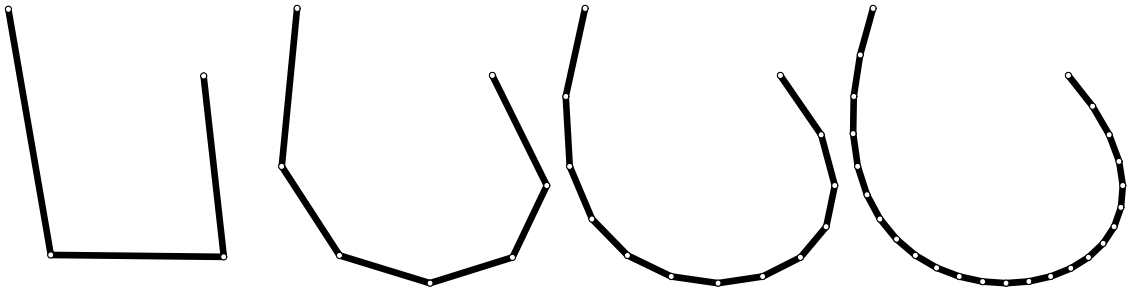


Figure 2.1: Example of subdivision for curves in the plane. On the left 4 points connected with straight line segments. To the right of it a refined version: 3 new points have been inserted “inbetween” the old points and again a piecewise linear curve connecting them is drawn. After two more steps of subdivision the curve starts to become rather smooth.

2.1 The Idea of Subdivision

We can summarize the basic idea of subdivision as follows:

> Subdivision defines a smooth curve or surface as the limit of a sequence of successive refinements.

Of course this is a rather loose description with many details as yet undetermined, but it captures the essence.

Figure 2.1 shows an example in the case of a curve connecting some number of initial points in the plane. On the left we begin with 4 points connected through straight line segments. Next to it is a refined version. This time we have the original 4 points and additionally 3 more points "inbetween" the old points. Repeating the process we get a smoother looking piecewise linear curve. Repeating once more the curve starts to look quite nice already. It is easy to see that after a few more steps of this procedure the resulting curve would be as well resolved as one could hope when using finite resolution such as that offered by a computer monitor or a laser printer.

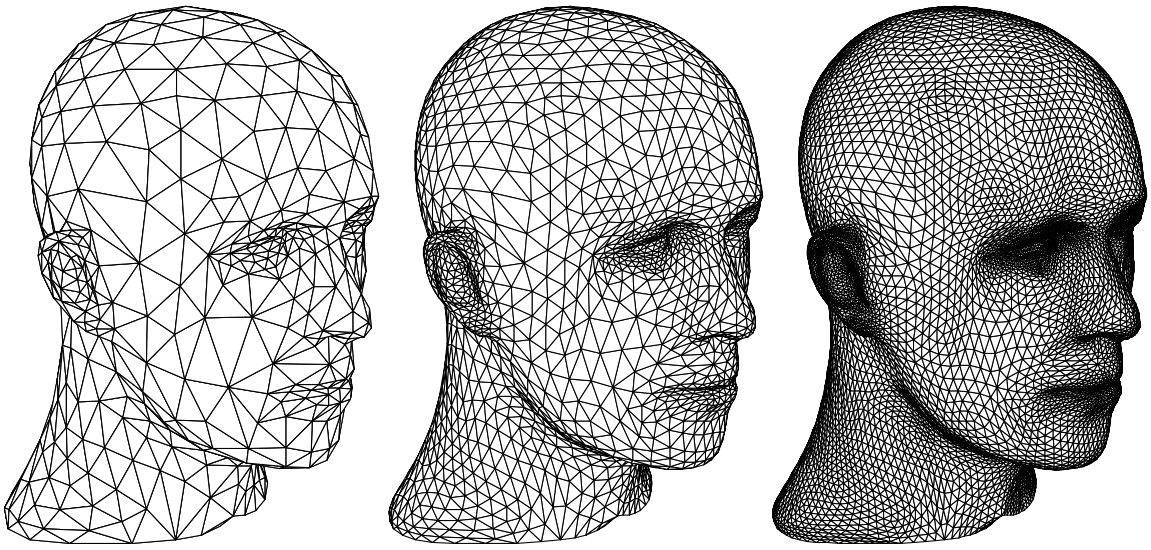


Figure 2.2: Example of subdivision for a surface, showing 3 successive levels of refinement. On the left an initial triangular mesh approximating the surface. Each triangle is split into 4 according to a particular subdivision rule (middle). On the right the mesh is subdivided in this fashion once again.

An example of subdivision for surfaces is shown in Figure 2.2. In this case each triangle in the original mesh on the left is split into 4 new triangles quadrupling the number of triangles in the mesh. Applying the same subdivision rule once again gives the mesh on the right.

Both of these examples show what is known as interpolating subdivision. The original points remain undisturbed while new points are inserted. We will see below that splines, which are generally not interpolating, can also be generated through subdivision. Albeit in that case new points are inserted *and* old points are moved in each step of subdivision.

How were the new points determined? One could imagine many ways to decide where the new points should go. Clearly, the shape and smoothness of the resulting curve or surface depends on the chosen rule. Here we list a number of properties that we might look for in such rules:

- **Efficiency:** the location of new points should be computed with a small number of floating point operations;
- **Compact support:** the region over which a point influences the shape of the final curve or surface should be small and finite;
- **Local definition:** the rules used to determine where new points go should not depend on "far away" places;
- **Affine invariance:** if the original set of points is transformed, e.g., translated, scaled, or rotated, the resulting shape should undergo the same transformation;
- **Simplicity:** determining the rules themselves should preferably be an offline process and there should only be a small number of rules;
- **Continuity:** what kind of properties can we prove about the resulting curves and surfaces, for example, are they differentiable?

For example, the rule used to construct the curve in Figure 2.1 computed new points by taking a weighted average of nearby old points: two to the left and two to the right with weights $1/16\,(-1,9,9,-1)$ respectively (we are ignoring the boundaries for the moment). It is very efficient since it only involves 4 multiplies and 3 adds (per coordinate); has compact support since only 2 neighbors on either side are involved; its definition is local since the weights do not depend on anything in the arrangement of the points; the rule is affinely invariant since the weights used sum to 1; it is very simple since only 1 rule is used (there is one more rule if one wants to account for the boundaries); finally the limit curves one gets by repeating this process ad infinitum are C^1.

Before delving into the details of how these rules are derived we quickly compare subdivision to other possible modeling approaches for smooth surfaces: traditional splines, implicit surfaces, and variational surfaces.

1. **Efficiency:** Computational cost is an important aspect of a modeling method. Subdivision is easy to implement and is computationally efficient. Only a small number of neighboring old points are used in the computation of the new points. This is similar to knot insertion methods found in spline modeling, and in fact many subdivision methods are simply generalization of knot insertion. On the other hand implicit surfaces, for example, are much more costly. An algorithm such as marching cubes is required to generate the polygonal approximation needed for rendering. Variational surfaces can be even worse: a global optimization problem has to be solved each time the surface is changed.

2. **Arbitrary topology:** It is desirable to build surfaces of arbitrary topology. This is a great strength of implicit modeling methods. They can even deal with *changing* topology during a modeling session. Classic spline approaches on the other hand have great difficulty with control meshes of arbitrary topology. Here, "arbitrary topology" captures two properties. First, the topological genus of the mesh and associated surface can be arbitrary. Second, the structure of the graph formed by the edges and vertices of the mesh can be arbitrary; specifically, each vertex may be of arbitrary degree.

   These last two aspects are related: for example, any control mesh for a sphere (topological genus 0) necessarily has vertices of different degrees. Only closed surfaces of genus 1 can be represented using a mesh with all vertices having the same degree.

   When rectangular spline patches are used in arbitrary control meshes, enforcing higher order continuity at extraordinary vertices becomes difficult and considerably increases the complexity of the representation (see Figure 2.3 for an example of points not having valence 4). Implicit surfaces can be of arbitrary topological genus, but the genus, precise location, and connectivity of a surface are typically difficult to control. Variational surfaces can handle arbitrary topology better than any other representation, but the computational cost can be high. Subdivision can handle arbitrary topology quite well without losing efficiency; this is one of its key advantages. Historically subdivision arose when researchers were looking for ways to address the arbitrary topology modeling challenge for splines.

3. **Surface features:** Often it is desirable to control the shape and size of features, such as creases, grooves, or sharp edges. Variational surfaces provide the most flexibility and exact control for creating features. Implicit surfaces, on the other hand, are very difficult to control, since all modeling is performed indirectly and there is much potential for undesirable interactions between different parts of the surface. Spline surfaces allow very precise control, but it is computationally expen-

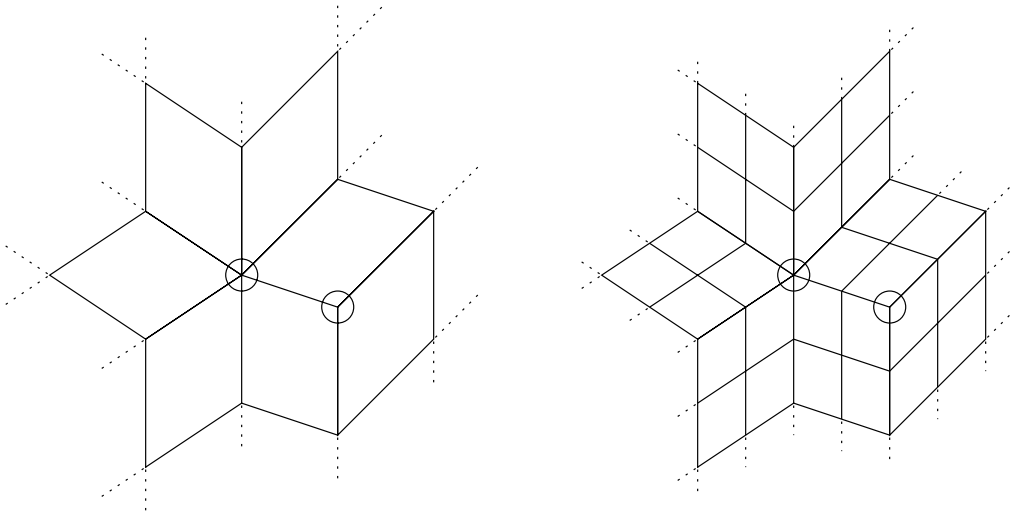


Figure 2.3: A mesh with two extraordinary vertices, one with valence 6 the other with valence 3. In the case of quadrilateral patches the standard valence is 4. Special efforts are required to guarantee high order of continuity between spline patches meeting at the extraordinary points; subdivision handles such situations in a natural way.

sive and awkward to incorporate features, in particular if one wants to do so in arbitrary locations. Subdivision allows more flexible controls than is possible with splines. In addition to choosing locations of control points, one can manipulate the coefficients of subdivision to achieve effects such as sharp creases or control the behavior of the boundary curves.

4. **Complex geometry:** For interactive applications, efficiency is of paramount importance. Because subdivision is based on repeated refinement it is very straightforward to incorporate ideas such as level-of-detail rendering and compression for the internet. During interactive editing locally adaptive subdivision can generate just enough refinement based on geometric criteria, for example. For applications that only require the visualization of fixed geometry, other representations, such as progressive meshes, are likely to be more suitable.

Since most subdivision techniques used today are based upon and generalize splines we begin with a quick review of some basic facts of splines which we will need to understand the connection between splines and subdivision.

2.2 Review of Splines

2.2.1 Piecewise Polynomial Curves

Splines are piecewise polynomial curves of some chosen degree. In the case of cubic splines, for example, each polynomial segment of the curve can be written as

$$\begin{aligned} x(t) &= a_3^i t^3 + a_2^i t^2 + a_1^i t + a_0^i \\ y(t) &= b_3^i t^3 + b_2^i t^2 + b_1^i t + b_0^i, \end{aligned}$$

where $(\mathbf{a}, \mathbf{b})$ are constant coefficients which control the shape of the curve over the associated segment. This representation uses monomials (t^3, t^2, t^1, t^0), which are restricted to the given segment, as basis functions.

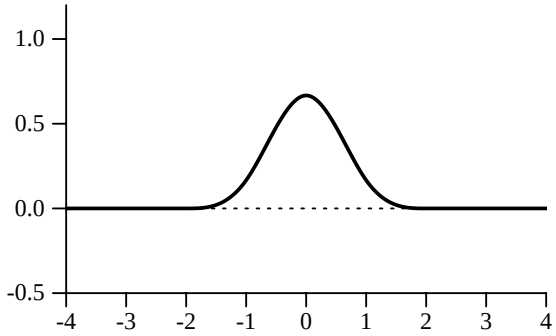


Figure 2.4: Graph of the cubic B-spline. It is zero for the independent parameter outside the interval $[-2, 2]$.

Typically one wants the curve to have some order of continuity along its entire length. In the case of cubic splines one would typically want C^2 continuity. This places constraints on the coefficients $(\mathbf{a}, \mathbf{b})$ of neighboring curve segments. Manipulating the shape of the desired curves through these coefficients, while maintaining the constraints, is very awkward and difficult. Instead of using monomials as the basic building blocks, we can write the spline curve as a linear combination of shifted *B-splines*, each with a coefficient known as a *control point*

$$\begin{aligned} x(t) &= \sum x_i B(t-i) \\ y(t) &= \sum y_i B(t-i). \end{aligned}$$

The new basis function $B(t)$ is chosen in such a way that the resulting curves are always continuous and that the influence of a control point is local. One way to ensure higher order continuity is to use basis

functions which are differentiable of the appropriate order. Since polynomials themselves are inifinitely smooth, we only have to make sure that derivatives match at the points where two polynomial segments meet. The higher the degree of the polynomial, the more derivatives we are able to match. We also want the influence of a control point to be maximal over a region of the curve which is close to the control point. Its influence should decrease as we move away along the curve and disappear entirely at some distance. Finally, we want the basis functions to be piecewise polynomial so that we can represent any piecewise polynomial curve of a given degree with the associated basis functions. B-splines are constructed to exactly satisfy these requirements (for a cubic B-spline see Figure 2.4) and in a moment we will show how they are constructed.

The advantage of using this representation rather than the earlier one of monomials, is that the continuity conditions at the segment boundaries are already "hardwired" into the basis functions. No matter how we move the control points, the spline curve will always maintain its continuity, for example, C^2 in the case of cubic B-splines.[1] Furthermore, moving a control point has the greatest effect on the part of the curve near that control point, and no effect whatsoever beyond a certain range. These features make B-splines a much more appropriate tool for modeling piecewise polynomial curves.

Note: When we talk about curves, it is important to distinguish the curve itself and the graphs of the coordinate functions of the curve, which can also be thought of as curves. For example, a curve can be described by equations $x(t) = sin(t)$, $y(t) = cos(t)$. The curve itself is a circle, but the coordinate functions are sinusoids. For the moment, we are going to concentrate on representing the coordinate functions.

2.2.2 Definition of B-Splines

There are many ways to derive B-splines. Here we choose repeated convolution, since we can see from it directly how splines can be generated through subdivision.

We start with the simplest case: piecewise constant coordinate functions. Any piecewise constant function can be written as

$$x(t) = \sum x_i B_0^i(t),$$

[1] The differentiability of the basis functions guarantees the differentiability of the coordinate functions of the curve. However, it does not guarantee the geometric smoothness of the curve. We will return to this distinction in our discussion of subdivision surfaces.

where $B_0(t)$ is the box function defined as

$$\begin{aligned} B_0(t) &= 1 \quad \text{if} \quad 0 \leq t < 1 \\ &= 0 \quad \text{otherwise,} \end{aligned}$$

and the functions $B_0^i(t) = B_0(t-i)$ are translates of $B_0(t)$. Furthermore, let us represent the continuous convolution of two functions $f(t)$and $g(t)$ with

$$(f \otimes g)(t) = \int f(s)g(t-s)ds.$$

A B-spline basis function of degree n can be obtained by convolving the basis function of degree $n-1$ with the box $B_0(t)$.[2] For example, the B-spline of degree 1 is defined as the convolution of $B_0(t)$ with itself

$$B_1(t) = \int B_0(s)B_0(t-s)ds.$$

Graphically (see Figure 2.5), this convolution can be evaluated by sliding one box function along the coordinate axis from minus to plus infinity while keeping the second box fixed. The value of the convolution for a given position of the moving box is the area under the product of the boxes, which is just the length of the interval where both boxes are non-zero. At first the two boxes do not have common support. Once the moving box reaches 0, there is a growing overlap between the supports of the graphs. The value of the convolution grows with t until $t = 1$. Then the overlap starts decreasing, and the value of the convolution decreases down to zero at $t = 2$. The function $B_1(t)$ is the linear hat function as shown in Figure 2.5.

We can compute the B-spline of degree 2 convolving $B_1(t)$ with the box $B_0(t)$ again

$$B_2(t) = \int B_1(s)B_0(t-s)ds.$$

In this case, the resulting curve consists of three quadratic segments defined on intervals $(0,1)$, $(1,2)$ and $(2,3)$. In general, by convolving l times, we can get a B-spline of degree l

$$B_l(t) = \int B_{l-1}(s)B_0(t-s)ds.$$

Defining B-splines in this way a number of important properties immediately follow. The first concerns the continuity of splines

[2]The *degree* of a polynomial is the highest order exponent which occurs, while the *order* counts the number of coefficients and is 1 larger. For example, a cubic curve is of degree 3 and order 4.

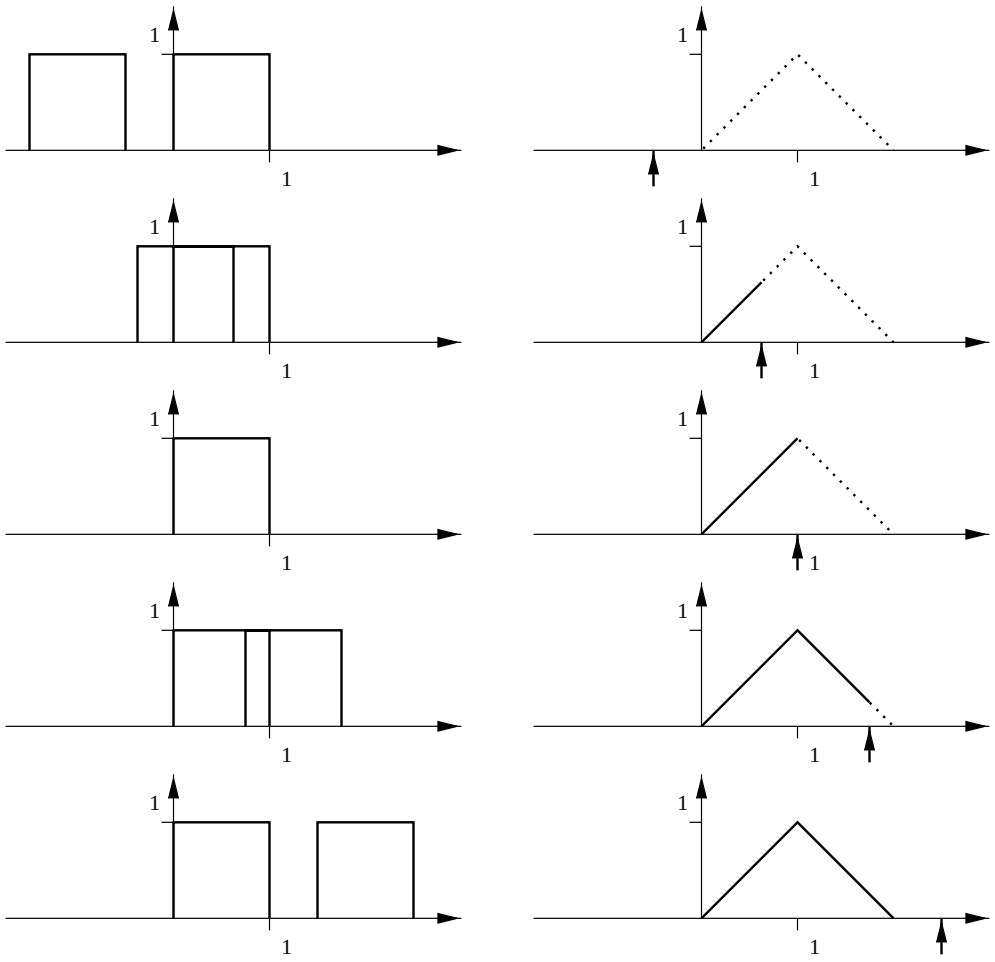


Figure 2.5: The definition of degree 1 B-Spline $B_1(t)$ (right side) through convolution of $B_0(t)$ with itself (left side).

Theorem 1 *If $f(t)$ is C^k-continuous, then $(B_0 \otimes f)(t)$ is C^{k+1}-continuous.*

This is a direct consequence of convolution with a box function. From this it follows that the B-spline of degree n is C^{n-1} continuous because the B-spline of degree 1 is C^0-continuous.

2.2.3 Refinability of B-splines

Another remarkable property of B-splines is that they obey a *refinement equation*. This is the key observation to connect splines and subdivision. The refinement equation for B-splines of degree l is given by

$$B_l(t) = \frac{1}{2^l} \sum_{k=0}^{l+1} \binom{l+1}{k} B_l(2t-k). \tag{2.1}$$

In other words, the B-spline of degree l can be written as a linear combination of *translated* (k) and *dilated* ($2t$) copies of itself. For a function to be refineable in this way is a rather special property. As an example of the above equation at work consider the hat function shown in Figure 2.5. It is easy to see that it can be written as a linear combination of dilated hat functions with weights $(1/2, 1, 1/2)$ respectively.

The property of refinability is the key to subdivision and so we will take a moment to prove it. We start by observing that the box function, i.e., the B-spline of degree 0 can be written in terms of dilates and translates of itself

$$B_0(t) = B_0(2t) + B_0(2t-1), \tag{2.2}$$

which is easily checked by direct inspection. Recall that we defined the B-spline of degree l as

$$B_l(t) = \bigotimes_{i=0}^{l} B_0(t) = \bigotimes_{i=0}^{l} (B_0(2t) + B_0(2t-1)) \tag{2.3}$$

This expression can be "multiplied" out by using the following properties of convolution for functions $f(t)$, $g(t)$, and $h(t)$

$$\begin{array}{rcll} f(t) \otimes (g(t) + h(t)) & = & f(t) \otimes g(t) + f(t) \otimes h(t) & \text{linearity} \\ f(t-i) \otimes g(t-k) & = & m(t-i-k) & \text{time shift} \\ f(2t) \otimes g(2t) & = & \frac{1}{2} m(2t) & \text{time scaling} \end{array}$$

where $m(t) = f(t) \otimes g(t)$. These properties are easy to check by substituting the definition of convolution and amount to simple change of variables in the integration.

For example, in the case of B_1 we get

$$\begin{aligned}
B_1(t) &= B_0(t) \otimes B_0(t) \\
&= (B_0(2t) + B_0(2t-1)) \otimes (B_0(2t) + B_0(2t-1)) \\
&= B_0(2t) \otimes B_0(2t) + B_0(2t) \otimes B_0(2t-1) + B_0(2t-1) \otimes B_0(2t) + B_0(2t-1) \otimes B_0(2t-1) \\
&= \frac{1}{2}B_1(2t) + \frac{1}{2}B_1(2t-1) + \frac{1}{2}B_1(2t-1) + \frac{1}{2}B_1(2t-1-1) \\
&= \frac{1}{2}(B_1(2t) + 2B_1(2t-1) + B_1(2t-2)) \\
&= \frac{1}{2^1}\sum_{k=0}^{2}\binom{2}{k}B_1(2t-k).
\end{aligned}$$

The general statement for B-splines of degree l now follows from the binomial theorem

$$(x+y)^{l+1} = \sum_{k=0}^{l+1}\binom{l+1}{k}x^{l+1-k}y^k,$$

with $B_0(2t)$ in place of x and $B_0(2t-1)$ in place of y.

2.2.4 Refinement for Spline Curves

With this machinery in hand let's revisit spline curves. Let

$$\gamma(t) = \begin{bmatrix} x(t) \\ y(t) \end{bmatrix} = \sum_i p_i B_l^i(t)$$

be such a spline curve of degree l with control points $(x_i, y_i)^T = p_i \in \mathbf{R}^2$. Since we don't want to worry about boundaries for now we leave the index set i unspecified. We will also drop the subscript l since the degree, whatever it might be, is fixed for all our examples. Due to the definition of $B^i(t) = B(t-i)$ each control point exerts influence over a small part of the curve with parameter values $t \in [i, i+l]$.

Now consider $\mathbf{p}$, the vector of control points of a given curve:

$$\mathbf{p} = \begin{bmatrix} \vdots \\ p_{-2} \\ p_{-1} \\ p_0 \\ p_1 \\ p_2 \\ \vdots \end{bmatrix}$$

and the vector $\mathbf{B}(t)$, which has as its elements the translates of the function B as defined above

$$\mathbf{B}(t) = \begin{bmatrix} \ldots & B(t+2) & B(t+1) & B(t) & B(t-1) & B(t-2) & \ldots \end{bmatrix}.$$

In this notation we can denote our curve as $\mathbf{B}(t)\mathbf{p}$.

Using the refinement relation derived earlier, we can rewrite each of the elements of $\mathbf{B}$ in terms of its dilates

$$\mathbf{B}(2t) = \begin{bmatrix} \ldots & B(2t+2) & B(2t+1) & B(2t) & B(2t-1) & B(2t-2) & \ldots \end{bmatrix},$$

using a matrix S to encode the refinement equations

$$\mathbf{B}(t) = \mathbf{B}(2t)S.$$

The entries of S are given by Equation 2.1

$$S_{2i+k,i} = s_k = \frac{1}{2^l}\binom{l+1}{k}.$$

The only non-zero entries in each column are the weights of the refinement equation, while successive columns are copies of one another save for a shift down by two rows.

We can use this relation to rewrite $\gamma(t)$

$$\gamma(t) = \mathbf{B}(t)\mathbf{p} = \mathbf{B}(2t)S\mathbf{p}.$$

It is still the same curve, but desribed with respect to dilated B-splines, i.e., B-splines whose support is half as wide and which are spaced twice as dense. We performed a change from the old basis $\mathbf{B}(t)$ to the new basis $\mathbf{B}(2t)$ and concurrently changed the old control points $\mathbf{p}$ to the appropriate new control points $S\mathbf{p}$. This process can be repeated

$$\begin{aligned} \gamma(t) &= \mathbf{B}(t)\mathbf{p}^0 \\ &= \mathbf{B}(2t)\mathbf{p}^1 &&= \mathbf{B}(2t)S\mathbf{p}^0 \\ &\quad\vdots \\ &= \mathbf{B}(2^j t)\mathbf{p}^j &&= \mathbf{B}(2^j t)S^j\mathbf{p}^0, \end{aligned}$$

from which we can define the relationship between control points at different levels of subdivision

$$\mathbf{p}^{j+1} = S\mathbf{p}^j,$$

where S is our infinite subdivision matrix.

Looking more closely at one component, i, of our control points we see that

$$p_i^{j+1} = \sum_l S_{i,l}\, p_l^j.$$

To find out exactly which s_k is affecting which term, we can divide the above into odd and even entries. For the odd entries we have

$$p_{2i+1}^{j+1} = \sum_l S_{2i+1,l}\, p_l^j = \sum_l s_{2(i-l)+1}\, p_l^j$$

and for the even entries we have

$$p_{2i}^{j+1} = \sum_l S_{2i,l}\, p_l^j = \sum_l s_{2(i-l)}\, p_l^j.$$

From which we essentially get two different subdivision rules one for the new *even* control points of the curve and one for the new *odd* control points. As examples of the above, let us consider a two concrete cases. For piecewise linear subdivision, the basis functions are hat functions. The odd coefficients are $\frac{1}{2}$ and $\frac{1}{2}$, and a lone 1 for the even point. For cubic splines the odd coefficients turn out to be $\frac{1}{2}$ and $\frac{1}{2}$, while the even coefficients are $\frac{1}{8}$, $\frac{6}{8}$, and $\frac{1}{8}$.

Another way to look at the distinction between even and odd is to notice that odd points at level $j+1$ are newly inserted, while even points at level $j+1$ correspond directly to the old points from level j. In the case of linear splines the even points are in fact the *same* at level $j+1$ as they were at level j. Subdivision schemes that have this property will later be called *interpolating*, since points, once they have been computed, will never move again. In contrast to this consider cubic splines. In that case even points at level $j+1$ are local averages of points at level j so that $p_{2i}^{j+1} \neq p_i^j$. Schemes of this type will later be called *approximating*.

2.2.5 Subdivision for Spline Curves

In the previous section we saw that we can refine the control point sequence for a given spline by multiplying the control point vector $\mathbf{p}$ by the matrix S, which encodes the refinement equation for the B-spline used in the definition of the curve. What happens if we keep repeating this process over and over, generating ever denser sets of control points? It turns out the control point sequence converges to the actual spline curve. The speed of convergence is geometric, which is to say that the difference between the curve and its control points decreases by a constant factor on every subdivision step. Loosely speaking this means that the actual curve is hard to distinguish from the sequence of control points after only a few subdivision steps.

We can turn this last observation into an algorithm and the core of the subdivision paradigm. Instead of drawing the curve itself on the screen we draw the control polygon, i.e., the piecewise linear curve through the control points. Applying the subdivision matrix to the control points defines a sequence of piecewise linear curves which quickly converge to the spline curve itself.

In order to make these observations more precise we need to introduce a little more machinery in the next section.

2.3 Subdivision as Repeated Refinement

2.3.1 Discrete Convolution

The coefficients s_k of the B-spline refinement equation can also be derived from another perspective, namely discrete convolution. This approach mimics closely the definition of B-splines through continuous convolution. Using this machinery we can derive and check many useful properties of subdivision by looking at simple polynomials.

Recall that the generating function of a sequence a_k is defined as

$$A(z) = \sum_k a_k z^k,$$

where $A(z)$ is the z-transform of the sequence a_k. This representation is closely related to the discrete Fourier transform of a sequence by restricting the argument z to the unit circle, $z = \exp(i\theta)$. For the case of two coefficient sequences a_k and b_k their convolution is defined as

$$c_k = (a \otimes b)_k = \sum_n a_{k-n} b_n.$$

In terms of generating functions this can be stated succinctly as

$$C(z) = A(z)B(z),$$

which comes as no surprise since convolution in the time domain is multiplication in the Fourier domain.

The main advantage of generating functions, and the reason why we use them here, is that manipulations of sequences can be turned into simple operations on the generating functions. A very useful example of this is the next observation. Suppose we have two functions that each satisfy a refinement equation

$$\begin{aligned} f(t) &= \sum_k a_k \, f(2t-k) \\ g(t) &= \sum_k b_k \, g(2t-k). \end{aligned}$$

In that case the convolution $h = f \otimes g$ of f and g also satisfies a refinement equation

$$h(t) = \sum_k c_k h(2t-k),$$

whose coefficients h_k^2 are given by the convolution of the coefficients of the individual refinement equations

$$c_k = \frac{1}{2}\sum_i a_{k-i} b_i.$$

With this little observation we can quickly find the refinement equation, and thus the coefficients of the subdivision matrix S, by repeated multiplication of generating functions. Recall that the box function $B_0(t)$ satisfies the refinement equation $B_0(t) = B_0(2t) + B_0(2t-1)$. The generating function of this refinement equation is $A(z) = (1+z)$ since the only non-zero terms of the refinement equation are those belonging to indices 0 and 1. Now recall the definition of B-splines of degree l

$$B_l(t) = \bigotimes_{k=0}^{l} B_0(t),$$

from which we immediately get the associated generating function

$$S(z) = \frac{1}{2^l}(1+z)^{l+1}.$$

The values s_k used for the definition of the subdivision matrix are simply the coefficients of the various powers of z in the polynomial $S(z)$

$$S(z) = \frac{1}{2^l}\sum_{k=0}^{l+1} \binom{l+1}{k} z^k,$$

where we used the binomial theorem to expand $S(z)$. Note how this matches the definition of s_k in Equation 2.1.

Recall Theorem 1, which we used to argue that B-splines of degree n are C^{n-1} continuous. That same theorem can now be expressed in terms of generating functions as follows

Theorem 2 *If $S(z)$ defines a convergent subdivision scheme yielding a C^k-continuous limit function then $\frac{1}{2}(1+z)S(z)$ defines a convergent subdivision scheme with C^{k+1}-continuous limit functions.*

We will put this theorem to work in analyzing a given subdivision scheme by peeling off as many factors of $\frac{1}{2}(1+z)$ as possible, while still being able to prove that the remainder converges to a continuous

limit function. With this trick in hand all we have left to do is establish criteria for the convergence of a subdivision scheme to a continuous function. Once we can verify such a condition for the subdivision scheme associated with B-spline control points we will be justified in drawing the piecewise linear approximations of control polygons as approximations for the spline curve itself. We now turn to this task.

2.3.2 Convergence of Subdivision

There are many ways to talk about the convergence of a sequence of functions to a limit. One can use different norms and different notions of convergence. For our purposes the simplest form will suffice, uniform convergence.

We say that a sequence of functions f_i defined on some interval $[a,b] \subset \mathbf{R}$ *converges uniformly* to a limit function f if for all $\varepsilon > 0$ there exists an $n_0 > 0$ such that for all $n > n_0$

$$\max_{t\in[a,b]} |f(t) - f_n(t)| < \varepsilon.$$

Or in words, as of a certain index (n_0) all functions in the sequence "live" within an ε sized tube around the limit function f. This form of convergence is sufficient for our purposes and it has the nice property that if a sequence of continuous functions converges uniformly to some limit function f, that limit function is itself continuous.

For later use we introduce some norm symbols

$$\begin{aligned} \|f(t)\| &= \sup_t |f(t)| \\ \|\mathbf{p}\| &= \sup_i |p_i| \\ \|S\| &= \sup_i \sum_k |S_{ik}|, \end{aligned}$$

which are compatible in the sense that, for example, $\|S\mathbf{p}\| \leq \|S\|\|\mathbf{p}\|$.

The sequence of functions we want to analyze now are the control polygons as we refine them with the subdivision rule S. Recall that the control polygon is the piecewise linear curve through the control points $\mathbf{p}^j$ at level j. Independent of the subdivision rule S we can use the linear B-splines to define the piecewise linear curve through the control points as $P^j(t) = \mathbf{B_1}(2^j t)\mathbf{p}^j$.

One way to show that a given subdivision scheme S converges to a continuous limit function is to prove that (1) the limit

$$P^\infty(t) = \lim_{j\to\infty} P^j(t)$$

exists for all t and (2) that the sequence $P^j(t)$ converges uniformly. In order to show this property we need to make the assumption that all rows of the matrix S sum to 1, i.e., the odd and even coefficients of the refinement relation separately sum to 1. This is a reasonable requirement since it is needed to ensure the affine invariance of the subdivision process, as we will later see. In matrix notation this means $S\mathbf{1} = \mathbf{1}$, or in other words, the vector of all 1's is an eigenvector of the subdivision matrix with eigenvalue 1. In terms of generating functions this means $S(-1) = 0$, which is easily verified for the generating functions we have seen so far.

Recall that the definition of continuity in the function setting is based on differences. We say $f(t)$ is continuous at t_0 if for any $\varepsilon > 0$ there exists a $\delta > 0$ so that $|f(t_0) - f(t)| < \varepsilon$ as long as $|t_0 - t| < \delta$. The corresponding tool in the subdivision setting is the difference between two adjacent control points $p^j_{i+1} - p^j_i = (\Delta \mathbf{p}^j)_i$. We will show that if the differences between neighboring control points shrink fast enough, the limit curve will exist and be continuous:

Lemma 3 *If $\|\Delta \mathbf{p}^j\| < c\gamma^j$ for some constant $c > 0$ and a shrinkage factor $0 < \gamma < 1$ for all $j > j_0 \geq 0$ then $P^j(t)$ converges to a continuous limit function $P^\infty(t)$.*

Proof: Let S be the subdivision rule at hand, $\mathbf{p}^1 = S\mathbf{p}^0$ and S_1 be the subdivision rule for B-splines of degree 1. Notice that the rows of $S - S_1$ sum to 0

$$(S - S_1)\mathbf{1} = S\mathbf{1} - S_1\mathbf{1} = \mathbf{1} - \mathbf{1} = \mathbf{0}.$$

This implies that there exists a matrix D such that $S - S_1 = D\Delta$, where Δ computes the difference of adjacent elements $(\Delta)_{ii} = -1$, $(\Delta)_{i,i+1} = 1$, and zero otherwise. The entries of D are given as $D_{ij} = -\sum_{k=i}^{j}(S - S_1)_{ik}$. Now consider the difference between two successive piecewise linear approximations of the control points

$$\begin{aligned}
\|P^{j+1}(t) - P^j(t)\| &= \|\mathbf{B_1}(2^{j+1}t)\mathbf{p}^{j+1} - \mathbf{B_1}(2^j t)\mathbf{p}^j\| \\
&= \|\mathbf{B_1}(2^{j+1}t)S\mathbf{p}^j - \mathbf{B_1}(2^{j+1}t)S_1\mathbf{p}^j\| \\
&= \|\mathbf{B_1}(2^{j+1}t)(S - S_1)\mathbf{p}^j\| \\
&< \|\mathbf{B_1}(2^{j+1}t)\|\|D\Delta\mathbf{p}^j\| \\
&< \|D\|\|\Delta\mathbf{p}^j\| \\
&< \|D\|c\gamma^j.
\end{aligned}$$

This implies that the telescoping sum $P^0(t) + \sum_{k=0}^{j}(P^{k+1} - P^k)(t)$ converges to a well defined limit function $P^\infty(t)$ as $j \to \infty$ and furthermore

$$\|P^\infty(t) - P^j(t)\| < \frac{\|D\|c}{1-\gamma}\gamma^j,$$

which implies uniform convergence and thus continuity of $P^\infty(t)$ as claimed.

How do we check such a condition for a given subdivision scheme? Suppose we had a derived subdivision scheme D for the differences themselves

$$\Delta \mathbf{p}^{j+1} = D\Delta \mathbf{p}^j,$$

defined as the scheme that satisfies

$$\Delta S = D\Delta.$$

Or in words, we are looking for a *difference scheme D* such that taking differences after subdivision is the same as applying the difference scheme to the differences. Does D always exist? The answer is yes if S is affinely invariant, i.e., $S(-1) = 0$. This follows from the following argument. Multiplying S by Δ computes a matrix whose rows are differences of adjacent rows in S. Since odd and even numbered rows of S each sum to one, the rows of ΔS must each sum to zero. Now the existence of a matrix D such that $\Delta S = D\Delta$ follows as in the argument above.

Given this difference scheme D all we would have to show is that some power $m > 0$ of D has norm less than 1, $\|D^m\| = \gamma < 1$. In that case $\|\Delta \mathbf{p}^j\| < c(\gamma^{1/m})^j$. (We will see in a moment that the extra degree of freedom provided by the parameter m is needed in some cases.)

As an example, let us check this condition for cubic B-splines. Recall that $B_3(z) = \frac{1}{8}(1+z)^4$, i.e.,

$$\begin{aligned} p_{2i+1}^{j+1} &= \frac{1}{8}(4p_i^j + 4p_{i+1}^j) \\ p_{2i}^{j+1} &= \frac{1}{8}(p_{i-1}^j + 6p_i^j + p_{i+1}^j). \end{aligned}$$

Taking differences we have

$$\begin{aligned} (\Delta \mathbf{p}^{j+1})_{2i} &= p_{2i+1}^{j+1} - p_{2i}^{j+1} = \frac{1}{8}(-p_{i-1}^j - 2p_i^j + 3p_{i+1}^j) \\ &= \frac{1}{8}(3(p_{i+1}^j - p_i^j) + 1(p_i^j - p_{i-1}^j)) = \frac{1}{8}(3(\Delta \mathbf{p}^j)_i + 1(\Delta \mathbf{p}^j)_{i-1}), \end{aligned}$$

and similarly for the odd entries so that $D(z) = \frac{1}{8}(1+z)^3$, from which we conclude that $\|D\| = \frac{1}{2}$, and that the subdivision scheme for cubic B-splines converges uniformly to a continuous limit function, namely the B-spline itself.

Another example, which is not a spline, is the so called 4 point scheme [4]. It was used to create the curve in Figure 2.1, which is interpolating rather than approximating as is the case with splines. The generating function for the 4 point scheme is

$$S(z) = \frac{1}{16}(-z^{-3} + 4z^{-2} - z^{-1})(1+z)^4$$

Recall that each additional factor of $\frac{1}{2}(1+z)$ in the generating function increases the order of continuity of the subdivision scheme. If we want to show that the limit function of the 4 point scheme is differentiable we need to show that $\frac{1}{8}(-z^{-3}+4z^{-2}-z^{-1})(1+z)^3$ converges to a continuous limit function. This in turn requires that $D(z)=\frac{1}{8}(-z^{-3}+4z^{-2}-z^{-1})(1+z)^2$ satisfy a norm estimate as before. The rows of D have non-zero entries of $(\frac{1}{4},\frac{1}{4})$, and $(\frac{-1}{8},\frac{6}{8},\frac{-1}{8})$ respectively. Thus $\|D\|=1$, which is not strong enough. However, with a little bit more work one can show that $\|D^2\|=\frac{3}{4}$, so that indeed the 4 point scheme is C^1.

In general, the difficult part is to find a set of coefficients for which subdivision converges. There is no general method to achieve this. Once a convergent subdivision scheme is found, one can always obtain a desired order of continuity by convolving with the box function.

2.3.3 Summary

So far we have considered subdivision only in the context of splines where the subdivision rule, i.e., the coefficients used to compute a refined set of control points, was fixed and everywhere the same. There is no pressing reason for this to be so. We can create a variety of different curves by manipulating the coefficients of the subdivision matrix. This could be done globally or locally. I.e., we could change the coefficients within a subdivision level and/or between subdivision levels. In this regard, splines are just a special case of the more general class of curves, subdivision curves. For example, at the beginning of this chapter we briefly outlined an interpolating subdivision method, while spline based subdivision is approximating rather than interpolating.

Why would one want to draw a spline curve by means of subdivision? In fact there is no sufficiently strong reason for using subdivision in one dimension and none of the commercial line drawing packages do so, but the argument becomes much more compelling in higher dimensions as we will see in later chapters.

In the next section we use the subdivision matrix to study the behavior of the resulting curve at a point or in the neighborhood of a point. We will see that it is quite easy, for example, to evaluate the curve exactly at a point, or to compute a tangent vector, simply from a deeper understanding of the subdivision matrix.

2.4 Analysis of Subdivision

In the previous section we have shown that uniform spline curves can be thought of as a special case of subdivision curves. So far, we have seen only examples for which we use a fixed set of coefficients to

compute the control points everywhere. The coefficients define the appearance of the curve, for example, whether it is differentiable or has sharp corners. Consequently it is possible to control the appearance of the curve by modifying the subdivision coefficients locally. So far we have not seen a compelling reason to do so in the 1D setting. However, in the surface setting it will be essential to change the subdivision rule locally around extraordinary vertices to ensure maximal order of continuity. But before studying this question we once again look at the curve setting first since the treatment is considerably easier to follow in that setting.

To study properties such as differentiability of the curve (or surface) we need to understand which of the control points influences the neighborhood of the point of interest. This notion is captured by the concept of invariant neighborhoods to which we turn now.

2.4.1 Invariant Neighborhoods

Suppose we want to study the limit curve of a given subdivision scheme in the vicinity of a particular control point.[3] To determine *local* properties of a subdivision curve, we do not need the whole infinite vector of control points or the infinite matrix describing subdivision of the entire curve. Differentiability, for example, is a local property of a curve. To study it we need consider only an arbitrarily small piece of the curve around the origin. This leads to the question of which control points influence the curve in the neighborhood of the origin?

As a first example consider cubic B-spline subdivision. There is one cubic segment to the left of the origin with parameter values $t \in [-1,0]$ and one segment to the right with parameter range $t \in [0,1]$. Figure 2.6 illustrates that we need 5 control points at the coarsest level to reach any point of the limit curve which is associated with a parameter value between -1 and 1. We say that the *invariant neighborhood* has size 5. This size depends on the number of non-zero entries in each row of the subdivision matrix, which is 2 for odd points and 3 for even points. The latter implies that we need one extra control point to the left of -1 and one to the right of 1.

The fact that the central 5 control points control the behavior of the limit curve at the origin holds independent of the level. With the central 5 control points at level j we can compute the central 5 control points at level $j+1$. This implies that in order to study the behavior of the curve at the origin all we have

[3] Here and in the following we assume that the point of interest is the origin. This can always be achieved through renumbering of the control points.

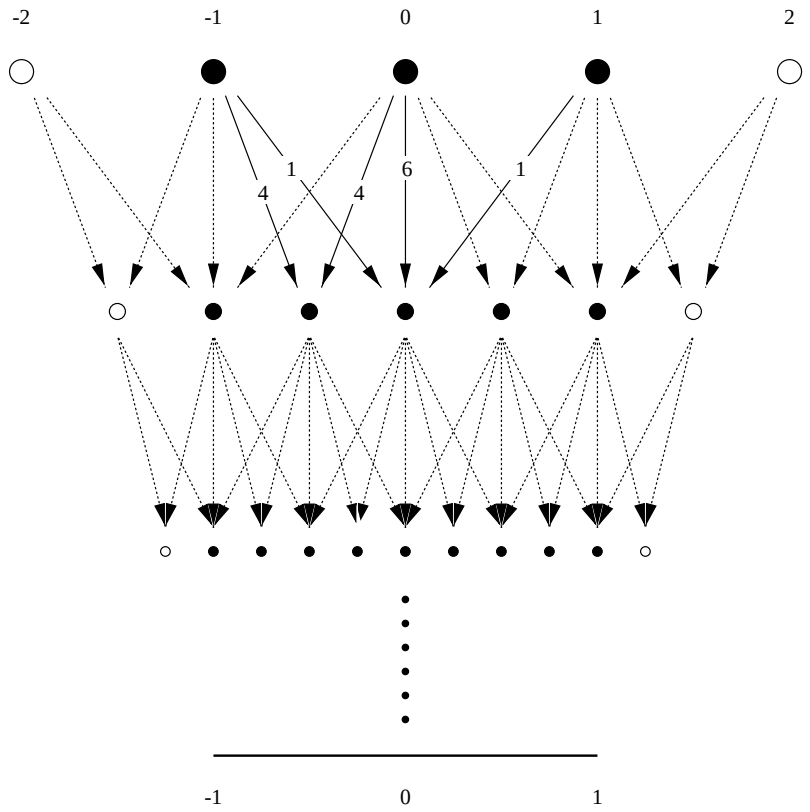


Figure 2.6: In the case of cubic B-spline subdivision the invariant neighborhood is of size 5. It takes 5 control points at the coarsest level to determine the behavior of the subdivision limit curve over the two segments adjacent to the origin. At each level we need one more control point on the outside of the interval $t \in [-1,1]$ in order to continue on to the next subdivision level. 3 initial control points for example would not be enough.

to analyze is a small 5×5 subblock of the subdivision matrix

$$\begin{pmatrix} p_{-2}^{j+1} \\ p_{-1}^{j+1} \\ p_0^{j+1} \\ p_1^{j+1} \\ p_2^{j+1} \end{pmatrix} = \frac{1}{8} \begin{pmatrix} 1 & 6 & 1 & 0 & 0 \\ 0 & 4 & 4 & 0 & 0 \\ 0 & 1 & 6 & 1 & 0 \\ 0 & 0 & 4 & 4 & 0 \\ 0 & 0 & 1 & 6 & 1 \end{pmatrix} \begin{pmatrix} p_{-2}^{j} \\ p_{-1}^{j} \\ p_0^{j} \\ p_1^{j} \\ p_2^{j} \end{pmatrix}.$$

Another example is the 4 point subdivision scheme. It has an invariant neighborhood of 7 (see Fig-

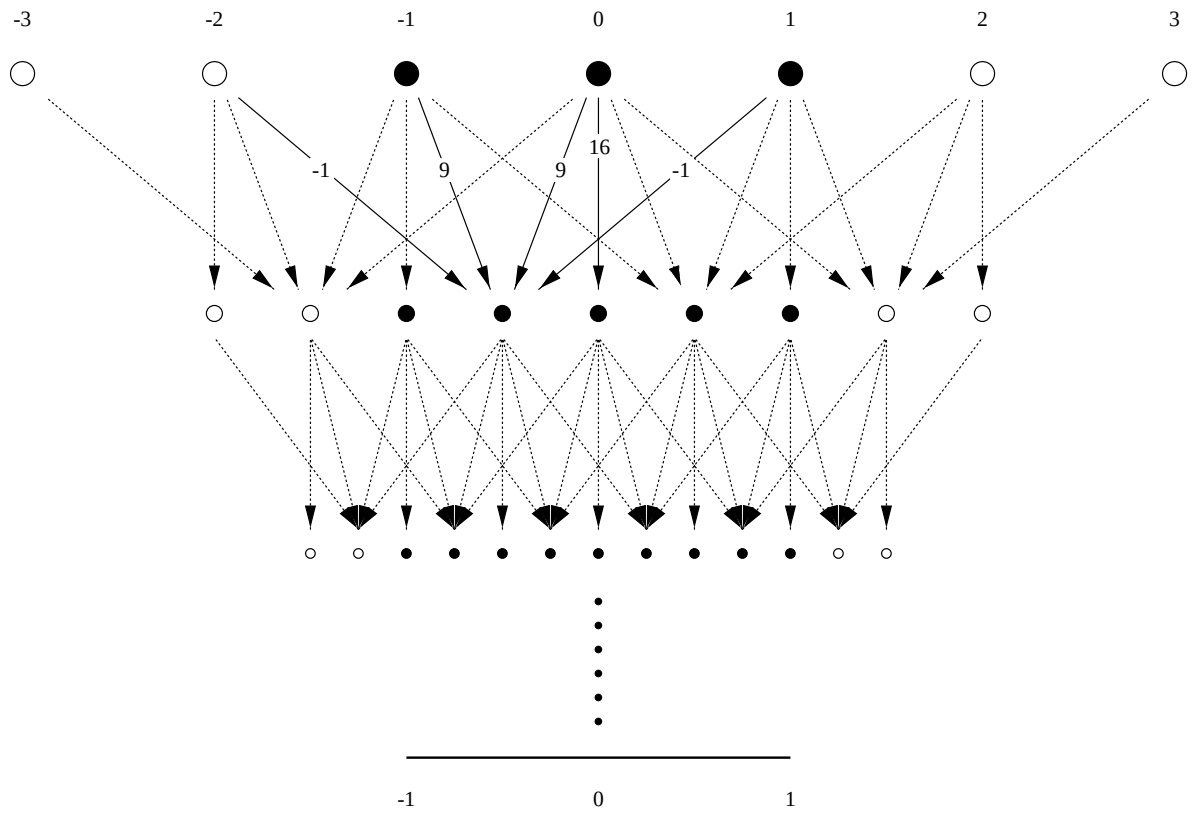


Figure 2.7: In the case of the 4 point subdivision rule the invariant neighborhood is of size 7. It takes 7 control points at the coarsest level to determine the behavior of the subdivision limit curve over the two segments adjacent to the origin. One extra point at p_2^j is needed to compute p_1^{j+1}. The other is needed to compute p_3^{j+1}, which requires p_3^j. Two extra points on the left and right result in a total of 7 in the invariant neighborhood.

ure 2.7). In this case the local subdivision matrix is

$$\begin{pmatrix} p_{-3}^{j+1} \\ p_{-2}^{j+1} \\ p_{-1}^{j+1} \\ p_0^{j+1} \\ p_1^{j+1} \\ p_2^{j+1} \\ p_3^{j+1} \end{pmatrix} = \frac{1}{16} \begin{pmatrix} -1 & 9 & 9 & -1 & 0 & 0 & 0 \\ 0 & 0 & 16 & 0 & 0 & 0 & 0 \\ 0 & -1 & 9 & 9 & -1 & 0 & 0 \\ 0 & 0 & 0 & 16 & 0 & 0 & 0 \\ 0 & 0 & -1 & 9 & 9 & -1 & 0 \\ 0 & 0 & 0 & 0 & 16 & 0 & 0 \\ 0 & 0 & 0 & -1 & 9 & 9 & -1 \end{pmatrix} \begin{pmatrix} p_{-3}^{j+1} \\ p_{-2}^{j+1} \\ p_{-1}^{j+1} \\ p_0^{j+1} \\ p_1^{j+1} \\ p_2^{j+1} \\ p_3^{j+1} \end{pmatrix}$$

Some properties of the curves generated by subdivision can be inferred from the properties of the local subdivision matrix. In particular, differentiability properties of the curve are related to the eigenstructure of the local subdivision matrix to which we now turn. From now on the symbol S will denote the *local* subdivision matrix.

2.4.2 Eigen Analysis

Recall from linear algebra that an *eigenvector* $\mathbf{x}$ of the matrix M is a non-zero vector such that $M\mathbf{x} = \lambda\mathbf{x}$, where λ is a scalar. We say that λ is the *eigenvalue* corresponding to the eigenvector $\mathbf{x}$.

Assume the local subdivision matrix S has size $n \times n$ and has real eigenvectors $\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{n-1}$, which form a basis, with corresponding real eigenvalues $\lambda_0 \geq \lambda_1 \geq \ldots \geq \lambda_{n-1}$. For example, in the case of cubic splines $n = 5$ and

$$\begin{aligned}(\lambda_0, \lambda_1, \lambda_2, \lambda_3, \lambda_4) &= (1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}) \\ (\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4) &= \begin{pmatrix} 1 & -1 & 1 & 1 & 0 \\ 1 & -\frac{1}{2} & \frac{2}{11} & 0 & 0 \\ 1 & 0 & -\frac{1}{11} & 0 & 0 \\ 1 & \frac{1}{2} & \frac{2}{11} & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 \end{pmatrix}.\end{aligned}$$

Note: not all subdivision schemes have only real eigenvalues or a complete set of eigenvectors. For example, the 4 point scheme has eigenvalues

$$(\lambda_0, \lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6) = (1, \frac{1}{2}, \frac{1}{4}, \frac{1}{4}, \frac{1}{8}, -\frac{1}{16}, -\frac{1}{16}),$$

but it does not have a complete set of eigenvectors. These degeneracies are the cause of much technical difficulty in the theory of subdivision. To keep our exposition simple and communicate the essential ideas we will ignore these cases and assume from now on that we have a complete set of eigenvectors.

In this setting we can write any vector $\mathbf{x}$ of length n as a linear combination of eigenvectors:

$$\mathbf{x} = \sum_{i=0}^{n-1} a_i \mathbf{x}_i$$

Similarly, we can write a decomposition of this type for a vector $\mathbf{p}$ of *n 2-D points* rather than single numbers. In this case each "coefficient" a_i is a 2-D point. The eigenvectors $\mathbf{x}_0, \ldots, \mathbf{x}_{n-1}$ are simply vectors of n real numbers.

In the basis of eigenvectors we can easily compute the result of application of the subdivision matrix to a vector of control points, that is, the control points on the next level

$$\begin{aligned} S\mathbf{p}^0 &= S\sum_{i=0}^{n-1} a_i\mathbf{x}_i \\ &= \sum_{i=0}^{n-1} a_i S\mathbf{x}_i \quad \text{by linearity of } S \\ &= \sum_{i=0}^{n-1} a_i\lambda_i\mathbf{x}_i \end{aligned}$$

Applying S j times, we obtain

$$\mathbf{p}^j = S^j\mathbf{p}^0 = \sum_{i=0}^{n-1} a_i\lambda_i^j\mathbf{x}_i.$$

2.4.3 Convergence of Subdivision

If $\lambda_0 > 1$, then $S^j\mathbf{x}^0$ would grow without bound as j increased and subdivision would not be convergent. Hence, we can see that in order for the sequence $S^j\mathbf{p}^0$ to converge at all, it is necessary that all eigenvalues are at most 1. It is also possible to show that only a single eigenvalue may have magnitude 1 [21].

A simple consequence of this analysis is that we can compute the limit position directly in the eigenbasis

$$P^\infty(0) = \lim_{j\to\infty} S^j\mathbf{p}^0 = \lim_{j\to\infty}\sum_{i=0}^{n-1} a_i\lambda_i^j\mathbf{x}_i = a_0,$$

since all eigen components $|\lambda_i| < 1$ decay to zero. For example, in the case of cubic B-spline subdivision we can compute the limit position of p_i^j as

$$p_i^\infty = a_0 = \frac{1}{6}(p_{i-1}^j + 4p_i^j + p_{i+1}^j).$$

Note that this expression is completely independent of the level j at which it is computed.

2.4.4 Invariance under Affine Transformations

If we moved all the control points simultaneously by the same amount, we would expect the curve defined by these control points to move in the same way as a rigid object. In other words, the curve should be *invariant under distance-preserving transformations, such as translation and rotation*. It follows from

linearity of subdivision that if subdivision is invariant with respect to distance-preserving transformations, it also should be invariant under any affine transformations. The family of affine transformations in addition to distance-preserving transformations, contains shears.

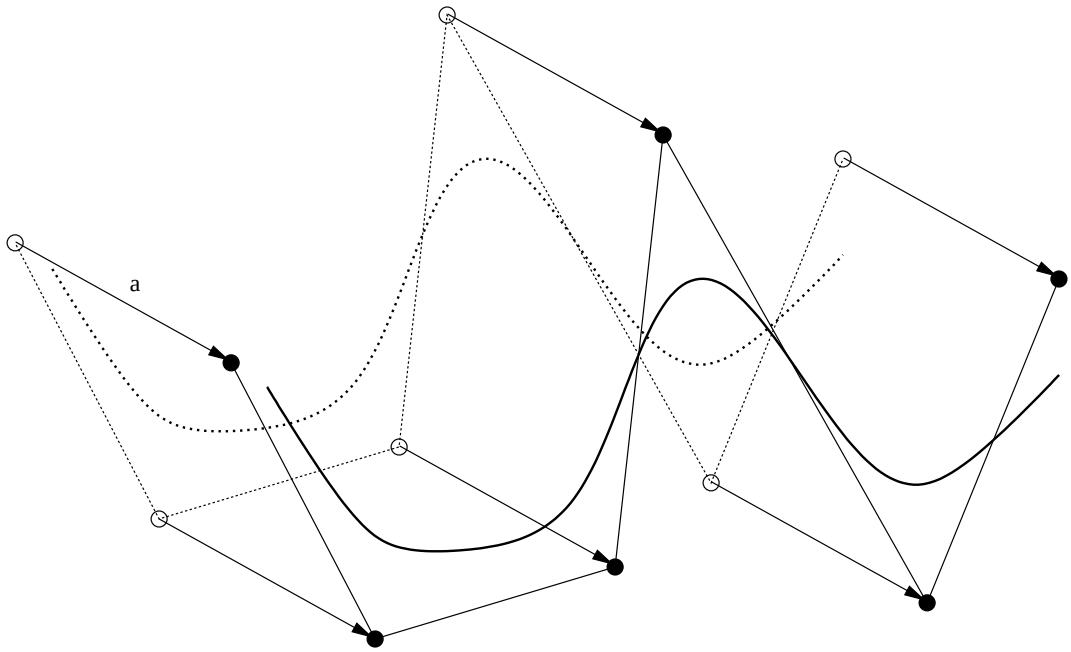


Figure 2.8: Invariance under translation.

Let $\vec{1}$ be an n-vector of 1's and $a \in \mathbf{R}^2$ a displacement in the plane (see Figure 2.8) Then $\vec{1} \cdot a$ represents a displacement of our seven points by a vector a. Applying subdivision to the transformed points, we get

$$\begin{aligned} S(\mathbf{p}^j + \vec{1} \cdot a) &= S\mathbf{p}^j + S(\vec{1} \cdot a) \quad \text{by linearity of } S \\ &= \mathbf{p}^{j+1} + S(\vec{1} \cdot a). \end{aligned}$$

From this we see that for translational invariance we need

$$S(\vec{1} \cdot a) = \vec{1} \cdot a$$

Therefore, $\vec{1}$ should be the eigenvector of S with eigenvalue $\lambda_0 = 1$.

Recall that when proving convergence of subdivision we assumed that **1** is an eigenvector with eigenvalue 1. We now see that this assumption satisfied by any reasonable subdivision scheme. It would be rather unnatural if the shape of the curve changed as we translate control points.

2.4.5 Geometric Behavior of Repeated Subdivision

If we assume that λ_0 is 1, and all other eigenvalues are less than 1, we can choose our coordinate system in such a way that a_0 is the origin in $\mathbf{R}^2$. In that case we have

$$\mathbf{p}^j = \sum_{i=1}^{n-1} a_i \lambda_i^j \mathbf{x}_i$$

Dividing both sides by λ_1^j, we obtain

$$\frac{1}{\lambda_1^j}\mathbf{p}^j = a_1\mathbf{x}_1 + \sum_{i=2}^{n-1} a_i \left(\frac{\lambda_i}{\lambda_1}\right)^j \mathbf{x}_i.$$

If we assume that $|\lambda_2|, \ldots, |\lambda_{n-1}| < |\lambda_1|$, the sum on the right approaches zero as $j \to \infty$. In other words the term corresponding to λ_1 will "dominate" the behavior of the vector of control points. In the limit, we get a set of n points arranged along the vector a_1. Geometrically, this is a vector tangent to our curve at the center point (see Figure 2.9).

Just as in the case of computing the limit point of cubic B-spline subdivision by computing a_0 we can compute the tangent vector at p_i^j by computing a_1

$$t_i^\infty = a_1 = p_{i+1}^j - p_{i-1}^j.$$

If there were two equal eigenvalues, say $\lambda_1 = \lambda_2$, as j increases, the points in the limit configuration will be linear combinations of two vectors a_1 and a_2, and in general would not be on the same line. This indicates that there will be no tangent vector at the central point. This leads us to the following condition, that, under some additional assumptions, is necessary for the existence of a tangent

All eigenvalues of S except $\lambda_0 = 1$ *should be less than* λ_1.

2.4.6 Summary

For our subdivision matrix S we desire the following characteristics

- the eigenvectors should form a basis;
- the first eigenvalue λ_0 should be 1;
- the second eigenvalue λ_1 should be less than 1;
- all other eigenvalues should be less than λ_1.

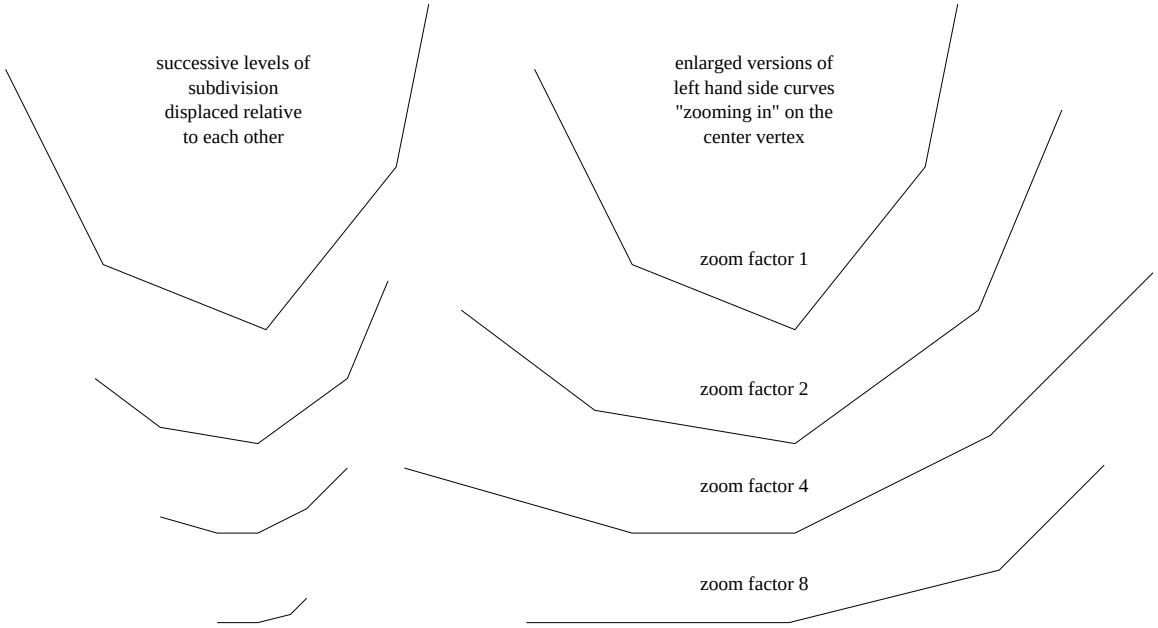


Figure 2.9: Repeatedly applying the subdivision matrix to our set of n control points results in the control points converging to a configuration aligned with the tangent vector. The various subdivision levels have been offset vertically for clarity.

Chapter 3

Subdivision Surfaces

Denis Zorin, Stanford University – New York University

In this chapter we review the basic theory of subdivision in two dimensions, and describe a number of subdivision schemes: Doo-Sabin, Catmull-Clark, Loop, Modified Butterfly, Kobbelt, Midedge.

Some of these schemes were around for a while: the 1978 papers of Doo and Sabin and Catmull and Clark were the first papers describing subdivision algorithms for surfaces. Other schemes are relatively new. Remarkably, during the period from 1978 until 1995 little progress was made in the area. In fact, until Reif's work on C^1-continuity of subdivision most basic questions about the behavior of subdivision surfaces near extraordinary vertices were not answered. Since then there was a steady stream of new theoretical and practical results: classical subdivision schemes were analyzed [20, 12], new schemes were proposed [26, 9, 7, 13], and general theory was developed for C^1 and C^k-continuity of subdivision [19, 14, 22, 24]. Smoothness analysis was performed in some form for all schemes described in this chapter; for all of them, definitive results were obtained during the last 2 years only.

One of the goals of this chapter is to provide an accessible introduction to the mathematics of subdivision surfaces (Sections 3.2 and 3.3. Building on the material of the first chapter, we concentrate on the few general concepts that we believe to be of primary importance: subdivision surfaces as parametric surfaces, C^1-continuity, eigenstructure of subdivision matrices, characteristic maps.

The developments of recent years convinced us of the importance of understanding the mathematical foundations of subdivision. A Computer Graphics professional who wishes to use subdivision in his code, probably is not interested in the subtle points of a theoretical argument. However, understanding the general concepts that are used to construct and analyze subdivision schemes would help her choose

the most appropriate subdivision algorithm or customize one for her needs.

We continue the chapter with an overview of the variety of existing surface subdivision schemes (Section 3.4). We attempt to describe various schemes in a systematic way to make it easier for the readers to understand the differences and similarities of different algorithms. Finally, we conclude with a discussion of the problems and limitations of subdivision (Section 3.5).

3.1 Subdivision Surfaces: an Example

One of the simplest subdivision schemes is the *Loop scheme*, invented by Charles Loop [10]. We will use this scheme as an example to introduce some basic features of subdivision for surfaces.

