

Stroke Surfaces: A Spatiotemporal Framework for Temporally Coherent Non-photorealistic Animations

J. P. Collomosse

P. M. Hall

Department of Computer Science, University of Bath, Bath, England.

Abstract

This paper outlines a novel framework for the automated synthesis of non-photorealistic animations from video sequences. Our approach is unique among such animation techniques in that we process the source video sequence as a spatiotemporal voxel volume. Video frames are segmented into homogeneous regions, and heuristic associations between regions formed over time to produce a collection of conceptually high level spatiotemporal objects. By manipulating objects in this representation we are able to synthesise a wide gamut of artistic effects, which we allow the user to stylise and influence through a parameterised “Video Paintbox”. The benefits of our approach are a versatile framework capable of producing animations exhibiting a high degree of temporal coherence; a property scarce in current non-photorealistic video rendering techniques which render video only on a per frame sequential (temporally local) basis.

1 Introduction

Processing video for use by the entertainment industry is just one aspect of contemporary Computer Vision, and forms part of a more general convergence trend between Computer Vision and Computer Graphics. This paper is concerned with the rendering of real video in artistic styles, and as such is broadly aligned with this convergence area and with the non-photorealistic rendering (NPR) community within Computer Graphics. In particular we are motivated by a desire to render video in cartoon-like styles, a problem which decomposes into two separable sub-goals: 1) producing temporally coherent stylised shading effects in the video; 2) emphasising motion within the image sequence. This paper addresses the former of these issues and complements work recently published by the authors in [1, 3], which addresses the latter issue.

A number of algorithms exist for rendering static images in non-photorealistic styles, which composit brush strokes upon a digital canvas to create painterly [5, 6] or sketchy [9] image stylisations. Researchers have found the extension of such techniques to video to be non-trivial. In particular, existing algorithms can not be applied directly to individual video frames without introducing aesthetically poor temporal coherence manifested as a flickering or *swimming* into the resulting animation. Attempts have been made to mitigate incoherence by translating strokes from frame to frame using an estimated optical flow vector field [7, 8]. We observe that such approaches operate at a *temporally low-level*; analysing and rendering video on a per frame sequential basis, and considering only content in the previous frame when rendering the next. Errors present in the estimated motion field quickly accumulate and propagate to subsequent frames. We argue that the problem of processing an image sequence for artistic effect is complex, and basing decisions upon a local, greedy (per frame) algorithm is unlikely to result in an optimal (in this context, temporally coherent) solution. A global analysis over all frames seems intuitively more likely to produce coherent renderings, yet all current techniques for the artistic rendering of image sequences operate on a per frame sequential

basis; none of these existing techniques achieve satisfactory levels of temporal coherence in practice without extensive manual correction [4].

Our approach is unique among NPR video techniques in that we process the source video sequence as a spatiotemporal voxel volume. Video frames are segmented into homogeneous regions, and heuristic associations between regions formed over time to produce a collection of conceptually high level spatiotemporal objects. These objects carve sub-volumes through the video volume delimited by continuous isosurface “stroke surface” patches. By manipulating objects in this representation we are able to synthesise a wide gamut of artistic effects, which we allow the user to stylise and influence through a parameterised “Video Paintbox”. The benefits of our approach are a versatile framework capable of producing animations exhibiting a higher degree of temporal coherence than current, temporally local video rendering algorithms.

2 Outline of the Video Paintbox

The Video Paintbox consists of a single rendering framework which may be broken into a front and back end. The front end is responsible for the parsing of the source video to create an intermediate representation (IR), and is largely automated through application of Computer Vision techniques. This abstracted video representation is then passed to the back end, where it is rendered in one of a range of artistic styles. The user is given control over the back end of the system via a set of high-level parameters which influence the style of the resulting animation.

2.1 Front end: Building the Representation

Frames are independently segmented into connected homogeneous regions using standard 2D Computer Vision techniques. The criterion for homogeneity we have chosen is colour, but one might equally well segment on the basis of texture or motion; the nature of the video content influences such a choice. We assume motion within the image sequence to be smooth, that is, free from scene cuts and cross-fades; we draw upon standard techniques to segment the video as such during pre-processing. For each frame, associations are created between that frame’s regions and those of frames adjacent to it. Associations are heuristically created based upon commonalities in colour, location and shape (using Fourier descriptors). The result is a graph of connected homogeneous sub-volumes carved from the spatiotemporal video volume, describing the trajectory of features in the video. These sub-volumes are smoothed via morphological filtering and fragmented into temporally convex video objects (Figure 1e). We observe that whilst a 3D segmentation of the video volume would be more in keeping with our spatiotemporal approach, our $2D + t$ scheme permits attributes such as shape or colour to evolve over time. Small, fast objects which may break voxel connectivity in the volume are also correctly grouped using our method. A 3D approach would be unsuitable in these cases.

