MULTI-LABEL PROPAGATION FOR COHERENT VIDEO SEGMENTATION AND ARTISTIC STYLIZATION

Tinghuai Wang[1]*, Jean-Yves Guillemaut, and John Collomosse*

Centre for Vision, Speech and Signal Processing
University of Surrey, Guildford, Surrey, UK.

ABSTRACT

We present a new algorithm for segmenting video frames into temporally stable colored regions, applying our technique to create artistic stylizations (e.g. cartoons and paintings) from real video sequences. Our approach is based on a multi-label graph cut applied to successive frames, in which the color data term and label priors are incrementally updated and propagated over time. We demonstrate coherent segmentation and stylization over a variety of home videos.

Index Terms— Video segmentation, Graph cut, Temporal coherence, Non-photorealistic (artistic) rendering / NPR.

1. INTRODUCTION

Artistic stylization of visual footage is a challenging problem to both Computer Vision and Graphics, requiring the automatic abstraction and depiction of scene structure. Video stylization remains an open problem; it is difficult to extract *temporally coherent* (stable) scene descriptions, and this typically leads to a distracting flicker within resulting animations.

In recent years, video segmentation and region tracking have been applied to yield *mid-level* models of scene structure [1] that have shown promise in artistic stylization [2, 3] and summarization [4]. Predominant approaches to video segmentation are frame-to-frame association ($2D + t$) and spatio-temporal clustering ($3D$) methods. The former independently segment 2D frames, and then create associations between regions over time to identify sporadic regions [5, 2, 6]. Although this filtering 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. Spatio-temporal approaches cluster pixels in (x, y, t), using unsupervised clustering techniques such as mean-shift [7, 3, 8], or Gaussian mixture models (GMM) [9] to group space-time pixels. However, these approaches become computationally infeasible for pixel counts in even moderate size videos, and often under-segment small or fast moving objects that form disconnected space-time volumes.

We propose a video segmentation algorithm, in which the region segmentation of each frame is guided by motion flow propagated priors estimated from the accumulated data of past frames. By exploiting historical information we demonstrate improvements in temporal coherence. Although recent interactive "video cut-out" systems [10, 11] track keypoints on region boundaries over time for matte extraction, 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, [10, 11] require regular manual correction, typically every 2-5 frames. Our algorithm requires no user interaction, beyond (optional) modification of the initial frame for aesthetics.

2. SEGMENTATION FRAMEWORK

The essence of our approach is to perform a multi-label graph cut on successive video frames, using information propagated forward from previous frames. This information comprises: i) a color distribution for each region represented via a Gaussian Mixture Model (GMM) built incrementally from past frames; ii) a subset of pixel-to-region labels from the previous frame. We check for region under-segmentation (e.g. due to the appearance of new objects) by comparing the historic and updated GMMs and introducing new labels accordingly. The region map of the first frame is boot-strapped using mean-shift, and may *optionally* be modified by the user for aesthetics. When all frames are segmented, we render the region maps in a variety of painterly or cartoon artistic styles.

2.1. Multi-label Graph cut

We define 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. finding 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 $l_i, i = \{1...L\}$. A subset of $\mathcal{L}$ are carried forward from the region map at $t-1$, via a propagation process described shortly (subsec. 2.2). This *prior labelling* of pixels ($\mathcal{O} \subseteq \mathcal{P}$) forms a hard constraint on the assignments of remaining pixels, 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:

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

[1] Funded by the Hewlett-Packard Labs IRP Programme.

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:

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

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 (subsec. 2.2).

The contrast term $V(.)$ encourages coherence in region labelling, and is computed using RGB color distance:

$$V(\mathcal{L},\mathcal{P}) = \gamma \sum_{(m,n)\in N} [l(m) \neq l(n)] e^{-\beta ||I(m)-I(n)||^2}. \quad (3)$$

where N is the set of pairs of 4-connected neighboring pixels in $\mathcal{P}$. β is chosen to be contrast adaptive as in [12]:

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

Constant γ is a versatile setting for a variety of images [13], and is set empirically to obtain satisfactory segmentation.

Motivated by the data term in [12] 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:

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

Optimizing (1) is NP-hard, but an approximate solution can be computed using the expansion move algorithm[14]. 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 [14]. Each α-expansion iteration can be solved exactly by performing a single graph-cut using min-cut/max-flow[15]. Convergence to a strong local optimum is usually achieved in 3-4 cycles of iterations over the label set.

2.2. Region propagation

Segmentation of I_t is dependent on the region map for I_{t-1}; specifically: i) the color models for regions Θ; ii) the set of pixels $\mathcal{O} \subseteq \mathcal{P}$ and corresponding label assignments at $t-1$. We now explain how this information is propagated.