The Loop scheme is defined for triangular meshes. The general pattern of refinement, which we call *vertex insertion*, is shown in Figure 3.1.

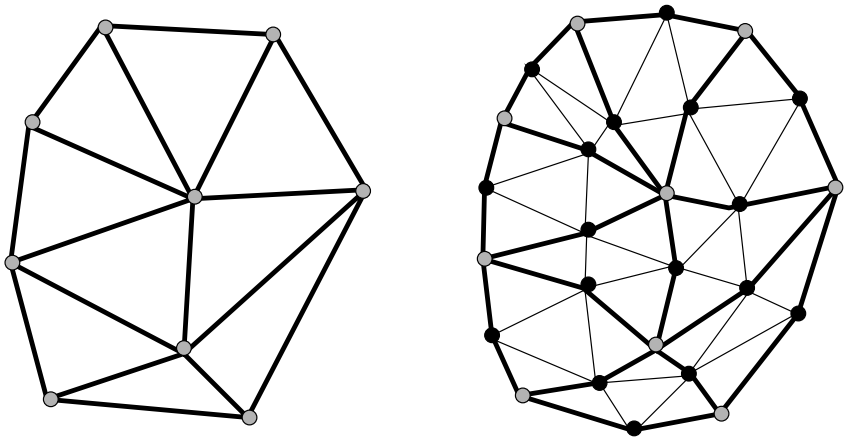


Figure 3.1: Refinement of a triangular mesh. New vertices are shown as black dots. Each edge of the the control mesh is split into two, and new vertices are reconnected to form 4 new triangles, replacing each triangle of the mesh.

Like most (but not all) other subdivision schemes, this scheme is based on a spline basis function, called the three-directional quartic box spline. Unlike more conventional splines, such as the bicubic spline, the three-directional box spline is defined on the regular *triangular* grid; the generating polynomial for this spline is

$$S(z_1,z_2) = \frac{1}{16}(1+z_1)^2(1+z_2)^2(1+z_1z_2)^2.$$

This spline basis function is C^2. Subdivision rules for it are shown in Figure 3.2.

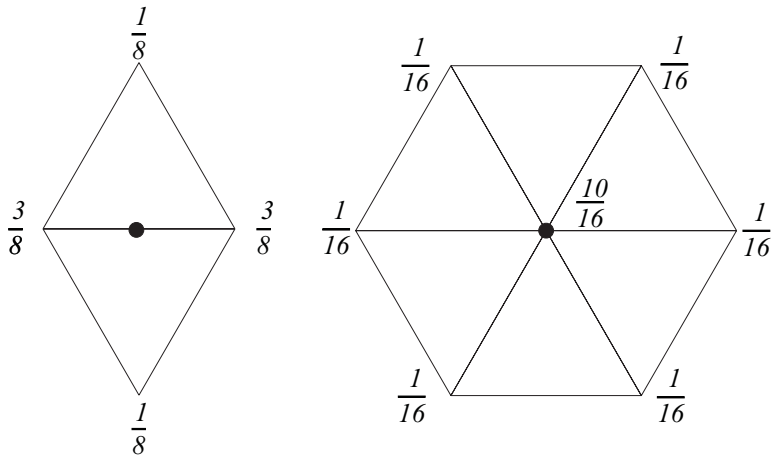


Figure 3.2: Subdivision coefficients for a three directional box spline.

In one dimension, once a spline basis is chosen, all the coefficients of the subdivision rules that are needed to generate a curve are completely determined. The situation is radically different and much more complex for surfaces. The structure of the control polygon for curves is always very simple: the vertices are arranged into a chain, and any two pieces of the chain of the same length always have identical structure. For two-dimensional meshes, the local structure of the mesh may vary. For example, the number of edges emanating from a vertex may be different from vertex to vertex. As a result the rules derived from the spline basis function may be applied only to parts of the mesh that are locally regular; that is, only to those vertices that have a valence of 6 (in the case of triangular schemes). In other cases, we have to design new rules for vertices with different valences. Such vertices are called *extraordinary*.

For the time being, we consider only meshes without a boundary. Note that in the example of the Loop subdivision scheme the rule used to compute the control point inserted at an edge can be applied anywhere. The only rule that needs modification is the rule used to compute new positions of control points inherited from the previous level.

Loop proposed to use coefficients shown in Figure 3.3. It turns out that this choice of coefficients guarantees that the limit surface of the scheme is “smooth.”

Note that these new rules only influence local behavior of the surface near extraordinary vertices. All vertices inserted in the course of subdivision are always regular, i.e., have valence 6.

This example demonstrates the main challenge in the design of subdivision schemes for surfaces:

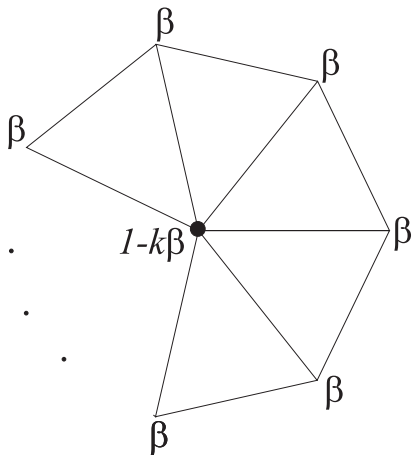


Figure 3.3: Loop scheme: coefficients for extraordinary vertices.

one has to define additional rules for irregular parts of the mesh in such a way that the limit surfaces have desired properties, in particular, are smooth. To understand why particular coefficients were chosen for Loop and other schemes, we have to define more precisely what a smooth surface is (Section 3.2), introducing two concepts of geometric smoothness—*tangent plane continuity* and C^1*-continuity*. Next, we explain how it is possible to understand local behavior of subdivision near extraordinary vertices using subdivision matrices (Section 3.3). Finally, in Section 3.4 we discuss a variety of subdivision rules in a systematic way.

3.2 Smoothness of Surfaces

Intuitively, we call a surface smooth, if, at a close distance, it becomes indistinguishable from a plane. Before discussing smoothness of subdivision surfaces in greater detail, we have to define more precisely what we mean by a surface, in a way that is convenient for analysis of subdivision.

The discussion in the section is somewhat informal; for a more rigorous treatment, see [19, 18, 22],

3.2.1 Natural Parameterization of Subdivision Surfaces

The Subdivision process produces a sequence of polyhedra with increasing numbers of faces and vertices. Intuitively, the subdivision surface is the limit of this sequence. The problem is that we have to define

what we mean by the limit more precisely. For this, and many other purposes, it is convenient to represent subdivision surfaces as functions defined on some parametric domain with values in $\mathbf{R}^3$. In the regular case, the plane or a part of the plane is the domain. However, for arbitrary control meshes, it might be impossible to parameterize the surface continuously over a planar domain.

Fortunately, there is a simple construction that allows one to use the *initial control mesh*, or more precisely, the corresponding polygonal complex, as the domain for the surface.

Parameterization over the initial control mesh We start with the simplest case: suppose the initial control mesh is a simple polyhedron, i.e., it does not have self-intersections.

Suppose each time we apply the subdivision rules to compute the finer control mesh, we also apply midpoint subdivision to a copy of the initial control polyhedron (see Figure 3.4). This means that we leave the old vertices where they are, and insert new vertices splitting each edge in two. Note that each control point that we insert in the mesh using subdivision corresponds to a point in the midpoint-subdivided polyhedron. Another important fact is that midpoint subdivision does not alter the control polyhedron regarded as a set of points; and no new vertices inserted by midpoint subdivision can possibly coincide.

We will use the second copy of the control polyhedron as our domain. We denote it as K, when it is regarded as a polyhedron with identified vertices, edges and faces, and $|K|$ when it is regarded simply as a subset of $\mathbf{R}^3$.

Important remark on notation: we will refer to the points computed by subdivision as **control points**; the word **vertex** is reserved for the vertices of the polyhedron that serves as the domain and new vertices added to it by midpoint subdivision. We will use the letter v to denote vertices, and $p^j(v)$ to denote the control point corresponding to v after j subdivision steps.

As we repeatedly subdivide, we get a mapping from a denser and denser subset of the domain to the control points of a finer and finer control mesh. At each step, we linearly interpolate between control vertices, and regard the mesh generated by subdivision as a piecewise linear function on the domain K. Now we have the same situation that we had for curves: a sequence of piecewise linear functions defined on a common domain. If this sequence of functions converges uniformly, the limit is a map f from $|K|$ into $\mathbf{R}^3$. This is the limit surface of subdivision.

An important fact about the parameterization that we have just constructed is that for a regular mesh the domain can be taken to be the plane with a regular triangular grid. If in the regular case the subdivision

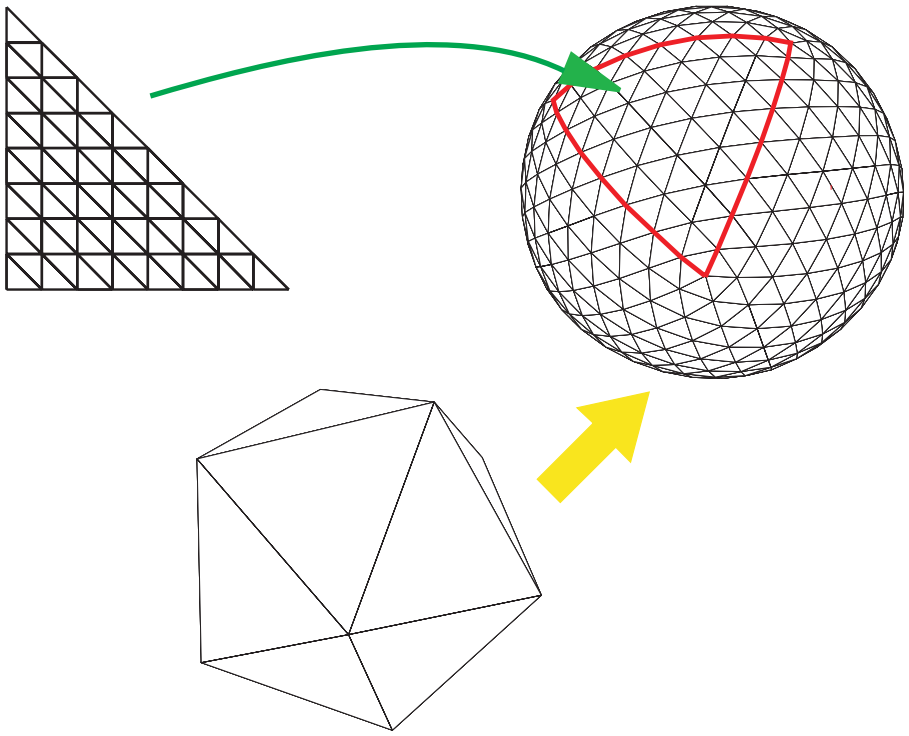


Figure 3.4: Natural parameterization of the subdivision surface

scheme reduces to spline subdivision, our parameterization is precisely the standard (u,v) parameterization of the spline, which is guaranteed to be smooth.

To understand the general idea, this definition is sufficient, and a reader not interested in the subtle details can proceed to the next section and assume from now on that the initial mesh has no self-intersections.

General case The crucial fact that we needed to parameterize the surface over its control polyhedron is the absence of self-intersections. Otherwise, it could happen that a vertex on the control polyhedron has more than one control point associated with it.

In general, we cannot rely on this assumption: quite often control meshes have self-intersections or coinciding control points. We can observe though that the positions of vertices of the control polyhedron are of no importance for our purposes: we can deform it in any way we want. In many cases, this is sufficient to eliminate the problem with self intersections; however, there are cases when the self-

intersection cannot be removed by any deformation (example: Klein bottle, Figure 3.5). It is always possible to do that if we place our mesh in a higher-dimensional space; in fact, 4 dimensions are always enough.

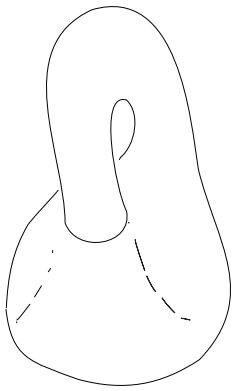


Figure 3.5: The surface (Klein bottle) has an intersection that cannot be removed in 3D.

This leads us to the following general choice of the domain: a polyhedron with no self-intersections, possibly in four-dimensional space. The polyhedron has to have the same structure as the initial control mesh of the surface, that is, there is a one-to-one correspondence between vertices, edges and faces of the domain and the initial control mesh. Note that now we are completely free to chose the control points of the initial mesh any way we like.

3.2.2 C^1-continuity and Tangent Plane Continuity

Now that we have defined the subdivision surface as a function $f : |K| \to \mathbf{R}^3$ on a polyhedron, we can formalize our intuitive notion of smoothness, namely local similarity to a piece of the plane. A surface is smooth at a point x of its domain $|K|$, if for a sufficiently small neighborhood U_x of that point the image $f(U_x)$ can be smoothly deformed into a a planar disk. More precisely,

Definition 1 *A surface $f : |K| \to \mathbf{R}^3$ is C^1-**continuous**, if for every point $x \in |K|$ there exists a regular parameterization $\pi : D \to f(U_x)$ of $f(U_x)$ over a unit disk D in the plane, where U_x is the neighborhood in $|K|$ of x. A **regular parameterization** π is one that is continuously differentiable, one-to-one, and has a Jacobi matrix of maximum rank.*

The condition that the Jacobi matrix of p has maximum rank is necessary to make sure that we have no degeneracies, i.e., that we really do have a surface, not a curve or point. If $p = (p_1, p_2, p_3)$ and the disc

is parameterized by x_1 and x_2, the condition is that the matrix

$$\begin{pmatrix} \frac{\partial p_1}{\partial x_1} & \frac{\partial p_1}{\partial x_2} \\ \frac{\partial p_2}{\partial x_1} & \frac{\partial p_2}{\partial x_2} \\ \frac{\partial p_3}{\partial x_1} & \frac{\partial p_3}{\partial x_2} \end{pmatrix}$$

have maximal rank (2).

There is another, weaker, definition of smoothness, which is often useful:

Definition 2 *A surface $f : |K| \to \mathbf{R}^3$ is **tangent plane continuous** at $x \in |K|$ if and only if surface normals are defined in a neighborhood around x and there exists a limit of normals at x.*

This is a useful definition, since it is easier to prove surfaces are tangent plane continuous (all that is required is to show the existence of a limit). Additionally the definition is very intuitive since it captures the notion that a surface is smooth if there exists a tangent plane. Tangent plane continuity, however, is weaker than C^1-continuity.

As a simple example of a surface that is tangent plane continuous but not C^1, consider the shape in Figure 3.6. Points in the vicinity of the central point are "wrapped around twice." There exists a tangent plane at that point, but the surface does not "locally look like a plane." Formally, there does not exist a regular parameterization from the unit disc onto the neighborhood of the center point, even though it has a well-defined tangent plane.

From the previous example, we see how the definition of tangent plane continuity must be strengthened to become C^1:

Lemma 4 *If a surface is tangent plane continuous at a point and the projection of the surface onto the tangent plane at that point is one-to-one, the surface is C^1.*

The proof can be found in [22].

3.3 Analysis of Subdivision Surfaces

In this section we discuss how to analyze smoothness of subdivision surfaces. Typically, it is known a priori that a scheme produces C^1-continuous (or better) surfaces in the regular setting. For local schemes in most cases this means that the surfaces generated on arbitrary meshes are C^1-continuous away from the extraordinary vertices. We start with a brief discussion of this fact, and than concentrate on analysis

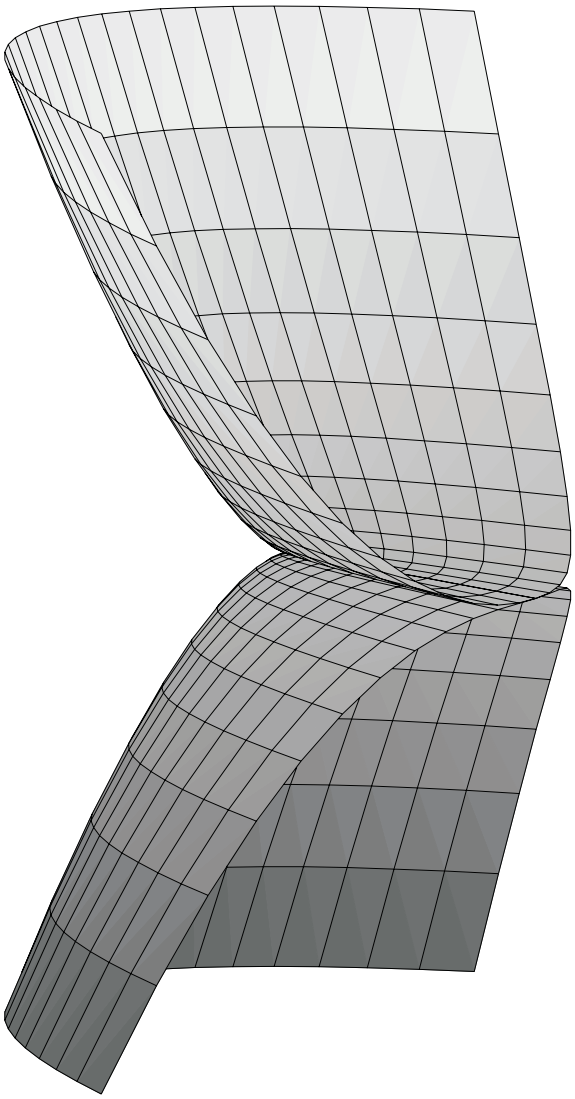


Figure 3.6: Example of a surface that is tangent plane continuous but not C^1.

of the behavior of the schemes near extraordinary vertices. Our goal is to formulate and provide some motivation for Reif's sufficient condition for C^1-continuity of subdivision.

We assume a subdivision scheme defined on a triangular mesh, with certain restrictions on the structure of the subdivision matrix. Similar derivations can be performed without these assumptions, but they

become significantly more complicated. We consider the simplest case so as not to obscure the main ideas of the analysis.

3.3.1 C^1-continuity of Subdivision away from Extraordinary Vertices

Most subdivision schemes are constructed from regular schemes, which are known to produce at least C^1-continuous surfaces in the regular setting for almost any initial configuration of control points. If our subdivision rules are local, we can take advantage of this knowledge to show that the surfaces generated by the scheme are C^1-continuous for almost any choice of control points anywhere *away from extraordinary vertices*. We call a subdivision scheme local, if only a finite number of control points is used to compute any new control point, and does not exceed a fixed number L for all subdivision levels and all control points.

One can demonstrate, as we did for the curves, that for any triangle T of the domain the surface $f(T)$ is completely determined by only a finite number of control points corresponding to vertices around T. For example, for the Loop scheme, we need only control points for vertices that are adjacent to the triangle. (see Figure 3.7). This is true for triangles at any subdivision level.

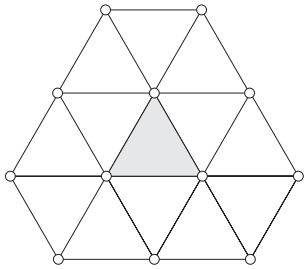


Figure 3.7: Control set for a triangle for the three-directional box spline.

To show this, fix a point x of the domain $|K|$ (not necessarily a vertex). For any level j, x is contained in a face of the domain; if x is a vertex, it is shared by several faces. Let $U^j(x)$ be the collection of faces on level j containing x, the *1-neighborhood* of x. The 1-neighborhood of a vertex can be identified with a k-gon in the plane, where k is the valence. We need j to be large enough so that all neighbors of triangles in $U^j(x)$ are free of extraordinary vertices. Unless x is an extraordinary vertex, this is easily achieved. $f(U^j(x))$ will be regular (see Figure 3.8).

This means that $f(U^j(x))$ is identical to a part of the surface corresponding to a regular mesh, and is

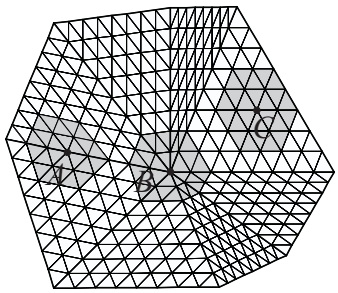


Figure 3.8: 2-neighborhoods (1-neighborhood of 1-neighborhood) of vertices *A*, *C* contain only regular vertices; this is not the case for *B*, which is an extraordinary vertex.

therefore C^1-continuous for almost any choice of control points by assumption.[1]

3.3.2 Smoothness Near Extraordinary Vertices

Now that we know that surfaces generated by our scheme are (at least) C^1-continuous away from the extraordinary vertices, all we have to do is find a a smooth parameterization near each extraordinary vertex, or establish that no such parameterization exists.

Consider the extraordinary vertex *B* in Figure 3.8. After sufficient number of subdivision steps, we will get a 1-neighborhood U^j of *B*, such that all control points defining $f(U^j)$ are regular, except *B* itself. This demonstrates that it is sufficient to determine if the scheme generates C^1-continuous surfaces for a very specific type of domains *K*: triangulations of the plane which have a single extraordinary vertex in their center, surrounded by regular vertices. We can assume all triangles of these triangulations to be identical (see Figure 3.9) and call such triangulations *k*-regular.

At first, the task still seems to be very difficult: for any configuration of control vertices, we have to find a parameterization of $f(U^j)$. However, it turns out that the problem can be further simplified.

We outline the idea behind a *sufficient* condition for C^1-continuity proposed by Reif [19]. This criterion tells us when the scheme is guaranteed to produce C^1-continuous surfaces, but if it fails, it is still possible that the scheme might be C^1.

We need two closely related tools to formulate the criterion: the *subdivision matrix* and the *characteristic map*. It turns out that rather than trying to consider all possible surfaces generated by subdivision, it is typically sufficient to look at a single map—the characteristic map. You have already seen subdivision

[1] Our argument is informal, and there are certain unusual cases when it fails; see [22] for details.

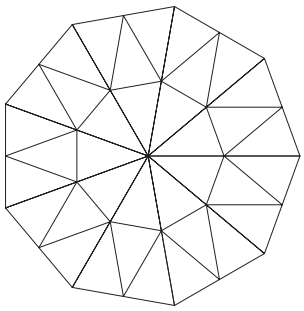


Figure 3.9: k-regular triangulation for $k = 9$.

matrices for one-dimensional schemes. Now we are going to extend this concept to two dimensions.

Subdivision matrix For the Loop scheme control points for only two rings of vertices around an extraordinary vertex B define $f(U)$ completely. We will call the set of vertices in these two rings the *control set* of U.

Let p_0^j be the value at level j of the control point corresponding to B. Assign numbers to the vertices in the two rings (there are $3k$ vertices). Note that U^j and U^{j+1} are similar: one can establish a one-to-one correspondence between the vertices simply by shrinking U^j by a factor of 2. Enumerate the vertices in

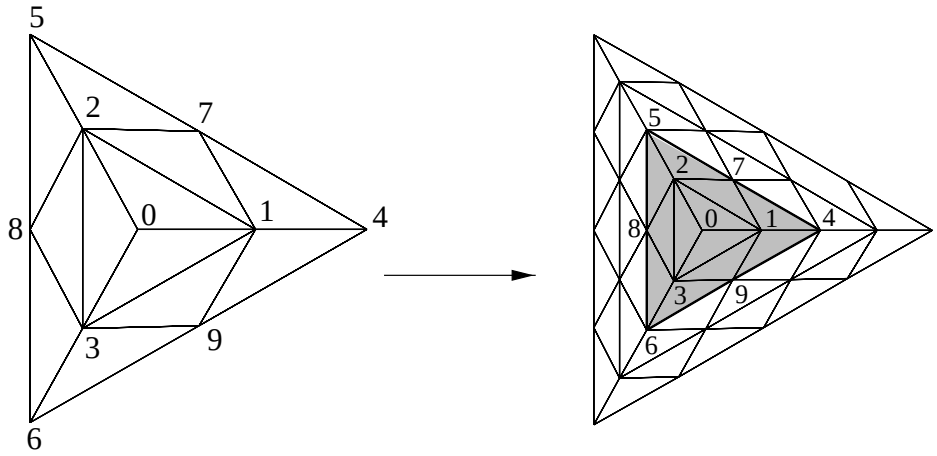


Figure 3.10: The Loop subdivision scheme near a vertex of degree 3. Note that $3 \times 3 + 1 = 10$ points in two rings are required.

the rings; there is total of $3k$ vertices, plus the vertex in the center. Let p_i^j, $i = 1 \ldots 3k$ be the corresponding

control points.

By definition of the control set, we can compute all values p_i^{j+1} from the values p_i^j. Because we only consider subdivision which computes finer levels by linear combination of points from the coarser level, the relation between the vectors of points $\mathbf{p}^{j+1}$ and $\mathbf{p}^j$ is given by a $3k+1 \times 3k+1$ matrix:

$$\begin{pmatrix} p_0^{j+1} \\ \vdots \\ p_{3k}^{j+1} \end{pmatrix} = S \begin{pmatrix} p_0^j \\ \vdots \\ p_{3k}^j \end{pmatrix}.$$

It is important to remember that each component of p^j is a point in the three-dimensional space. The matrix S is the subdivision matrix, which, in general, can change from level to level. We consider only schemes for which it is fixed. Such schemes are called *stationary*.

We can now rewrite each of the coordinate vectors in terms of the eigenvectors of the matrix S (compare to the use of eigen vectors in the 1D setting). Thus,

$$\mathbf{p}^0 = \sum_i a_i x_i$$

and

$$\mathbf{p}^j = (\mathrm{S})^j \mathbf{p}^0 = \sum_i (\lambda_i)^j a_i x_i$$

where the x_i are the eigenvectors of S, and the λ_i are the corresponding eigenvalues, arranged in nonincreasing order. As discussed for the one-dimensional case, λ_0 has to be 1 for all subdivision schemes, in order to guarantee invariance with respect to translations and rotations. Furthermore, all stable, converging subdivision schemes will have all the remaining λ_i less than 1.

Subdominant eigenvalues and eigenvectors It is clear that as we subdivide, the behavior of $\mathbf{p}^j$, which determines the behavior of the surface in the immediate vicinity of our point of interest, will depend only on the eigenvectors corresponding to the largest eigenvalues of S.

To proceed with the derivation, we will assume for simplicity that $\lambda = \lambda_1 = \lambda_2 > \lambda_3$. We will call λ_1 and λ_2 *subdominant eigenvalues*. Furthermore, we let $a_0 = 0$; this corresponds to choosing the origin of our coordinate system in the limit position of the vertex of interest (just as we did in the 1D setting). Then we can write

$$\frac{\mathbf{p}^j}{(\lambda)^j} = a_1 x_1 + a_2 x_2 + \left(\frac{\lambda_3}{\lambda}\right)^j x_3 \ldots \tag{3.1}$$

where the higher-order terms disappear in the limit.

This formula is very important, and deserves careful consideration. Recall that $\mathbf{p}^j$ is a vector of $3k+1$ 3D points, while x_i are vectors of $3k+1$ numbers. Hence the coefficients a_i in the decomposition above have to be 3D points.

This means that, up to a scaling by $(\lambda)^j$, the control set for $f(U)$ approaches a fixed configuration. This configuration is determined by x_1 and x_2, which depend only on the subdivision scheme, and on a_1 and a_2 which depend on the initial control mesh.

Each vertex in $\mathbf{p}^j$ for sufficiently large j is a linear combination of a_1 and a_2, up to a vanishing term. This indicates that a_1 and a_2 span the tangent plane. Also note that if we apply an affine transform A, taking a_1 and a_2 to coordinate vectors e_1 and e_2 in the plane, then, up to a vanishing term, the scaled configuration will be independent of the initial control mesh. The transformed configuration consists of 2D points with coordinates $(x_{1,i}, x_{2,i})$, $i = 0 \ldots 3k$, which depend on the subdivision matrix.

Informally, this indicates that up to a vanishing term, all subdivision surfaces generated by a scheme differ near an extraordinary point only by an affine transform. In fact, this is not quite true: it may happen that a particular configuration $(x_{1,i}, x_{2,i})$, $i = 0 \ldots 3k$ does not generate a surface patch, but, say, a curve. In that case, the vanishing terms will have influence on the smoothness of the surface.

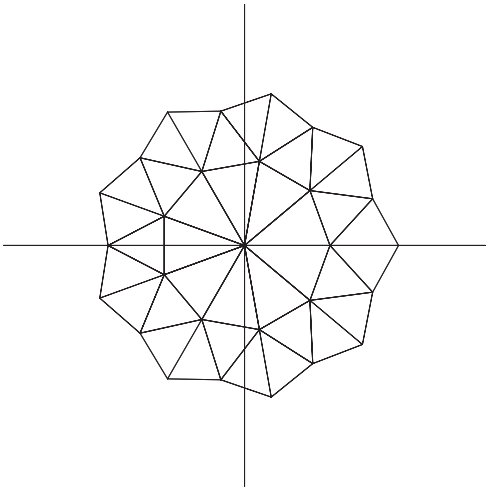


Figure 3.11: Control set of the characteristic map for $k = 9$.

3.3.3 Characteristic Map

Our observations motivate the definition of the *characteristic map*. Informally speaking, any subdivision surface generated by a scheme looks near an extraordinary vertex of valence k like the characteristic map of that scheme for valence k.

Note that when we described subdivision as a function from the plane to $\mathbf{R}^3$, we may use control vertices not from $\mathbf{R}^3$, but from $\mathbf{R}^2$; clearly, subdivision rules can be applied in the plane rather then in space. Then in the limit we obtain a map from the plane into the plane. The characteristic map is a map of this type.

As we have seen, the configuration of control points near an extraordinary vertex approaches $a_1x_1 + a_2x_2$, up to a scaling transformation. This means that the part of the surface defined on the k-gon U^j as $j \to \infty$, and scaled by the factor $1/\lambda^j$, approaches the surface defined by the vector of control points $a_1x_1 + a_2x_2$. Let $f[\mathbf{p}] : U \to \mathbf{R}^3$ be the limit surface generated by subdivision on U from the control set p.

Definition 3 *The characteristic map of a subdivision scheme for a valence k is the map $\Phi : U \to \mathbf{R}^2$ generated by the vector of 2D control points $e_1x_1 + e_2x_2$: $\Phi = f[e_1x_1 + e_2x_2]$, where e_1 and e_2 are unit coordinate vectors, and x_1 and x_2 are subdominant eigenvectors.*

Regularity of the characteristic map Inside each triangle of the k-gon U, the map is C^1: the argument of Section 3.3.1 can be used to show this. Moreover, the map has one-sided derivatives on the boundaries of the triangles, except at the extraordinary vertex, so we can define one-sided Jacobians on the boundaries of triangles too. We will say that the characteristic map is *regular* if its Jacobian is not zero anywhere on U excluding the extraordinary vertex but including the boundaries between triangles.

The regularity of the characteristic map has a geometric meaning: any subdivision surface can be written, up to a scale factor λ^j, as

$$f[\mathbf{p}^j](t) = A\Phi(t) + a(t)O\left((\lambda_3/\lambda)^j\right),$$

$t \in U^j$, $a(t)$ a bounded function $U^j \to \mathbf{R}^3$, and A is a linear transform taking the unit coordinate vectors in the plane to a_1 and a_2. Differentiating along the two coordinate directions t_1 and t_2 in the parametric domain U^j, and taking a cross product, after some calculations, we get the expression for the normal to the surface:

$$\mathbf{n}(t) = (a_1 \times a_2)J[\Phi(t)] + O\left((\lambda_3/\lambda)^{2j}\right)\tilde{a}(t)$$

where $J[\Phi]$ is the Jacobian, and $\tilde{a}(t)$ some bounded vector function on U^j.

The fact that the Jacobian does not vanish for Φ means that the normal is guaranteed to converge to $a_1 \times a_2$; therefore, the surface is tangent plane continuous.

Now we need to take only one more step. If, in addition to regularity, we assume that Φ is injective, we can invert it and parameterize any surface as $f(\Phi^{-1}(s))$, where $s \in \Phi(U)$. Intuitively, it is clear that up to a vanishing term this map is just an affine map, and is differentiable. We omit a rigorous proof here. For a complete treatment see [19]; for more recent developments, see [22] and [24].

We arrive at the following condition, which is the basis of smoothness analysis of all subdivision schemes considered in these notes.

Reif's sufficient condition for smoothness. Suppose the eigenvectors of a subdivision matrix form a basis, the largest three eigenvalues are real and satisfy

$$\lambda_0 = 1 > \lambda_1 = \lambda_2 > |\lambda_3|$$

If the characteristic map is regular, then almost all surfaces generated by subdivision are tangent plane continuous; if the characteristic map is also injective, then almost all surfaces generated by subdivision are C^1-continuous.

Note: Reif's original condition is somewhat different, because he defines the characteristic map on an annular region, rather than on a k-gon. This is necessary for applications, but makes it somewhat more difficult to understand.

In the next section, we will discuss the most popular stationary subdivision schemes, all of which have been proved to be C^1-continuous at extraordinary vertices. These proofs are far from trivial: checking the conditions of Reif's criterion is quite difficult, especially checking for injectivity. In most cases calculations are done in symbolic form and use closed-form expressions for the limit surfaces of subdivision [20, 7, 12, 13]. In [23] an interval-based approach is described, which does not rely on closed-form expressions for limit surfaces, and can be applied, for example, to interpolating schemes.

3.3.4 Tangents and Limit Positions

Similar to the one-dimensional case, the coefficients a_0 a_1 and a_2 in the decomposition 3.1 are the limit position of the control point for the central vertex v_0, and two tangents respectively. To compute these coefficients, we need corresponding left eigenvectors. In the descriptions of subdivision schemes in the next section we describe these calculations whenever information is available.

3.4 Subdivision Zoo

3.4.1 Overview of Subdivision Schemes

In this section we describe most known stationary subdivision schemes generating C^1-continuous surfaces on arbitrary meshes. Without doubt, our discussion is not exhaustive even as far as stationary schemes are concerned. There are even wholly different classes of subdivision schemes, most importantly variational schemes, that we do not discuss here. Variational subdivision is described in the part of the notes written by Leif Kobbelt.

At a first glance, the variety of existing schemes might appear chaotic. However, there is a straightforward way to classify most of the schemes based on three criteria:

- the type of refinement rule (vertex insertion or corner-cutting);
- the type of generated mesh (triangular or quadrilateral);
- whether the scheme is approximating or interpolating.

The following table shows this classification:

Vertex insertion		
	Triangular meshes	*Quadrilateral meshes*
Approximating	Loop	Catmull-Clark
Interpolating	Modified Butterfly	Kobbelt

Corner-cutting
Doo-Sabin
Midedge

It can be seen from this table that there is little replication in functionality: most schemes produce substantially different types of surfaces. Now we consider our classification criteria in greater detail.

First, we note that each subdivision scheme defined on meshes of arbitrary topology is based on a *regular subdivision scheme* such as a subdivision schemes for splines, for example. Our classification is primarily a classification of regular subdivision schemes—once such a scheme is fixed, additional rules have to be specified only for extraordinary vertices or faces that cannot be part of a regular mesh.

Mesh type Regular subdivision schemes act on regular control meshes, that is, vertices of the mesh correspond to regularly spaced points in the plane. However, the faces of the mesh can be formed in different ways. For a regular mesh, it is natural to use faces that are identical. If, in addition, we assume

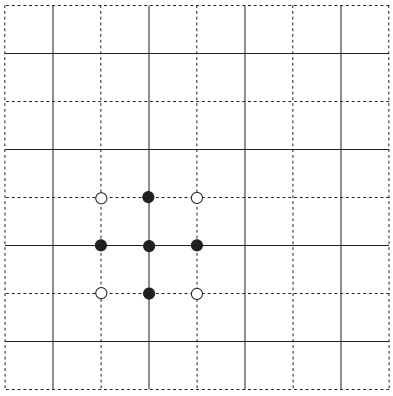
that the faces are regular polygons, it turns out that there are only three ways to choose the face polygons: we can use only squares, equilateral triangles and regular hexagons. Meshes consisting of hexagons are not very common, and the first two types of tiling are the most convenient for practical purposes. These leads to two types of regular subdivision schemes: those defined for quadrilateral tilings, and those defined for triangular tilings.

Vertex insertion and corner-cutting Once the tiling of the plane is fixed, we have to define how a refined tiling generated by the scheme is related to the original tiling. There are two main approaches that are used to generate a refined tiling: one is *vertex insertion* and the other is *corner cutting* (see Figure 3.12). The schemes using the first method are often called *primal*, and the schemes using the second method are called *dual*. In the first case, each edge of a triangular or a quadrilateral mesh is split into two, old vertices of the mesh are retained, and new vertices inserted on edges are connected. For quadrilaterals, an additional vertex is inserted for each face.

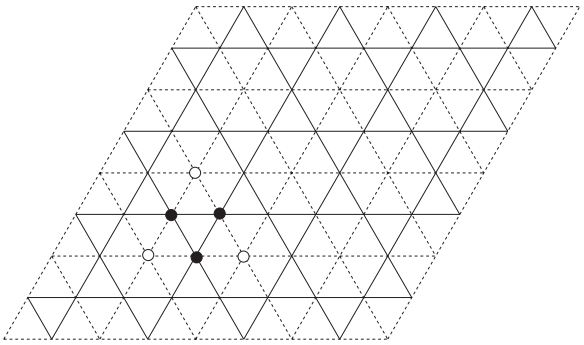
In the second case, for each old face, a new similar face is created inside of it and the newly created faces are connected. As a result, we get four new vertices for each old edge, a new face for each edge and each vertex. The old vertices are discarded. Geometrically, one can think about this process as first cutting off the vertices, and then cutting off the edges of a polyhedron. For quadrilateral tilings, this can be done in such a way that the refined tiling has only quadrilateral faces. For triangles, we can get only a hexagonal tiling. Thus, a regular corner-cutting algorithm for triangles would have to alternate between triangular and hexagonal tilings.

Approximation vs. Interpolation Vertex insertion schemes can be interpolating or approximating: as the vertices of the coarser tiling are also vertices of the refined tiling, for each vertex a sequence of control points, corresponding to different subdivision levels, is defined. If all points in the sequence are the same, we say that the scheme is interpolating. Otherwise, we call it approximating. Interpolation is an attractive feature in more than one way. First, the original control points defining the surface are also points of the limit surface, which allows one to control it in a more intuitive manner. Second, many algorithms can be considerably simplified, and many calculations can be performed "in place." Unfortunately, the quality of these surfaces is not as high as the quality of surfaces produced by approximating schemes, and the schemes do not converge as fast to the limit surface as the approximating schemes.

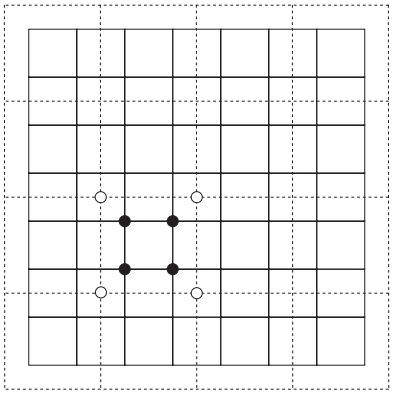
We concentrate primarily on vertex insertion schemes; we discuss one corner-cutting scheme, Doo-Sabin. The Midedge subdivision scheme, proposed by Habib and Warren [6], and independently discovered by Peters and Reif [13], is discussed in the part of these notes written by Jörg Peters.



Vertex insertion for quadrilaterals



Vertex insertion for triangles



Corner-cutting for quadrilaterals

Figure 3.12: Enumeration of vertices of a mesh near an extraordinary vertex; for a boundary vertex, the $0-th$ sector is adjacent to the boundary.

3.4.2 Notation and Terminology

Here we summarize the notation that we use in subsequent sections. Some of it was already introduced earlier.

Regular and extraordinary vertices. We have already seen that subdivision schemes defined on triangular meshes create new vertices only of valence 6 in the interior. On the boundary, the newly created vertices have valence 4. Similarly, on quadrilateral meshes both vertex-insertion and

corner-cutting schemes create only vertices of valence 4 in the interior, and 3 on the boundary. Hence, after several subdivision steps, most vertices in a mesh will have one of these valences (6 in the interior, 4 on the boundary for triangular meshes, 4 in the interior, 3 on the boundary for quadrilateral). The vertices with these valences are called *regular*, and vertices of other valences *extraordinary*.

Notation for vertices near a fixed vertex. In Figure 3.13 we show the notation that we use for vertices of quadrilateral and triangular subdivision schemes near a fixed vertex. Typically, we need it for extraordinary vertices; we also use it for regular vertices, to describe calculations of limit positions and tangent vectors. Note that this notation is for a fixed level; the names of vertices changes from one level to the next. For brevity, we denote the value $p^j(v^j_{i,l})$ by $p^j_{i,l}$.

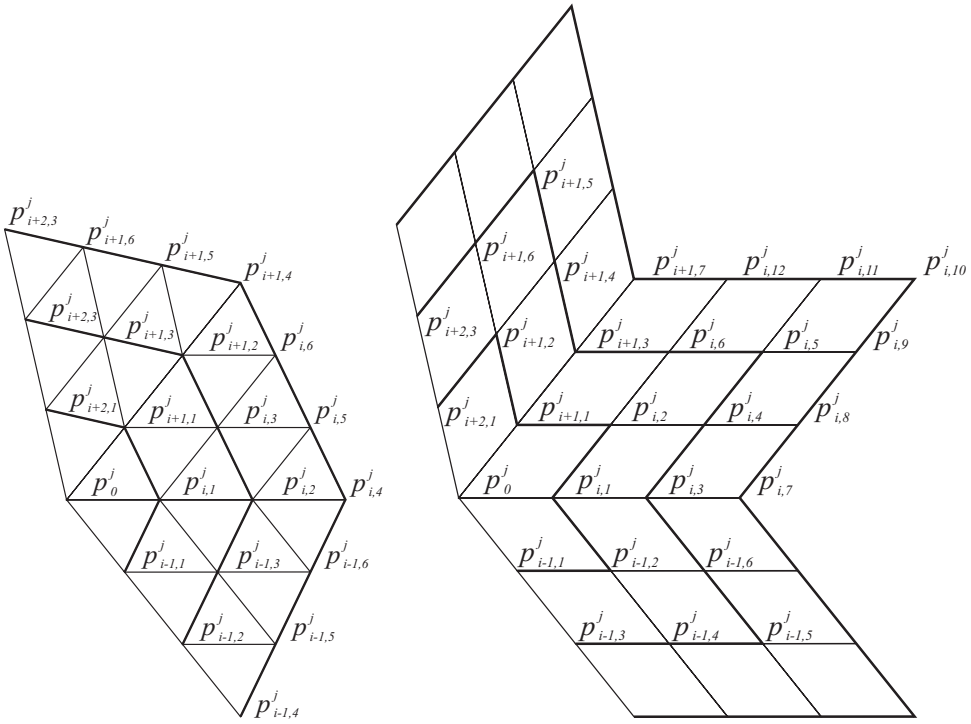


Figure 3.13: Enumeration of vertices of a mesh near an extraordinary vertex; for a boundary vertex, the $0-th$ sector is adjacent to the boundary.

Odd and even vertices. For vertex insertion (primal) schemes, the vertices of the coarser mesh are also

vertices of the refined mesh. For any subdivision level, we call all new vertices that are created at that level, *odd vertices*. This term comes from the one-dimensional case, when vertices of the control polygons can be enumerated sequentially and on any level the newly inserted vertices are assigned odd numbers. The vertices inherited from the previous level are called *even*. (See also Chapter 2).

Face and edge vertices. For triangular schemes (Loop and Modified Butterfly), there is only one type of odd vertex. For quadrilateral schemes, some vertices are inserted when edges of the coarser mesh are split, other vertices are inserted for a face. These two types of odd vertices are called *edge* and *face* vertices respectively.

Boundaries and creases. Typically, special rules have to be specified on the boundary of a mesh. These rules are commonly chosen in such a way that the boundary curve of the limit surface does not depend on any interior control vertices, and is smooth or piecewise smooth (C^1 or C^2). The same rules can be used to introduce sharp features into C^1-surfaces: some interior edges can be *tagged* as crease edges, and boundary rules are applied for all vertices that are inserted on such edges.

Masks. We often specify a subdivision rule by providing its *mask*. The mask is a picture showing which control points are used to compute a new control point, which we denote with a black dot. The numbers are the coefficients of the subdivision rule. For example, if p_1, p_2 are vertices of an edge, and v_3 and v_4 are the other two vertices of the triangles that share this edge, then the Loop subdivision rule for an interior odd vertex v depicted in Figure 3.14, can be written as

$$p^{j+1}(v) = \frac{3}{8}p^j(v_1) + \frac{3}{8}p^j(v_2) + \frac{1}{8}p^j(v_3) + \frac{1}{8}p^j(v_4)$$

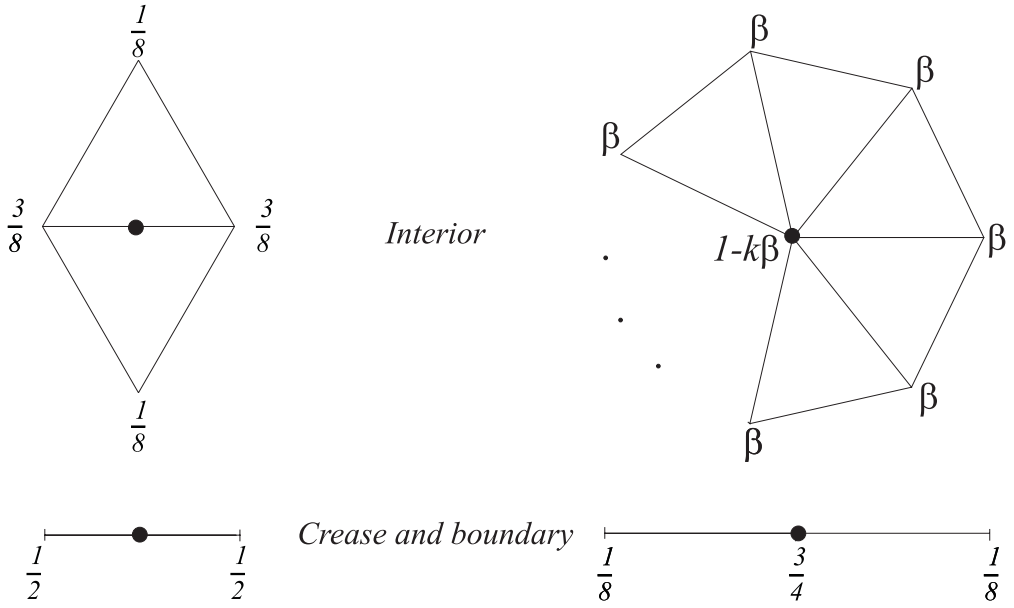
3.4.3 Loop Scheme

The Loop scheme is a simple approximating vertex insertion scheme for triangular meshes proposed by Charles Loop [10]. C^1-continuity of this scheme for valences up to 100, including the boundary case, was proved by Schweitzer [20]. The proof for all valences can be found in [22].

The scheme is based on the *three-directional box spline*, which produces C^2-continuous surfaces on the regular meshes. The Loop scheme produces surfaces that are C^2-continuous everywhere except at extraordinary vertices, where they are C^1-continuous. Hoppe, DeRose, Duchamp et al. [8] proposed a piecewise C^1-continuous extension of the Loop scheme, with special rules defined for edges.

The scheme can be applied to arbitrary polygonal meshes, after the mesh is converted to a triangular mesh, for example, by triangulating each polygonal face.

Subdivision rules The masks for the Loop scheme are shown in Figure 3.14. For boundaries and edges tagged as *crease* edges, special rules are used. These rules produce a cubic spline curve along the boundary/crease. The curve does not depend on the control points which correspond to vertices that are outside the boundary/crease.



a. Masks for odd vertices *b. Masks for even vertices*

Figure 3.14: The Loop subdivision. In the picture above, β can be chosen to be either $\frac{1}{n}(5/8-(\frac{3}{8}+\frac{1}{4}\cos\frac{2\pi}{n})^2)$ (original choice of Loop [10]), or, for $n>3$, $\beta=\frac{3}{8n}$ as proposed by Warren [21]. For $n=3$, $\beta=3/16$ can be used.

In [8], the rules for extraordinary crease vertices and their neighbors on the crease were modified to produce tangent plane continuous surfaces on either side of the crease (or on one side of the boundary). In practice, for reasons discussed in [25], this modification does not lead to a significant difference in the appearance of the surface. At the same time, as a result of this modification, the boundary curve becomes dependent on the valences of vertices on the curve. This is a disadvantage in situations when two surfaces have to be joined together along a boundary. It appears that in practically all cases it is safe to use the rules shown in Figure 3.14. Although the surface will not be formally C^1 near vertices of valence greater than 7, the result will be visually indistinguishable from a C^1-surface obtained with

modified rules, with the additional advantage of independence of the boundary from the interior.

If it is necessary to ensure C^1-continuity, we propose a different modification. Rather than modifying the rules for the boundary curve, and making it dependent on the valence of vertices, we modify rules for interior odd vertices adjacent to an extraordinary vertex. For $n < 7$, no modification is necessary. For $n > 7$, it is sufficient to use the mask shown in Figure 3.15. Then the limit surface can be shown to be C^1 on the boundary.

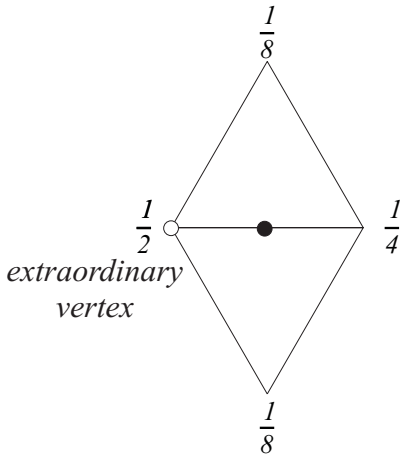


Figure 3.15: Modified rule for odd vertices adjacent to a boundary extraordinary vertex.

Tangent vectors The rules for computing tangent vectors for Loop's scheme are especially simple. To compute a pair of tangent vectors at an interior vertex, use

$$\begin{aligned} t_1 &= \sum_{i=0}^{k-1} \cos\frac{2\pi}{k} p(v_{i,1}) \\ t_2 &= \sum_{i=0}^{k-1} \sin\frac{2\pi}{k} p(v_{i,1}). \end{aligned} \tag{3.2}$$

These formulas can be applied to the control points at any subdivision level.

Quite often, the tangent vectors are used to compute a normal. The normal obtained as the cross product $t_1 \times t_2$ can be interpreted geometrically. This cross product can be written as a weighted sum of normals to all possible triangles with vertices $p(v)$, $p(v_{i,1})$, $p(v_{l,1})$, $i,l = 0 \ldots k-1$, $i \neq l$. The standard way of obtaining vertex normals for a mesh by averaging the normals of triangles adjacent to a vertex,

can be regarded as a first approximation to the normals given by the formulas above. At the same time, it is worth observing that computing normals as $t_1 \times t_2$ is less expensive than averaging the normals of triangles. The geometric nature of the normals obtained in this way suggests that they can be used to compute approximate normals for other schemes, even if the precise normals require more complicated expressions.

At a boundary vertex, the tangent along the curve is computed using $t_{along} = p(v_{0,1}) - p(v_{k-1,1})$. The tangent across the boundary/crease is computed as follows [8]:

$$\begin{aligned} t_{across} &= p(v_{0,1}) + p(v_{1,1}) - 2p(v_0) \quad \text{for } k = 2 \\ t_{across} &= p(v_{2,1}) - p(v_0) \quad \text{for } k = 3 \\ t_{across} &= \sin\theta\,(p(v_{0,1}) + p(v_{k-1,1})) + (2\cos\theta - 2)\sum_{i=1}^{k-2} \sin i\theta\, p(v_{i,1}) \quad \text{for } k \geq 4 \end{aligned} \tag{3.3}$$

where $\theta = \pi/(k-1)$. These formulas apply whenever the scheme is tangent plane continuous at the boundary; it does not matter which method was used to ensure tangent plane continuity.

Limit positions Another set of simple formulas allows one to compute limit positions of control points for a fixed vertex, that is, the limit $\lim_{j\to\infty} p^j(v)$ for a fixed v. For interior vertices, the mask for computing the limit value at an interior vertex is the same as the mask for computing the value on the next level, with β replaced by $\chi = \frac{1}{3/8\beta + n}$.

For boundary vertices, the formula is always

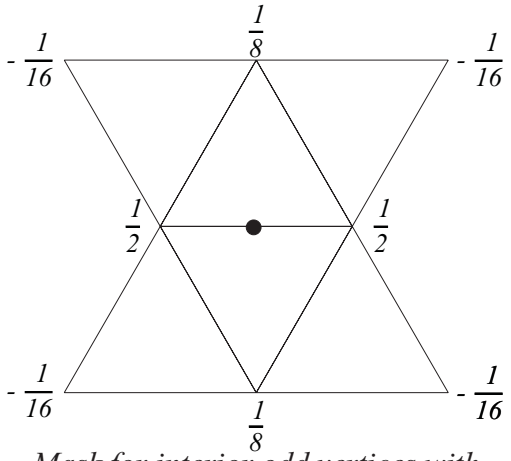
$$p^\infty(v_0) = \frac{1}{5}p(v_{0,1}) + \frac{3}{5}p(v_0) + \frac{1}{5}p(v_{1,k-1})$$

This expression is similar to the rule for even boundary vertices, but with different coefficients. However, different formulas have to be used if the rules on the boundary are modified as in [8].

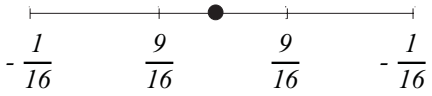
3.4.4 Modified Butterfly Scheme

The Butterfly scheme was proposed by Dyn, Gregory and Levin in [5]. However, although the original Butterfly scheme is defined on arbitrary triangular meshes, the limit surface is not C^1-continuous at extraordinary points of valence $k = 3$ and $k > 7$ [22]. It is C^1 on regular meshes.

Unlike approximating schemes based on splines, this scheme does not produce piecewise polynomial surfaces in the limit. In [26] a modification of the Butterfly scheme was proposed, which guarantees that the scheme produces C^1-continuous surfaces for arbitrary meshes (for a proof see [22]). The scheme is known to be C^1 but not C^2 on regular meshes. The masks for the the scheme are shown in Figure 3.16.

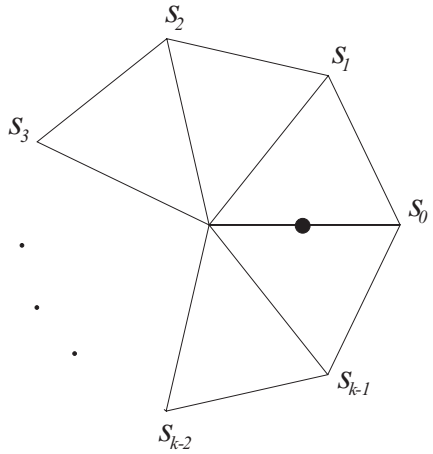


Mask for interior odd vertices with regular neighbors



Mask for crease and boundary vertices

a. Masks for odd vertices



b. Mask for odd vertices adjacent to an extraordinary vertex

Figure 3.16: Modified Butterfly subdivision. The coefficients s_i are $\frac{1}{k}\left(\frac{1}{4}+\cos\frac{2i\pi}{k}+\frac{1}{2}\cos\frac{4i\pi}{k}\right)$ for $k>5$. For $k=3$, $s_0=\frac{5}{12}$, $s_{1,2}=-\frac{1}{12}$; for $k=4$, $s_0=\frac{3}{8}$, $s_2=-\frac{1}{8}$, $s_{1,3}=0$.

The tangent vectors at extraordinary interior vertices can be computed using the same rules as for the Loop scheme. For regular vertices, the formulas are more complex: in this case, we have to use control points in a 2-neighborhood of a vertex. If the control points are arranged into a vector $p=[p_0,p_{0,1},p_{0,2},p_{0,3},\ldots,p_{0,5},p_{5,3}]$ of length 19, than the tangents are given by scalar products $(l_1\cdot p)$ and

$(l_2 \cdot p)$, where the vectors l_1 and l_2 are

$$l_1 = \left[0, 16, -8, -8, 16, -8, -8, -\frac{8\sqrt{3}}{3}, \frac{4\sqrt{3}}{3}, \frac{4\sqrt{3}}{3}, -\frac{8\sqrt{3}}{3}, \frac{4\sqrt{3}}{3}, \frac{4\sqrt{3}}{3}, 1, -\frac{1}{2}, -\frac{1}{2}, 1, -\frac{1}{2}, -\frac{1}{2}\right]$$
$$l_2 = \left[0, 0, 8, -8, 0, 8, -8, 0, \frac{4\sqrt{3}}{3}, -\frac{4\sqrt{3}}{3}, 0, \frac{4\sqrt{3}}{3}, -\frac{4\sqrt{3}}{3}, 0, -\frac{1}{2}, \frac{1}{2}, 0, -\frac{1}{2}, \frac{1}{2}\right] \tag{3.4}$$

Because the scheme is interpolating, no formulas are needed to compute the limit positions: all control points are on the surface. On the boundary, the four point subdivision scheme is used [4]. To achieve C^1-continuity on the boundary, special coefficients have to be used (see [25] for details).

3.4.5 Catmull-Clark Scheme

The Catmull-Clark scheme was described in [2]. It is based on the tensor product bicubic spline. The masks are shown in Figure 3.17. The scheme produces surfaces that are C^2 everywhere except at extraordinary vertices, where they are C^1. The tangent plane continuity of the scheme was analyzed by Ball and Storry [1], and C^1-continuity by Peters and Reif [12]. The values of α and β can be chosen from a wide range (see Figure 3.18).

On the boundary, using the coefficients for the cubic spline produces acceptable results. See [25] for further details about the behavior on the boundary.

The rules of Catmull-Clark scheme are defined for meshes with quadrilateral faces. Arbitrary polygonal mesh can be reduced to a quadrilateral mesh using a more general form of Catmull-Clark rules [2]:

- a face control point for an n-gon is computed as the average of the corners of the polygon;
- an edge control point as the average of the endpoints of the edge and newly computed face control points of adjacent faces;
- the formula for even control points can be chosen in different ways; the original formula is

$$p^{j+1}(v) = \frac{k-2}{k} p^j(v) + \frac{1}{k^2} \sum_{i=0}^{k-1} p^j(v_i) + \frac{1}{k^2} \sum_{i=0}^{k-1} p^{j+1}(v_i^f)$$

  where v_i are the vertices adjacent to v on level j, and v_i^f are face vertices on level $j+1$ corresponding to faces adjacent to v.

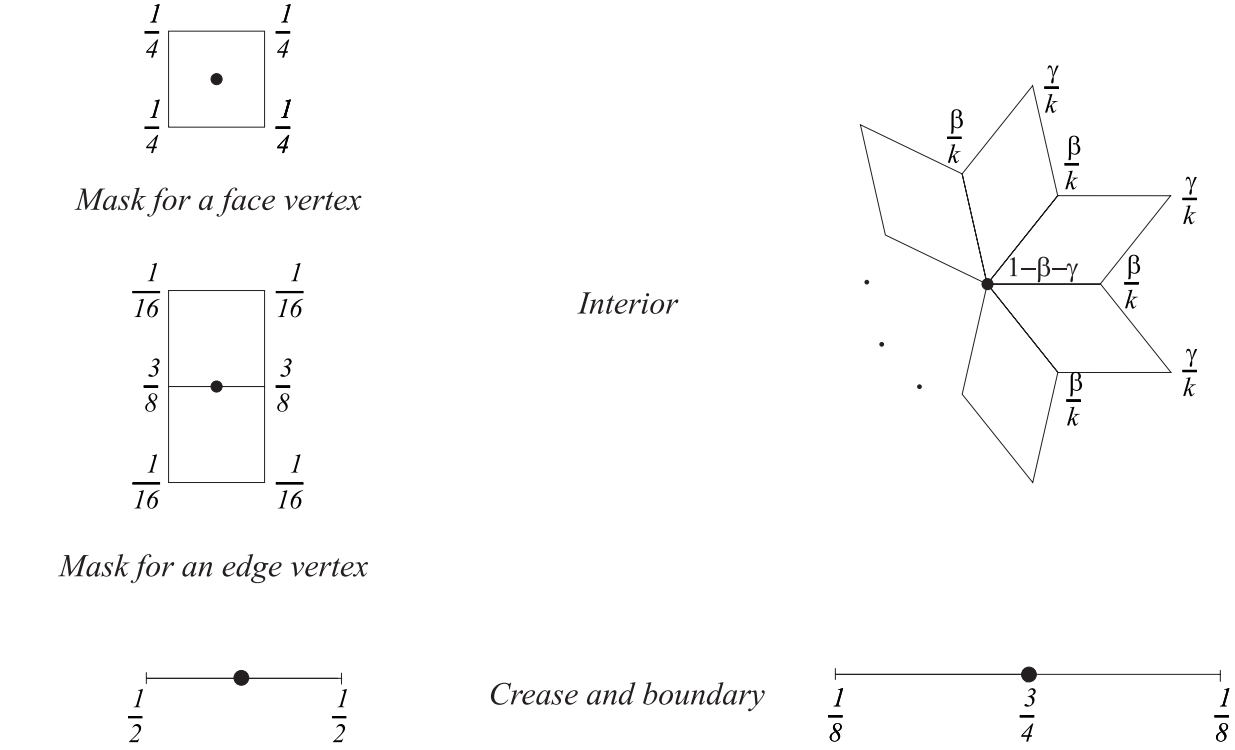


Figure 3.17: Catmull-Clark subdivision. Catmull and Clark [2] suggest the following coefficients for rules at extraordinary vertices: $\beta = \frac{3}{2k}$ and $\gamma = \frac{1}{4k}$

3.4.6 Kobbelt Scheme

This interpolating scheme was described by Kobbelt in [9]. For regular meshes, it reduces to the tensor product of four point schemes. C^1-continuity of this scheme for interior vertices for all valences is proven in [23].

Crucial for the construction of this scheme is the observation (valid for any tensor-product scheme) that the face control points can be computed in two steps: first, all edge control points are computed.

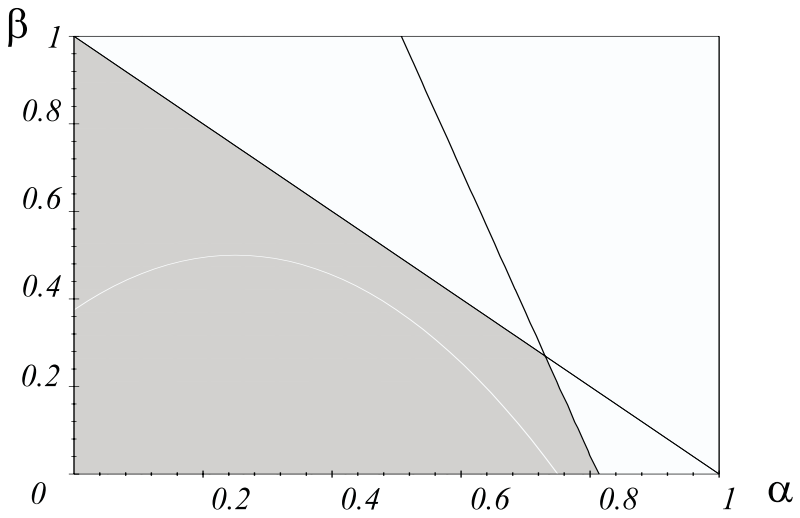


Figure 3.18: Ranges for coefficients α and β of the Catmull-Clark scheme; $\alpha = 1-\gamma-\beta$ is the coefficient of the central vertex.

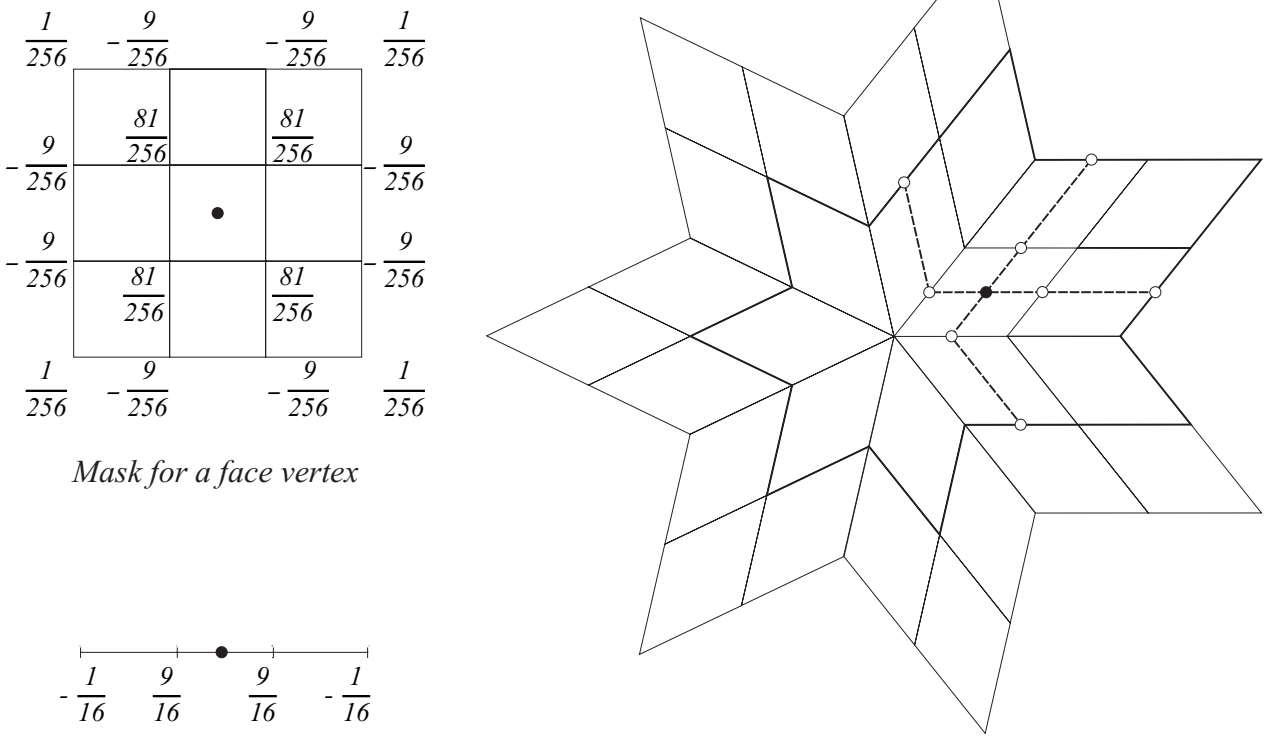
Next, face vertices are computed using the *edge rule* applied to a sequence of edge control points on the same level. As shown in Figure 3.19, there are two ways to compute a face vertex in this way. In the regular case, the result is the same. Assuming this method of computing all face control points, only one rule of the regular scheme is modified: the edge odd control points adjacent to an extraordinary vertex are computed differently. Specifically,

$$p_{i,1}^{j+1} = (\frac{1}{2}-w)p_0^j + (\frac{1}{2}-w)p_{i,1}^j + wp_i^j + wp_{i,3}^j$$
$$v_i^j = \frac{4}{k}\sum_{i=0}^{k-1} p_{i,1}^j - (p_{i-1,1}^j + p_{i,1}^j + p_{i+1,1}^j) - \frac{w}{1/2-w}(p_{i-2,2}^j + p_{i-1,2}^j + p_{i,2}^j + p_{i+1,2}^j) + \frac{4w}{(1/2-w)k}\sum_{i=0}^{k-1} p_{i,2}^j \tag{3.5}$$

where $w = -1/16$ (also, see Figure 3.13 for notation). On the boundaries and creases, the four point subdivision rule is used.

