Video Stylization for Digital Ambient Displays of Home Movies

Tinghuai Wang and John Collomosse*
Centre for Vision, Speech and Signal Processing,
University of Surrey, UK

David Slatter, Phil Cheatle and Darryl Greig†
Multimedia Interaction and Understanding Lab,
Hewlett-Packard Labs, Bristol UK.

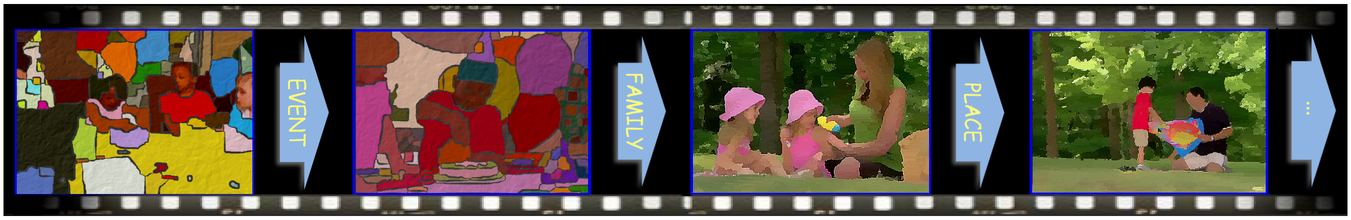

Figure 1: *Video clips are stylized into cartoons or paintings, and sequenced according to semantic and visual similarity (view at 200% zoom).*

Abstract

Falling hardware costs have prompted an explosion in casual video capture by domestic users. Yet, this video is infrequently accessed post-capture and often lies dormant on users' PCs. We present a system to breathe life into home video repositories, drawing upon artistic stylization to create a "Digital Ambient Display" that automatically selects, stylizes and transitions between videos in a semantically meaningful sequence. We present a novel algorithm based on multi-label graph cut for segmenting video into temporally coherent region maps. These maps are used to both stylize video into cartoons and paintings, and measure visual similarity between frames for smooth sequence transitions. We demonstrate coherent segmentation and stylization over a variety of home videos.

CR Categories: I.4.6 [Image Processing and Computer Vision]: Segmentation—Region growing, partitioning I.3.4 [Computer Graphics]: Graphics Utilities—Paint systems

Keywords: Video stylization, segmentation, graph cut, temporal coherence, composition, ambient displays.

1 Introduction

The serendipitous rediscovery of memories whilst browsing physical media archives (e.g. a box of photos in the attic), can trigger enjoyable reminiscence over past events. Digital media collections are intrinsically more accessible than physical archives, yet the focus on the PC as the main portal to these collections poses a convenience barrier to realizing their value.

In this work we consider a genre of content consumption experience which we call *ambient experiences*. These are distinguished from compelling or intense experiences in that they are able to co-exist harmoniously with other activities such as conversations, shared meals and so forth. An ambient experience does not demand the full attention of the user but is able to play out in a pleasing, unobtrusive way such that fresh and interesting content is available in the attention spaces of everyday life.

The problem of displaying still images in an ambient way is much explored and fairly well understood. Existing solutions range from low-tech, such as framed photographs, through to digital picture frames, slide shows and more elaborate multi-media presentations based on still images ([Xiao et al. 2008; Slatter et al. 2010]). However there are very few solutions for ambient experiences involving *video content*. With the expected proliferation of large format video displays around the home there has been much interest in interaction methods [Arksey 2007; You et al. 2008]; we believe there is a complementary need for minimal interaction ambient technologies on the same displays when full attention is either not possible or not desirable.

We propose a solution for *video* collections in the form of *Digital Ambient Displays* (DADs); always-on displays for living spaces that enable users to effortlessly visualize and rediscover their video collections. Rather than simply stitching videos together, we harness artistic stylization to depict video in a more abstract sense; creating a flowing, temporal composition that conveys the essence of users' experiences through their videos.

Creating a DAD requires that home video is automatically parsed into an underlying scene representation that enables both:

- Coherent stylization of video into aesthetically pleasing forms
- Generation of appropriate transition effects and sequencing decisions, to create an appealing temporal composition.

Following [DeCarlo and Santella 2002; Collomosse et al. 2005] we identify a color region segmentation as being an appropriate "mid-level" scene abstraction, and in Section 3 contribute a novel algorithm for segmenting video frames into a deforming set of temporally coherent regions. We demonstrate how these regions may be stylized via either shading or stroke-based rendering, to produce coherent cartoon and painterly video styles (Section 3.5). We describe our stochastic approach to video selection and composition in Section 4, presenting a gallery of results in Section 5 and video.

1.1 Related Work

Our video compositions are driven by the stochastic selection of video clips (Section 4) in a concatenative manner reminiscent of Motion Graphs [Kovar et al. 2002]. Stochastic transitions between video frames were first proposed in [Schodl et al. 2000], within the

*e-mail: { tinghuai.wang | j.collomosse} @surrey.ac.uk
†e-mail: {david.slatter | phil.cheatle | darryl.greig} @hp.com

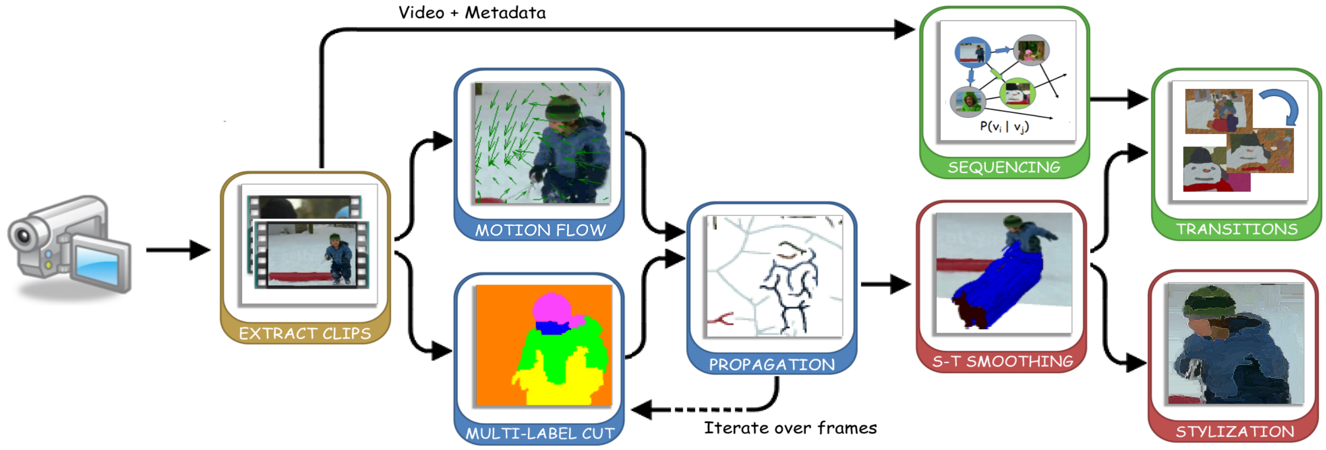

Figure 2: *System overview. Videos are segmented into coherent region maps encoding visual structure. Maps drive stylization and transitions.*

scope of a single video and based upon visual similarity only. Others have studied composition of photos for abstract [Collomosse and Hall 2003] or video artwork [Slatter et al. 2010; Bizzocchi 2008]. By contrast, little work explores compositions of video clips or the use of artistic video stylization in ambient displays.

