From functional animation to sprite-based display (Expanded Version)

Conal Elliott
http://www.research.microsoft.com/~conal

October 23, 1998

Technical Report
MSR-TR-98-28

Microsoft Research
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052

A shorter version of this report will appear in the proceedings of First International Workshop on Practical Aspects of Declarative Languages (PADL'99), and will be © Springer-Verlag.

From functional animation to sprite-based display (Expanded Version)

Conal Elliott
http://www.research.microsoft.com/~conal

October 23, 1998

Abstract

Functional animation encourages a highly modular programming style, by supplying a set of arbitrarily composable functions for building up animations. In contrast, libraries for sprite-based display impose rigid structure, in order to allow acceleration by hardware and low level software. This paper presents a method to bridge the gap between functional specification and stateful, sprite-based presentation of animation. The method's correctness is proved informally by derivation from a simple non-effective specification, exploiting algebraic properties of the animation data types that are made explicit in the functional approach. We have implemented this method in the *Fran* system, which is freely available.

1 Introduction

The functional approach to animation offers considerable flexibility and modularity [1, 5]. Animations are first-class values—elements of a data type consisting of a set of constants and combining operators. The data type allows great flexibility in composing these basic building blocks into either directly useful or attractive animations, or new building blocks, parameterized as desired. Moreover, animation is a polymorphic notion, applying to 2D images, 3D geometry, and consituent types like colors, points, vectors, numbers, booleans, etc. Consequently, there is not just one animation type, but a collection of types and type constructors. In a well-designed set of data types, the type system imposes just enough discipline to rule out nonsensical compositions (such as rotating by the angle "purple"), without inhibiting the author's creativity. In this way, the data types are designed to serve the needs of the author (and readers) of an animation.

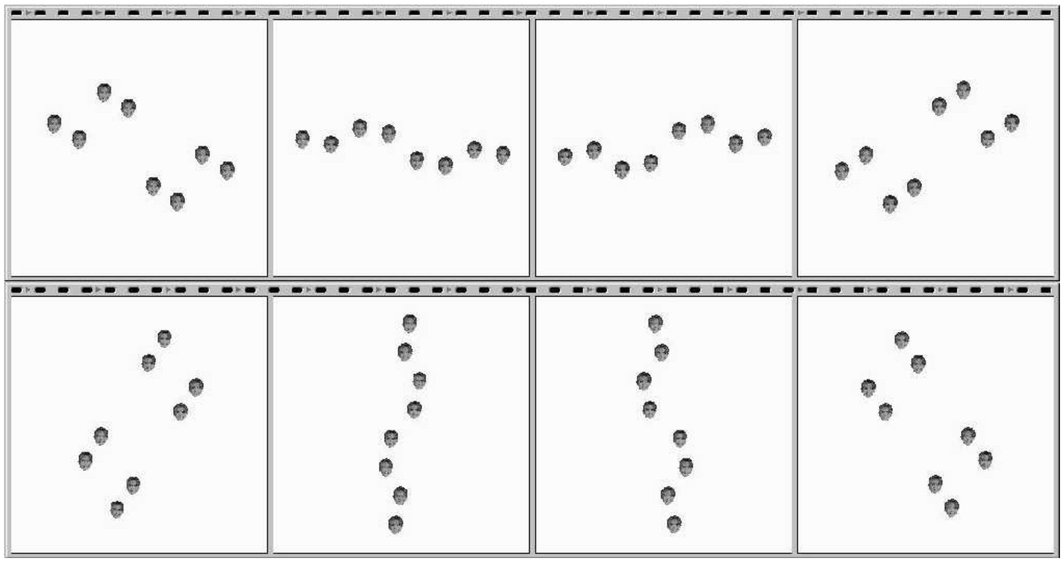


Figure 1: `repSpinner (-1) 0.5 charlotte 3`

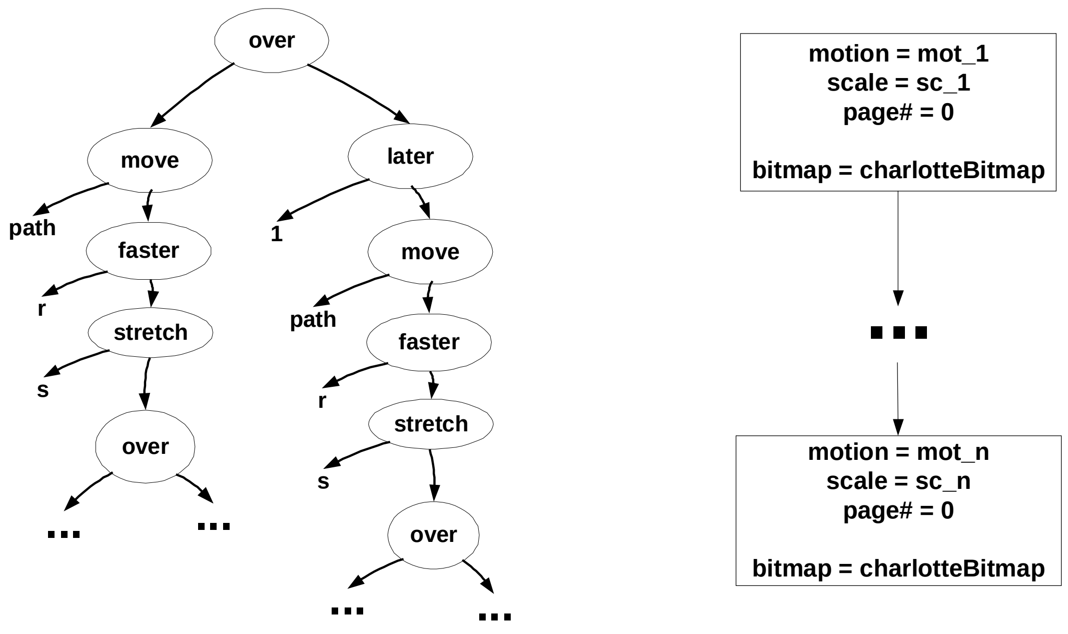
Lower level graphics presentation libraries are designed not for convenience of a program's author or readers, but rather for efficient execution on anticipated hardware. Programs written directly on top of these libraries must adapt to relatively inflexible representations and tend to be relatively non-modular.

To illustrate the convenience of functional animation, consider the animation in Figure 1.[1] The function `repSpinner` repeatedly transforms an animation `im`, stretching `im` by `r`, speeding it up by `s`, and putting two copies into circular orbit.[2]

```
repSpinner :: RealB -> RealB -> ImageB -> Int -> ImageB
repSpinner r s im n = iterate spinner im !! n
 where
   spinner im = orbit `over` later 1 orbit
    where
      orbit = move path (faster r (stretch s im))
      path  = vector2Polar 0.5 (pi*time)
```

[1]`ftp://ftp.research.microsoft.com/pub/tr/tr-98-28/animations.htm` contains running versions of this example and a few others. Those versions were recorded at 20 fps (frames per second), but run in Fran at 60 fps.