Segmentation and association produces a series of voxel volumes from which we generate the IR to be passed to the back end. In

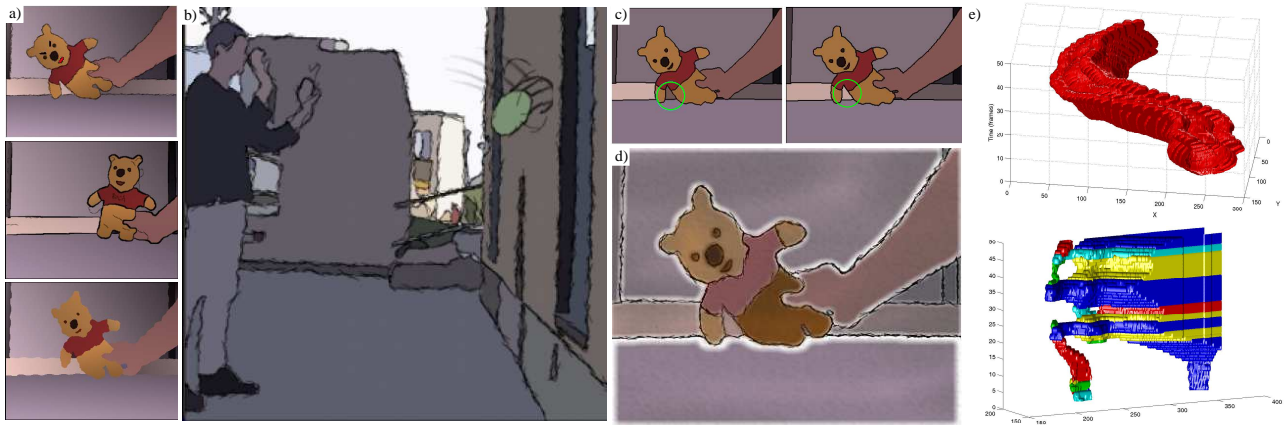


Figure 1 Demonstrating our Video Paintbox. (a) Examples of coherent roto-scoping (upper), gradient-shaded cartoon with internal detail (middle), and coherent wobbling effect created by periodic distortion of stroke surfaces (lower). (b) Combining the framework with our earlier motion emphasis work [1, 3] to produce cartoon-style video. (c) Temporal incoherence (left) present when shading on a per frame basis, circled in green. Our spatiotemporal representation allows us to mitigate such artifacts (right). (d) A coherent sketchy effect has been generated by shattering and jittering of the spatiotemporal stroke surfaces which bound each video object. (e) Video volumes corresponding to the bear’s head (upper) and right-hand skirting board (lower); individual video objects are falsely coloured to illustrate their temporal convexity.

our framework, the spatiotemporal locations of objects are represented in terms of their interfacing surfaces. This is preferable to storing each object in terms of its own bounding surface, leading to a more compact and manipulable representation (which is useful later when we smooth or deform object boundaries). When two objects abut in the video volume, their interface may be represented in piecewise form by a number of (possibly discontinuous) surface patches. Each of these patches we term a “stroke surface”, and store the complete set of stroke surfaces for the video as one half of the IR. Each stroke surface holds an additional *winged edge structure* which contains two pointers corresponding to the two objects which it separates. A supplementary database is maintained as the second half of the IR, containing one record per object in the video volume. This counterpart database is referenced by the pointers held in the winged edge structure and stores various attributes about each object, such as mean colour over time, and a record of temporal children and parents (thus encapsulating the object association graph). Internal edge detail within the regions may also be encapsulated in the stroke surface representation; in such cases both pointers in the winged edge structure reference the same video object.

This representation facilitates easy temporal manipulation of the video sequence; we may smooth stroke surfaces to produce temporally coherent segmentations, or introduce high frequency components to produce coherent wobbling effects (Figure 1a, lower).

2.2 Back end: Rendering the Representation

To render a particular frame at time t , the set of stroke surfaces embedded in the video volume $\mathbb{R}^3 = [x, y, z]$ are intersected with the plane $z = t$. Intersected surfaces are then rendered in a two stage process comprising: 1) shading of the object’s interior region; 2) rendering of the object’s outline (the *holding line*) and any interior cue lines also present. This separation allows us to create many novel video effects, such allowing interior shading to spill outside of the holding lines.