Unlike other schemes, eigenvectors of the subdivision matrix cannot be computed explicitly; hence, there are no precise expressions for tangents. In any case, the effective support of this scheme is too large for such formulas to be of practical use: typically, it is sufficient to subdivide several times and then use, for example, the formulas for the Loop scheme (see discussion in the section on the Loop scheme).

For more details on this scheme, see the part of the notes written by Leif Kobbelt.



Mask for a face vertex

Mask for edge, crease and boundary vertices

a. Regular masks

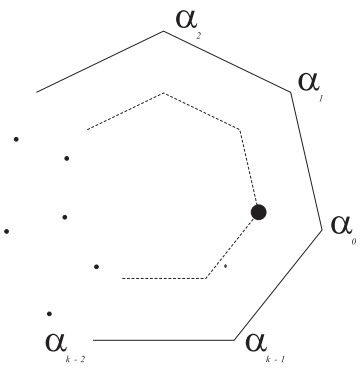
b. Computing a face vertex adjacent to an extraordinary vertex

Figure 3.19: Kobbelt subdivision.

3.4.7 Doo-Sabin Scheme

The Doo-Sabin subdivision is quite simple conceptually: there is no distinction between odd and even vertices, and a single mask is sufficient to define the scheme. A special rule is required only for the boundaries, where the limit curve is a quadratic spline. It was observed by Doo that this can also be achieved by replicating the boundary edge, i.e., creating a quadrilateral with two coinciding pairs of vertices. Nasri [11] describes other ways of defining rules for boundaries.

C^1-continuity for schemes similar to the Doo-Sabin schemes was analyzed by Peters and Reif [12].



Mask for interior vertices

Mask for boundary vertices

Figure 3.20: The Doo-Sabin subdivision. The coefficients are defined by the formulas $\alpha_0 = 1/4 + 5/4k$ and $\alpha_i = (3 + 2\cos(2i\pi/k))/4k$, for $i = 1 \ldots k-1$

3.5 Limitations of Stationary Subdivision

Stationary subdivision, while overcoming certain problems inherent in spline representations, still has a number of limitations. Most problems are much more apparent for interpolating schemes than for approximating schemes. In this section we briefly discuss a number of these problems.

Problems with curvature continuity While it is possible to obtain subdivision schemes which are C^2-continuous, there are indications that such schemes either have very large support [17, 15], or necessarily have zero curvature at extraordinary vertices. A compromise solution was recently proposed by Umlauf [16]. Nevertheless, this limitation is quite fundamental: degeneracy or discontinuity of curvature typically leads to visible defects of the surface.

Decrease of smoothness with valence For some schemes, as the valence increases, the magnitude of the third largest eigenvalue approaches the magnitude of the subdominant eigenvalues. As an example we consider surfaces generated by the Loop scheme near vertices of high valence.

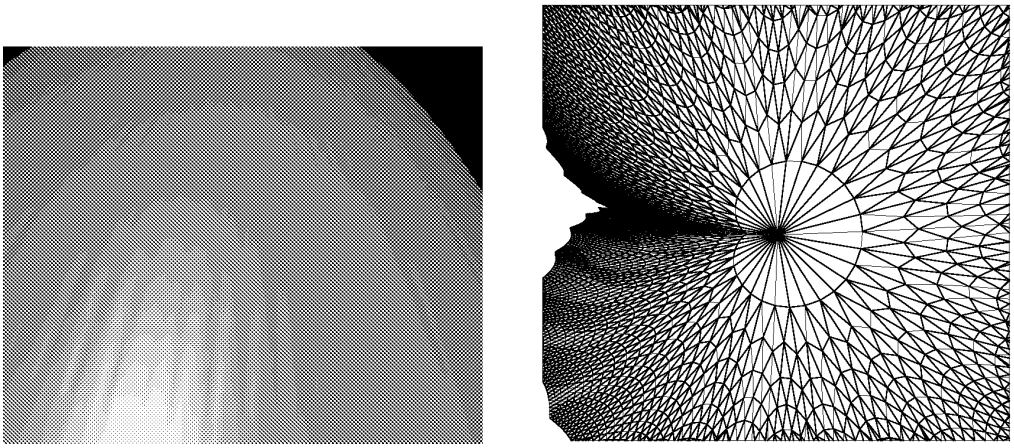


Figure 3.21: Left: ripples on a surface generated by the Loop scheme near a vertex of large valence; Right: mesh structure for the Loop scheme near an extraordinary vertex with a significant "high-frequency" component; a crease starting at the extraordinary vertex appears.

In Figure 3.21 (right side), one can see a typical problem that occurs because of "eigenvalue clustering:" a crease might appear, abruptly terminating at the vertex. In some cases this behavior may be desirable, but our goal is to make it controllable rather than let the artifacts appear by chance.

Ripples Another problem, presence of ripples in the surface close to an extraordinary point, is also shown in Figure 3.21. It is not clear whether this artifact can be eliminated. It is closely related to the curvature problem.

Uneven structure of the mesh On regular meshes, subdivision matrices of C^1-continuous schemes always have subdominant eigenvalue $1/2$. When the eigenvalues of subdivision matrices near extraordinary vertices significantly differ from $1/2$, the structure of the mesh becomes uneven: the ratio of the size of triangles on finer and coarser levels adjacent to a given vertex is roughly proportional to the magnitude of the subdominant eigenvalue. This effect can be seen clearly in Figure 3.23.

Optimization of subdivision rules It is possible to eliminate eigenvalue clustering, as well as the difference in eigenvalues of the regular and extraordinary case by prescribing the eigenvalues of the subdivision matrix and deriving suitable subdivision coefficients. This approach was used to derive coefficients of the Butterfly scheme.

As expected, the meshes generated by the modified scheme have better structure near extraordinary points (Figure 3.22). However, the ripples become larger, so one kind of artifact is traded for another. It is, however, possible to seek an optimal solution or one close to optimal; alternatively, one may resort to a family of schemes that would provide for a controlled tradeoff between the two artifacts.

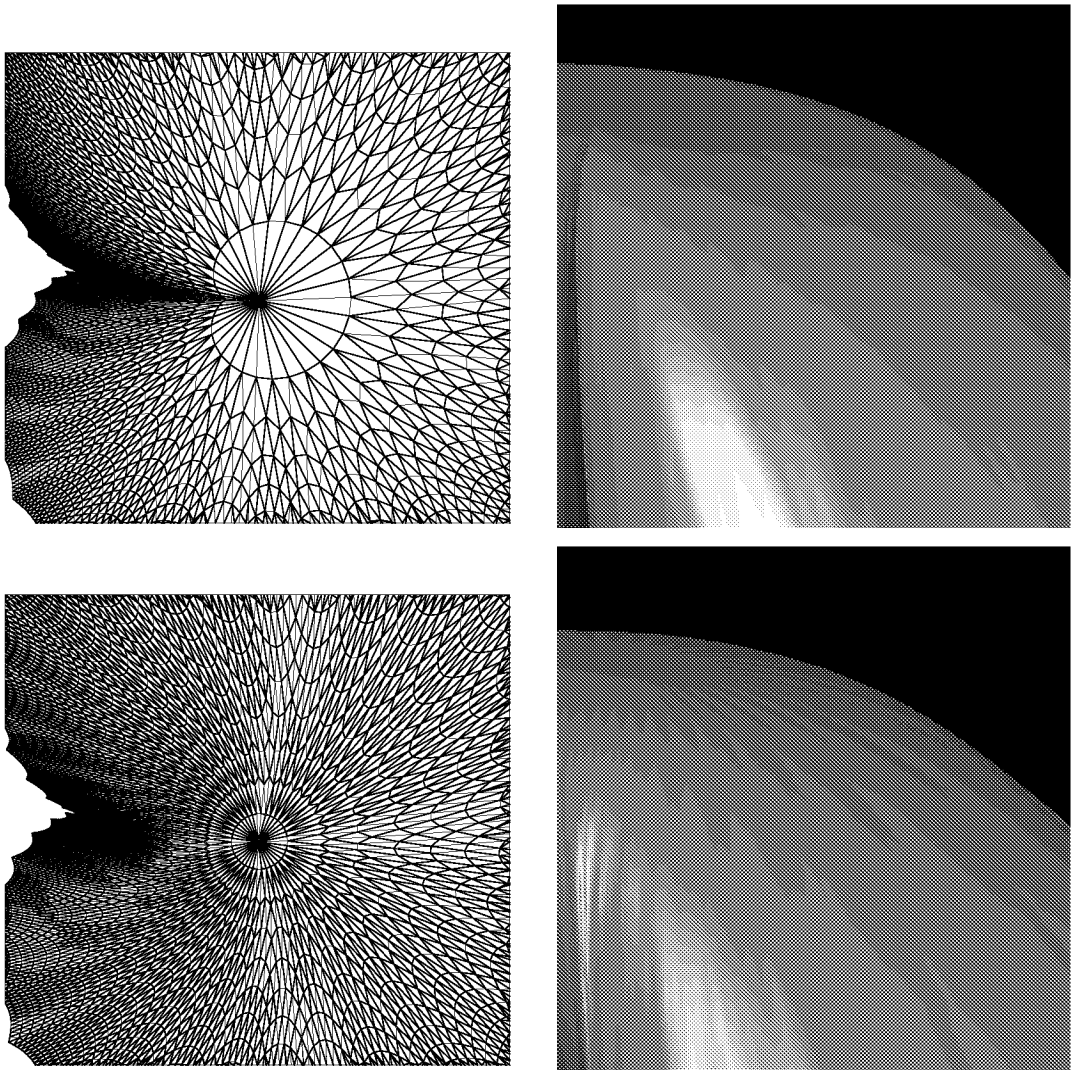


Figure 3.22: Left: mesh structure for the Loop scheme and the modified Loop scheme near an extraordinary vertex; a crease does not appear for the modified Loop. Right: shaded images of the surfaces for Loop and modified Loop; ripples are more apparent for modified Loop.

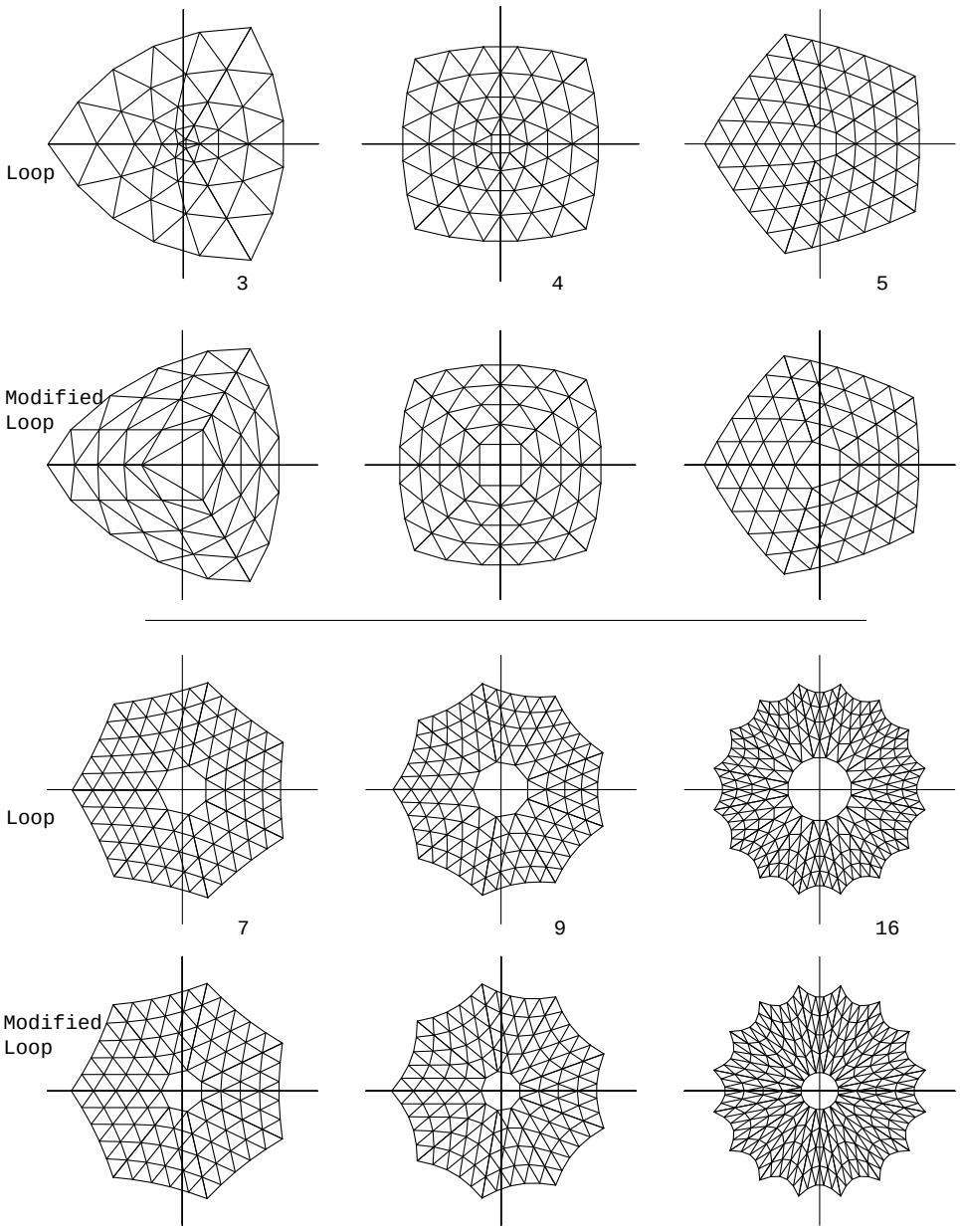


Figure 3.23: Comparison of control nets for Loop scheme and modified Loop scheme. Note that for Loop scheme the size of the hole in the ring (1-neighborhood removed) is very small relatively to the surrounding triangles for valence 3 and becomes larger as k grows. For modified Loop scheme this size remains constant.

Bibliography

[1] BALL, A. A., AND STORRY, D. J. T. Conditions for Tangent Plane Continuity over Recursively Generated B-Spline Surfaces. *ACM Trans. Gr. 7*, 2 (1988), 83–102.

[2] CATMULL, E., AND CLARK, J. Recursively Generated B-Spline Surfaces on Arbitrary Topological Meshes. *Computer Aided Design 10*, 6 (1978), 350–355.

[3] DOO, D., AND SABIN, M. Analysis of the Behaviour of Recursive Division Surfaces near Extraordinary Points. *Computer Aided Design 10*, 6 (1978), 356–360.

[4] DYN, N., GREGORY, J. A., AND LEVIN, D. A Four-Point Interpolatory Subdivision Scheme for Curve Design. *Comput. Aided Geom. Des. 4* (1987), 257–268.

[5] DYN, N., LEVIN, D., AND GREGORY, J. A. A Butterfly Subdivision Scheme for Surface Interpolation with Tension Control. *ACM Trans. Gr. 9*, 2 (April 1990), 160–169.

[6] HABIB, A., AND WARREN, J. Edge and Vertex Insertion for a Class of C^1 Subdivision Surfaces. presented at 4th SIAM COnference on Geometric Design, November 1995.

[7] HABIB, A., AND WARREN, J. Edge and Vertex Insertion for a Class of Subdivision Surfaces. Preprint. Computer Science, Rice University, 1996.

[8] HOPPE, H., DEROSE, T., DUCHAMP, T., HALSTEAD, M., JIN, H., MCDONALD, J., SCHWEITZER, J., AND STUETZLE, W. Piecewise Smooth Surface Reconsruction. In *Computer Graphics Proceedings*, Annual Conference Series, 295–302, 1994.

[9] KOBBELT, L. Interpolatory Subdivision on Open Quadrilateral Nets with Arbitrary Topology. In *Proceedings of Eurographics 96*, Computer Graphics Forum, 409–420, 1996.

[10] LOOP, C. Smooth Subdivision Surfaces Based on Triangles. Master's thesis, University of Utah, Department of Mathematics, 1987.

[11] NASRI, A. H. Polyhedral Subdivision Methods for Free-Form Surfaces. *ACM Trans. Gr. 6*, 1 (January 1987), 29–73.

[12] PETERS, J., AND REIF, U. Analysis of generalized B-spline subdivision algorithms. *SIAM Jornal of Numerical Analysis* (1997).

[13] PETERS, J., AND REIF, U. The simplest subdivision scheme for smoothing polyhedra. *ACM Trans. Gr. 16(4)* (October 1997).

[14] PRAUTZSCH, H. Analysis of C^k-subdivision surfaces at extraordianry points. Preprint. Presented at Oberwolfach, June, 1995, 1995.

[15] PRAUTZSCH, H., AND REIF, U. Necessary Conditions for Subdivision Surfaces. 1996.

[16] PRAUTZSCH, H., AND UMLAUF, G. Improved Triangular Subdivision Schemes. In *Proceedings of the CGI '98*, 1998.

[17] REIF, U. A Degree Estimate for Polynomial Subdivision Surface of Higher Regularity. Tech. rep., Universität Stuttgart, Mathematisches Institut A, 1995. preprint.

[18] REIF, U. Some New Results on Subdivision Algorithms for Meshes of Arbitrary Topology. In *Approximation Theory VIII*, C. K. Chui and L. Schumaker, Eds., vol. 2. World Scientific, Singapore, 1995, pp. 367–374.

[19] REIF, U. A Unified Approach to Subdivision Algorithms Near Extraordinary Points. *Comput. Aided Geom. Des. 12* (1995), 153–174.

[20] SCHWEITZER, J. E. *Analysis and Application of Subdivision Surfaces*. PhD thesis, University of Washington, Seattle, 1996.

[21] WARREN, J. Subdivision Methods for Geometric Design. Unpublished manuscript, November 1995.

[22] ZORIN, D. *Subdivision and Multiresolution Surface Representations*. PhD thesis, Caltech, Pasadena, 1997.

[23] ZORIN, D. A method for analysis of C^1-continuity of subdivision surfaces. submitted to SIAM Journal of Numerical Analysis, 1998.

[24] ZORIN, D. Smoothness of subdivision on irregular meshes. submitted to Constructive Approximation, 1998.

[25] ZORIN, D. Smoothness of subdivision surfaces on the boundary. preprint, Computer Science Department, Stanford University, 1998.

[26] ZORIN, D., SCHRÖDER, P., AND SWELDENS, W. Interpolating Subdivision for Meshes with Arbitrary Topology. *Computer Graphics Proceedings (SIGGRAPH 96)* (1996), 189–192.

Chapter 4

Some Properties of Subdivision Derived from Splines

Author: Jörg Peters

This section discusses some properties of dual subdivision schemes such as Doo-Sabin and generalized 4 direction box spline subdivision, as well as some properties of the Catmull-Clark scheme. In a second part (co-authored with Ahmad Nasri) it is shown how to compute the volume of a closed subdivision surface.

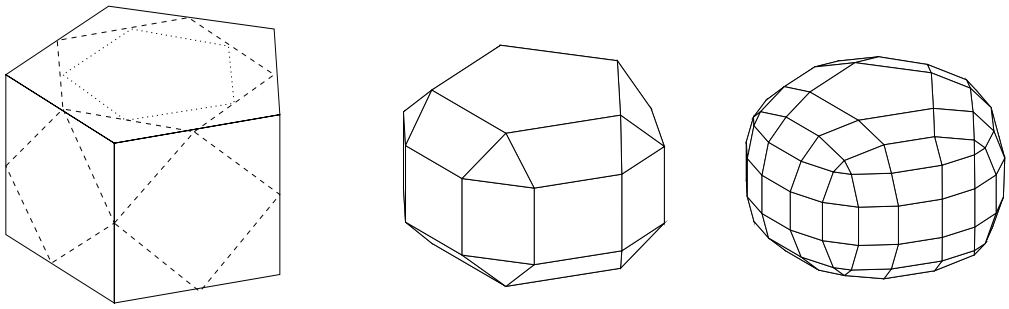


Figure 1: Midedge Subdivision (double steps)

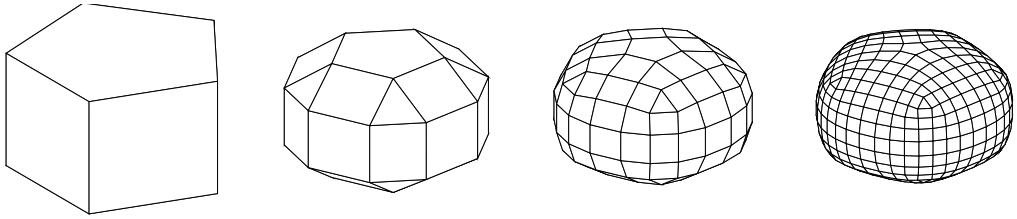


Figure 2: Doo-Sabin subdivision

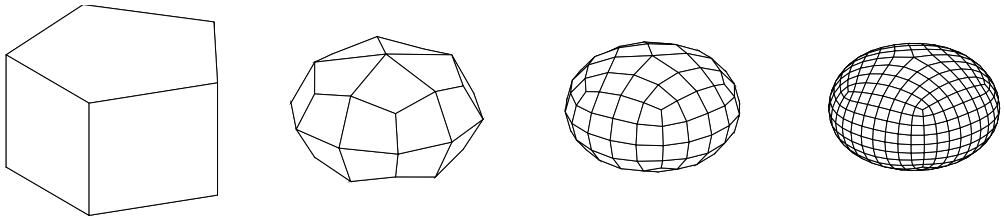


Figure 3: Catmull-Clark subdivision.

1 Subdivision schemes generalizing box spline subdivision

The idea of generating smooth free-form surfaces of arbitrary topology by iterated mesh refinement dates back at least to 1978, when two papers [2], [3] appeared back to back in the same issue of Computer Aided Design. Named after their inventors, the Doo-Sabin and the Catmull-Clark algorithm represent generalizations of the refinement schemes for biquadratic and bicubic B-splines, respectively. By combining a construction principle of striking simplicity with high fairness of the generated surfaces, both algorithms have since become standard tools in graphics (cf. Figures 2 and 3).

An even simpler subdivison scheme, midedge subdivision Figure 1,*left*, was exhibited and analyzed in [6]: given an input polyhedron with not necessarily planar facets, midedge subdivision *connects every edge-midpoint to the four midpoints of the edges that share both a vertex and a face with the current edge.* Once all midpoints are linked, the old mesh is discarded. Since the subdivision mask consists of only two points and there is only one rule regardless of the connectivity of the polyhedron, the subdivision scheme is as simple as it can be. Figure 1 shows every second subdivision step.

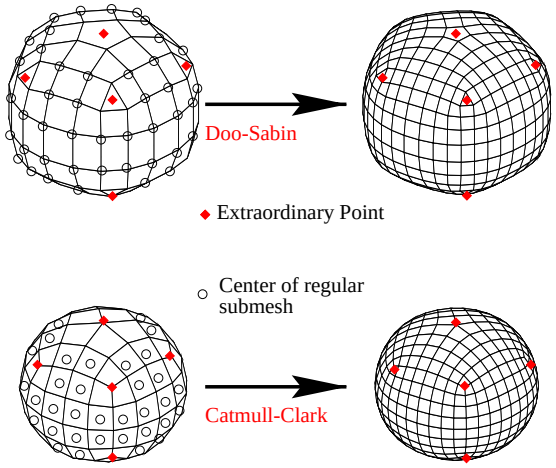


Figure 4: Diamonds mark extraordinary points on the subdivision polyhedra. The centers of regular submeshes in the less refined polyhedra on the *left* are marked by circles. A regular submesh of the Doo-Sabin scheme (*top*) consists of four quadrilaterals surrounding a vertex. A regular submesh of the Catmull-Clark scheme (*bottom*) consists of eight quadrilaterals surrounding a quadrilateral.

2 Polynomial heritage

The three schemes displayed have in common that they generate increasing regions whose points all have the same valence and whose facets all have the same number of edges. A submesh of points and facets with this standard valence and number of edges is called *regular.* Specific subdivision rules like those of Doo and Sabin, Catmull and Clark, or midedge subdivision derive their appeal from the fact that the limit surface is explicitly known for regular meshes; the limit surfaces are respectively biquadratic tensor-product spline surfaces, bicubic tensor-product spline surfaces and surfaces formed as a linear combination of shifts of a 4-direction piecewise quadratic box spline. Loop's scheme [4] generalizes a 3-direction box spline.

To understand the heritage and implications of this regularity property, we focus on the simplest subdivision scheme. In midedge subdivision every new non-boundary point has

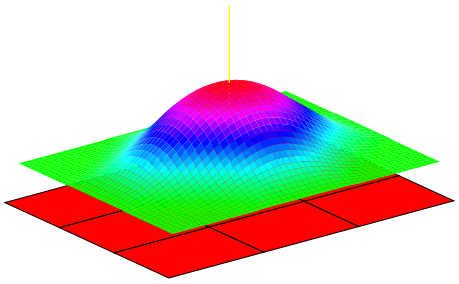


Figure 5: The four-direction box spline.

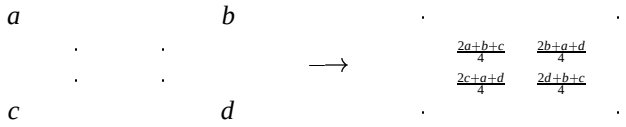


Figure 6: Four-direction box spline subdivision.

exactly four neighbors, and every mesh point is replaced by a quadrilateral. A regular submesh consists therefore of quadrilaterals and 4-valent points.

We want to see that midedge subdivision rule on a regular submesh generates a piecewise quadratic surface. To this end we compare the subdivision process with the subdivision process of a 4-direction box spline surface.

A 4-direction box spline surface is a piecewise quadratic surface formed as a linear combination of the shifts of a basis function called the Zwart-Powell element ([11],[8]). The function is shown in Figure 5; [6] describes it in terms of the Bernstein-Bézier form.

The 4-direction box spline surface is alternatively defined by the following subdivision process (c.f. [1]). *At each step replicate each coefficient with index* (i, j) *in a new array at positions* $(2i, 2j)$, $(2i+1, 2j)$, $(2i, 2j+1)$, *and* $(2i+1, 2j+1)$. *Then average in the new array first all entries* i, j *with* $i+1, j+1$ *and then all entries* i, j *with* $i-1, j+1$. Figure 6 shows the result of the averaging process: for each quadrilateral each of four new points is obtained by taking $1/2$ of one point and $1/4$ of each of its two neighbors. Thus two steps of the midedge subdivision equal one step of the 4-direction box-spline subdivision. (see also [10] p280).

That is, *On a regular mesh, midedge subdivision converges to a surface parametrized by shifts of the 4-direction box-spline.*

Since we know the limit surface for regular mesh regions, we will in the following always combine two steps of midedge subdivision and refer to it as a double-step. The double-step rule is efficiently represented by the *mask*

$$\begin{array}{ccc} & 1 & \\ {}^{1/4} & 2 & 1 \end{array}$$

which applies to a vertex and its two neighbors in the same facet. Note that in contrast to the Doo-Sabin and other subdivision masks, the (double) midedge subdivision mask is not parametrized by the valence n of the particular point, but is uniform for all configurations.

3 Tangent continuity at extraordinary points – an example

We will look at the tangent continuity at extraordinary points in the context of the simplest subdivision scheme and summarize similar findings for the Doo-Sabin and the Catmull Clark algorithm at the end of the section.

Since each midedge subdivision double-step replaces points and edges by quadrilaterals, the new mesh at each step is regular except for a fixed number of increasingly separated non-four-sided mesh facets (c.f. Figure 4). We consider one such facet (cf. Figure 7). At each step the polygon defining the facet contracts, converging towards the average of the points. To confirm the existence of a well-defined surface at this extraordinary point, we analyze the local *subdivision matrix* that maps k layers of old points surrounding the extraordinary point to k layers of new points. For midedge subdivision $k = 3$. For almost all inputs, the eigenvalues and eigenvectors of the subdivision matrix determine the smoothness of the limit surface [9].

The labeling used here is shown in Figure 8. The vector of control points $\mathbf{B}_m$ that define the mth surface layer consists of n blocks of 9 elements, each and the subdivision matrix A transforms

$$\mathbf{B}_{m+1} = A\mathbf{B}_m = A^{m+1}\mathbf{B}_0.$$

The eigenvalues of the subdivision matrix A are real-valued

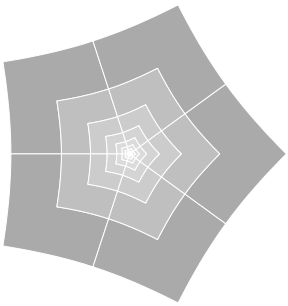


Figure 7: Union of surface layers at an extraordinary point.

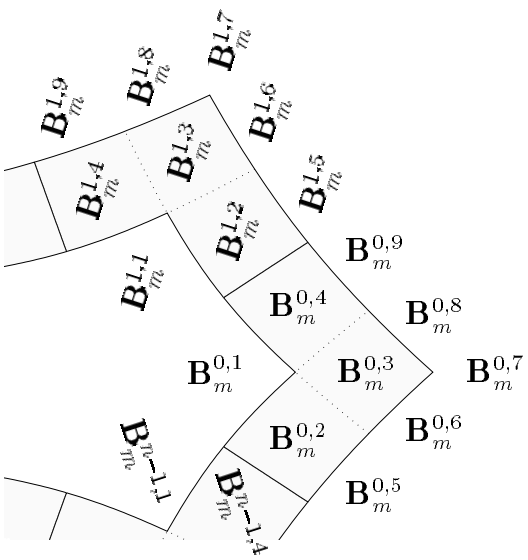


Figure 8: Labeling of control points transformed by the subdivision matrix.

and form a decreasing sequence starting with

$$1, \lambda, \lambda, \text{ where } \lambda := \frac{1+\cos(\frac{2\pi}{n})}{2}.$$

The single maximal eigenvalue 1 assures convergence to an affinely invariant surface. To verify its smoothness, it suffices to show that the eigenvectors corresponding to the two subdominant eigenvalues λ define a regular and injective map, namely the characteristic map that parametrizes the tangent plane in the limit point. For the details see [6].

A similar analysis of the Doo-Sabin and the Catmull Clark algorithm in [7] settled the properties of these schemes 20 years after their inception. Here is the summary.

- The Doo-Sabin algorithm in its general form uses weights $\alpha = [\alpha^0, \ldots, \alpha^{n-1}]$ for computing a new n-gon from an old one (c.f. Figure 11). Affine invariance and symmetry, i.e. $\sum_j \alpha^j = 1$, $\alpha^j = \alpha^{n-j}$, $j \in \mathbb{Z}_n$ imply that the discrete Fourier transform $\hat{\alpha} = [1, \hat{\alpha}^1, \hat{\alpha}^2, \ldots, \hat{\alpha}^2, \hat{\alpha}^1]$ is real. If $\lambda := \hat{\alpha}^1$ is greater in absolute value than the other entries except for 1 and if

  $$1/4 < \lambda < \lambda_{\max}(n)$$

  for values $\lambda_{\max}(n) < 1$ then the limit surface is smooth. The bound $\lambda_{\max}(n)$ is computed explicitly in [7]. If $1 > \lambda > \lambda_{\max}(n)$ then the limit is a continuous, non-smooth surface.

- The Doo-Sabin algorithm in its original form [3] generates smooth limit surfaces.

- The Catmull-Clark algorithm in its general form uses three weights α, β, γ summing up to one for computing the new location of an extraordinary vertex from its predecessor and the centers of its neighbors. If

  $$2\,|4\alpha - 1 \pm \sqrt{(4\alpha-1)^2 + 8\beta - 4}| < c_n + 5 + \sqrt{(c_n+9)(c_n+1)}$$

  with $c_n := \cos(2\pi/n)$, then the limit surface is smooth. If one of the both values on the left hand side exceeds the right hand side, then the limit surface is not smooth.

- The Catmull-Clark algorithm in its original form complies with the conditions and generates smooth limit surfaces.

4 Convergence improvement at irregular points

The need for an improvement in the convergence of the midedge subdivision is illustrated by Figure 9. The input data form a cylinder by extruding a 16-gon that is regular except that one of the vertices has been moved above the plane to form a peak. Convergence of the central 16-edge facet is slow. The subdominant eigenvalue $\lambda = (1+\cos(\frac{2\pi}{16}))/2 \approx .962$ implies that the distance of a vertex to the centroid of the facet shrinks by less than 4%. Conversely the convergence at the peak is so fast that it seems pointed. This fast convergence of triangular facets is already apparent in Figure 1 where the central triangle shrinks almost to a point in just 2 double-steps. To even out the speed of convergence, we modify the subdivision mask (averaging weights for creating new mesh points from old ones) for n-sided facets to have weights

$$\alpha_j = 2\sum_{i=0}^{\bar{n}} 2^{-i}\cos(ji\frac{2\pi}{n}) \text{ where } \bar{n} := \left\lfloor \frac{n-1}{2} \right\rfloor$$

resulting in eigenvalues

$$\hat{\alpha}_i = \hat{\alpha}_{n-i} = 2^{-i}, \quad i = 0..\bar{n}$$

with $\hat{\alpha}_{n/2} = 0$ for n even, consistent with the regular case. The improved convergence is illustrated by three examples of modified midedge subdivision in Figure 12. To distribute the change of normal and keep some features sharp, we parametrize the mask in the first subdivision double-step only as follows (cf. [5]):

$$\begin{matrix} \frac{1-\gamma}{2} \\ \gamma \end{matrix} \quad \frac{1-\gamma}{2} \qquad \gamma \in [0.5, 1].$$

In the extreme case, setting $\gamma = 1$ at a vertex results in a sharp point and setting $\gamma = 1$ for the two vertices of an edge results in a sharp edge. Conversely, setting $\gamma = 0.5$ at a point, smoothes maximally. The checker pattern on the objects in Figure 12 highlights individual triangles on the surfaces.

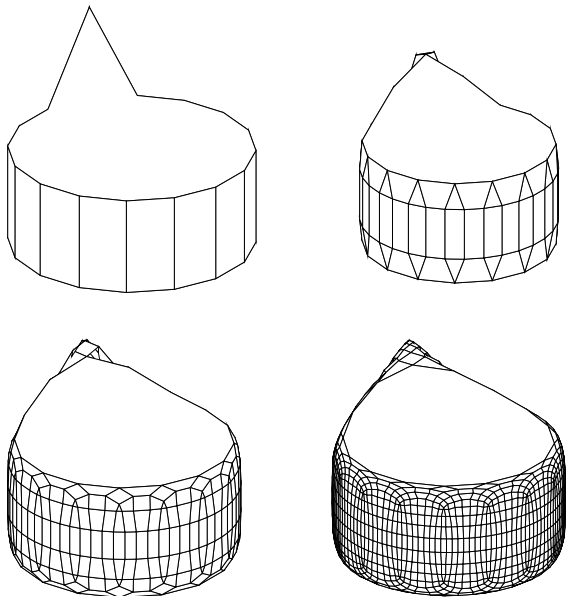


Figure 9: Fast convergence of the midedge subdivision for 3-sided facets and slow convergence for large facets.

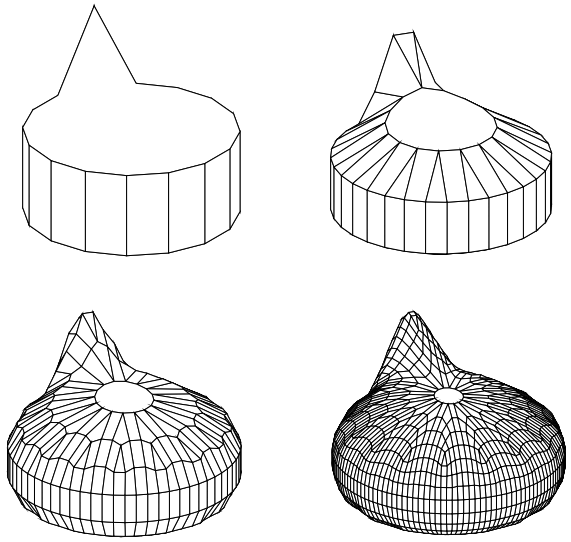


Figure 10: Balanced convergence of the modified midedge subdivision.

References

[1] BOOR, C. D., HÖLLIG, K., AND RIEMENSCHNEIDER, S. *Box splines*. Springer Verlag, 1994.

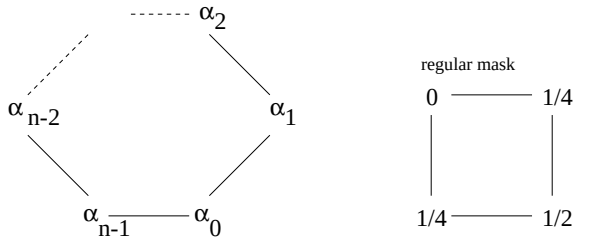


Figure 11: Subdision masks: rules for creating new mesh points from old ones.

[2] CATMULL, E., AND CLARK, J. Recursively generated B-spline surfaces on arbitrary topological meshes. *Computer Aided Design 10* (1978), 350–355.

[3] DOO, D., AND SABIN, M. A. Behaviour of recursive subdivision surfaces near extraordinary points. *Computer Aided Design 10* (1978), 356–360.

[4] LOOP, C. Smooth subdivision for surfaces based on triangles. Master's thesis, University of Utah, 1987.

[5] PETERS, J. Smooth free-form surfaces over irregular meshes generalizing quadratic splines. *Computer-Aided Geometric Design 10* (1993), 347–361.

[6] PETERS, J., AND REIF, U. The simplest subdivision scheme for smoothing polyhedra. *ACM Transactions on Graphics 16(4)* (October 1997).

[7] PETERS, J., AND REIF, U. Analysis of generalized B-spline subdivision algorithms. *SIAM Journal on Numerical Analysis 35*, 2 (April 1998).

[8] POWELL, M. Piecewise quadratic surface fitting for contour plotting. In *Software for Numerical Mathematics*, D. Evans, Ed. Academic Press, 1969.

[9] REIF, U. A unified approach to sudivision algorithms near extraordinary vertices. *Computer Aided Geometric Design 12* (1995), 153–174.

[10] SABIN, M. Recursive subdivision. In *The Mathematics of Surfaces*. Clarendon Press, Oxford, 1986, pp. 269–282.

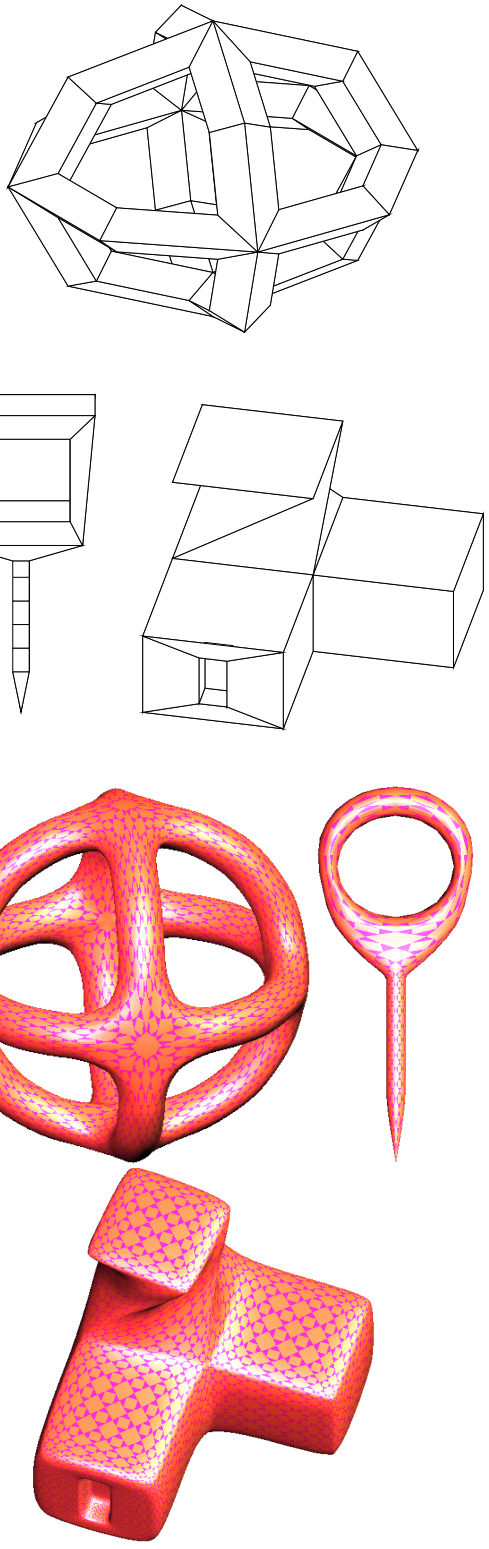


Figure 12: Modified midedge subdivision

[11] ZWART, P. Multivariate splines with non-degenerate partitions. *SIAM J. of Numer. Analysis 10* (1973), 665–673.

Chapter 5

Interpolatory Subdivision for Quad Meshes

Author: Leif Kobbelt

III Interpolatory Subdivison for Quad-Meshes

A simple interpolatory subdivision scheme for quadrilateral nets with arbitrary topology is presented which generates C^1 surfaces in the limit. The scheme satisfies important requirements for practical applications in computer graphics and engineering. These requirements include the necessity to generate smooth surfaces with local creases and cusps. The scheme can be applied to open nets in which case it generates boundary curves that allow a C^0-join of several subdivision patches. Due to the local support of the scheme, adaptive refinement strategies can be applied. We present a simple device to preserve the consistency of such adaptively refined nets.

The original paper has been published in:

L. Kobbelt
Interpolatory Subdivision on Open Quadrilateral Nets with Arbitrary Topology,
Computer Graphics Forum 15 (1996), Eurographics '96 issue, pp. 409–420

3.1 Introduction

The problem we address in this paper is the generation of smooth interpolating surfaces of arbitrary topological type in the context of practical applications. Such applications range from the design of free-form surfaces and scattered data interpolation to high quality rendering and mesh generation, e.g., in finite element analysis. The standard set-up for this problem is usually given in a form equivalent to the following:

A *net* $N = (V,F)$ representing the input is to be mapped to a *refined net* $N' = (V',F')$ which is required to be a sufficiently close approximation of a smooth surface. In this notation the sets V and V' contain the *data points* $\mathbf{p}_i, \mathbf{p}'_i \in \mathbb{R}^3$ of the input or output respectively. The sets F and F' represent the *topological information* of the nets. The elements of F and F' are finite sequences of points $s_k \subset V$ or $s'_k \subset V'$ each of which enumerates the corners of one not necessarily planar *face* of a net.

If all elements $s_k \in F$ have length four then N is called a *quadrilateral net*. To achieve interpolation of the given data, $V \subset V'$ is required. Due to the geometric background of the problem we assume N to be *feasible*, i.e., at each point $\mathbf{p}_i$ there exists a plane T_i such that the projection of the faces meeting at $\mathbf{p}_i$ onto T_i is injective. A net is *closed* if every edge is part of exactly two faces. In *open* nets, boundary edges occur which belong to one face only.

There are two major 'schools' for computing N' from a given N. The first or classic way of doing this is to explicitely find a collection of local (piecewise polynomial) parametrizations (*patches*) corresponding to the faces of N. If these patches smoothly join at common boundaries they form an overall smooth patch complex. The net N' is then obtained by sampling each patch on a sufficiently fine grid. The most important step in this approach is to find smoothly joining patches which represent a surface of arbitrary topology. A lot of work has been done in this field, e.g., [16], [15], [17] ...

Another way to generate N' is to define a *refinement operator* S which directly maps nets to nets without constructing an explicit parametrization of a surface. Such an operator performs both, a *topological* refinement of the net by splitting the faces and a *geometric* refinement by determining the position of the new points in order to reduce the angles between adjacent faces (*smoothing*). By iteratively applying S one produces a sequence of nets N_i with $N_0 = N$ and $N_{i+1} = S\,N_i$. If S has certain properties then the sequence $S^i N$ converges to a smooth limiting surface and we can set $N' := S^k N$ for some sufficiently large k. Algorithms of this kind are proposed in [2], [4], [14], [7], [10], and [11]. All these schemes are either non-interpolatory or defined on *triangular* nets which is not appropriate for some engineering applications.

The scheme which we present here is a *stationary refinement scheme* [9], [3], i.e., the rules to compute the positions of the new points use simple affine combinations of points from the unrefined net. The term *stationary* implies that these rules are the same on every refinement level. They are derived from a modification of the well-known four-point scheme [6]. This scheme refines polgons by $S : (\mathbf{p}_i) \mapsto (\mathbf{p}'_i)$ with

$$\begin{aligned} \mathbf{p}'_{2i} &:= \mathbf{p}_i \\ \mathbf{p}'_{2i+1} &:= \frac{8+\omega}{16}(\mathbf{p}_i + \mathbf{p}_{i+1}) - \frac{\omega}{16}(\mathbf{p}_{i-1} + \mathbf{p}_{i+2}) \end{aligned} \tag{11}$$

where $0 < \omega < 2(\sqrt{5}-1)$ is sufficient to ensure convergence to a smooth limiting curve [8]. The standard value is $\omega = 1$ for which the scheme has cubic precision. In order to minimize the number of special cases, we restrict ourselves to the refinement of quadrilateral nets. The faces are split as shown in Fig. 10 and hence, to complete the definition of the operator S, we need rules for new points corresponding to edges and/or faces of the unrefined net. To generalize the algorithm for interpolating arbitrary nets, a precomputing step is needed (cf. Sect. 3.2).

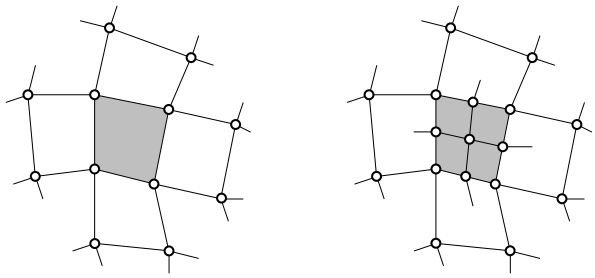


Figure 10: The refinement operator splits one quadrilateral face into four. The new vertices can be associated with the edges and faces of the unrefined net. All new vertices have valency four.

The major advantages that this scheme offers, are that it has the interpolation property *and* works on quadrilateral nets. This seems to be most appropriate for engineering applications (compared to non-interpolatory schemes or triangular nets), e.g., in finite element analysis since quadrilateral (bilinear) elements are less stiff than triangular (linear) elements [19]. The scheme provides the maximum flexibility since it can be applied to *open* nets with *arbitrary* topology. It produces smooth surfaces and yields the possibility to generate local creases and cusps. Since the support of the scheme is local, adaptive refinement strategies can be applied. We present a technique to keep adaptively refined nets C^0-consistent (cf. Sect. 3.6) and shortly describe an appropriate data structure for the implementation of the algorithm.

3.2 Precomputing: Conversion to Quadrilateral Nets

It is a fairly simple task to convert a given arbitrary net $\tilde{N}$ into a quadrilateral net N. One straightforward solution is to apply one single *Catmull-Clark-type* split C [2] to every face (cf. Fig. 11). This split operation divides every *n*-sided face into *n* quadrilaterals and needs the position of newly computed *face-points* and *edge-points* to be well-defined. The vertices of $\tilde{N}$ remain unchanged.

The number of faces in the modified net N equals the sum of the lengths of all sequences $s_k \in \widetilde{F}$.

The number of faces in the quadrilateralized net N can be reduced by half if the net $\tilde{N}$ is closed, by not applying C but rather its (topological) square root $\sqrt{C}$, i.e., a refinement operator whose double application is equivalent to one application of C (cf. Fig. 11). For this split, only new *face-points* have to be computed. For open nets, the $\sqrt{C}$-split modifies the boundary polygon in a non-intuitive way. Hence, one would have to handle several special cases with boundary triangles if one is interested in a well-behaved boundary curve of the resulting surface.

3.3 Subdivision Rules for Closed Nets with Arbitrary Topology

The topological structure of any quadrilateral net after several applications of a uniform refinement operator consists of large regular regions with isolated singularities which correspond to the non-regular vertices of the initial net (cf. Fig. 12). By *topological regularity* we mean a tensor product structure with four faces meeting at every vertex. The natural way to define refinement operators for quadrilateral nets is therefore to modify a tensor product scheme such that special rules for the vicinity of non-regular vertices are found. In this paper we will use the interpolatory four-point scheme [6] in its tensor product version as the basis for the modification.

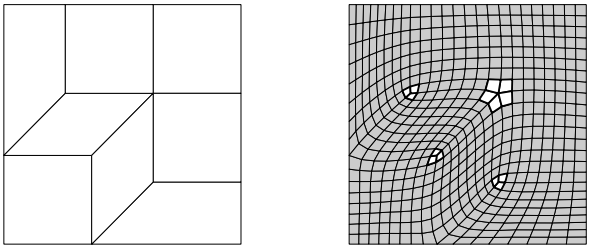


Figure 12: Isolated singularities in the refined net.

Consider a portion of a regular quadrilateral net with vertices $\mathbf{p}_{i,j}$. The vertices can be indexed locally such that each face is represented by a sequence $s_{i,j} = \{\mathbf{p}_{i,j}, \mathbf{p}_{i+1,j}, \mathbf{p}_{i+1,j+1}, \mathbf{p}_{i,j+1}\}$. The points $\mathbf{p}'_{i,j}$ of the refined net can be classified into three disjunct groups. The *vertex-points* $\mathbf{p}'_{2i,2j} := \mathbf{p}_{i,j}$ are fixed due to the interpolation requirement. The *edge-points* $\mathbf{p}'_{2i+1,2j}$ and $\mathbf{p}'_{2j,2i+1}$ are computed by applying the four-point rule (11) in the corresponding grid direction, e.g.,

$$\mathbf{p}'_{2i+1,2j} := \frac{8+\omega}{16}(\mathbf{p}_{i,j}+\mathbf{p}_{i+1,j}) - \frac{\omega}{16}(\mathbf{p}_{i-1,j}+\mathbf{p}_{i+2,j}). \quad (12)$$

Finally, the *face-points* $\mathbf{p}'_{2i+1,2j+1}$ are computed by applying the four-point rule to either four consecutive edge-points $\mathbf{p}'_{2i+1,2j-2}, \ldots, \mathbf{p}'_{2i+1,2j+4}$ or to $\mathbf{p}'_{2i-2,2j+1}, \ldots, \mathbf{p}'_{2i+4,2j+1}$. The resulting weight coefficient masks for these rules are shown in Fig. 13. The symmetry of the *face*-mask proves the equivalence of both alternatives to compute the face-points. From the differentiability of the limiting curves generated by the four-point scheme, the smoothness of the limiting surfaces generated by infinitely refining a regular quadrilateral net, follows immediately. This is a simple tensor product argument.

For the refinement of irregular quadrilateral nets, i.e., nets which include some vertices where other than four faces meet, a consistent indexing which allows the application of the above rules is impossible. If other than four edges meet at one vertex, it is not clear how to choose the four points to which one can apply the above rule

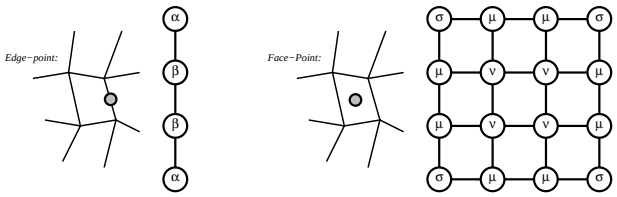


Figure 13: Subdivision masks for regular regions with $\alpha = -\frac{\omega}{16}$, $\beta = \frac{8+\omega}{16}$ and $\sigma = \alpha^2$, $\mu = \alpha\beta$, $\nu = \beta^2$.

for computing the edge-points. However, once all the edge-points are known, there always are exactly two possibilities to choose four consecutive edge-points when computing a certain face-point since the net is quadrilateral. It is an important property of tensor product schemes on regular nets that both possibilities lead to the same result (commuting univariant refinement operators). In order to modify the tensor product scheme as little as possible while generalizing it to be applicable for nets with arbitrary topology, we want to conserve this property. Hence, we will propose a subdivision scheme which only needs *one* additional rule: the one for computing edge-points corresponding to edges adjacent to a non-regular vertex. All other edge-points and all face-points are computed by the application of the original four-point scheme and the additional rule will be such that both possibilities for the face-points yield the same result.

We use the notation of Fig. 14 for points in the neighborhood of a singular vertex $\mathbf{p}$. The index i is taken to be *modulo n* where n is the number of edges meeting at $\mathbf{p}$. Applying the original four-point rule wherever possible leaves only the points $\mathbf{x}_i$ and $\mathbf{y}_i$ undefined. If we require that both possible ways to compute $\mathbf{y}_i$ by applying the standard four-point rule to succeeding edge-points lead to the same result, we get a dependence relating $\mathbf{x}_{i+1}$ to $\mathbf{x}_i$

$$\mathbf{x}_{i+1} = \mathbf{x}_i + \frac{w}{8}(\mathbf{h}_i - \mathbf{h}_{i+1}) + \frac{w^2}{8(4+w)}(\mathbf{k}_{i-2} - \mathbf{k}_{i+2}) + \frac{w}{8}(\mathbf{l}_{i+2} - \mathbf{l}_{i-1}) + \frac{4+w}{8}(\mathbf{l}_{i+1} - \mathbf{l}_i),$$

which can be considered as compatibility condition. In the regular case, this condition is satisfied for any tensor product rule. The compatibility uniquely defines the cyclic differences $\triangle\mathbf{x}_i = \mathbf{x}_{i+1} - \mathbf{x}_i$ which sum to zero (*telescoping sums*). Hence, there always exists a solution and even one degree of freedom is left for the definition of the $\mathbf{x}_i$.

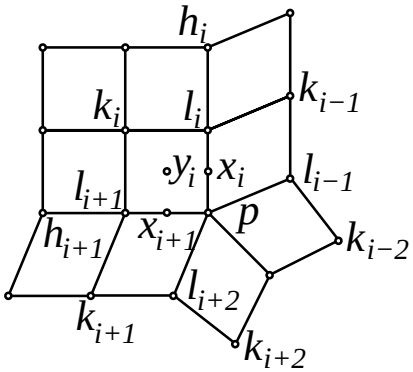


Figure 14: Notation for vertices around a singular vertex P.

The points $\mathbf{x}_i$ will be computed by rotated versions of the same subdivision mask. Thus, the vicinity of $\mathbf{p}$ will become more and more symmetric while refinement proceeds. Hence, the distance

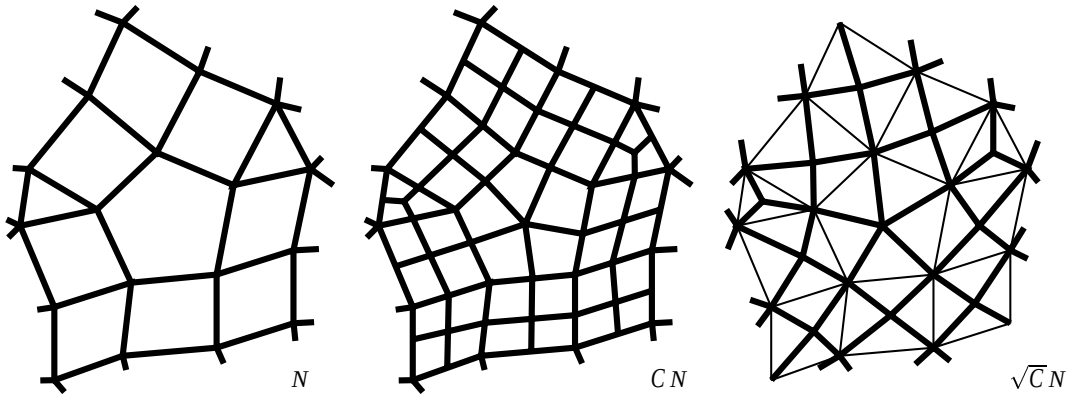


Figure 11: Transformation of an arbitrary net $\tilde{N}$ into a quadrilateral net N by one Catmull-Clark-split C (middle) or by its square root (right, for closed nets).

between $\mathbf{p}$ and the center of gravity of the $\mathbf{x}_i$ will be a good measure for the roughness of the net near $\mathbf{p}$ and the rate by which this distance tends to zero can be understood as the 'smoothing rate'. The center of gravity in the regular ($n = 4$) case is:

$$\frac{1}{n}\sum_{i=0}^{n-1}\mathbf{x}_i = \frac{4+w}{8}\mathbf{p} + \frac{1}{2n}\sum_{i=0}^{n-1}\mathbf{l}_i - \frac{w}{8n}\sum_{i=0}^{n-1}\mathbf{h}_i. \tag{13}$$

In the non-regular case, we have

$$\frac{1}{n}\sum_{i=0}^{n-1}\mathbf{x}_i \;=\; \mathbf{x}_j + \frac{1}{n}\sum_{i=0}^{n-2}(n-1-i)\triangle\mathbf{x}_{i+j}, \qquad j \in \{0,\ldots,n-1\}. \tag{14}$$

Combining common terms in the telescoping sum and equating the right hand sides of (13) and (14) leads to

$$\mathbf{x}_j = -\frac{w}{8}\mathbf{h}_j + \frac{4+w}{8}\mathbf{l}_j + \frac{4+w}{8}\mathbf{p} - \frac{w}{8}\mathbf{v}_j, \tag{15}$$

where we define the *virtual point*

$$\mathbf{v}_j := \frac{4}{n}\sum_{i=0}^{n-1}\mathbf{l}_i - (\mathbf{l}_{j-1}+\mathbf{l}_j+\mathbf{l}_{j+1}) + \frac{w}{4+w}(\mathbf{k}_{j-2}+\mathbf{k}_{j-1}+\mathbf{k}_j+\mathbf{k}_{j+1}) - \frac{4w}{(4+w)n}\sum_{i=0}^{n-1}\mathbf{k}_i. \tag{16}$$

Hence, the $\mathbf{x}_j$ can be computed by applying (11) to the four points $\mathbf{h}_j, \mathbf{l}_j, \mathbf{p}$ and $\mathbf{v}_j$. The formula also holds in the case $n = 4$ where $\mathbf{v}_j = \mathbf{l}_{j+2}$. Such a virtual point $\mathbf{v}_j$ is defined for every edge and both of its endpoints. Hence to refine an edge which connects two singular vertices $\mathbf{p}_1$ and $\mathbf{p}_2$, we first compute the two virtual points $\mathbf{v}_1$ and $\mathbf{v}_2$ and then apply (11) to $\mathbf{v}_1$, $\mathbf{p}_1$, $\mathbf{p}_2$ and $\mathbf{v}_2$. If all edge-points $\mathbf{x}_j$ are known, the refinement operation can be completed by computing the face-points $\mathbf{y}_j$. These are well defined since the auxillary edge-point rule is constructed such that both possible ways lead to the same result.

3.4 Convergence Analysis

The subdivision scheme proposed in the last section is a stationary scheme and thus the convergence criteria of [1] and [18] can be applied. In the regular regions of the net (which enlarge during refinement), the smoothness of the limiting surface immediately follows from the smoothness of the curves generated by the univariate four-point scheme. Hence to complete the convergence analysis, it is sufficient to look at the vicinities of the finitely many isolated singular vertices (cf. Fig. 12).

Let $\mathbf{p}_0, \ldots, \mathbf{p}_k$ be the points from a fixed neighborhood of the singular vertex $\mathbf{p}_0$. The size of the considered neighborhood depends on the support of the underlying tensor product scheme and contains 5 'rings' of faces around $\mathbf{p}_0$ in our case. The collection of all rules to compute the new points $\mathbf{p}'_0, \ldots, \mathbf{p}'_k$ of the same 'scaled' (5-layer-) neighborhood of $\mathbf{p}_0 = \mathbf{p}'_0$ in the refined net can be represented by a block-circulant matrix $\mathbf{A}$ such that $(\mathbf{p}'_i)_i = \mathbf{A}(\mathbf{p}_i)_i$. This matrix is called the *refinement matrix*. After [1] and [18] the convergence analysis can be reduced to the analysis of the eigenstructure of $\mathbf{A}$. For the limiting surface to have a unique tangent plane at $\mathbf{p}_0$ it is sufficient that the leading eigenvalues of $\mathbf{A}$ satisfy

$$\lambda_1 = 1, \quad 1 > \lambda_2 = \lambda_3, \quad |\lambda_2| > |\lambda_i|, \forall i \geq 4.$$

Table 2 shows theses eigenvalues of the refinement matrix $\mathbf{A}$ for vertices with n adjacent edges in the standard case $\omega = 1$. The computation of the spectrum can be done by exploiting the block-circulant structure of $\mathbf{A}$. We omit the details here, because the dimension of $\mathbf{A}$ is $k \times k$ with $k = 30n+1$.

n	λ_1	λ_2	λ_3	$\lambda_{i\geq 4} \leq$
3	1.0	0.42633	0.42633	0.25
4	1.0	0.5	0.5	0.25
5	1.0	0.53794	0.53794	0.36193
6	1.0	0.55968	0.55968	0.42633
7	1.0	0.5732	0.5732	0.46972
8	1.0	0.58213	0.58213	0.5
9	1.0	0.58834	0.58834	0.52180

Table 2: Leading eigenvalues of the subdivision matrix

In addition to a uniquely defined tangent plane we also have to have local injectivity in order to guarantee the regularity of the surface. This can be checked by looking at the natural parametrization of the surface at $\mathbf{p}_0$ which is spanned by the eigenvectors of $\mathbf{A}$ corresponding to the subdominant eigenvalues λ_2 and λ_3. The injectivity of this parametrization is a sufficient condition. The details can be found in [18]. Fig. 15 shows meshes of 'isolines' of these characteristic maps which are well-behaved.

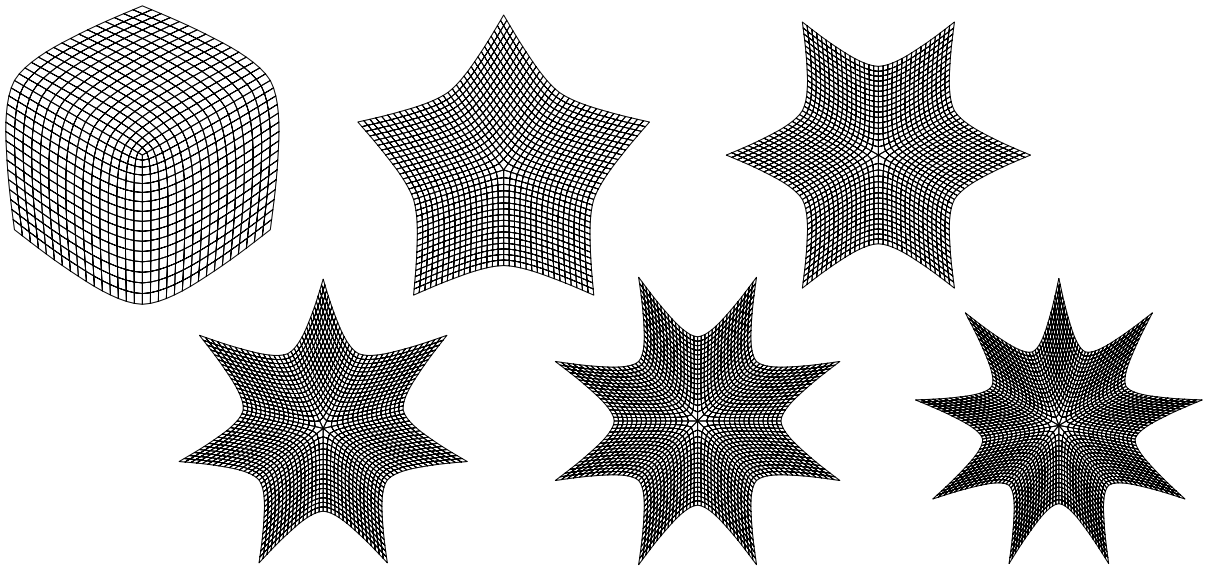


Figure 15: Sketch of the characteristic maps in the neighborhood of singular vertices with $n = 3, 5, \ldots, 9$.

3.5 Boundary Curves

If a subdivision scheme is supposed to be used in practical modeling or reconstruction applications, it must provide features that allow the definition of creases and cusps [12]. These requirements can be satisfied if the scheme includes special rules for the refinement of *open* nets which yield well-behaved boundary curves that interpolate the boundary polygons of the given net. Having such a scheme, creases can be modeled by joining two separate subdivision surfaces along a common boundary curve and cusps result from a topological hole in the initial net which geometrically shrinks to a single point, i.e., a face $s = \{\mathbf{p}_1, \ldots, \mathbf{p}_n\}$ of a given net is deleted to generate a hole and its vertices are moved to the same location $\mathbf{p}_i = \mathbf{p}$ (cf. Fig. 16).

To allow a C^0-join between two subdivision patches whose initially given nets have a common boundary polygon, it is necessary that their limiting boundary curves only depend on these common points, i.e., they must not depend on any interior point. For our scheme, we achieve this by simply applying the original univariate four-point rule to boundary polygons. Thus, the boundary curve of the limiting surface is exactly the four-point curve which is defined by the initial boundary polygon. Further, it is necessary to not only generate *smooth* boundary curves but rather to allow *piecewise* smooth boundary curves, e.g., in cases where more than two subdivision patches meet at a common point. In this case we have to cut the boundary polygon into several segments by marking some vertices on the boundary as being *corner vertices*. Each segment between two corner vertices is then treated separately as an open polygon.