Video stylization was first addressed by [Litwinowicz 1997], who produces painterly video by pushing brush strokes from frame to frame in the direction of optical flow motion vectors. This approach was later extended by [Hays and Essa 2004] who similarly move strokes within independent motion layers, and by [Bousseau et al. 2007] who harness flow to advect strokes between adjacent frames. Complementary work by [Hertzmann and Perlin 2000] use differences between consecutive frames of video, painting over areas of the new frame that differ significantly from the previous frame. While these methods can produce impressive painterly video, the errors in the estimated per-pixel motion field can quickly accumulate and propagate to subsequent frames, resulting in increasing temporal incoherence. This can lead to a distraction scintillation or “flicker” when strokes of the stylized output no longer match object motion [Meier 1996].

More recently, image segmentation techniques have been applied to yield *mid-level* models of scene structure [Wang et al. 2004; Collomosse 2004] that can be rendered in artistic styles. By extending the mean-shift based stylization approach of [DeCarlo and Santella 2002] on images, [Collomosse et al. 2005] create spatio-temporal volumes from video by associating 2D segmentations over time and fitting *stroke surfaces* to voxel objects. Although this geometric smoothing improves stability, temporal coherence is not ensured because the region map for each frame is formed independently without knowledge of the adjacent frames. Furthermore, association is confounded by the poor repeatability of 2D segmentation algorithms between similar frames, causing variations in the shape and photometric properties of regions that require manual correction. [Wang et al. 2004] also transforms video into spatio-temporal volumes by clustering space-time pixels using a mean-shift operator. However, this approach becomes computationally infeasible for pixel counts in even moderate size videos, and often under-segments small or fast moving objects that form disconnected volumes. This also requires manual correction and frequent grouping of space-time volumes. [Winnemoller et al. 2006] presents a method to abstract video using a bilateral filter, attenuating detail in low-contrast regions while preserving sharp edges. Anisotropic filtering was also proposed in [Kyprianidis et al. 2009] using the Kuwahara filter. Such approaches do not seek to parse a description of scene structure, making them useful for scenes that are difficult to segment, but limited to a characteristic soft-shaded artistic style.

We adopt a scene segmentation approach; convenient both for diverse video stylization and for creating the structural correspondences between frames for transition animations. We propose a new video segmentation algorithm, in which the segmentation of each frame is guided by motion flow propagated priors estimated from the region labels of past frames. In doing so we combine the automation of early optical flow stylization algorithms with the robustness and coherence of region segmentation approaches; propagating labels with flow, and resolving ambiguities using a graph-cut optimization to create coherent region maps. Some recent interactive “video cut-out” systems are similar in spirit [Bai et al. 2009; Price et al. 2009]; tracking key-points on region boundaries over time for matte segmentation. However we differ in several ways. First, we propagate label priors and data forward with motion flow within regions, rather than tracking 2D windows on region boundaries that contain clutter from adjacent regions. Second, we are more general, producing a multi-label (region) map rather than a binary matte. Third, [Bai et al. 2009; Price et al. 2009] require regular manual correction, typically every ~ 5 frames. Our algorithm requires no user interaction, beyond (optional) modification of the initial frame for aesthetics.

2 System Overview

The Digital Ambient Display (DAD) requires the video collection to be ingested as short, visually interesting clips that form the atomic unit of composition. Obtaining such clips differs from classical shot detection as raw home footage tends to consist of a few lengthy shots. An existing algorithm [Wang et al. 2009] performs this pre-processing. The DAD then sequences a subset of clips to create a temporal composition that flow smoothly both visually and semantically (Figure 1). For example, the DAD might select a clip of the family in the garden, and follow this with a family clip in visually similar alternate environment such as a park or at a birthday. The sequencing algorithm is described in Section 4.

Presentation of video in the DAD is underpinned by a novel algorithm for segmenting video frames into temporally coherent colored regions (sub-secs. 3.1-3.3). These region maps form a stable representation of visual structure in the scene that is used both to drive artistic rendering algorithms for stylization (sub-Sec. 3.5), and to perform matching of scene elements between frames in order to generate animated clip transitions (sub-Sec. 4.2). Figure 2 provides an overview of our system; the segmentation, sequencing and stylization subsystems are blue, green and red respectively.

3 Video Stylization

We describe a new coherent video segmentation algorithm which performs a multi-label graph cut on successive video frames, using both photometric properties of the current frame and prior information propagated forward from previous frames. This information consists of:

i. an incrementally built Gaussian Mixture Model (GMM) encoding the color distribution of each region over past frames;

ii. a subset of pixel-to-region labels from the previous frame.

We check for region under-segmentation (e.g. the appearance of new objects, or objects emerging from occlusion) by comparing the historic and updated GMM color models for each region, and introducing new labels into that region if the color model appears to be temporally inconsistent. The region map of the first frame is boot-strapped using mean-shift segmentation, and may *optionally* be modified by the user for aesthetics e.g. to abstract away background detail by merging regions.

We first describe our segmentation algorithm (sub-Secs 3.1-3.3) and then describe how the coherent region maps are applied to stylize video (sub-Secs 3.4-3.5) and create the animations used to transition between successive clips in the DAD video sequence.

3.1 Multi-label Graph cut

We formulate video segmentation as the problem of assigning region labels existing in frame I_{t-1} to each pixel $p \in \mathcal{P}$ in frame $I_t(p)$; i.e. seeking the best mapping $l : \mathcal{P} \rightarrow \mathcal{L}$ where $\mathcal{L} = (l(1), \ldots, l(p), \ldots, l(|\mathcal{P}|))$ is the set assignments of labels $l_i, i = \{1...L\}$, and $\mathcal{P}$ is an 8-connected lattice of pixels.

A subset of $\mathcal{L}$ are carried forward from the region map at $t-1$, via a propagation process described shortly (sub-Sec. 3.2). This *prior labelling* of pixels ($\mathcal{O} \subseteq \mathcal{P}$) forms a hard constraint on the assignments of remaining pixels in I_t, which are labelled to minimize a global energy function encouraging both temporal consistency of color distribution between frames, and spatial homogeneity of contrast within each frame. This is captured by the data and pairwise terms of the Gibbs energy function:

$$E(\mathcal{L}, \Theta, \mathcal{P}) = U(\mathcal{L}, \Theta, \mathcal{P}) + V(\mathcal{L}, \mathcal{P}). \tag{1}$$

The data term $U(.)$ exploits the fact that different color homogeneous regions tend to follow different color distributions. This encourages assignment of pixels to the labelled region following the most similar color model (we write the parameters of such models Θ). The data term is defined as:

$$\begin{aligned} U(\mathcal{L}, \Theta, \mathcal{P}) &= \sum_{p \in \mathcal{P}} -\log P_g(I_t(p)|l(p); \Theta). \\ P_g(I(p)|l(p) = l_i; \Theta) &= \sum_{k=1}^{K_i} w_{ik} \mathcal{N}(I(p); \mu_{ik}, \Sigma_{ik}). \end{aligned} \tag{2}$$

i.e. the data model of the i^{th} label l_i is represented by a mixture of Gaussians (GMM), with parameters w_{ik}, μ_{ik} and Σ_{ik} representing the weight, the mean and the covariance of the k^{th} component. The parameters of all GMMs ($\Theta = \{w_{ik}, \mu_{ik}, \Sigma_{ik}, i = 1, \ldots, L, k = 1, \ldots, K_i\}$) are learned from historical observations of each region's color distribution (sub-Sec. 3.2).

