Probabilistic Motion Diffusion of Labeling Priors for Coherent Video Segmentation

Tinghuai Wang, *Student Member, IEEE*, and John Collomosse, *Member, IEEE*

***Abstract*—We present a robust algorithm for temporally coherent video segmentation. Our approach is driven by multi-label graph cut applied to successive frames, fusing information from the current frame with an appearance model and labeling priors propagated forwarded from past frames. We propagate using a novel motion diffusion model, producing a per-pixel motion distribution that mitigates against cumulative estimation errors inherent in systems adopting "hard" decisions on pixel motion at each frame. Further, we encourage spatial coherence by imposing label consistency constraints within image regions (super-pixels) obtained via a bank of unsupervised frame segmentations, such as mean-shift. We demonstrate quantitative improvements in accuracy over state-of-the-art methods on a variety of sequences exhibiting clutter and agile motion, adopting the Berkeley methodology for our comparative evaluation.**

***Index Terms*—Computer vision, image segmentation, image sequences.**

I. Introduction

VIDEO segmentation aims to partition pixels into spatio-temporal groups exhibiting coherence and consistency in both appearance and motion. Stable and accurate video segmentation is fundamental to many multimedia tasks, such as video summarization [1], content based retrieval [2], matteing [3] and video stylization [4].

A key challenge is the production of *temporally coherent* segmentations; regions whose shape and neighborhood topology evolve smoothly over time while tracking the underlying video content. Although recent years have delivered significant advances, coherent segmentation remains challenging for real-world video of even moderate complexity. Changes in illumination, viewpoint, and occlusion relationships introduce ambiguities that in turn induce instability in boundaries and the potential for localized under- or over-segmentation. Temporal correlation between consecutive frames via motion estimation (e.g., optical flow) can alleviate these difficulties, however inter-frame motion estimation is often inaccurate introducing further ambiguity to the process. Given the approximate nature of boundary and motion estimation, it is natural to formulate these motion ambiguities in a probabilistic framework.

This paper contributes a novel video segmentation algorithm, in which the segmentation of each frame is guided by motion-flow propagated label priors from previous frames, where flow is estimated via a new probabilistic motion diffusion model. Our approach builds upon the success of multi-label graph-cut approaches to image and video segmentation. The core novel contributions are our motion propagation model, and the combination of this propagated prior information with per-frame estimates of super-pixel boundaries; a growing trend in the image segmentation literature [5]–[9].

In contrast to previous techniques based on flow vectors, our diffusion model produces a new probabilistic motion estimate modelling the *distribution of motion vectors for each pixel*. This distribution guides the diffusion of information from pixel labeling in prior frames, to influence segmentation of the current frame. To decide the segmentation of a given frame, we incorporate not only motion propagated soft labeling constraints at the pixel-level but also propose a soft higher-order constraint by imposing label consistency within image regions (super-pixels [10], [11]) obtained via several unsupervised segmentations of the frame (e.g., mean-shift). These resemble the form of unary potentials commonly used in pairwise conditional random fields (CRFs) for different image labeling problems [12], [13]. This formulation enable the use of powerful graph cut based move making algorithms for performing inference in the framework. By enforcing labeling consistency in this way, we show inaccuracies in boundaries and region over-segmentation to be alleviated. We quantify this improvement through comparison to three state of the art methods; a spatio-temporal method [14] and a "hard" CRF-based motion propagation method that relies upon a single flow vector for each pixel rather than our novel "soft" motion diffusion, and recent graph based video segmentation method based on dense optical flow propagation [21].

We describe our proposed video segmentation algorithm in Section III, presenting the motion diffusion model for propagation of labeling priors in Section IV and describing the supporting energy terms for the CRF in Section V. We evaluate our approach over several challenging video clips exhibiting clutter and agile motion, adopting the methodology of the Berkeley Segmentation Benchmark [15] to provide a quantitative comparative evaluation to state-of-the-art techniques (Section VI). We show our approach to be quantitatively closer to manually annotated ground-truth segmentations of our footage, and release these results at http://personal.ee.surrey.ac.uk/Personal/Tinghuai.Wang/TMM2011.

Manuscript received May 22, 2011; revised September 14, 2011; accepted November 08, 2011. Date of publication November 22, 2011; date of current version March 21, 2012. This work was supported under the Hewlett-Packard Labs IRP Programme. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Alan Hanjalic.