When dealing with open polygons, it is not possible to refine the first or the last edge by the original four-point scheme since rule (11) requires a well-defined 2-neighborhood. Therefore, we have to find another rule for the point $\mathbf{p}_1^{m+1}$ which subdivides the edge $\overline{\mathbf{p}_0^m \mathbf{p}_1^m}$. We define an *extrapolated* point $\mathbf{p}_{-1}^m := 2\mathbf{p}_0^m - \mathbf{p}_1^m$. The point $\mathbf{p}_1^{m+1}$ then results from the application of (11) to the subpolygon $\mathbf{p}_{-1}^m, \mathbf{p}_0^m, \mathbf{p}_1^m, \mathbf{p}_2^m$. Obviously, this additional rule can be expressed as a stationary linear combination of points from the non-extrapolated open polygon:

$$\mathbf{p}_1^{m+1} := \frac{8-w}{16}\mathbf{p}_0^m + \frac{8+2w}{16}\mathbf{p}_1^m - \frac{w}{16}\mathbf{p}_2^m \qquad (17)$$

The rule to compute the point $\mathbf{p}_{2n-1}^{m+1}$ subdividing the last edge $\overline{\mathbf{p}_{n-1}^m \mathbf{p}_n^m}$ is defined analogously.

This modification of the original scheme does not affect the convergence to a continuously differentiable limit, because the estimates for the contraction rate of the maximum second forward difference used in the convergence proof of [6] remain valid. This is obvious since the extrapolation only adds the zero component $\triangle^2 p_{-1}^m$ to the sequence of second order forward differences. The main convergence criterion of [13] also applies.

It remains to define refinement rules for inner edges of the net which have one endpoint on the boundary and for faces including at least one boundary vertex. To obtain these rules we use the same heuristic as in the univariate case. We extrapolate the unrefined net over every boundary edge to get an additional layer of faces. When computing the egde- and face-points refining the original net by the rules from Sect. 3.3, these additional points can be used. To complete the refinement step, the extrapolated faces are finally deleted.

Let $\mathbf{q}_1, \ldots, \mathbf{q}_r$ be the *inner* points of the net which are connected to the boundary point $\mathbf{p}$ then the extrapolated point will be

$$\mathbf{p}^* := 2\mathbf{p} - \frac{1}{r}\sum_{i=1}^{r}\mathbf{q}_i.$$

If the boundary point $\mathbf{p}$ belongs to the face $s = \{\mathbf{p}, \mathbf{q}, \mathbf{u}, \mathbf{v}\}$ and is not connected to any inner vertex then we define $\mathbf{p}^* := 2\mathbf{p} - \mathbf{u}$. For every boundary edge $\overline{\mathbf{p}\,\mathbf{q}}$ we add the extrapolated face $s^* = \{\mathbf{p}, \mathbf{q}, \mathbf{q}^*, \mathbf{p}^*\}$.

Again, the tangent-plane continuity of the resulting limiting surface can be proved by the sufficient criteria of [1] and [18]. This is obvious since for a fixed number of interior edges adjacent to some boundary vertex $\mathbf{p}$, the refinement of the extrapolated net can be rewritten as a set of stationary refinement rules which define the new points in the vicinity of $\mathbf{p}$ as linear combinations of points from the non-extrapolated net. However the refinement matrix is no longer block-circulant.

At every surface point lying on the boundary of a tangent plane continuous surface, one tangent direction is determined by the tangent of the boundary curve (which in this case is a four-point curve that does not depend on inner vertices). On boundaries, we can therefore drop the requirement of [18] that the leading eigenvalues of the refinement matrix have to be equal. This symmetry is only a consequence of the assumption that the rules to compute the new points around a singular vertex are identical modulo

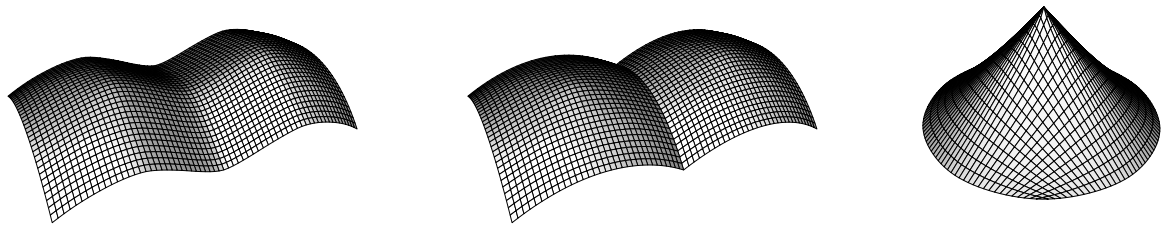


Figure 16: Modeling sharp features (piecewise smooth boundary, crease, cusp)

rotations (block-circulant refinement matrix). Although $\lambda_2 \neq \lambda_3$ causes an increasing local distortion of the net, the smoothness of the limiting surface is not affected. This effect can be viewed as a reparametrization in one direction. (Compare this to the distortion of a regular net which is refined by binary subdivision in one direction and trinary in the other.)

We summarize the different special cases which occur when refining an open net by the given rules. In Fig. 17 the net to be refined consists of the solid white faces while the extrapolated faces are drawn transparently. The dark vertex is marked as a corner vertex. We have to distinguish five different cases:

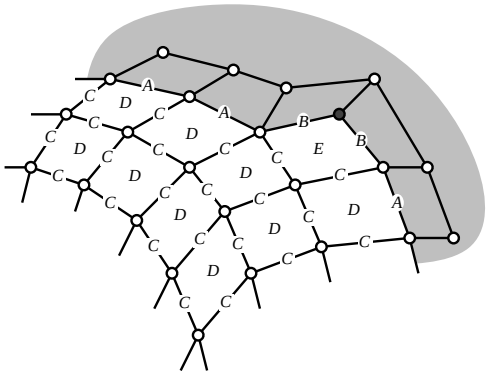


Figure 17: Occurences of the different special cases.

A: Within boundary segments, we apply (11) to four succeeding boundary vertices.

B: To the first and the last edge of an open boundary segment, we apply the special rule (17).

C: Inner edge-points can be computed by application of (15). If necessary, extrapolated points are involved.

D: For every face-point of this class, at least one sequence of four C-points can be found to which (11) can be applied. If there are two possibilities for the choice of these points then both lead to the same result which is guaranteed by the construction of (15).

E: In this case no appropriate sequence of four C-points can be found. Therefore, one has to apply (17) to a B-point and the two C-points following on the opposite side of the corner face. In order to achieve independence of the grid direction, even in case the corner vertex is not marked, we apply (17) in both directions and compute the average of the two results.

3.6 Adaptive Refinement

In most numerical applications, the exponentially increasing number of vertices and faces during the iterative refinement only allows a small number of refinement steps to be computed. If high acuracy is needed, e.g., in finite element analysis or high quality rendering, it is usually sufficient to perform a high resolution refinement in regions with high curvature while 'flat' regions may be approximated rather coarsely. Hence, in order to keep the amount of data reasonable, the next step is to introduce adaptive refinement features.

The decision *where* high resolution refinement is needed, strongly depends on the underlying application and is not discussed here. The major problem one always has to deal with when adaptive refinement of nets is performed is to handle or eliminate C^{-1}-inconsistencies which occur when faces from different refinement levels meet. A simple trick to repair the resulting triangular holes is to split the bigger face into three quadrilaterals in an Y-fashion (cf. Fig 18). However this Y-split does not repair the hole. Instead it shifts the hole to an adjacent edge. Only combining several Y-elements such that they build a 'chain' connecting two inconsistencies leads to an overall consistent net. The new vertices necessary for the Y-splits are computed by the rules of Sect. 3.3. The fact that every Y-element contains a singular ($n = 3$) vertex causes no problems for further refinement because this Y-element is only of temporary nature, i.e., if any of its three faces or any neighboring face is to be split by a following local refinement adaption, then first the Y-split is undone and a proper Catmull-Clark-type split is performed before proceeding. While this simple technique seems to be known in the engineering community, the author is not aware of any reference where the theoretical background for this technique is derived. Thus, we sketch a simple proof that shows under which conditions this technique applies.

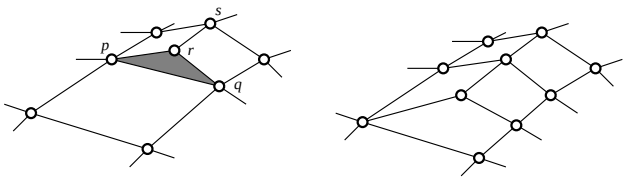


Figure 18: A hole in an adaptively refined net and an Y-element to fill it.

First, in order to apply the Y-technique we have to restrict the considered nets to *balanced* nets. These are adaptively refined nets (without Y-elements) where the refinement levels of neighboring faces differ at most by one. Non-balanced inconsistencies can not be handled by the Y-technique. Hence, looking at a particular face s from the n-th refinement level, all faces having at least one vertex in common with s are from the levels $(n-1)$, n, or $(n+1)$. For the proof we can think of first repairing all inconsistencies between level $n-1$ and n and then proceed with higher levels. Thus, without loss of generality, we can restrict our considerations to a situation where all relevant faces are from level $(n-1)$ or n.

A *critical* edge is an edge, where a triangular hole occurs due to different refinement levels of adjacent faces. A sequence of Y-elements can always be arranged such that two critical edges are connected, e.g., by surrounding one endpoint of the critical edge with a 'corona' of Y-elements until another critical edge is reached (cf. Fig. 19). Hence, on closed nets, we have to require the number of critical edges to be even. (On open nets, any boundary edge can

stop a chain of Y-elements.) We show that this is always satisfied, by induction over the number of faces from the n-th level within an environment of $(n-1)$-faces. Faces from generations $> n$ or $< (n-1)$ do not affect the situation since we assume the net to be balanced.

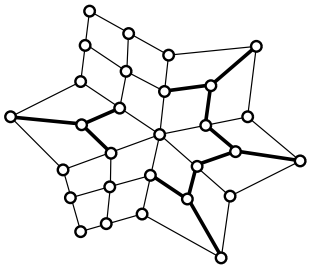


Figure 19: Combination of Y-elements

The first adaptive Catmull-Clark-type split on a uniformly refined net produces four critical edges. Every succeeding split changes the number of critical edges by an even number between -4 and 4, depending on the number of direct neighbors that have been split before. Thus the number of critical edges is always even. However, the n-faces might form a ring having in total an even number of critical edges which are separated into an odd number 'inside' and an odd number 'outside'. It turns out that this cannot happen: Let the inner region surrounded by the ring of n-faces consist of r quadrilaterals having a total number of $4r$ edges which are candidates for being critical. Every edge which is shared by two such quadrilaterals reduces the number of candidates by two and thus the number of boundary edges of this inner region is again even.

The only situation where the above argument is not valid, occurs when the considered net is open and has a hole with an odd number of boundary edges. In this case, every loop of n-faces enclosing this hole will have an odd number of critical edges on each side. Hence, we have to further restrict the class of nets to which we can apply the Y-technique to *open balanced nets which have no hole with an odd number of edges*. This restriction is not serious because one can transform any given net in order to satisfy this requirement by applying an *initial uniform refinement step* before adaptive refinement is started. Such an initial step is needed anyway if a given arbitrary net has to be transformed into a quadrilateral one (cf. Sect. 3.2).

It remains to find an *algorithm* to place the Y-elements correctly, i.e., to decide which critical edges should be connected by a corona. This problem is not trivial because interference between the Y-elements building the 'shores' of two 'islands' of n-faces lying close to each other, can occur. We describe an algorithm which only uses local information and decides the orientation separately for each face instead of 'marching' around the islands.

The initially given net (level 0) has been uniformly refined once before the adaptive refinement begins (level 1). Let every vertex of the adaptively refined net be associated with the generation in which it was introduced. Since all faces of the net are the result of a Catmull-Clark-type split (no Y-elements have been placed so far), they all have the property that three of its vertices belong to the same generation g and the fourth vertex belongs to a generation $g' < g$. This fact yields a unique orientation for every face. The algorithm starts by marking all vertices of the net which are endpoints of a critical edge, i.e. if a $(n-1)$-face $\{\mathbf{p},\mathbf{q},\ldots\}$ meets two n-faces $\{\mathbf{p},\mathbf{r},\mathbf{s},\ldots\}$ and $\{\mathbf{q},\mathbf{r},\mathbf{s},\ldots\}$ then $\mathbf{p}$ and $\mathbf{q}$ are marked (cf. Fig. 18). After the *marking-phase*, the Y-elements are placed. Let $s = \{\mathbf{p},\mathbf{q},\mathbf{u},\mathbf{v}\}$ be a face of the net where $\mathbf{p}$ is the unique vertex which belongs to an elder generation than the other three. If neither $\mathbf{q}$ nor $\mathbf{v}$ are marked then no Y-element has to be placed within this face. If only one of them is marked then the Y-element has to be oriented as shown in Fig. 20 and if both are marked this face has to be refined by a proper Catmull-Clark-type split.

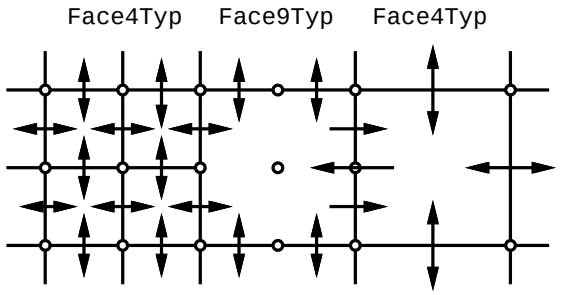


Figure 21: References between different kinds of faces.

The correctness of this algorithm is obvious since the vertices which are marked in the first phase are those which are common to faces of different levels. The second phase guarantees that a corona of Y-elements is built around each such vertex (cf. Fig. 19).

3.7 Implementation and Examples

The described algorithm is designed to be useful in practical applications. Therefore, besides the features for creating creases and cusps and the ability to adaptively refine a given quadrilateral net, efficiency and compact implementation are also important. Both can be achieved by this algorithm. The crucial point of the implementation is the design of an appropriate data structure which supports an efficient navigation through the neighborhood of the vertices. The most frequently needed access operation to the data structure representing the balanced net, is to enumerate all faces which lie around one vertex or to enumerate all the neighbors of one vertex. Thus every vertex should be associated with a linked list of the objects that constitute its vicinity. We propose to do this implicitely by storing the topological information in a data structure `Face4Typ` which contains all the information of one quadrilateral face, i.e., references to its four corner points and references to its four directly neighboring faces. By these references, a doubly linked list around every vertex is available.

Since we have to maintain an adaptively refined net, we need an additional datatype to consistently store connections between faces from different refinement levels. We define another structure `Face9Typ` which holds references to nine vertices and eight neighbors. These *multi-faces* can be considered as 'almost' split faces, where the geometric information (the new edge- and face-points) is already computed but the topological split has not yet been performed. If, during adaptive refinement, some n-face is split then all its neighbors which are from the same generation are converted into `Face9Typ`'s. Since these faces have pointers to eight neighbors, they can mimic faces from different generations and therefore connect them correctly. The `Face9Typ`'s are the candidates for the placement of Y-elements in order to re-establish consistency. The various references between the different kinds of faces are shown in Fig. 21.

To relieve the application program which decides where to adaptively refine, from keeping track of the balance of the net, the implementation of the refinement algorithm should perform recursive refinement operations when necessary, i.e., if a n-face s is to be refined then first all $(n-1)$-neighbors which have at least one vertex in common with s must be split.

The following pictures are generated by using our experimental implementation. The criterion for adaptive refinement is a discrete approximation of the Gaussian curvature. The running time of the algorithm is directly proportional to the number of computed points, i.e., to the complexity of the output-net. Hence, since the number of regions where deep refinement is necessary usually is

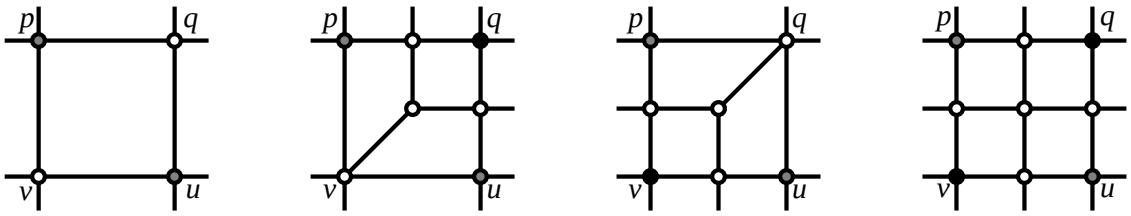


Figure 20: The orientation of the Y-elements depends on whether the vertices **q** and **v** are marked (black) or not (white). The status of vertices **p** and **u** does not matter (gray).

fixed, we can reduce the space- and time-complexity from exponential to linear (as a function of the highest occurring refinement level in the output).

References

[1] A. Ball / D. Storry, *Conditions for Tangent Plane Continuity over Recursively Generated B-Spline Surfaces*, ACM Trans. Graph. 7 (1988), pp. 83–102

[2] E. Catmull, J. Clark, *Recursively generated B-spline surfaces on arbitrary topological meshes*, CAD 10 (1978), pp. 350–355

[3] A. Cavaretta / W. Dahmen / C. Micchelli, *Stationary Subdivision*, Memoirs of the AMS 93 (1991), pp. 1-186

[4] D. Doo / M. Sabin, *Behaviour of Recursive Division Surfaces Near Extraordinary Points*, CAD 10 (1978), pp. 356–360

[5] S. Dubuc, *Interpolation Through an Iterative Scheme*, Jour. of Mathem. Anal. and Appl., 114 (1986), pp. 185–204

[6] N. Dyn / J. Gregory / D. Levin, *A 4-Point Interpolatory Subdivision Scheme for Curve Design*, CAGD 4(1987), pp. 257–268

[7] N. Dyn / J. Gregory / D. Levin, *A Butterfly Subdivision Scheme for Surface Interpolation with Tension Controll*, ACM Trans. Graph. 9 (1990), pp. 160–169

[8] N. Dyn / D. Levin, *Interpolating subdivision schemes for the generation of curves and surfaces*, Multivar. Approx. and Interp., W. Hauŝmann and K. Jetter (eds.), 1990 Birkhäuser Verlag, Basel

[9] N. Dyn, *Subdivision Schemes in Computer Aided Geometric Design*, Adv. in Num. Anal. II, Wavelets, Subdivisions and Radial Functions, W.A. Light ed., Oxford Univ. Press, 1991, pp. 36–104.

[10] N. Dyn / D. Levin / D. Liu, *Interpolatory Convexity-Preserving Subdivision Schemes for Curves and Surfaces*, CAD 24 (1992), pp. 221–216

[11] M. Halstead / M. Kass / T. DeRose, *Efficient, fair interpolation using Catmull-Clark surfaces*, Computer Graphics 27 (1993), pp. 35–44

[12] H. Hoppe, *Surface Reconstruction from unorganized points*, Thesis, University of Washington, 1994

[13] L. Kobbelt, *Using the Discrete Fourier-Transform to Analyze the Convergence of Subdivision Schemes*, Appl. Comp. Harmonic Anal. 5 (1998), pp. 68–91

[14] C. Loop, *Smooth Subdivision Surfaces Based on Triangles*, Thesis, University of Utah, 1987

[15] C. Loop, *A G^1 triangular spline surface of arbitrary topological type*, CAGD 11 (1994), pp. 303–330

[16] J. Peters, *Smooth mesh interpolation with cubic patches*, CAD 22 (1990), pp. 109–120

[17] J. Peters, *Smoothing polyhedra made easy*, ACM Trans. on Graph., Vol 14 (1995), pp. 161–169

[18] U. Reif, *A unified approach to subdivision algorithms near extraordinary vertices*, CAGD 12 (1995), pp. 153–174

[19] K. Schweizerhof, Universität Karlsruhe *private communication*

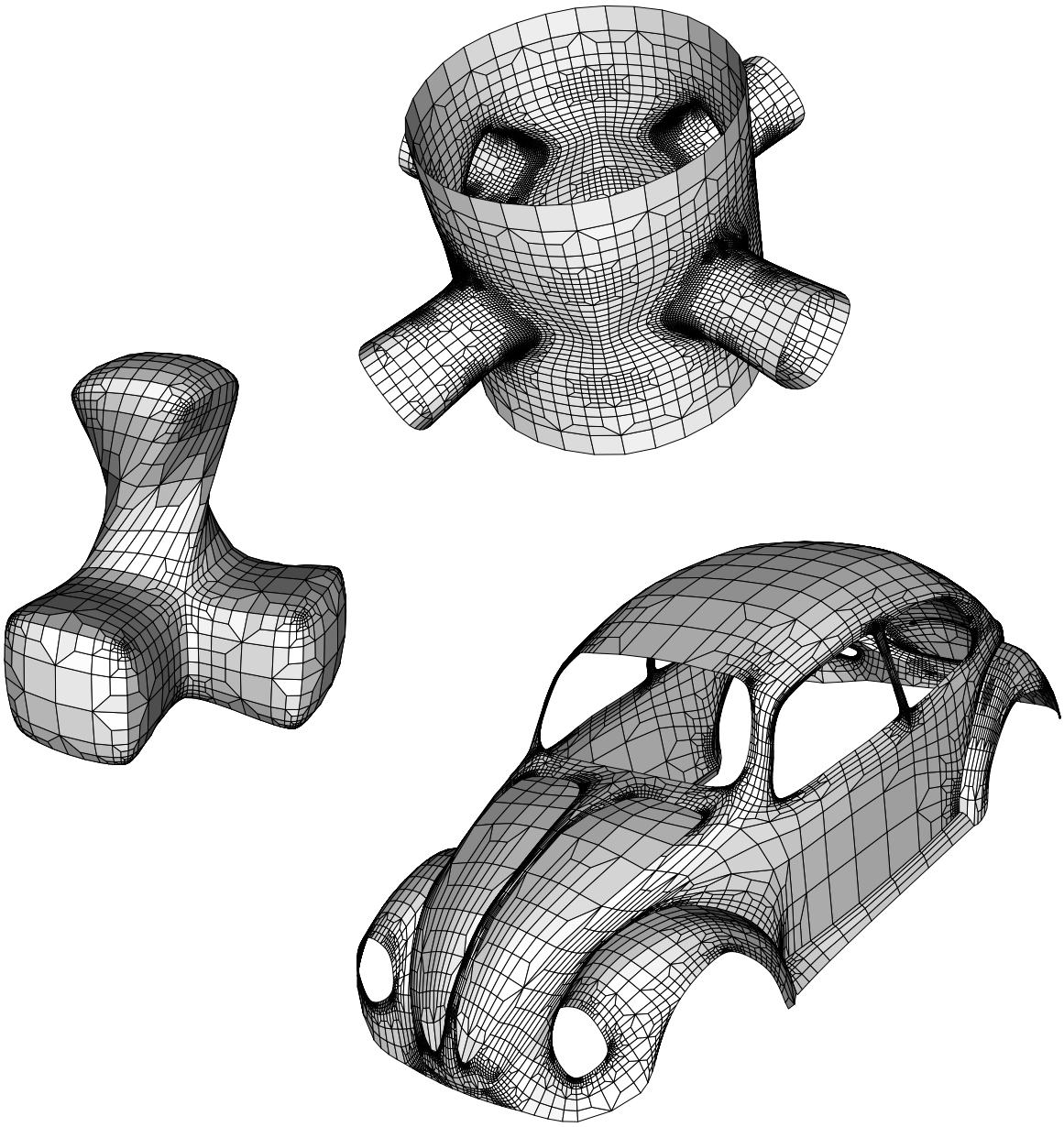


Figure 22: Examples for adaptively refined nets.

Chapter 6

Interactive Multiresolution Mesh Editing

Speaker: Denis Zorin

Interactive Multiresolution Mesh Editing

Denis Zorin* Caltech
Peter Schröder† Caltech
Wim Sweldens‡ Bell Laboratories

Abstract

We describe a multiresolution representation for meshes based on subdivision, which is a natural extension of the existing patch-based surface representations. Combining subdivision and the smoothing algorithms of Taubin [26] allows us to construct a set of algorithms for interactive multiresolution editing of complex hierarchical meshes of arbitrary topology. The simplicity of the underlying algorithms for refinement and coarsification enables us to make them local and adaptive, thereby considerably improving their efficiency. We have built a scalable interactive multiresolution editing system based on such algorithms.

1 Introduction

Applications such as special effects and animation require creation and manipulation of complex geometric models of arbitrary topology. Like real world geometry, these models often carry detail at many scales (cf. Fig. 1). The model might be constructed from scratch (ab initio design) in an interactive modeling environment or be scanned-in either by hand or with automatic digitizing methods. The latter is a common source of data particularly in the entertainment industry. When using laser range scanners, for example, individual models are often composed of high resolution meshes with hundreds of thousands to millions of triangles.

Manipulating such fine meshes can be difficult, especially when they are to be edited or animated. Interactivity, which is crucial in these cases, is challenging to achieve. Even without accounting for any computation on the mesh itself, available rendering resources alone, may not be able to cope with the sheer size of the data. Possible approaches include mesh optimization [15, 13] to reduce the size of the meshes.

Aside from considerations of economy, the choice of representation is also guided by the need for multiresolution editing semantics. The representation of the mesh needs to provide control at a large scale, so that one can change the mesh in a broad, smooth manner, for example. Additionally designers will typically also want control over the minute features of the model (cf. Fig. 1). Smoother approximations can be built through the use of patches [14], though at the cost of loosing the high frequency details. Such detail can be reintroduced by combining patches with displacement maps [17]. However, this is difficult to manage in the arbitrary topology setting and across a continuous range of scales and hardware resources.

Figure 1: Before the Armadillo started working out he was flabby, complete with a double chin. Now he exercises regularly. The original is on the right (courtesy Venkat Krischnamurthy). The edited version on the left illustrates large scale edits, such as his belly, and smaller scale edits such as his double chin; all edits were performed at about 5 frames per second on an Indigo R10000 Solid Impact.

For reasons of efficiency the algorithms should be highly adaptive and dynamically adjust to available resources. Our goal is to have a single, simple, uniform representation with scalable algorithms. The system should be capable of delivering multiple frames per second update rates even on small workstations taking advantage of lower resolution representations.

In this paper we present a system which possesses these properties

- **Multiresolution control:** Both broad and general handles, as well as small knobs to tweak minute detail are available.
- **Speed/fidelity tradeoff:** All algorithms dynamically adapt to available resources to maintain interactivity.
- **Simplicity/uniformity:** A single primitive, triangular mesh, is used to represent the surface across all levels of resolution.

Our system is inspired by a number of earlier approaches. We mention multiresolution editing [11, 9, 12], arbitrary topology subdivision [6, 2, 19, 7, 28, 16], wavelet representations [21, 24, 8, 3], and mesh simplification [13, 17]. Independently an approach similar to ours was developed by Pulli and Lounsbery [23].

It should be noted that our methods rely on the finest level mesh having subdivision connectivity. This requires a remeshing step before external high resolution geometry can be imported into the editor. Eck et al. [8] have described a possible approach to remeshing arbitrary finest level input meshes fully automatically. A method that relies on a user's expertise was developed by Krishnamurthy and Levoy [17].

1.1 Earlier Editing Approaches

H-splines were presented in pioneering work on hierarchical editing by Forsey and Bartels [11]. Briefly, H-splines are obtained by adding finer resolution B-splines onto an existing coarser resolution B-spline patch relative to the coordinate frame induced by the

*dzorin@gg.caltech.edu
†ps@cs.caltech.edu
‡wim@bell-labs.com

coarser patch. Repeating this process, one can build very complicated shapes which are entirely parameterized over the unit square. Forsey and Bartels observed that the hierarchy induced coordinate frame for the offsets is essential to achieve correct editing semantics.

H-splines provide a uniform framework for representing both the coarse and fine level details. Note however, that as more detail is added to such a model the internal control mesh data structures more and more resemble a fine polyhedral mesh.

While their original implementation allowed only for regular topologies their approach could be extended to the general setting by using surface splines or one of the spline derived general topology subdivision schemes [18]. However, these schemes have not yet been made to work adaptively.

Forsey and Bartels' original work focused on the ab initio design setting. There the user's help is enlisted in defining what is meant by different levels of resolution. The user decides where to add detail and manipulates the corresponding controls. This way the levels of the hierarchy are hand built by a human user and the representation of the final object is a function of its editing history.

To edit an a priori given model it is crucial to have a general procedure to define coarser levels and compute details between levels. We refer to this as the *analysis* algorithm. An H-spline analysis algorithm based on weighted least squares was introduced [10], but is too expensive to run interactively. Note that even in an ab initio design setting online analysis is needed, since after a long sequence of editing steps the H-spline is likely to be overly refined and needs to be consolidated.

Wavelets provide a framework in which to rigorously define multiresolution approximations and fast analysis algorithms. Finkelstein and Salesin [9], for example, used B-spline wavelets to describe multiresolution editing of curves. As in H-splines, parameterization of details with respect to a coordinate frame induced by the coarser level approximation is required to get correct editing semantics. Gortler and Cohen [12], pointed out that wavelet representations of detail tend to behave in undesirable ways during editing and returned to a pure B-spline representation as used in H-splines.

Carrying these constructions over into the arbitrary topology surface framework is not straightforward. In the work by Lounsbery et al. [21] the connection between wavelets and subdivision was used to define the different levels of resolution. The original constructions were limited to piecewise linear subdivision, but smoother constructions are possible [24, 28].

An approach to surface modeling based on variational methods was proposed by Welch and Witkin [27]. An attractive characteristic of their method is flexibility in the choice of control points. However, they use a global optimization procedure to compute the surface which is not suitable for interactive manipulation of complex surfaces.

Before we proceed to a more detailed discussion of editing we first discuss different surface representations to motivate our choice of synthesis (refinement) algorithm.

1.2 Surface Representations

There are many possible choices for surface representations. Among the most popular are polynomial patches and polygons.

Patches are a powerful primitive for the construction of coarse grain, smooth models using a small number of control parameters. Combined with hardware support relatively fast implementations are possible. However, when building complex models with many patches the preservation of smoothness across patch boundaries can be quite cumbersome and expensive. These difficulties are compounded in the arbitrary topology setting when polynomial parameterizations cease to exist everywhere. Surface splines [4, 20, 22] provide one way to address the arbitrary topology challenge.

As more fine level detail is needed the proliferation of control points and patches can quickly overwhelm both the user and the most powerful hardware. With detail at finer levels, patches become less suited and polygonal meshes are more appropriate.

Polygonal Meshes can represent arbitrary topology and resolve fine detail as found in laser scanned models, for example. Given that most hardware rendering ultimately resolves to triangle scan-conversion even for patches, polygonal meshes are a very basic primitive. Because of sheer size, polygonal meshes are difficult to manipulate interactively. Mesh simplification algorithms [13] provide one possible answer. However, we need a mesh simplification approach, that is hierarchical and gives us shape handles for smooth changes over larger regions while maintaining high frequency details.

Patches and fine polygonal meshes represent two ends of a spectrum. Patches efficiently describe large smooth sections of a surface but cannot model fine detail very well. Polygonal meshes are good at describing very fine detail accurately using dense meshes, but do not provide coarser manipulation semantics.

Subdivision connects and unifies these two extremes.

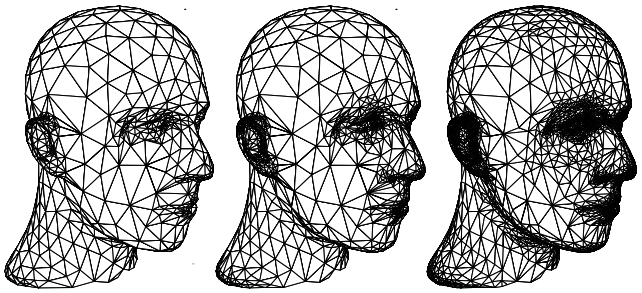


Figure 2: Subdivision describes a smooth surface as the limit of a sequence of refined polyhedra. The meshes show several levels of an adaptive Loop surface generated by our system (dataset courtesy Hugues Hoppe, University of Washington).

Subdivision defines a smooth surface as the limit of a sequence of successively refined polyhedral meshes (cf. Fig. 2). In the regular patch based setting, for example, this sequence can be defined through well known knot insertion algorithms [5]. Some subdivision methods generalize spline based knot insertion to irregular topology control meshes [2, 6, 19] while other subdivision schemes are independent of splines and include a number of interpolating schemes [7, 28, 16].

Since subdivision provides a path from patches to meshes, it can serve as a good foundation for the unified infrastructure that we seek. A single representation (hierarchical polyhedral meshes) supports the patch-type semantics of manipulation *and* finest level detail polyhedral edits equally well. The main challenge is to make the basic algorithms fast enough to escape the exponential time and space growth of naive subdivision. This is the core of our contribution.

We summarize the main features of subdivision important in our context

- **Topological Generality:** Vertices in a triangular (resp. quadrilateral) mesh need not have valence 6 (resp. 4). Generated surfaces are smooth everywhere, and efficient algorithms exist for computing normals and limit positions of points on the surface.
- **Multiresolution:** because they are the limit of successive refinement, subdivision surfaces support multiresolution algorithms, such as level-of-detail rendering, multiresolution editing, compression, wavelets, and numerical multigrid.

- **Simplicity:** subdivision algorithms are simple: the finer mesh is built through insertion of new vertices followed by *local* smoothing.
- **Uniformity of Representation:** subdivision provides a single representation of a surface at all resolution levels. Boundaries and features such as creases can be resolved through modified rules [14, 25], reducing the need for trim curves, for example.

1.3 Our Contribution

Aside from our perspective, which unifies the earlier approaches, our major contribution—and the main challenge in this program—is the design of highly adaptive and dynamic data structures and algorithms, which allow the system to function across a range of computational resources from PCs to workstations, delivering as much interactive fidelity as possible with a given polygon rendering performance. Our algorithms work for the class of 1-ring subdivision schemes (definition see below) and we demonstrate their performance for the concrete case of Loop's subdivision scheme.

The particulars of those algorithms will be given later, but Fig. 3 already gives a preview of how the different algorithms make up the editing system. In the next sections we first talk in more detail about subdivision, smoothing, and multiresolution transforms.

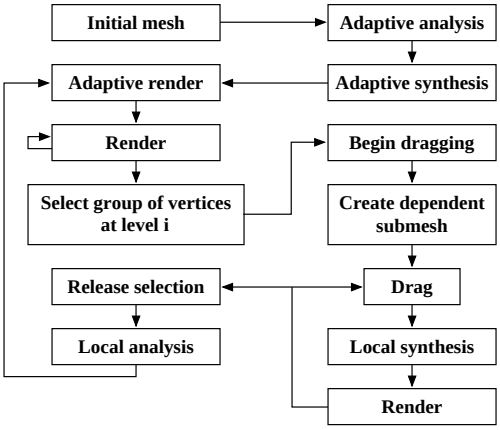


Figure 3: The relationship between various procedures as the user moves a set of vertices.

2 Subdivision

We begin by defining subdivision and fixing our notation. There are 2 points of view that we must distinguish. On the one hand we are dealing with an abstract *graph* and perform topological operations on it. On the other hand we have a *mesh* which is the geometric object in 3-space. The mesh is the image of a map defined on the graph: it associates a *point* in 3D with every *vertex* in the graph (cf. Fig. 4). A *triangle* denotes a face in the graph or the associated polygon in 3-space.

Initially we have a triangular graph T^0 with vertices V^0. By recursively *refining* each triangle into 4 subtriangles we can build a sequence of finer triangulations T^i with vertices V^i, $i > 0$ (cf. Fig. 4). The superscript i indicates the *level* of triangles and vertices respectively. A triangle $t \in T^i$ is a triple of indices $t = \{v_a, v_b, v_c\} \subset V^i$.

The vertex sets are nested as $V^j \subset V^i$ if $j < i$. We define *odd* vertices on level i as $M^i = V^{i+1} \setminus V^i$. V^{i+1} consists of two disjoint sets: *even* vertices (V^i) and *odd* vertices (M^i). We define the *level* of a vertex v as the smallest i for which $v \in V^i$. The level of v is $i+1$ if and only if $v \in M^i$.

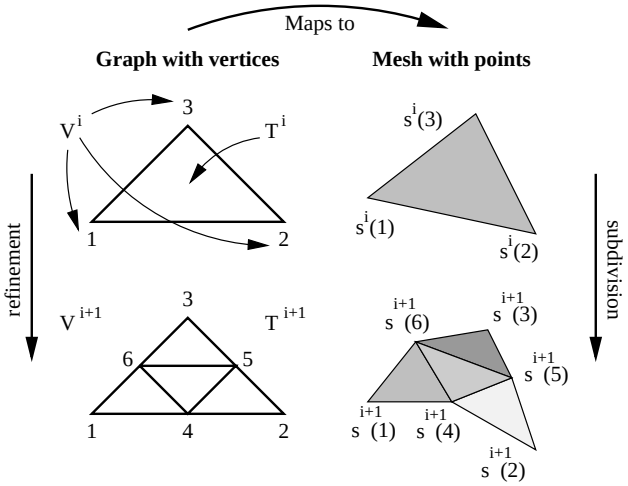


Figure 4: Left: the abstract graph. Vertices and triangles are members of sets V^i and T^i respectively. Their index indicates the level of refinement when they first appeared. Right: the mapping to the mesh and its subdivision in 3-space.

With each set V^i we associate a map, i.e., for each vertex v and each level i we have a 3D point $s^i(v) \in \mathbf{R}^3$. The set s^i contains all points on level i, $s^i = \{s^i(v) \mid v \in V^i\}$. Finally, a *subdivision scheme* is a linear operator S which takes the points from level i to points on the *finer* level $i+1$: $s^{i+1} = S\, s^i$

Assuming that the subdivision converges, we can define a limit surface σ as

$$\sigma = \lim_{k \to \infty} S^k\, s^0.$$

$\sigma(v) \in \mathbf{R}^3$ denotes the point on the limit surface associated with vertex v.

In order to define our offsets with respect to a local frame we also need tangent vectors and a normal. For the subdivision schemes that we use, such vectors can be defined through the application of linear operators Q and R acting on s^i so that $q^i(v) = (Qs^i)(v)$ and $r^i(v) = (Rs^i)(v)$ are linearly independent tangent vectors at $\sigma(v)$. Together with an orientation they define a local orthonormal frame $F^i(v) = (n^i(v), q^i(v), r^i(v))$. It is important to note that in general it is not necessary to use precise normals and tangents during editing; as long as the frame vectors are affinely related to the positions of vertices of the mesh, we can expect intuitive editing behavior.

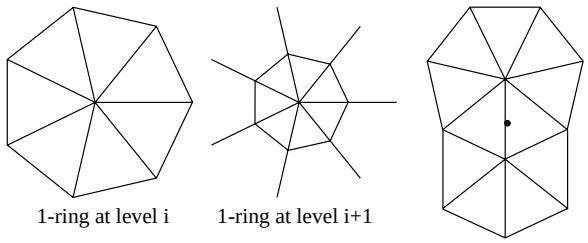


Figure 5: An even vertex has a 1-ring of neighbors at each level of refinement (left/middle). Odd vertices—in the middle of edges—have 1-rings around each of the vertices at either end of their edge (right).

Next we discuss two common subdivision schemes, both of which belong to the class of *1-ring schemes*. In these schemes points at level $i+1$ depend only on 1-ring neighborhoods of points

at level i. Let $v \in V^i$ (v even) then the point $s^{i+1}(v)$ is a function of only those $s^i(v_n)$, $v_n \in V^i$, which are immediate neighbors of v (cf. Fig. 5 left/middle). If $m \in M^i$ (m odd), it is the vertex inserted when splitting an edge of the graph; we call such vertices *middle vertices* of edges. In this case the point $s^{i+1}(m)$ is a function of the 1-rings around the vertices at the ends of the edge (cf. Fig. 5 right).

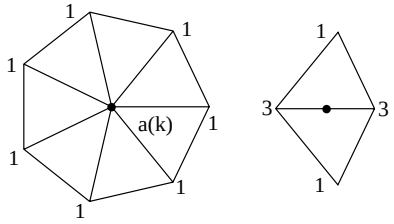


Figure 6: Stencils for Loop subdivision with unnormalized weights for even and odd vertices.

Loop is a non-interpolating subdivision scheme based on a generalization of quartic triangular box splines [19]. For a given even vertex $v \in V^i$, let $v_k \in V^i$ with $1 \le k \le K$ be its K 1-ring neighbors. The new point $s^{i+1}(v)$ is defined as $s^{i+1}(v) = (a(K)+K)^{-1}(a(K)\, s^i(v) + \sum_{k=1}^{K} s^i(v_k))$ (cf. Fig. 6), $a(K) = K(1-\alpha(K))/\alpha(K)$, and $\alpha(K) = 5/8-(3+2\cos(2\pi/K))^2/64$. For odd v the weights shown in Fig. 6 are used. Two independent tangent vectors $t_1(v)$ and $t_2(v)$ are given by $t_p(v) = \sum_{k=1}^{K} \cos(2\pi(k+p)/K)\, s^i(v_k)$.

Features such as boundaries and cusps can be accommodated through simple modifications of the stencil weights [14, 25, 29].

Butterfly is an interpolating scheme, first proposed by Dyn et al. [7] in the topologically regular setting and recently generalized to arbitrary topologies [28]. Since it is interpolating we have $s^i(v) = \sigma(v)$ for $v \in V^i$ even. The exact expressions for odd vertices depend on the valence K and the reader is referred to the original paper for the exact values [28].

For our implementation we have chosen the Loop scheme, since more performance optimizations are possible in it. However, the algorithms we discuss later work for any 1-ring scheme.

3 Multiresolution Transforms

So far we only discussed subdivision, i.e., how to go from coarse to fine meshes. In this section we describe analysis which goes from fine to coarse.

We first need *smoothing*, i.e., a linear operation H to build a smooth coarse mesh at level $i-1$ from a fine mesh at level i:

$$s^{i-1} = H\, s^i.$$

Several options are available here:

- **Least squares:** One could define analysis to be optimal in the least squares sense,

$$\min_{s^{i-1}} \|s^i - S\, s^{i-1}\|^2.$$

  The solution may have unwanted undulations and is too expensive to compute interactively [10].

- **Fairing:** A coarse surface could be obtained as the solution to a global variational problem. This is too expensive as well. An alternative is presented by Taubin [26], who uses a *local* non-shrinking smoothing approach.

Because of its computational simplicity we decided to use a version of Taubin smoothing. As before let $v \in V^i$ have K neighbors $v_k \in V^i$. Use the average, $\overline{s}^i(v) = K^{-1} \sum_{k=1}^{K} s^i(v_k)$, to define the discrete Laplacian $\mathcal{L}(v) = \overline{s}^i(v) - s^i(v)$. On this basis Taubin gives a Gaussian-like smoother which does not exhibit shrinkage

$$H := (I + \mu\, \mathcal{L})\, (I + \lambda\, \mathcal{L}).$$

With subdivision and smoothing in place, we can describe the transform needed to support multiresolution editing. Recall that for multiresolution editing we want the difference between successive levels expressed with respect to a frame induced by the coarser level, i.e., the offsets are relative to the smoother level.

With each vertex v and each level $i > 0$ we associate a *detail vector*, $d^i(v) \in \mathbf{R}^3$. The set d^i contains all detail vectors on level i, $d^i = \{d^i(v) \mid v \in V^i\}$. As indicated in Fig. 7 the detail vectors are defined as

$$d^i = (F^i)^t\, (s^i - S\, s^{i-1}) = (F^i)^t\, (I - S\, H)\, s^i,$$

i.e., the detail vectors at level i record how much the points at level i differ from the result of subdividing the points at level $i-1$. This difference is then represented with respect to the local frame F^i to obtain coordinate independence.

Since detail vectors are sampled on the fine level mesh V^i, this transformation yields an overrepresentation in the spirit of the Burt-Adelson Laplacian pyramid [1]. The only difference is that the smoothing filters (Taubin) are not the dual of the subdivision filter (Loop). Theoretically it would be possible to subsample the detail vectors and only record a detail per odd vertex of M^{i-1}. This is what happens in the wavelet transform. However, subsampling the details severely restricts the family of smoothing operators that can be used.

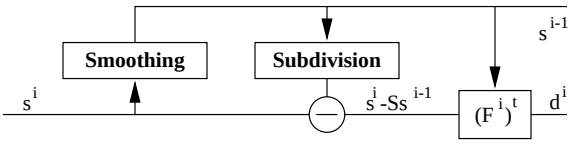


Figure 7: Wiring diagram of the multiresolution transform.

4 Algorithms and Implementation

Before we describe the algorithms in detail let us recall the overall structure of the mesh editor (cf. Fig 3). The analysis stage builds a succession of coarser approximations to the surface, each with fewer control parameters. Details or offsets between successive levels are also computed. In general, the coarser approximations are not visible; only their control points are rendered. These control points give rise to a *virtual surface* with respect to which the remaining details are given. Figure 8 shows wireframe representations of virtual surfaces corresponding to control points on levels 0, 1, and 2.

When an edit level is selected, the surface is represented internally as an approximation at this level, plus the set of all finer level details. The user can freely manipulate degrees of freedom at the edit level, while the finer level details remain unchanged relative to the coarser level. Meanwhile, the system will use the synthesis algorithm to render the modified edit level with all the finer details added in. In between edits, analysis enforces consistency on the internal representation of coarser levels and details (cf. Fig. 9).

The basic algorithms `Analysis` and `Synthesis` are very simple and we begin with their description.

Let $i = 0$ be the coarsest and $i = n$ the finest level with N vertices. For each vertex v and all levels i finer than the first level

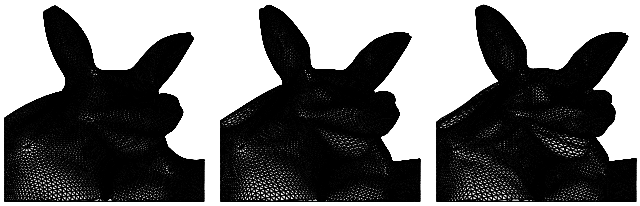


Figure 8: Wireframe renderings of virtual surfaces representing the first three levels of control points.

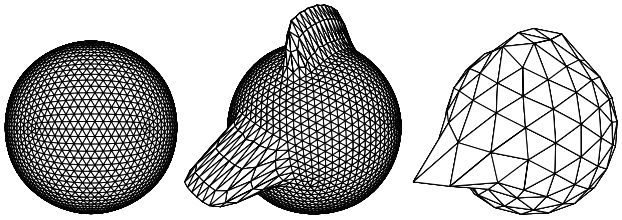


Figure 9: Analysis propagates the changes on finer levels to coarser levels, keeping the magnitude of details under control. Left: The initial mesh. Center: A simple edit on level 3. Right: The effect of the edit on level 2. A significant part of the change was absorbed by higher level details.

where the vertex v appears, there are storage locations $v.s[i]$ and $v.d[i]$, each with 3 floats. With this the total storage adds to $2 * 3 * (4N/3)$ floats. In general, $v.s[i]$ holds $s^i(v)$ and $v.d[i]$ holds $d^i(v)$; temporarily, these locations can be used to store other quantities. The local frame is computed by calling $v.F(i)$.

Global analysis and synthesis are performed level wise:

```
Analysis
  for i = n downto 1
    Analysis(i)
```

```
Synthesis
  for i = 1 to n
    Synthesis(i)
```

With the action at each level described by

```
Analysis(i)
  ∀v ∈ V^{i-1} : v.s[i-1] := smooth(v, i)
  ∀v ∈ V^i     : v.d[i] := v.F(i)^t * (v.s[i] - subd(v, i-1))
```

and

```
Synthesis(i)
  ∀v ∈ V^i : s.v[i] := v.F(i) * v.d[i] + subd(v, i-1)
```

Analysis computes points on the coarser level $i-1$ using smoothing (`smooth`), subdivides s^{i-1} (`subd`), and computes the detail vectors d^i (cf. Fig. 7). Synthesis reconstructs level i by subdividing level $i-1$ and adding the details.

So far we have assumed that all levels are uniformly refined, i.e., all neighbors at all levels exist. Since time and storage costs grow exponentially with the number of levels, this approach is unsuitable for an interactive implementation. In the next sections we explain how these basic algorithms can be made memory and time efficient.

Adaptive and *local* versions of these generic algorithms (cf. Fig. 3 for an overview of their use) are the key to these savings. The underlying idea is to use lazy evaluation and pruning based on thresholds. Three thresholds control this pruning: ϵ_A for adaptive analysis, ϵ_S for adaptive synthesis, and ϵ_R for adaptive rendering. To make lazy evaluation fast enough several caches are maintained explicitly and the order of computations is carefully staged to avoid recomputation.

4.1 Adaptive Analysis

The generic version of analysis traverses entire levels of the hierarchy starting at some finest level. Recall that the purpose of analysis is to compute coarser approximations and detail offsets. In many regions of a mesh, for example, if it is flat, no significant details will be found. *Adaptive analysis* avoids the storage cost associated with detail vectors below some threshold ϵ_A by observing that small detail vectors imply that the finer level almost coincides with the subdivided coarser level. The storage savings are realized through *tree pruning*.

For this purpose we need an integer $v.finest := \max_i\{\|v.d[i]\| \geq \epsilon_A\}$. Initially $v.finest = n$ and the following precondition holds before calling `Analysis(i)`:

- The surface is uniformly subdivided to level i,
- $\forall v \in V^i : v.s[i] = s^i(v)$,
- $\forall v \in V^i \mid i < j \leq v.finest : v.d[j] = d^j(v)$.

Now `Analysis(i)` becomes:

```
Analysis(i)
  ∀v ∈ V^{i-1} : v.s[i-1] := smooth(v, i)
  ∀v ∈ V^i :
    v.d[i] := v.s[i] - subd(v, i-1)
    if v.finest > i or ||v.d[i]|| ≥ ε_A then
      v.d[i] := v.F(i)^t * v.d[i]
    else
      v.finest := i - 1
  Prune(i-1)
```

Triangles that do not contain details above the threshold are unrefined:

```
Prune(i)
  ∀t ∈ T^i : If all middle vertices m have m.finest = i-1
             and all children are leaves, delete children.
```

This results in an adaptive mesh structure for the surface with $v.d[i] = d^i(v)$ for all $v \in V^i$, $i \leq v.finest$. Note that the resulting mesh is not restricted, i.e., two triangles that share a vertex can differ in more than one level. Initial analysis has to be followed by a synthesis pass which enforces restriction.

4.2 Adaptive Synthesis

The main purpose of the general synthesis algorithm is to rebuild the finest level of a mesh from its hierarchical representation. Just as in the case of analysis we can get savings from noticing that in flat regions, for example, little is gained from synthesis and one might as well save the time and storage associated with synthesis. This is the basic idea behind *adaptive synthesis*, which has two main purposes. First, ensure the mesh is restricted on each level, (cf. Fig. 10). Second, refine triangles and recompute points until the mesh has reached a certain measure of local flatness compared against the threshold ϵ_S.

The algorithm recomputes the points $s^i(v)$ starting from the coarsest level. Not all neighbors needed in the subdivision stencil of a given point necessarily exist. Consequently adaptive synthesis

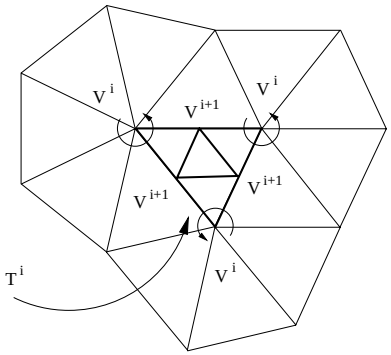


Figure 10: A restricted mesh: the center triangle is in T^i and its vertices in V^i. To subdivide it we need the 1-rings indicated by the circular arrows. If these are present the graph is restricted and we can compute s^{i+1} for all vertices and middle vertices of the center triangle.

lazily creates all triangles needed for subdivision by temporarily refining their parents, then computes subdivision, and finally deletes the newly created triangles unless they are needed to satisfy the restriction criterion. The following precondition holds before entering `AdaptiveSynthesis`:

- $\forall t \in T^j \mid 0 \le j \le i$: t is restricted
- $\forall v \in V^j \mid 0 \le j \le v.depth$: $v.s[j] = s^j(v)$

where $v.depth := \max_i\{s^i(v)\text{has been recomputed}\}$.

```
AdaptiveSynthesis

  ∀v ∈ V^0 : v.depth := 0
  for i = 0 to n − 1
    temptri := {}
    ∀t ∈ T^i :
      current := {}
      Refine(t, i, true)
    ∀t ∈ temptri : if not t.restrict then
      Delete children of t
```

The list $temptri$ serves as a cache holding triangles from levels $j < i$ which are temporarily refined. A triangle is appended to the list if it was refined to compute a value at a vertex. After processing level i these triangles are unrefined unless their $t.restrict$ flag is set, indicating that a temporarily created triangle was later found to be needed permanently to ensure restriction. Since triangles are appended to $temptri$, parents precede children. Deallocating the list tail first guarantees that all unnecessary triangles are erased.

The function `Refine`(t, i, dir) (see below) creates children of $t \in T^i$ and computes the values $Ss^i(v)$ for the vertices and middle vertices of t. The results are stored in $v.s[i+1]$. The boolean argument dir indicates whether the call was made directly or recursively.

```
Refine(t, i, dir)

  if t.leaf then  Create children for t
  ∀v ∈ t : if v.depth < i + 1 then
    GetRing(v, i)
    Update(v, i)
    ∀m ∈ N(v, i + 1, 1) :
      Update(m, i)
      if m.finest ≥ i + 1 then
        forced := true
  if dir and Flat(t) < ε_S and not forced then
    Delete children of t
  else
    ∀t ∈ current : t.restrict := true

Update(v, i)
  v.s[i + 1] := subd(v, i)
  v.depth := i + 1
  if v.finest ≥ i + 1 then
    v.s[i + 1] += v.F(i + 1) ∗ v.d[i + 1]
```

The condition $v.depth = i + 1$ indicates whether an earlier call to `Refine` already recomputed $s^{i+1}(v)$. If not, call `GetRing`(v, i) and `Update`(v, i) to do so. In case a detail vector lives at v at level i ($v.finest \ge i + 1$) add it in. Next compute $s^{i+1}(m)$ for middle vertices on level $i + 1$ around v ($m \in N(v, i + 1, 1)$, where $N(v, i, l)$ is the l-ring neighborhood of vertex v at level i). If m has to be calculated, compute `subd`(m, i) and add in the detail if it exists and record this fact in the flag $forced$ which will prevent unrefinement later. At this point, all s^{i+1} have been recomputed for the vertices and middle vertices of t. Unrefine t and delete its children if `Refine` was called directly, the triangle is sufficiently flat, and none of the middle vertices contain details (i.e., $forced =$ `false`). The list $current$ functions as a cache holding triangles from level $i - 1$ which are temporarily refined to build a 1-ring around the vertices of t. If after processing all vertices and middle vertices of t it is decided that t will remain refined, none of the coarser-level triangles from $current$ can be unrefined without violating restriction. Thus $t.restrict$ is set for all of them. The function `Flat`(t) measures how close to planar the corners and edge middle vertices of t are.

Finally, `GetRing`(v, i) ensures that a complete ring of triangles on level i adjacent to the vertex v exists. Because triangles on level i are restricted triangles all triangles on level $i - 1$ that contain v exist (precondition). At least one of them is refined, since otherwise there would be no reason to call `GetRing`(v, i). All other triangles could be leaves or temporarily refined. Any triangle that was already temporarily refined may become permanently refined to enforce restriction. Record such candidates in the $current$ cache for fast access later.

```
GetRing(v, i)

  ∀t ∈ T^{i−1} with v ∈ t :
    if t.leaf then
      Refine(t, i − 1, false); temptri.append(t)
      t.restrict := false; t.temp := true
    if t.temp then
      current.append(t)
```

4.3 Local Synthesis

Even though the above algorithms are adaptive, they are still run everywhere. During an edit, however, not all of the surface changes. The most significant economy can be gained from performing analysis and synthesis only over submeshes which require it.

Assume the user edits level l and modifies the points $s^l(v)$ for $v \in V^{*l} \subset V^l$. This invalidates coarser level values s^i and d^i for certain subsets $V^{*i} \subset V^i$, $i \leq l$, and finer level points s^i for subsets $V^{*i} \subset V^i$ for $i > l$. Finer level detail vectors d^i for $i > l$ remain correct by definition. Recomputing the coarser levels is done by *local incremental analysis* described in Section 4.4, recomputing the finer level is done by *local synthesis* described in this section.

The set of vertices V^{*i} which are affected depends on the support of the subdivision scheme. If the support fits into an m-ring around the computed vertex, then all modified vertices on level $i+1$ can be found recursively as

$$V^{*i+1} = \bigcup_{v \in V^{*i}} N(v, i+1, m).$$

We assume that $m = 2$ (Loop-like schemes) or $m = 3$ (Butterfly type schemes). We define the *subtriangulation* T^{*i} to be the subset of triangles of T^i with vertices in V^{*i}.

`LocalSynthesis` is only slightly modified from `AdaptiveSynthesis`: iteration starts at level l and iterates only over the submesh T^{*i}.

4.4 Local Incremental Analysis

After an edit on level l *local incremental analysis* will recompute $s^i(v)$ and $d^i(v)$ locally for coarser level vertices ($i \leq l$) which are affected by the edit. As in the previous section, we assume that the user edited a set of vertices v on level l and call V^{*i} the set of vertices affected on level i. For a given vertex $v \in V^{*i}$ we define

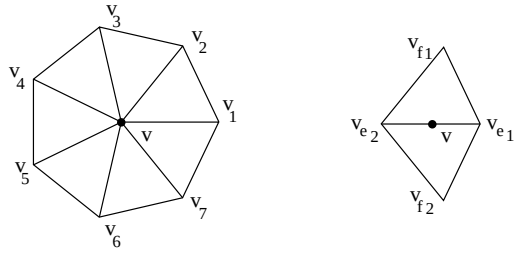


Figure 11: Sets of even vertices affected through smoothing by either an even v or odd m vertex.

$R^{i-1}(v) \subset V^{i-1}$ to be the set of vertices on level $i-1$ affected by v through the smoothing operator H. The sets V^{*i} can now be defined recursively starting from level $i = l$ to $i = 0$:

$$V^{*i-1} = \bigcup_{v \in V^{*i}} R^{i-1}(v).$$

The set $R^{i-1}(v)$ depends on the size of the smoothing stencil and whether v is even or odd (cf. Fig. 11). If the smoothing filter is 1-ring, e.g., Gaussian, then $R^{i-1}(v) = \{v\}$ if v is even and $R^{i-1}(m) = \{v_{e1}, v_{e2}\}$ if m is odd. If the smoothing filter is 2-ring, e.g., Taubin, then $R^{i-1}(v) = \{v\} \cup \{v_k \mid 1 \leq k \leq K\}$ if v is even and $R^{i-1}(m) = \{v_{e1}, v_{e2}, v_{f1}, v_{f2}\}$ if v is odd. Because of restriction, these vertices always exist. For $v \in V^i$ and $v' \in R^{i-1}(v)$ we let $c(v, v')$ be the coefficient in the analysis stencil. Thus

$$(H\, s^i)(v') = \sum_{v \,|\, v' \in R^{i-1}(v)} c(v, v') s^i(v).$$

This could be implemented by running over the v' and each time computing the above sum. Instead we use the dual implementation, iterate over all v, accumulating (`+=`) the right amount to $s^i(v')$ for $v' \in R^{i-1}(v)$. In case of a 2-ring Taubin smoother the coefficients are given by

$$\begin{aligned} c(v,v) &= (1-\mu)(1-\lambda) + \mu\lambda/6 \\ c(v,v_k) &= \mu\lambda/6K \\ c(m,v_{e1}) &= ((1-\mu)\lambda + (1-\lambda)\mu + \mu\lambda/3)/K \\ c(m,v_{f1}) &= \mu\lambda/3K, \end{aligned}$$

where for each $c(v, v')$, K is the outdegree of v'.

The algorithm first copies the old points $s^i(v)$ for $v \in V^{*i}$ and $i \leq l$ into the storage location for the detail. If then propagates the incremental changes of the modified points from level l to the coarser levels and adds them to the old points (saved in the detail locations) to find the new points. Then it recomputes the detail vectors that depend on the modified points.

We assume that before the edit, the old points $s^l(v)$ for $v \in V^{*l}$ were saved in the detail locations. The algorithm starts out by building V^{*i-1} and saving the points $s^{i-1}(v)$ for $v \in V^{*i-1}$ in the detail locations. Then the changes resulting from the edit are propagated to level $i-1$. Finally $S\, s^{i-1}$ is computed and used to update the detail vectors on level i.

```
LocalAnalysis(i)

  ∀v ∈ V^{*i} : ∀v' ∈ R^{i-1}(v) :
    V^{*i-1} ∪= {v'}
    v'.d[i-1] := v'.s[i-1]
  ∀v ∈ V^{*i} : ∀v' ∈ R^{i-1}(v) :
    v'.s[i-1] += c(v,v') * (v.s[i] - v.d[i])
  ∀v ∈ V^{*i-1} :
    v.d[i] = v.F(i)^t * (v.s[i] - subd(v, i-1))
    ∀m ∈ N(v,i,1) :
      m.d[i] = m.F(i)^t * (m.s[i] - subd(m, i-1))
```

Note that the odd points are actually computed twice. For the Loop scheme this is less expensive than trying to compute a predicate to avoid this. For Butterfly type schemes this is not true and one can avoid double computation by imposing an ordering on the triangles. The top level code is straightforward:

```
LocalAnalysis

  ∀v ∈ V^{*l} : v.d[l] := v.s[l]
  for i := l downto 0
    LocalAnalysis(i)
```

It is difficult to make incremental local analysis adaptive, as it is formulated purely in terms of vertices. It is, however, possible to adaptively clean up the triangles affected by the edit and (un)refine them if needed.

4.5 Adaptive Rendering

The *adaptive rendering* algorithm decides which triangles will be drawn depending on the rendering performance available and level of detail needed.

The algorithm uses a flag $t.draw$ which is initialized to `false`, but set to `true` as soon as the area corresponding to t is drawn. This can happen either when t itself gets drawn, or when a set of its descendents, which cover t, is drawn. The top level algorithm loops through the triangles starting from the level $n-1$. A triangle

is always responsible for drawing its children, never itself, unless it is a coarsest-level triangle.

```
AdaptiveRender

  for i = n − 1 downto 0
    ∀t ∈ T^i : if not t.leaf then
      Render(t)
  ∀t ∈ T^0 : if not t.draw then
    displaylist.append(t)
```

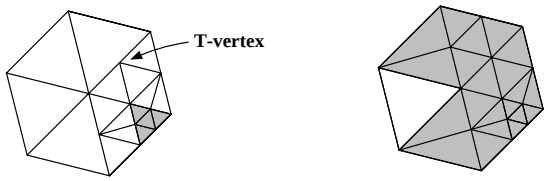


Figure 12: Adaptive rendering: On the left 6 triangles from level i, one has a covered child from level $i+1$, and one has a T-vertex. On the right the result from applying `Render` to all six.

The `Render`(t) routine decides whether the children of t have to be drawn or not (cf. Fig.12). It uses a function `edist`(m) which measures the distance between the point corresponding to the edge's middle vertex m, and the edge itself. In the when case any of the children of t are already drawn or any of its middle vertices are far enough from the plane of the triangle, the routine will draw the rest of the children and set the draw flag for all their vertices and t. It also might be necessary to draw a triangle if some of its middle vertices are drawn because the triangle on the other side decided to draw its children. To avoid cracks, the routine `cut`(t) will cut t into 2, 3, or 4, triangles depending on how many middle vertices are drawn.

```
Render(t)

  if (∃ c ∈ t.child | c.draw = true
   or ∃ m ∈ t.mid_vertex | edist(m) > ε_D) then
    ∀c ∈ t.child :
      if not c.draw then
        displaylist.append(c)
        ∀v ∈ c : v.draw := true
    t.draw := true
  else if ∃ m ∈ t.mid_vertex | m.draw = true
    ∀t' ∈ cut(t) : displaylist.append(t')
    t.draw := true
```

4.6 Data Structures and Code

The main data structure in our implementation is a forest of triangular quadtrees. Neighborhood relations within a single quadtree can be resolved in the standard way by ascending the tree to the least common parent when attempting to find the neighbor across a given edge. Neighbor relations between adjacent trees are resolved explicitly at the level of a collection of roots, i.e., triangles of a coarsest level graph. This structure also maintains an explicit representation of the boundary (if any). Submeshes rooted at any level can be created on the fly by assembling a new graph with some set of triangles as roots of their child quadtrees. It is here that the explicit representation of the boundary comes in, since the actual trees are never copied, and a boundary is needed to delineate the actual submesh.

The algorithms we have described above make heavy use of container classes. Efficient support for sets is essential for a fast implementation and we have used the C++ Standard Template Library. The mesh editor was implemented using OpenInventor and OpenGL and currently runs on both SGI and Intel PentiumPro workstations.

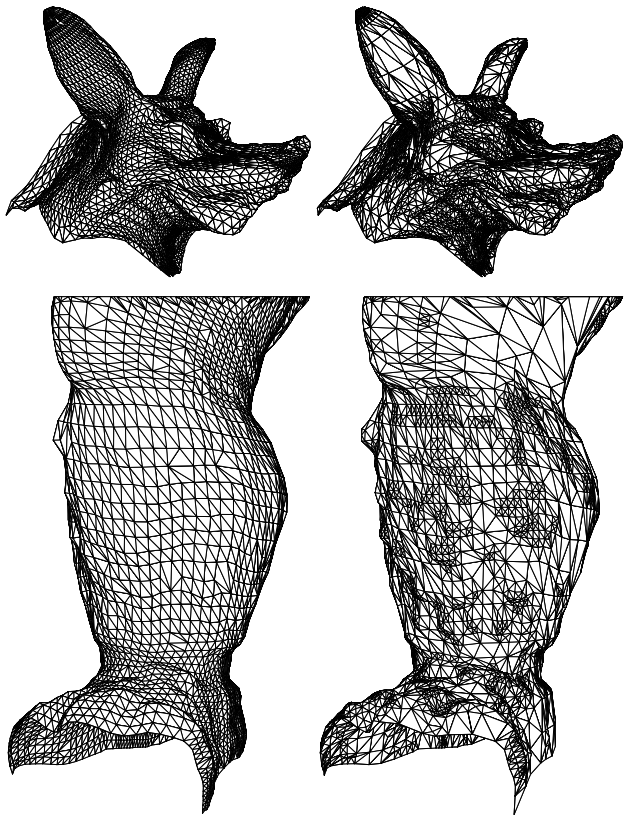


Figure 13: On the left are two meshes which are uniformly subdivided and consist of 11k (upper) and 9k (lower) triangles. On the right another pair of meshes mesh with approximately the same numbers of triangles. Upper and lower pairs of meshes are generated from the same original data but the right meshes were optimized through suitable choice of ϵ_S. See the color plates for a comparison between the two under shading.

5 Results

In this section we show some example images to demonstrate various features of our system and give performance measures.

Figure 13 shows two triangle mesh approximations of the Armadillo head and leg. Approximately the same number of triangles are used for both adaptive and uniform meshes. The meshes on the left were rendered uniformly, the meshes on the right were rendered adaptively. (See also color plate 15.)