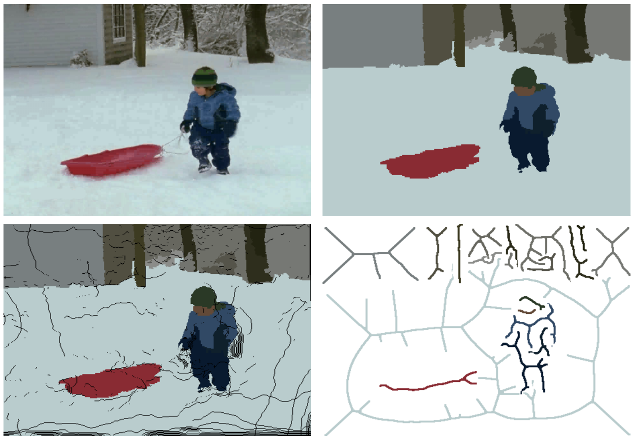

Figure 3: *Prior propagation: (Top-left) Video frame I_{t-1}; (Top-right) region labelling of I_{t-1} following multi-label graph cut; (Bot.-left) region labels warped according to per-pixel motion flow field $I'_{t-1} \rightarrow I_t$ —for example, note the shift of the boy's left glove. (Bot.-right) Thinning yields prior labels for the segmentation of I_t.*

The pair-wise term $V(.)$ encourages coherence in region labelling and discontinuities to occur at high contrast locations, which is computed using RGB color distance:

$$V(\mathcal{L}, \mathcal{P}) = \gamma \sum_{(m,n) \in N} [l(m) \neq l(n)] e^{-\beta}. \tag{3}$$

where N is the set of pairs of 8-connected neighboring pixels in $\mathcal{P}$, and $[l(m) \neq l(n)]$ is a binary term measuring label equivalence of pixels m and n. β is chosen to be contrast adaptive as in [Boykov and Funka-Lea 2006]:

$$\beta = \frac{1}{2} \langle ||I(m) - I(n))||^2 \rangle^{-1}. \tag{4}$$

Constant γ is a versatile setting for a variety of images [Blake et al. 2004], and is set empirically to obtain satisfactory segmentation.

Motivated by the data term in [Boykov and Funka-Lea 2006] we enforce hard constraints on the motion propagated prior labels assigned to label l_i, by setting the data term of $p \in \mathcal{O}$ to be:

$$U_{p:\{p \in \mathcal{O}\}} = \begin{cases} 0 & \text{if } l(p) = l_i; \\ \infty & \text{if } l(p) \neq l_i. \end{cases} \tag{5}$$

Optimizing (1) to yield an appropriate assignment of labels to pixels is NP-hard, but an approximate solution can be computed by treating the optimization as a multi-label graph cut and solving this using the expansion move algorithm of [Boykov et al. 2001]. An α-expansion iteration is a change of labeling such that p either retains its current value or takes the new label l_α. The expansion move proceeds by cycling the set of labels and performing an α-expansion iteration for each label until (1) cannot be decreased [Boykov et al. 2001]. Each α-expansion iteration can be solved exactly by performing a single graph-cut using the min-cut/max-flow as described in [Boykov and Kolmogorov 2004]. Convergence to a strong local optimum is usually achieved in 3-4 cycles of iterations over our label set. We improve the computation and memory efficiency of each iteration by dynamically reusing the flow at each iteration of the min-cut/max-flow algorithm (after [Alahari et al. 2008]). This results in a speed-up of an order of two.

3.2 Region propagation

The segmentation of I_t described in sub-Sec 3.1 is dependent on the information propagated from the previous frame at $t-1$; specifically: i) the color models for regions Θ; ii) the set of pixels $\mathcal{O} \subseteq \mathcal{P}$ and their corresponding label assignments at $t-1$. We now explain the propagation process in detail.

Our approach is to estimate the motion of pixels in frame I_{t-1}, and translate those pixels and their respective label assignments from the previous frame to the current frame (I_t). Motion is estimated using a model of *rigid motion plus deformation*.

We first estimate a global affine transform between successive frames I_{t-1} and I_t, using a RANSAC search based on SIFT features [Lowe 2004] matched between the frames. Performing an affine warp on I_t and the corresponding region map compensates for large rigid (e.g. camera) motion, resulting in a new image $I^{'}_{t-1}$. Local deformations are captured by estimating smoothed optical flow [Black and Anandan 1993] between $I^{'}_{t-1}$ and I_t, independently within each region. Note that we do not assume or require accurate motion estimation at this stage. Figure 3 (bot.-left) provides an example region map from the BOY sequence $t-1$ warped according to motion field $I^{'}_{t-1} \rightarrow I_t$.

We select a subset of the motion propagated pixels (written $\mathcal{O}$), and their corresponding region assignments, as prior labels to influence the segmentation of I_t. To mitigate the impact of imprecise motion estimation, we form $\mathcal{O}$ by sampling from a morphologically thinned skeleton of the motion propagated regions (Figure 3, bot.-right). This approach is inspired by the “scribbles” used in the interactive Grab-Cut system of [Blake et al. 2004], but note that we perform an automatic and multi-region (as opposed to binary) labelling. The skeleton emphasizes geometrical and topological properties of the motion propagated region map, such as its connectivity, topology, length, direction, and width. To further deal with the uncertainties in positions which are closer to the estimated region boundary, we use only the skeletons whose distance to the boundary exceeds a pre-set confidence. Figure 3 illustrates the complete process, which we find to be tolerant to moderate misalignments caused by inaccurate motion estimation.

We build a GMM color model for each region l_i, sampling the historical colors of labelled pixels over recent frames. To cope with variations in luminance often present in the sequence, the proportion of samples $S_{l_i,t-d} \in [0,1]$ ($d > 0$) drawn from all l_i-labeled pixels from historical frame I_{t-d} decreases exponentially as the temporal distance d from the current frame I_t increases (Figure 4):

$$S_{l,t-d} \propto e^{-d^2/\sigma_d^2}. \qquad (6)$$

Our system selects a smaller σ_d when luminance variance is large, contributing more recent data to the GMM, otherwise the historical data contributes more to increase robustness.

3.3 Refining region labels

The method of sub-Sec. 3.1 labels I_t with some or all of the region labels in use in the region map at $t-1$. However, new objects may appear in the sequence over time I_t due to occlusion effects of objects moving into shot. This is most apparent in clips such as DRAMA (Fig 11). These objects may warrant introduction of a new region label, should they differ in color from existing regions. In such a situation, pixels comprising the object are erroneously labelled from the existing label set by the graph cut optimization, which in turn perturbs the color distribution of the region. We can detect this by measuring the χ^2 distance (as defined in [Hall and Hicks 2004]) between the GMM of a region at time t and the historical GMM built over time (Figure 4).

