Feature Fusion and Label Propagation for Textured Object Video Segmentation

V. B. Surya Prasath\textsuperscript{a}, Rengarajan Pelapur\textsuperscript{a}, Kannappan Palaniappan\textsuperscript{a}, Gunasekaran Seetharaman\textsuperscript{b}

\textsuperscript{a} Department of Computer Science, University of Missouri, Columbia MO 65211 USA
\textsuperscript{b} Air Force Research Lab, Rome, NY USA

ABSTRACT

We study an efficient texture segmentation model for multichannel videos using a local feature fitting based active contour scheme. We propose a flexible motion segmentation approach using fused features computed from texture and intensity components in a globally convex continuous optimization and fusion framework. A fast numerical implementation is demonstrated using an efficient dual minimization formulation. The novel contributions include the fusion of local feature density functions including luminance-chromaticity and local texture in a globally convex active contour variational method, combined with label propagation in scale space using noisy sparse object labels initialized from long term optical flow-based point trajectories. We provide a proof-of-concept demonstration of this novel multi-scale label propagation approach to video object segmentation using synthetic textured video objects embedded in a noisy background and starting with sparse label set trajectories for each object.

1. INTRODUCTION

The proliferation of full motion video (FMV) realtime data streams and forensic archives presents a great challenge for automatic processing, exploitation and dissemination. For example, 72 hours of new video is uploaded every minute to the popular video sharing site YouTube. The effective analysis and mining of this exponentially expanding volume of video data can be categorized as a Big Data problem for which a variety of tools from disparate fields such as computer vision, machine learning and multimodal fusion are needed. Object segmentation in videos is a critical enabling technology in video understanding and developing content-based video retrieval systems with practical capabilities. Recently there have been efforts to leverage emerging image and video segmentation techniques from computer vision. We can classify these segmentation methods into three broad categories: (a) supervised where the user is required to scribble or select object labels, (b) semi-supervised that uses a small collection of labeled training data in combination with a large corpus of unlabeled data using active learning methods, and (c) unsupervised where the segmentation is completely automatic.

In what follows we review some of the main methods which show promise in object segmentation from videos and refer to\textsuperscript{1} for other related schemes. Under the category of unsupervised methods: Dense motion segmentation by extending super-pixels (named supervoxels),\textsuperscript{2} graph based methods.\textsuperscript{3} These methods can lead to over segmentations and leakage during propagation of labels. To avoid over-segmentation one can use supervised or semi-supervised methods.\textsuperscript{4,5} Though these supervised methods obtain better results in general, they usually require labeling from the user which can take considerable effort.

In this work we consider a variational approach for label propagation using dense optical flow combined with a multi-scale segmentation model. This unsupervised method uses texture, intensity and color features in a coherent variational formulation for object segmentation from natural videos. For this purpose we utilize the fully unsupervised long distance optical flow to obtain initial labels.\textsuperscript{6} Since in inhomogeneous regions there are no reliable structures which can be captured by motion estimation techniques, the object motion vector field is usually sparse. Moreover, computational constraints require reducing the number of motion vectors being analyzed. In this paper we study how to obtain dense segmentations from sparse clusters of object motion vectors. For this purpose we use the recently studied color, texture feature fusion based globally convex active

Further author information: (Send correspondence to V. B. S. Prasath).: E-mail: prasaths@missouri.edu
contour segmentation method which works well for a variety of natural images. Experimental results with synthetic cases illustrating different scenarios are used to show the advantage of using our variational approach.

The rest of the paper is organized as follows. Section 2 introduces the variational segmentation model along with the label propagation using point trajectories from dense optical flow and provides some synthetic sequences and their corresponding segmentations based on our scheme. Finally, Section 3 concludes the paper.

2. SEGMENTATION USING SPATIAL FEATURE FUSION COMBINED WITH MULTISCALE LABEL PROPAGATION

Before embarking upon the detailed explanation of our proposed model for video segmentation we briefly recall the globally convex version of the active contour method for 2D color image segmentation and refer to\textsuperscript{7,8} for the detailed analysis.