[2]`iterate f x` produces the infinite list `[x, f x, f (f x), ...]`, and `l!!n` extracts the `n`-th element of the list `l`. Their use together here results in applying `spinner` to `im`, `n` times.



(a) Image hierarchy (b) Sprite list

Figure 2: Image hierarchies vs. sprite lists

In this example, time- and space-transforming functions (`faster` and `stretch`) are applied to `over`lays of animations that contain more time- and space-transformed animations (when $n > 1$), as shown in Figure 2(a). In contrast, sprite-based subsystems impose rigid structuring on an animation. It must be a sequence of implicitly overlaid "sprites", where each sprite is the result of applying possibly time-varying motion and scaling to a "flip-book" of bitmaps, as in Figure 2(b). (In our example, the flip-book has only one page.) It is tedious and error-prone for a programmer or artist to work with such restrictions, when the animated objects of interest have natural hierarchical structure, as in `repSpinner`. The required sprite motion and scaling vectors (`mot_`*i* and `sc_`*i*) are quite complex, because each one is affected by a number of space- and time-transforms above them in the original hierarchy.

This paper presents an algorithm to convert functional animations into a "sprite normal form" amenable to a fast sprite-based presentation library. As such, it bridges the gap between the high level flexibility desired for specification of animations and the low level regularity imposed by the presentation library. This algorithm is used in the implementation of *Fran* ("Functional Reactive Animation") [5, 3, 4], which is written in the purely functional language Haskell [8, 7] and runs in conjunction with a fast "temporal sprite engine," implemented in C++ and running in a separate thread. The sprite engine, which is described in this paper, manages a list of sprites. While some sprites are rendered on the fly, the sprite engine also supports fast flip-book animation. Typically, Fran updates information maintained by the sprite engine ten times per second, while the sprite engine interpolates motion, scaling, and indexing, and updates the screen, all at sixty times per second.

In order to transform flexibly specified animations into the form required by the sprite engine, we exploit several algebraic properties. In addition to time- and space-transformation, Fran's animation algebra supports cropping, conditionals and reactivity, each requiring special care for conversion to sprite normal form.

Fran is freely available, including the sprite engine, in full source code at

`http://www.research.microsoft.com/~conal/fran`

2 Animation data types

Fran's animation data types have been described elsewhere (e.g., [5], and for reference [13]). The animation-level types other than animated images, sounds, and 3D geometry come from applying the `Behavior` type constructor to these static types, and lifting operations on the static types to operations on behaviors. Behaviors are functions of continuous time. As a convention, the behavior types have synonyms made by adding a "`B`" to the static type name, e.g., "`Transform2B`" is a synonym for "`Behavior Transform2`".

As an illustrative example, and because it is particularly relevant to the sprite-based implementation of 2D animation, we will describe one such type next.

2.1 Static 2D transforms

The `Transform2` type represents 2D geometric transformation on images, points, or vectors. It also defines a type class `Transformable2` containing types of 2D transformable objects.

```
module Transform2 where

identity2  :: Transform2
translate2 :: Vector2 -> Transform2
rotate2    :: RealVal -> Transform2
uscale2    :: RealVal -> Transform2  -- only uniform scaling
compose2   :: Transform2 -> Transform2 -> Transform2
inverse2   :: Transform2 -> Transform2

class Tranformable2 a where
```

```
    (*%)    :: Transform2 -> a -> a   -- Applies a transform

  instance Transformable2B Point2B
  instance Transformable2B Vector2B
```

Intuitively, 2D transforms are mappings from 2D space to itself. Because of the restricted vocabulary above, however, one can take advantage of widely available low level rendering algorithms and hardware acceleration.

2.2 Behaviors

The type `Behavior` α represents time-varying values of type α, where time is continuous (real-valued). See [5] for the semantics of behaviors, but we will mention here a new formulation of formulation of the some of the primitives.

```
  ($*)      :: Behavior (a -> b) -> Behavior a -> Behavior b
  constantB :: a -> Behavior a
  time      :: Behavior Time
```

Semantically, these operators correspond to type-specialized versions of the classic SKI combinators.

```
-- Semantics
type Behavior a = Time -> a
type Time = RealVal
(fb $* xb)  t = (fb t) (xb t)  -- S
constantB x t = x              -- K
time        t = t              -- I
```

The name "`$*`" comes from the fact that it is the lifted version of function application, whose Haskell infix operator is called "`$`".

The lifting operators from [5] are

```
  lift0 :: a -> Behavior a
  lift1 :: (a -> b) ->
           Behavior a -> Behavior b
  lift2 :: (a -> b -> c) ->
           Behavior a -> Behavior b -> Behavior c
  -- etc
```

and are defined simply in terms of `constantB` and "`$*`", as follows. (Note that semantically `lift1` is the classic B combinator.)

```
  lift0            = constantB
  lift1 f b1       = lift0 f $* b1
  lift2 f b1 b2    = lift1 f b1 $* b2
  lift3 f b1 b2 b3 = lift2 f b1 b2 $* b3
  -- etc
```

Thus lifting promotes an n-ary function to an n-ary function. Semantically, from the definitions above, it follows that

```
  (liftn f b1 ... bn) t == f (b1 t) ... (bn t)
```

Efficient implementation of behaviors is a rather tricky matter and is discussed is a separate paper [2]. Briefly, the current representation of behaviors is as a data structure that contains two aspects. The first is a structural representation intended for analysis and optimization. The second is a sampling representation, which is a lazy memo function [9] from time streams to value streams.[3]

```
data Behavior a = Behavior (BStruct a) ([T] -> [a])

data BStruct a
   = ConstantB a
   | NoStructureB
   | TimeTransB (Behavior a) (Behavior Time)
   | UntilB (Behavior a) (Event (Behavior a))
  deriving Show
```

It is not currently possible to capture the SKI structure in a Haskell datatype, because doing so would require an existential type for S (i.e., "`$*`"), but fortunately, existential types are in the process of being added to Haskell implementations. (Another trick is needed for I – i.e., `time`. While S is too polymorphic, I is not polymorphic enough. The solution to this problem is to generalize the constructor to be polymorphic, and then define the desired specialization.[4])

In addition to behaviors, there are some other "behavior-like" types, described by the following type class of "generalized behaviors".[5]

```
class GBehavior bv where
  untilB        :: bv -> Event bv -> bv
  timeTransform :: bv -> TimeB    -> bv
  condBUnOpt    :: BoolB -> bv    -> bv    -> bv
```

These operations support reactivity, time transformation, and conditional behaviors, respectively. The last of these is an unoptimized conditional, from which the optimized conditional is defined as follows.

```
condB :: GBehavior bv => BoolB -> bv -> bv -> bv
condB (Behavior (ConstantB True ) _) imb _    = imb
condB (Behavior (ConstantB False) _) _   imb' = imb'
condB c imb imb' = condBUnOpt c imb imb'
```

Naturally, behaviors are in the `GBehavior` class.

```
instance GBehavior (Behavior a) where
  condBUnOpt = lift3 (\ a b c -> if a then b else c)
  ...
```

Most animated types are defined very simply, by lifting. For instance, here is the type `Transform2B` of animated 2D transforms.