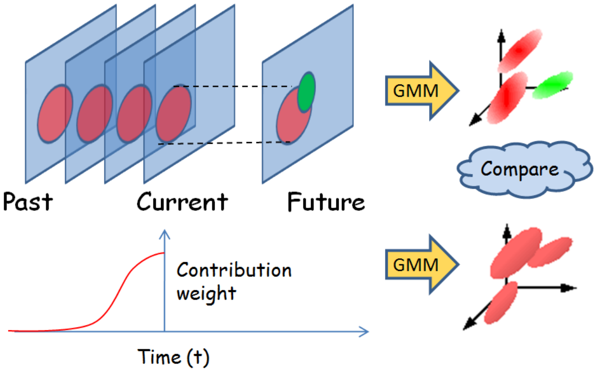

Figure 4: *A GMM color model of each region is built incrementally over time, with contributions biased toward more recent observations. If the GMM of a region abruptly changes color distribution (χ^2 metric) then the region is re-segmented (sub-Sec 3.3).*

For successive frames, we keep two sets of color models for each label l in frame I_t being processed: (1) Historical color models associated with each label $M^h_{l:\{l\in\mathcal{L}\}} := G_l(I_{t-4}, I_{t-3})$ and (2) an updated color model $M^u_{l:\{l\in\mathcal{L}\}} := G_l(I_t)$. We set a guard interval of two frames between those two models to detect a significant change. If the χ^2 distance between these two models exceeds a threshold, new objects are deemed present.

To build color models for the new objects we extract the dominant modes of colors within the region. We apply mean-shift to perform unsupervised clustering on the spatial-color modes (XY+RGB) of pixels in the region. This yields a localized segmentation of pixels in the region. We extend our label set to accommodate each new region arising from the mean-shift segmentation, and for each new region also compute GMM color models and region skeletons as in sub-Sec. 3.2. Re-applying the graph cut optimization locally within the region, using these new labels and constraints, yields an improved segmentation for I_t that is carried to successive frames.

3.4 Smoothing and filtering

Our segmentation algorithm produces stable region maps, but due to visual ambiguities in poor contrast areas, the location of region boundaries tend to oscillate in position by a few pixels. We can attenuate this effect by performing spatio-temporal smoothing. Specifically, by coherently labelling regions in adjacent frames, we have formed a set of space-time volumes. Applying a fine scale (3×3×3) Gaussian filter removes boundary noise. We avoid removing detail by only filtering volumes above a certain size.

We inspect the duration $d_{l,k}$ of the disconnected video objects k ($k = 1 \ldots K_{obj_l}$) with the same label l, in a time window of 24 frames (1 second). If the duration of any of these disconnected video object within this time window is shorter than a length

$$D_{l:\{l\in\mathcal{L}\}} = \min\{\max_{k\in\{1\ldots K_{obj_l}\}} d_{l,k}, \tau_r\}. \qquad (7)$$

this video object is removed. τ_r is set to be six frames (about 1/4 second). The effect of this process is that the spurious volumes due to false segmentation and short-lived objects are removed, as shown in Figure 5. The “holes” left by filtering and smoothing are filled by extrapolating region labels from immediate space-time neighbors on a nearest-neighbor basis.

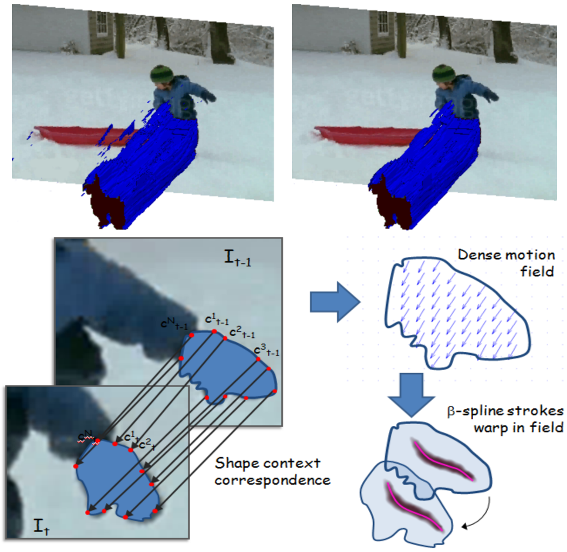

Figure 5: *Above: co-labelled regions are smoothed in space-time to remove any spurious regions. Below: Brush strokes are painted on a stable reference frame, created by corresponding co-labelled regions in adjacent frames and interpolating a dense motion field.*

3.5 Stroke Placement and Shading

Our segmentation algorithm ensures regions not only deform in a coherent manner, but are also labelled consistently between frames. This space-time description of scene structure may be rendered in a variety of artistic styles; here we give an example of one shading and one stroke based style.

3.5.1 Cartooning

Superimposing black edges over regions shaded with their mean pixel color can produce coherent cartoon effects (Figure 8). In our cartoon examples, a mask of inter-region boundaries is produced for each frame. We identify "junction" points on region boundaries by identifying 3×3 pixel windows containing > 2 region labels - and remove the corresponding boundary fragments from the mask. This results in a series of connected pixel chains that we transform into β-spline strokes by sampling knots at equi-distant intervals. The strokes are rendered as dark brush strokes, with thickness proportional to stroke length (after [Wang et al. 2004]) tapering toward the stroke ends. We render frames independently without further post-processing; this is both for simplicity and to demonstrate the temporal coherence of our segmentation output.

We can also exploit the temporally corresponded region labelling to differentially render regions of interest. For instance, users are particularly sensitive to over-abstraction of detail in faces; commonly present in home video footage. We run human face detection [Viola and Jones 2004] over frames to identify labelled regions likely to contain faces. Internal detail in these regions may be restored by blending in a posterized image of underlying video footage, and detail further enhanced by subtracting a Laplacian of Gaussian (LoG) filtered image from the result.

3.5.2 Painterly Rendering

Alternatively we can paint β-spline brush strokes inside regions, coherently deforming those splines by warping their control points to match the motion of the region boundary (similar to the manually bootstraped rotoscoping system of [Agarwala et al. 2004]). Boundary correspondences are computed between temporally adjacent, co-labelled regions using Shape Contexts [Belongie et al. 2002]. The set of N corresponded boundary locations $\{< c^1_{t-1}, c^1_t >, < c^2_{t-1}, c^2_t >, ..., < c^N_{t-1}, c^N_t >\}$ is used to derive the motion vector for a control point p at time t as:

$$p = \frac{1}{N}\sum_{i=1}^{N}\omega(p, c^i_t)|c^i_t - c^i_{t-1}|. \qquad (8)$$

where $\omega(.)$ is a Gaussian weighted function of the shortest distance between two points within the region (see Figure 5, below). Our coherent segmentation promotes smooth deformation of region shape, and so flicker-free motion of brush strokes.