The authors are with the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford GU2 7XH, U.K. (e-mail: Tinghuai.Wang@surrey.ac.uk; J.Collomosse@surrey.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2011.2177078

II. Related Work

Video segmentation has received considerable attention in recent years, with the majority of research effort categorized into two fundamental strategies; spatio-temporal (3-D) analysis and frame-to-frame segmentation (2-D+t).

Methods in first category tackle video segmentation as a spatio-temporal (x,y,t) clustering problem. For example, Dementhon [16] proposes a spatio-temporal approach in which hierarchical mean shift clustering is applied on pixels of 3-D space-time video stack, which are mapped to 7-dimensional feature points, i.e., three color components and 4 motion components derived from inter-frame flow estimates. Anisotropic [17] and causal spatio-temporal kernels [14] have also been explored to refine mean-shift approaches to space-time segmentation. Shi and Malik [18] propose a pairwise graph based model to describe the spatio-temporal relations in the 3-D video data and have employed the spectral clustering analysis to solve the video segmentation problem. Ristivojevic and Konrad [19] derive active surfaces through the space-time volume, which compete iteratively to deliniate object boundaries. Greenspan *et al.* [20] present an approach to extracting coherent space-time regions in feature space via GMM unsupervised clustering. Grundmann *et al.* [21] present an efficient and scalable approach to spatio-temporal segmentation of long video sequences using a hierarchical graph-based algorithm, combining a volumetric over-segmentation with a hierarchical re-segmentation. However, these approaches usually 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.

The second category of approach segments 2-D frames independently, and then creates associations between regions over time to identify and prune sporadic regions [22]–[24]. Moscheni *et al.* [22] process two consecutive frames at a time by iteratively merging over-segmented regions together based on their mutual spatio-temporal similarity. Collomosse *et al.* [23] create spatio-temporal volumes from video by associating 2-D segmentations over time and fitting *stroke surfaces* to voxel objects. Brendel and Todorovic [24] adopt a region-tracking approach in which similar regions are transitively matched and clustered across the video and temporal coherence is forced by incorporating contour cues to allow splitting and merging of regions. These methods are inspired from the observation that pixels constituting a particular segment often belong to the same object or may share common appearance properties. Furthermore, it becomes much more efficient as inference only needs to be performed over a small number of segments rather than all the pixels. Although the stability is improved in these methods, lack of temporal information from adjacent frames during over-segmentation may cause jitter across frames and the temporal coherence is not ensured; the poor repeatability of 2-D segmentation algorithms between similar frames, causing variations in the shape and photometric properties of regions. Wang *et al.* [25] propose a video segmentation algorithm to apply multi-label graph cut on successive frames, in which the segmentation of each frame is driven by motion flow propagated labeling priors and incrementally updated data model estimated from the past frames to improve the temporal coherence. However the flow-propagated labels in this work are assumed to be hard constraints, i.e., perfect estimates, which is often an unsafe assumption for optical flow in general sequences. In our proposed system we also follow a flow-propagation strategy, but avoid imposing hard constraints on motion propagated priors.

Interactive $2-D+t$ video object segmentation systems have also been proposed in recent years [26], [27] tracking region boundaries over time for matte segmentation. However our algorithm propagates label priors and data forward with motion flow within a subset of pixels in regions, rather than tracking 2-D windows on region boundaries that contain clutter from adjacent regions. Furthermore we tackle the more general problem of multi-label segmentation rather than a binary matte, and do so automatically with no manual correction.

In addition to motion propagation, our algorithm utilizes conceptually higher level soft constraints defined via multiple unsupervised over-segmentations of the video frame. This approach has also been widely adopted for image segmentation [5]–[7] using a single over-segmentation. In contrast to these that use multiple super-pixels as a hard constraint (i.e., assuming that all pixels constituting a particular region belong to the same label), more recent work integrates a higher-order region consistency potential with conventional unary and pairwise constraints by using CRFs in a soft framework [8], [9]). We adapt the latter approach in our video framework, but differentiate ourselves in several ways. First, we adopt over-segmented super-pixels from multiple unsupervised segmentation algorithms rather than a single segmentation algorithm—after [28]–[30] but using the soft framework of [8], [9]. Second, rather than computing a penalty via the number of pixels in the super-pixel not taking the dominant label, our method considers the region consistency potential as a even *softer* constraint which is similar to the data prior present in pairwise CRFs [12], [13], and thus can be solved efficiently. Third, to the best of our knowledge, we are the first to apply higher-order spatial constraints to address the video segmentation problem. This is interesting because the temporal incoherence of the per-frame segmentations is nevertheless shown to improve the spatial and temporal coherence of our video segmentation.