[3]The representation given here is somewhat simplified from the one used in Fran. See [2] for details.

[4]My thanks to Alastair Reid and Mark Jones for pointing out this trick.

[5]By convention, the name of the lifted version of a type is formed by adding a "`B`" to the end of the unlifted type's name. Thus the types `TimeB` and `BoolB` mentioned in the `GBehavior` type class refer to time- and boolean-valued behaviors respectively.

```
emptyImage  :: ImageB                 -- transparent everywhere
solidImage  :: ImageB                 -- solid color image
flipImage   :: HFlipBook -> RealB -> ImageB -- flipbook-based
renderImage :: Renderer -> ImageB     -- text, 2D & 3D geometry
soundImage  :: SoundB -> ImageB            -- embedded sound
over        :: ImageB -> ImageB -> ImageB  -- overlay
withColor   :: ColorB -> ImageB -> ImageB  -- colored image
crop        :: RectB  -> ImageB -> ImageB  -- cropped image

(*%)        :: Transform2B -> ImageB -> ImageB  -- apply 2D transform
untilB      :: ImageB -> Event ImageB -> ImageB     -- reactive
condB       :: BoolB -> ImageB -> ImageB -> ImageB  -- conditional
timeTransform :: ImageB -> TimeB -> ImageB
```

Figure 3: Abstract interface to the `ImageB` type

```
module Transform2B where

import qualified Transform2 as T

type Transform2B = Behavior T.Transform2

identity2  = lift0 T.identity2
translate2 = lift1 T.translate2
rotate2    = lift1 T.rotate2
uscale2    = lift1 T.uscale2
compose2   = lift2 T.compose2
inverse2   = lift1 T.inverse2

class Transformable2B a where
  (*%)  ::  Transform2B -> a -> a

instance T.Transformable2 a => Transformable2B (Behavior a) where
  (*%) =  lift2 (T.*%)
```

2.3 Image animation

The type `ImageB` represents image animations that are spatially and temporally continuous and infinite. The primitives used in the construction of `ImageB` values are shown in Figure 3.[6]

The `renderImage` function is used for all of the various kind of synthetic images, based on text, 2D geometry, 3D geometry, etc. The type `Renderer` maps time-varying cropping rectangle, color, scale factor, rotation angle, and time transform to a time-varying bitmap:

```
type Renderer = RectB -> Maybe ColorB -> RealB -> RealB
```

[6]The type `Event a`, for an arbitrary type `a`, represents a stream of time-stamped values of type `a`, and is the basis of reactivity in behaviors.

```
   (crop rectB           $
    mbWithColor mbColorB $
    stretch scaleB       $
    turn    angleB       $
    (`timeTransform` tt) $
    RenderImage renderer )
 ==
    surfaceIm (renderer rectB mbColorB scaleB angleB tt)

  -- Possibly apply a color
  mbWithColor :: Maybe ColorB -> ImageB -> ImageB
  mbWithColor Nothing  imB = imB
  mbWithColor (Just c) imB = withColor c imB

  -- The ImageB contained on a discrete image (for specification).
  surfaceIm :: SurfaceULB -> ImageB
```

Figure 4: Property of a "renderer"

```
              -> TimeB -> SurfaceULB

  -- Bitmap plus upper-left corner location
  data SurfaceUL  = SurfaceUL HDDSurface Point2
  type SurfaceULB = Behavior SurfaceUL
```

Displaying an animated image also plays the sounds it contains. Spatial transformation has audible effect; the horizontal component of translation becomes left/right balance, and the scale becomes volume adjustment. Cropping silences sounds outside of the crop region.

By definition, a *renderer* not only has the type above, but must also perform cropping, coloring, scaling, rotation, and time transformation, according to its given parameters, as expressed in Figure 4:[7]

A "flip book" is a sequence "pages", each of which is a bitmap. There is also a simple utility that loads a named file, makes a one-page flip-book and applies `flipImage`: `importBitmap :: String -> ImageB`

For notational convenience, Fran also provides a few functions for directly transforming `ImageB` values.

```
  move :: Transformable2B bv => Vector2B -> bv -> bv
  move dp thing = translate2 dp *% thing

  stretch :: Transformable2B bv => RealB -> bv -> bv
  stretch sc thing = uscale2 sc *% thing
```

[7]The "`$`" operator is an alternative notation for function application. Because it is right-associative and has low syntactic precedence, it is sometimes used to eliminate cascading parentheses. Also, an infix operator (here `` `timeTransform` ``) with one argument but missing the other denotes a function that takes the missing argument (here an `ImageB` value) and fills it in.

```
turn :: Transformable2B bv => RealB -> bv -> bv
turn angle thing = rotate2 angle *% thing
```

In the early implementations of Fran, `ImageB` was simply defined to be `Behavior Image`, for a type `Image` of static images. This representation made for a very simple implementation, but it has a fundamental problem: the image structure of an `ImageB` value cannot be determined at the behavior level. It must first be sampled with some time t to extract a static image, whose structure can then be examined. To display an animation then, one must repeatedly sample it, and display the resulting static images. Consequently, the display computation cannot build and maintain data structures and system resources for parts of an animation. In particular, it cannot allocate one sprite for each `flipImage` and `renderImage`, set them moving, and then update the motion paths incrementally. That is, the implementation cannot take advantage of a sprite engine.

It appears then that the modularity imposed by lifting requires an underlying presentation library to work in "immediate mode" (draw this now), rather than "retained mode" (build a model and repeatedly edit and redisplay it), to borrow terms from 3D graphics programming. In designing Fran, however, we targeted the upcoming generation of graphics accelerator cards, which we believe to be increasingly oriented toward retained mode. The sprite-based Talisman architecture [16] was of special interest. It performs transformation and overlaying of sprites with α-blending and high-quality anisotropic image filtering, without requiring a screen-size frame buffer. In fact, we designed and implemented Fran's underlying sprite engine to be a reasonable approximation to a partial Talisman interface, before real cards were available.