2.1 Globally Convex Chan and Vese Spatial Segmentation Model

The convex formulation of the traditional vector valued Chan and Vese model,\textsuperscript{9} amounts to solving the minimization problem:\textsuperscript{7,8}

$$\min_{0 \leq \phi \leq 1} \left\{ \mu \int_{\Omega} |\nabla \phi| \, dx + \sum_{i=1}^{3} \int_{\Omega} \lambda_i R_i^i(x, c^i) \phi(x) \, dx, \right\}$$

(1)

where $\phi : \Omega \to [0, 1]$ is a function of bounded variation, $u \in BV(\Omega)$ and $R_i^i(x, c^i) = (I^i - c_{in}^i)^2 - (I^i - c_{out}^i)^2$ ($i = 1, 2, 3$) which is known as the image fitting term, $x \in \Omega$ represents the pixel $x = (x_1, x_2)$ and $I := (I^1, I^2, I^3) : \Omega \to \mathbb{R}^3$ represents the input RGB image. The vector $c^i = (c_{in}^i, c_{out}^i)$ represents the averages (mean values) of each channel $I^i$, respectively inside and outside of the segmentation contour. The parameters $\lambda_i$ are positive scalars weighting the fitting terms $R_i^i(x, c^i)$ and can be chosen depending upon the imaging modality. The total variation term (first term in Eqn.(1)) keeps the segmentation curve regular. It can be shown that for any $\lambda_i \geq 0$, there exists a minimizer $\phi$ of (1), which is a global minimum.\textsuperscript{8} The energy functional is homogeneous of degree one in $u$ and has a global minimizer by restricting $u$ such that $0 \leq u(x) \leq 1$. Moreover, the above model (1) does not require a level set based implementation as in the original formulation\textsuperscript{9} and can be optimized very efficiently.\textsuperscript{10}

Since the Chan and Vese model relies only on the mean (piecewise constant) gray values of the input channel $I^i$ alone (for each channel $i$), it can neither capture the spatially varying intensity regions nor textured regions with fidelity. In this work, we use the extension of the globally convex Chan-Vese model (1) to support more general fitting terms and to handle textured regions across multiple channels. For this purpose we use the following brightness-chromaticity decomposition based globally convex active contour segmentation model.\textsuperscript{7} First the input image $I$ is decomposed into four components (channels), namely the intensity of brightness $I := |I|$ and three chromaticity components $(C^1, C^2, C^3) := (I^1/|I|, I^2/|I|, I^3/|I|)$. This mapping $(I^1, I^2, I^3) \to (I, C^1, C^2, C^3)$ separates the intensity and color information and is found to be useful for image denoising,\textsuperscript{11} colorization,\textsuperscript{12} and segmentation.\textsuperscript{7} Following the globally convex version of the Chan and Vese model (1), we extend it to the color (RGB) image segmentation as follows.

$$\min_{0 \leq \phi \leq 1, \rho} \mathcal{E}(\phi, \rho) = \mu \int_{\Omega} |\nabla \phi| \, dx + \int_{\Omega} \left( \lambda_T F_T + \lambda_C F_C \right) \phi \, dx$$

(2)

where the multichannel edge-based texture feature fitting term is given by,

$$F_T = (\omega \times |G_\sigma \ast I| - c_{in})^2 - (\omega \times |G_\sigma \ast I| - c_{out})^2$$

(3)

with $\omega$ being the local cumulative intensity histogram weight and the intensity based localized chromaticity fitting term,

$$F_C = \sum_{i=1}^{3} |G_\sigma \ast C^i - c_{in}^i|^2 - |G_\sigma \ast C^i - c_{out}^i|^2.$$
These two terms utilize the texture (gradient), and intensity and chromaticity information from the input image respectively. The parameter vector is \( \rho := (c_{in}, c_{out}, c_{in}^{2}, c_{out}^{2}, c_{in}^{3}, c_{out}^{3}) \), the components are the mean values for image \( I \). Figure 1 shows an example result using the energy minimization function (2) for a natural image along with their fitting terms defined in Eqns. (3-4). The energy minimization in Eqn. (2) is solved using a fast dual minimization scheme\(^7\) and will be utilized in our label propagation scheme which we describe next.