III. Preliminaries

Consider a discrete random field consisting of an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ without loop edges, a finite set $\mathcal{L} = \{l_1, l_2, \ldots, l_L\}$ of labels, and a probability distribution P on the space $\mathcal{X} = L^{\mathcal{V}}$ of label assignments. $\mathrm{x} \in \mathcal{X}$ is a map that assigns to each vertex v a label x_v in $\mathcal{L}$. Let N_v denote the set of neighbors $\{u \in \mathcal{V} \,|\, (u,v) \in E\}$ of vertex v. A clique c is a set of vertices in $\mathcal{G}$ in which every vertex has an edge to every other vertex. A random field is said to be Markov if and only if it satisfies the relation property: $P(\mathrm{x}) > 0 \quad \forall\, \mathrm{x} \in \mathcal{L}^{\mathcal{V}}$, and the Markovian property:

$$P(x_v \,|\, x_{\mathcal{V}\setminus v}) = P(x_v \,|\, x_{N_v}). \tag{1}$$

This property states that the assignment of a label to a vertex is conditionally dependent on the assignment to other vertices only through its neighbors.

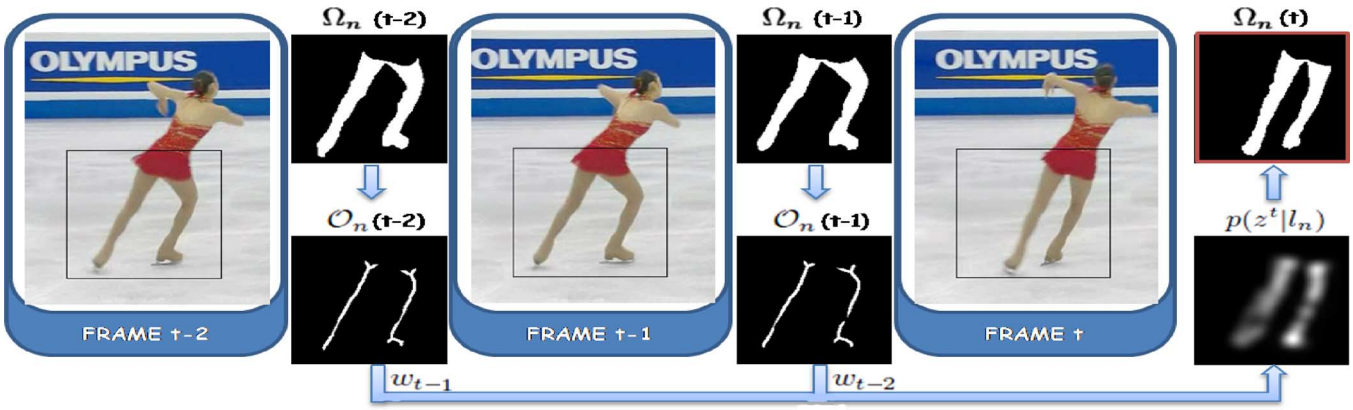


Fig. 1. Illustration of our motion diffusion process over two frames of "YUNAKIM". A subset of pixels $\mathcal{O}_n(t-s)$ from each region $\Omega_n(t-s)$ in each frame I_{t-s} is propagated to frame I_t based on motion estimation and diffused to its close vicinity following a Gaussian distribution. Labeling prior probability $p(z^t \mid l_n)$ is formulated as the merged diffusion probability from previous frame I_{t-s} by weight w_{t-s}.

An energy function $E : \mathcal{L}^{\mathcal{V}} \rightarrow \mathbb{R}$ maps any labeling $\mathrm{x} \in \mathcal{L}^{\mathcal{V}}$ to a real number $E(\mathrm{x})$ called its energy. Energy functions are formed as the negative logarithm of the posterior probability distribution of the label assignment. Minimizing the energy function is equivalent to maximize the posterior probability. The maximum a posteriori probability (MAP) x^* of a random field is defined as

$$\mathrm{x}^* = \mathrm{argmin}_{\mathrm{x} \in \mathcal{L}} E(\mathrm{x}). \tag{2}$$

The posterior distribution over the labelings of the conditional random field is a Gibbs distribution and the corresponding Gibbs energy is given by

$$E(\mathrm{x}) = \sum_{c \in \mathcal{C}} \psi_c(\mathrm{x}_c) \tag{3}$$

where $\mathcal{C}$ is the set of all cliques [31], and $\psi_c(x_c)$ is known as the potential function of the clique c and $\mathrm{x}_c = \{x_i, i \in c\}$.