Locally changing threshold parameters can be used to resolve an area of interest particularly well, while leaving the rest of the mesh at a coarse level. An example of this “lens” effect is demonstrated in Figure 14 around the right eye of the Mannequin head. (See also color plate 16.)

We have measured the performance of our code on two platforms: an Indigo R10000@175MHz with Solid Impact graphics, and a PentiumPro@200MHz with an Intergraph Intense 3D board.

We used the Armadillo head as a test case. It has approximately 172000 triangles on 6 levels of subdivision. Display list creation took 2 seconds on the SGI and 3 seconds on the PC for the full model. We adjusted ϵ_R so that both machines rendered models at 5 frames per second. In the case of the SGI approximately 113,000 triangles were rendered at that rate. On the PC we achieved 5 frames per second when the rendering threshold had been raised enough so that an approximation consisting of 35000 polygons was used.

The other important performance number is the time it takes to recompute and re-render the region of the mesh which is changing as the user moves a set of control points. This submesh is rendered in immediate mode, while the rest of the surface continues to be rendered as a display list. Grabbing a submesh of 20-30 faces (a typical case) at level 0 added 250 mS of time per redraw, at level 1 it added 110 mS and at level 2 it added 30 mS in case of the SGI. The corresponding timings for the PC were 500 mS, 200 mS and 60 mS respectively.

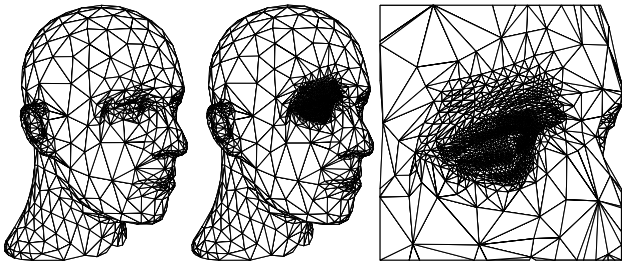


Figure 14: It is easy to change ϵ_S locally. Here a "lens" was applied to the right eye of the Mannequin head with decreasing ϵ_S to force very fine resolution of the mesh around the eye.

6 Conclusion and Future Research

We have built a scalable system for interactive multiresolution editing of arbitrary topology meshes. The user can either start from scratch or from a given fine detail mesh *with subdivision connectivity*. We use smooth subdivision combined with details at each level as a uniform surface representation across scales and argue that this forms a natural connection between fine polygonal meshes and patches. Interactivity is obtained by building both local and adaptive variants of the basic analysis, synthesis, and rendering algorithms, which rely on fast lazy evaluation and tree pruning. The system allows interactive manipulation of meshes according to the polygon performance of the workstation or PC used.

There are several avenues for future research:

- Multiresolution transforms readily connect with compression. We want to be able to store the models in a compressed format and use progressive transmission.
- Features such as creases, corners, and tension controls can easily be added into our system and expand the users' editing toolbox.
- Presently no real time fairing techniques, which lead to more intuitive coarse levels, exist.
- In our system coarse level edits can only be made by dragging coarse level vertices. Which vertices live on coarse levels is currently fixed because of subdivision connectivity. Ideally the user should be able to dynamically adjust this to make coarse level edits centered at arbitrary locations.
- The system allows topological edits on the coarsest level. Algorithms that allow topological edits on all levels are needed.
- An important area of research relevant for this work is generation of meshes with subdivision connectivity from scanned data or from existing models in other representations.

Acknowledgments

We would like to thank Venkat Krishnamurthy for providing the Armadillo dataset. Andrei Khodakovsky and Gary Wu helped beyond the call of duty to bring the system up. The research was supported in part through grants from the Intel Corporation, Microsoft, the Charles Lee Powell Foundation, the Sloan Foundation, an NSF CAREER award (ASC-9624957), and under a MURI (AFOSR F49620-96-1-0471). Other support was provided by the NSF STC for Computer Graphics and Scientific Visualization.

References

[1] BURT, P. J., AND ADELSON, E. H. Laplacian Pyramid as a Compact Image Code. *IEEE Trans. Commun. 31*, 4 (1983), 532–540.

[2] CATMULL, E., AND CLARK, J. Recursively Generated B-Spline Surfaces on Arbitrary Topological Meshes. *Computer Aided Design 10*, 6 (1978), 350–355.

[3] CERTAIN, A., POPOVIĆ, J., DEROSE, T., DUCHAMP, T., SALESIN, D., AND STUETZLE, W. Interactive Multiresolution Surface Viewing. In *SIGGRAPH 96 Conference Proceedings*, H. Rushmeier, Ed., Annual Conference Series, 91–98, Aug. 1996.

[4] DAHMEN, W., MICCHELLI, C. A., AND SEIDEL, H.-P. Blossoming Begets B-Splines Bases Built Better by B-Patches. *Mathematics of Computation 59*, 199 (July 1992), 97–115.

[5] DE BOOR, C. *A Practical Guide to Splines*. Springer, 1978.

[6] DOO, D., AND SABIN, M. Analysis of the Behaviour of Recursive Division Surfaces near Extraordinary Points. *Computer Aided Design 10*, 6 (1978), 356–360.

[7] DYN, N., LEVIN, D., AND GREGORY, J. A. A Butterfly Subdivision Scheme for Surface Interpolation with Tension Control. *ACM Trans. Gr. 9*, 2 (April 1990), 160–169.

[8] ECK, M., DEROSE, T., DUCHAMP, T., HOPPE, H., LOUNSBERY, M., AND STUETZLE, W. Multiresolution Analysis of Arbitrary Meshes. In *Computer Graphics Proceedings*, Annual Conference Series, 173–182, 1995.

[9] FINKELSTEIN, A., AND SALESIN, D. H. Multiresolution Curves. *Computer Graphics* Proceedings, Annual Conference Series, 261–268, July 1994.

[10] FORSEY, D., AND WONG, D. Multiresolution Surface Reconstruction for Hierarchical B-splines. Tech. rep., University of British Columbia, 1995.

[11] FORSEY, D. R., AND BARTELS, R. H. Hierarchical B-Spline Refinement. *Computer Graphics (SIGGRAPH '88 Proceedings)*, Vol. 22, No. 4, pp. 205–212, August 1988.

[12] GORTLER, S. J., AND COHEN, M. F. Hierarchical and Variational Geometric Modeling with Wavelets. In *Proceedings Symposium on Interactive 3D Graphics*, May 1995.

[13] HOPPE, H. Progressive Meshes. In *SIGGRAPH 96 Conference Proceedings*, H. Rushmeier, Ed., Annual Conference Series, 99–108, August 1996.

[14] HOPPE, H., DEROSE, T., DUCHAMP, T., HALSTEAD, M., JIN, H., MCDONALD, J., SCHWEITZER, J., AND STUETZLE, W. Piecewise Smooth Surface Reconstruction. In *Computer Graphics Proceedings*, Annual Conference Series, 295–302, 1994.

[15] HOPPE, H., DEROSE, T., DUCHAMP, T., MCDONALD, J., AND STUETZLE, W. Mesh Optimization. In *Computer Graphics (SIGGRAPH '93 Proceedings)*, J. T. Kajiya, Ed., vol. 27, 19–26, August 1993.

[16] KOBBELT, L. Interpolatory Subdivision on Open Quadrilateral Nets with Arbitrary Topology. In *Proceedings of Eurographics 96*, Computer Graphics Forum, 409–420, 1996.

Figure 15: Shaded rendering (OpenGL) of the meshes in Figure 13.

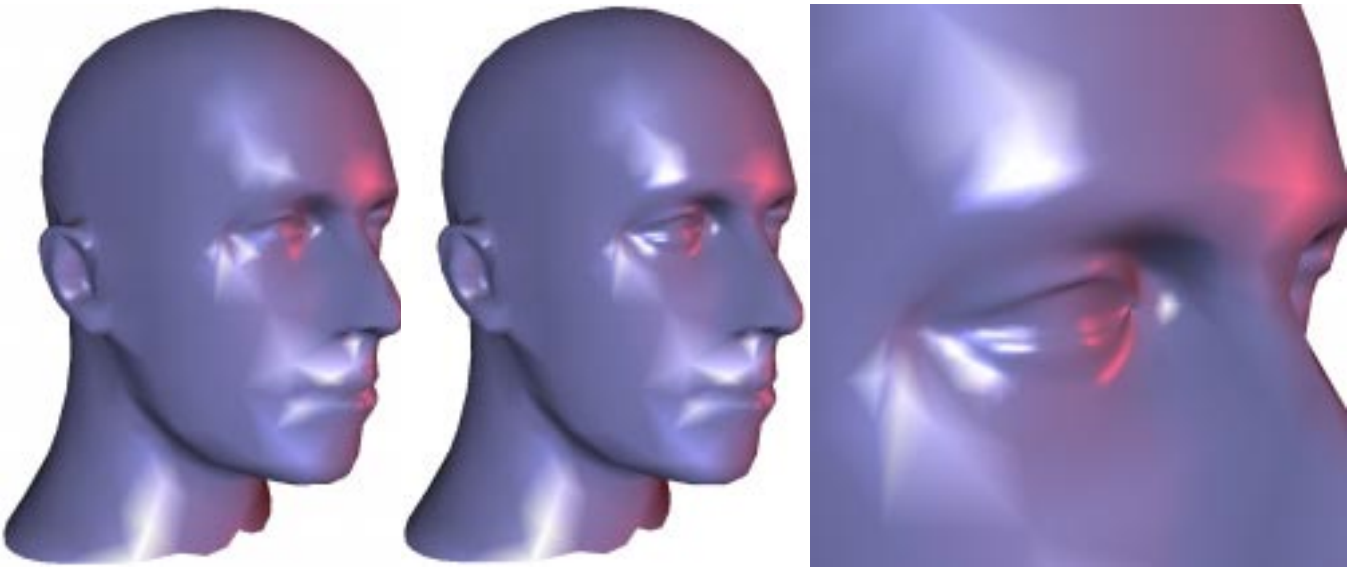


Figure 16: Shaded rendering (OpenGL) of the meshes in Figure 14.

[17] KRISHNAMURTHY, V., AND LEVOY, M. Fitting Smooth Surfaces to Dense Polygon Meshes. In *SIGGRAPH 96 Conference Proceedings*, H. Rushmeier, Ed., Annual Conference Series, 313–324, August 1996.

[18] KURIHARA, T. Interactive Surface Design Using Recursive Subdivision. In *Proceedings of Communicating with Virtual Worlds*. Springer Verlag, June 1993.

[19] LOOP, C. Smooth Subdivision Surfaces Based on Triangles. Master's thesis, University of Utah, Department of Mathematics, 1987.

[20] LOOP, C. Smooth Spline Surfaces over Irregular Meshes. In *Computer Graphics Proceedings*, Annual Conference Series, 303–310, 1994.

[21] LOUNSBERY, M., DEROSE, T., AND WARREN, J. Multiresolution Analysis for Surfaces of Arbitrary Topological Type. *Transactions on Graphics 16*, 1 (January 1997), 34–73.

[22] PETERS, J. C^1 Surface Splines. *SIAM J. Numer. Anal. 32*, 2 (1995), 645–666.

[23] PULLI, K., AND LOUNSBERY, M. Hierarchical Editing and Rendering of Subdivision Surfaces. Tech. Rep. UW-CSE-97-04-07, Dept. of CS&E, University of Washington, Seattle, WA, 1997.

[24] SCHRÖDER, P., AND SWELDENS, W. Spherical wavelets: Efficiently representing functions on the sphere. *Computer Graphics Proceedings, (SIGGRAPH 95)* (1995), 161–172.

[25] SCHWEITZER, J. E. *Analysis and Application of Subdivision Surfaces*. PhD thesis, University of Washington, 1996.

[26] TAUBIN, G. A Signal Processing Approach to Fair Surface Design. In *SIGGRAPH 95 Conference Proceedings*, R. Cook, Ed., Annual Conference Series, 351–358, August 1995.

[27] WELCH, W., AND WITKIN, A. Variational surface modeling. In *Computer Graphics (SIGGRAPH '92 Proceedings)*, E. E. Catmull, Ed., vol. 26, 157–166, July 1992.

[28] ZORIN, D., SCHRÖDER, P., AND SWELDENS, W. Interpolating Subdivision for Meshes with Arbitrary Topology. *Computer Graphics Proceedings (SIGGRAPH 96)* (1996), 189–192.

[29] ZORIN, D. N. *Subdivision and Multiresolution Surface Representations*. PhD thesis, Caltech, Pasadena, California, 1997.

Chapter 7

Subdivision Surfaces and Wavelets

Speaker: Michael Lounsbery

The material in this section was co-authored by Michael Lounsbery, Tony DeRose and Joe Warren and is based on *Multiresolution Analysis for Surfaces of Arbitrary Type*; ACM TOG, 16(1), 1997.

1 Introduction

Multiresolution analysis and wavelets have received considerable attention in recent years fueled largely by the diverse collection of problems that benefit from their use. The basic idea behind multiresolution analysis is to decompose a complicated function into a "simpler" low-resolution part, together with a collection of perturbations called wavelet coefficients that is necessary to recover the original function. For many of the functions encountered in practice, a large percentage of the wavelet coefficients are small, meaning that good approximations can be obtained by using only a few of the largest coefficients. Impressive "lossy" compression rates for images have been achieved using this type of approximation [8].

Two-dimensional wavelets are important for a variety of applications, including image compression. They are generally constructed by forming tensor products of univariate wavelets [7], in much the same way that tensor-product B-spline surfaces are formed by products of univariate B-splines. Unfortunately, tensor-product constructions require the functions to be decomposed be defined on $\mathbb{R}^2$ or on a periodic version of $\mathbb{R}^2$, such as the cylinder or torus. There also exist nontensor-product constructions for wavelets on $\mathbb{R}^2$ [7, 13], but none of these methods are applicable to functions defined on more general topological domains, such as spheres.

In this chapter, we show that by building them on subdivision surfaces, wavelets may be extended to functions defined on domains of arbitrary topological type[1]. (For a more complete description of the work, see Lounsbery *et al.*[16].)

1.1 Applications.

The construction of wavelets on subdivision surfaces considerably extends the class of applications to which multiresolution analysis can be applied, including,

- *Continuous level-of-detail control for animation.* When a complex shape is rendered in an animation, a fully detailed representation of the shape contains much more detail than is required for all but the closest view. Using a compressed subdivision wavelet representation of complex objects, it is possible to greatly reduce the number of polygons in a scene without significantly reducing visible details. Moreover, it is possible to smoothly vary the level of detail, avoiding discontinuous jumps that occur as a result of sudden switching between separate models.

- *Compression of functions defined on surfaces.* Consider the example of Figure 8, where a globe (a geometric sphere) is pseudo-colored according to the elevation. The pseudo-coloring can be thought of as a function that maps each point on the sphere to an *RGB* triple. A straightforward method for storing the function is to store its value at a large number of regularly distributed points; in this case more than one million points were used. The methods in this paper can be used to create compressed wavelet approximations of varying complexity.

- *Multiresolution editing of surfaces.* Hierarchical B-splines, as introduced by Forsey and Bartels [11], provide a powerful mechanism for editing shapes at various levels of detail. However, hierarchical B-splines can only represent a restricted class of surface topologies.

[1] The topological type of a surface refers to its genus, presence of boundary curves, etc.

Subdivision wavelets provide an alternative to hierarchical B-splines, and are capable of representing tangent-plane smooth multiresolution surfaces of arbitrary topological type. Editing at fractional levels of detail can also be achieved using the methods developed by Finkelstein and Salesin [10].

- *Surface optimization.* The multiple levels of approximation produced by wavelet techniques offer a sort of multigrid technique for optimization. Pentland [21] uses wavelet methods to implement multigrid optimization for surface interpolation over regular grids. Meyers [19, 20] shows how wavelets can accelerate the reconstruction of surfaces from contour data. The wavelet techniques described in this chapter can be used to accelerate optimization over subdivision surfaces.

1.2 Wavelets and Multiresolution Analysis.

Wavelets have roots in approximation theory, signal processing, and physics. The formal development of multiresolution analysis as a rigorous theory dates to Mallat in 1989 [17], although related pyramidal techniques for image analysis [4] precede his work. In the brief time since their development, wavelets have found use in signal analysis [17, 22], image processing [8, 18], physics [1], and numerical analysis [3]. More recently, wavelets have been applied to a wide range of computer graphics problems, including curve modeling [10], image editing [2], improved global illumination [23, 12, 5], optimization for surface interpolation [21], creating surfaces from contours [19, 20], and animation control [14]. An excellent tutorial on the use of wavelets in computer graphics is presented by Stollnitz *et al.* [25].

1.3 Intuitive Application to Surfaces.

Although the mathematical underpinnings of multiresolution analysis of surfaces are rather involved, the resulting algorithms are relatively simple. Before delving into the mathematical details, let us first understand how the method can be applied to decompose the polyhedral object shown in Figure 1(a).

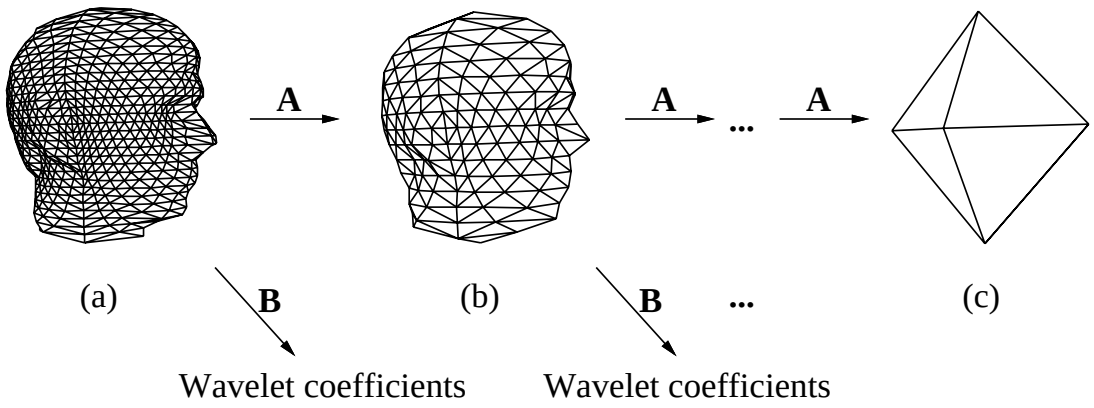


Figure 1: Decomposition of a polyhedral surface.

The main idea behind multiresolution analysis is the decomposition of a function (in this case, a polyhedron expressed as a parametric function on a subdivided sphere) into a low-resolution part

and a "detail" part. The low-resolution part of the polyhedron in Figure 1(a) is shown in Figure 1(b); the vertices in (b) are computed as certain weighted averages of the vertices in (a). These weighted averages essentially implement a *low pass filter* (a filter that eliminates fine-level detail) denoted as **A**. The detail part consists of a collection of fairly abstract coefficients, called *wavelet coefficients*, computed as weighted differences of the vertices in (a). These differencing weights form a high-pass filter **B**. The decomposition process, technically called *analysis*, can be used to further split (b) into an even lower-resolution version and corresponding wavelet coefficients. This cascade of analysis steps, referred to as a *filter bank* algorithm, culminates with the coarsest-level representation in (c), together with wavelet coefficients at each level.

The analysis filters **A** and **B** are constructed so that the original polyhedron can be recovered exactly from the low-resolution version and the wavelet coefficients. Recovery, technically called *synthesis*, reconstructs (a) from (b) and the finest-level wavelet coefficients. Recovery refines each triangle of (b) into four subtriangles by introducing new vertices at edge midpoints, followed by perturbing the resulting collection of vertices according to the wavelet coefficients. The refining and perturbing steps are described by two other filters **P** (the refining filter) and **Q** (the perturbing filter), collectively called synthesis filters.

The trick is to develop the four analysis and synthesis filters so that: (1) the low-resolution versions are good approximations of the original object (in a least-squares sense); (2) the magnitude of a wavelet coefficient reflects a coefficient's importance by measuring the error introduced when the coefficient is set to zero; and (3) analysis and synthesis filter banks should have time complexity that is linear in the number of vertices.

1.4 Multiresolution Analysis on the Real Line.

To formulate multiresolution analysis for surfaces of arbitrary topological type, we must first use a fairly general, but unfortunately abstract, view of multiresolution analysis. There are two basic ingredients for a multiresolution analysis: an infinite chain of nested linear function spaces $V^0 \subset V^1 \subset V^2 \subset \cdots$ and an inner product $\langle f, g \rangle$ defined on any pair of functions $f, g \in V^j$, for some $j < \infty$. Intuitively, V^j contains functions of resolution j, with the detail increasing as j increases.

The inner product is used to define the orthogonal complement spaces W^j as

$$W^j := \{ f \in V^{j+1} \mid \langle f, g \rangle = 0 \ \ \forall g \in V^j \}.$$

Orthogonal complements are often written as $V^{j+1} = V^j \oplus W^j$ because any function $f^{j+1} \in V^{j+1}$ can be written uniquely as an orthogonal decomposition

$$f^{j+1} = f^j + h^j,$$

where $f^j \in V^j$ and $h^j \in W^j$. Orthogonal decompositions are important for approximation purposes: it is easy to show that f^j is the best approximation to f^{j+1} in that it is the unique function in V^j that minimizes the least-squares residual $\langle f^{j+1} - f^j, f^{j+1} - f^j \rangle$. Thus, given a high-resolution function f^{j+1}, its low-resolution part is f^j, and its detail part is h^j.

The following terminology is now standard: *scaling functions* refers to bases for the spaces V^j, and *wavelets* refers to bases for the orthogonal complement spaces. As shown in Section 2.5.4, the analysis and synthesis filters are determined by considering various ways of changing bases between scaling functions and wavelets.

1.5 Subdivision Rules: Split and Average.

Intuitively speaking, subdivision surfaces are defined by iteratively refining a control mesh M^0 so that the sequence of increasingly faceted meshes $M^1, M^2, \ldots$ converges to some limit surface $S = M^\infty$. In each subdivision step, the vertices of M^{j+1} are computed as affine combinations of the vertices of M^j. Thus, if $\mathbf{V}^j$ is a matrix whose i-th row consists of the $x, y,$ and z coordinates of vertex i of M^j, there exists a nonsquare matrix of constants $\mathbf{P}^j$ such that

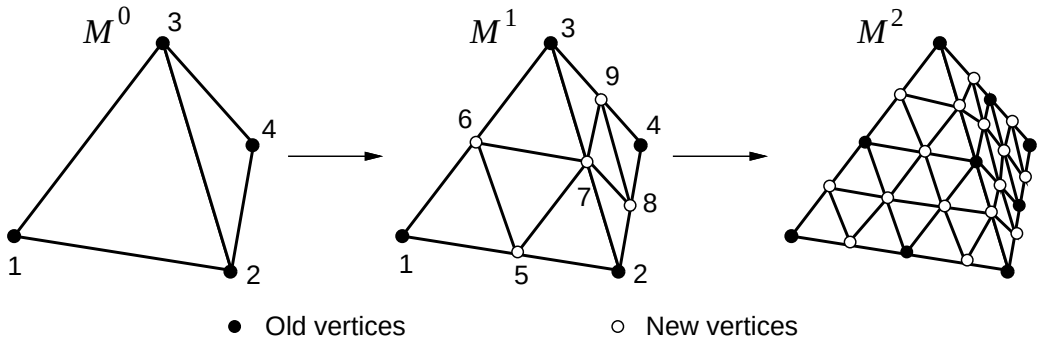
$$\mathbf{V}^{j+1} = \mathbf{P}^j \, \mathbf{V}^j. \tag{1}$$

The matrix $\mathbf{P}^j$ therefore characterizes the subdivision method. The beauty of subdivision surface schemes is that the entries of $\mathbf{P}^j$ depend only on the connectivity of the vertices in M^0, not on the geometric positions of the vertices.

The simplest example of such a scheme is polyhedral subdivision. Its splitting step is the same as that of other primal schemes, but its averaging step is simply the identity — that is, no vertices are repositioned. Thus, given a polyhedron M^0 with triangular faces, a new polyhedron M^1 is built directly by splitting each triangular face of M^0 into four subfaces as in Figure 2. The matrix $\mathbf{P}^0$ characterizing the first polyhedral subdivision step on the tetrahedron is also shown in Figure 2. Running this subdivision scheme for j steps on an initial triangular mesh M^0 produces a mesh M^j. M^j includes the vertices of M^0 together with new vertices introduced through subdivision. The valence of the vertices of M^j that correspond to the original vertices in M^0 remains fixed. The new vertices introduced through subdivision however are always of valence six, corresponding to a regular triangular tiling of the surface. As the mesh is further subdivided, the extraordinary points become increasingly isolated in an otherwise regular tiling of the mesh.

Polyhedral subdivision converges to the original polyhedron covering M^0, that is, to a C^0 surface. Other schemes have also been developed that converge to tangent-plane smooth limit surfaces that either approximate or interpolate the vertices of M^0.

Although they can be somewhat tricky to analyze, subdivision surfaces provide a versatile and efficient tool for modeling surfaces of arbitrary topological type. In the next section, subdivision surfaces are also shown to have the refinability property essential for constructing a multiresolution analysis.



$$\mathbf{P}^0 = \left(\frac{\mathbf{O}^0}{\mathbf{N}^0}\right) = \left(\begin{array}{cccc} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ \hline \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} \end{array}\right) \quad \mathbf{Q}^0 = \left(\begin{array}{cccccc} -\frac{3}{8} & -\frac{3}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & -\frac{3}{8} \\ -\frac{3}{8} & \frac{1}{8} & -\frac{3}{8} & -\frac{3}{8} & \frac{1}{8} & \frac{1}{8} \\ \frac{1}{8} & -\frac{3}{8} & -\frac{3}{8} & \frac{1}{8} & -\frac{3}{8} & \frac{1}{8} \\ \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & -\frac{3}{8} & -\frac{3}{8} & -\frac{3}{8} \\ \frac{5}{8} & -\frac{1}{8} & -\frac{1}{8} & -\frac{1}{8} & \frac{1}{8} & -\frac{1}{8} \\ -\frac{1}{8} & \frac{5}{8} & -\frac{1}{8} & \frac{1}{8} & -\frac{1}{8} & -\frac{1}{8} \\ -\frac{1}{8} & -\frac{1}{8} & \frac{5}{8} & -\frac{1}{8} & -\frac{1}{8} & \frac{1}{8} \\ -\frac{1}{8} & \frac{1}{8} & -\frac{1}{8} & \frac{5}{8} & -\frac{1}{8} & -\frac{1}{8} \\ \frac{1}{8} & -\frac{1}{8} & -\frac{1}{8} & -\frac{1}{8} & \frac{5}{8} & -\frac{1}{8} \\ -\frac{1}{8} & -\frac{1}{8} & \frac{1}{8} & -\frac{1}{8} & -\frac{1}{8} & \frac{5}{8} \end{array}\right)$$

$$\left(\frac{\mathbf{A}^0}{\mathbf{B}^0}\right) = \left(\begin{array}{cccccccccc} \frac{7}{16} & -\frac{1}{16} & -\frac{1}{16} & -\frac{1}{16} & \frac{3}{8} & \frac{3}{8} & -\frac{1}{8} & -\frac{1}{8} & -\frac{1}{8} & \frac{3}{8} \\ -\frac{1}{16} & \frac{7}{16} & -\frac{1}{16} & -\frac{1}{16} & \frac{3}{8} & -\frac{1}{8} & \frac{3}{8} & \frac{3}{8} & -\frac{1}{8} & -\frac{1}{8} \\ -\frac{1}{16} & -\frac{1}{16} & \frac{7}{16} & -\frac{1}{16} & -\frac{1}{8} & \frac{3}{8} & \frac{3}{8} & -\frac{1}{8} & \frac{3}{8} & -\frac{1}{8} \\ -\frac{1}{16} & -\frac{1}{16} & -\frac{1}{16} & \frac{7}{16} & -\frac{1}{8} & -\frac{1}{8} & -\frac{1}{8} & \frac{3}{8} & \frac{3}{8} & \frac{3}{8} \\ \hline -\frac{1}{2} & -\frac{1}{2} & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ -\frac{1}{2} & 0 & -\frac{1}{2} & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & -\frac{1}{2} & -\frac{1}{2} & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & -\frac{1}{2} & 0 & -\frac{1}{2} & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -\frac{1}{2} & -\frac{1}{2} & 0 & 0 & 0 & 0 & 1 & 0 \\ -\frac{1}{2} & 0 & 0 & -\frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right)$$

$$\mathbf{I}^0 = \left(\begin{array}{cccc} 1 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & 1 & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & 1 & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 1 \end{array}\right) \quad \boldsymbol{\alpha}^0 = \left(\begin{array}{cccccc} \frac{3}{8} & \frac{3}{8} & -\frac{1}{8} & -\frac{1}{8} & -\frac{1}{8} & \frac{3}{8} \\ \frac{3}{8} & -\frac{1}{8} & \frac{3}{8} & \frac{3}{8} & -\frac{1}{8} & -\frac{1}{8} \\ -\frac{1}{8} & \frac{3}{8} & \frac{3}{8} & -\frac{1}{8} & \frac{3}{8} & -\frac{1}{8} \\ -\frac{1}{8} & -\frac{1}{8} & -\frac{1}{8} & \frac{3}{8} & \frac{3}{8} & \frac{3}{8} \end{array}\right)$$

Figure 2: Polyhedral subdivision of a tetrahedron and various associated filters

2 Multiresolution Analysis from Subdivision

2.1 Nested linear spaces through subdivision

When formulating multiresolution analysis on the entire real line, the nested sequence of linear spaces required by multiresolution analysis are generally obtained by defining a single scaling fucntion $\phi(x)$ that satisies a refinement equation of the form

$$\phi(x) = \sum_i p_i \phi(2x - i)$$

for some fixed constants p_i. The refinement equation (sometimes called a two-scale relation) guarantees that the spaces defined as

$$V^j := \text{Span}\{\phi(2^j x - i) | i = -\infty, \ldots, \infty\}$$

are nested. In other words, the nested spaces are generated by translations and dilations of a single refinable function $\phi(x)$.

To generalize these ideas to domains of arbitrary topological type, one could attempt to make definitions for what it means to dilate and translate a function on an arbitrary topological domain. One could then try to find a refinable scaling function and proceed as before to define orthogonal complements, wavelets, and so on. We have instead chosen what appears to be a simpler approach, motivated by subdivision.

In this section, we use recursive subfivision to define a collection of functions $\phi_i^j(\mathbf{x})$ that are refinable in the sense that each function with superscript j lies in the span of the functions with superscript $j+1$; the argument $\mathbf{x}$ is a point that ranges over a domain 2-manifold of arbitrary topological type. In one respect, this is a generalization of the approach taken by Daubechies in that her locally supported orthogonal scaling functions are also defined through a recursive subdivision procedure. Although in general the ϕ^{j+1} are not simple dilates of the $\phi^j(\mathbf{x})$, we can nevertheless use them to deine a sequence of nested spaces.

2.2 Refinable Basis Functions from Subdivision.

All subdivision schemes that we have mentioned converge to either C^0 or C^1 surfaces. In such a case, it is possible to show that the resulting limit surface can be parametrized using a function $S(\mathbf{x})$, where $\mathbf{x}$ is a point that ranges over the faces of the initial mesh, M^0 [15]. The initial mesh M^0 therefore serves as the parameter domain for the limit surface. Specifically, it can be shown [15] that for any $j \geq 0$, and any point $\mathbf{x}$ on some face of M^0, $S(\mathbf{x})$ can be written as

$$S(\mathbf{x}) = \sum_i v_i^j \phi_i^j(\mathbf{x}), \tag{2}$$

where $v^j + i$ denotes the i-th vertex of M^j. The scaling functions depend on the subdivision rules used, and are defined through a limiting procedure, or using the evaluation method of Stam[24].

It is convenient to rewrite Equation 2 in matrix form as

$$S(\mathbf{x}) = \mathbf{\Phi}^j(\mathbf{x})\, \mathbf{V}^j, \tag{3}$$

where $\Phi^j(\mathbf{x})$ denotes the row matrix of scaling functions $\phi_i^j(\mathbf{x})$, and where $\mathbf{V}^j$ is as in Equation 1. Equation 3 shows an analogy with B-splines, where the $\phi_i^j(\mathbf{x})$ are comparable to the B-spline basis functions, and the $\mathbf{V}^j$ are akin to the control points.

These scaling functions can also be shown to satisfy the *refinement relation*

$$\Phi^j(\mathbf{x}) = \Phi^{j+1}(\mathbf{x})\mathbf{P}^j. \tag{4}$$

This equation establishes refinability because it states that each of the functions $\phi_i^j(\mathbf{x})$ can be written as a linear combination of the functions $\phi_i^{j+1}(\mathbf{x})$. □

It is convenient to write Equation 4 in block matrix form by writing $\Phi^{j+1}(\mathbf{x})$ as

$$\Phi^{j+1}(\mathbf{x}) = (O^{j+1}(\mathbf{x}) \;\; N^{j+1}(\mathbf{x})). \tag{5}$$

In this equation, $O^{j+1}(\mathbf{x})$ consists of all scaling functions $\phi_i^{j+1}(\mathbf{x})$ associated with the old vertices of M^j (the black vertices in Figure 2) and $N^{j+1}(\mathbf{x})$ consists of the remaining scaling functions associated with the new vertices added when obtaining M^{j+1} from M^j (the white vertices in Figure 2). Equation 4 can now be expressed in block matrix form:

$$\Phi^j(\mathbf{x}) = (O^{j+1}(\mathbf{x}) \;\; N^{j+1}(\mathbf{x})) \begin{pmatrix} \mathbf{O}^j \\ \mathbf{N}^j \end{pmatrix}, \tag{6}$$

where $\mathbf{O}^j$ and $\mathbf{N}^j$ represent the portions of the subdivision matrix $\mathbf{P}^j$ which weight the "old" and "new" vertices, respectively. The block matrix decomposition of $\mathbf{P}^0$ for the example tetrahedron appears in Figure 2.

2.3 Nested Linear Spaces.

Given these relations, a chain of nested linear spaces $V^j(M^0)$ associated with a mesh M^0 can now be defined as follows:

$$V^j(M^0) := \text{Span}(\Phi^j(\mathbf{x})),$$

where the $V^j(M^0)$ are spaces of scalar-valued functions.

Equation 4 implies that these spaces are indeed nested; that is,

$$V^0(M^0) \subset V^1(M^0) \subset \cdots.$$

The notation $V^j(M^0)$ emphasizes that the linear spaces are adapted to M^0 in that they consist of functions having M^0 as the domain.

2.4 Inner Products over Subdivision Surfaces.

Given a chain of nested linear spaces, the other necessary ingredient for the creation of a multiresolution analysis is the existence of an inner product on these spaces. In this section, we define an inner product and sketch a method for computing the inner product of functions defined through subdivision. A detailed treatment is given in Lounsbery[15].

The inner product values are used to build linear systems defining the wavelets. For an efficient implementation of subdivision wavelets, the actual values of neither the inner products nor the wavelets need to be computed at run time, if the base mesh M^0 is known in advance.

2.4.1 Definition.

Given two functions $f, g \in V^j(M^0)$, $j < \infty$, (with some foresight) we define their inner product to be

$$\langle f, g \rangle := \sum_{\tau \in \Delta(M^0)} \frac{1}{\text{Area}(\tau)} \int_{\mathbf{x}' \in \tau} f(\mathbf{x}')\, g(\mathbf{x}')\, d\mathbf{x}',$$

where $\Delta(M^0)$ denotes the set of triangular faces of M^0, and where $d\mathbf{x}'$ is the usual Euclidean area form for the triangle τ in $\mathbb{R}^3$.

This definition of inner product implies that triangles of different geometric size and shape are weighted equally; that is, the inner product is independent of the geometric positions of the vertices of M^0. This inner product definition has two important consequences.

First, in the process of constructing the least-squares best wavelet approximation to a function, each approximated triangle is weighted equally, independent of its true geometric size. The effect of this weighting depends on the particular application, but we have found no problems for the real-world examples that are described in Section 3.1.

Moreover, this measure has an important practical benefit. Because the orthogonal complement spaces are invariant of the geometry of the mesh, a significant amount of precomputation of inner products and wavelets can be performed — which allows the wavelet algorithms to be implemented much more efficiently.

An alternative is to define the inner product so as to weight the integral by the areas of triangles in M^0. Whether such a definition has enough important practical benefit to offset its much increased computation may be an interesting topic for future research.

2.4.2 Computation.

For piecewise linear subdivision leading to polyhedral surfaces, the scaling functions $\phi_i^j(\mathbf{x})$ are simply the *hat functions* over M^0. (The hat function at a vertex **p** in a mesh is the piecewise linear function which is 1 at **p**, blends smoothly to 0 at the neighbors of **p**, and is 0 for all other vertices. An example is shown in Figure 3.) When functions f and g are combinations of piecewise linear scaling functions, it is fairly simple to directly compute the integral.

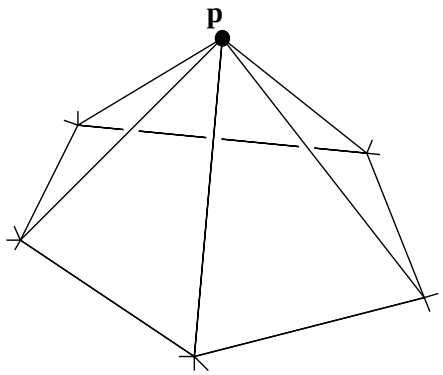


Figure 3: A piecewise linear hat function for a vertex **p** of valence 5.

In general, however, the nature of the scaling functions can make explicit integration difficult. In these cases, one could estimate the inner product $\langle f, g \rangle$ by subdividing the scaling functions

some number of times and directly integrating the resulting piecewise linear approximation. In this section, we will see that such estimation is unnecessary, and that exact integration is still possible for the general case.

Indeed, one may compute $\langle f, g \rangle$ exactly whenever f and g are given as expansions in ϕ_i^j:

$$f(\mathbf{x}) = \sum_i f_i^j \phi_i^j(\mathbf{x}) \qquad\qquad g(\mathbf{x}) = \sum_i g_i^j \phi_i^j(\mathbf{x}).$$

Bilinearity of the inner product allows $\langle f, g \rangle$ to be written in matrix form as

$$\langle f, g \rangle = \mathbf{g}^T \mathbf{I}^j \mathbf{f}$$

where $\mathbf{f}$ and $\mathbf{g}$ are column matrices consisting of the coefficients of f and g, respectively, and where $\mathbf{I}^j$ is the square matrix whose i,i'-th entry is $(\mathbf{I}^j)_{i,i'} = \langle \phi_i^j, \phi_{i'}^j \rangle$. The inner product matrix $\mathbf{I}^0$ for the example tetrahedron appears in Figure 2.

When the scaling functions in $\Phi^j(\mathbf{x})$ are locally supported, the subdivision matrices $\mathbf{P}^j$ are sparse, meaning that $\mathbf{I}^j$ is also sparse. $\mathbf{I}^j$ can be computed exactly by solving a system of linear equations; hence, $\langle f, g \rangle$ can also be computed exactly. This result is somewhat surprising because it does not require a closed-form expression for the scaling functions – only a subdivision relation.

Without going too far into the details, the basic idea between the exact integration procedure is to establish the following linear recurrence that relates $\mathbf{I}^j$ to $\mathbf{I}^{j+1}$; $\mathbf{I}^j$ can be written as

$$\mathbf{I}^j = (\mathbf{P^j})^\mathbf{T} \mathbf{I^{j+1}} \mathbf{P^j}, \mathbf{j} = \mathbf{0}, \ldots$$

Since $\mathbf{P^j}$ is known, the only unknowns in the above equation are the entries of the $\mathbf{I}$s. It turns out that the infinite chain of recurrences has only a finite number of distinct entries (up to a common scale factor), meaning that a finite linear system can be solved to determine these entries.

Once an absolute scale for the homogeneous system is chosen, an arbitrary integral such as $\langle f, g \rangle$ can be computed exactly.

2.5 Wavelets from Subdivision.

We have established nested linear spaces and an inner product. We are now in a position to define our wavelet spaces as

$$W^j(M^0) \;:=\; \{ f \in V^{j+1}(M^0) \;|\; \langle f, g \rangle = 0 \;\; \forall g \in V^j(M^0) \},$$

and to construct wavelets, that is, a set of functions $\Psi^j(\mathbf{x}) = (\psi_1^j(\mathbf{x}), \psi_2^j(\mathbf{x}), \ldots)$ that span the orthogonal complement space $W^j(M^0)$. (The elements of $\Psi^j(\mathbf{x})$ are not mutually orthogonal. Some authors, including Chui [6], refer to such functions as *pre-wavelets*.)

It is of significant practical importance that the decomposition and reconstruction filters associated with these wavelets are constructed and applied in linear time. This practical concern drives much of the development in this section.

2.5.1 The Construction.

Our wavelet construction consists of two steps. First, we build a basis for $V^{j+1}(M^0)$ using the scaling functions $\Phi^j(\mathbf{x})$ and the new scaling functions $N^{j+1}(\mathbf{x})$ in $V^{j+1}(M^0)$. It is straightforward to

show that $\Phi^j(\mathbf{x})$ and $N^{j+1}(\mathbf{x})$ together span $V^{j+1}(M^0)$ if, and only if, the matrix $\mathbf{O}^j$ (encoding the subdivision rule around the old vertices) is invertible. Most primal subdivision methods, including polyhedral subdivision and the butterfly method, have this property.[2] Given a function $S^{j+1}(\mathbf{x})$ in $\mathbf{V}^{j+1}(M^0)$ expressed as an expansion in the basis $(\Phi^j(\mathbf{x})\ N^{j+1}(\mathbf{x}))$, an approximation in $V^j(M^0)$ can be obtained by *restriction to* $\Phi^j(\mathbf{x})$; that is, by setting to zero the coefficients corresponding to $N^{j+1}(\mathbf{x})$. However, this method generally does not produce the best least-squares approximation.

To ensure the best least-squares approximation after restriction to $\Phi^j(\mathbf{x})$, we may orthogonalize the new basis functions $N^{j+1}(\mathbf{x})$ by computing their projection into $W^j(M^0)$. The resulting functions $\Psi^j(\mathbf{x})$ are wavelets because they form a basis for $W^j(M^0)$. Expressed in matrix form:

$$N^{j+1}(\mathbf{x}) = \Psi^j(\mathbf{x}) + \Phi^j(\mathbf{x})\,\boldsymbol{\alpha}^j. \tag{7}$$

Figure 4 is a plot of one such wavelet for the case of polyhedral subdivision. If $S^{j+1}(x)$ is expanded in terms of $\Phi^j(\mathbf{x})$ and $\Psi^j(\mathbf{x})$, then the restriction of $S^{j+1}(\mathbf{x})$ to $\Phi^j(\mathbf{x})$ is guaranteed to be the best approximation to $S^{j+1}(\mathbf{x})$ in $V^j(M^0)$ in a least-squares sense.

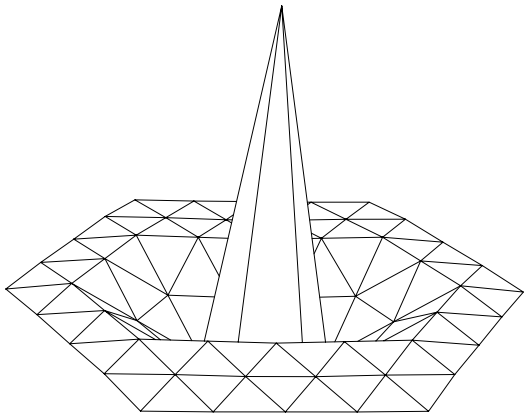


Figure 4: A polyhedral wavelet centered on a regular vertex of valence 6.

2.5.2 Computation of Wavelets.

The coefficients $\boldsymbol{\alpha}^j$ are solutions to the linear system formed by taking the inner product of each side of Equation 7 with $\Phi^j(\mathbf{x})$:

$$\begin{aligned} \langle \Phi^j(\mathbf{x}), \Phi^j(\mathbf{x}) \rangle\, \boldsymbol{\alpha}^j &= \langle \Phi^j(\mathbf{x}), N^{j+1}(\mathbf{x}) \rangle \\ &= (\mathbf{P}^j)^T\, \langle \Phi^{j+1}(\mathbf{x}), N^{j+1}(\mathbf{x}) \rangle \end{aligned} \tag{8}$$

where $\langle \mathbf{F}, \mathbf{G} \rangle$ stands for the matrix whose i,i'-th entry is $\langle (\mathbf{F})_i, (\mathbf{G})_{i'} \rangle$. The matrix denoted by $\langle \Phi^j(\mathbf{x}), \Phi^j(\mathbf{x}) \rangle$ is therefore simply $\mathbf{I}^j$, and the matrix $\langle \Phi^{j+1}(\mathbf{x}), N^{j+1}(\mathbf{x}) \rangle$ is a submatrix of $\mathbf{I}^{j+1}$ that consists of the columns that correspond to members of $N^{j+1}(\mathbf{x})$. The matrix $\boldsymbol{\alpha}^0$ for the example tetrahedron appears in Figure 2.

This system of equations may be solved for the coefficients $\boldsymbol{\alpha}^j$. Following Equation 7, the values $\boldsymbol{\alpha}^j$ are sufficient to construct the wavelets centered around the vertex at the center of each scaling function in $N^{j+1}(\mathbf{x})$.

[2] One notable exception is Catmull-Clark subdivision for vertices of valence three. However, the subdivision rule for such vertices can be easily modified to produce an invertible matrix.

2.5.3 Locally Supported Approximations to the Wavelets.

Although this construction produces wavelets orthogonal to the scaling functions at level j, there are some important practical difficulties that result. First, the inner product matrix $\mathbf{I}^j$ used to solve Equation 8 must be inverted. Second, the inverse of $\mathbf{I}^j$ is dense even though $\mathbf{I}^j$ is sparse. As a consequence, the resulting wavelets are globally supported on M^0, implying that filter bank decomposition and reconstruction algorithms using these wavelets require quadratic time. In order to apply these operations in linear time, we must build wavelets that are locally supported.

Constructing wavelets of local support is a common problem in the wavelet literature and has been handled in various ways. Chui [6] is able to develop unique locally supported wavelets for B-spline curves. Mallat [17] builds wavelets that, in addition to being orthogonal to the scaling functions, are mutually orthogonal. Unlike Chui, Mallat's conditions do not allow the construction of locally supported wavelets. Instead, in practice Mallat approximates his wavelets with locally supported truncations that are not strictly orthogonal to their scaling functions. Because the values in Mallat's wavelets decay exponentially away from the center, the deviation from orthogonality is bounded by choosing the size of the local support.

When truncation iof this sort s used, decomposition and reconstruction are not inverses. This is a disadvantage, because it means that decomposition followed by reconstruction does not exactly reproduce the original surface. Nevertheless, Mallat has found the resulting nonorthogonal approximation perfectly adequate for practical use.

Like those of Mallat, the wavelets resulting from Equation 8 are orthogonal, and are globally supported over the entire domain mesh M^0, with values that appear to decay exponentially. We currently do not know of a construction leading to unique locally supported versions of subdivision wavelets, nor whether such a construction always exists. We will therefore obtain locally supported functions by relaxing the condition that the $\psi_i^j(\mathbf{x})$'s lie in $W^j(M^0)$. Instead, one may construct them to span a space V_*^j that is some (nonorthogonal) complement of $V^j(M^0)$ in $V^{j+1}(M^0)$. This approach, which works well in practice, improves on Mallat's approach of merely truncating coefficients. Instead, we build functions in V_*^j that are the least-squares projections of $\psi_i^j(\mathbf{x})$ into W^j. This practice ensures that they are as close as possible to orthogonal, given their restricted support. Additionally, this construction ensures that the analysis and synthesis filters are inverses, which truncation alone cannot achieve. We will see below that it is possible to make these approximations $V_*^j(M^0)$ arbitrarily close to $W^j(M^0)$, at the expense of increasing their support.

These approximations are constructed by selecting their supports *a priori*. For each $\psi_i^j(\mathbf{x})$, those members of $\Phi^j(\mathbf{x})$ whose supports are sufficiently distant from the support of $(N^{j+1})_i$ have their corresponding coefficients in the i-th column of $\boldsymbol{\alpha}^j$ set to zero. The remaining nonzero coefficients can be found by solving a smaller, local variant of Equation 8. By allowing more of the coefficients of $\boldsymbol{\alpha}^j$ to be nonzero, the supports grow. The wavelets we have constructed have values which are observed to decay exponentially. Therefore, as the support grows, the local approximation $V_*^0(M^0)$ built for them quickly approaches $V^0(M^0)$.

2.5.4 A Filter Bank Algorithm.

Analysis and synthesis on the infinite real line are based on an assumption of spatial invariance — every place looks like every other place. This means that standard analysis and synthesis filters can be represented by a convolution kernel, that is, by a sequence of real numbers. This is not the case for multiresolution analysis on arbitrary topological domains. The filter coefficients in

general must vary over the mesh, so the filters are represented by (hopefully sparse) matrices.

The analysis and synthesis filters can be conveniently expressed using block matrix equations. Let $\Psi^j(\mathbf{x})$ denote the row matrix of the locally supported wavelet approximations spanning $V_*^j(M^0)$. For any multiresolution analysis the synthesis filters are defined by the relation

$$\left(\Phi^j(\mathbf{x}) \;\; \Psi^j(\mathbf{x})\right) = \Phi^{j+1}(\mathbf{x}) \begin{pmatrix} \mathbf{P}^j & \mathbf{Q}^j \end{pmatrix}, \tag{9}$$

and the analysis filters are obtained from the inverse relation

$$\begin{pmatrix} \mathbf{A}^j \\ \mathbf{B}^j \end{pmatrix} = \begin{pmatrix} \mathbf{P}^j & \mathbf{Q}^j \end{pmatrix}^{-1}. \tag{10}$$

For the construction, it is again convenient to write $\Phi^{j+1}(x)$ in block form as

$$(O^{j+1}(\mathbf{x}) \; N^{\,j+1}(\mathbf{x})).$$

It then follows from Equation 7 that the synthesis filters can be written in block form as

$$\begin{pmatrix} \mathbf{P}^j & \mathbf{Q}^j \end{pmatrix} = \begin{pmatrix} \mathbf{O}^j & -\mathbf{O}^j \boldsymbol{\alpha}^j \\ \mathbf{N}^j & \mathbf{1} - \mathbf{N}^j \boldsymbol{\alpha}^j \end{pmatrix}, \tag{11}$$

where $\mathbf{1}$ denotes the identity matrix. The analysis filters are obtained from Equation 10. (Examples for both are shown for the tetrahedron in Figure 2.)

From a practical standpoint, it is critical that the analysis and synthesis matrices are sparse. To achieve linear time decomposition and reconstruction, they must each have a constant number of nonzero entries in each row. If $\mathbf{P}^j$ and $\boldsymbol{\alpha}^j$ are sparse, then $\mathbf{Q}^j$ is sparse. Unfortunately, the analysis filters derived from Equation 10 need not be sparse. For interpolating subdivision schemes such as polyhedral subdivision and the G^1 "butterfly" scheme of Dyn *et al.* [9], the situation is much improved. Such interpolating schemes have the property that $\mathbf{O}^j$ is the identity matrix. Equation 11 in this case is greatly simplified; the resulting filters are

$$\begin{pmatrix} \mathbf{P}^j & \mathbf{Q}^j \end{pmatrix} = \begin{pmatrix} \mathbf{1} & -\boldsymbol{\alpha}^j \\ \mathbf{N}^j & \mathbf{1} - \mathbf{N}^j \boldsymbol{\alpha}^j \end{pmatrix} \qquad \begin{pmatrix} \mathbf{A}^j \\ \mathbf{B}^j \end{pmatrix} = \begin{pmatrix} \mathbf{1} - \boldsymbol{\alpha}^j \mathbf{N}^j & \boldsymbol{\alpha}^j \\ -\mathbf{N}^j & \mathbf{1} \end{pmatrix}.$$

If $\mathbf{P}^j$ and $\boldsymbol{\alpha}^j$ are sparse, then all four filters are also sparse, leading to linear time decomposition and reconstruction. The situation is less desirable for methods related to B-splines, such as Loop's scheme and Catmull-Clark surfaces. For these subdivision schemes, the synthesis filters are sparse, but the analysis filters are dense. This implies that decomposition is still possible in O(n) time, but that the speed of reconstruction depends on the time to invert the sparse decomposition matrix.

The analysis filters can be used to decompose a surface $S^{j+1}(\mathbf{x})$ in $V^{j+1}(M^0)$ given by

$$S^{j+1}(\mathbf{x}) = \sum_i v_i^{j+1} \phi_i^{j+1}(\mathbf{x}) \tag{12}$$

into a lower resolution part in $V^j(M^0)$ plus a detail part in $V_*^j(M^0)$

$$S^{j+1}(\mathbf{x}) = \sum_i v_i^j \phi_i^j(\mathbf{x}) + \sum_i w_i^j \psi_i^j(\mathbf{x})$$

as follows. Let $\mathbf{V}^j$ be as in Equation 3, and let $\mathbf{W}^j$ denote the corresponding matrix of wavelet coefficients w_i^j. We can rewrite Equation 12 in matrix form and substitute the definition of the analysis filters. Thus:

$$\begin{aligned} S^{j+1}(\mathbf{x}) &= \Phi^{j+1}(\mathbf{x})\,\mathbf{V}^{j+1} \\ &= \left(\Phi^j(\mathbf{x})\,\Psi^j(\mathbf{x})\right)\begin{pmatrix}\mathbf{A}^j\\ \mathbf{B}^j\end{pmatrix}\mathbf{V}^{j+1} \\ &= \Phi^j(\mathbf{x})\mathbf{A}^j\,\mathbf{V}^{j+1} + \Psi^j(\mathbf{x})\mathbf{B}^j\,\mathbf{V}^{j+1} \end{aligned}$$

therefore,

$$\mathbf{V}^j = \mathbf{A}^j\,\mathbf{V}^{j+1} \qquad\qquad \mathbf{W}^j = \mathbf{B}^j\,\mathbf{V}^{j+1}.$$

Of course, the analysis filters $\mathbf{A}^{j-1}$ and $\mathbf{B}^{j-1}$ can now be applied to $\mathbf{V}^j$ to yield $\mathbf{V}^{j-1}$, $\mathbf{W}^{j-1}$, etc. A similar argument shows that $\mathbf{V}^{j+1}$ can be recovered from $\mathbf{V}^j$ and $\mathbf{W}^j$ using the synthesis filters:

$$\mathbf{V}^{j+1} = \mathbf{P}^j\,\mathbf{V}^j + \mathbf{Q}^j\,\mathbf{W}^j.$$

3 Wavelet Compression of Surfaces

3.1 Polyhedral Examples.

In this section, our theory is applied to two compression problems: the compression of a polyhedral model consisting of over 32,000 triangles, and compression of a piecewise linear representation of a color function defined on over one million points on the globe. The data for these examples were resampled from cylindrical sections onto a subdivided octahedron.

3.1.1 Geometric Data.

The input for the first example (shown in Figure 5(a)) is a polyhedral mesh consisting of 32,768 triangles whose vertices were resampled from laser range data originally provided through the courtesy of Cyberware, Inc. The triangulation was created by recursively subdividing an octahedron six times. The octahedron therefore serves as the domain mesh M^0, with the input triangulation considered as a parametric function $S(\mathbf{x}), \mathbf{x} \in M^0$ lying in $V^6(M^0)$. More precisely, if v_i^6 denotes the vertices of the input mesh, $S(\mathbf{x})$ can be written as

$$S(\mathbf{x}) = \Phi^6(\mathbf{x})\mathbf{V}^6, \qquad \mathbf{x} \in M^0$$

where the scaling functions $\Phi^6(\mathbf{x})$ are the (piecewise linear) functions defined through polyhedral subdivision.

The locally supported wavelet approximations $\psi_i^j(\mathbf{x})$ for this example are chosen to be supported on 2-discs. (The k-disc around a vertex v of a triangulation is defined to be the set of all triangles whose vertices are reachable from v by following k or fewer edges of the triangulation.) The filter bank process outlined in Section 2.5.4 can be applied in linear time to rewrite $S(\mathbf{x})$ in the form

$$S(\mathbf{x}) = \Phi^0(\mathbf{x})\ \mathbf{V}^0 + \sum_{j=0}^{5} \Psi^j(\mathbf{x})\ \mathbf{W}^j.$$

The first term describes a base shape as the projection of $S(\mathbf{x})$ into $V^0(M^0)$, which in this case is an approximating octahedron with vertex positions given by the six rows of $\mathbf{V}^0$. For this data, the decomposition stage of the filter bank runs in about 14 seconds on an SGI Indigo^2 Extreme.

Approximations to the original mesh $S(\mathbf{x})$ can be easily obtained from the wavelet expansion using only coefficients greater than a pre-selected threshold. The models in Figure 6(b), (d), and (f) are compressed to 1%, 13%, and 32%, respectively. Notice that thresholding causes the mesh to refine in areas of high detail, while leaving large triangles in areas of relatively low detail. An unoptimized implementation of the entire decomposition and reconstruction process that produces Figure 6(d) runs in about 53 seconds from input to output.

Figure 6 also illustrates the use of wavelet approximations for automatic level-of-detail control in rendering. The original Spock polyhedron is shown in full resolution in Figure 5(a). Views of the full-resolution model from various distances are shown in the rest of Figure 5. When viewing the input polyhedron at these distances, it is inefficient and unnecessary to render all 32,000 triangles. The approximations shown in the left column of Figure 6 may instead be used without significantly degrading image quality.

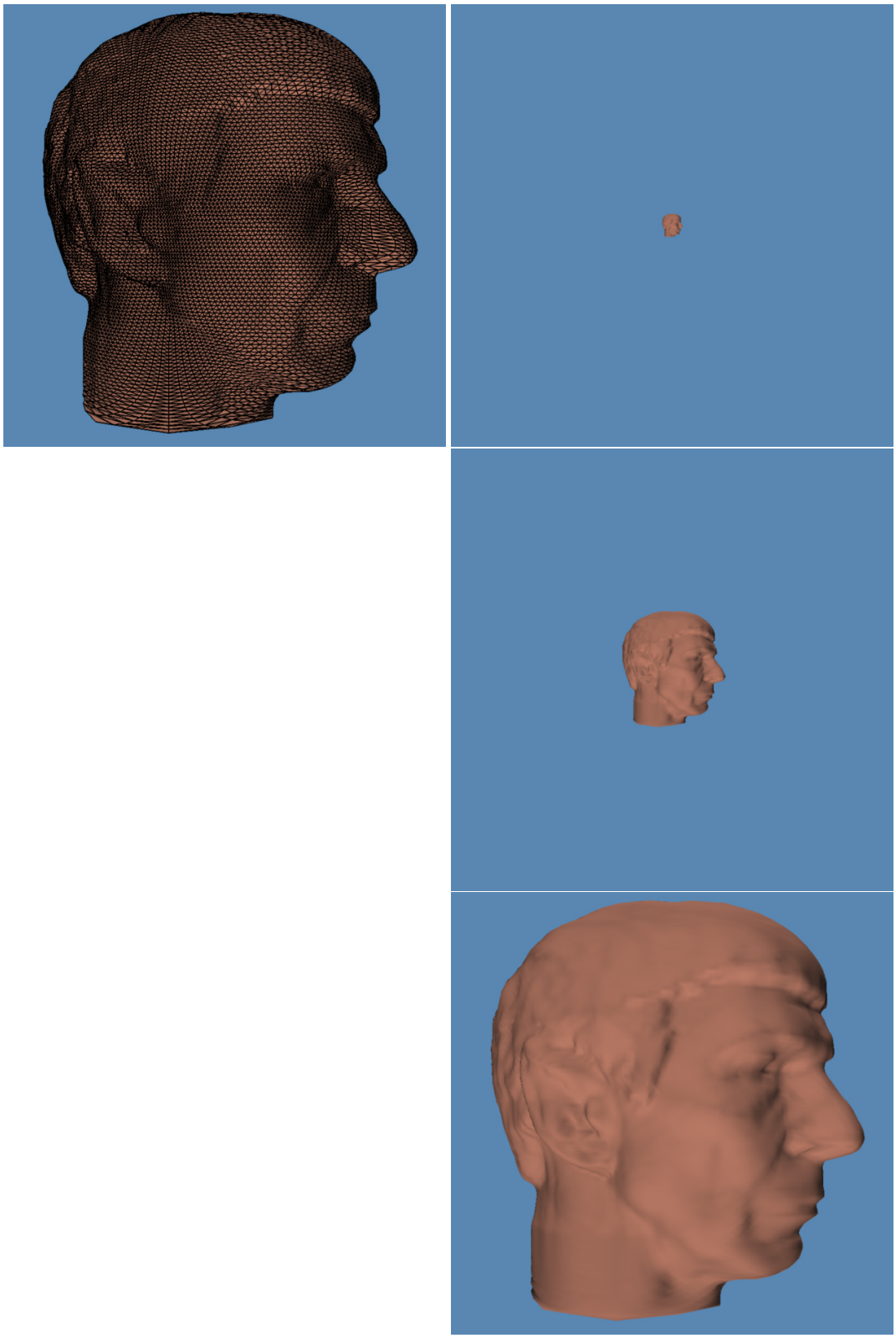


Figure 5: The full-resolution Spock polyhedron. Left: the full mesh (16,386 points, 32,768 triangles). Right: Gouraud-shaded views of the full mesh at various distances.

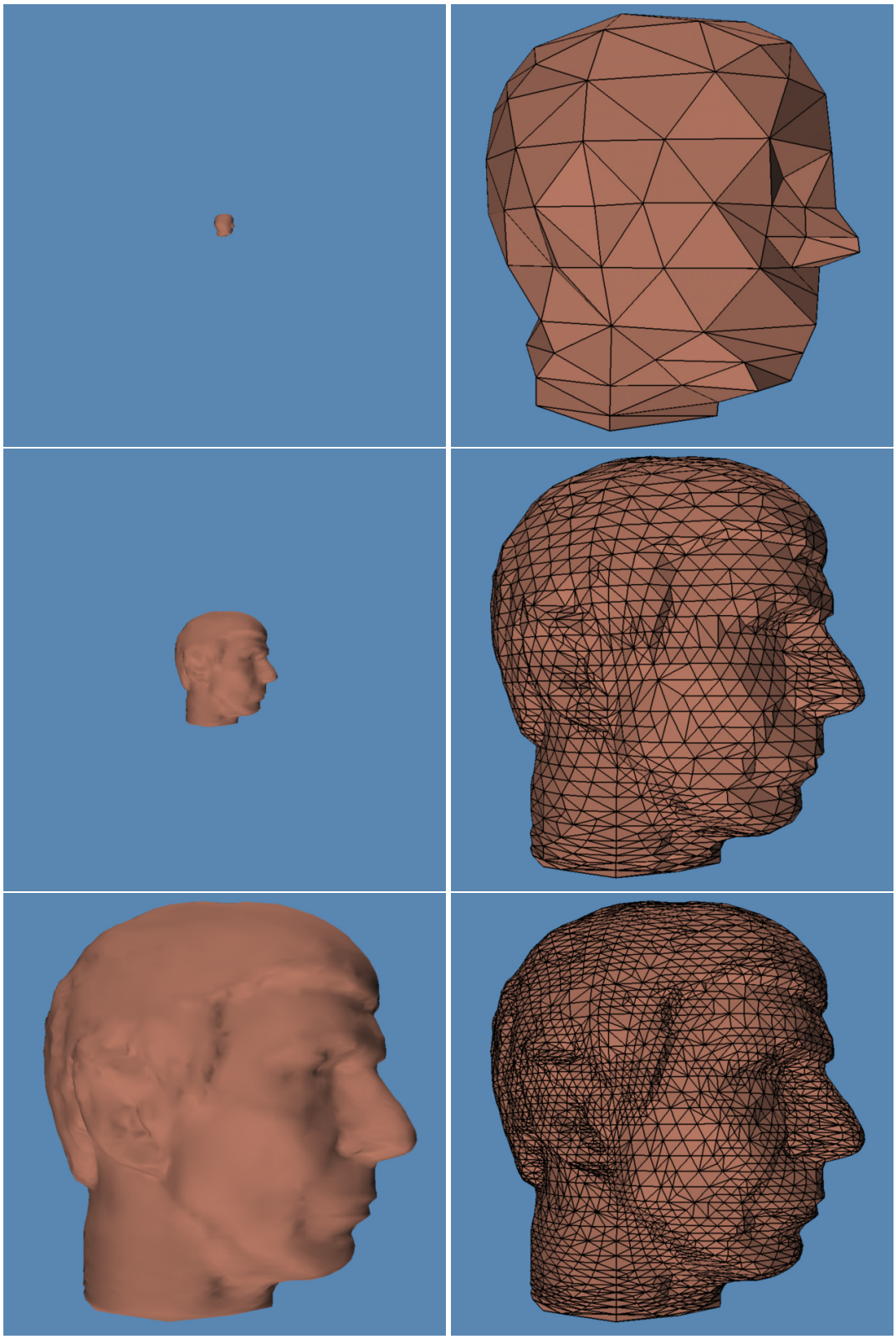


Figure 6: Wavelet approximations of the Spock polyhedron.

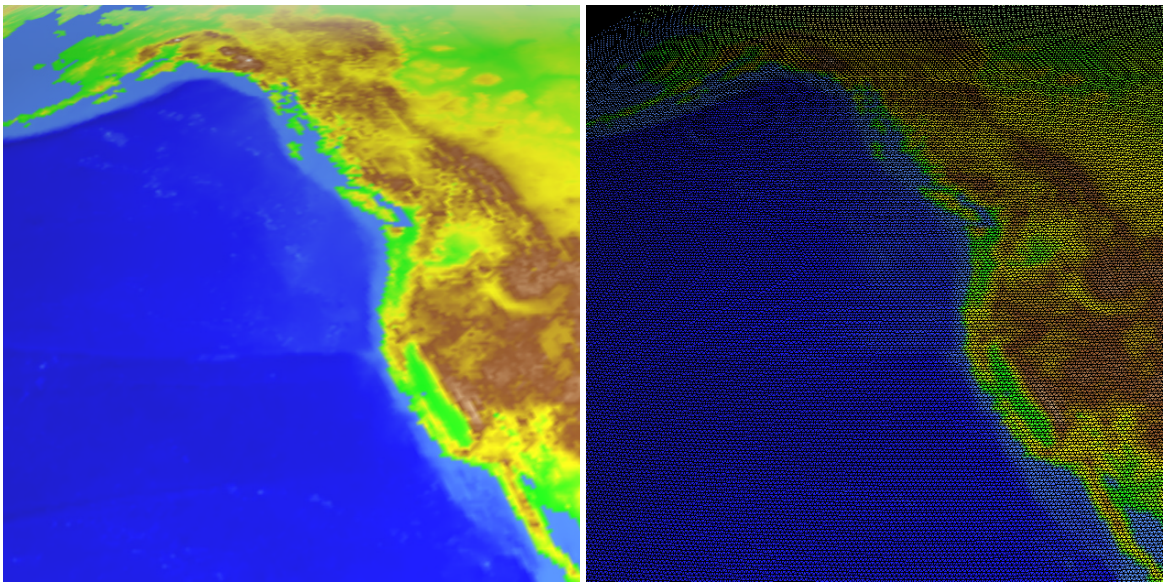


Figure 7: Close-up of Earth data before compression

3.1.2 Color Data.

Subdivision wavelets may be applied to more general functions over surfaces than geometric data. Figure 8 demonstrates another application — that of compressing a function on the sphere. In this example, elevation and bathymetry data obtained from the U.S. National Geophysical Data Center was used to create a piecewise linear pseudocoloring of the sphere. The resulting color function was represented by 2,097,152 triangles and 1,048,578 vertices. The full-resolution pseudocoloring was too large to be rendered on an SGI Indigo^2 Extreme with 128 MB, and is therefore not shown in its entirety in Figure 8. Instead, Figure 7 shows a close-up of a region that is compressed in Figure 8. An appreciation for the density of the data can be obtained from Figure 7(b), where even at close range the mesh lines of the original uncompressed data are so dense that the image appears almost completely black.