We paint the β-spline strokes within a region using Hertzmann's bidirectional stroke growth algorithm [Hertzmann 1998]. In the original algorithm, strokes are grown from random seed points using the orientation of an intensity gradient field computed from the underlying image. However, computing such orientation directly from video footage typically promotes incoherence. Instead, we interpolate an orientation field from the shape of the region. Orientations are locally obtained at points of correspondence on the boundary $\theta[x, y] \mapsto \mathrm{atan}(c^{i-1}_t - c^i_t)$. We define a dense orientation field Θ_Ω over all coordinates within the region $\Omega \in \Re^2$, minimizing:

$$\underset{\Theta}{\mathrm{argmin}} \int\int_{\Omega} (\bigtriangledown\Theta - \theta)^2 \quad s.t. \;\; \Theta|_{\delta\Omega} = \theta|_{\delta\Omega}. \qquad (9)$$

i.e. $\triangle\Theta = 0$ over Ω s.t. $\Theta|_{\delta\Omega} = \theta|_{\delta\Omega}$ for which a discrete solution was presented in [Perez et al. 2003] solving Poisson's equation with Dirichlet boundary conditions. Examples of painterly output are given in Figure 12.

4 Video Sequencing

We next explain the process for sequencing videos from the user's collection, and creating the animated transitions between clips.

4.1 Stochastic Composition

Our DAD visualizes the users' video collection by selecting and rendering video clips to form a perpetual sequence of stylized video. We desire a temporal composition in which successive clips have some semantic connection - for example a clip of a birthday in a particular year follows a birthday from a previous year. Or, a clip of the family at a birthday party is followed by another clip of the family, perhaps on vacation.

We decide the semantic relevance of two clips using keyword ("tags"); textual meta-data that is manually annotated against each clip at the time of capture. These tags are drawn from the ontology depicted in Figure 6 (top), which relates these tags through high-level "concepts" relevant to household video collections (such as Event, Place or People present in the clip). We define the semantic distance between two tags as the number of concept nodes "hops" apart those tags are in our ontology. Our experimental video collection contains of 23 home videos covering a variety of topics, tagged as "Christmas", "Birthday", "Childhood", "Family", "Beach" and "Countryside" (Figure 6, bottom). The semantic relevance of two clips $d_s(A, B)$ is determined by computing the semantic distance between all pairings of tags (where one tag is from A, and one from B), and finding the minimum distance from that set of pairings.

We also wish to transition between visually similar clips, to avoid sharp changes in color that may prove distracting in the ambient setting. We measure this by assessing the similarity of spatial color

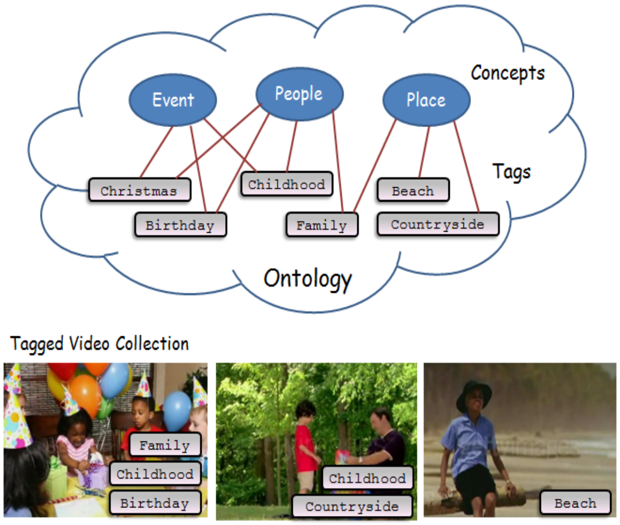

Figure 6: *Ontology: video tags are categorized into high level concepts. User videos are annotated with tags from the ontology.*

distribution between the final and initial frames of the current and next clip respectively. This is achieved using the image retrieval algorithm of [Jacobs et al. 1995], which ranks images according to similarity of significant wavelet coefficients within each color channel. In our system we treat the final frame of the previous frame i as a "query", and use the retrieval algorithm to assign a rank $d_v(i, j)$ to each clip j in the collection according to the similarity of its start frame.

Given a collection of videos $\mathcal{V} = \{v_1, v_2, ..., v_n\}$ we generate a sequence of clips stochastically, as a random walk over a directed graph $\{\mathcal{V}, \mathcal{E}\}$ comprising nodes $\mathcal{V}$ with edge weight $e_{ij} = P_(v_i|v_j)$ s.t. $\sum_{k=1}^{n} e_{kj} = 1$ indicating the transition probability between clip subject to semantic constraints and visual similarity:

$$P_(v_i|v_j) \propto \frac{1}{1 + d_s(v_i, v_j)} e^{-d_v(v_i,v_j)^2/n^2}. \quad (10)$$

As in [Kovar et al. 2002; Schodl et al. 2000] we introduce procedures to avoid dead-ends in the graph. To mitigate against the reappearance of recently displayed videos, as each node is visited, it is removed from the graph. When the system finds itself in a dead end i.e. at node v_i we observe $\sum_{k=1}^{n} e_{ki} = 0$, all nodes in the graph are restored. In our experiments we thresholded d_s at 1 to constrain successive clips to belong to the same ontological concept, with a bias to successive clips sharing tags. The use of larger ontologies and more diverse tag vocabularies may benefit from removal of this constraint.

4.2 Rendering Transitions

Having established a mechanism to stochastically select the next video clip during visualization, we animate the transition between the current and next clips according to scene structure.

We first establish a mapping between each region R^j_{t-1} and R^i_t corresponding, respectively, to the final and initial frames of the two clips. This is performed in a greedy manner, iteratively pairing off regions that minimize:

$$\operatorname*{argmin}_{\{i,j\}} \left[\begin{array}{ccc} w_1 & w_2 & w_3 \end{array} \right] \left[\begin{array}{c} C(R^j_{t-1}, R^i_t) \\ A(R^j_{t-1}, R^i_t) \\ S(R^j_{t-1}, R^i_t) \end{array} \right] \quad (11)$$

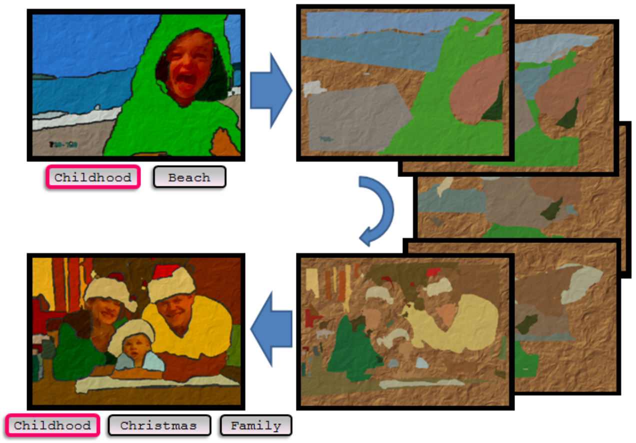

Figure 7: *Frames from the transition animation between two clips.*

where the normalized functions: $C(.)$ indicates mean color similarity; $A(.)$ indicates relative area; $S(.)$ indicates shape similarity in terms of region compactness. We bias weights ω_{1-3} empirically to 0.5, 0.4, 0.1. The greedy assignment continues until (11) falls below a threshold. Unassigned regions in the mapping are animated to "disappear" (shrink to a point at the centroid) or "appear" (grow from the centroid); whereas regions mapped between frames are animated to morph into one another.