From successive frames I_{t-1} and I_t, we first estimate a global affine transform using RANSAC and SIFT features. Affine warping both I_{t-1} and the corresponding region map compensates for large rigid (e.g. camera) motion, resulting in a new image I'_{t-1}. Local deformation is captured by estimating smoothed optical flow [16] 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.

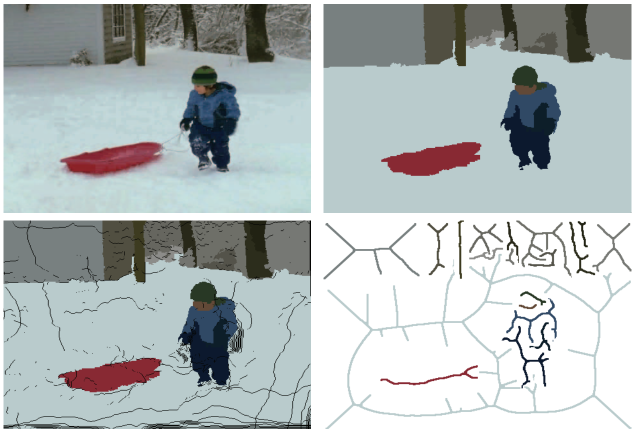

Fig. 1. Illustrating segmentation and prior propagation: (TL) Video frame I_{t-1}; (TR) region labelling of I_{t-1}; (BL) labels warped according to motion flow field $I_{t-1} \rightarrow I_t$ — note the boy's left glove. (BR) Extracted priors for segmentation of I_t.

We select a subset of the motion propagated pixels $\mathcal{O}$, and their corresponding region assignments, as prior labels to influence the segmentation of I_t. To account for the impact from imprecise motion estimation, we form $\mathcal{O}$ by sampling from a morphologically dilated skeleton of each region. This is inspired by the "scribbles" used in interactive GrabCut [13], but note that we perform automatic, multi-region (as opposed to binary) labelling. The skeleton emphasizes geometrical and topological properties of the region, 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. Fig. 1 illustrates our process, which is tolerant to moderate misalignment caused by inaccurate motion estimates.

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 increases:

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

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.

2.3. Refining region labels

The method of subsec. 2.1 labels I_t with some or all of the region labels in use at $t-1$. However, new objects may appear in the sequence over time I_t due to occlusion effects; This is most apparent in DRAMA (Fig 4). 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, perturbing the color distribution of the region. We can detect this by measuring the χ^2 distance (as defined in [17]) between the GMM of a region at time t and the historical GMM built over time. If the χ^2 distance 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 localised segmentation of pixels in the region. We extend our label set to accommodate each new region, and for each region also obtain color models and region skeletons as in subsec. 2.2. Re-applying graph cut within the region, using these new constraints, yields an improved segmentation for I_t.

3. VIDEO STYLIZATION

Following coherent segmentation (Sec. 2), we may post-process the region map to render a variety of stylized effects. Region boundaries in poor contrast areas may oscillate in position by a couple of pixels. However, by coherently labelling regions in adjacent frames, we have formed a set of space-time volumes. Applying a fine scale ($3\times3\times3$) Gaussian filter removes this noise; we avoid removing detail by only filtering volumes above a certain size (Fig. 2, right).

Superimposing black edges over regions shaded with their mean pixel color can produce coherent cartoon effects (Fig. 2, bottom). Edges may be drawn around region boundaries (left), or at maxima of a DoG field [18] (middle). Eroded regions and a textured canvas give the impression of watercolor (right). We can also exploit the "tracked" regions to create more sophisticated *rotoscoped* effects. In Fig. 2 (left) we paint β-spline strokes within each region and transform control points to match region deformation. We achieve this by registering temporally adjacent, co-labelled regions (via Shape Contexts) to obtain a sparse vector field $\mathbf{v}$ from boundary correspondences. Stroke control points are translated according to a dense vector field $f \in \Re^2$, obtained by minimizing $\int\int |\bigtriangledown f - \mathbf{v}|^2$ via Poisson's equation [19]. Our coherent segmentation promotes smooth deformation of region shape, and so flicker-free motion of brush strokes.

4. RESULTS AND DISCUSSION

Figs. 3,4 show coherent segmentation over five video sequences. Fig 3 compares our automatic approach to two leading methods; for per-frame [20] and spatio-temporal [21] segmentation. The region boundaries from our proposed method exhibit improved stability over time. Fig. 4 (top) shows correct handling of regions that disappear and appear within sequences. Fig. 4 (bottom) tests on fast moving footage containing small objects (BEAR is from [1, 2]). Unlike previous work, fine scale features (e.g. the bear's face) are retained when present in the initial 'boot-strap' frame.