2.2 Object Segmentation Based On Label Propagation

We study a variational model that propagates labels across scale space within homogeneous object regions sparsely sampled by clustered point trajectories. We show that we can obtain dense accurate segmentations from sparse initial labels and multiscale label propagation. Our method overcomes the over-segmentation issue associated with static single frame image segmentation applied to video streams. We first utilize optical flow to extract motion vectors between a pair of frames to be used in our dense segmentation. Due to its efficiency we use the dense optical flow method studied by Brox and Malik known as Large Displacement Optical Flow (LDOF).\(^6\) Clustering analysis of point trajectories is used as a robust tool to initialize (moving) objects and extract object regions from video shots in a fully unsupervised manner. The main question is how to obtain dense or filled segmentations from sparse clusters of point trajectories. For this purpose we utilize the globally convex segmentation model in (2) as an additional grouping step with an extension to support label propagation across scale space to increase robustness and accuracy. We start with the video shot as input, compute point trajectories.\(^14\) This yields a discrete sparse set of labels (roughly around \( \sim 30\% \) pixels are labelled), which are consistent over the whole shot.

Let \( \tilde{u} = (\tilde{u}_1, \tilde{u}_2, \ldots, \tilde{u}_n) : \Omega \to \{0,1\}^n, n \in \mathbb{N} \) point trajectory labels, i.e.,

\[
\tilde{u}_i = \begin{cases} 
1 & \text{if } x \in L_i, \\
0 & \text{else},
\end{cases}
\]

where \( L_i \) is the set of co-ordinates occupied by a trajectory with label \( i \). The objective is to find a function \( u = (u_1, u_2, \ldots, u_n) : \Omega \to \{0,1\}^n \) that remains close to the given labels for points in \( L := \cup_{i=1}^{n} L_i \). We can minimize,

\[
E_{data}(u) = \frac{1}{2} \int_{\Omega} \chi_L \sum_{i=1}^{n} (u_i - \tilde{u}_i)^2 dx
\]

where \( \chi_L : \Omega \to \{0,1\} \) is the label indicator function, a characteristic function with value 1 on \( L \) and 0 elsewhere. At all other points, the minimizer can take any value, To force other image locations to take specific labels (interpolation or label propagation property), we require a regularizer which can be given in terms of the

![Figure 1. Globally convex feature fusion based active contour method for color image segmentation.](image-url)

(a) Input RGB image of size 321 \times 321 (b) Multichannel edge-based texture feature fitting term \( F_T \), see Eqn. (3) (c) Total intensity based localized chromaticity feature fitting term \( F_C \), see Eqn. (4) (d-e) Switching off one of the feature fitting terms gives a different object segmentation. The image is taken from the Berkeley Segmentation DataSet (BSDS500).\(^13\)
label functions that are relaxed to the continuous domain \( u : \Omega \rightarrow [0,1] \),

\[
\mathcal{E}_{reg}(u) = \int_\Omega g(|\nabla I|) \psi\left(\sum_{i=0}^n |\nabla u_i|^2\right) dx,
\]

(7)

where \( \psi : \mathbb{R}_+ \rightarrow \mathbb{R}_+ \) is the regularization function imposing the smoothness of the label field; for example,

\[
\psi(s^2) = \sqrt{s^2 + \epsilon^2},
\]

(8)

the regularized total variation with \( \epsilon = 10^{-4} \) is used in our experiments. The diffusivity function \( g : \Omega \rightarrow \mathbb{R}_+ \) is used to indicate possible edges (jump discontinuities) and here we set it to the traditional edge indicator function from the image segmentation literature,

\[
g(|\nabla I|^2) = \frac{1}{\sqrt{|\nabla I|^2 + \epsilon^2}}.
\]

(9)

Thus, we minimize the following energy,

\[
\min_{u \in [0,1]} \mathcal{E}(u) = \frac{1}{2} \int_\Omega \chi L \sum_{i=1}^n (u_i - \tilde{u}_i)^2 dx + \int_\Omega g(|\nabla I|) \psi\left(\sum_{i=0}^n |\nabla u_i|^2\right) dx.
\]

(10)

We further utilize the multi-scale nature of our segmentation model (2) to propagate labels through scale space. Let \( u = (u_0^0, \ldots, u_n^0, u_1^1, \ldots, u_n^1, \ldots, u_m^K) \) be the label function across the full scale space hierarchy for \( n \) point trajectories (or objects) and \( K \) scales. In practice we use \( K = 3 \) levels for multiscale segmentation and we minimize the following energy which takes into account coupled label propagation between scales (or level hierarchies),

\[
\min_{u \in [0,1]} \mathcal{E}(u) = \frac{1}{2} \int_\Omega \chi L \sum_{i=1}^n (u_i^0 - \tilde{u}_i)^2 dx + \sum_{k=0}^K \int_\Omega g^k \psi\left(\sum_{i=0}^n |\nabla u_i^k|^2\right) dx + \sum_{k=0}^K \int_\Omega g^k \psi\left(\sum_{i=0}^n |u_i^k - u_i^{k-1}|^2\right) dx,
\]