For reasons given above, the `ImageB` is not represented in terms of a static `Image` type, but rather as a recursively defined data type, as shown in Figure 5. The functions in Figure 3 are defined as simple optimizations of the `ImageB` constructors from Figure 5. For instance, transforming the empty or solid image has no effect:

```
xf *% EmptyImage = EmptyImage
xf *% SolidImage = SolidImage
xf *% im         = TransformI xf im
```

Although a programmer would not be likely to transform empty or solid images explicitly, such compositions arise at runtime due to modular programming style, as well as some of the reactivity optimizations discussed in [2]. The other `ImageB` operators are defined similarly. The constructors `CondI`, `UntilI`, and `TimeTransI` are used to define the overloaded functions `condB`, `untilB`, and `timeTransform` that apply to sound, 3D geometry, and all behaviors, as well as to image animations. The "`*%`" operator is overloaded to apply to several types, as well.

Image animations are generalized behaviors:

```
data ImageB
 = EmptyImage                         -- transparent everywhere
 | SolidImage                         -- solid color
 | FlipImage  HFlipBook RealB         -- page # behavior
 | RenderImage Renderer               -- text, 2D & 3D geometry
 | SoundI     SoundB                  -- embedded sound
 | Over       ImageB   ImageB         -- overlay
 | TransformI Transform2B ImageB      -- 2D transformed
 | WithColorI ColorB ImageB           -- colored
 | CropI      RectB ImageB            -- cropped
 | CondI      BoolB ImageB ImageB     -- conditional
 | UntilI     ImageB (Event ImageB)   -- reactivity
 | TimeTransI ImageB  TimeB           -- time transformed
```

Figure 5: Data type representing 2D image animations

```
instance  GBehavior ImageB  where
  untilB        = UntilI
  timeTransform = TimeTransI
  condBUnOpt    = CondI
```

2.4 Sound

Fran has a fairly simple type `SoundB` of "animated sounds." Its building blocks are similar to those of `ImageB`.

```
module SoundB where

silence    :: SoundB
importWave :: String -> Bool   -> SoundB -- file name, repeat?
mix        :: SoundB -> SoundB -> SoundB
volume     :: RealB  -> SoundB -> SoundB
pitch      :: RealB  -> SoundB -> SoundB
pan        :: RealB  -> SoundB -> SoundB
```

Representation of `SoundB` via lifting is problematic both conceptually and pragmatically. Conceptually, it is unclear what is a useful and implementable notion of static sound. Pragmatically, as with `ImageB`, a lifting-based implementation would require an immediate mode presentation library, but the libraries we know of, such as Microsoft's DirectSound [12, 11] are retained mode. DirectSound provides interfaces to allocate sound buffers in either main memory or audio card memory, and direct the sound buffers to be mixed and played. The client program can then optionally adjust the sound buffers' volume, frequency, and left/right balance attributes. A retained mode architecture is especially desirable for sound, because it is difficult for an application program to keep the hardware's sound buffers filled and perform or direct mixing and output in a timely enough fashion to avoid buffer overflow or underflow, both of which have easily discernible effects. A lazy functional program would have particular difficulty because of unpredictable interruptions due to garbage collection and evaluation of postponed computations. DirectSound creates a few highest-priority threads that feed buffers and adjust registers in the sound hardware.

The representation of `SoundB` is as a recursive data type.

```
data SoundB = SilentS
            | BufferS HDSBuffer Bool  -- for importWave
            | MixS    SoundB SoundB
            | VolumeS RealB SoundB
            | PanS    RealB SoundB
            | PitchS  RealB SoundB
            | UntilS  SoundB (Event SoundB)
            | TimeTransS SoundB TimeB
  deriving Show
```

Animated sounds are generalized behaviors as well. Conditional is implemented by mixing two sounds with applied volumes that silence one or the other, depending on the boolean behavior.[8]

```
instance  GBehavior SoundB  where
  untilB        = UntilS
  timeTransform = TimeTransS
  condBUnOpt c snd snd' =
    volume v snd `mix` volume v' snd'
   where
     v  = condB c 1 0
     v' = 1 - v
```

2.5 3D animation

Many 3D graphics presentations libraries also have a retained-mode structure, e.g., Silicon Graphics' Inventor [15] and Performer [14], and Microsoft's Direct3DRM [6, 11]. Fran uses Direct3DRM to display its type `GeometryB` of animated 3D geometry. Its structure is much like `ImageB` and `SoundB`.

3 A temporal sprite engine

The primary purpose of the sprite engine is to scale, move, and overlay a sequence of images, and to do so at a very high and regular rate.[9] Because anima-

[8]With α-blending for transparency, an analogous implementation of conditional image animations would be possible. Alternatively, one could use `stretch`. Fran does not use this technique, because scaling an image moves all of its features closer to or further from the origin, and due to the sprite engine's automatic linear interpolation this motion is visible for 1/10 second. Fast α-blending would fix this problem.

[9]Ideally redisplay is done exactly at the video refresh rate and occurs during the video blank interval. We cannot attain this ideal, due to lack of real-time operating system support, and because the video card hardware interface does not generate interrupts to signal the beginning of the video blank interval. However, in practice we do fairly well.

tions may involve arbitrarily complex behaviors, and because garbage collection and lazy evaluation cause unpredictable delays, we could not meet this goal if the window refresh were implemented in Haskell or even invoked by a Haskell program. For these reasons, the sprite engine is implemented in C++, does no memory allocation, and runs in its own thread.

3.1 Sprites and sprite trees

The sprite engine maintains an ordered collection of sprites, represented via a C++ class hierarchy. The classes representing individual sprites are as follows.

- `FlipSprite` has a bitmap that is selected from a "flip book" object. Each flip book contains a pointer to a bitmap (a DirectDraw surface [12, 11]) stored in video memory, and information describing a rectangular array of images contained somewhere within the bitmap. This array is taken to be a linear sequence of consecutive "pages". Flip books and the bitmap images to which they refer are immutable and may be shared among any number of flip sprites. The current page number, spatial transform, and cropping rectangle are all stored in the unshared sprites rather than the shared flip book.

- `RenderedSprite` has its own drawing surface and a method for replacing the surface with a new one. This sprite is used for all "rendered" images, such as text, 2D geometry, and camera-viewed 3D geometry.

- `SolidSprite` is a uniformly colored sprite, cropped but not spatially transformed.

- `SoundSprite` is an embedded sound, and is neither cropped nor transformed. It contains a DirectSound "duplicate" sound buffer and volume, frequency, and left/right balance attributes. A duplicate buffer is a DirectSound object that contains a pointer to (typically immutable) shareable sound data, plus its own unshared attributes for volume, frequency, and left/right balance.