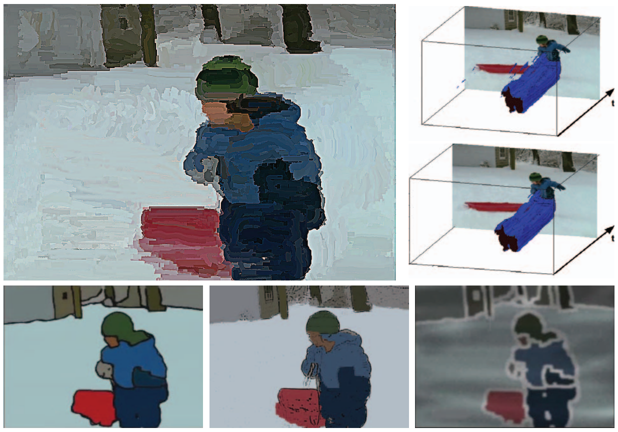

Fig. 2. Artistic stylization: (t-l) paint strokes move coherently within regions; (t-r) before/after temporal smoothing; (bottom) cartoons with edge detail, and a watercolor effect.

Similarly, PANDA shows the aesthetic ability to selectively abstract detail (bushes) from the stylized video, when interactively removed by the user in the initial frame.

In summary we have introduced a novel framework for video segmentation driven by multi-label graph-cut. Frame segmentation is influenced by the propagation of labels from previous frames using optical flow. We have demonstrated temporally coherent segmentation of the video into colored regions, and that this can be exploited to create stylized region-based effects such as painterly and cartoon rendering. Extensions will explore the backward propagation of labels to further improve coherence. We would also like to differentiate 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.

Acknowledgement: We thank Phil Cheatle, David Slatter and Darryl Greig at HP Labs for many insightful discussions.

5. REFERENCES

[1] Collomosse J., "Higher level techniques for the artistic rendering of images and video," *Ph.D. thesis, Univ. Bath*, 2004.

[2] Collomosse J.P., Rowntree D., and Hall P.M., "Stroke surfaces: Temporally coherent artistic animations from video," *TVCG*, vol. 11, pp. 540–549, 2005.

[3] Wang J., Xu Y., Shum H., and Cohen M., "Video tooning," in *SIGGRAPH*, 2004, vol. 23, pp. 574–583.

[4] Goldman D.B., Gonterman C., Curless B., Salesin D., and Seitz S.M., "Video object annotation, navigation, and composition," in *UIST*, 2008, pp. 3–12.

[5] Moscheni F., Bhattacharjee S., and Kunt M., "Spatiotemporal segmentation based on region merging," *PAMI*, vol. 20, pp. 897–915, 1998.

[6] Brendel W. and Todorovic S., "Video object segmentation by tracking regions," in *ICCV*, 2009.

[7] DeMenthon D. and Megret R., "Spatio-temporal segmentation of video by hierarchical mean shift analysis," in *CVPR*, 2000, pp. 142–151.

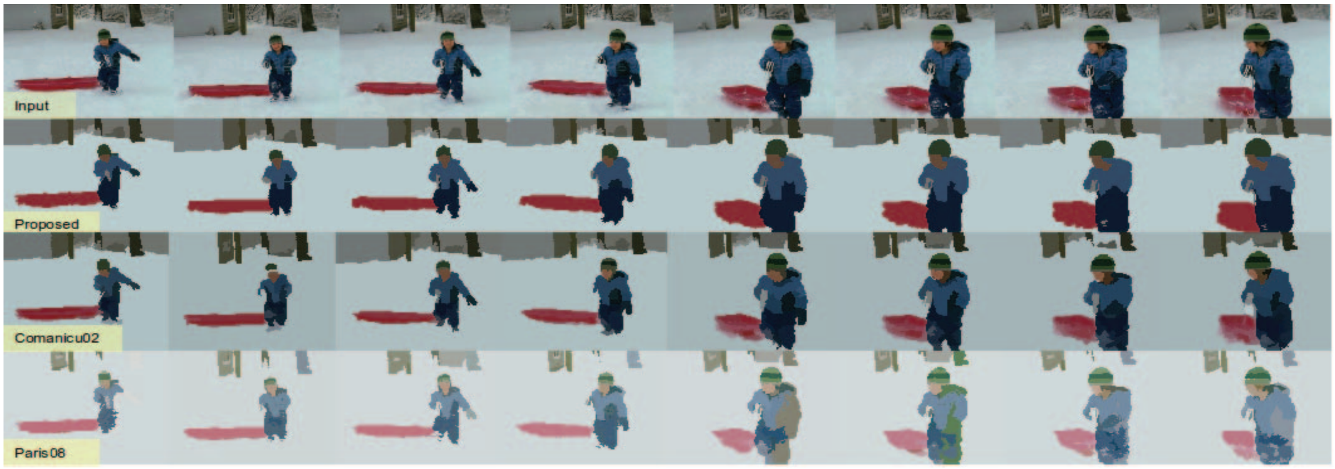

Fig. 3. Comparing the accuracy and coherence of the proposed approach on the BOY sequence, to 'synergistic' mean-shift + edge [20] and a state of the art spatio-temporal method [21]. Boundaries are less prone to variation in shape and topology.