A. Segmentation Framework

We formulate video segmentation as a pixel-labeling problem of assigning each pixel $i \in \mathcal{V}$ in frame I_t with a value from the existing label set $\mathcal{L}^{\mathcal{V}}$ in frame I_{t-1}.

A subset of $\mathcal{L}$ are carried forward from the region map at $t-1$, via a propagation process described shortly (Section IV). After the propagation process, each pixel in frame I_t bears a set of *prior* probabilities of observing a pixel propagated from different label regions in frame I_{t-1}. The *prior* labeling probabilities of pixels form a soft constraint on the assignments of pixels in I_t, which are labeled to minimize a global energy function. This energy function is adapted from the Gibbs energy function typically used in computer vision and consists of unary, pairwise and higher order cliques as

$$E(x) = \sum_{i \in \mathcal{V}} \psi_i(x_i) + \sum_{i \in \mathcal{V}, j \in \mathcal{N}_i} \psi_{ij}(x_i, x_j) + \sum_{i \in \mathcal{V}} \sum_{c \in \mathcal{S}} \psi_c(x_i) \tag{4}$$

where $\mathcal{V}$ corresponds to the set of all pixels in frame I_t, $\mathcal{S}$ represents the set of super-pixels from over-segmentations (Section V-C). This energy function encourages both temporal consistency of appearance between frames, and spatial homogeneity of contrast within each frame. Moreover, it incorporates a third potential partly enforcing the label consistency inside the regions generated by unsupervised image segmentation algorithms. We describe in detail how each of these potentials are defined, and the optimization of (4) in Section V, but first describe the process by which labels are propagated over time in our framework.

IV. Label Diffusion for Coherent Segmentation

We introduce a motion diffusion model which combines motion estimates made over several time intervals (*frames*) under a probabilistic framework, and accounts for the estimation errors by adaptively refining the internal parameters of this framework. The purpose of the motion diffusion model is to propagate forward the labels of past frames—so forming a distribution of priors for segmentation of the current frame. We bootstrap the first frame of segmentation using mean-shift [10], with a bias to over-segmentation that is resolved via merging of regions due to their similar appearance in subsequent frames.

A. Single-Frame Motion Diffusion

We first compute the SIFT flow [32] from frame I_{t-1} to I_t. The SIFT flow consists of matching densely sampled SIFT features between the two images, while preserving spatial discontinuities. The use of SIFT features allows robust matching across different scene/object appearances and the discontinuity-preserving spatial model allows matching of objects located at different parts of the scene. Although there some discontinuities in the flow field caused by matching errors, we do not assume or require accurate motion estimation at this stage. Indeed our motion diffusion framework is proposed on the assumption that there will be inaccuracies.

Let Ω_n be a region of interest in frame I_{t-1} labeled as l_n. Propagating the whole region to the successive frame I_t by SIFT flow often involves erroneous estimation, especially in positions close to boundary. We only select a subset of pixels $\mathcal{O}_n \subset \Omega_n$ for propagation (Fig. 1). To account for the impact from imprecise motion estimation close to boundary, we form $\mathcal{O}_n$ by

sampling from a morphologically dilated skeleton of each region. The skeleton preserves geometrical and topological properties of the region. To further deal with the uncertainties in positions which are close to the region boundary, we use only the skeletons whose distance to the boundary exceeds a confidence, measured by a distance transform. A skeleton based propagation scheme was first proposed in [25] for similar reasons. However rather than propagating each label using just one flow vector from a single frame [26], [25], our approach diffuses labels across a distribution of directions (derived from multiple frames, Section IV-B) as we now explain.

$\mathcal{O}_n$ contains pixels $J_k^{t-1}(k = 1, 2, 3, \ldots, |\mathcal{O}_n|)$, where $|\mathcal{O}_n|$ is the cardinality of $\mathcal{O}_n$. The position of each pixel is denoted as z_k^{t-1}. For each pixel $J_k^{t-1} \in \mathcal{O}_n$ we predict its position z_k^t in frame I_t based on the motion vector from SIFT flow. As a perfect motion estimation is not available, the proposed model only assumes the motion estimation to be probabilistic. The *diffusion process* diffuses the propagated subset of pixels to close vicinity, treating the predicted position as the center of a Gaussian distribution

$$p\left(z^t; J_k^{t-1}\right) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{\|z^t - z_k^t\|^2}{2\sigma_k^2}\right) \tag{5}$$

where z^t is a position in frame I_t. The variance σ_k reflects the error in motion estimation which is adaptively set for each pixel J_k^{t-1}. For motion estimation which is likely to contain large prediction errors, we set σ_k to large values.