The approximations shown in Figure 8(a)-(f) were produced using subdivision wavelet compression. Figure 8(a) shows a distant view of the Earth using an approximation of only 0.1% [the mesh is shown in (b)]. Likewise, Figures 8(c) and (d) show the result of compression to 2% for a medium-range view. At close range the compression to 16% in (e) is nearly indistinguishable from the full-resolution model in Figure 7(a). A comparison of the meshes shown in Figure 7(b) and Figure 8(f) reveals the striking degree of compression achieved in this case.

3.1.3 Run Times.

Example run times for the Spock and Earth data sets appear Tables 1 and 2. All times reflect compression done on an SGI Indigo^2 Extreme with a 100 MHz MIPS R4000 CPU and 128 MB of physical memory.

The first two lines of Table 1 give information about the size of the Spock input. *Decomposition time* gives the filter bank time needed for any compression of the Spock data, and *Selection time* is the time required to select wavelet coefficients by thresholding.

Table 2 is broken down to compare side by side reconstruction times for different coefficient

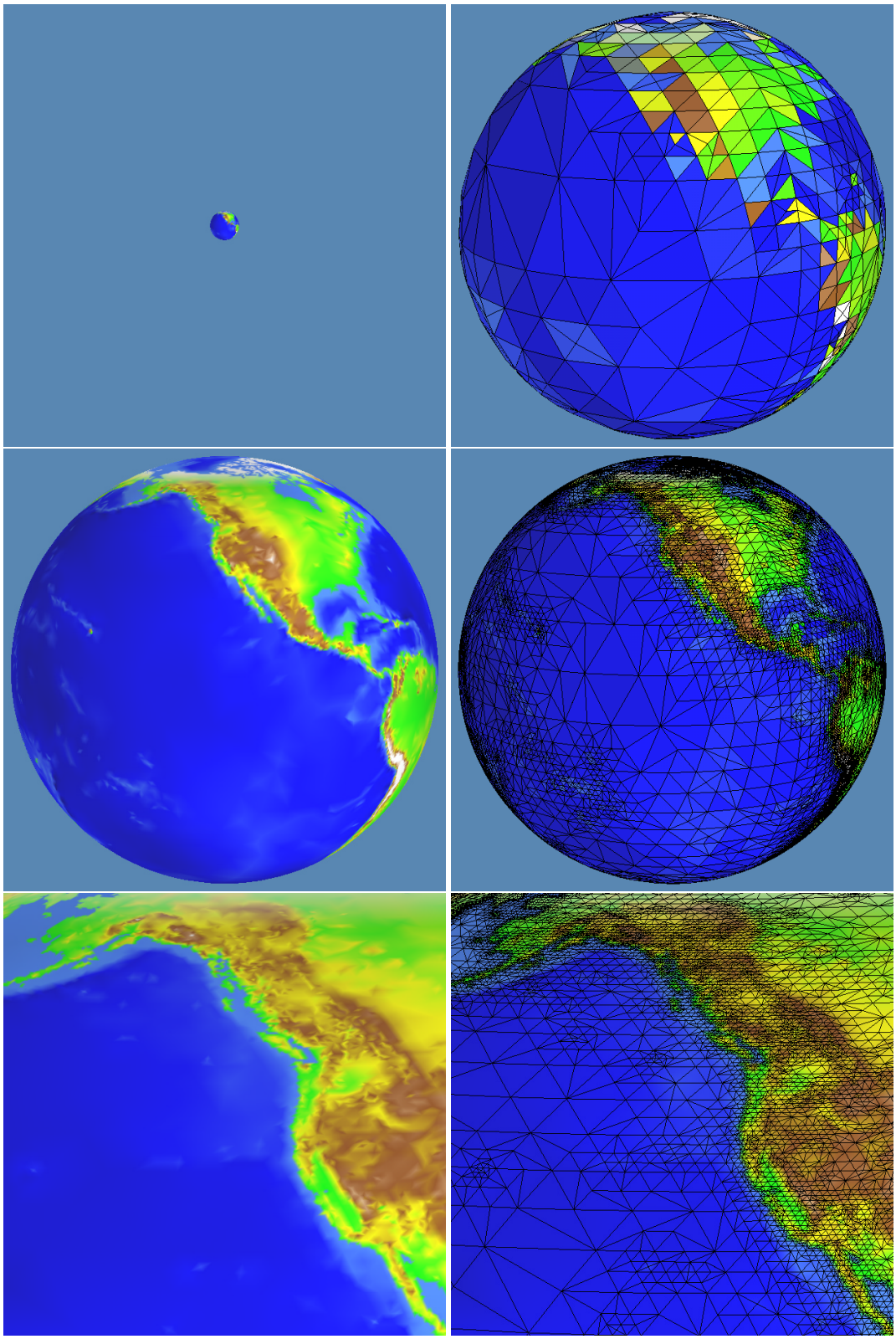


Figure 8: Approximating color as a function over the sphere.

thresholds. The thresholds chosen were used to generate the compressions shown in Figure 6(b), (d), and (f). In the table, *Coefficients selected* tells how many coefficients were selected using the indicated coefficient threshold. Information about the size of each reconstruction is given in *Triangles in reconstruction* and *Fraction of original*. *Reconstruction time* includes time for adding back wavelets and recursive expansion, but not for triangulation. Reconstruction and decomposition times are given separately. When separate compressions of the same data set are run, decomposition needs to be run only once, but selection and reconstruction need to be run for each compression. The times for coefficient selection and for triangulation are much less than for decomposition and reconstruction; their contribution is included in *Total processing time*, which gives the beginning-to-end time for a single decomposition and reconstruction session. This value is the sum of decomposition, selection, and reconstruction time, plus all additional operations, including memory management and disk I/O. (Disk I/O is particularly time-consuming for the file of Earth input, which takes up more than 51 MB of ASCII data before compression.)

Tables 3 and 4 give information about the Earth data set, using the same format as the Spock tables.

Table 1: Statistics for Spock data.

INPUT VERTICES	16,386
INPUT TRIANGLES	32,768
SUBDIVISION DEPTH	6
DECOMPOSITION TIME	14
SELECTION TIME	1

Table 2: Spock compression results

FIGURE	6(b)	6(d)	6(f)
COEFFICIENT THRESHOLD	2.0	0.06	0.017
COEFFICIENTS SELECTED	107	1,966	5,054
TRIANGLES IN RECONSTRUCTION	240	4,272	10,704
FRACTION OF ORIGINAL	0.007	0.13	0.33
RECONSTRUCTION TIME	4	29	63
TOTAL PROCESSING TIME	25	53	87

Table 3: General Earth data.

INPUT VERTICES	1,048,578
INPUT TRIANGLES	2,097,152
SUBDIVISION DEPTH	9
DECOMPOSITION TIME	588
SELECTION TIME	50

Table 4: Earth compression results

FIGURE	8(b)	8(d)	8(f)
COEFFICIENT THRESHOLD	0.02	0.002	0.0005
COEFFICIENTS SELECTED	741	15,101	138,321
TRIANGLES IN RECONSTRUCTION	1,968	39,408	328,058
FRACTION OF ORIGINAL	0.001	0.019	0.156
RECONSTRUCTION TIME	75	310	1,230
TOTAL PROCESSING TIME	1,315	1,754	2,904

3.2 Smoothly Blending between Approximations.

Sudden switching between models of different detail in an animation may produce a noticeable jump. This problem is easily mended by using a wavelet expansion, where the wavelet coefficients are treated as continuous functions of the viewing distance. With this simple technique, the object geometry can smoothly change its appearance as the viewing distance changes.

Consider two multiresolution approximations: $A(0)$ is shown at time t_0, and a subsequent finer-resolution approximation $A(1)$ is depicted at time t_1. The set of coefficients C present in $A(1)$ but not in $A(0)$ may be smoothly blended over the period of time between t_0 and t_1 using a linear interpolation of the new wavelet coefficients. More precisely, the approximation $A(t)$ presented at time t, $(t_0 < t \leq t_1)$ scales each coefficient c_i of C by

$$\frac{(t-t_0)}{(t_1-t_0)}.$$

The result is a smooth linear blend between $A(0)$ and $A(1)$. Note that a triangulation of the intermediate approximation $A(t)$ has exactly the connectivity of $A(1)$ along the entire blend; hence, there is a discontinuous jump in the connectivity of the underlying mesh used to represent the geometry as the time moves past t_0. However, the actual geometric effect of the transition is truly gradual, and can be shown to the viewer free of any noticeable jumps.

Using subdivision wavelets for smooth blending has proven successful in the production of frame-by-frame animations of complex models. In these animations, both color and geometry vary smoothly as the distance from the eye to the model varies.

References

[1] A. Arneodo, E. Bacry, and J.-F. Muzy. Solving the inverse fractal problem from wavelet analysis. *Europhysics letters*, 25(7):479–484, 1 March 1994.

[2] Deborah Berman, Jason Bartell, and David Salesin. Multiresolution painting and compositing. *Computer Graphics Annual Conference Series*, pages 85–90, July 1994.

[3] G. Beylkin, R. Coifman, and V. Rokhlin. Fast wavelet transforms and numerical algorithms. *Communications on Pure and Applied Mathematics*, 44:141–183, 1991.

[4] P. J. Burt and E. H. Adelson. The Laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, 31(4):532–540, April 1983.

[5] Per H. Christensen, Eric J. Stollnitz, David H. Salesin, and Tony D. DeRose. Wavelet radiance. In *Proceedings of the Fifth Eurographics Workshop on Rendering*, pages 287–302, Darmstadt, Germany, 1994.

[6] Charles K. Chui. *Wavelet Analysis and its Applications*, volume 1. Academic Press, Inc., Boston, 1992.

[7] Ingrid Daubechies. *Ten lectures on wavelets*. SIAM, Philadelphia, 1992.

[8] R. DeVore, B. Jawerth, and B. Lucier. Image compression through wavelet transform coding. *IEEE Transactions on Information Theory*, 38(2):719–746, March 1992.

[9] Nira Dyn, David Levin, and John Gregory. A butterfly subdivision scheme for surface interpolation with tension control. *ACM Transactions on Graphics*, 9(2):160–169, April 1990.

[10] Adam Finkelstein and David Salesin. Multiresolution curves. *Computer Graphics Annual Conference Series*, pages 261–268, July 1994.

[11] D. Forsey and R. Bartels. Hierarchical B-spline refinement. *Computer Graphics*, 22(4):205–212, 1988.

[12] Steven Gortler, Peter Schröder, Michael F. Cohen, and Pat Hanrahan. Wavelet radiosity. *Computer Graphics Annual Conference Series*, pages 221–230, August 1993.

[13] Rong-Qing Jia and Charles A. Micchelli. Using the refinement equations for the construction of pre-wavelets II: Powers of two. In P. J. Laurent, A. LeMéhauté, and L. L. Schumaker, editors, *Curves and Surfaces*, pages 209–246. Academic Press, 1991.

[14] Zicheng Liu, Steven J. Gortler, and Michael F. Cohen. Hierarchical spacetime control. *Computer Graphics Annual Conference Series*, pages 35–22, July 1994.

[15] M. Lounsbery. *Multiresolution analysis for surfaces of arbitrary topological type*. PhD thesis, Department of Computer Science and Engineering, Univ. of Washington, September 1994.

[16] Michael Lounsbery, Tony DeRose, and Joe Warren. Multiresolution analysis for surfaces of arbitrary topological type. *ACM Transactions on Graphics*, pages 34–73, January 1997.

[17] Stephane Mallat. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, July 1989.

[18] Stephane Mallat and Wen Liang Hwang. Singularity detection and processing with wavelets. *IEEE Transactions on Information Theory*, 38(2):617–643, March 1992.

[19] David Meyers. Multiresolution tiling. In *Proceedings of Graphics Interface*, pages 25–32, May 1994.

[20] David Meyers. *Reconstruction of Surfaces from Planar Contours*. PhD thesis, University of Washington, Seattle, Washington, July 1994.

[21] A. P. Pentland. Fast solutions to physical equilibrium and interpolation problems. *Visual Computer*, 8(5-6):303–314, June 1992.

[22] Olivier Rioul and Martin Vetterli. Wavelets and signal processing. *IEEE Signal Processing Magazine*, 8(4):14–38, October 1991.

[23] Peter Schröder, Steven J. Gortler, Michael F. Cohen, and Pat Hanrahan. Wavelet projections for radiosity. In *Proceedings of the Fourth Eurographics Workshop on Rendering*, pages 105–114, Paris, France, 1993.

[24] J. Stam. Fast evaluation of Catmull-Clark subdivision surfaces at arbitrary parameter values. *Submitted for publication*, 1998.

[25] Eric J. Stollnitz, Tony D. DeRose, and David H. Salesin. Wavelets for computer graphics: A primer. In preparation, 1994.

Chapter 8

A Variational Approach to Subdivision

Speaker: Leif Kobbelt

Variational Subdivision Schemes

Leif Kobbelt*

University of Erlangen–Nürnberg

Preface

The generic strategy of subdivision algorithms which is to define smooth curves and surfaces *algorithmically* by giving a set of simple rules for refining control polygons or meshes is a powerful technique to overcome many of the mathematical difficulties emerging from (polynomial) spline-based surface representations. In this section we highlight another application of the subdivision paradigm in the context of high quality surface generation.

From CAGD it is known that the technical and esthetic quality of a curve or a surface does not only depend on infinitesimal properties like the C^k differentiability. Much more important seems to be the *fairness* of a geometric object which is usually measured by curvature based energy functionals. A surface is hence considered optimal if it minimizes a given energy functional subject to auxiliary interpolation or approximation constraints.

Subdivision and fairing can be effectively combined into what is often refered to as *variational subdivision* or *discrete fairing*. The resulting algorithms inherit the simplicity and flexibility of subdivision schemes and the resulting curves and surfaces satisfy the sophisticated requirements for high end design in geometric modeling applications.

The basic idea that leads to variational subdivision schemes is that one subdivision step can be considered as a *topological split operation* where new vertices are introduced to increase the number of degrees of freedom, followed by a *smoothing operation* where the vertices are shifted in order to increase the overall smoothness. From this point of view is is natural to ask for the maximum smoothness that can be achieved on a given level of refinement while observing prescribed interpolation constraints.

We use an energy functional as a mathematical criterion to rate the smoothness of a polygon or a mesh. In the continuous setting, such scalar valued fairing functionals are typically defined as an integral over a combination of (squared) derivatives. In the discrete setting, we approximate such functionals by a sum over (squared) divided differences.

In the following we reproduce a few papers where this approach is described in more detail. In the univariate setting we consider interpolatory variational subdivision schemes which perform a greedy optimization in the sense that when computing the polygon $\mathbf{P}_{m+1}$ from $\mathbf{P}_m$ the new vertices' positions are determined by an energy minimization process but when proceeding with $\mathbf{P}_{m+2}$ the vertices of $\mathbf{P}_{m+1}$ are not adjusted.

In the bivariate setting, i.e., the subdivision and optimization of triangle meshes, we start with a given control mesh $\mathbf{P}_0$ whose vertices are to be interpolated by the resulting mesh. In this case it turns out that the mesh quality can be improved significantly if we use all the vertices from $\mathbf{P}_m \setminus \mathbf{P}_0$ for the optimization in the mth subdivision step.

Hence the algorithmic structure of variational subdivision degenerates to an alternating refinement and (constrained) global optimization. In fact, from a different viewing angle the resulting algorithms perform like a multi-grid solver for the discretized optimization problem. This observation provides the mathematical justification for the *discrete fairing approach*.

For the efficient fairing of continuous parameteric surfaces, the major difficulties arise from the fact that geometrically meaningful energy functionals depend on the control vertices in a highly non-linear fashion. As a consequence we have to either do non-linear optimization or we have to approximate the true functional by a linearized version. The reliability of this approximation usually depends on how close to isometric the surface's parameterization is. Alas, spline-patch-based surface representations often do not provide enough flexibility for an appropriate re-parameterization which would enable a feasible linearization of the geometric energy functional. Figure 1 shows two surfaces which are both optimal with respect to the same energy functional but for different parameterizations.

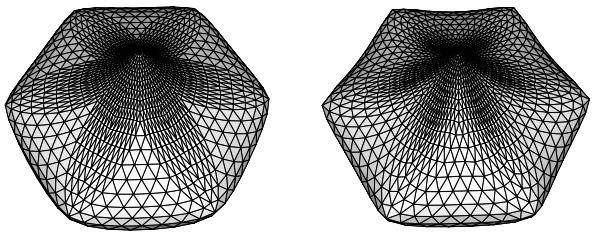


Figure 1: Optimal surfaces with respect to the same functional and interpolation constraints but for different parameterizations (isometric left, uniform right).

With the discrete fairing approach, we can exploit the auxiliary freedom to define an individual local parameterization for every vertex in the mesh. By this we find an isometric parameterization for each vertex and since the vertices are in fact the only points where the surface is evaluated, the linearized energy functional is a good approximation to the original one.

The discrete fairing machinery turns out to be a powerful tool which can facilitate the solution of many problems in the area of surface generation and modeling. The overall objective behind the presented applications will be the attempt to avoid, bypass, or at least delay the mathematically involved generation of spline CAD-models whenever it is appropriate.

*Computer Sciences Department (IMMD9), University of Erlangen-Nürnberg, Am Weichselgarten 9, 91058 Erlangen, Germany, Leif.Kobbelt@informatik.uni-erlangen.de

I Univariate Variational Subdivision

In this paper a new class of interpolatory refinement schemes is presented which in every refinement step determine the new points by solving an optimization problem. In general, these schemes are global, i.e., every new point depends on all points of the polygon to be refined. By choosing appropriate quadratic functionals to be minimized iteratively during refinement, very efficient schemes producing limiting curves of high smoothness can be defined. The well known class of stationary interpolatory refinement schemes turns out to be a special case of these variational schemes.

The original paper which also contains the omitted proofs has been published in:

L. Kobbelt
A Variational Approach to Subdivision,
CAGD 13 (1996) pp. 743–761, Elsevier

1.1 Introduction

Interpolatory refinement is a very intuitive concept for the construction of interpolating curves or surfaces. Given a set of points $\mathbf{p}_i^0 \in \mathbb{R}^d$ which are to be interpolated by a smooth curve, the first step of a refinement scheme consists in connecting the points by a piecewise linear curve and thus defining a polygon $\mathbf{P}_0 = (\mathbf{p}_0^0, \ldots, \mathbf{p}_{n-1}^0)$.

This initial polygon can be considered as a very coarse approximation to the final interpolating curve. The approximation can be improved by inserting new points between the old ones, i.e., by subdividing the edges of the given polygon. The positions of the new points $\mathbf{p}_{2i+1}^1$ have to be chosen appropriately such that the resulting (refined) polygon $\mathbf{P}_1 = (\mathbf{p}_0^1, \ldots, \mathbf{p}_{2n-1}^1)$ looks *smoother* than the given one in some sense (cf. Fig. 2). Interpolation of the given points is guaranteed since the old points $\mathbf{p}_i^0 = \mathbf{p}_{2i}^1$ still belong to the finer approximation.

By iteratively applying this interpolatory refinement operation, a sequence of polygons $(\mathbf{P}_m)$ is generated with vertices becoming more and more dense and which satisfy the interpolation condition $\mathbf{p}_i^m = \mathbf{p}_{2i}^{m+1}$ for all i and m. This sequence may converge to a smooth limit $\mathbf{P}_\infty$.

Many authors have proposed different schemes by explicitly giving particular rules how to compute the new points $\mathbf{p}_{2i+1}^{m+1}$ as a function of the polygon $\mathbf{P}_m$ to be refined. In (Dubuc, 1986) a simple refinement scheme is proposed which uses four neighboring vertices to compute the position of a new point. The position is determined in terms of the unique cubic polynomial which uniformly interpolates these four points. The limiting curves generated by this scheme are smooth, i.e., they are differentiable with respect to an equidistant parametrisation.

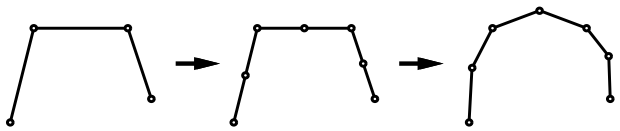


Figure 2: Interpolatory refinement

In (Dyn et al., 1987) this scheme is generalized by introducing an additional design or tension parameter. Replacing the interpolating cubic by interpolating polynomials of arbitrary degree leads to the *Lagrange-schemes* proposed in (Deslauriers & Dubuc, 1989). Raising the degree to $(2k+1)$, every new point depends on $(2k+2)$ old points of its vicinity. In (Kobbelt, 1995a) it is shown that at least for moderate k these schemes produce C^k-curves.

Appropriate formalisms have been developed in (Cavaretta et al., 1991), (Dyn & Levin, 1990), (Dyn, 1991) and elsewhere that allow an easy analysis of such *stationary schemes* which compute the new points by applying fixed banded convolution operators to the original polygon. In (Kobbelt, 1995b) simple criteria are given which can be applied to convolution schemes without any band limitation as well (cf. Theorem 2).

(Dyn et al., 1992) and (Le Méhauté & Utreras, 1994) propose non-linear refinement schemes which produce smooth interpolating (C^1-) curves and additionally preserve the convexity properties of the initial data. Both of them introduce constraints which locally define areas where the new points are restricted to lie in. Another possibility to define interpolatory refinement schemes is to dualize corner-cutting algorithms (Paluszny et al., 1994). This approach leads to more general necessary and sufficient convergence criteria.

In this paper we want to define interpolatory refinement schemes in a more systematic fashion. The major principle is the following: We are looking for refinement schemes for which, given a polygon $\mathbf{P}_m$, the refined polygon $\mathbf{P}_{m+1}$ is *as smooth as possible*. In order to be able to compare the "smoothness" of two polygons we define functionals $E(\mathbf{P}_{m+1})$ which measure the total amount of (discrete) strain energy of $\mathbf{P}_{m+1}$. The refinement operator then simply chooses the new points $\mathbf{p}_{2i+1}^{m+1}$ such that this functional becomes a minimum.

An important motivation for this approach is that in practice good approximations to the final interpolating curves should be achieved with little computational effort, i.e., maximum smoothness after a minimal number of refinement steps is wanted. In non-discrete curve design based, e.g., on splines, the concept of defining interpolating curves by the minimization of some energy functional (*fairing*) is very familiar (Meier & Nowacki, 1987), (Sapidis, 1994).

This basic idea of making a variational approach to the definition of refinement schemes can also be used for the definition of schemes which produce smooth surfaces by refining a given triangular or quarilateral net. However, due to the global dependence of the new points from the given net, the convergence analysis of such schemes strongly depends on the topology of the net to be refined and is still an open question. Numerical experiments with such schemes show that this approach is very promising. In this paper we will only address the analysis of univariate schemes.

1.2 Known results

Given an arbitrary (open/closed) polygon $\mathbf{P}_m = (\mathbf{p}_i^m)$, the *difference polygon* $\triangle^k \mathbf{P}_m$ denotes the polygon whose vertices are the vectors

$$\triangle^k \mathbf{p}_i^m := \sum_{j=0}^{k} \binom{k}{j} (-1)^{k+j} \mathbf{p}_{i+j}^m.$$

In (Kobbelt, 1995b) the following characterization of sequences of polygons $(\mathbf{P}_m)$ generated by the iterative application of an interpolatory refinement scheme is given:

Lemma 1 *Let $(\mathbf{P}_m)$ be a sequence of polygons. The scheme by which they are generated is an interpolatory refinement scheme (i.e., $\mathbf{p}_i^m = \mathbf{p}_{2i}^{m+1}$ for all i and m) if and only if for all $m, k \in \mathbb{N}$ the condition*

$$\triangle^k \mathbf{p}_i^m = \sum_{j=0}^{k} \binom{k}{j} \triangle^k \mathbf{p}_{2i+j}^{m+1}$$

holds for all indices i of the polygon $\triangle^k \mathbf{P}_m$.

Also in (Kobbelt, 1995b), the following sufficient convergence criterion is proven which we will use in the convergence analysis in the next sections.

Theorem 2 *Let $(\mathbf{P}_m)$ be a sequence of polygons generated by the iterative application of an arbitrary interpolatory refinement scheme. If*

$$\sum_{m=0}^{\infty} \|2^{km} \triangle^{k+l} \mathbf{P}_m\|_\infty < \infty$$

for some $l \in \mathbb{N}$ then the sequence $(\mathbf{P}_m)$ uniformly converges to a k-times continuously differentiable curve $\mathbf{P}_\infty$.

This theorem holds for all kinds of interpolatory schemes on open and closed polygons. However, in this paper we will only apply it to linear schemes whose support is global.

1.3 A variational approach to interpolatory refinement

In this and the next two sections we focus on the refinement of *closed* polygons, since this simplifies the description of the refinement schemes. Open polygons will be considered in Section 1.6.

Let $\mathbf{P}_m = (\mathbf{p}_0^m, \ldots, \mathbf{p}_{n-1}^m)$ be a given polygon. We want $\mathbf{P}_{m+1} = (\mathbf{p}_0^{m+1}, \ldots, \mathbf{p}_{2n-1}^{m+1})$ to be the smoothest polygon for which the interpolation condition $\mathbf{p}_{2i}^{m+1} = \mathbf{p}_i^m$ holds. Since the roughness at some vertex $\mathbf{p}_i^{m+1}$ is a local property we measure it by a an operator

$$K(\mathbf{p}_i^{m+1}) := \sum_{j=0}^{k} \alpha_j \mathbf{p}_{i+j-r}^{m+1}.$$

The coefficients α_j in this definition can be an arbitrary finite sequence of real numbers. The indices of the vertices $\mathbf{p}_i^{m+1}$ are taken *modulo* $2n$ according to the topological structure of the closed polygon $\mathbf{P}_{m+1}$. To achieve full generality we introduce the shift r such that $K(\mathbf{p}_i^{m+1})$ depends on $\mathbf{p}_{i-r}^{m+1}, \ldots, \mathbf{p}_{i+k-r}^{m+1}$. Every discrete measure of roughness K is associated with a characteristic polynomial

$$\alpha(z) = \sum_{j=0}^{k} \alpha_j z^j.$$

Our goal is to minimize the total strain energy over the whole polygon $\mathbf{P}_{m+1}$. Hence we define

$$E(\mathbf{P}_{m+1}) := \sum_{i=0}^{2n-1} K(\mathbf{p}_i^{m+1})^2 \tag{1}$$

to be the energy functional which should become minimal. Since the points $\mathbf{p}_{2i}^{m+1}$ of $\mathbf{P}_{m+1}$ with even indices are fixed due to the interpolation condition, the points $\mathbf{p}_{2i+1}^{m+1}$ with odd indices are the only free parameters of this optimization problem. The unique minimum of the quadratic functional is attained at the common root of all partial derivatives:

$$\begin{aligned} \frac{\partial}{\partial \mathbf{p}_{2l+1}^{m+1}} E(\mathbf{P}_{m+1}) &= \sum_{i=0}^{k} \frac{\partial}{\partial \mathbf{p}_{2l+1}^{m+1}} K(\mathbf{p}_{2l+1+r-i}^{m+1})^2 \\ &= 2 \sum_{i=0}^{k} \alpha_i \sum_{j=0}^{k} \alpha_j \mathbf{p}_{2l+1-i+j}^{m+1} \qquad (2) \\ &= 2 \sum_{i=-k}^{k} \beta_i \mathbf{p}_{2l+1+i}^{m+1} \end{aligned}$$

with the coefficients

$$\beta_{-i} = \beta_i = \sum_{j=0}^{k-i} \alpha_j \alpha_{j+i}, \qquad i = 0, \ldots, k. \tag{3}$$

Hence, the strain energy $E(\mathbf{P}_{m+1})$ becomes minimal if the new points $\mathbf{p}_{2i+1}^{m+1}$ are the solution of the linear system

$$\begin{pmatrix} \beta_0 & \beta_2 & \beta_4 & \cdots & \beta_2 \\ \beta_2 & \beta_0 & \beta_2 & \cdots & \beta_4 \\ \vdots & \vdots & \vdots & \ddots & \end{pmatrix} \begin{pmatrix} \mathbf{p}_1^{m+1} \\ \mathbf{p}_3^{m+1} \\ \vdots \\ \mathbf{p}_{2n-1}^{m+1} \end{pmatrix} = \begin{pmatrix} -\beta_1 & -\beta_1 & -\beta_3 & \cdots & -\beta_3 \\ -\beta_3 & -\beta_1 & -\beta_1 & \cdots & -\beta_5 \\ \vdots & \vdots & \vdots & \ddots & \end{pmatrix} \begin{pmatrix} \mathbf{p}_0^m \\ \mathbf{p}_1^m \\ \vdots \\ \mathbf{p}_{n-1}^m \end{pmatrix} \tag{4}$$

which follows from (2) by separation of the fixed points $\mathbf{p}_{2i}^{m+1} = \mathbf{p}_i^m$ from the variables. Here, both matrices are circulant and (almost) symmetric. A consequence of this symmetry is that the new points do not depend on the orientation by which the vertices are numbered (left to right or vice versa).

To emphasize the analogy between curve fairing and interpolatory refinement by variational methods, we call the equation

$$\sum_{i=-k}^{k} \beta_i \mathbf{p}_{2l+1+i}^{m+1} = 0, \qquad l = 0, \ldots, n-1 \tag{5}$$

the *Euler-Lagrange-equation.*

Theorem 3 *The minimization of $E(\mathbf{P}_{m+1})$ has a well-defined solution if and only if the characteristic polynomial $\alpha(z)$ for the local measure K has no diametric roots $z = \pm\omega$ on the unit circle with $\mathrm{Arg}(\omega) \in \pi\mathbb{N}/n$. (Proof to be found in the original paper)*

Remark The set $\pi\mathbb{N}/2^m$ becomes dense in $\mathbb{R}$ for increasing refinement depth $m \to \infty$. Since we are interested in the smoothness properties of the limiting curve $\mathbf{P}_\infty$, we should drop the restriction that the diametric roots have to have $\mathrm{Arg}(\omega) \in \pi\mathbb{N}/n$. For *stability* reasons we require $\alpha(z)$ to have no diametric roots on the unit circle at all.

The optimization by which the new points are determined is a geometric process. In order to obtain meaningful schemes, we have to introduce more restrictions on the energy functionals E or on the measures of roughness K.

For the expression $K^2(\mathbf{p}_i)$ to be valid, K has to be vector valued, i.e., the sum of the coefficients α_j has to be zero. This is equivalent to $\alpha(1) = 0$. Since

$$\sum_{i=-k}^{k} \beta_i = \sum_{i=0}^{k} \sum_{j=0}^{k} \alpha_i \alpha_j = \left(\sum_{j=0}^{k} \alpha_j \right)^2$$

the sum of the coefficients β_i also vanishes in this case and *affine invariance* of the (linear) scheme is guaranteed because constant functions are reproduced.

1.4 Implicit refinement schemes

In the last section we showed that the minimization of a quadratic energy functional (1) leads to the conditions (5) which determine the solution. Dropping the variational background, we can more generally prescribe arbitrary real coefficients $\beta_{-k}, \ldots, \beta_k$ (with

$\beta_{-i} = \beta_i$ to establish symmetry and $\sum \beta_i = 0$ for affine invariance) and define an interpolatory refinement scheme which chooses the new points $\mathbf{p}_{2i+1}^{m+1}$ of the refined polygon $\mathbf{P}_{m+1}$ such that the homogeneous constraints

$$\sum_{i=-k}^{k} \beta_i \mathbf{p}_{2l+1+i}^{m+1} = 0, \qquad l = 0, \ldots, n-1 \tag{6}$$

are satisfied. We call these schemes: *implicit refinement schemes* to emphasize the important difference to other refinement schemes where usually the new points are computed by one or two *explicitly* given rules (cf. the term *implicit curves* for curves represented by $f(x,y) = 0$). The stationary refinement schemes are a special case of the implicit schemes where $\beta_{2j} = \delta_{j,0}$. In general, the implicit schemes are non-stationary since the resulting weight coefficients by which the new points $\mathbf{p}_{2i+1}^{m+1}$ are computed depend on the number of vertices in $\mathbf{P}_m$.

In (Kobbelt, 1995b) a general technique is presented which allows to analyse the smoothness properties of the limiting curve generated by a given implicit refinement scheme.

The next theorem reveals that the class of implicit refinement schemes is not essentially larger than the class of variational schemes.

Theorem 4 *Let $\beta_{-k}, \ldots, \beta_k$ be an arbitrary symmetric set of real coefficients ($\beta_{-i} = \beta_i$). Then there always exists a (potentially complex valued) local roughness measure K such that (6) is the Euler-Lagrange-equation corresponding to the minimization of the energy functional (1). (Proof to be found in the original paper)*

Remark We do not consider implicit refinement schemes with complex coefficients β_i since then (6) in general has no real solutions.

Example To illustrate the statement of the last theorem we look at the 4-point scheme of (Dubuc, 1986). This is a stationary refinement scheme where the new points $\mathbf{p}_{2i+1}^{m+1}$ are computed by the rule

$$\mathbf{p}_{2i+1}^{m+1} = \frac{9}{16}(\mathbf{p}_i^m + \mathbf{p}_{i+1}^m) - \frac{1}{16}(\mathbf{p}_{i-1}^m + \mathbf{p}_{i+2}^m).$$

The scheme can be written in implicit form (6) with $k = 3$ and $\beta_{\pm 3} = 1$, $\beta_{\pm 2} = 0$, $\beta_{\pm 1} = -9$, $\beta_0 = 16$ since the common factor $\frac{1}{16}$ is not relevant. The roots of $\beta(z)$ are $z_1 = \ldots = z_4 = 1$ and $z_{5,6} = -2 \pm \sqrt{3}$. From the construction of the last proof we obtain

$$\alpha(z) = (2+\sqrt{3}) - (3+\sqrt{12})z + \sqrt{3}z^2 + z^3$$

as one possible solution. Hence, the quadratic strain energy which is minimized by the 4-point scheme is based on the local roughness estimate

$$K(\mathbf{p}_i) = (2+\sqrt{3})\mathbf{p}_i - (3+\sqrt{12})\mathbf{p}_{i+1} + \sqrt{3}\mathbf{p}_{i+2} + \mathbf{p}_{i+3}.$$

1.5 Minimization of differences

Theorem 2 asserts that a fast contraction rate of some higher differences is sufficient for the convergence of a sequence of polygons to a (k times) continuously differentiable limit curve. Thus it is natural to look for refinement schemes with a maximum contraction of differences. This obviously is an application of the variational approach. For the quadratic energy functional we make the ansatz

$$E_k(\mathbf{P}_{m+1}) := \sum_{i=0}^{2n-1} \|\triangle^k \mathbf{p}_i^{m+1}\|^2. \tag{7}$$

The partial derivatives take a very simple form in this case

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{p}_{2l+1}^{m+1}} E_k(\mathbf{P}_{m+1}) &= \sum_{i=0}^{k} \frac{\partial}{\partial \mathbf{p}_{2l+1}^{m+1}} \|\triangle^k \mathbf{p}_{2l+1-i}^{m+1}\|^2 \\
&= 2\sum_{i=0}^{k} (-1)^{k+i} \binom{k}{i} \triangle^k \mathbf{p}_{2l+1-i}^{m+1} \\
&= 2(-1)^k \triangle^{2k} \mathbf{p}_{2l+1-k}^{m+1}.
\end{aligned}$$

and the corresponding Euler-Lagrange-equation is

$$\triangle^{2k} \mathbf{p}_{2l+1-k}^{m+1} = 0, \qquad l = 0, \ldots, n-1 \tag{8}$$

where, again, the indices of the $\mathbf{p}_i^{m+1}$ are taken *modulo* $2n$. The characteristic polynomial of the underlying roughness measure K is $\alpha(z) = (z-1)^k$ and thus solvability and affine invariance of the refinement scheme are guaranteed. The solution of (8) only requires the inversion of a banded circulant matrix with bandwidth $2\lfloor \frac{k}{2} \rfloor + 1$.

Theorem 5 *The refinement scheme based on the minimization of E_k in (7) produces at least C^k-curves. (Proof to be found in the original paper)*

In order to prove even higher regularities of the limiting curve one has to combine more refinement steps. In (Kobbelt, 1995b) a simple technique is presented that allows to do the convergence analysis of such multi-step schemes numerically. Table 1 shows some results where r denotes the number of steps that have to be combined in order to obtain these differentiabilities.

In analogy to the non-discrete case where the minimization of the integral over the squared k-th derivative has piecewise polynomial C^{2k-2} solutions (B-splines), it is very likely that the limiting curves generated by iterative minimization of E_k are actually in C^{2k-2} too. The results given in Table 1 can be improved by combining more than r steps. For $k = 2, 3$, however, sufficiently many steps have already been combined to verify $\mathbf{P}_\infty \in C^{2k-2}$.

k	r	diff'ty	k	r	diff'ty
2	2	C^2	7	6	C^{10}
3	11	C^4	8	4	C^{11}
4	2	C^5	9	6	C^{13}
5	7	C^7	10	4	C^{14}
6	3	C^8	11	6	C^{16}

Table 1: Lower bounds on the differentiability of $\mathbf{P}_\infty$ generated by iterative minimization of $E_k(\mathbf{P}_m)$.

For illustration and to compare the quality of the curves generated by these schemes, some examples are given in Fig. 3. The curves result from applying different schemes to the initial data $\mathbf{P}_0 = (\ldots, 0, 1, 0, \ldots)$. We only show the middle part of one periodic interval of $\mathbf{P}_\infty$. As expected, the decay of the function becomes slower as the smoothness increases.

Remark Considering Theorem 2 it would be more appropriate to minimize the maximum difference $\|\triangle^k \mathbf{P}_m\|_\infty$ instead of $\|\triangle^k \mathbf{P}_m\|_2$. However, this leads to non-linear refinement schemes which are both, hard to compute and difficult to analyse. Moreover, in (Kobbelt, 1995a) it is shown that a contraction rate of

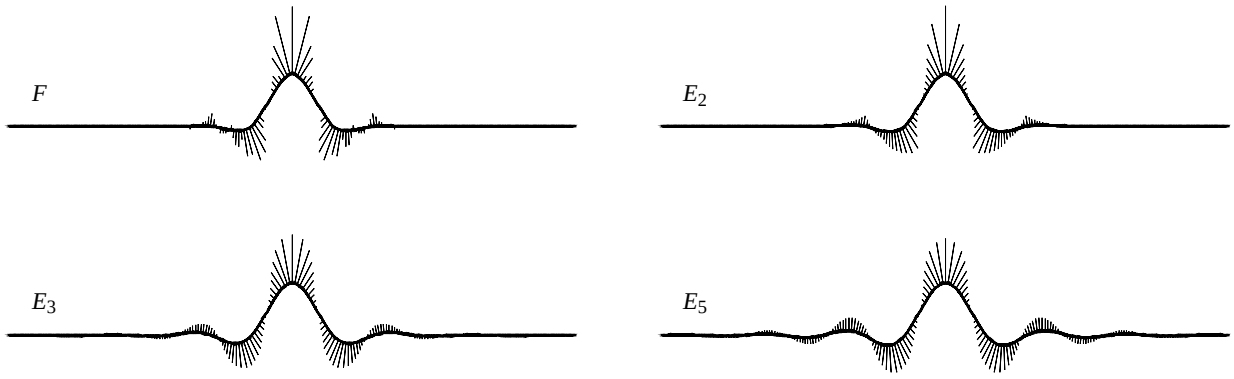


Figure 3: Discrete curvature plots of finite approximations to the curves generated by the four-point scheme F ($\mathbf{P}_\infty \in C^1$) and the iterative minimization of E_2 ($\mathbf{P}_\infty \in C^2$), E_3 ($\mathbf{P}_\infty \in C^4$) and E_5 ($\mathbf{P}_\infty \in C^7$).

$\|\triangle^{2k}\mathbf{P}_m\|_\infty = O(2^{-mk})$ implies $\|\triangle^k\mathbf{P}_m\|_\infty = O(2^{-m(k-\varepsilon)})$ for every $\varepsilon > 0$. It is further shown that $\|\triangle^k\mathbf{P}_m\|_\infty = O(2^{-mk})$ is the theoretical fastest contraction which can be achieved by interpolatory refinement schemes. Hence, the minimization of $\|\triangle^k\mathbf{P}_m\|_\infty$ cannot improve the asymptotic behavior of the contraction.

1.6 Interpolatory refinement of open polygons

The convergence analysis of variational schemes in the case of open finite polygons is much more difficult than it is in the case of closed polygons. The problems arise at both ends of the polygons $\mathbf{P}_m$ where the regular topological structure is disturbed. Therefore, we can no longer describe the refinement operation in terms of Toeplitz matrices but we have to use matrices which are Toeplitz matrices almost everywhere except for a finite number of rows, i.e., except for the first and the last few rows.

However, one can show that in a middle region of the polygon to be refined the smoothing properties of an implicit refinement scheme applied to an open polygon do not differ very much from the same scheme applied to a closed polygon. This is due to the fact that in both cases the influence of the old points $\mathbf{p}_i^m$ on a new point $\mathbf{p}_{2j+1}^{m+1}$ decrease exponentially with increasing topological distance $|i-j|$ for all asymptotically stable schemes (Kobbelt, 1995a).

For the refinement schemes which iteratively minimize forward differences, we can at least prove the following.

Theorem 6 *The interpolatory refinement of open polygons by iteratively minimizing the* $2k$*-th differences, generates at least* C^{k-1}*-curves. (Proof to be found in the original paper)*

The statement of this theorem only gives a lower bound for the differentiability of the limiting curve $\mathbf{P}_\infty$. However, the author conjects that the differentiabilities agree in the open and closed polygon case. For special cases we can prove better results.

Theorem 7 *The interpolatory refinement of open polygons by iteratively minimizing the second differences, generates at least* C^2*-curves. (Proof to be found in the original paper)*

1.7 Local refinement schemes

By now we only considered refinement schemes which are based on a *global* optimization problem. In order to construct local refinement schemes we can restrict the optimization to some local subpolygon. This means a new point $\mathbf{p}_{2l+1}^{m+1}$ is computed by minimizing some energy functional over a *window* $\mathbf{p}_{l-r}^m, \ldots, \mathbf{p}_{l+1+r}^m$. As the index l varies, the window is shifted in the same way.

Let E be a given quadratic energy functional. The solution of its minimization over the window $\mathbf{p}_{l-r}^m, \ldots, \mathbf{p}_{l+1+r}^m$ is computed by solving an Euler-Lagrange-equation

$$B\,(\mathbf{p}_{2l+1+2i}^{m+1})_{i=-r}^{r} = C\,(\mathbf{p}_{l+i}^m)_{i=-r}^{r+1}. \tag{9}$$

The matrix of $B^{-1}C$ can be computed explicitly and the weight coefficients by which a new point $\mathbf{p}_{2l+1}^{m+1}$ is computed, can be read off from the corresponding row in $B^{-1}C$. Since the coefficients depend on E and r only, this construction yields a stationary refinement scheme.

For such local schemes the convergence analysis is independent from the topological structure (open/closed) of the polygons to be refined. The formalisms of (Cavaretta et al., 1991), (Dyn & Levin, 1990) or (Kobbelt, 1995b) can be applied.

Minimizing the special energy functional $E_k(\mathbf{P})$ from (7) over open polygons allows the interesting observation that the resulting refinement scheme has polynomial precision of degree $k-1$. This is obvious since for points lying equidistantly parameterized on a polynomial curve of degree $k-1$, all k-th differences vanish and $E_k(\mathbf{P}) = 0$ clearly is the minimum of the quadratic functional.

Since the $2r+2$ points which form the subpolygon $\mathbf{p}_{l-r}^m, \ldots, \mathbf{p}_{l+1+r}^m$ uniquely define an interpolating polynomial of degree $2r+1$, it follows that the local schemes based on the minimization of $E_k(\mathbf{P})$ are identical for $k \geq 2r+2$. These schemes coincide with the Lagrange-schemes of (Deslauriers & Dubuc, 1989). Notice that $k \leq 4r+2$ is necessary because higher differences are not possible on the polygon $\mathbf{p}_{2(l-r)}^{m+1}, \ldots, \mathbf{p}_{2(l+1+r)}^{m+1}$ and minimizing $E_k(\mathbf{P}) \equiv 0$ makes no sense.

The local variational schemes provide a nice feature for practical purposes. One can use the refinement rules defined by the coefficients in the rows of $B^{-1}C$ in (9) to compute points which subdivide edges near the ends of open polygons. Pure stationary refinement schemes do not have this option and one therefore has to deal with *shrinking ends*. This means one only subdivides those edges which allow the application of the given subdivision mask and cuts off the remaining part of the unrefined polygon.

If $k \geq 2r+2$ then the use of these auxiliary rules causes the limiting curve to have a polynomial segment at both ends. This can be seen as follows. Let $\mathbf{P}_0 = (\mathbf{p}_0^0, \ldots, \mathbf{p}_n^0)$ be a given polygon and denote the polynomial of degree $2r+1 \leq k-1$ uniformly interpolating the points $\mathbf{p}_0^0, \ldots, \mathbf{p}_{2r+1}^0$ by $f(x)$.

The first vertex of the refined polygon $\mathbf{P}_1$ which not necessarily lies on $f(x)$ is $\mathbf{p}_{2r+3}^1$. Applying the same refinement scheme iteratively, we see that if $\mathbf{p}_{\delta_m}^m$ is the first vertex of $\mathbf{P}_m$ which does not lie

on $f(x)$ then $\mathbf{p}^{m+1}_{\delta_{m+1}} = \mathbf{p}^{m+1}_{2\delta_m - 2r-1}$ is the first vertex of $\mathbf{P}_{m+1}$ with this property. Let $\delta_0 = 2r+2$ and consider the sequence

$$\lim_{m\to\infty} \frac{\delta_m}{2^m} \;=\; (2r+2) - (2r+1) \lim_{m\to\infty} \sum_{i=1}^{m} 2^{-i} \;=\; 1.$$

Hence, the limiting curve $\mathbf{P}_\infty$ has a polynomial segment $f(x)$ between the points $\mathbf{p}^0_0$ and $\mathbf{p}^0_1$. An analog statement holds at the opposite end between $\mathbf{p}^0_{n-1}$ and $\mathbf{p}^0_n$.

This feature also arises naturally in the context of Lagrange-schemes where the new points near the ends of an open polygon can be chosen to lie on the first or last well-defined polynomial. It can be used to exactly compute the derivatives at the endpoints $\mathbf{p}^0_0$ and $\mathbf{p}^0_n$ of the limiting curve and it also provides the possibility to smoothly connect refinement curves and polynomial splines.

1.8 Computational Aspects

Since for the variational refinement schemes the computation of the new points $\mathbf{p}^{m+1}_{2i+1}$ involves the solution of a linear system, the algorithmic structure of these schemes is slightly more complicated than it is in the case of stationary refinement schemes. However, for the refinement of an open polygon $\mathbf{P}_m$ the computational complexity is still linear in the length of $\mathbf{P}_m$. The matrix of the system that has to be solved, is a banded Toeplitz-matrix with a small number of pertubations at the boundaries.

In the closed polygon case, the best we can do is to solve the circulant system in the Fourier domain. In particular, we transform the initial polygon $\mathbf{P}_0$ once and then perform m refinement steps in the Fourier domain where the convolution operator becomes a diagonal operator. The refined spectrum $\widehat{\mathbf{P}}_m$ is finally transformed back in order to obtain the result $\mathbf{P}_m$. The details can be found in (Kobbelt, 1995c). For this algorithm, the computational costs are dominated by the discrete Fourier transformation of $\widehat{\mathbf{P}}_m$ which can be done in $O(n\log(n)) = O(2^m m)$ steps. This is obvious since the number $n = 2^m n_0$ of points in the refined polygon $\mathbf{P}_m$ allows to apply m steps of the fast Fourier transform algorithm.

The costs for computing $\mathbf{P}_m$ are therefore $O(m)$ per point compared to $O(1)$ for stationary schemes. However, since in practice only a small number of refinement steps are computed, the constant factors which are hidden within these asymptotic estimates are relevant. Thus, the fact that implicit schemes need a smaller bandwidth than stationary schemes to obtain the same differentiability of the limiting curve (cf. Table 1) equalizes the performance of both.

In the implementation of these algorithms it turned out that all these computational costs are dominated by the 'administrative' overhead which is necessary, e.g., to build up the data structures. Hence, the differences in efficiency between stationary and implicit refinement schemes can be neglected.

References

[Cavaretta et al., 1991] Cavaretta, A. and Dahmen, W. and Micchelli, C. (1991), Stationary Subdivision, Memoirs of the AMS 93, 1–186

[Clegg, 1970] Clegg, J. (1970), *Variationsrechnung*, Teubner Verlag, Stuttgart

[Deslauriers & Dubuc, 1989] Deslauriers, G. and Dubuc, S. (1989), Symmetric iterative interpolation processes, Constructive Approximation 5, 49–68

[Dubuc, 1986] Dubuc, S. (1986), Interpolation through an iterative scheme, Jour. of Mathem. Anal. and Appl. 114, 185–204

[Dyn et al., 1987] Dyn, N. and Gregory, J. and Levin, D. (1987), A 4-point interpolatory subdivision scheme for curve design, CAGD 4, 257–268

[Dyn & Levin, 1990] Dyn, N. and Levin, D. (1990), Interpolating subdivision schemes for the generation of curves and surfaces, in: Haußmann W. and Jetter K. eds., *Multivariate Approximation and Interpolation*, Birkhäuser Verlag, Basel

[Dyn et al., 1992] Dyn, N. and Levin, D. and Liu, D. (1992), Interpolatory convexity-preserving subdivision schemes for curves and surfaces, CAD 24, 221–216

[Dyn, 1991] Dyn, N. (1991), Subdivision schemes in computer aided geometric design, in: Light, W. ed., *Advances in Numerical Analysis II, Wavelets, Subdivisions and Radial Functions*, Oxford University Press

[Golub & Van Loan, 1989] Golub, G. and Van Loan, C. (1989), *Matrix Computations*, John Hopkins University Press

[Kobbelt, 1995a] Kobbelt, L. (1995a), *Iterative Erzeugung glatter Interpolanten*, Universität Karlsruhe

[Kobbelt, 1995b] Kobbelt, L. (1995b), Using the Discrete Fourier-Transform to Analyze the Convergence of Subdivision Schemes, Appl. Comp. Harmonic Anal. 5 (1998), pp. 68–91

[Kobbelt, 1995c] Kobbelt, L. (1995c), Interpolatory Refinement is Low Pass Filtering, in Daehlen, M. and Lyche, T. and Schumaker, L. eds., Math. Meth in CAGD III

[Meier & Nowacki, 1987] Meier, H. and Nowacki, H. (1987), Interpolating curves with gradual changes in curvature, CAGD 4, 297–305

[Le Méhauté & Utreras, 1994] Le Méhauté A. and Utreras, F. (1994), Convexity-preserving interpolatory subdivision, CAGD 11, 17–37

[Paluszny et al., 1994] Paluszny M. and Prautzsch H. and Schäfer, M. (1994), Corner cutting and interpolatory refinement, Preprint

[Sapidis, 1994] Sapidis, N. (1994), *Designing Fair Curves and Surfaces*, SIAM, Philadelphia

[Widom, 1965] Widom, H. (1965), Toeplitz matrices, in: Hirschmann, I. ed., *Studies in Real and Complex Analysis*, MAA Studies in Mathematics 3

II Discrete Fairing

Many mathematical problems in geometric modeling are merely due to the difficulties of handling piecewise polynomial parameterizations of surfaces (e.g., smooth connection of patches, evaluation of geometric fairness measures). Dealing with polygonal meshes is mathematically much easier although infinitesimal smoothness can no longer be achieved. However, transferring the notion of fairness to the discrete setting of triangle meshes allows to develop very efficient algorithms for many specific tasks within the design process of high quality surfaces. The use of discrete meshes instead of continuous spline surfaces is tolerable in all applications where (on an intermediate stage) explicit parameterizations are not necessary. We explain the basic technique of discrete fairing *and give a survey of possible applications of this approach.*

The original paper has been published in:

L. Kobbelt
Variational Design with Parametric Meshes of Arbitrary Topology,
in Creating fair and shape preserving curves and surfaces, Teubner, 1998

2.1 Introduction

Piecewise polynomial spline surfaces have been the standard representation for free form surfaces in all areas of CAD/CAM over the last decades (and still are). However, although B-splines are optimal with respect to certain desirable properties (differentiability, approximation order, locality, ...), there are several tasks that cannot be performed easily when surface parameterizations are based on piecewise polynomials. Such tasks include the construction of globally smooth closed surfaces and the shape optimization by minimizing intrinsically geometric fairness functionals [5, 12].

Whenever it comes to involved numerical computations on free form surfaces — for instance in finite element analysis of shells — the geometry is usually sampled at discrete locations and converted into a piecewise linear approximation, i.e., into a polygonal mesh.

Between these two opposite poles, i.e., the *continuous* representation of geometric shapes by spline patches and the *discrete* representation by polygonal meshes, there is a compromise emerging from the theory of *subdivision surfaces* [9]. Those surfaces are defined by a *base mesh* roughly describing its shape, and a *refinement rule* that allows one to split the edges and faces in order to obtain a finer and smoother version of the mesh.

Subdivision schemes started as a generalization of *knot insertion* for uniform B-splines [11]. Consider a control mesh $[\mathbf{c}_{i,j}]$ and the knot vectors $[u_i] = [i\,h_u]$ and $[v_i] = [i\,h_v]$ defining a tensor product B-spline surface S. The same surface can be given with respect to the refined knot vectors $[\hat{u}_i] = [i\,h_u/2]$ and $[\hat{v}_i] = [i\,h_v/2]$ by computing the corresponding control vertices $[\hat{\mathbf{c}}_{i,j}]$, each $\hat{\mathbf{c}}_{i,j}$ being a simple linear combination of original vertices $\mathbf{c}_{i,j}$. It is well known that the iterative repetition of this process generates a sequence of meshes C_m which converges to the spline surface S itself.

The generic subdivision paradigm generalizes this concept by allowing arbitrary rules for the computation of the new control vertices $\hat{\mathbf{c}}_{i,j}$ from the given $\mathbf{c}_{i,j}$. The generalization also includes that we are no longer restricted to tensor product meshes but can use rules that are adapted to the different topological special cases in meshes with arbitrary connectivity. As a consequence, we can use any (manifold) mesh for the base mesh and generate smooth surfaces by iterative refinement.

The major challenge is to find appropriate rules that guarantee the convergence of the meshes C_m generated during the subdivision process to a smooth limit surface $S = C_\infty$. Besides the classical stationary schemes that exploit the piecewise regular structure of iteratively refined meshes [2, 4, 9], there are more complex geometric schemes [15, 8] that combine the subdivision paradigm with the concept of optimal design by energy minimization (*fairing*).

The technical and practical advantages provided by the representation of surfaces in the form of polygonal meshes stem from the fact that we do not have to worry about infinitesimal inter-patch smoothness and the refinement rules do not have to rely on the existence of a globally consistent parameterization of the surface. In contrast to this, spline based approaches have to introduce complicated non-linear geometric continuity conditions to achieve the flexibility to model closed surfaces of arbitrary shape. This is due to the topologically rather rigid structure of patches with triangular or quadrilateral parameter domain and fixed polynomial degree of cross boundary derivatives. The non-linearity of such conditions makes efficient optimization difficult if not practically impossible. On discrete meshes however, we can derive *local* interpolants according to local parameterizations (*charts*) which gives the freedom to adapt the parameterization individually to the local geometry and topology.

In the following we will shortly describe the concept of *discrete fairing* which is an efficient way to characterize and compute dense point sets on high quality surfaces that observe prescribed interpolation or approximation constraints. We then show how this approach can be exploited in several relevant fields within the area of free form surface modeling.

The overall objective behind all the applications will be the attempt to avoid, bypass, or at least delay the mathematically involved generation of spline CAD-models whenever it is appropriate. Especially in the early design stages it is usually not necessary to have an explicit parameterization of a surface. The focus on polygonal mesh representations might help to free the creative designer from being confined by mathematical restrictions. In later stages the conversion into a spline model can be based on more reliable information about the intended shape. Moreover, since technical engineers are used to performing numerical simulations on polygonal approximations of the true model anyway, we also might find short-cuts that allow to speed up the turn-around cycles in the design process, e.g., we could alter the shape of a mechanical part by modifying the FE-mesh directly without converting back and forth between different CAD-models.

2.2 Fairing triangular meshes

The observation that in many applications the global fairness of a surface is much more important than infinitesimal smoothness motivates the *discrete fairing* approach [10]. Instead of requiring G^1 or G^2 continuity, we simply approximate a surface by a plain triangular C^0– mesh. On such a mesh we can think of the (discrete) curvature being located at the vertices. The term *fairing* in this context means to minimize these local contributions to the total (discrete) curvature and to equalize their distribution across the mesh.

We approximate local curvatures at every vertex $\mathbf{p}$ by divided differences with respect to a locally isometric parameterization $\mu_{\mathbf{p}}$. This parameterization can be found by estimating a tangent plane $T_{\mathbf{p}}$ (or the normal vector $\mathbf{n}_{\mathbf{p}}$) at $\mathbf{p}$ and projecting the neighboring vertices $\mathbf{p}_i$ into that plane. The projected points yield the parameter values (u_i, v_i) if represented with respect to an orthonormal basis $\{\mathbf{e}_u, \mathbf{e}_v\}$ spanning the tangent plane

$$\mathbf{p}_i - \mathbf{p} \;=\; u_i\,\mathbf{e}_u + v_i\,\mathbf{e}_v + d_i\,\mathbf{n}_{\mathbf{p}}.$$

Another possibility is to assign parameter values according to the lengths and the angles between adjacent edges (*discrete exponential*

map) [15, 10].

To obtain reliable curvature information at $\mathbf{p}$, i.e., second order partial derivatives with respect to the locally isometric parameterization $\mu_{\mathbf{p}}$, we solve the normal equation of the Vandermonde system

$$V^T V \left[\tfrac{1}{2} f_{uu},\, f_{uv},\, \tfrac{1}{2} f_{vv} \right]^T = V^T [d_i]_i$$

with $V = [u_i^2, u_i v_i, v_i^2]_i$ by which we get the best approximating quadratic polynomial in the least squares sense. The rows of the inverse matrix $(V^T V)^{-1} V^T =: [\alpha_{i,j}]$ by which the Taylor coefficients f_* of this polynomial are computed from the data $[d_i]_i$, contain the coefficients of the corresponding divided difference operators Γ_*.

Computing a weighted sum of the squared divided differences is equivalent to the discrete sampling of the corresponding continuous fairness functional. Consider for example

$$\int_S \kappa_1^2 + \kappa_2^2 \, dS$$

which is approximated by

$$\sum_{\mathbf{p}_i} \omega_i \left(\|\Gamma_{uu}(\mathbf{p}_j - \mathbf{p}_i)\|^2 + 2\,\|\Gamma_{uv}(\mathbf{p}_j - \mathbf{p}_i)\|^2 + \|\Gamma_{vv}(\mathbf{p}_j - \mathbf{p}_i)\|^2 \right). \tag{10}$$

Notice that the value of (10) is independent of the particular choices $\{\mathbf{e}_u, \mathbf{e}_v\}$ for each vertex due to the rotational invariance of the functional. The discrete fairing approach can be understood as a generalization of the traditional finite difference method to parametric meshes where divided difference operators are defined with respect to locally varying parameterizations. In order to make the weighted sum (10) of local curvature values a valid quadrature formula, the weights ω_i have to reflect the local area element which can be approximated by observing the relative sizes of the parameter triangles in the local charts $\mu_{\mathbf{p}} : \mathbf{p}_i - \mathbf{p} \mapsto (u_i, v_i)$.

Since the objective functional (10) is made up of a sum over squared local linear combinations of vertices (in fact, of vertices being direct neighbors of one central vertex), the minimum is characterized by the solution of a global but sparse linear system. The rows of this system are the partial derivatives of (10) with respect to the movable vertices $\mathbf{p}_i$. Efficient algorithms are known for the solution of such systems [6].

2.3 Applications to free form surface design

When generating fair surfaces from scratch we usually prescribe a set of interpolation and approximation constraints and fix the remaining degrees of freedom by minimizing an energy functional. In the context of discrete fairing the constraints are given by an initial triangular mesh whose vertices are to be approximated by a fair surface being topologically equivalent. The necessary degrees of freedom for the optimization are obtained by uniformly subdividing the mesh and thus introducing new *movable* vertices.

The discrete fairing algorithm requires the definition of a local parameterization $\mu_{\mathbf{p}}$ for each vertex $\mathbf{p}$ including the newly inserted ones. However, projection into an estimated tangent plane does not work here, because the final positions of the new vertices are obviously not known a priori. In [10] it has been pointed out that in order to ensure solvability and stability of the resulting linear system, it is appropriate to define the local parameterizations (local metrics) for the new vertices by *blending* the metrics of nearby vertices from the original mesh. Hence, we only have to estimate the local charts covering the original vertices to set-up the linear system which characterizes the optimal surface. This can be done prior to actually computing a solution and we omit an additional optimization loop over the parameterization.

When solving the sparse linear system by iterative methods we observe rather slow convergence. This is due to the low-pass filter characteristics of the iteration steps in a Gauß-Seidel or Jacobi scheme. However since the mesh on which the optimization is performed came out of a uniform refinement of the given mesh (*subdivision connectivity*) we can easily find nested grids which allow the application of highly efficient multi-grid schemes [6].

Moreover, in our special situation we can generate sufficiently smooth starting configurations by midpoint insertion which allows us to neglect the pre-smoothing phase and to reduce the V-cycle of the multi-grid scheme to the alternation of binary subdivision and iterative smoothing. The resulting algorithm has linear complexity in the number of generated triangles.

The advantage of this discrete approach compared to the classical fair surface generation based on spline surfaces is that we do not have to approximate a geometric functional that uses true curvatures by one which replaces those by second order partial derivatives with respect to the fixed parameterization of the patches. Since we can use a custom tailored parameterization for each point evaluation of the second order derivatives, we can choose this parameterization to be isometric — giving us access to the true geometric functional.

Figure 4 shows an example of a surface generated this way. The implementation can be done very efficiently. The shown surface consists of about 50K triangles and has been generated on a SGI R10000 (195MHz) within 10 seconds. The scheme is capable of generating an arbitrarily dense set of points on the surface of minimal energy. It is worth to point out that the scheme works completely automatic: no manual adaption of any parameters is necessary, yet the scheme produces good surfaces for a wide range of input data.

2.4 Applications to interactive modeling

For subdivision schemes we can use any triangular mesh as a control mesh roughly describing the shape of an object to be modeled. The flexibility of the schemes with respect to the connectivity of the underlying mesh allows very intuitive modifications of the mesh. The designer can move the control vertices just like for Bezier-patches but she is no longer tied to the common restrictions on the connectivity which is merely a consequence of the use of tensor product spline bases.

When modeling an object by Bezier-patches, the control vertices are the handles to influence the shape and the de Casteljau algorithm associates the control mesh with a smooth surface patch. In our more general setting, the designer can work on an *arbitrary* triangle mesh and the connection to a smooth surface is provided by the discrete fairing algorithm. The advantages are that control vertices are interpolated which is a more intuitive interaction metaphor and the topology of the control structure can adapt to the shape of the object.

Figure 5 shows the model of a mannequin head. A rather coarse triangular mesh allows already to define the global shape of the head (left). If we add more control vertices in the areas where more detail is needed, i.e., around the eyes, the mouth and the ears, we can construct the complex surface at the far right. Notice how the discrete fairing scheme does not generate any artifacts in regions where the level of detail changes.

2.5 Applications to mesh smoothing

In the last sections we saw how the discrete fairing approach can be used to generate fair surfaces that interpolate the vertices of a given triangular mesh. A related problem is to smooth out high frequency noise from a given *detailed* mesh without further refinement. Consider a triangulated surface emerging for example from 3D laser scanning or iso-surface extraction out of CT volume data. Due to measurement errors, those surfaces usually show oscillations that do not stem from the original geometry.

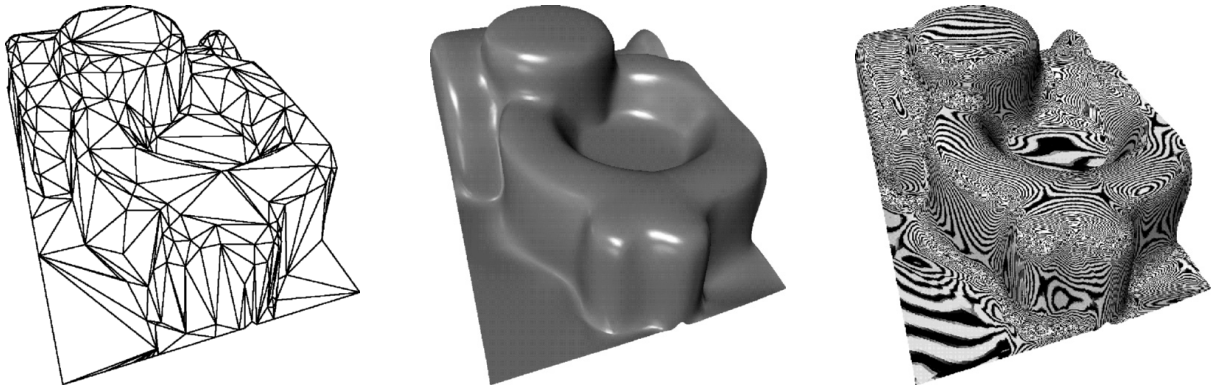


Figure 4: A fair surface generated by the discrete fairing scheme. The flexibility of the algorithm allows to interpolate rather complex data by high quality surfaces. The process is completely automatic and it took about 10 sec to compute the refined mesh with 50K triangles. On the right you see the reflection lines on the final surface.