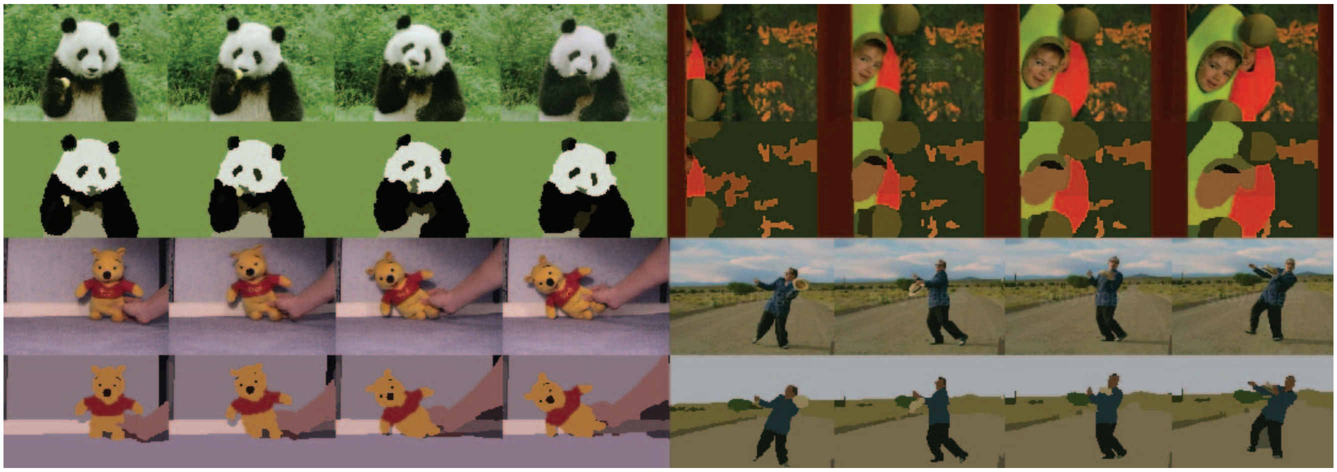

Fig. 4. Top: Illustrating removal (PANDA) and insertion (DRAMA) of regions over time. Bottom: Illustrating the stable tracking of fast, small regions (BEAR, DANCE). Videos: **http://personal.ee.surrey.ac.uk/Personal/Tinghuai.Wang/project.html**.

[8] Paris S. and Durand F., "A topological approach to hierarchical segmentation using mean shift," in *CVPR*, 2007, pp. 1–8.

[9] Greenspan H., Goldberger J., and Mayer A., "A probabilistic framework for spatio-temporal video representation," in *ECCV*, 2002, pp. 461–475.

[10] X. Bai, J. Wang, D. Simons, and G. Saprio, "Video snapcut: Robust video object cutout using localized classifiers," in *ACM SIGGRAPH*, 2009.

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

[12] Boykov Y. and Funka-Lea G., "Graph cuts and efficient n-d image segmentation," *IJCV*, vol. 2, no. 70, pp. 109–131, 2006.

[13] Blake A, Rother C., Brown M., Prez P., and Torr P., "Interactive image segmentation using an adaptive gmmrf model," in *ECCV*, 2004, pp. 428–441.

[14] Boykov Y., Veksler O., and Zabih R., "Fast approximate energy minimization via graph cuts," *PAMI*, vol. 23, pp. 1222–1239, 2001.

[15] Boykov Y. and Kolmogorov V., "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *PAMI*, vol. 26, pp. 1124–1137, 2004.

[16] M. Black and P. Anandan, "A framework for the robust estimation of optical flow," in *ICCV*, 1993, pp. 231–236.

[17] Hall P. M. and Hicks Y., "CSBU-2004-03: A method to add gaussian mixture models," Tech. Rep., Univ. Bath, 2004.

[18] H. Winnemoller, S.C. Olsen, and B. Gooch, "Real-time video abstraction," in *ACM SIGGRAPH*, 2006, pp. 1221–1226.

[19] Perez P., Gangnet M., and Blake A., "Poisson image editing," in *ACM SIGGRAPH*, 2003, pp. 313–318.

[20] Comaniciu D. and Meer P., "Mean shift: A robust approach toward feature analysis," *PAMI*, vol. 24, pp. 603–619, 2002.

[21] Paris S., "Edge-preserving smoothing and mean-shift segmentation of video streams," in *ECCV*, 2008, pp. 460–473.