Although Gaussian diffusion is frequently used to model uncertainty in tracking it has been explored only recently in the context of interactive video segmentation, for binary matteing [27]. The key to the robustness of our new multi-label diffusion approach is to propagate only a subset of pixels in regions to account for the imprecise motion estimates close to boundaries typically observed during our early experiments. Furthermore, using local motion coherence to encode the motion estimation error (as opposed to a global measurement of motion alignment error [27]) accommodates the per-pixel local motion estimation errors (σ_k) that cannot necessarily be reflected by a global measurement or single propagation.

We now explain how to determine σ_k. The error in estimating the motion of a region of interest often causes discontinuities in the flow field. Such discontinuities are often referred to as *motion non-coherence*. A small portion of an moving object with rigid shape in a sequence often exhibits coherent motions. We correlate the prediction error with local motion non-coherence. For each pixel J_k^{t-1}, we consider the motion vectors in a 5×5 window centered at J_k^{t-1}. All motion vectors within this window are firstly quantized as N angles $(2\pi)/(N), (4\pi)/(N), \ldots 2\pi$. A quantized motion vector histogram h_k^{t-1} is computed across the local motion vectors. We define a motion coherence factor M_k^{t-1} by measuring the entropy of h_k^{t-1}

$$M_k^{t-1} = \min\left\{1, \frac{\log(N)}{-\sum_{i=1}^{N} H_k^{t-1}(i) \log\left(H_k^{t-1}(i)\right)}\right\} \tag{6}$$

where

$$H_k^{t-1}(i) = \frac{h_k^{t-1}(i) + \epsilon}{\sum_{i=1}^{N} h_k^{t-1}(i)} \tag{7}$$

and ϵ is a small constant (10^{-5} in the system). In information theory, entropy is a measure of the uncertainty associated with a random variable. Higher entropy of h_k^{t-1} indicates lower local motion coherence in the window, and thus smaller M_k^{t-1}. σ_k is computed as

$$\sigma_k = \theta_\gamma \exp\left(\theta_\mu M_k^{t-1}\right) \tag{8}$$

where θ_γ and θ_μ are constant parameters.

The probability of observing a pixel propagated from $\mathcal{O}_n$ (labeled as l_i) at location z^t on frame I_t is

$$p^{t-1}(z^t \mid \mathcal{O}_n) = \sum_{k=1}^{|\mathcal{O}_n|} p\left(X_k^{t-1}\right) p\left(z^t \left| J_k^{t-1}\right.\right) \tag{9}$$

where $p(X_k^{t-1}) = 1/|\mathcal{O}_n|$, assuming equal priors for every pixel in $\mathcal{O}_n$. As motions of l_n-labeled pixels are predicted based on $\mathcal{O}_n$, $p^{t-1}(z^t \mid \mathcal{O}_n)$ can be approximated as the labeling *prior* probability of label l_n at pixel z^t, i.e., $p^{t-1}(z^t \mid l_n)$.

B. Multi-Frame Motion Diffusion

We build a single-frame probabilistic motion diffusion model in Section IV-A taking into account the estimation errors. As we later show, our diffusion model greatly enhances the coherence of skeleton based motion propagation [25] during occlusion and rapid movement. However, gross SIFT matching errors occasionally occur and may result in amplified errors in the propagation process.

To mitigate gross prediction errors, we adopt a multi-frame fusion scheme. We perform single-frame diffusion process on multiple successive frames I_{t-T}, $I_{t-T+1}, \ldots I_{t-1}$ in the sequence to acquire multiple diffusion probabilities $p^{t-T}(z^t \mid l_n)$, $p^{t-T+1}(z^t \mid l_n), \ldots p^{t-1}(z^t \mid l_n)$ and $p^1(z^t \mid l_n)$ regarding label l_i. Merging multiple frames' diffusion probabilities we have

$$p(z^t \mid l_n) = \sum_{s=1}^{T} w_{t-s} p^{t-s}(z^t \mid l_n) \tag{10}$$

where each frame contributes to the final fusion with weight w ($\sum_{s=1}^{T} w_{t-s} = 1$), which is inversely proportional to the alignment error in the scope of the region of interest Ω_n on each frame

$$w_{t-s} = 1 \Big/ \sqrt{\frac{1}{|\Omega_n|} \sum_{z \in \Omega_n} \|I_{t-s}(z) - I'_{t-s}(z)\|^2} \tag{11}$$

where I'_{t-s} is the warped color image from frame I_t to I_{t-s} by the SIFT flow. Accurate alignment generally indicates reliable SIFT flow and such frames thus contribute more to the probability fusion.