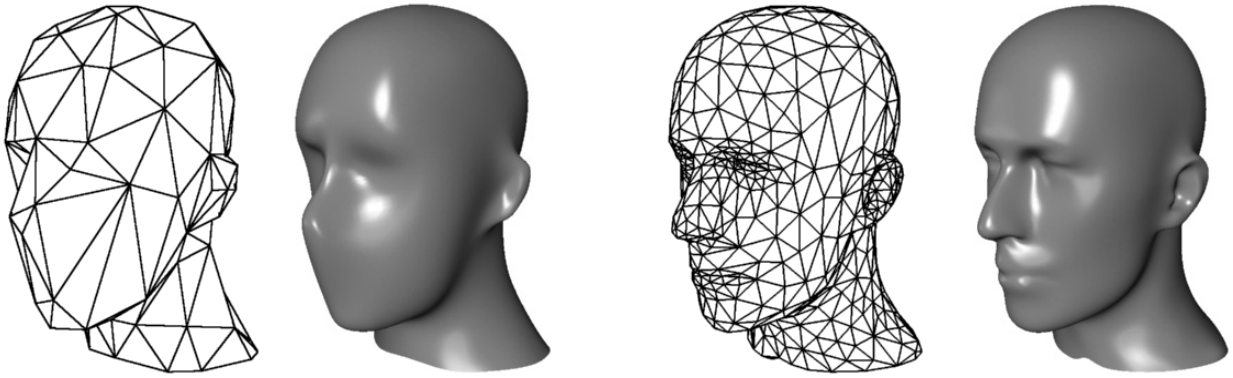


Figure 5: Control meshes with arbitrary connectivity allow to adapt the control structure to the geometry of the model. Notice that the influence of one control vertex in a tensor product mesh is always rectangular which makes it difficult to model shapes with non-rectangular features.

Constructing the above mentioned local parameterizations, we are able to quantify the noise by evaluating the local curvature. Shifting the vertices while observing a maximum tolerance can reduce the total curvature and hence smooth out the surface. From a signal processing point of view, we can interpret the iterative solving steps for the global sparse system as the application of recursive digital low-pass filters [13]. Hence it is obvious that the process will reduce the high frequency noise while maintaining the low frequency shape of the object.

Figure 6 shows an iso-surface extracted from a CT scan of an engine block. The noise is due to inexact measurement and instabilities in the extraction algorithm. The smoothed surface remains within a tolerance which is of the same order of magnitude as the diagonal of one voxel in the CT data.

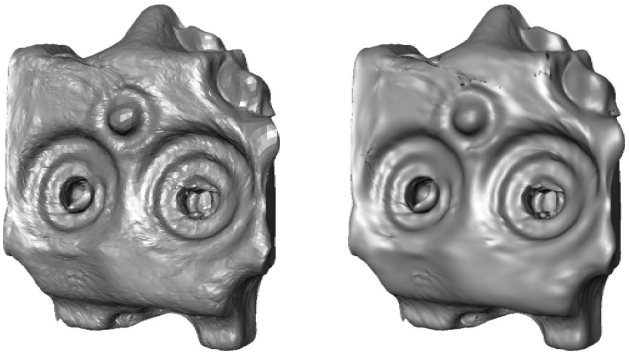


Figure 6: An iso-surface extracted from a CT scan of an engine block. On the left, one can clearly see the noise artifacts due to measurement and rounding errors. The right object was smoothed by minimizing the discrete fairing energy. Constraints on the positional delocation were imposed.

2.6 Applications to surface interrogation

Deriving curvature information on a discrete mesh is not only useful for fair interpolation or post-processing of measured data. It can also be used to visualize artifacts on a surface by plotting the color coded discrete curvature directly on the mesh. Given for example the output of the numerical simulation of a physical process: since deformation has occurred during the simulation, this output typically consists merely of a discrete mesh and no continuous surface description is available.

Using classical techniques from differential geometry would require to fit an interpolating spline surface to the data and then visualize the surface quality by curvature plots. The availability of

samples of second order partial derivatives with respect to locally isometric parameterizations at every vertex enables us to show this information directly without the need for a continuous surface.

Figure 7 shows a mesh which came out of the FE-simulation of a loaded cylindrical shell. The shell is clamped at the boundaries and pushed down by a force in normal direction at the center. The deformation induced by this load is rather small and cannot be detected by looking, e.g., at the reflection lines. The discrete mean curvature plot however clearly reveals the deformation. Notice that histogram equalization has been used to optimize the color contrast of the plot.

2.7 Applications to hole filling and blending

Another area where the discrete fairing approach can help is the filling of undefined regions in a CAD model or in a measured data set. Of course, all these problems can be solved by fairing schemes based on spline surfaces as well. However, the discrete fairing approach allows one to split the overall (quite involved) task into simple steps: we always start by constructing a triangle mesh defining the global topology. This is easy because no G^1 or higher boundary conditions have to be satisfied. Then we can apply the discrete fairing algorithm to generate a sufficiently dense point set on the objective surface. This part includes the refinement and energy minimization but it is almost completely automatic and does not have to be adapted to the particular application. In a last step we fit polynomial patches to the refined data. Here we can restrict ourselves to pure fitting since the fairing part has already been taken care of during the generation of the dense data. In other words, the discrete fairing has recovered enough information about an optimal surface such that staying as close as possible to the generated points (in a least squares sense) is expected to lead to high quality surfaces. To demonstrate this methodology we give two simple examples.

First, consider the point data in Figure 8. The very sparsely scattered points in the middle region make the task of interpolation rather difficult since the least squares matrix for a locally supported B-spline basis might become singular. To avoid this, fairing terms would have to be included into the objective functional. This however brings back all the problems mentioned earlier concerning the possibly poor quality of parameter dependent energy functionals and the prohibitive complexity of non-linear optimization.

Alternatively, we can connect the points to build a spatial triangulation. Uniform subdivision plus discrete fairing recovers the missing information under the assumption that the original surface was sufficiently fair. The un-equal distribution of the measured data points and the strong distortion in the initial triangulation do not cause severe instabilities since we can define individual parameterizations for every vertex. These allow one to take the local geometry into account.

Another standard problem in CAD is the *blending* or *filleting* between surfaces. Consider the simple configuration in Figure 9 where several plane faces (dark grey) are to be connected smoothly. We first close the gap by a simple coarse triangular mesh. Such a mesh can easily be constructed for any reasonable configuration with much less effort than constructing a piecewise polynomial representation. The boundary of this initial mesh is obtained by sampling the surfaces to be joined.

We then refine the mesh and, again, apply the discrete fairing machinery. The smoothness of the connection to the predefined parts of the geometry is guaranteed by letting the blend surface mesh overlap with the given faces by one row of triangles (all necessary information is obtained by sampling the given surfaces). The vertices of the triangles belonging to the original geometry are not allowed to move but since they participate in the global fairness functional they enforce a smooth connection. In fact this technique allows to define Hermite-type boundary conditions.

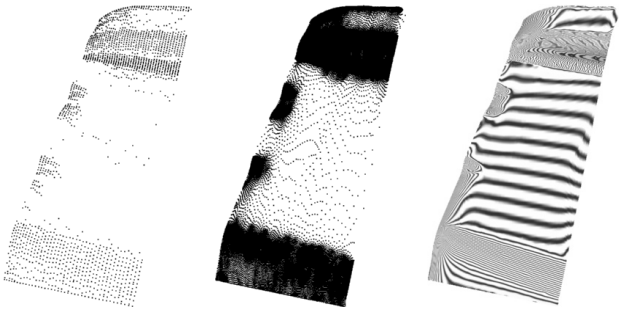


Figure 8: The original data on the left is very sparse in the middle region of the object. Triangulating the points in space and discretely fairing the iteratively refined mesh recovers more information which makes least squares approximation much easier. On the right, reflection lines on the resulting surface are shown.

2.8 Conclusion

In this paper we gave a survey of currently implemented applications of the discrete fairing algorithm. This general technique can be used in all areas of CAD/CAM where an approximation of the actual surface by a reasonably fine triangular mesh is a sufficient representation. If compatibility to standard CAD formats matters, a spline fitting post-process can always conclude the discrete surface generation or modification. This fitting step can rely on more information about the intended shape than was available in the original setting since a *dense* set of points has been generated.

As we showed in the previous sections, mesh smoothing and hole filling can be done on the discrete structure *before* switching to a continuous representation. Hence, the bottom line of this approach is to do most of the work in the discrete setting such that the mathematically more involved algorithms to generate piecewise polynomial surfaces can be applied to enhanced input data with most common artifacts removed.

We do not claim that splines could ever be completely replaced by polygonal meshes but in our opinion we can save a considerable amount of effort if we use spline models only where it is really necessary and stick to meshes whenever it is possible. There seems to be a huge potential of applications where meshes do the job if we find efficient algorithms.

The major key to cope with the genuine complexity of highly detailed triangle meshes is the introduction of a hierarchical structure. Hierarchies could emerge from classical multi-resolution techniques like subdivision schemes but could also be a by-product of mesh simplification algorithms.

An interesting issue for future research is to find efficient and numerically stable methods to enforce convexity preservation in the fairing scheme. At least local convexity can easily be maintained by introducing non-linear constraints at the vertices.

Prospective work also has to address the investigation of explicit and reliable techniques to exploit the discrete curvature information for the detection of feature lines in the geometry in order to split a given mesh into geometrically coherent segments. Further, we can try to identify regions of a mesh where the value of the curvature is approximately constant — those regions correspond to special geometries like spheres, cylinders or planes. This will be the topic of a forthcoming paper.

References

[1] E. Catmull, J. Clark, *Recursively generated B-spline surfaces on arbitrary topological meshes*, CAD 10 (1978), pp. 350–355

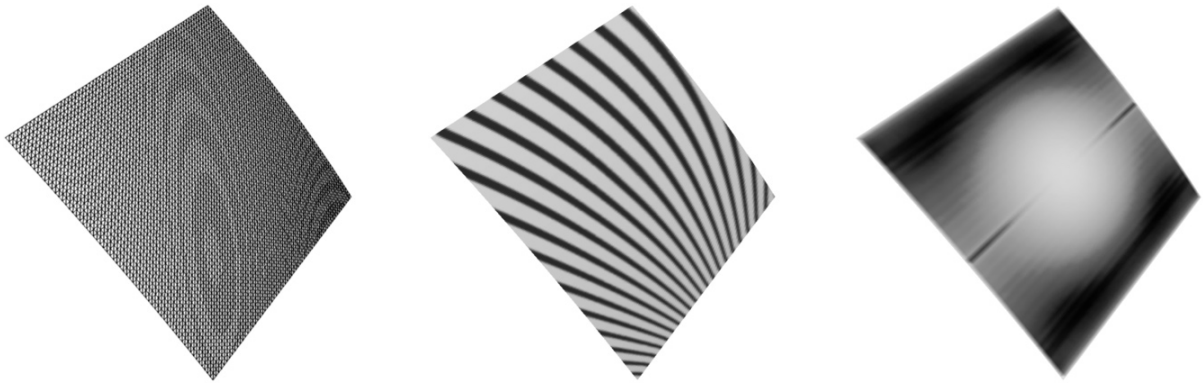


Figure 7: Visualizing the discrete curvature on a finite element mesh allows to detect artifacts without interpolating the data by a continuous surface.

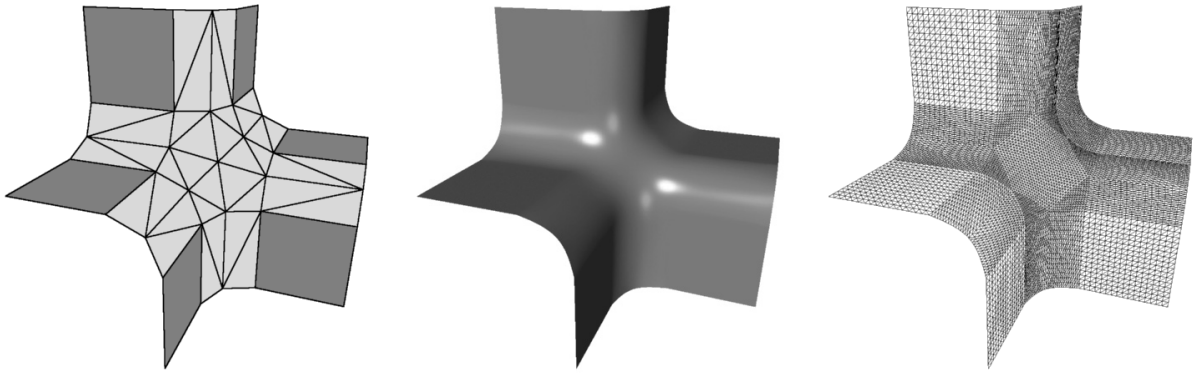


Figure 9: Creating a "monkey saddle" blend surface to join six planes. Any blend surface can be generated by closing the gap with a triangular mesh first and then applying discrete fairing.

[2] Celniker G. and D. Gossard, *Deformable curve and surface finite elements for free-form shape design*, ACM Computer Graphics **25** (1991), 257–265.

[3] D. Doo and M. Sabin, *Behaviour of Recursive Division Surfaces Near Extraordinary Points*, CAD 10 (1978), pp. 356–360

[4] N. Dyn, *Subdivision Schemes in Computer Aided Geometric Design*, Adv. Num. Anal. II, Wavelets, Subdivisions and Radial Functions, W.A. Light ed., Oxford Univ. Press, 1991, pp. 36–104.

[5] Greiner G., *Variational design and fairing of spline surfaces*, Computer Graphics Forum **13** (1994), 143–154.

[6] Hackbusch W., *Multi-Grid Methods and Applications*, Springer Verlag 1985, Berlin.

[7] Hagen H. and G. Schulze, *Automatic smoothing with geometric surface patches*, CAGD **4** (1987), 231–235.

[8] Kobbelt L., *A variational approach to subdivision*, CAGD **13** (1996), 743–761.

[9] Kobbelt L., *Interpolatory subdivision on open quadrilateral nets with arbitrary topology*, Comp. Graph. Forum **15** (1996), 409–420.

[10] Kobbelt L., *Discrete fairing*, Proceedings of the Seventh IMA Conference on the Mathematics of Surfaces, 1997, pp. 101–131.

[11] J. Lane and R. Riesenfeld, *A Theoretical Development for the Computer Generation and Display of Piecewise Polynomial Surfaces*, IEEE Trans. on Pattern Anal. and Mach. Int., 2 (1980), pp. 35–46

[12] Moreton H. and C. Séquin, *Functional optimization for fair surface design*, ACM Computer Graphics **26** (1992), 167–176.

[13] Taubin G., *A signal processing approach to fair surface design*, ACM Computer Graphics **29** (1995), 351–358

[14] Welch W. and A. Witkin, *Variational surface modeling*, ACM Computer Graphics **26** (1992), 157–166

[15] Welch W. and A. Witkin, *Free-form shape design using triangulated surfaces*, ACM Computer Graphics **28** (1994), 247–256

Chapter 9

Exploiting Subdivision in Modeling and Animation

Speaker: David Forsey

Some material in this section is co-authored by Cristin Barghiel, Richard Bartels, and David Forsey.

Pasting Spline Surfaces

Cristin Barghiel, Richard Bartels and David Forsey

Abstract. Details can be added to spline surfaces by applying displacement maps. However, the computational intensity of displacement mapping prevents its use for interactive design. In this paper we explore a form of simulated displacement that can be used for interactive design by providing a preview of true displacements at low computational cost.

§Introduction

Displacement mapping is a standard and useful tool for realistic rendering [3,8], and it has been suggested for application to surface approximation [6]. Because of its computational cost, however, it has not yet been fully explored as a design tool. In [4] a restricted form of displacement mapping was investigated for use in interactive design. Composite surfaces were constructed by layering a sequence of *detail surfaces* onto a *base surface*. Each detail was formulated as a vector displacement field and mapped onto a region of the base. The process was restricted in that, in order to achieve continuity between detail and base, it was necessary that the detail share a common parametric domain with the base and possess a knot structure derivable from the base by refinement, a setting studied recently by Weller and Hagen [9]. These restrictions permitted the computations for displacement mapping to be simplified to such an extent that interactive design was feasible. Here we extend the techniques of [4] to remove some of those restrictions.

True displacement mapping involves a vector-valued function $\mathbf{d}(r,s)$ that is nonzero over a compact domain $(r,s) \in \mathcal{D}$, a 1-1 transformation $\mathbf{T}$ of that domain into the domain $(u,v) \in \mathcal{S}$ of a surface (point-valued) $\mathrm{S}_0(u,v)$, and the resulting composition $\mathrm{S}_1(u,v) = \mathrm{S}_0(u,v) + \mathbf{d}(\mathbf{T}^{-1}(u,v))$. (Bold, lower case letters denote vectors; bold upper case letters are transformations; Roman and Greek lower case letters are scalars; Roman upper case letters are points, and calligraphic letters are domains.) The continuity of S_1 is controlled by the continuity of S_0, $\mathbf{T}$ and $\mathbf{d}$. In the surface design setting that we wish to consider, the intent of employing a displacement map is to achieve the "pasting" of a detail surface $\mathrm{D}(r,s)$ onto a base surface $\mathrm{S}_0(u,v)$. The function $\mathbf{d}$ is

Mathematical Methods in CAGD III
M. Dæhlen, T. Lyche, and L. L. Schumaker (eds.), pp. 1–3.

ISBN xxx.

constructed from the difference between D and a *reference surface*. For convenience, the domain of D may be embedded into the range of D; that is, the points on D are identified with points in the space $\{(r, s, t)\}$, and the domain of D is identified with the plane $(r, s, 0)$ by some convenient homeomorphism. This plane is often taken as the reference surface: $\mathbf{d}(r, s) = \mathrm{D}(r, s) - (r, s, 0)$. The composition yielding S_1 must be implemented with respect to a coordinate frame $\{(r, s, 0), \mathbf{i}, \mathbf{j}, \mathbf{k}\}$ appropriate to $\mathbf{d}$, D and some manifold coordinate frame $\{\mathrm{S}_0(u, v), \mathbf{l}(u, v), \mathbf{m}(u, v), \mathbf{n}(u, v)\}$ defined over S_0 (with $\mathbf{n}(u, v)$ taken as normal to $\mathrm{S}(u, v)$). To achieve this, the mapping $\mathbf{T}$ should not only map (u, v) smoothly into (r, s) but it should also provide a smooth mapping from $\{(r, s, 0), \mathbf{i}, \mathbf{j}, \mathbf{k}\}$ to $\{\mathrm{S}_0(u, v), \mathbf{l}(u, v), \mathbf{m}(u, v), \mathbf{n}(u, v)\}$. The elements of this setting are shown in Figure 1, and except for the third dimension, this is exactly the setting for texture mapping [7].

$$\mathbf{d} = \alpha\mathbf{i} + \beta\mathbf{j} + \gamma\mathbf{k} \qquad \mathrm{S}_1 = \mathrm{S}_0 + \alpha\mathbf{l} + \beta\mathbf{m} + \gamma\mathbf{n}$$
$$\mathrm{D} = (r, s, 0) + \mathbf{d}$$

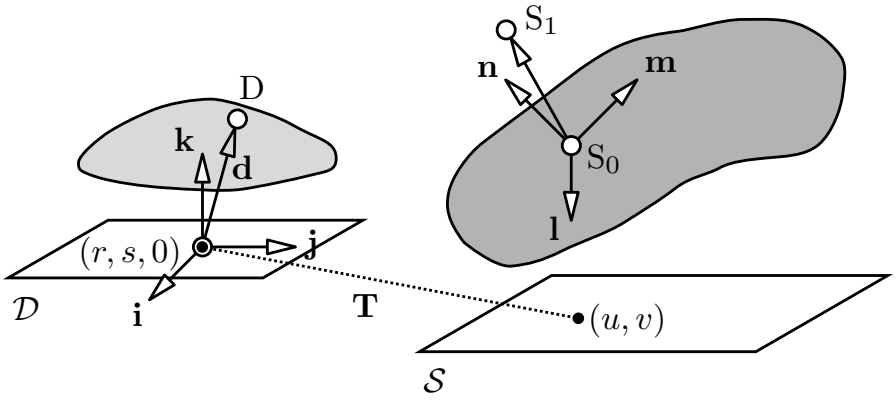


Fig. 1. Displacement mapping.

§Simulated Displacement

Since the process just outlined must be carried out at every surface point to be considered, a large amount of computation is involved. In high-quality image rendering, texture and displacement mappings are supportable because they are comparable to other computationally intensive aspects of the rendering process (e.g. ray tracing) and because we do not expect to produce high-quality images in real time. For interactive design, however, costs must be cut. We shall do this by limiting the mapping process to a selected, small number of sites, and choose the sites and the geometric entities to be mapped in a way that provides us with a reasonable preview of what a true displacement mapping will finally produce. The simulation must support such a preview in a dynamic as well as a static sense. That is; if the detail surface D is modified in its interior, that modification should be possible either before or after the pasting, and if the base surface S is modified after pasting, the pasted detail

should conform to the change. Figure 2 illustrates this in a schematic way, where the detail is in grey, the base is in black, and the control points of the detail are shown as circles.

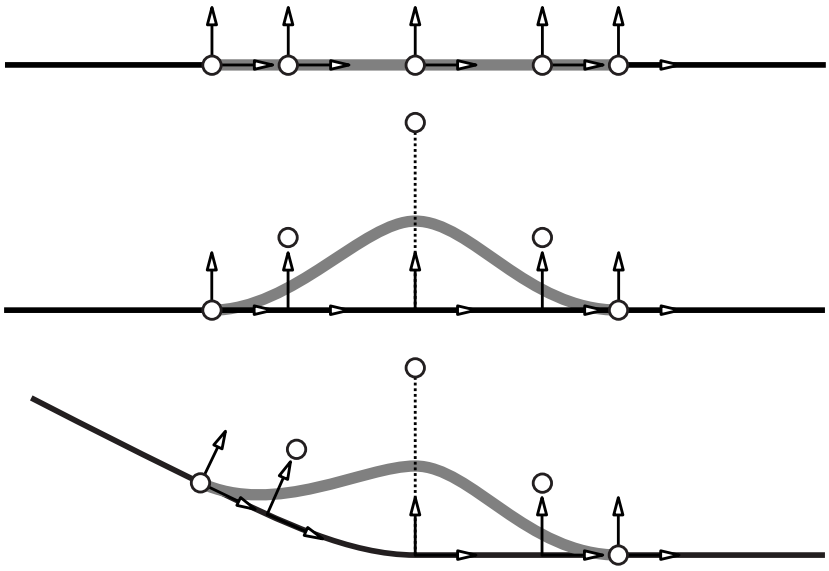


Fig. 2. Schema for displacement simulation.

Figure 2 also reveals further elements of the approach. The simulation will involve only mappings of the control points of the detail surface. We lock each control point of D to some individual origin site chosen on the base spline S_0, and the displacements $\mathbf{d}$ of these control points are measured with respect to coordinate vectors that are sensitive to the orientation of the base spline at the locality of each origin site. This is simpler to describe in terms of curves; having an extra variable adds nothing of substance to the ideas.

The detail spline is $\mathrm{D}(r) = \sum_\ell D_\ell c_\ell(r)$, $r \in \mathcal{D}$ for some basis splines c_ℓ. The displacement to be formed from $\mathrm{D}(r)$ is $\mathbf{d}(r) = \sum_\ell [D_\ell - (r_\ell, 0)]\, c_\ell(r)$ where $D_\ell - (r_\ell, 0)$ can be represented as $(r_\ell, 0) + \alpha_\ell \mathbf{i}_\ell + \beta_\ell \mathbf{j}_\ell$ in terms of coordinate frames $\{(r_\ell, 0), \mathbf{i}_\ell, \mathbf{j}_\ell\}$ (the orthonormal coordinate vectors $\mathbf{i}_\ell$,$\mathbf{j}_\ell$ are usually all chosen the same, for convenience). The base spline is $\mathrm{S}_0(u) = \sum_k \mathrm{S}_k b_k(u)$, $u \in \mathcal{S}$, and for all $u \notin \mathbf{T}(\mathcal{D})$ the composite spline $\mathrm{S}_1(u)$ is taken to be identical to $\mathrm{S}_0(u)$. For $\mathbf{T}(\mathcal{D}) \subset \mathcal{S}$ the composite spline $\mathrm{S}_1(u)$ is taken to be

$$\mathrm{S}_1(u) = \sum_\ell \left[\mathrm{S}_0(\mathbf{T}r_\ell) + \alpha_\ell \mathbf{m}(\mathbf{T}r_\ell) + \beta_\ell \mathbf{n}(\mathbf{T}r_\ell)\right] c_\ell(\mathbf{T}^{-1}u)$$

where $\mathbf{m}(u)$ and $\mathbf{n}(u)$ are unit tangent and normal, respectively, to $\mathrm{S}_0(u)$. Figure 3 provides an overview. (To avoid computing with $\mathbf{T}^{-1}$ one may regard this portion of S_1 to be a separate, free-standing spline in the variable r and display it juxtaposed to S_0 to provide an impression of S_1 in its entirety. On workstations that support *trimming*, S_0 can be trimmed against this free-standing spline for a better visual impression. Finally, note that if the *sense*

of the normal $\mathbf{n}$ is changed, the result will be to map the displacement onto the opposite side of the curve. Thus, one may choose to map a displacement *positively* or *negatively*.)

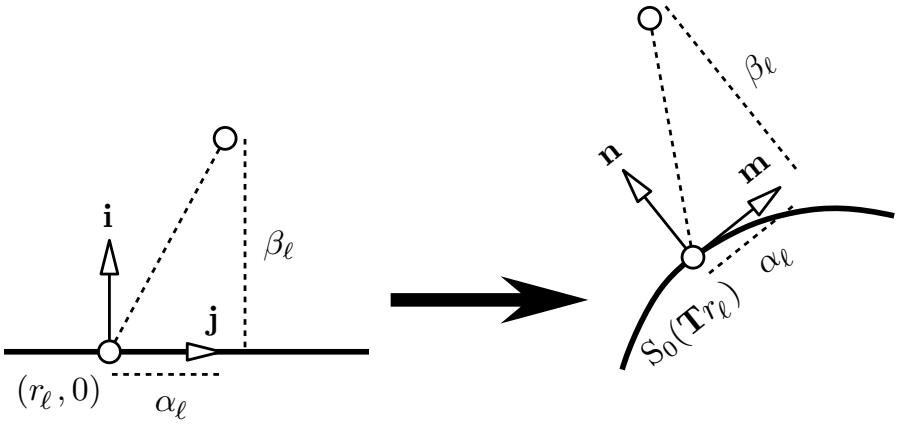


Fig. 3. Mapping displacement to base.

If modifications are made to D, these modifications are recorded through additional displacements for the control points of D

$$\mathbf{d}(r) \to \overline{\mathbf{d}}(r) = \sum_\ell \left[(r_\ell, 0) + (\alpha_\ell + \sigma_\ell)\mathbf{i}_\ell + (\beta_\ell + \tau_\ell)\mathbf{j}_\ell\right] c_\ell(r)$$

and these modifications may be applied to S_1 without the need to recompute the mapping

$$\mathrm{S}_1(u) = \sum_\ell \left[\mathrm{S}_0(\mathbf{T}r_\ell) + (\alpha_\ell + \sigma_\ell)\mathbf{m}(\mathbf{T}r_\ell) + (\beta_\ell + \tau_\ell)\mathbf{n}(\mathbf{T}r_\ell)\right] c_\ell(\mathbf{T}^{-1}u)$$

If any change is made retroactively to the base, $\mathrm{S}_0 \to \overline{\mathrm{S}}_0$, then the altered version of S_1 becomes

$$\mathrm{S}_1(u) \to \overline{\mathrm{S}}_1(u) = \sum_\ell \left[\overline{\mathrm{S}}_0(\mathbf{T}r_\ell) + \alpha_\ell \overline{\mathbf{m}}(\mathbf{T}r_\ell) + \beta_\ell \overline{\mathbf{n}}(\mathbf{T}r_\ell)\right] c_\ell(\mathbf{T}^{-1}u)$$

where, again, the mapping need not be recomputed. The only interactive design step that requires a recomputation of the mapping $\mathbf{T}$ is the step that revises the pasting of a detail onto a new base position.

The sites $(r_\ell, 0)$ chosen for this simulated displacement mapping are based on the *nodal points* (also called *Greville points*) for the basis $c_\ell(r)$; that is, the nodal point r_ℓ corresponding to c_ℓ is the value of r found by averaging all but the leftmost and rightmost knots on the support of c_ℓ. Several arguments make this a sensible choice. Under any reasonable refinement of D by knot insertion, each control point D_ℓ exhibits a higher order of convergence to the point $\mathrm{D}(r_\ell)$ than it does to any neighboring point $\mathrm{D}(r)$ [2], which provides this

simulated version of displacement mapping a good limiting correspondence with true displacement mapping under refinement. Further, under the view that the simulated displacement mapping is designed to lay the domain $\mathcal{D}$ over the spline S_0 and thereby transmit the shape of S_0 in some meaningful way to the detail D, this association of nodal points with control points is known to be effective in providing such a geometric transmission of shape [5]. Finally, to take the most naive view, if D and S_0 both simply constituted flat planes, as suggested by the first part of Figure 2, and the mapping constituted the identity, $\mathbf{T}r \equiv r \equiv u$ then $\mathrm{S}_0(u) = \mathrm{S}_1$, and the most convenient values to take for α_ℓ and β_ℓ would be zero. In this setting our mapping would simplify to,

$$\mathrm{S}_0(u) = \mathrm{S}_1(u) = (u, 0) = \sum_\ell \left[\mathrm{S}_0(u_\ell)\right] c_\ell(u) = \sum_\ell (u_\ell, 0) c_\ell(u)$$

In order for this to be true, we must have $u = \sum_\ell u_\ell c_\ell(u)$ which is another property satisfied by the nodal points [2].

The above discussion extends to surfaces if (r) is replaced by (r, s), (u) is replaced by (u, v), the tangential curve coordinate vector $\mathbf{m}(u)$ is replaced by a suitable pair of tangential surface coordinate vectors $\mathbf{l}(u, v)$, $\mathbf{m}(u, v)$, a bivariate spline basis is used, and some equivalent of the nodal points can be found. For tensor product surfaces the extension is immediate and obvious.

§Hierarchical Application and Overlapping Domains

The use of displacement maps need not be restricted to a single detail $\mathrm{D} \equiv \mathrm{D}_0$. The surface S_1 can serve as the base for a further detail D_1, and the process can continue through a sequence of displacements

$$\langle \mathrm{S}_0, \mathrm{D}_0 \rangle \to \mathrm{S}_1 \,,\ \langle \mathrm{S}_1, \mathrm{D}_1 \rangle \to \mathrm{S}_2 \,,\ \langle \mathrm{S}_2, \mathrm{D}_2 \rangle \to \mathrm{S}_3 \,,\ \ldots$$

In this context it is more meaningful to speak of a *base composition*, consisting of an original base and zero or more details pasted upon it, and a *detail composition* consisting of another original base and zero or more details. The only thing distinguishing a base composition from a detail composition is the *pasting order*: the composite detail will be applied as a displacement upon the composite base.

Figure 5, for example, shows the domains of a base composition $\mathcal{B} = \{\mathcal{B}_1, \mathcal{B}_2, \mathcal{B}_3, \mathcal{B}_4, \mathcal{B}_5\}$ and a detail composition $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2\}$. The components of each set are listed in pasted order. Domain $\mathcal{B}_1$ is the *root* of the composition, since it is not pasted onto any domain and is sufficiently large to accommodate all its pasted details. (*Every* surface composition is required to have a root domain – we have not investigated the possibility of having portions of a detail "hanging free" from any base.)

The domain dependencies are shown in Figure 6, where the nodes of the graph represent domains, and edges indicate portions of domains that overlap. Although both $\mathcal{B}_5$ and $\mathcal{D}_2$ are subdomains of $\mathcal{D}_1$, they lie on opposite sides: $\mathcal{B}_5$, a base component, lies *under* $\mathcal{D}_1$, while detail surface $\mathcal{D}_2$ is pasted *on* $\mathcal{D}_1$.

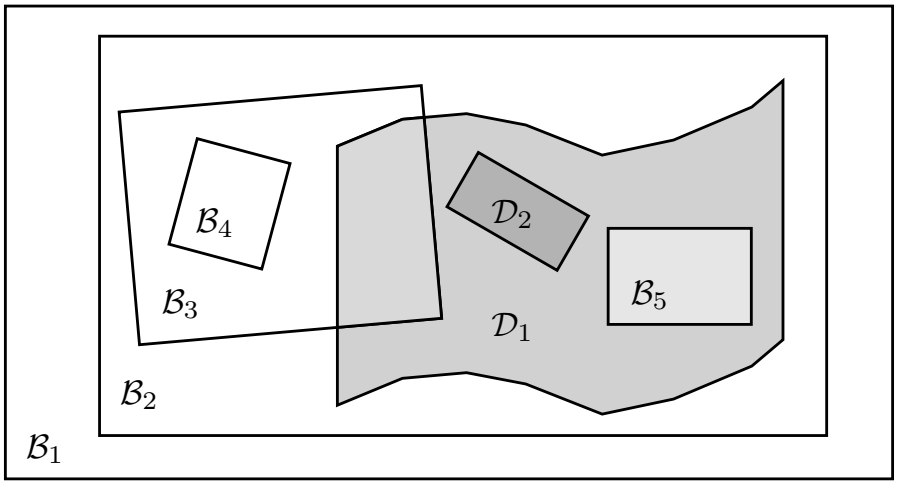


Fig. 5. Overlapping domains: the polygonal view.

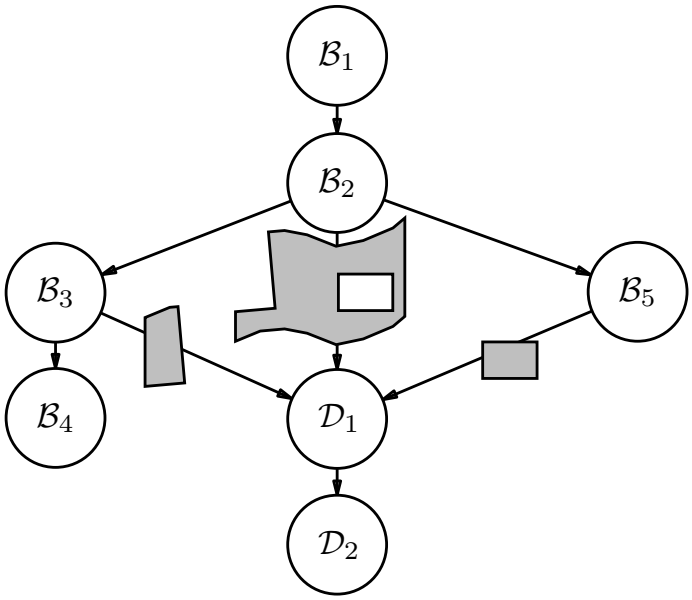


Fig. 6. Overlapping domains: the structural view.

This overlapping structure may change during interactive design as a detail composition is added to a base composition, moved about upon the base composition, or taken away from the base composition. When the structure of a composition changes, two steps precede the surface mapping operation: (1) recomputation of domain intersections, which requires a fast polygon-polygon clipping algorithm, and (2) reassociation of each Greville point on the detail domain composition with its underlying base-domain intersections, which requires a sequence of point-in-polygon operations, each one of which is followed by the construction of a surface-coordinate frame $\{\mathrm{S}, \mathbf{l}, \mathbf{m}, \mathbf{n}\}$ and the mapping of control vertices. This second step constitutes revising the mapping $\mathbf{T}$ between base composition and detail composition and represents the major expense for interactive updating. (If the domain dependency structure does

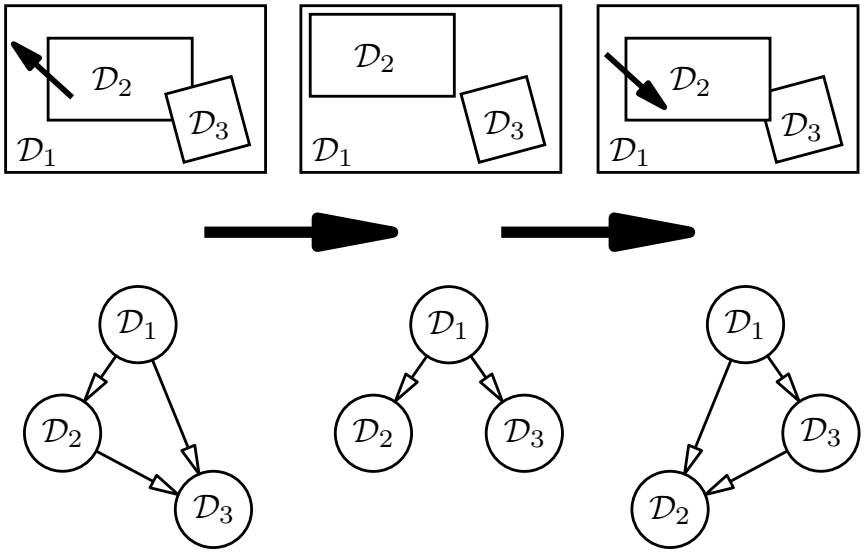


Fig. 7. Domain shuffling in a DAG.

not change but the *shape* of any component of surface – base or detail – is modified through the change of one or more control points, the two steps just mentioned are not necessary, but the control points for any detail surface affected must be re-mapped in pasted order, which requires evaluation of the affected surfaces at the nodal points, reconstruction of coordinate frames, and the transcription of control-point components into the new frames.)

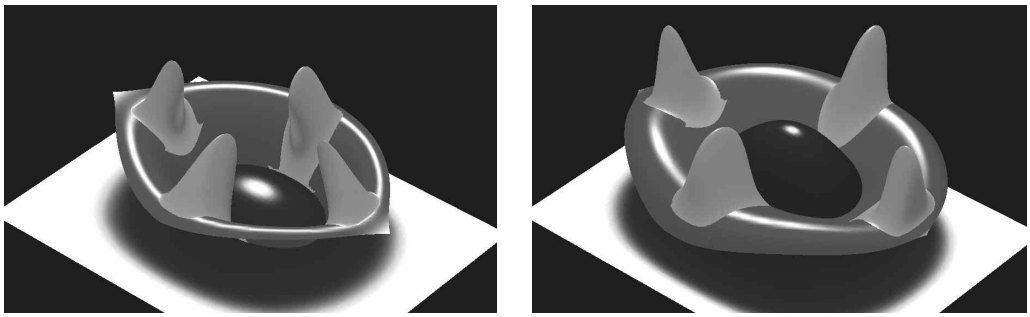


Fig. 8. Detail changes as base is modified.

The domain dependencies form a directed graph with interconnected nodes (Figure 6). Once some surface slides completely from under its detail surfaces, it ceases to influence the shape of these detail surfaces. If subsequently shifted back across any ex-detail surface domain, it is pasted *on top of* this ex-detail surface (Figure 7). This mechanism acts to prevent loop formation in the graph and provides an intuitive rule for structural motion. Thus, the layout of pasted domains may be maintained as a directed acyclic graph (DAG). The technical details of maintaining this DAG, together with some shortcuts involving the use and updating of a spanning tree for the DAG are covered in [1].

A surface editor, *PasteMaker*, has been written in C++ at the Computer Graphics Laboratory in the University of Waterloo to illustrate the feasibility

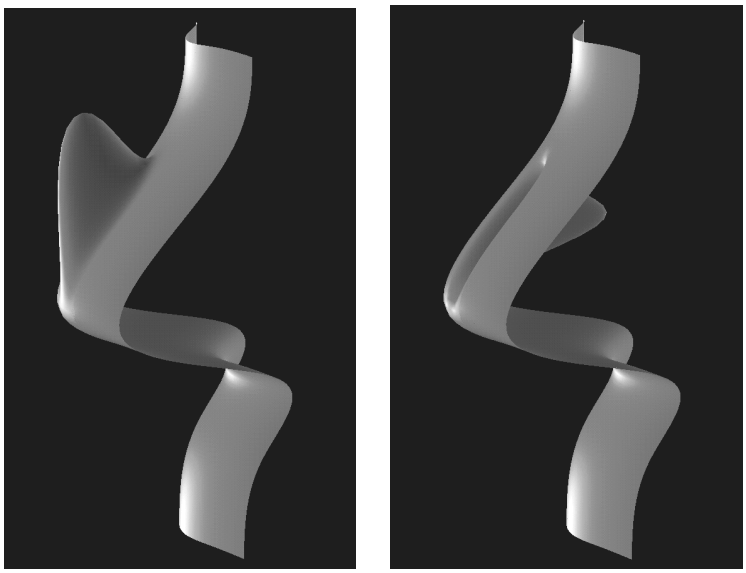


Fig. 9. Positive and negative pasting of the same detail.

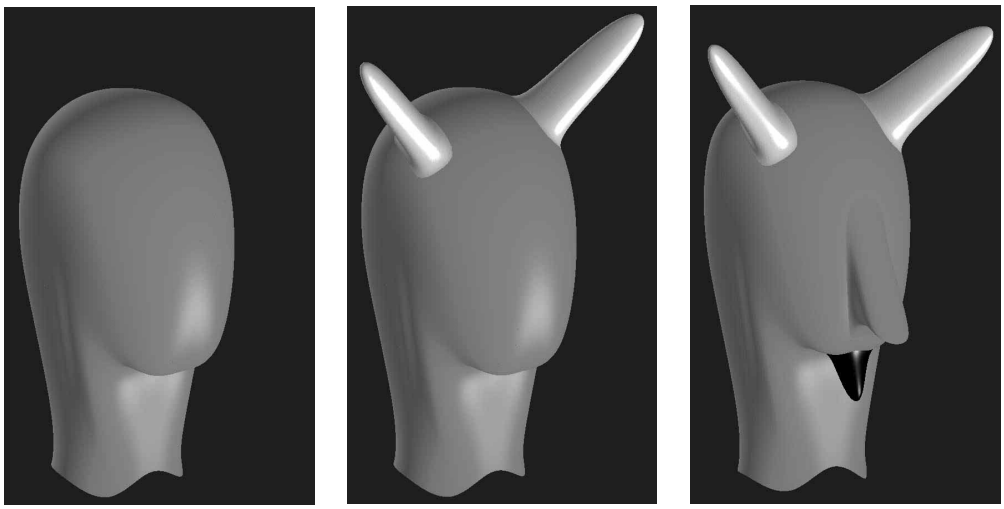


Fig. 10. Pasting some features on a head.

of interactive design using displacements. Editing may proceed at any level of a composition. The surfaces existing later in pasted order, thus serving as a detail composition, react in a natural way to the modification of their base composition (Figure 8). As we have mentioned previously, pasting may be *positive* or *negative*; that is, a detail may be attached so as to provide a bump or a depression by adjusting the sense of the surface normal (Figure 9). Figure 10 provides an illustration of three stages in the design of a cartoon character for animation. The head surface was imported from another spline editor, while the beard, nose, and horns were designed within *PasteMaker*. Finally, Figure 11 shows a comparison of simulated displacement and true displacement. The upper left image in the figure shows a base and detail surface before pasting. The upper right image shows the result of pasting the detail at an angle across the base. The detail's 10 patches in the lengthwise

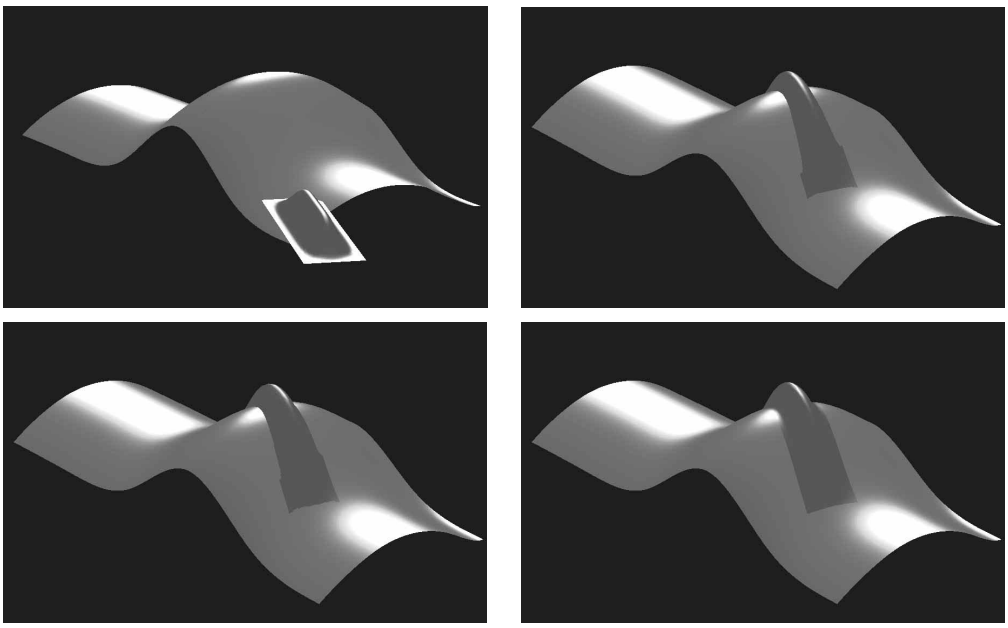


Fig. 11. Simulated vs. true displacement mapping.

direction are not quite able to "follow" the bulge in the base surface, resulting in an intrusion of the base into the detail. The lower left image shows the detail surface after refinement (naively, by midpoint insertion in the parameter corresponding to the lengthwise direction). The final image shows the situation after one more naive refinement. This image is indistinguishable from that produced by actual displacement mapping.

Acknowledgements. Funding was contributed by NSERC, ITRC, and GM. The authors are grateful for the help of Greg Veres, Milan Sreckovic, Bruce Hickey, Alex Nicolaou, and Al Vermeulen, who provided a number of C++ classes. Thanks are also due to Barry Fowler for helpful discussions.

References

1. Barghiel, Cristin, Feature oriented composition of B-spline surfaces, Master's thesis, Department of Computer Science, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1 (Available as Computer Science Department Technical Report CS-94-13), March 1994.
2. de Boor, Carl, *A Practical Guide to Splines*. Springer-Verlag, New York, New York, 1978.
3. Crow, Frank, Texture, in *Computer Graphics Techniques: Theory and Practice*, D. F. Rogers and R. A. Earnshaw (eds.), Springer-Verlag, New York, New York, 1990, 159–187.
4. Forsey, David, and Richard Bartels, Hierarchical B-Spline Refinement, *Proceedings of SIGGRAPH '88*, Computer Graphics **22**, 4 (1988) 205–212.
5. Goodman, Tim, Bernstein-Schoenberg Operators, submitted to *Mathematical Methods in Computer Aided Geometric Design III*, T. Lyche and

L. L. Schumaker (eds.), Academic Press, New York, New York, 1995.
6. Kaasa, Johannes, and Geir Westgaard, The 'Face Lift' Algorithm, in *Wavelets, Images, and Surface Fitting*, P-J. Laurent, A. Le Méhauté, and L. L. Schumaker (eds.), A K Peters, Wellesley, Massachusetts, 1994, 303–310.
7. Rogers, David, *Procedural Elements for Computer Graphics*, McGraw-Hill, New York, New York, 1985.
8. Upstill, Steve, *The RenderMan Companion* Addison-Wesley, Reading, Massachusetts, 1990.
9. Weller, Frank, and Hans Hagen, Tensor Product Spline Spaces with Knot Segments, submitted to *Mathematical Methods in Computer Aided Geometric Design III*, T. Lyche and L. L. Schumaker (eds.), Academic Press, New York, New York, 1995.

Cristin Barghiel
Side Effects Software Inc.
20 Maud Street, Suite 300
Toronto, Ontario CANADA M5V 2M5
cb@sidefx.com

Richard H. Bartels
Computer Science Department
University Of Waterloo
200 University Ave. W.
Waterloo, Ontario CANADA N2L 3G1
rhbartel@uwaterloo.ca

David R. Forsey
Department of Computer Science
University of British Columbia
Vancouver, BC CANADA V6T 1Z2
drforsey@cs.ubc.ca

Exploiting Subdivision in Modeling and Animation

David R. Forsey

The University of British Columbia
&
Radical Entertainment Inc.

Topics

Multiresolution Animation

- deformation
- How to use it.

Retrospective

- ***"Hierarchical B-Spline Refinement", SIGGRAPH'88***

  Forsey/Bartels.

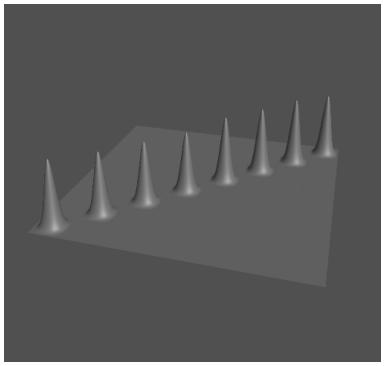
Motivations for H-Splines

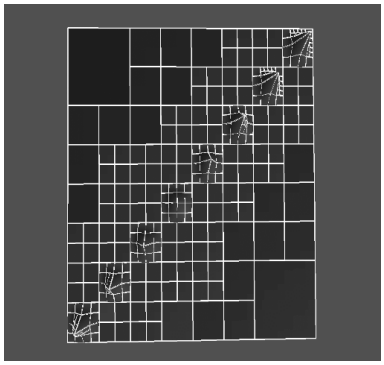
- ***H-Splines were developed specifically for animation.***
- ***Address problems with traditional B-splines***
  - *non-local knot insertion*
  - *the "degree-of-freedom" problem.*

Local Refinement

- Models rapidly become “heavy”
  - *patches in locations they are not needed*
  - *extra rendering overhead*
  - *may affect previously defined animation*
  - *even more points to animate*
- Local refinement adds patches to a restricted area

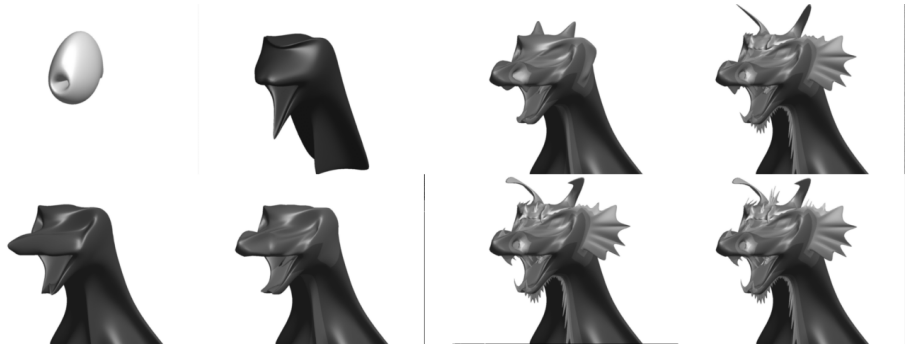
Local Refinement





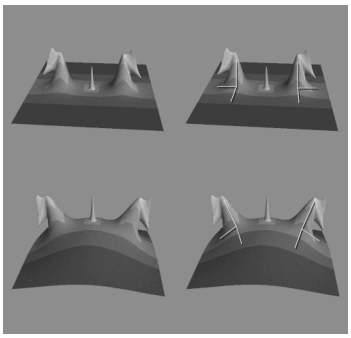
Local Refinement

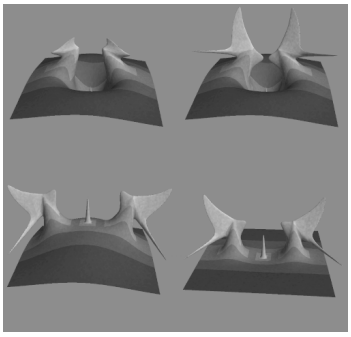
- Created using a multiresolution representation
- Can dramatically reduce the storage requirements



Multiresolution Editing

- Each resolution procedurally related to the next coarsest level
  - *arbitrary procedure (offset operator) for each control point.*
  - *creates a "reactive surface"*





H-Splines

Applicable to any surface formulation with:

- Local support
- Subdivision that does not alter the shape of the limit surface
  - *polygons*
  - *other splines (including curves & deformation lattices)*
  - *various subdivision surface formulations*
  - *wavelets (depending upon basis used)*

H-Splines

Advantages

- Top-down modelling (more like sculpting)
  - *details maintained during broad-scale editing*
  - *faster/easier to animate complex surfaces*
- Efficient representation (compression)
- Refinement only where needed

Multiresolution Animation

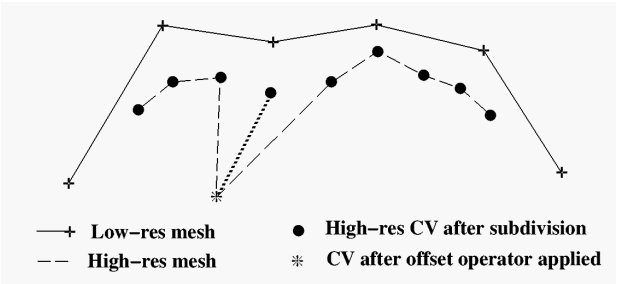
Available in a limited form commercially:

- MetaNURBs (Lightwave)
- Meshsmooth (3DMax)
- Symbolics L-surf (now Nicheman)
- various polygon smoothing packages

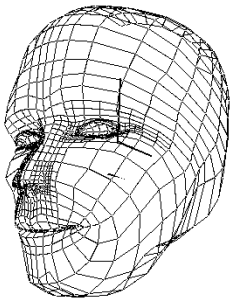
Animating with H-Splines

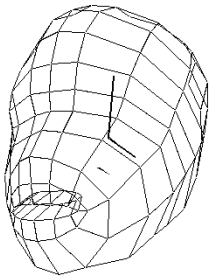
Offset Operators:

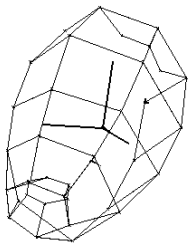
- tangent plane offset
- “frame” offset (attaches a control vertex to a skeletal segment)
- others described later



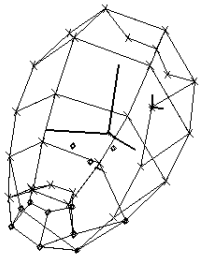
Simple Head

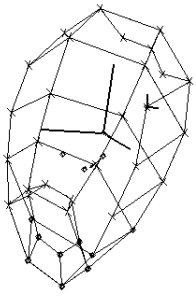


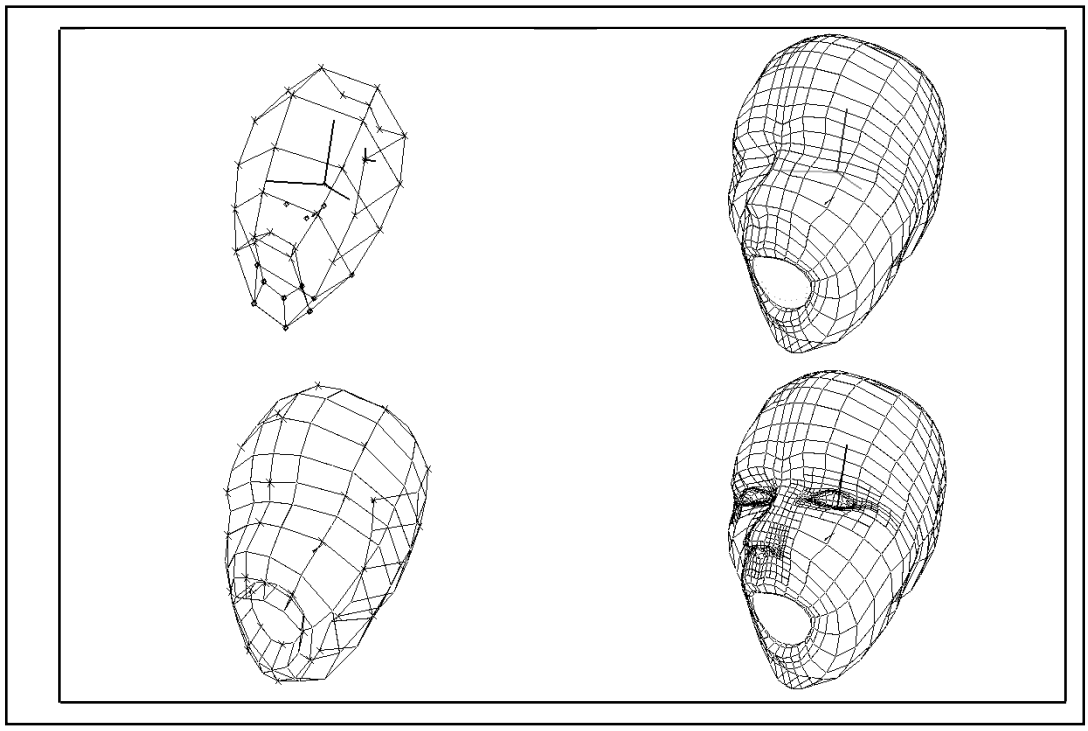


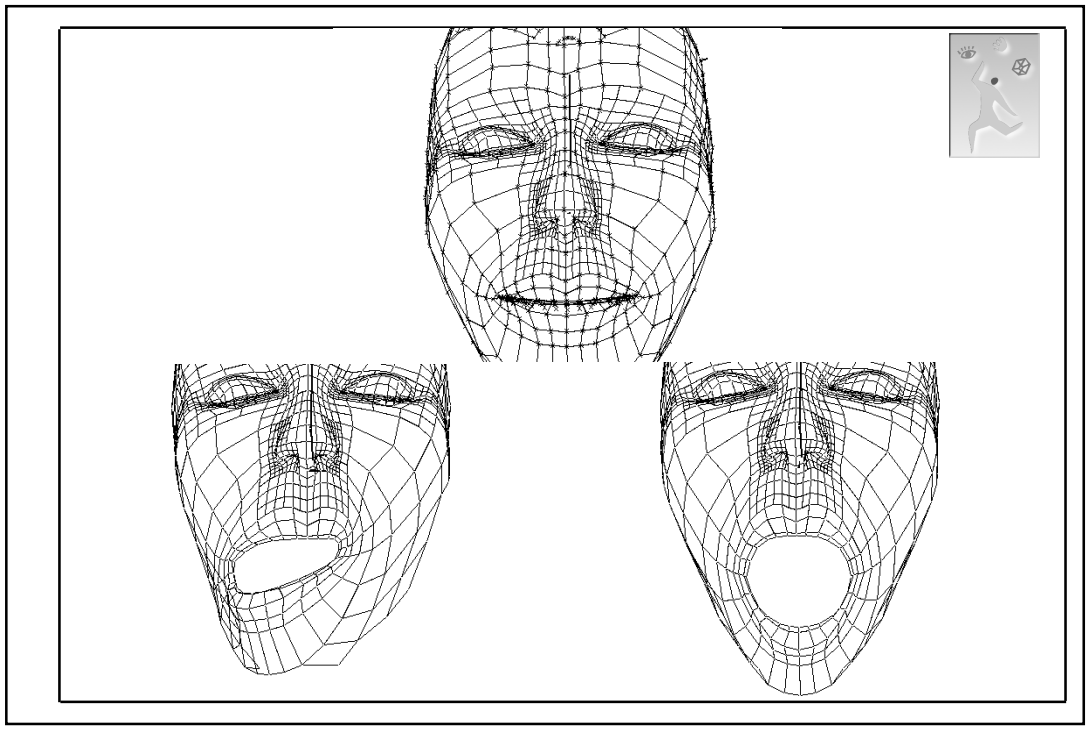


Low-Res Attachment



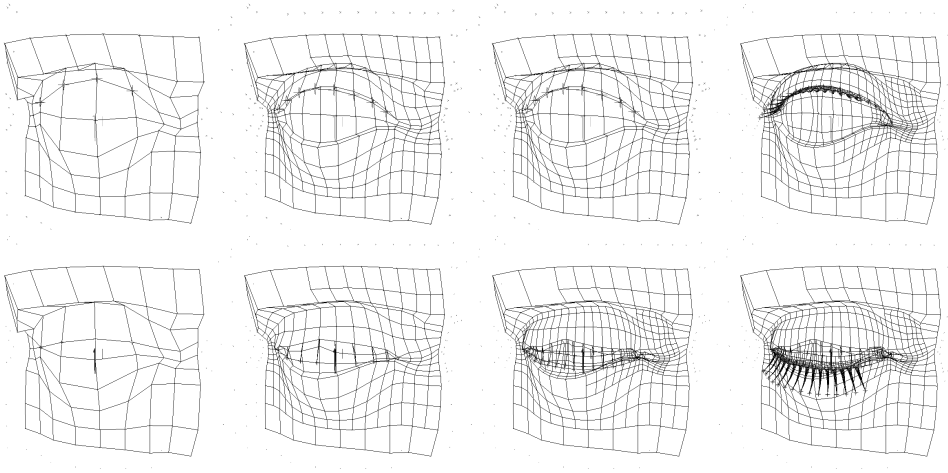




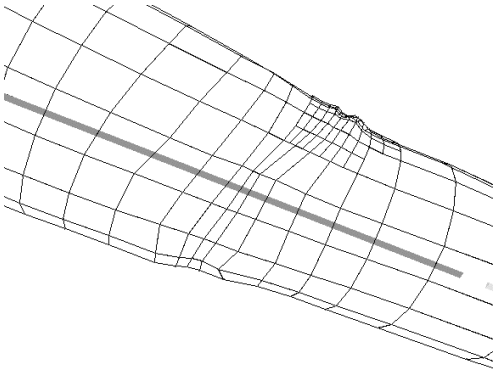


Example Animation

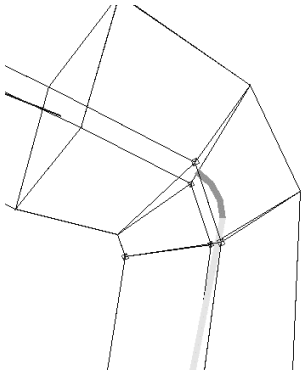
“Locking down” geometry



Locking Down Geometry, Part II

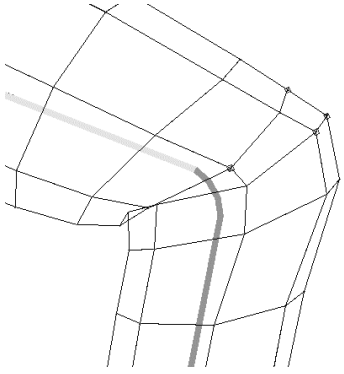


Full Detail

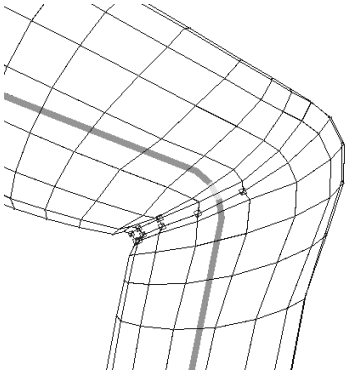


Lowest Resolution

Locking Down Geometry, Part II

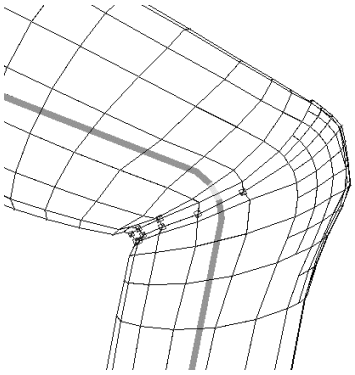


Level 1

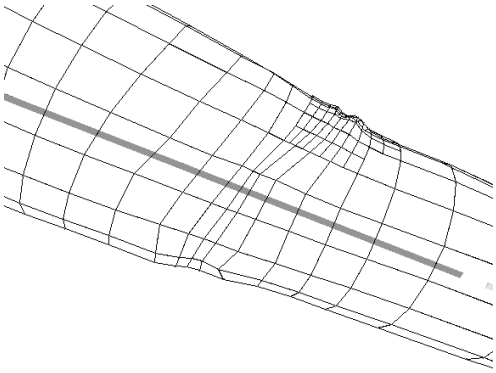


Level 2

Locking Down Geometry, Part II

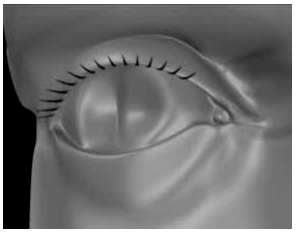


Full Resolution Bent



Full Resolution Straight

Example Animation

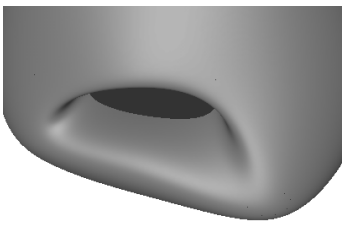


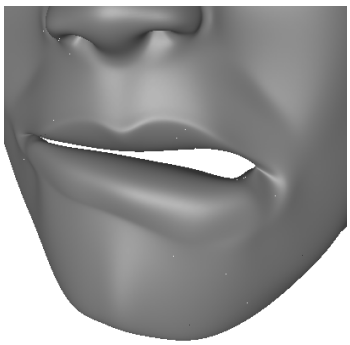
Multiresolution Animation

Advantages

- allows top-down approach to surface animation
- easy to mix broad-scale and fine scale effects
- faster to animate
  - *low-res model always available (for interactive speed)*
  - *less worry about high-res model behaviour*

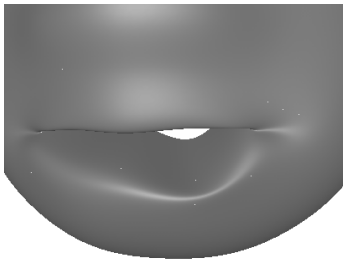
Layered Animation

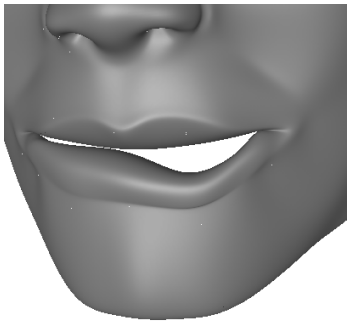




Low-res modification ***High-res effect***

Layered Animation

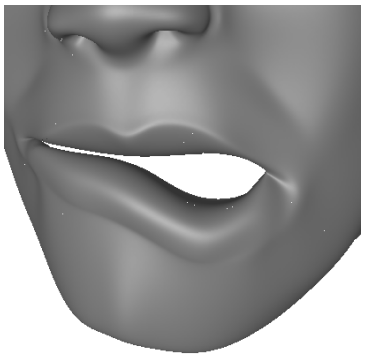


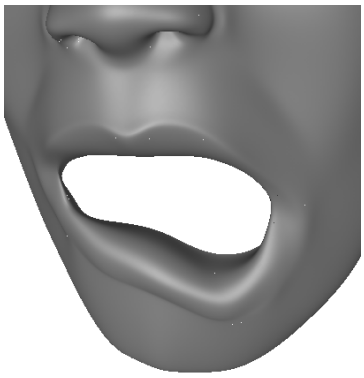


Medium-res modification

High-res effect

Layered Animation

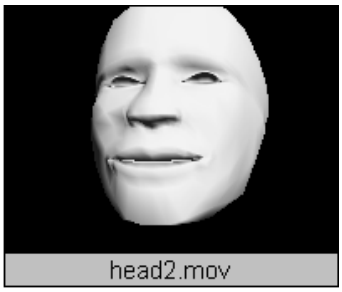


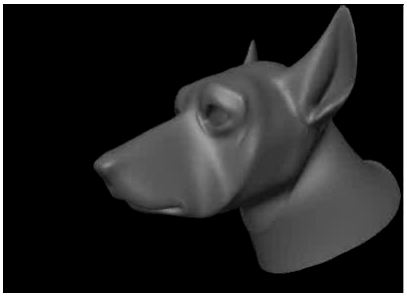


Both modifications

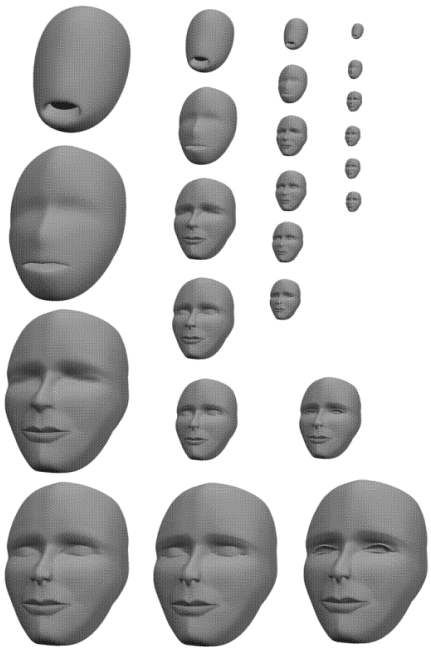
Both modifications, jaw open

Example Animation





Level of Detail



- Because the animation occurs at multiple levels of detail, low-res models still animate when used as low resolution geometry

Layered Animation

- Animation at each resolution layers deformations rather than blending them.
- Makes it easy to combine vertex animation with shape interpolation.
- Can combine relative (offset animation) with absolute (skeletal attachments) effects easily.
- Provides LOD for animation as well as geometry

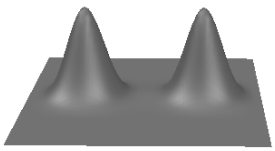
Editing Operations

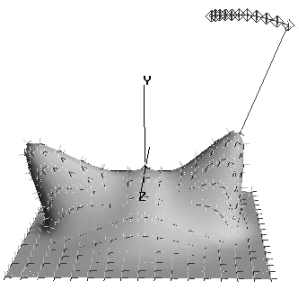
- Any spline-based tools can be used with H-Splines
- Additional operations possible:
  - *copy level, region (rubber stamp details)*
  - *non-hierarchical editing*
  - *modify offset:*
    - » move cv normal/tangent/along parent surface
    - » move cv along offset
    - » rotate offset
    - » move offset origin

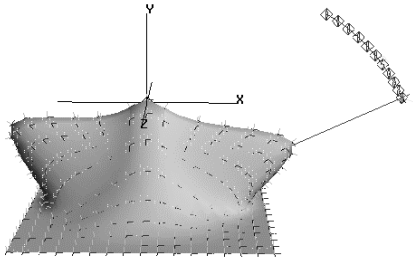
Secondary Motion using Offsets

- Typical spring/mass simulations produce “jello” effects
- Can provide more structure by treating each offsets as a rod/mass system.
  - *prevents motion into surface*
  - *surface details preserved*
  - *fast, easy to calculate*

Secondary Motion using Offsets







Secondary Motion using Offsets

Sample animation

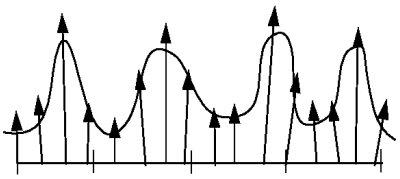
Secondary Motion using Offsets

Wrinkles

- Uses a “behaviour map” indicating location of wrinkle and dynamic properties (if combined with rod/mass simulation).
  - *value from map used to determine how to alter the offset parameterized by change in distance to neighbour*
  - *low-res modifications to surface increase/decrease size of wrinkle*

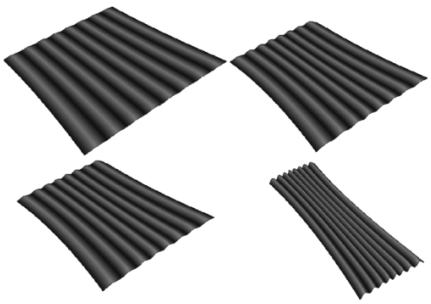
Secondary Motion using Offsets

Wrinkles



Secondary Motion using Offsets

Wrinkles



Surface Approximation

Problem

- Animators need to control placement of isoparms
- Draw lines directly onto sculptures for digitization as a spline control mesh
- Non-locality of knots creates heavier mesh, crowded lines

H-Spline Approach

- Allows mesh to be drawn with areas of higher detail

Variations

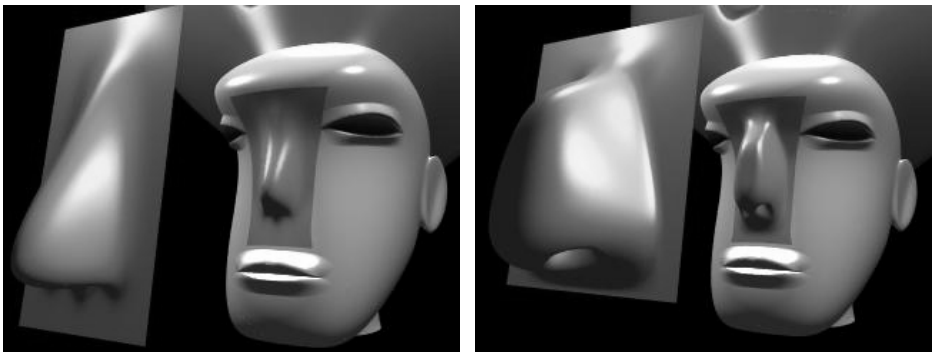
Link to paper

Surface Pasting

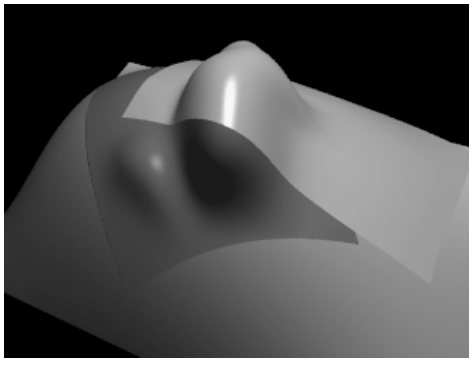
- Diagonal features are difficult to model
- Such features are modeled separately and “pasted” onto the low-res surface
- Generalizes H-Splines, but are more computationally expensive.