A displayed animation could be represented by a list of individual sprites, in back-to-front order, so that upper (later) sprites are painted over lower (earlier) sprites. This representation is the simplest for display, but is awkward for editing. When an event occurs in an animation, an animated object may be introduced, removed, or replaced. Since the appearance of such an object may contain any number of sprites, the event may introduce, remove, or replace a whole contiguous subsequence of sprites. The sprite engine makes it easy and fast to edit the sprite list at a conceptual level by using a list of *sprite trees* rather than a list of individual sprites. The internal nodes of these trees correspond exactly to the points of structural (as opposed to parametric) mutability of the animation being displayed. (For Fran's use, the points of mutability are generated from the reactivity construct “`untilB`”.) Correspondingly, there is a `SpriteTree` subclass `SpriteGroup` that contains a list of sprite trees and supports a method to replace the list, in its entirety, with another one. The sprite engine requires that there be no structure sharing, so the old list and all of its elements are recursively deallocated. Lack of sharing also allows direct linking of sprite trees. Each sprite tree contains a `next` pointer.

For conditional animations, such as generated by Fran’s “`condB`”, there is a `SpriteTree` subclass `CondSpriteTree`, which contains an externally mutable boolean and two sprite tree lists.

The sprite engine expects its sprite trees to be updated less often that they are redisplayed. For example, Fran tries to sample behaviors at roughly ten times per second, but the sprite engine redisplays at 60 times per second. Between updates, the sprite engine performs linear interpolation of all sprite attributes, which are are actually represented as linear functions rather than constant values. For smoothness, updates must then be given for times in the future, so that the a new linear path may be set from the current time and value to the given one. These linear behaviors form a compromise between arbitrary behaviors, which have unpredictable sampling requirements, and mere constant values, which fail to provide smooth motion. Other possibile compromises include quadratic and cubic curves, for which fast incremental sampling algorithms may be used.

The main loop of the sprite engine is then to traverse its sprite tree list recursively and in depth-first, back-to-front order. Each image sprite in turn is told to paint itself, and in doing so samples its attributes’ linear functions and then does a very fast video memory “blit” (copy). Sound sprites are “painted” by updating the DirectSound buffer attributes. Updating is double buffered to avoid visible image tearing. Thus the blitting is done to a “back buffer”, which is a video memory surface being the same size as the display window. When all of the sprites have been painted, the back buffer is blitted to the front (visible) buffer. Given a modern video card with adequate video RAM and hardware stretching, these operations are extremely fast.

3.2 Interpretation of sprite trees

The goal of `ImageB` display is first to “spritify”, i.e., convert an abstract `ImageB` value to a initial list of sprite trees, and then update the trees iteratively. Concurrently, the sprite engine traverses and displays the sprite trees. Although the sprite trees are really implemented in C++, for the purpose of exposition, we will express them here in as the Haskell type definitions in Figure 6.

In order to define “correct” conversion, we need an interpretation of sprite tree lists. We specify this interpretation by mapping sprite tree lists to `ImageB` values. Note that this mapping is hypothetical, serving to (a) specify what the sprite engine does, and (b) justify our implementation in Section 4 of the *reverse* mapping, i.e., from `ImageB` values to `[SpriteTree]`.

```
data SpriteTree =
   SoundSprite RealB          -- volume adjust
               RealB          -- left/right pan
               RealB          -- pitch adjust
               Bool           -- whether to auto-repeat
               SoundBuffer
 | RenderedSprite Vector2B    -- translation vector
                  SurfaceULB
 | SolidSprite RectB          -- cropping region
               ColorB         -- solid color
 | FlipSprite RectB           -- cropping region
              Vector2B        -- translation vector
              RealB           -- scale factor
              HFlipBook
              RealB           -- page number
 | UntilT [SpriteTree] (Event [SpriteTree])
 | CondT   BoolB [SpriteTree] [SpriteTree]
```

Figure 6: Sprite trees as Haskell types

Note that `[SpriteTree]` type almost fits the GBehavior style, but not quite. Instead, it matches the sprite engine's data structure.

The interpretation functions are given in Figure 7. The function `treesIm` is the main interpretation function, and says that a list of sprite trees represents an overlay of the images represented by the member trees, in reverse order. This reversal reflects the fact that the sprite tree lists are in back-to-front order so that the sprite engine can paint the sprite trees in list order. Later (upper) sprite trees are then painted over earlier (lower) sprite trees. The first four clauses of the `treeIm` function interpret individual sprites. Note the rigidity of order of applied operations imposed by the sprite constructors, as contrasted with the flexibility afforded by the `ImageB` type (Figure 3). It is exactly because of this difference that, while interpretation of sprite tree lists is straightforward, generation is not.

The `UntilT` case says that a reactive sprite tree represents a reactive image animation. This animation is initially the one represented by the initial sprite tree list. When the event occurs, yielding a new sprite tree list, the animation switches to the one represented by these new trees.[10] The conditional case is similar.

[10]The event `e ==> treesIm` occurs whenever `e` does. At each occurrence, the function `treesIm` is applied to the the event data from `e`. Thus the event combinator "`==>`" is a `map` on events.

```
treesIm :: [SpriteTree] -> ImageB
treesIm []            = emptyImage
treesIm (bot : above) = treesIm above `over` treeIm bot

treeIm (RenderedSprite motionB surfaceB) =
  move motionB (surfaceIm surfaceB)

treeIm (SolidSprite rectB colorB) =
  crop rectB (withColor colorB solidImage)

treeIm (FlipSprite rectB motionB scaleB book pageB) =
  crop rectB     $
  move motionB   $
  stretch scaleB $
  flipImage book pageB

treeIm (SoundSprite volB panB pitchB soundB) =
  volume volB  $
  pan panB     $
  pitch pitchB $
  soundImage soundB


treeIm (trees `UntilT` e) =
  treesIm trees `untilB` (e ==> treesIm)

treeIm (CondT c trees trees') =
  condB c (treesIm trees) (treesIm trees')
```

Figure 7: Interpreting sprite trees

4 From abstract animations to sprite trees

The interpretation of sprite tree lists as image animations given above helps to specify the reverse process, which we must implement in order to use the sprite engine to display animations: The `spritify` algorithm takes sprite trees to overlay, a cropping rectangle, optional color, a space- and a time-transformation, all applied to the given `ImageB` while spritifying. (Recall `mbWithColor` from Section 2.3.)

```
spritify :: [SpriteTree] -> RectB -> Maybe ColorB -> Transform2B
         -> TimeB -> ImageB -> [SpriteTree]
treesIm (spritify above rectB mbColorB xfB tt imB) ==
  treesIm above `over`
  (crop rectB           $
   mbWithColor mbColorB $
   (xfB *%)             $
   (`timeTransform` tt) $
   imB)
```

The algorithm works by recursive traversal of `ImageB` values, accumulating operations that are found out of order. To get the algorithm started, Fran has just an `ImageB`, `imB`, and a cropping rectangle, `windowRectB`, based on the display window's size (which may vary with time), and so invokes `spritify` as `spritify [] windowRectB Nothing identity2 time imB`, where `identity2` is the identity 2D transform, and `time` serves as the identity time transform.