$p(z^t \mid l_n)$ reveals the likelihood of the pixel at z^t being assigned with label l_i propagated from previous frame in the sequence. This probability is encoded directly in the *unary* term of our energy function (4), which comprises a sum of appearance and labeling potentials (described in Section V, (13)).

C. Incrementally Updated Color Model

As we explain shortly (Section V), the segmentation of I_t is dependent on the unary term of (4) comprising a per label appearance model built incrementally over time. A component of this model is a Gaussian mixture model (GMM), the parameters of which are written $\Theta^n_{\rm col}$ for each label l_n, and which is initially built by sampling l_n-labeled pixels in starting keyframe I_1. We sample in the RGB color space. To avoid possible sampling errors caused by the imperfect region boundaries, we only use pixels whose spatial distance to the region boundary is larger than a confidence distance (3 pixels in our system) as the training data for the GMMs. As the color distribution is normally simple in each appearance homogeneous region, the number of components in each GMM is set to 3.

To cope with luminance variations in the sequence, we update the color model to achieve good segmentation by sampling the historical colors of labeled pixels over recent frames. As the segmentation in frame I_t is not necessarily perfect and may contain errors, the updated color model might gradually drift away from capturing the correct color distribution, resulting in amplified errors in the segmentation. We stably update the color model by sampling the pixels within Ω_n in previous frames, with a decreasing proportion $\theta^d_p \in [0,1]$ $(d>0)$ as the temporal distance d from the current frame I_t increases, using $\theta^d_p \propto e^{-d^2/\sigma_d^2}$. 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.

D. Label Management

If a region labeled n in I_t deviates significantly from its corresponding historic appearance model (determined via a threshold on the χ^2 distance between $\Theta^n_{\rm col}$ at time t and $t-1$), then it is likely that the labeling is in error. Given that pixels matching the appearance of labels in the set are likely to be assigned correctly, we assume that significant changes are due to appearance of a new semantic region in the sequence. We therefore run our bootstrap procedure (e.g., mean-shift) over pixels putatively labeled n to create a new set of labels that are merged into $\mathcal{L}^{\mathcal{V}}$. The frame I_t is then re-segmented using the enriched label set. Any superfluous region labels generated by this process are immediately merged into other similar labels via the graph-cut labeling process.

The related problem of label deletion is accommodated naturally within our framework as, depending on the pixel data, the multi-label graph cut may not assign a propagated label to the current frame.

E. Smoothing and Filtering

Due to visual ambiguities in low contrast areas, some pixels might be mis-labeled which results in unsatisfactory temporal coherence. We improve the temporal coherence by performing spatio-temporal smoothing. Specifically, we create a set of space-time volumes by coherently labeling regions in adjacent frames, and apply a fine scale $(3\times3\times3)$ Gaussian filter to remove boundary noise. We only filter volumes above a certain size to avoid removing salient detail.

To further remove short-lived volumes which persist beyond the Gaussian filtering step, we inspect the duration $t_{l,k}$ of the kth $(k=1\ldots K_l)$ disconnected space-time volume labeled as l in a time window of 24 frames (1 s). We remove any disconnected volumes within this time window which are shorter than a length

$$T_{l:\{1\ldots L\}} = \min\left\{\max_{k\in\{1\ldots K_l\}} t_{l,k}, \tau_r\right\} \tag{12}$$

where τ_r is set to be six frames (about 1/4 second). We fill the "holes" left by filtering and smoothing by extrapolating region labels from immediate space-time neighbors on a nearest-neighbor basis using a distance transform.

V. Definition of Energy Potentials

We now describe how the diffused labeling priors are integrated into the unary, pair-wise and super-pixel consistency terms as defined respectively in (4). We illustrate the importance of each in Fig. 4 where various terms are disabled to qualitatively demonstrate their contribution to segmentation coherence.