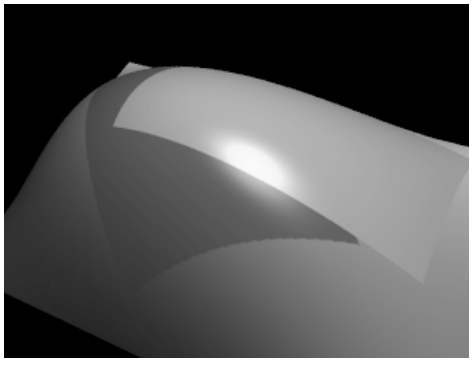
(Note: commercially available in the Houdini animation system)

Surface Pasting



Surface Pasting



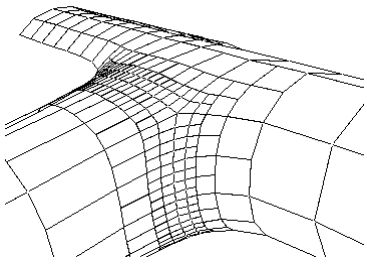


Overlapping features ***Underlying parameter space***

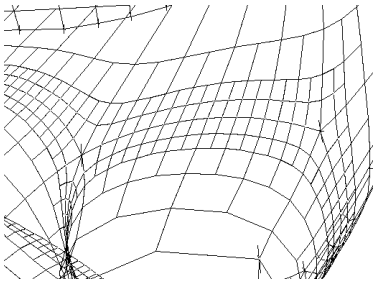
Grafting Surfaces

- Arbitrary topology surfaces difficult with tensor-product splines
  - *use Catmull-Clark subdivision surfaces to join spline surfaces*
  - *maintain compatibility with B-splines*
  - *minimizes the number of extraordinary points*
  - *maintains a simple U/V mapping for textures and attribute maps*

Grafting

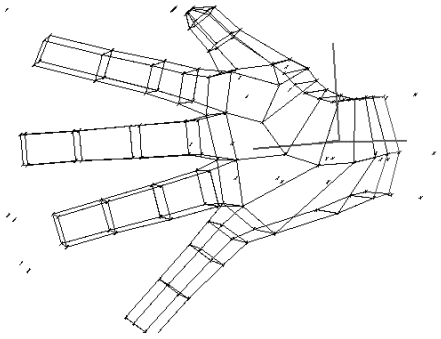


Two grafted cylinders

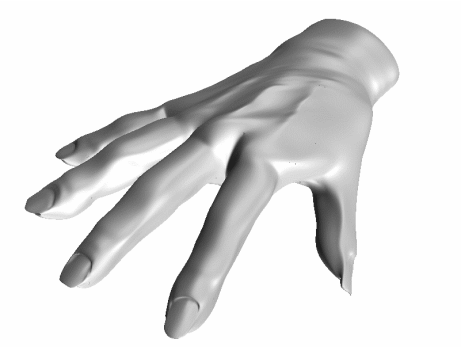


Finger grafts

Grafting



Low-res hand with grafts



Rendered hand

Multiresolution Simulation

Problem

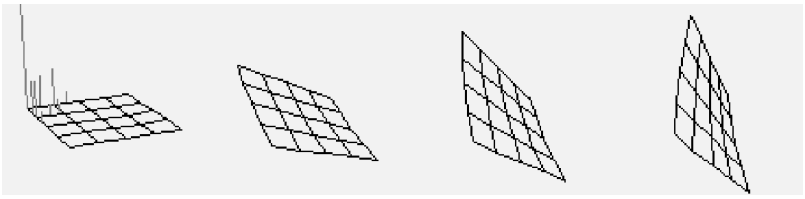
- simulations based on simple finite elements (spring/mass meshes) are difficult to control
  - *unclear how to set the parameters for a particular behaviour*
  - *stiff systems suffer from numerical instability or small step sizes*
  - *very few formulations that deal with micro/macro effects in the same model*

Multiresolution Simulation

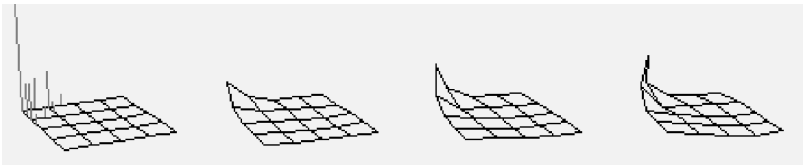
link to paper

Animator's Solution:

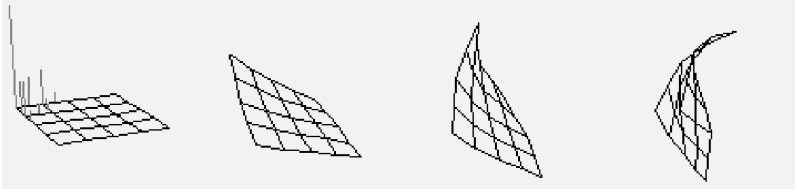
- use a multiresolution representation of surface
- propagate a proportion of forces to each successive resolution
  - *similar to multigrid methods, but different equation solved at each resolution*
  - *can retain surface details using an offset representation*



100 % of forces applied at lowest resolution



100 % of forces applied at highest resolution

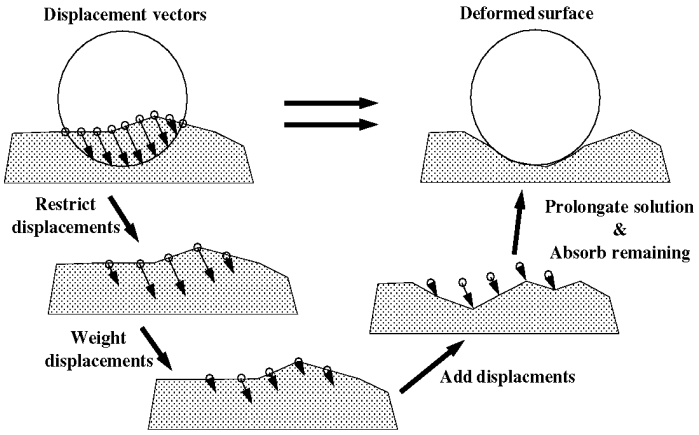


Forces applied at multiple resolutions

Multiresolution Deformations

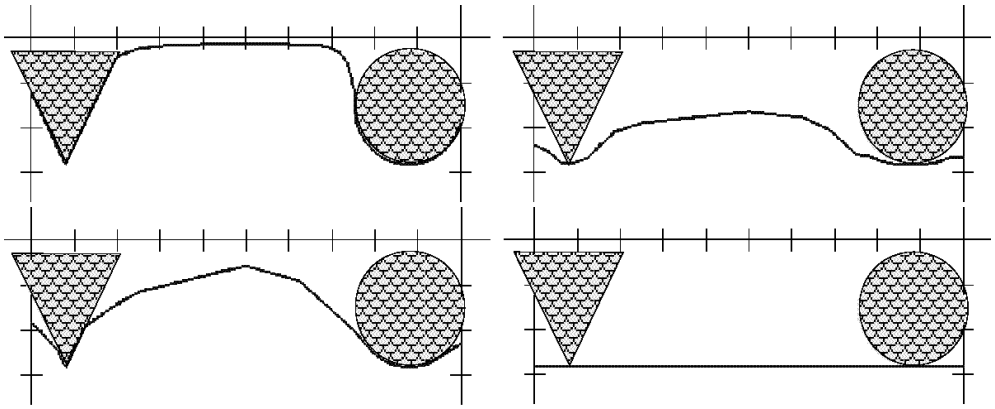
link to paper

- Uses similar approach, but uses kinematics rather than dynamics

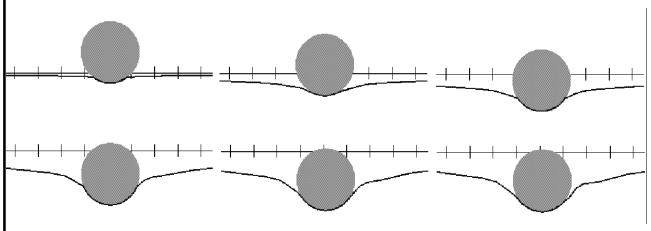


Multiresolution Deformations

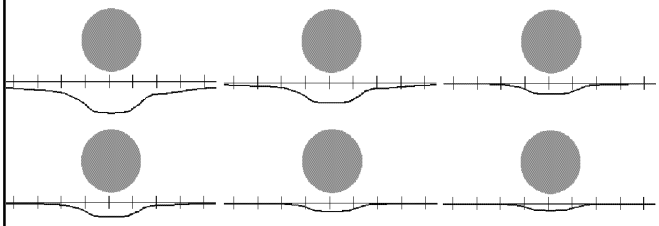
- Different behaviours are created by varying the amount each resolution accomodates the deformation



Multiresolution deformations



Behaviour during insertion of sphere



Behaviour during withdrawal of sphere

Time varying effects are added by delaying the change to the position and orientation of the offsets

Summary

Multiresolution Modeling and Animation

- provides a number of additional tools based on offset operators
- fast, easy to use on complex models
- provides good control over complex models
- compatible with tools or simulations that operate on spline surfaces
- computations simple enough for real-time

Chapter 10

Subdivision Surfaces in the Making of Geri's Game

Speaker: Tony DeRose

Subdivision Surfaces in Character Animation

Tony DeRose Michael Kass Tien Truong

Pixar Animation Studios

Figure 1: Geri.

Abstract

The creation of believable and endearing characters in computer graphics presents a number of technical challenges, including the modeling, animation and rendering of complex shapes such as heads, hands, and clothing. Traditionally, these shapes have been modeled with NURBS surfaces despite the severe topological restrictions that NURBS impose. In order to move beyond these restrictions, we have recently introduced subdivision surfaces into our production environment. Subdivision surfaces are not new, but their use in high-end CG production has been limited.

Here we describe a series of developments that were required in order for subdivision surfaces to meet the demands of high-end production. First, we devised a practical technique for constructing provably smooth variable-radius fillets and blends. Second, we developed methods for using subdivision surfaces in clothing simulation including a new algorithm for efficient collision detection. Third, we developed a method for constructing smooth scalar fields on subdivision surfaces, thereby enabling the use of a wider class of programmable shaders. These developments, which were used extensively in our recently completed short film *Geri's game*, have become a highly valued feature of our production environment.

1 Motivation

The most common way to model complex smooth surfaces such as those encountered in human character animation is by using a patchwork of trimmed NURBS. Trimmed NURBS are used primarily because they are readily available in existing commercial systems such as Alias-Wavefront and SoftImage. They do, however, suffer from at least two difficulties:

1. Trimming is expensive and prone to numerical error.

2. It is difficult to maintain smoothness, or even approximate smoothness, at the seams of the patchwork as the model is animated. As a case in point, considerable manual effort was required to hide the seams in the face of Woody, a principal character in *Toy Story*.

Subdivision surfaces have the potential to overcome both of these problems: they do not require trimming, and smoothness of the model is automatically guaranteed, even as the model animates.

The use of subdivision in animation systems is not new, but for a variety of reasons (several of which we address in this paper), their use has not been widespread. In the mid 1980s for instance, Symbolics was possibly the first to use subdivision in their animation system as a means of creating detailed polyhedra. The LightWave 3D modeling and animation system from NewTek also uses subdivision in a similar fashion.

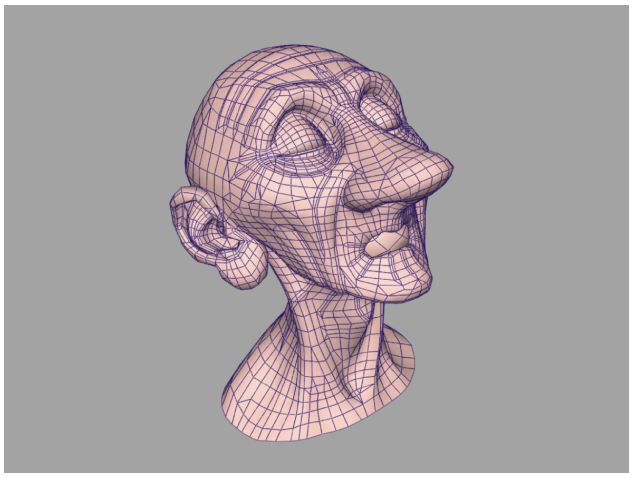


Figure 2: The control mesh for Geri's head, created by digitizing a full-scale model sculpted out of clay.

This paper describes a number of issues that arose when we added a variant of Catmull-Clark [2] subdivision surfaces to our animation and rendering systems, Marionette and RenderMan [17], respectively. The resulting extensions were used heavily in the creation of Geri (Figure 1), a human character in our recently completed short film *Geri's game*. Specifically, subdivision surfaces were used to model the skin of Geri's head (see Figure 2), his hands, and his clothing, including his jacket, pants, shirt, tie, and shoes.

In contrast to previous systems, such as those mentioned above, that use subdivision as a means to embellish polygonal models, our system uses subdivision as a means to define piecewise smooth surfaces. Since our system reasons about the limit surface itself, polygonal artifacts are never present, no matter how the surface animates or how closely it is viewed.

The use of subdivision surfaces posed new challenges throughout the production process, from modeling and animation to rendering. In modeling, subdivision surfaces free the designer from worrying about the topological restrictions that haunt NURBS modelers, but they simultaneously prevent the use of special tools that have been developed over the years to add features such as variable radius fillets to NURBS models. In Section 3, we describe an approach for introducing similar capabilities into subdivision surface models. The basic idea is to generalize the infinitely sharp creases of Hoppe *et. al.* [10] to obtain semi-sharp creases – that is, creases whose sharpness can vary from zero (meaning smooth) to infinite.

Once models have been constructed with subdivision surfaces, the problems of animation are generally easier than with corresponding NURBS surfaces. As alluded to earlier, the reason is that subdivision surface models are seamless, meaning that the surface is guaranteed to remain smooth as the model is animated.

Using subdivision surfaces for physically-based animation of clothing, however, poses its own difficulties which we address in Section 4. First, it is necessary to express the energy function of the clothing on subdivision meshes in such a way that the resulting motion does not inappropriately reveal the structure of the subdivision control mesh. Second, in order for a physical simulator to make use of subdivision surfaces it must compute collisions very efficiently. While collisions of NURBS surfaces have been studied in great detail, little work has been done previously with subdivision surfaces.

Having modeled and animated subdivision surfaces, some formidable challenges remain before they can be rendered. The topological freedom that makes subdivision surfaces so attractive for modeling and animation means that they generally do not admit parametrizations suitable for texture mapping. Solid textures [12, 13] and projection textures [9] can address some production needs, but Section 5.1 shows that it is possible to go a good deal further by using programmable shaders in combination with smooth scalar fields defined over the surface.

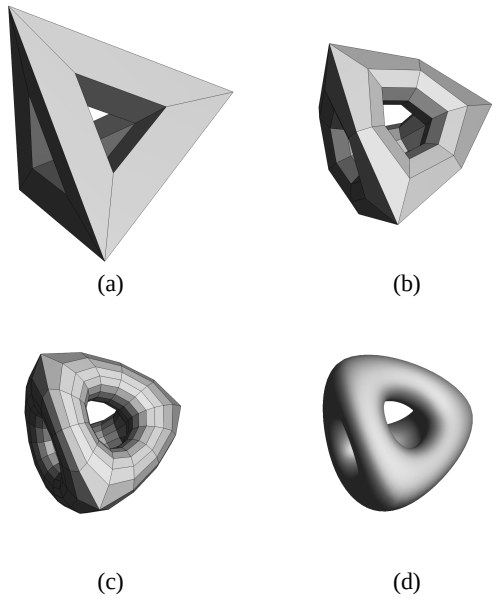


Figure 3: Recursive subdivision of a topologically complicated mesh: (a) the control mesh; (b) after one subdivision step; (c) after two subdivision steps; (d) the limit surface.

The combination of semi-sharp creases for modeling, an appropriate and efficient interface to physical simulation for animation, and the availability of scalar fields for shading and rendering have made subdivision surfaces an extremely effective tool in our production environment.

2 Background

A single NURBS surface, like any other parametric surface, is limited to representing surfaces which are topologically equivalent to a sheet, a cylinder or a torus. This is a fundamental limitation for any surface that imposes a global planar parameterization. A single subdivision surface, by contrast, can represent surfaces of arbitrary topology. The basic idea is to construct a surface from an arbitrary polyhedron by repeatedly subdividing each of the faces, as illustrated in Figure 3. If the subdivision is done appropriately, the limit of this subdivision process will be a smooth surface.

Catmull and Clark [2] introduced one of the first subdivision schemes. Their method begins with an arbitrary polyhedron called the control mesh. The control mesh, denoted M^0 (see Figure 3(a)), is subdivided to produce the mesh M^1 (shown in Figure 3(b)) by splitting each face into a collection of quadrilateral subfaces. A face having n edges is split into n quadrilaterals. The vertices of M^1 are computed using certain weighted averages as detailed below. The same subdivision procedure is used again on M^1 to produce the mesh M^2 shown in Figure 3(c). The subdivision surface is defined to be the limit of the sequence of meshes $M^0, M^1, \ldots$ created by repeated application of the subdivision procedure.

To describe the weighted averages used by Catmull and Clark it is convenient to observe that each vertex of M^{i+1} can be associated with either a face, an edge, or a vertex of M^i; these are called face, edge, and vertex points, respectively. This association is indicated in Figure 4 for the situation around a vertex v^0 of M^0. As indicated

in the figure, we use f's to denote face points, e's to denote edge points, and v's to denote vertex points. Face points are positioned at the centroid of the vertices of the corresponding face. An edge point e_j^{i+1}, as indicated in Figure 4 is computed as

$$e_j^{i+1} = \frac{v^i + e_j^i + f_{j-1}^{i+1} + f_j^{i+1}}{4}, \tag{1}$$

where subscripts are taken modulo the valence of the central vertex v^0. (The valence of a vertex is the number of edges incident to it.) Finally, a vertex point v^i is computed as

$$v^{i+1} = \frac{n-2}{n} v^i + \frac{1}{n^2} \sum_j e_j^i + \frac{1}{n^2} \sum_j f_j^{i+1} \tag{2}$$

Vertices of valence 4 are called ordinary; others are called extraordinary.

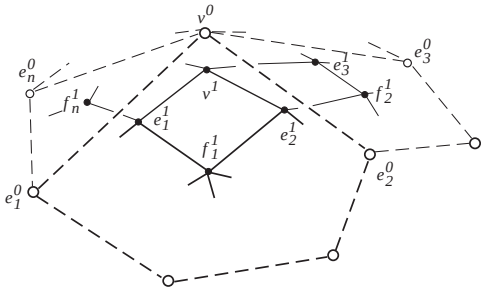


Figure 4: The situation around a vertex v^0 of valence n.

These averaging rules — also called subdivision rules, masks, or stencils — are such that the limit surface can be shown to be tangent plane smooth no matter where the control vertices are placed [14, 19].[1]

Whereas Catmull-Clark subdivision is based on quadrilaterals, Loop's surfaces [11] and the Butterfly scheme [6] are based on triangles. We chose to base our work on Catmull-Clark surfaces for two reasons:

1. They strictly generalize uniform tensor product cubic B-splines, making them easier to use in conjunction with existing in-house and commercial software systems such as Alias-Wavefront and SoftImage.

2. Quadrilaterals are often better than triangles at capturing the symmetries of natural and man-made objects. Tube-like surfaces — such as arms, legs, and fingers — for example, can be modeled much more naturally with quadrilaterals.

Following Hoppe *et. al.* [10] it is possible to modify the subdivision rules to create piecewise smooth surfaces containing infinitely sharp features such as creases and corners. This is illustrated by Geri's hand, shown in Figure 5. Infinitely sharp creases were used to separate the skin of the hand from the finger nails. Sharp creases can be modeled by marking a subset of the edges of the control mesh as being sharp, then using specially designed rules in the neighborhood of sharp edges. Appendix A describes the necessary special rules and when to use them.

Again following Hoppe *et. al.*, we deal with boundaries of the control mesh by tagging the boundary edges as sharp. We have also found it convenient to tag boundary vertices of valence 2 as corners, even though they would normally be treated as crease vertices since they are incident to two sharp edges. We do this to mimic the behavior of endpoint interpolating tensor product uniform cubic B-spline surfaces, as illustrated in Figure 6.

[1] Technical caveat for the purist: The surface is guaranteed to be smooth except for control vertex positions in a set of measure zero.

Figure 5: Geri's hand as a piecewise smooth Catmull-Clark surface. Infinitely sharp creases are used between the skin and the finger nails.

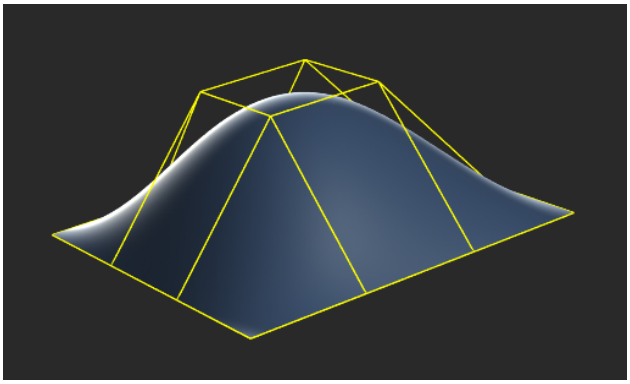


Figure 6: A surface where boundary edges are tagged as sharp and boundary vertices of valence two are tagged as corners. The control mesh is yellow and the limit surface is cyan.

3 Modeling fillets and blends

As mentioned in Section 1 and shown in Figure 5, infinitely sharp creases are very convenient for representing piecewise-smooth surfaces. However, real-world surfaces are never infinitely sharp. The corner of a tabletop, for instance, is smooth when viewed sufficiently closely. For animation purposes it is often desirable to capture such tightly curved shapes.

To this end we have developed a generalization of the Catmull-Clark scheme to admit semi-sharp creases – that is, creases of controllable sharpness, a simple example of which is shown in Figure 7.

One approach to achieve semi-sharp creases is to develop subdivision rules whose weights are parametrized by the sharpness s of the crease. This approach is difficult because it can be quite hard to discover rules that lead to the desired smoothness properties of the limit surfaces. One of the roadblocks is that subdivision rules around a crease break a symmetry possessed by the smooth rules: typical smooth rules (such as the Catmull-Clark rules) are invariant under cyclic reindexing, meaning that discrete Fourier transforms can be used to prove properties for vertices of arbitrary valence (cf. Zorin [19]). In the absence of this invariance, each valence must currently be considered separately, as was done by Schweitzer [15].

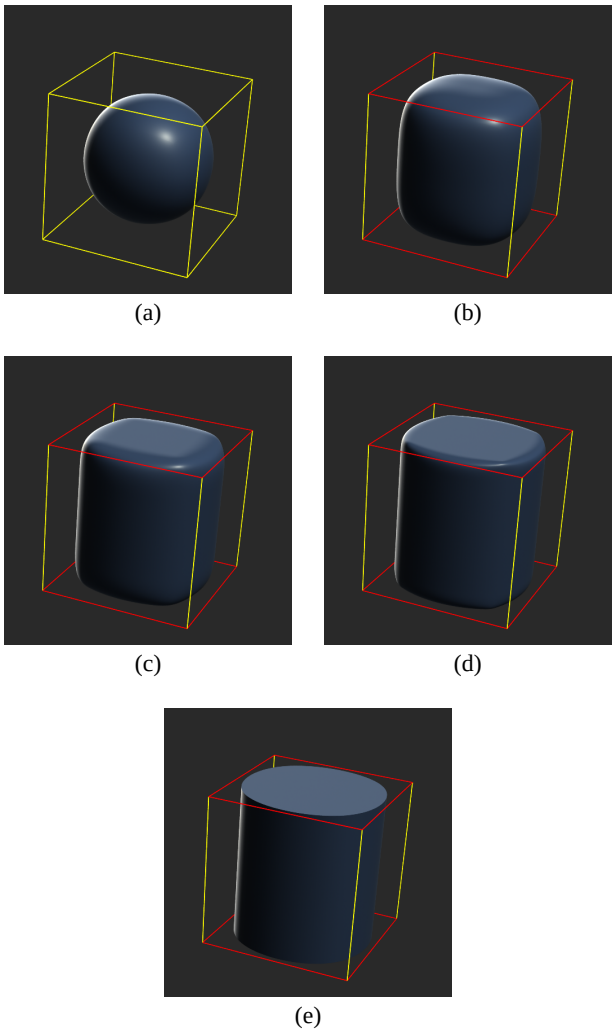


Figure 7: An example of a semi-sharp crease. The control mesh for each of these surfaces is the unit cube, drawn in wireframe, where crease edges are red and smooth edges are yellow. In (a) the crease sharpness is 0, meaning that all edges are smooth. The sharpnesses for (b), (c), (d), and (e) are 1, 2, 3, and infinite, respectively.

Another difficulty is that such an approach is likely to lead to a zoo of rules depending on the number and configuration of creases through a vertex. For instance, a vertex with two semi-sharp creases passing through it would use a different set of rules than a vertex with just one crease through it.

Our approach is to use a very simple process we call hybrid subdivision. The general idea is to use one set of rules for a finite but arbitrary number of subdivision steps, followed by another set of rules that are applied to the limit. Smoothness therefore depends only on the second set of rules. Hybrid subdivision can be used to obtain semi-sharp creases by using infinitely sharp rules during the first few subdivision steps, followed by use of the smooth rules for subsequent subdivision steps. Intuitively this leads to surfaces that are sharp at coarse scales, but smooth at finer scales.

Now the details. To set the stage for the general situation where the sharpness can vary along a crease, we consider two illustrative special cases.

Case 1: A constant integer sharpness s crease: We subdivide s times using the infinitely sharp rules, then switch to the smooth rules. In other words, an edge of sharpness $s > 0$ is subdivided using the sharp edge rule. The two subedges created each have sharpness $s - 1$. A sharpness $s = 0$ edge is considered smooth, and it stays smooth for remaining subdivisions. In the limit where $s \to \infty$ the sharp rules are used for all steps, leading to an infinitely sharp crease. An example of integer sharpness creases is shown in Figure 7. A more complicated example where two creases of different sharpnesses intersect is shown in Figure 8.

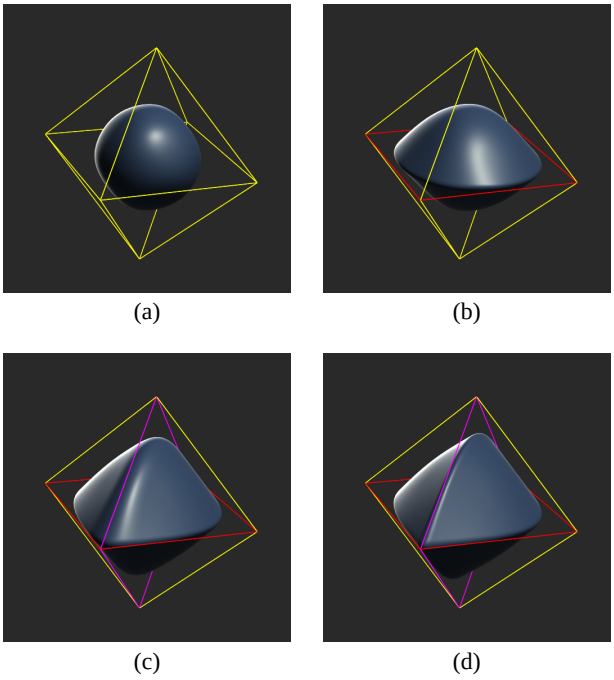


Figure 8: A pair of crossing semi-sharp creases. The control mesh for all surfaces is the octahedron drawn in wire frame. Yellow denotes smooth edges, red denotes the edges of the first crease, and magenta denotes the edges of the second crease. In (a) the crease sharpnesses are both zero; in (b), (c), and (d) the sharpness of the red crease is 4. The sharpness of the magenta crease in (b), (c), and (d) is 0, 2, and 4, respectively.

Case 2: A constant, but not necessarily integer sharpness s: the main idea here is to interpolate between adjacent integer sharpnesses. Let $s\downarrow$ and $s\uparrow$ denote the floor and ceiling of s, respectively. Imagine creating two versions of the crease: the first obtained by subdividing $s\downarrow$ times using the sharp rules, then subdividing one additional time using the smooth rules. Call the vertices of this first version $v\downarrow_0, v\downarrow_1, \ldots$. The second version, the vertices of which we denote by $v\uparrow_0, v\uparrow_1, \ldots$, is created by subdividing $s\uparrow$ times using the sharp rules. We take the $s\uparrow$-times subdivided semi-sharp crease to have vertex positions $v_i^{s\uparrow}$ computed via simple linear interpolation:

$$v_i^{s\uparrow} = (1 - \sigma)v\downarrow_i + \sigma v\uparrow_i \tag{3}$$

where $\sigma = (s - s\downarrow)/(s\uparrow - s\downarrow)$. Subsequent subdivisions are done using the smooth rules. In the case where all creases have the same non-integer sharpness s, the surface produced by the above process is identical to the one obtained by linearly interpolating between the integer sharpness limit surfaces corresponding to $s\downarrow$ and $s\uparrow$. Typically, however, crease sharpnesses will not all be equal, meaning that the limit surface is not a simple blend of integer sharpness surfaces.

The more general situation where crease sharpness is non-integer and varies along a crease is presented in Appendix B. Figure 9 de-

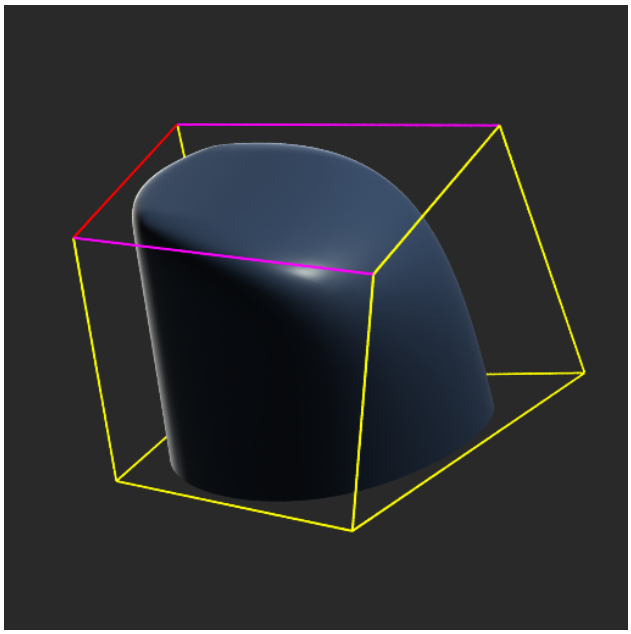


Figure 9: A simple example of a variable sharpness crease. The edges of the bottom face of the cubical control mesh are infinitely sharp. Three edges of the top face form a single variable sharpness crease with edge sharpnesses set to 2 (the two magenta edges), and 4 (the red edge).

Figure 10: A more complex example of variable sharpness creases. This model, inspired by an Edouard Lanteri sculpture, contains numerous variable sharpness creases to reduce the size of the control mesh. The control mesh for the model made without variable sharpness creases required 840 faces; with variable sharpness creases the face count dropped to 627. Model courtesy of Jason Bickerstaff.

picts a simple example. A more complex use of variable sharpness is shown in Figure 10.

4 Supporting cloth dynamics

The use of simulated physics to animate clothing has been widely discussed in the literature (cf. [1, 5, 16]). Here, we address the issues that arise when interfacing a physical simulator to a set of geometric models constructed out of subdivision surfaces. It is not our intent in this section to fully detail our cloth simulation system – that would require an entire paper of its own. Our goal is rather to highlight issues related to the use of subdivision surfaces to model both kinematic and dynamic objects.

In Section 4.1 we define the behavior of the cloth material by constructing an energy functional on the subdivision control mesh. If the material properties such as the stiffness of the cloth vary over the surface, one or more scalar fields (see Section 5.1) must be defined to modulate the local energy contributions. In Section 4.2 we describe an algorithm for rapidly identifying potential collisions involving the cloth and/or kinematic obstacles. Rapid collision detection is crucial to achieving acceptable performance.

4.1 Energy functional

For physical simulation, the basic properties of a material are generally specified by defining an energy functional to represent the attraction or resistance of the material to various possible deformations. Typically, the energy is either specified as a surface integral or as a discrete sum of terms which are functions of the positions of surface samples or control vertices. The first type of specification typically gives rise to a finite-element approach, while the second is associated more with finite-difference methods.

Finite-element approaches are possible with subdivision surfaces, and in fact some relevant surface integrals can be computed analytically [8]. In general, however, finite-element surface integrals must be estimated through numerical quadrature, and this gives rise to a collection of special cases around extraordinary points. We chose to avoid these special cases by adopting a finite-difference approach, approximating the clothing with a mass-spring model [18] in which all the mass is concentrated at the control points.

Away from extraordinary points, Catmull-Clark meshes under subdivision become regular quadrilateral grids. This makes them ideally suited for representing woven fabrics which are also generally described locally by a gridded structure. In constructing the energy functions for clothing simulation, we use the edges of the subdivision mesh to correspond with the warp and weft directions of the simulated woven fabrics.

Since most popular fabrics stretch very little along the warp or weft directions, we introduce relatively strong fixed rest-length springs along each edge of the mesh. More precisely, for each edge from p_1 to p_2, we add an energy term $k_s E_s(p_1, p_2)$ where

$$E_s(p_1, p_2) = \frac{1}{2}\left(\frac{|p_1 - p_2|}{|p_1^* - p_2^*|} - 1\right)^2. \tag{4}$$

Here, p_1^* and p_2^* are the rest positions of the two vertices, and k_s is the corresponding spring constant.

With only fixed-length springs along the mesh edges, the simulated clothing can undergo arbitrary skew without penalty. One way to prevent the skew is to introduce fixed-length springs along the diagonals. The problem with this approach is that strong diagonal springs make the mesh too stiff, and weak diagonal springs allow the mesh to skew excessively. We chose to address this problem by introducing an energy term which is proportional to the product of the energies of two diagonal fixed-length springs. If p_1 and p_2 are vertices along one diagonal of a quadrilateral mesh face and p_3 and p_4 are vertices along the other diagonal, the energy is given by

$k_d E_d(p_1, p_2, p_3, p_4)$ where k_d is a scalar parameter that functions analagously to a spring constant, and where

$$E_d(p_1, p_2, p_3, p_4) = E_s(p_1, p_2) E_s(p_3, p_4). \quad (5)$$

The energy $E_d(p_1, p_2, p_3, p_4)$ reaches its minimum at zero when either of the diagonals of the quadrilateral face are of the original rest length. Thus the material can fold freely along either diagonal, while resisting skew to a degree determined by k_d. We sometimes use weak springs along the diagonals to keep the material from wrinkling too much.

With the fixed-length springs along the edges and the diagonal contributions to the energy, the simulated material, unlike real cloth, can bend without penalty. To add greater realism to the simulated cloth, we introduce an energy term that establishes a resistance to bending along virtual threads. Virtual threads are defined as a sequence of vertices. They follow grid lines in regular regions of the mesh. When a thread passes through an extraordinary vertex of valence n, it continues by exiting along the edge $\lfloor n/2 \rfloor$-edges away in the clockwise direction. If $p_1, p_2,$ and p_3 are three points along a virtual thread, the anti-bending component of the energy is given by $k_p E_p(p_1, p_2, p_3)$ where

$$E_p(p_1, p_2, p_3) = \frac{1}{2} * \left(\frac{|p_3 - p_2|}{|p_3^* - p_2^*|} - \frac{|p_2 - p_1|}{|p_2^* - p_1^*|} \right)^2 \quad (6)$$

and $p_1^*, p_2^*,$ and p_3^* are the rest positions of the three points.

By adjusting k_s, k_d and k_p both globally and locally, we have been able to simulate a reasonably wide variety of cloth behavior. In the production of *Geri's game*, we found that Geri's jacket looked a great deal more realistic when we modulated k_p over the surface of the jacket in order to provide more stiffness on the shoulder pads, on the lapels, and in an area under the armpits which is often reinforced in real jackets. Methods for specifying scalar fields like k_p over a subdivision surface are discussed in more detail in section 5.1.

4.2 Collisions

The simplest approach to detecting collisions in a physical simulation is to test each geometric element (i.e. point, edge, face) against each other geometric element for a possible collision. With N geometric elements, this would take N^2 time, which is prohibitive for large N. To achieve practical running times for large simulations, the number of possible collisions must be culled as rapidly as possible using some type of spatial data structure. While this can be done in a variety of different ways, there are two basic strategies: we can distribute the elements into a two-dimensional surface-based data structure, or we can distribute them into a three-dimensional volume-based data structure. Using a two-dimensional structure has several advantages if the surface connectivity does not change. First, the hierarchy can be fixed, and need not be regenerated each time the geometry is moved. Second, the storage can all be statically allocated. Third, there is never any need to rebalance the tree. Finally, very short edges in the surface need not give rise to deep branches in the tree, as they would using a volume-based method.

It is a simple matter to construct a suitable surface-based data structure for a NURBS surface. One method is to recursively subdivide the (s, t) parameter plane into an quadtree. Since each node in the quadtree represents a subsquare of the parameter plane, a bounding box for the surface restricted to the subsquare can constructed. An efficient method for constructing the hierarchy of boxes is to compute bounding boxes for the children using the convex hull property; parent bounding boxes can then be computed in a bottom up fashion by unioning child boxes. Having constructed the quadtree, we can find all patches within ϵ of a point p as follows. We start at the root of the quadtree and compare the bounding box of the root node with a box of size 2ϵ centered on p. If there is no intersection, then there are no patches within ϵ of p. If there is an intersection, then we repeat the test on each of the children and recurse. The recursion terminates at the leaf nodes of the quadtree, where bounding boxes of individual subpatches are tested against the box around p.

Subdivision meshes have a natural hierarchy for levels finer than the original unsubdivided mesh, but this hierarchy is insufficient because even the unsubdivided mesh may have too many faces to test exhaustively. Since there is there is no global (s, t) plane from which to derive a hierarchy, we instead construct a hierarchy by "unsubdividing" or "coarsening" the mesh: We begin by forming leaf nodes of the hierarchy, each of which corresponds to a face of the subdivision surface control mesh. We then hierarchically merge faces level by level until we finish with a single merged face corresponding to the entire subdivision surface.

The process of merging faces proceeds as follows. In order to create the ℓth level in the hierarchy, we first mark all non-boundary edges in the $\ell - 1$st level as candidates for merging. Then until all candidates at the ℓth level have been exhausted, we pick a candidate edge e, and remove it from the mesh, thereby merging the two faces f_1 and f_2 that shared e into a "superface" f^*. The hierarchy is extended by creating a new node to represent f^*, the children of which are set to be the nodes corresponding to f_1 and f_2. If f^* were to immediately participate in another merge the hierarchy could quickly become poorly balanced. To ensure that the hierarchy is well balanced, we next remove all edges of f^* from the candidate list. When all the candidate edges at one level have been exhausted, we begin the next level by marking non-boundary edges as candidates. Hierarchy construction halts when only a single superface remains in the mesh.

The coarsening hierarchy is constructed once in a preprocessing phase. During each iteration of the simulation, control vertex positions change so the bounding boxes stored in the hierarchy must be updated. Updating the boxes is again a bottom up process: the current control vertex positions are used to update the bounding boxes at the leaves of the hierarchy. We do this efficiently by storing with each leaf in the hierarchy a set of pointers to the vertices used to construct its bounding box. Bounding boxes are then unioned up the hierarchy. A point can be "tested against" a hierarchy to find all faces within ϵ of the point by starting at the root of the hierarchy and recursively testing bounding boxes, just as is done with the NURBS quadtree.

We build a coarsening hierarchy for each of the cloth meshes, as well as for each of the kinematic obstacles. To determine collisions between a cloth mesh and a kinematic obstacle, we test each vertex of the cloth mesh against the hierarchy for the obstacle. To determine collisions between a cloth mesh and itself, we test each vertex of the mesh against the hierarchy for the same mesh.

5 Rendering subdivision surfaces

In this section, we introduce the idea of smoothly varying scalar fields defined over subdivision surfaces and show how they can be used to apply parametric textures to subdivision surfaces. We then describe a collection of implementation issues that arose when subdivision surfaces and scalar fields were added to RenderMan.

5.1 Texturing using scalar fields

NURBS surfaces are textured using four principal methods: parametric texture mapping, procedural texture, 3D paint [9], and solid texture [12, 13]. It is straightforward to apply 3D paint and solid texturing to virtually any type of primitive, so these techniques can readily be applied to texture subdivision surfaces. It is less clear, however, how to apply parametric texture mapping, and more

generally, procedural texturing to subdivision surfaces since, unlike NURBS, they are not defined parametrically.

With regard to texture mapping, subdivision surfaces are more akin to polygonal models since neither possesses a global (s,t) parameter plane. The now-standard method of texture mapping a polygonal model is to assign texture coordinates to each of the vertices. If the faces of the polygon consist only of triangles and quadrilaterals, the texture coordinates can be interpolated across the face of the polygon during scan conversion using linear or bilinear interpolation. Faces with more than four sides pose a greater challenge. One approach is to pre-process the model by splitting such faces into a collection of triangles and/or quadrilaterals, using some averaging scheme to invent texture coordinates at newly introduced vertices. One difficulty with this approach is that the texture coordinates are not differentiable across edges of the original or pre-processed mesh. As illustrated in Figures 11(a) and (b), these discontinuities can appear as visual artifacts in the texture, especially as the model is animated.

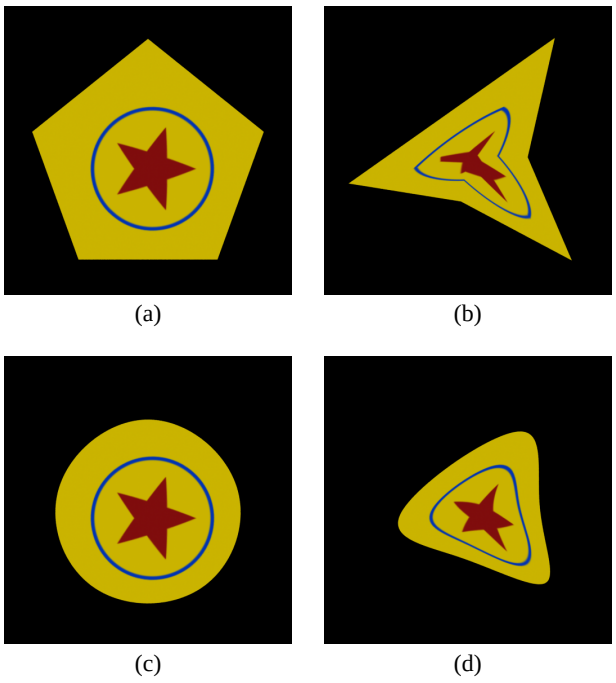


Figure 11: (a) A texture mapped regular pentagon comprised of 5 triangles; (b) the pentagonal model with its vertices moved; (c) A subdivision surface whose control mesh is the same 5 triangles in (a), and where boundary edges are marked as creases; (d) the subdivision surface with its vertices positioned as in (b).

Fortunately, the situation for subdivision surfaces is profoundly better than for polygonal models. As we prove in Appendix C, smoothly varying texture coordinates result if the texture coordinates (s,t) assigned to the control vertices are subdivided using the same subdivision rules as used for the geometric coordinates (x,y,z). (In other words, control point positions and subdivision can be thought of as taking place in a 5-space consisting of (x,y,z,s,t) coordinates.) This is illustrated in Figure 11(c), where the surface is treated as a Catmull-Clark surface with infinitely sharp boundary edges. A more complicated example of parametric texture on a subdivision surface is shown in Figure 12.

As is generally the case in real productions, we used a combination of texturing methods to create Geri: the flesh tones on his head and hands were 3D-painted, solid textures were used to add fine detail to his skin and jacket, and we used procedural texturing (described more fully below) for the seams of his jacket.

The texture coordinates s and t mentioned above are each instances of a scalar field; that is, a scalar-valued function that varies over the surface. A scalar field f is defined on the surface by assigning a value f_v to each of the control vertices v. The proof sketch in Appendix C argues that the function $f(p)$ created through subdivision (where p is a point on the limit surface) varies smoothly wherever the subdivision surface itself is smooth.

Scalar fields can be used for more than just parametric texture mapping — they can be used more generally as arbitrary parameters to procedural shaders. An example of this occurs on Geri's jacket. A scalar field is defined on the jacket such that it takes on large values for points on the surface near a seam, and small values elsewhere. The procedural jacket shader used the value of the this field to add the apparent seams to the jacket. We used other scalar fields to darken his nostril and ear cavities, and to modulate various physical parameters of the cloth in the cloth simulator.

We assign scalar field values to the vertices of the control mesh in a variety of ways, including manual assignment, smoothing, and painting. By smoothing we mean that the field values f_v are manually specified at a subset of the control points. The remaining field values are then determined using Laplacian smoothing. An example of relatively careful manual assignment and smoothing is illustrated in Figure 12. By painting we mean the specification of a scalar field by painting an intensity map on one or more rendered images of the surface. We then use a least squares solver to determine the field values that best reproduce the painted intensities.

5.2 Implementation issues

We have implemented subdivision surfaces, specifically semi-sharp Catmull-Clark surfaces, as a new geometric primitive in RenderMan.

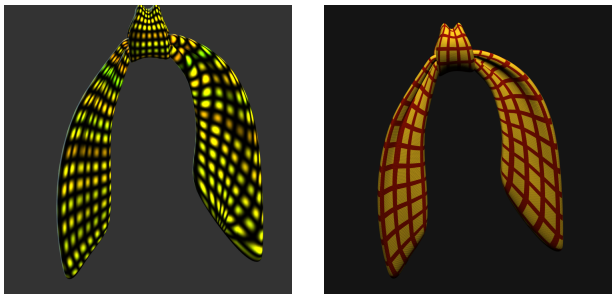
Our renderer, built upon the REYES architecture [4], demands that all primitives be convertible into grids of micropolygons (i.e. half-pixel wide quadrilaterals). Consequently, each type of primitive must be capable of splitting itself into a collection of subpatches, bounding itself (for culling and bucketing purposes), and dicing itself into a grid of micropolygons.

Each face of a Catmull-Clark control mesh can be associated with a patch on the surface, so the first step in rendering a Catmull-Clark surface is to split it in into a collection of individual patches. The control mesh for each patch consists of a face of the control mesh together with neighboring faces and their vertices. To bound each patch, we use the knowledge that a Catmull-Clark surface lies within the convex hull of its control mesh. We therefore take the bounding box of the mesh points to be the bounding box for the patch. Once bounded, the primitive is tested to determine if it is diceable; it is not diceable if dicing would produce a grid with too many micropolygons or a wide range of micropolygon sizes. If the patch is not diceable, then we split each patch by performing a subdivision step to create four new subpatch primitives. If the patch is diceable, it is repeatedly subdivided until it generates a grid with the required number of micropolygons. Finally, we move each of the grid points to its limit position using the method described in Halstead *et. al.* [8].

An important property of Catmull-Clark surfaces is that they give rise to bicubic B-splines patches for all faces except those in the neighborhood of extraordinary points or sharp features. Therefore, at each level of splitting, it is often possible to identify one or more subpatches as B-spline patches. As splitting proceeds, more of the surface can be covered with B-spline patches. Exploiting this fact has three advantages. First, the fixed 4×4 size of a B-spline patch allows for efficiency in memory usage because there is no need to store information about vertex connectivity. Second, the fact that a B-spline patch, unlike a Catmull-Clark patch, can be

split independently in either parametric direction makes it possible to reduce the total amount of splitting. Third, efficient and well understood forward differencing algorithms are available to dice B-spline patches [7].

We quickly learned that an advantage of semi-sharp creases over infinitely sharp creases is that the former gives smoothly varying normals across the crease, while the latter is guaranteed not to. This implies that if the surface is displaced in the normal direction in a creased area, it will tear at an infinitely sharp crease but not at a semi-sharp one.



(a) (b)

Figure 12: Gridded textures mapped onto a bandanna modeled using two subdivision surfaces. One surface is used for the knot, the other for the two flaps. In (a) texture coordinates are assigned uniformly on the right flap and nonuniformly using smoothing on the left to reduce distortion. In (b) smoothing is used on both sides and a more realistic texture is applied.

6 Conclusion

Our experience using subdivision surfaces in production has been extremely positive. The use of subdivision surfaces allows our model builders to arrange control points in a way that is natural to capture geometric features of the model (see Figure 2), without concern for maintaining a regular gridded structure as required by NURBS models. This freedom has two principal consequences. First, it dramatically reduces the time needed to plan and build an initial model. Second, and perhaps more importantly, it allows the initial model to be refined locally. Local refinement is not possible with a NURBS surface, since an entire control point row, or column, or both must be added to preserve the gridded structure. Additionally, extreme care must be taken either to hide the seams between NURBS patches, or to constrain control points near the seam to create at least the illusion of smoothness.

By developing semi-sharp creases and scalar fields for shading, we have removed two of the important obstacles to the use of subdivision surfaces in production. By developing an efficient data structure for culling collisions with subdivisions, we have made subdivision surfaces well suited to physical simulation. By developing a cloth energy function that takes advantage of Catmull-Clark mesh structure, we have made subdivision surfaces the surfaces of choice for our clothing simulations. Finally, by introducing Catmull-Clark subdivision surfaces into our RenderMan implementation, we have shown that subdivision surfaces are capable of meeting the demands of high-end rendering.

A Infinitely Sharp Creases

Hoppe *et. al.* [10] introduced infinitely sharp features such as creases and corners into Loop's surfaces by modifying the subdivision rules in the neighborhood of a sharp feature. The same can be done for Catmull-Clark surfaces, as we now describe.

Face points are always positioned at face centroids, independent of which edges are tagged as sharp. Referring to Figure 4, suppose the edge $v^i\ e^i_j$ has been tagged as sharp. The corresponding edge point is placed at the edge midpoint:

$$e^{i+1}_j = \frac{v^i + e^i_j}{2}. \tag{7}$$

The rule to use when placing vertex points depends on the number of sharp edges incident at the vertex. A vertex with one sharp edge is called a dart and is placed using the smooth vertex rule from Equation 2. A vertex v^i with two incident sharp edges is called a crease vertex. If these sharp edges are $e^i_j v^i$ and $v^i e^i_k$, the vertex point v^{i+1} is positioned using the crease vertex rule:

$$v^{i+1} = \frac{e^i_j + 6v^i + e^i_k}{8}. \tag{8}$$

The sharp edge and crease vertex rules are such that an isolated crease converges to a uniform cubic B-spline curve lying on the limit surface. A vertex v^i with three or more incident sharp edges is called a corner; the corresonding vertex point is positioned using the corner rule

$$v^{i+1} = v^i \tag{9}$$

meaning that corners do not move during subdivision. See Hoppe *et. al.* [10] and Schweitzer [15] for a more complete discussion and rationale for these choices.

Hoppe *et. al.* found it necessary in proving smoothness properties of the limit surfaces in their Loop-based scheme to make further distinctions between so-called regular and irregular vertices, and they introduced additional rules to subdivide them. It may be necessary to do something similar to prove smoothness of our Catmull-Clark based method, but empirically we have noticed no anamolies using the simple strategy above.

B General semi-sharp creases

Here we consider the general case where a crease sharpness is allowed to be non-integer, and to vary along the crease. The following procedure is relatively simple and strictly generalizes the two special cases discussed in Section 3.

We specify a crease by a sequence of edges $e_1, e_2, ...$ in the control mesh, where each edge e_i has an associated sharpness $e_i.s$. We associate a sharpness per edge rather than one per vertex since there is no single sharpness that can be assigned to a vertex where two or more creases cross.[2]

During subdivision, face points are always placed at face centroids. The rules used when placing edge and vertex points are determined by examining edge sharpnesses as follows:

- An edge point corresponding to a smooth edge (i.e, $e.s = 0$) is computed using the smooth edge rule (Equation 1).
- An edge point corresponding to an edge of sharpness $e.s >= 1$ is computed using the sharp edge rule (Equation 7).
- An edge point corresponding to an edge of sharpness $e.s < 1$ is computed using a blend between smooth and sharp edge rules: specifically, let v_{smooth} and v_{sharp} be the edge points computed using the smooth and sharp edge rules, respectively. The edge point is placed at

$$(1 - e.s)v_{smooth} + e.s v_{sharp}. \tag{10}$$

[2] In our implementation we do not allow two creases to share an edge.

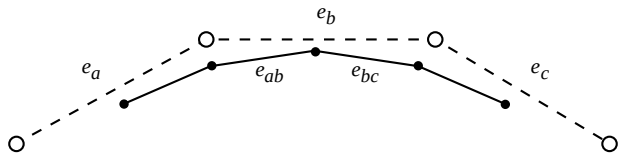


Figure 13: Subedge labeling.

- A vertex point corresponding to a vertex adjacent to zero or one sharp edges is computed using the smooth vertex rule (Equation 2).

- A vertex point corresponding to a vertex v adjacent to three or more sharp edge is computed using the corner rule (Equation 9).

- A vertex point corresponding to a vertex v adjacent to two sharp edges is computed using the crease vertex rule (Equation 8) if $v.s \geq 1$, or a linear blend between the crease vertex and corner masks if $v.s < 1$, where $v.s$ is the average of the incidence edge sharpnesses.

When a crease edge is subdivided, the sharpnesses of the resulting subedges is determined using Chaikin's curve subdivision algorithm [3]. Specifically, if e_a, e_b, e_c denote three adjacent edges of a crease, then the subedges e_{ab} and e_{bc} as shown in Figure 13 have sharpnesses

$$\begin{aligned} e_{ab}.s &= \max(\frac{e_a.s + 3e_b.s}{4} - 1, 0) \\ e_{bc}.s &= \max(\frac{3e_b.s + e_c.s}{4} - 1, 0) \end{aligned}$$

A 1 is subtracted after performing Chaikin's averaging to account for the fact that the subedges (e_{ab}, e_{bc}) are at a finer level than their parent edges (e_a, e_b, e_c). A maximum with zero is taken to keep the sharpnesses non-negative. If either e_a or e_b is infinitely sharp, then e_{ab} is; if either e_b or e_c is infinitely sharp, then e_{bc} is. This relatively simple procedure generalizes cases 1 and 2 described in Section 3. Examples are shown in Figures 9 and 10.

C Smoothness of scalar fields

In this appendix we wish to sketch a proof that a scalar field f is smooth as a function on a subdivision surface wherever the surface itself is smooth. To say that a function on a smooth surface S is smooth to first order at a point p on the surface is to say that there exists a parametrization $S(s,t)$ for the surface in the neighborhood of p such that $S(0,0) = p$, and such that the function $f(s,t)$ is differentiable and the derivative varies continuously in the neighborhood of $(0,0)$.

The characteristic map, introduced by Reif [14] and extended by Zorin [19], provides such a parametrization: the characteristic map allows a subdivision surface S in three space in the neighborhood of a point p on the surface to be written as

$$S(s,t) = (x(s,t), y(s,t), z(s,t)) \tag{11}$$

where $S(0,0) = p$ and where each of $x(s,t)$, $y(s,t)$, and $z(s,t)$ is once differentiable if the surface is smooth at p. Since scalar fields are subdivided according to the same rules as the x, y, and z coordinates of the control points, the function $f(s,t)$ must also be smooth.

Acknowledgments

The authors would like to thank Ed Catmull for creating the *Geri's game* project, Jan Pinkava for creating Geri and for writing and directing the film, Karen Dufilho for producing it, Dave Haumann and Leo Hourvitz for leading the technical crew, Paul Aichele for building Geri's head, Jason Bickerstaff for modeling most of the rest of Geri and for Figure 10, and Guido Quaroni for Figure 12. Finally, we'd like to thank the entire crew of *Geri's game* for making our work look so good.

References

[1] David E. Breen, Donald H. House, and Michael J. Wozny. Predicting the drape of woven cloth using interacting particles. In Andrew Glassner, editor, *Proceedings of SIGGRAPH '94 (Orlando, Florida, July 24–29, 1994)*, Computer Graphics Proceedings, Annual Conference Series, pages 365–372. ACM SIGGRAPH, ACM Press, July 1994. ISBN 0-89791-667-0.

[2] E. Catmull and J. Clark. Recursively generated B-spline surfaces on arbitrary topological meshes. *Computer Aided Design*, 10(6):350–355, 1978.

[3] G. Chaikin. An algorithm for high speed curve generation. *Computer Graphics and Image Processing*, 3:346–349, 1974.

[4] Robert L. Cook, Loren Carpenter, and Edwin Catmull. The Reyes image rendering architecture. In Maureen C. Stone, editor, *Computer Graphics (SIGGRAPH '87 Proceedings)*, pages 95–102, July 1987.

[5] Martin Courshesnes, Pascal Volino, and Nadia Magnenat Thalmann. Versatile and efficient techniques for simulating cloth and other deformable objects. In Robert Cook, editor, *SIGGRAPH 95 Conference Proceedings*, Annual Conference Series, pages 137–144. ACM SIGGRAPH, Addison Wesley, August 1995. held in Los Angeles, California, 06-11 August 1995.

[6] Nira Dyn, David Leven, and John Gregory. A butterfly subdivision scheme for surface interpolation with tension control. *ACM Transactions on Graphics*, 9(2):160–169, April 1990.

[7] James D. Foley, Andries van Dam, Steven K. Feiner, and John F. Hughes. *Computer Graphics: Principles and Practice*. Prentice-Hall, 1990.

[8] Mark Halstead, Michael Kass, and Tony DeRose. Efficient, fair interpolation using Catmull-Clark surfaces. *Computer Graphics*, 27(3):35–44, August 1993.

[9] Pat Hanrahan and Paul E. Haeberli. Direct WYSIWYG painting and texturing on 3D shapes. In Forest Baskett, editor, *Computer Graphics (SIGGRAPH '90 Proceedings)*, volume 24, pages 215–223, August 1990.

[10] H. Hoppe, T. DeRose, T. Duchamp, M. Halstead, H. Jin, J. McDonald, J. Schweitzer, and W. Stuetzle. Piecewise smooth surface reconstruction. *Computer Graphics*, 28(3):295–302, July 1994.

[11] Charles T. Loop. Smooth subdivision surfaces based on triangles. Master's thesis, Department of Mathematics, University of Utah, August 1987.

[12] Darwyn R. Peachey. Solid texturing of complex surfaces. In B. A. Barsky, editor, *Computer Graphics (SIGGRAPH '85 Proceedings)*, volume 19, pages 279–286, July 1985.

[13] Ken Perlin. An image synthesizer. In B. A. Barsky, editor, *Computer Graphics (SIGGRAPH '85 Proceedings)*, volume 19, pages 287–296, July 1985.

[14] Ulrich Reif. A unified approach to subdivision algorithms. Mathematisches Institute A 92-16, Universitaet Stuttgart, 1992.

[15] Jean E. Schweitzer. *Analysis and Application of Subdivision Surfaces*. PhD thesis, Department of Computer Science and Engineering, University of Washington, 1996.

[16] Demetri Terzopoulos, John Platt, Alan Barr, and Kurt Fleischer. Elastically deformable models. In Maureen C. Stone, editor, *Computer Graphics (SIGGRAPH '87 Proceedings)*, volume 21, pages 205–214, July 1987.

[17] Steve Upstill. *The RenderMan Companion*. Addison-Wesley, 1990.

[18] Andrew Witkin, David Baraff, and Michael Kass. An introduction to physically based modeling. SIGGRAPH Course Notes, Course No. 32, 1994.

[19] Denis Zorin. *Stationary Subdivision and Multiresolution Surface Representations*. PhD thesis, Caltech, Pasadena, 1997.