(11)

where \( g^k(I^k, I^{k-1}) \) are multi-scale diffusivity functions similar to the single scale case given above and are defined as

\[
g^k(I^k, I^{k-1}) = \frac{1}{\sqrt{|I^k - I^{k-1}|^2 + \epsilon^2}}.
\]

(12)

Here

\[
\tilde{I}^k = \frac{1}{|\Omega_m^k|} \int_{\Omega_m^k} I(x) dx
\]

(13)

represents mean color of the obtained segmentations (using scheme (2)) and \( \Omega_m^k \) are partitioned regions for the frame at scale or level \( k \), \( m = 1, \ldots, M^k \). The above functional (11) incorporates multi-scale spatial label propagation on a frame by frame basis and are summed across channels (not shown for clarity). Initial sparse noisy labels are extracted based on a separate clustering process of point trajectories. The functional provides a regularization framework for propagating labels both across space and across scale, assuming that the sparse clustered point trajectories over time are available and sufficiently accurate for a good initialization but not necessarily error-free. So in this functional model there is no explicit regularization over time and the multi-scale regularization is used to produce spatially dense smoothly interpolated labels. Following the success of dual minimization implementation in other domains we use it here to implement our label propagation across scale space; we refer to the work of Chambolle for more details.
Figure 2. Synthetic video object segmentation using the variational minimization method in (11). We show a scenario where two objects moving with same magnitude but opposing direction, and each object has a different texture than the background and occlude each other at the end. From top to bottom we show frames #1, 21, 26, 28, 60. (a) Input frame (b) Ground-truth labels, color coded for better visualization (c) Sparse labels obtained by dropping 90% of the real labels from the objects (d) Segmentation using the globally convex variational minimization model combined with multi-scale label propagation and optical flow-based label initialization. When the objects overlap then two different shades of blue color are used for clarity. See the online version for the color figures.
Figure 3. Synthetic video object segmentation using the variational minimization method in (11). We show a scenario where two objects moving towards each other at the same velocity and one of the object has the same texture as the background (highlighted by a white boundary in the top row). From top to bottom we show frames #1, 28, 52, 55, 60. (a) Input frame (b) Ground-truth labels, color coded for better visualization (c) Sparse labels obtained by dropping 75% of the real labels from the objects (d) Segmentation using the globally convex variational minimization model combined with multi-scale label propagation and optical flow-based label initialization. When the objects overlap then two different shades of blue color are used for clarity. See the online version for the color figures.
2.3 Preliminary Experimental Results

To test our variational label propagation method we utilize several synthetic cases depicting different scenarios. Figures 2 and 3 illustrate our scheme along with ground truth labels and segmentations for videos (60 frames of size 600 × 1024). The ground truth labels (Figure 2(b) and 3(b) columns), and simulated sparse noisy labels (Figure 2(c) and 3(c) columns) are shown after dropping 90% and 75% of the labels from the foreground objects respectively. Figure 2 shows sample frames from the first synthetic video which contains two objects having different texture patterns than the white Gaussian noise background moving with the same speed in opposite directions. Figure 3 shows another set of sample frames from the second synthetic video where one of the objects (right side) has the same texture as the background (shown as white boundary in Figure 3(a) top row) which is a harder segmentation problem since motion is necessary to distinguish the moving foreground object from the background. In both cases our motion-based clustered point trajectory initialized variational scale space label propagation method works well and captures moving objects with improved precision.

3. CONCLUSION

In this paper, we studied a variational feature fusion based scale space label propagation for segmenting textured moving objects in multichannel video data. By utilizing the large displacement optical flow clustered point trajectories as a bottom-up temporal cue we obtain segmentation based on a texture, intensity and color feature fusion using a globally convex active contour formulation. Synthetic examples are given to illustrate the concept and initial results are promising for obtaining moving object segmentations from videos and currently we are working on wide area\textsuperscript{16,17,23} and full motion videos with people\textsuperscript{19} and objects.\textsuperscript{20–22}

Acknowledgment

This research was partially supported by U.S. Air Force Research Laboratory (AFRL) under agreement AFRL FA8750-14-2-0072. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of AFRL or the U.S. Government.

REFERENCES