Regions are morphed using simple linear blending. Each region is vectorized into a polygon and a series of regularly spaced control vertices established on the boundary. A correspondence is established between vertices of R^j_{t-1} and R^i_t to minimize distance between corresponded vertices. The position of control vertices are linearly blended over time (typically $\frac{1}{4}$ second) to animate the region from one shape to another. Region color is similarly blended. Although more complex vertex correspondence approaches were investigated [Belongie et al. 2002], these lacked stability when presented with moderate changes in region shape. The resulting transitions are shown in Figure 7 and the accompanying video.

5 Results and Discussion

We demonstrate the DAD using a manually tagged collection of 23 videos. Figure 8 shows representative frames of the stylized footage in both cartoon and painterly styles; 6 minutes of the perpetual animated display is also included in the supplementary material. The sequence of clips for the first few transitions is indicated in Figure 9 and an example transition animation given in Figure 7. In most cases videos sharing semantically similar tags are selected, although variation within an ontological concept is also evident e.g. 'countryside' to 'beach'. Sequencing decisions are also influenced by visual similarity, biased primarily toward background color. The resulting clip transition animations match large, similarly colored regions between frames producing a pleasing smooth transition effect evident throughout the DAD sequence.

To demonstrate the advantages of the proposed multi-label video segmentation algorithm, we compare the approach proposed in Section 3 to two leading segmentation methods for per-frame [Comaniciu and Meer 2002] and spatio-temporal [Paris 2008] segmentation (Figure 10). We observe the region boundaries in the proposed method to exhibit improved stability over time. Figure 11 indicates the region maps produced by the segmentation algorithm over four video sequences. We test our algorithm on fast moving

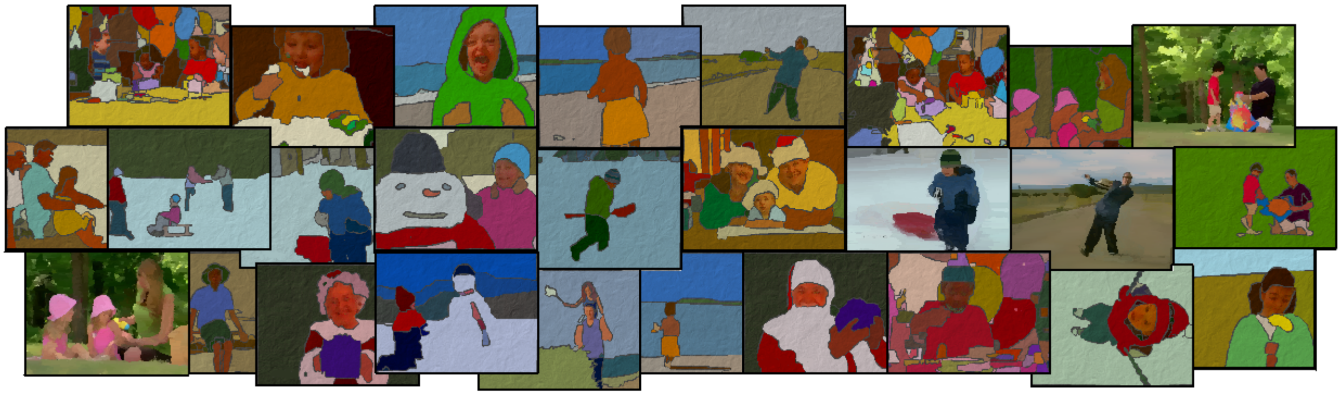

Figure 8: *A collage of stylized frames sampled from the user video collection studied in this paper.*

footage containing small objects ("BEAR" from [Collomosse et al. 2005]). Unlike previous work, fine scale features (e.g. the bear's eyes and nose) are retained. Similarly, "DANCE" demonstrates the ability to cope with fast motion and partial occlusions. "DRAMA" shows correct handling of regions that disappear and appear within sequences, the latter detected by changes in the region color distribution and addressed as out-lined in sub-Sec 3.3. The "KITE" sequences shows the aesthetic ability to selectively abstract detail (trees) from the stylized video, when interactively removed by the user in the initial frame. In all cases our segmentations appear flicker-free; some flicker is occasionally present the bottom-left of clips due to the frame identifier which could be manually abstracted away by modifying the initial frame in a similar way.

Following coherent segmentation, Figs. 12 shows frames of painterly renderings over four video sequences in natural scenes. The smooth deformation of regions enables stable and flicker-free motion of brush strokes, which produces an aesthetically pleasing painterly effect over the input video sequences.

6 Conclusion

We have presented a *Digital Ambient Display* (DAD) that harnesses artistic stylization to create an abstraction of user's experiences through their home video collections. The DAD automatically selects, stylizes and transitions between clips enabling users to passively consume their video collections and rediscover past memories.

The main technical contribution of our paper is a novel algorithm for coherent video segmentation based on multi-label graph cut, and its applications to stylized animation in the DAD. By parsing the video into coherent spatial segments, we are able to represent scene structure. This representation allows us to establish correspondence between frames, enabling the coherent stylization of video objects with both shading and painterly effects. The latter is possible by painting brush strokes on a smoothly deforming reference frame. We are also able to create interesting transition effects between different video clips using region correspondence.

The system can be further enhanced by exploring the backward propagation of region labels to further improve coherence of segmentation. We would also like to improve the painterly rendering by differentiating between region motion caused by occlusion vs. object deformation, to more closely align the movement of painted strokes to the perceived structure in the scene.

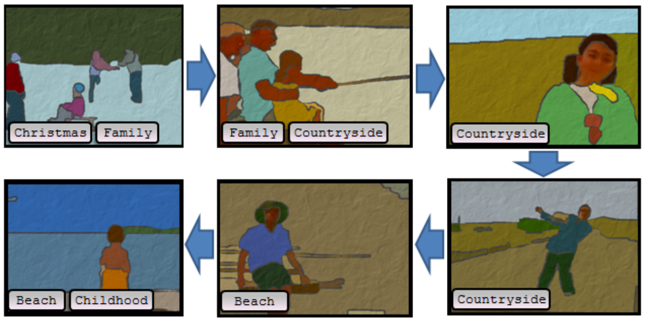

Figure 9: *Illustrating video sequencing based on semantic and visual similarity (full sequence in supplementary video).*

Regarding the temporal composition of clips into a DAD sequence, we currently take a first order approach to clip selection; considering only the previous clip when predicting the next. We would like to explore graph optimization algorithms similar to [Kovar et al. 2002] to plan routes through a subset of clips e.g. to encompass a theme such as "family vacations" rather than traversing the whole database. This may improve scalability of the semantic similarity to larger collections. Other extensions might exploring automatic meta-data annotation on users' video collection, e.g. by extending the photo categorization method presented in [Ruiz et al. 2003].

Acknowledgement

This work was supported by HP Labs IRP funded studentship #474: "Artistic Rendering of Consumer Video".

References

AGARWALA, A., HERTZMANN, A., SALESIN, D., AND SEITZ, S. 2004. Keyframe-based tracking for rotoscoping and animation. In *Proc. ACM SIGGRAPH*, 294–302.

ALAHARI, K., KOHLI, P., AND TORR. 2008. Reduce, reuse & recycle: Efficiently solving multi-label MRFs. In *CVPR*, 1–8.

ARKSEY, N. 2007. *Exploring the Design Space for Concurrent Use of Personal and Large Displays for In-Home Collaboration*. Master's thesis, University of British Columbia.