A. Appearance Model

The unary term $\psi_i(x_i)$ exploits the fact that different appearance homogeneous regions tend to follow different appearance models. This encourages assignment of pixels to the label following the most similar appearance model (we write the parameters of such models Θ). The unary term is defined as the negative logarithm of the likelihood of a label being assigned to pixel i. It can be computed from the appearance model for each label. To provide more discriminative power for accurate segmentation, the unary term incorporates color and texture features as well as *prior* labeling probabilities. The unary term is defined as

$$\psi_i(x_i) = \theta_{\rm col}\psi_{\rm col}(x_i) + \theta_{\rm tex}\psi_{\rm tex}(x_i) + \theta_{\rm lab}\psi_{\rm lab}(x_i) \tag{13}$$

where $\theta_{\rm col}$, $\theta_{\rm tex}$ and $\theta_{\rm lab}$ are weights of color potential $\psi_{\rm col}(x_i)$, texture potential $\psi_{\rm tex}(x_i)$ and *prior* labeling potential $\psi_{\rm lab}(x_i)$, respectively.

1) Color Potential: Color potential is defined as

$$\begin{aligned}\psi_{\rm col}(x_i) &= -\log P_g(I_t(i)\,|\,x_i;\Theta_{\rm col})\\ P_g(I_t(i)\,|\,x_i=l_n;\Theta_{\rm col}) &= \sum_{k=1}^{K_n} w_{nk}\mathcal{N}(I_t(i);\mu_{nk},\Sigma_{nk})\end{aligned} \tag{14}$$

i.e., the color model of the nth label l_n is represented by a mixture of Gaussians (GMM), with parameters w_{nk}, μ_{nk} and Σ_{nk} representing the weight, the mean and the covariance of the kth component. The parameters of all GMMs $(\Theta_{\rm col}=\{w_{nk},\mu_{nk},\Sigma_{nk}, n=1,\ldots,|L|,\quad k=1,\ldots,K_n\})$ are learned from historical observations of each region's color distribution (Section IV-C).

2) Texture Potential: Color potential alone is not very discriminative and we incorporate texture potential to achieve more accurate segmentation. To this end, we adopt textons [33]

which have been proven effective in categorizing materials [34] and generic object classes [35], [8], [36].

For extracting texton histograms, we use a filter bank made of 36 bar and edge filters, 1 Laplacian of Gaussian (LoG) and 1 Gaussian filter. The 36 bar and edge filters (6 orientations and 3 scales for each) are applied to the L channel only, producing 36 filter responses. The Gaussian filter is applied to each CIELab channel, thus producing 3 filter responses. The LoG is also applied to the L channel only, thus producing 1 filter response. We quantize filter responses to 200 textons by running K-means clustering and each pixel in I_t is assigned to the nearest cluster center to generate the texton map T_t. We define texture potential as

$$\begin{aligned} \psi_{\text{tex}}(x_i) &= -\log P_g(T_t(i) \,|\, x_i; \Theta_{\text{tex}}) \\ P_g(T_t(i) \,|\, x_i = l_n; \Theta_{\text{tex}}) &= \mathcal{H}^n(T_t(i)). \end{aligned} \tag{15}$$

The texture model Θ_{tex} of the nth label l_n is represented by a discrete probability model given the normalized texton histogram $\mathcal{H}^n$ learned from the textons map in the starting keyframe.

3) Labeling Potential: The labeling *prior* potential exploits the fact that pixels with a higher probability propagated from particular labeled region tend to have consistent label assignment. Unlike other interactive or automatic segmentation algorithms which use the labeling prior as a hard constraint, we incorporate labeling prior as a soft constraint which is inferred from a probabilistic motion estimation framework which inherently takes into account the motion estimation errors. The labeling potential $\psi_{\text{lab}}(x_i) = p(i \,|\, x_i)$ maps directly to $p(z^t \,|\, l_n)$ derived for each pixel, given a label, as defined in Section IV-B.

B. Encouraging Spatial Coherence

The pairwise term encourages coherence in region labeling and discontinuities to occur at high contrast locations, which is computed using RGB color distance as in Grab-Cut [37]:

$$\psi_{ij}(x_i, x_j) = \begin{cases} 0 & \text{if } x_i = x_j \\ \theta_\lambda e^{-\theta_\beta \|I_t(i) - I_t(j)\|^2} & \text{if } x_i \neq x_j \end{cases} \tag{16}$$

where θ_β is chosen to be contrast adaptive [38]:

$$\theta_\beta = \frac{1}{2} \langle \|I_t(i) - I_t(j)\|^2 \rangle^{-1} \tag{17}$$

where $\langle \cdot \rangle$ denotes expectation over an image sample.