Our claim that the algorithm given below satisfies this specification may be proved by induction on the `ImageB` representation. Rather than give the algorithm first and then the proof, however, we instead derive the algorithm from the specification so that it is correct by construction. We will rely on many simple algebraic properties of our data types. These properties may be proved from semantic models of the data types, but we leave this matter to a future paper. Meanwhile they may be taken as axioms.

4.1 Solid images

Solid images are unaffected by space- or time- transformations, and are of a default color, as expressed by the following properties:

```
xfB *% SolidImage == SolidImage

SolidImage `timeTransform` tt == SolidImage

withColor defaultColor SolidImage == SolidImage
```

These facts simplify the specification of `spritify`:

```
treesIm (spritify above rectB mbColorB xfB tt SolidImage) ==
  treesIm above `over`
  crop rectB (withColor (chooseColorB mbColorB) SolidImage)
```

where

```
chooseColorB :: Maybe ColorB -> ColorB
chooseColorB Nothing  = defaultColor
chooseColorB (Just c) = c
```

Now considering the interpretation of `SolidSprite` and the definition of `treesIm` given in Figure 7, the specification simplifies further:

```
treesIm (spritify above rectB mbColorB xfB tt SolidImage) ==
  treesIm (SolidSprite rectB (chooseColorB mbColorB) : above)
```

This final simplification then directly justifies the `SolidImage` case in the `spritify` algorithm.

```
spritify above rectB mbColorB xfB tt SolidImage =
  SolidSprite rectB (chooseColorB mbColorB) : above
```

4.2 Rendered images

Because of their spatially continuous nature, `ImageB` values may be thought of as being constructed bottom-up, whereas their implementation via `spritify` is top-down. For instance given "`stretch 10 circle`", we would not want to render a circle to a (discrete) bitmap and then stretch the bitmap, because the resulting quality would be poor. Similarly, for "`stretch 0.1 circle`", it would be wasteful to render and then shrink. Instead, scaling transforms are postponed and incorporated into the rendering, which can then be done at an appropriate resolution. Similarly, rotation is an expensive operation on bitmaps, but can generally be moved into rendering.

For this reason, we want to factor the transform behavior into translation, scale, and rotation. Fran's `factorTransform2` function does just this:

```
xfB == translate motionB `compose2`
       uscale    scaleB  `compose2`
       rotate    angleB
 where
   (motionB, scaleB, angleB) = factorTransform2 xfB
```

Moreover, the meaning of transform composition is function composition.

```
(xfB `compose2` xfB') *% thing == xfB *% (xfB' *% thing)
```

Factoring the spatial transform and turning transform composition into function composition, our specification becomes the following.

```
treesIm (spritify above rectB mbColorB xfB tt
                  (RenderImage renderer)) ==
  treesIm above `over`
  (crop rectB           $
   mbWithColor colorB   $
   move    motionB      $
   stretch scaleB       $
   turn    angleB       $
   (`timeTransform` tt)  $
   RenderImage renderer)
 where
   (motionB, scaleB, angleB) = factorTransform2 xfB
```

Next, since renderers do not perform motion, we must extract the `move` from within applications of `withColor` and `crop`. Coloring commutes with spatial transformation:

```
withColor c (xfb *% imb) == xfb *% (withColor c imb)
```

Cropping is trickier, requiring that the cropping rectangle be inversely transformed:

```
crop rectB (xfb *% imb) == xfb *% crop (inverse2 xfb *% rectB) imb
```

For example, doubling the size of an image and then cropping with a rectangle it is equivalent to cropping with a half-size rectangle and then doubling in size.

The definition of a *renderer* in Section 2.3 then leads to the following rule for spritifying rendered images.[11]

```
spritify above rectB mbColorB xfB tt (RenderImage renderer) ==
  RenderedSprite motionB
      (renderer (move (-motionB) rectB)
                mbColorB scaleB angleB tt)
  : above
 where
   (motionB, scaleB, angleB) = factorTransform2 xfB
```

4.3 Flip-book animation

An analysis similar to the one for `SolidImage` above applies to flip book animation. Fran imposes two restrictions on flip-book animations (and the special case of constant bitmaps). They may not be colored or rotated. Simplifying the `spritify` specification under these assumptions yields the following.

```
treesIm (spritify above rectB Nothing xfB tt
                  (FlipImage book pageB)) ==
  treesIm above `over`
  (crop rectB          $
   mbWithColor Nothing  $
   move    motionB      $
   stretch scaleB       $
   turn (constantB 0)   $
   (`timeTransform` tt) $
   FlipImage book pageB )
 where
   (motionB, scaleB, constantB 0) = factorTransform2 xfB
```

We can then simplify further, using the following facts.

```
mbWithColor Nothing imb == imb

turn (constantB 0) imb  == imb

(flipImage book pageB) `timeTransform` tt ==
flipImage book (pageB `timeTransform` tt)
```

[11]The cropping and scaling behaviors should really be added to the `RenderImage` constructor and have the sprite engine use them, even though these operations are already performed during rendering. The reason is that the sprite engine can apply them incrementally at a much higher rate, for smoothness.

After applying these properties, the specification matches the interpretation of `FlipSprite`, and we get the following rule for `flipImage`.

```
spritify above rectB _ xfB tt (FlipImage book pageB) =
  FlipSprite rectB motionB scaleB
             book (pageB `timeTransform` tt)
  : above
 where
   (motionB, scaleB, _) = factorTransform2 xfB
```

4.4 The empty image

The empty image is trivially spritified, because (a) time transformation, space transformation, coloring, and cropping all map the empty image to itself, and (b) the empty image is the identity for `over`.

```
spritify above _ _ _ _ EmptyImage = above
```

4.5 Overlays

The treatment of overlays follows from the distribution of time transformation, space transformation, coloring, and cropping over the `over` operation, plus the associativity of `over`.

```
spritify above rectB mbColor xfB tt (top `Over` bot) =
  spritify (spritify above rectB mbColor xfB tt top)
           rectB mbColor xfB tt bot
```

4.6 Time transformation

Spritifying a time transformed animation is relatively simple because of the following composition property.

```
(imb `timeTransform` tt') `timeTransform` tt  ==
imb `timeTransform` (tt' `timeTransform` tt)  ==
```

Thus:

```
spritify above rectB mbColor xfB tt (TimeTransI imb tt') =
  spritify above rectB mbColor xfB (tt' `timeTransform` tt) imb
```

4.7 Space transformation

Time transformation distributes over space transformation:

```
(xfB *% imb) `timeTransform` tt ==
(xfB `timeTransform` tt) *% (imb `timeTransform` tt)
```

Space transforms compose:

```
xfB *% (xfB' *% imB) == (xfB `compose2` xfB') *% imB
```