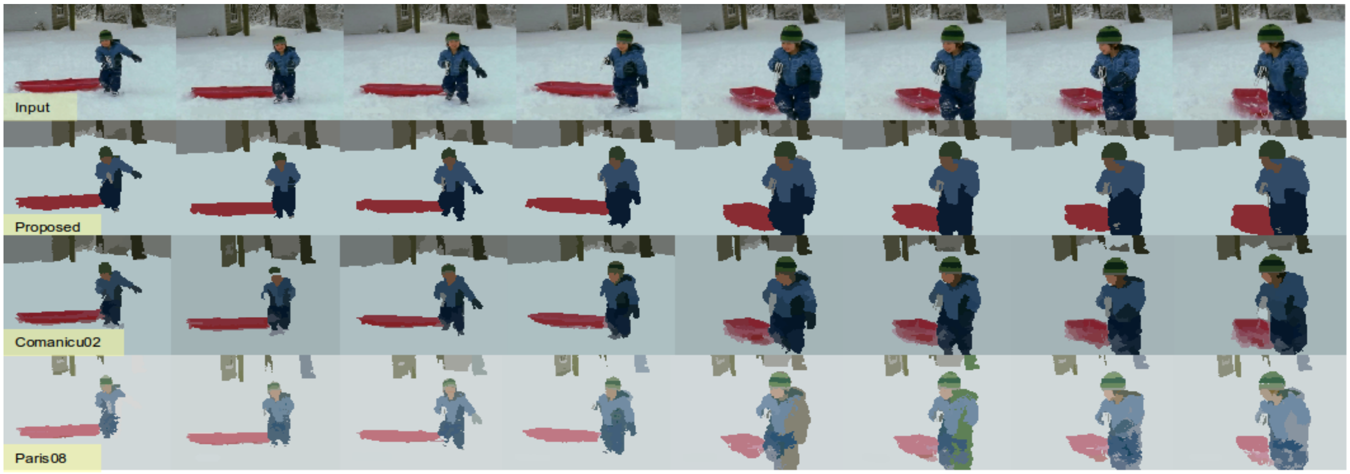

Figure 10: *Comparing the accuracy and coherence of our segmentation algorithm on the BOY sequence, to 'synergistic' mean-shift + edge (Comaniciu, 2002) and a state of the art spatio-temporal method (Paris, 2008). Boundaries are less prone to variation in shape and topology.*

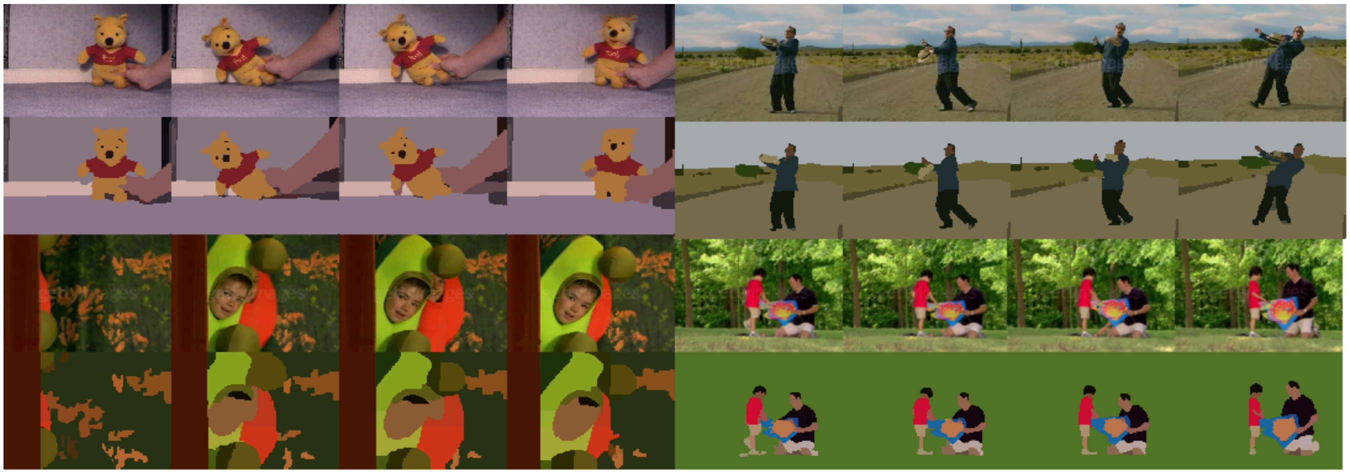

Figure 11: *Illustrating the coherent region maps produced by our segmentation method. Top: BEAR and DANCE contain small regions moving quickly over time. Bottom: The DRAMA sequence shows correct handling of of regions appearance. The KITE sequence indicates how background detail may (optionally) be abstracted by modifying the initial frame segmentation to merge unwanted detailed regions.*

Bai, X., Wang, J., Simons, D., and Saprio, G. 2009. Video snapcut: Robust video object cutout using localized classifiers. In *Proc. ACM SIGGRAPH*.

Belongie, S., Malik, J., and Puzicha, J. 2002. Shape matching and object recognition using shape contexts. *IEEE Trans. Pattern Anal. Machine Intel. (PAMI) 24*, 509–521.

Bizzocchi, J. 2008. Winterscape and ambient video – an intermedia border zone. In *Proc. ACM Multimedia*.

Black, M., and Anandan, P. 1993. A framework for the robust estimation of optical flow. In *ICCV*, 231–236.

Blake, A., Rother, C., Brown, M., Prez, P., and Torr, P. 2004. Interactive image segmentation using an adaptive gmmrf model. In *ECCV*, 428–441.

Bousseau, A., Neyret, F., Thollot, J., and Salesin, D. 2007. Video watercolorization using bidirectional texture advection. In *Proc. ACM SIGGRAPH*, 1–7.

Boykov, Y., and Funka-Lea, G. 2006. Graph cuts and efficient n-d image segmentation. *IJCV 2*, 70, 109–131.

Boykov, Y., and Kolmogorov, V. 2004. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Trans. Pattern Anal. Machine Intel. (PAMI) 26*, 1124–1137.

Boykov, Y., Veksler, O., and Zabih, R. 2001. Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Anal. Machine Intel. (PAMI) 23*, 1222–1239.

Collomosse, J., and Hall, P. 2003. Cubist style rendering from photographs. *IEEE Trans. Visualization and Comp. Graphics (TVCG) 9*, 4, 443–453.

Collomosse, J., Rowntree, D., and Hall, P. 2005. Stroke surfaces: Temporally coherent artistic animations from video. *IEEE Trans. Visualization and Comp. Graphics (TVCG) 11*, 540–549.