With the coherent segmentation provided by the stroke surface representation, cartoon-like effects can be created by flat shading regions with their mean colour for each frame. However as objects divide and merge their mean colour can change significantly causing flickering (Figure 1c). This is symptomatic of the general problem of assigning region attributes in a coherent manner, and we draw upon our spatiotemporal representation to mitigate such incoherence. Recall that objects are associated via an association graph

structure. We may reconstruct the feature sub-volume in a small temporal window around the current frame using this graph. By smoothly varying database attributes, such as colour, over the volume we can create a smooth transition of those attributes over time even if objects appear disjoint in the current frame but connect at some other instant in time. Such coherence could not be obtained using the per frame sequential analysis performed by current NPR video methods.

Similarly, we may estimate inter-frame motion (via homography) for each feature, and smooth these motion parameters over the feature sub-volume; users may draw upon a feature in a particular key-frame, and have their illustration move coherently with that feature. This is an automated example of roto-scoping, a technique pioneered in the 1970s, in which an animator manually traces over stills to give a stylised effect in a cartoon. In a sense, static image-based NPR methods (for example stroke-based renderers) may also be considered a form of modern day roto-scoping. By placing brush strokes on a feature in a similar manner, we can create coherent painterly effects in the video.

Holding lines may also be rendered in a range of styles. The splines produced during time-plane/surface intersection may be interpreted as brush trajectories, and simulated brushes used to paint feature outlines. Spatiotemporal effects are also possible, for example by shattering the stroke surfaces into multiple fragments prior to rendering, then applying stochastic affine variations to each fragment which results in a coherent sketchy effect (Figure 1d).

3 Discussion

Our work demonstrates that the non-photorealistic rendering of video as a spatiotemporal entity, rather than on a per frame basis, proves beneficial in terms of 1) enhancing temporal coherence of animation, 2) diversifying the gamut of video NPR effects through extension of static stroke-based NPR techniques. Furthermore, the compact nature of the IR is an area worthy of further investigation. We have successfully combined the proposed framework with our previous motion emphasis work to create highly stylised cartoon-like animations from video (Figure 1b).

A full description of the Video Paintbox framework is available in a technical report [2] at <http://www.cs.bath.ac.uk/~vision/cartoon>. Rendered video clips are also available for download at this site.

References

- [1] J. P. Collomosse, D. Rowntree, and P. M. Hall. Cartoon-style rendering of motion from video. In *Vision, Video and Graphics*, pages 117–124, July 2003.
- [2] J. P. Collomosse, D. Rowntree, and P. M. Hall. Stroke surfaces: A spatio-temporal framework for temporally coherent non-photorealistic animations. Technical Report 2003–01, University of Bath, UK, June 2003.
- [3] J. P. Collomosse, D. Rowntree, and P. M. Hall. Video analysis for cartoon-like special effects. In R. Harvey and A. Bangham, editors, *Proceedings BMVC*, volume 2, pages 749–758, Norwich, September 2003.
- [4] S. Green, D. Salesin, S. Schofield, A. Hertzmann, and P. Litwinowicz. Non-photorealistic rendering. *SIGGRAPH '99 Non-Photorealistic Rendering Course Notes*, 1999.
- [5] P. Haeblerli. Paint by numbers: abstract image representations. In *Proceedings Computer Graphics (ACM SIGGRAPH)*, volume 4, pages 207–214, 1990.
- [6] A. Hertzmann. Painterly rendering with curved brush strokes of multiple sizes. In *Proceedings Computer Graphics (ACM SIGGRAPH)*, pages 453–460, 1998.
- [7] A. Hertzmann and K. Perlin. Painterly rendering for video and interaction. In *Proceedings 1st International Symposium on Non-photorealistic Animation and Rendering (NPAR)*, pages 7–12, 2000.
- [8] P. Litwinowicz. Processing images and video for an impressionist effect. In *Proceedings Computer Graphics (ACM SIGGRAPH)*, pages 407–414, Los Angeles, USA, 1997.
- [9] M. P. Salisbury, M. T. Wong, J. F. Hughes, and D. H. Salesin. Orientable textures for image-based pen-and-ink illustration. In *Proceedings Computer Graphics (ACM SIGGRAPH)*, pages 401–406, Los Angeles, USA, 1997.