The rule for `TransformI` then follows easily.

```
spritify above rectB mbColor xfB tt (xfB' `TransformI` imb) =
  spritify above rectB mbColor
           (xfB `compose2` (xfB' `timeTransform` tt)) tt imb
```

4.8 Coloring

The treatment of `withColor` follows from the following commutativity-like properties and combination rule.[12]

```
(withColor c imb) `timeTransform` tt ==
withColor (c `timeTransform` tt) (imb `timeTransform` tt)

xf *% withColor c imb == withColor c (xf *% imb)

withColor c (withColor c' imb) == withColor c imb
```

Then

```
spritify above rectB mbColor xfB tt (WithColorI color' imb) =
  spritify above rectB mbColor'' xfB tt imb
 where
   mbColor'' = mbColor ++ Just (color' `timeTransform` tt)
```

The Haskell operation "`++`" on `Maybe` is defined as follows.

```
Nothing ++ mb  = mb
Just x  ++ _   = Just x
```

4.9 Cropping

Cropping works very much like space transformation and coloring, justified by analogous properties.

```
(crop rectB imb) `timeTransform` tt ==
crop (rectB `timeTransform` tt) (imb `timeTransform` tt)

xf *% crop rectB imb == crop (xf *% rectB) (xf *% imb)

crop rectB (crop rectB' imb)             ==
crop (rectB `intersectRect` rectB') imb
```

Then

```
spritify above rectB mbColor xfB tt (CropI rectB' imb) =
  spritify above rectB'' mbColor xfB tt imb
 where
   rectB'' = (rectB `intersectRect`) $
             (xfB *%)                $
             (`timeTransform` tt)    $
             rectB'
```

[12]Note that in the case of nested colorings, the *outer* ones win, in contrast to many graphics libraries. The reason is that our informal model of coloring is to simply paint over what is already present. The conventional, innermost-wins policy is simpler to implement, but has arguably more complex semantics.

4.10 Conditional animation

Time- and space-transformation, coloring, and cropping, all distribute over conditionals:

```
   (condB boolB thenB elseB) `timeTransform` tt
== condB (boolB `timeTransform` tt)
         (thenB `timeTransform` tt)
         (elseB `timeTransform` tt)

   xf *% (condB boolB thenB elseB)
== condB boolB (xf *% thenB) (xf *% elseB)

   withColor c (condB boolB thenB elseB)
== condB boolB (withColor c thenB) (withColor c elseB)

   crop rectB (condB boolB thenB elseB)
== condB boolB (crop rectB thenB) (crop rectB elseB)
```

The rule then follows:

```
spritify above rectB mbColor xfB tt (CondI c imb imb') =
  CondT (c `timeTransform` tt)
        (spritify [] rectB mbColor xfB tt imb)
        (spritify [] rectB mbColor xfB tt imb)
  : above
```

Note that we spritify `imb` and `imb'` with an empty `above` list, and then place the whole conditional sprite tree under the given `above` sprites. Alternatively, one might spritify `imb` and `imb'` each with the given `above` list, and make a singleton `CondT` sprite tree list. We chose the implementation above because it preserves the invariant that sprite tree structure is never shared (Section 3.1).

4.11 Reactive animation

Reactive animations are spritified in much the same way as conditional animations, justified by analogous properties. The rule:

```
spritify above rectB mbColor xfB tt (imb `UntilI` e) =
  UntilT (spritify [] rectB mbColor xfB tt imb)
         (e `afterE` (rectB, mbColor, xfB, tt) ==>
           \ (imb', (rectB', mbColor', xfB', tt')) ->
            spritify [] rectB' mbColor' xfB' tt' imb')
  : above
```

The `afterE` combinator gives access to the "residual", or "aged" version, of a value of a `GBehavior` type upon the occurrence of an event. Using the unaged behaviors instead would cause a time-space leak, since the behaviors would be held onto from their beginning while waiting for the event to occur. See [2] for details.

4.12 Embedded sound

Image-embedded sounds are spritified by generating a collection of sound sprites. In Fran, the pan (left/right balance), and the volume are influenced by the spatial transform being applied. If the sound's location is outside of the cropping rectangle then it is silenced. (The operator "`.+^`" is point/vector addition.)

```
spritify above rectB _ xfB tt (SoundI sound) =
 spritifySoundB above vol pan pitch tt sound
  where
    (motion, scale, _) = factorTransform2 xfB
    -- Pan based on x coordinate.  The 7.0 is empirical.
    pan = 7.0 * fst (vector2XYCoords motion)
    -- Volume is based on scale, possibly silenced by cropping
    vol = ifB (rectContains rectB (origin2 .+^ motion))
              (abs scale) 0
    pitch = 1
```

The new sound conversion function `spritifySoundB` is specified and implemented much like spritifying image animations. The specification:

```
spritifySoundB :: [SpriteTree] -> RealB -> RealB -> RealB
               -> TimeB -> SoundB -> [SpriteTree]

treesIm (spritifySoundB above volB panB pitchB tt sound) ==
  treesIm above `over`
  soundImage (
     volume volB          $
     pan    panB          $
     pitch  pitchB        $
     (`timeTransform` tt) $
     sound)
```

The implementation, given in Figure 8, is somewhat simpler than that of `spritify`, because of the independence of the operations for adjusting pan, volume, and pitch.

4.13 State and concurrent updating

The spritifying algorithm above is idealized in that it constructs immutable sprite tree lists containing Fran behaviors. In fact, the sprite engine's data structures are mutable, both parametrically and structurally, and only accomodate linear behaviors. The actual implementation of `spritify` and `spritifySound` creates the mutable sprite trees with initial values for position, scale, pitch, etc., and then iteratively updates these attributes, while the sprite engine runs concurrently. For simplicity of implementation, every active sprite is managed by its own Haskell thread, using the Concurrent Haskell primitives [10]. Each such thread is fueled by a request channel, with each request saying either to continue or to quit, and puts status messages into a response channel. Each sprite thread iteratively samples the appropriate behaviors (slightly into the future) and invokes an update method on the sprite object. In response, the sprite engine charts a new linear course for each attribute.

```
spritifySoundB above _ _ _ SilentS = above

spritifySoundB above volB panB pitchB tt (BufferS buff repeat) =
  SoundSprite volB panB pitchB repeat buff : above

spritifySoundB above volB panB pitchB tt (sound `MixS` sound') =
  spritifySoundB above' volB panB pitchB tt sound'
 where
   above' = spritifySoundB above volB panB pitchB tt sound

spritifySoundB above volB panB pitchB tt (sound `TimeTransS` tt') =
  spritifySoundB above volB panB pitchB (tt' `timeTransform` tt)
                 sound

spritifySoundB above volB panB pitchB tt (VolumeS v sound) =
  spritifySoundB above (volB * v `timeTransform` tt)
                 panB pitchB tt sound

spritifySoundB above volB panB pitchB tt (PanS p sound) =
  spritifySoundB above volB
                 (panB + timeTransform p tt)
                 pitchB tt sound

spritifySoundB above volB panB pitchB tt (PitchS p sound) =
  spritifySoundB above volB panB
                 (pitchB * p `timeTransform` tt)
                 tt sound

spritifySoundB above volB panB pitchB tt (sound `UntilS` e) =
 UntilT (spritifySoundB [] volB panB pitchB tt sound)
        (e `afterE` (volB, panB, pitchB, tt) ==>
          \ (sound', (volB', panB', pitchB', tt')) ->
            spritifySoundB [] volB' panB' pitchB' tt' sound')
 : above
```