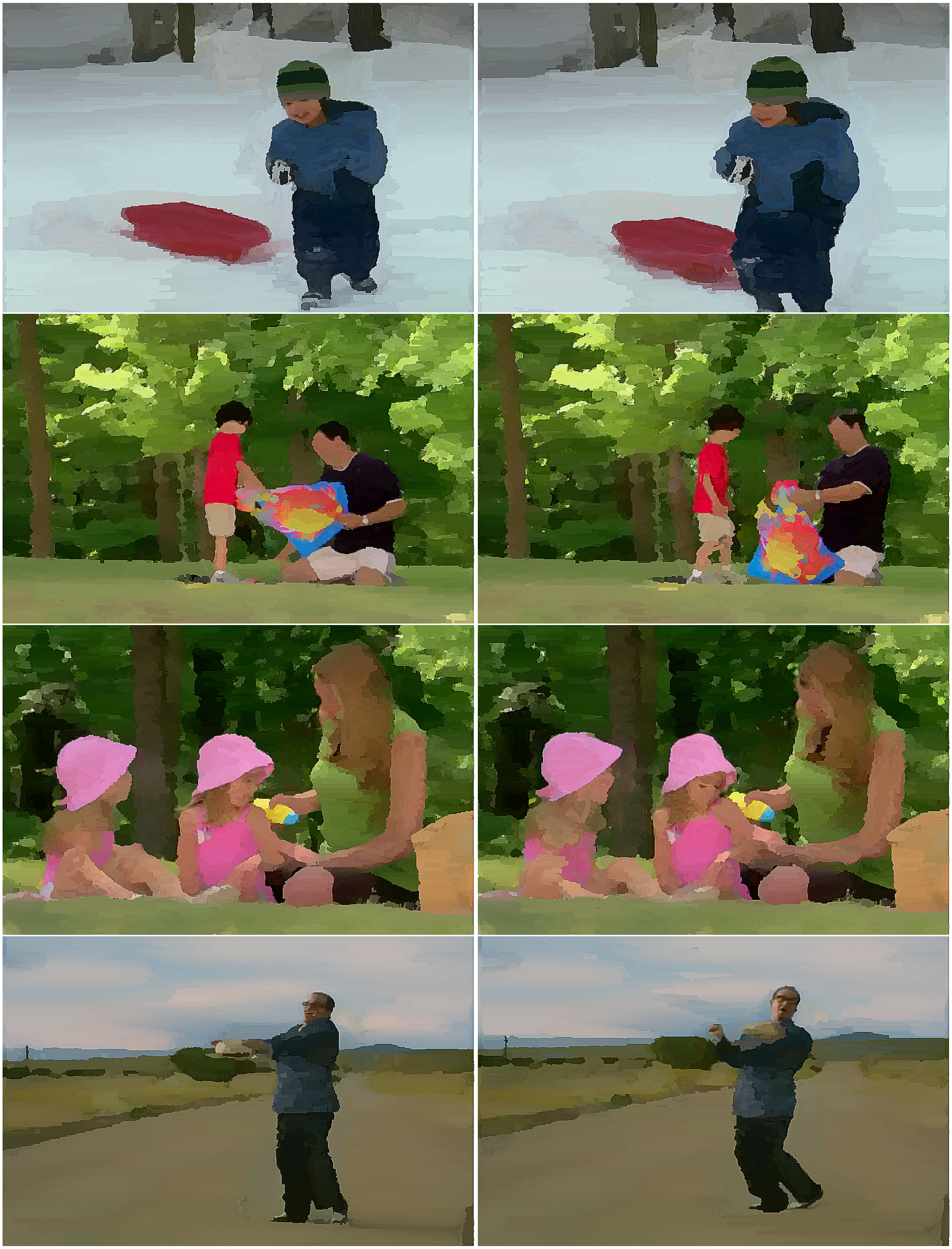

Figure 12: *Examples of coherent painterly renderings produced from the BOY, KITE, PICNIC and DANCE videos (top to bottom).*

Collomosse, J. 2004. *Higher Level Techniques for the Artistic Rendering of Images and Video*. PhD thesis, University of Bath.

Comaniciu, D., and Meer, P. 2002. Mean shift: A robust approach toward feature analysis. *IEEE Trans. PAMI 24*, 603–619.

DeCarlo, D., and Santella, A. 2002. Abstracted painterly renderings using eye-tracking data. In *Proc. ACM SIGGRAPH*, 769–776.

Hall, P. M., and Hicks, Y. 2004. CSBU-2004-03: A method to add gaussian mixture models. Tech. rep., Univ. Bath.

Hays, J., and Essa, I. A. 2004. Image and video based painterly animation. In *Proc. ACM NPAR*, 113–120.

Hertzmann, A., and Perlin, K. 2000. Painterly rendering for video and interaction. In *Proc. ACM NPAR*, 7–12.

Hertzmann, A. 1998. Painterly rendering with curved brush strokes of multiple sizes. In *Proc. ACM SIGGRAPH*, 453–460.

Jacobs, C., Finkelstein, A., and Salesin, S. 1995. Fast multiresolution image querying. In *Proc. ACM SIGGRAPH*, 277–286.

Kovar, L., Gleicher, M., and Pighin, F. 2002. Motion graphs. In *Proc. ACM SIGGRAPH*, 473–482.

Kyprianidis, J.-E., Kang, H., and Doellner, J. 2009. Image and video abstraction by anisotropic kuwahara filtering. In *Proc. Pacific Graphics*, vol. 28.

Litwinowicz, P. 1997. Processing images and video for an impressionist effect. In *Proc. ACM SIGGRAPH*, 407–414.

Lowe, D. 2004. Distinctive image features from scale-invariant keypoints. *IJCV 60*, 91–110.

Meier, B. J. 1996. Painterly rendering for animation. In *Proc. ACM SIGGRAPH*, 477–484.

Paris, S. 2008. Edge-preserving smoothing and mean-shift segmentation of video streams. In *ECCV*, 460–473.

Perez, P., Gangnet, M., and Blake, A. 2003. Poisson image editing. In *Proc. ACM SIGGRAPH*, 313–318.

Price, B., Morse, B., and Cohen, S. 2009. Livecut: Learning-based interactive video segmentation by evaluation of multiple propogated cues. In *ICCV*.

Ruiz, D., Takahashi, H., and Nakajima, M. 2003. Image categorization using color blobs in a mobile environment. In *Proc. Eurographics*, 427–432.

Schodl, A., Skeliski, R., Salesin, D., and Essa, H. 2000. Video textures. In *Proc. ACM SIGGRAPH*, 489–498.

Slatter, D., Cheatle, P., and Greig, D. 2010. Faces from the web: Automatic selection and composition of media for casual screen consumption and printed art-work. In *Proc. SPIE*.

Viola, P., and Jones, M. 2004. Robust real-time object detection. *Intl. Journal. Computer vision (IJCV) 57*, 2, 137–154.

Wang, J., Xu, Y., Shum, H., and Cohen, M. 2004. Video tooning. In *Proc. ACM SIGGRAPH*, vol. 23, 574–583.

Wang, T., Mansfield, A., Hu, R., and Collomosse, J. 2009. An evolutionary approach to automatic video editing. In *Proc. 6th European Conf. on Visual Media Production (CVMP)*.

Winnemoller, H., Olsen, S., and Gooch, B. 2006. Realtime video abstraction. In *Proc. ACM SIGGRAPH*, 1221–1226.

Xiao, J., Zhang, X., Cheatle, P., Gao, Y., and Atkins, C. 2008. Mixed-initiative photo collage authoring. In *Proc. ACM Multimedia*, 509–518.

You, W., Feis, S., and Lea, R. 2008. Studying vision-based multiple-user interaction with in-home large displays. In *Proc. 3rd ACM workshop on Human-Centred Computing (HCC)*, 19–26.