Figure 8: Spritifying SoundB values

```
spritify :: SpriteTree  -> RectB  -> Maybe ColorB
         -> Transform2B -> ImageB -> Maybe TimeB
         -> [Time] -> MVar Bool -> MVar Bool -> IO SpriteTree

spritify above rectB mbColorB xfB tt SolidImage
         ts requestV replyV = do
  sprite <- newMonochromeSprite above
  let update ~(t:ts')
             ~(RectLLUR (Point2XY llx lly) (Point2XY urx ury) : rects')
             ~(ColorRGB r g b : colors') = do
         continue <- takeMVar requestV
         if continue then do
           updateMonochromeSprite sprite t llx lly urx ury r g b
           putMVar replyV True
           update ts' rects' colors'
          else
           putMVar replyV False
  forkIO $ update ts (rectB `ats` ts) (chooseColorB mbColorB `ats` ts)
  return (toSpriteTree sprite)
```

Figure 9: `spritify` case with concurrent updating

For example, Figure 9 is a slightly simplified implementation of the `SolidImage` case. A monochrome sprite is allocated in the sprite engine, chaining to the sprite trees given as "`above`". The time stream `ts` is used to sample the cropping rectangle and color, using `ats` (which is memoized as described in [2]). The resulting sample values are used to interactively update the sprite attributes, as long as the request channel `requestV` says to continue. The iteration is placed in a new Haskell thread via `forkIO`.

The handling of a reactive image animation, `imb `untilB` e`, is somewhat tricky. The initial animation `imb` is spritified, starting any number of threads. (`OverI`-based animations lead to multiple threads.) The resulting sprite tree list is wrapped up in a `SpriteGroup` object (corresponding to the `UntilT` constructor above). One more thread runs to watch for event occurrences and causes the update work corresponding to `imb` to continue until the first occurrence of the event `e`, and then to stop. At that point, a new animation is available and is spritified, generating a new sprite tree list, which is then passed to a method on the `SpriteGroup` object that recursively deletes its old list and installs the new one. The number of running threads thus varies in response to events. Because every thread has accompanying overhead for controlling its work, a useful future optimization would be to create many fewer threads.

5 Conclusions

The implementation techniques described in this paper bridge the gap between functional animation and retained-mode display. Functional animation serves the needs of composability, while retained-mode promotes efficient use of hardware resources. A recurring theme in this work is the application of algebraic properties of our animation data types, in order to normalize animations to the relatively restrictive form imposed by retained-mode presentation libraries. An area of future work is to develop rigorous semantic models for these data types. The models would form the interface between the informal mental models taught to and used by the everyday animation programmer and the correct and efficient implementation of the animation library itself. This semantic orientation has driven our work from the start, but some work remains to make it complete and precise.

References

[1] Kavi Arya. A functional animation starter-kit. *Journal of Functional Programming*, 4(1):1–18, January 1994.

[2] Conal Elliott. Functional implementations of continuous modeled animation. To appear in PLILP/ALP '98. `http://www.research.microsoft.-com/~conal/papers/plilpalp98/short.ps`.

[3] Conal Elliott. Modeling interactive 3D and multimedia animation with an embedded language. In *The Conference on Domain-Specific Languages*, pages 285–296, Santa Barbara, California, October 1997. USENIX. WWW version at `http://www.research.microsoft.com/-~conal/papers/dsl97/dsl97.html`.

[4] Conal Elliott. Composing reactive animations. *Dr. Dobb's Journal*, July 1998. Extended version with animations at `http://www.research.-microsoft.com/~conal/fran/{tutorial.htm,tutorialArticle.zip}`.

[5] Conal Elliott and Paul Hudak. Functional reactive animation. In *Proceedings of the 1997 ACM SIGPLAN International Conference on Functional Programming*, pages 263–273, Amsterdam, The Netherlands, 9–11 June 1997.

[6] Rob Glidden. *Graphics Programming With Direct3D : Techniques and Concepts*. Addison-Wesley, 1996.

[7] Paul Hudak and Joseph Fasel. A gentle introduction to Haskell. *SIGPLAN Notices*, 27(5), May 1992. See `http://haskell.org/tutorial/-index.html` for latest version.

[8] Paul Hudak, Simon L. Peyton Jones, and (editors) Philip Wadler. Report on the programming language Haskell, A non-strict purely functional language (Version 1.2). *SIGPLAN Notices*, Mar, 1992. See `http://haskell.org/report/index.html` for latest version.

[9] J. Hughes. Lazy memo functions. In J. P. Jouannaud, editor, *Functional Programming Languages and Computer Architecture*, volume 201 of *LNCS*, pages 129–146. Springer Verlag, September 1985.

[10] Simon Peyton Jones, Andrew Gordon, and Sigbjorn Finne. Concurrent Haskell. In *Conference Record of POPL '96: The* 23rd *ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*, pages 295–308, St. Petersburg Beach, Florida, 21–24 January 1996.

[11] Microsoft. DirectX home page. `http://www.research.microsoft.com/-directx`.

[12] Microsoft and Peter Donnelly. *Inside DirectX*. Microsoft Press, 1998.

[13] John Peterson, Conal Elliott, and Gary Shu Ling. Fran user's manual, Revised February, 1998. `http://www.research.microsoft.com/-~conal/Fran/UsersMan.htm`.

[14] John Rohlf and James Helman. IRIS performer: A high performance multiprocessing toolkit for real-time 3D graphics. In Andrew Glassner, editor, *Proceedings of SIGGRAPH '94 (Orlando, Florida, July 24–29, 1994)*, pages 381–395, July 1994.

[15] Paul S. Strauss and Rikk Carey. An object-oriented 3D graphics toolkit. *Computer Graphics*, 26(2):341–349, July 1992.

[16] Jay Torborg and Jim Kajiya. Talisman: Commodity real-time 3D graphics for the PC. In *SIGGRAPH 96 Conference Proceedings*, pages 353–364, August 1996. held in New Orleans, Louisiana, 04-09 August 1996.