C. Super-Pixel Consistency Term

The super-pixel consistency term encourages the pixels belonging to a super-pixel to be assigned with the same label. We define this spatially "higher order" term as

$$\psi_c(x_i) = \begin{cases} 0 & \text{if } i \notin c \\ \frac{\theta_c}{|c|} \sum_{j \in c} \psi_j(x_i) & \text{if } i \in c \end{cases} \tag{18}$$

after [8], where θ_c is the parameter weighting the label consistency partly enforced by super-pixel c, and $|c|$ is the cardinality of super-pixel c. The expression $\sum_{j \in c} \psi_j(x_i)$ gives the label consistency cost, i.e., the cost if all pixels constituting super-pixel c are labeled as x_i (pixel i). $\psi_c(x_i)$ is thus defined as the weighted average unary potential of pixels in super-pixel c against label x_i. The indication is that an optimal label assignment to pixel i should also fit all pixels in super-pixel c as long as c has good homogeneity of visual appearance.

In practice, due to the non-homogeneity of visual appearance and parameter settings, the shapes of super-pixels may not always be consistent with the real object boundaries in only one over-segmentation or one unsupervised segmentation algorithm. Some super-pixels may quite often contain pixels belonging to multiple labels and will encourage an incorrect labeling. Therefore, following [39], multiple segmentations resulted from with different parameter sets of different unsupervised segmentation algorithms [10], [11] per frame are generated, so that although some super-pixels may fail to agree with object boundaries, the others would be good super-pixels that correspond to coherent boundaries. Different unsupervised segmentation algorithms promote differently featured homogeneous regions. Mean shift segmentation [10] generates regions with homogeneous colors, whereas *Super-pixel* [11] produces segmentations incorporating various Gestalt cues, i.e., contour, texture, brightness and good continuation.

Each super-pixel partly enforces the label consistency of regions with a weight. We correlate the weight with the quality of super-pixel from the over-segmentations. We adopt the super-pixel quality measure presented in [9], using the variance of unary potentials of all constituent pixels of a super-pixel as

$$\sigma_c = \exp\left(-\frac{\theta_s}{|c|} \sum_{j \in c} \left(\psi_j(x_i) - \frac{\sum_{j \in c} \psi_j(x_i)}{|c|} \right)^2 \right) \tag{19}$$

and θ_c is defined as the normalized σ_c:

$$\theta_c = \frac{\sigma_c}{\sum_{c \in \mathcal{S}} \sigma_c}. \tag{20}$$

As opposed to other segmentation algorithms which use the hard label consistency in regions assuming that all pixels constituting a particular region are assigned with the same label, we use it as a soft label consistency constraint, similar to the Robust P^n model and non-parametric approaches of [8] and [9]. Unlike the Robust P^n model which is based on the number of pixels in the super-pixel not taking the dominant label, we use the spatial constraint imposed by each super-pixel as soft constraint and naturally incorporate it to the unary term, and thus simplify the optimization without explicitly performing higher-order optimization (see Section V-D).

D. Optimization

Although the proposed energy function (4) takes the similar form of the Robust P^n model in [8], the super-pixel consistency term is not based on the count of the number of labeled pixels within a single super-pixel. Rather, we define a *soft* constraint to reflect the label consistency enforced by different over-segmentations. We define this as the weighted average unary potential of pixels in each super-pixel. This definition is convenient as this spatially "higher order" term does not take multiple numbers of variables in the clique, and so can effectively be further

TABLE I
SUMMARY OF VIDEO CLIPS USED IN OUR EVALUATION, ANNOTATED AS TO MOTION AND OCCLUSION CONDITIONS PRESENT

Clip	Motion	Occlusion
BOY (Fig. 5)	Slow	None
DANCE (Fig. 5)	Agile	Light
MONKEYBAR (Fig. 5)	Agile	Heavy
GARDEN (Fig. 8)	Slow	Light
WALKDOG (Fig. 9)	Slow	Heavy
YUNAKIM (Fig. 9)	Agile	Heavy
SKATEBOARD (Fig. 9)	Fast	Light
COWGIRL (Fig. 9)	Slow	Light
BASEBALL (Fig. 9)	Fast	Heavy

merged to unary term and the energy function (4) can be simplified to

$$E(x) = \sum_{i \in \mathcal{V}} \left(\psi_i(x_i) + \frac{\theta_c}{|c|} \sum_{j \in c} \psi_j(x_i) \right) + \sum_{i \in \mathcal{V}, j \in \mathcal{N}_i} \psi_{ij}(x_i, x_j). \quad (21)$$

As the pairwise potentials of the energy function (21) is of the form of a Potts model it can be minimized using the α-expansion and $\alpha\beta$-swap algorithms [40]. 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 (4) cannot be decreased [40]. Each α-expansion iteration can be solved exactly by performing a single graph-cut using the min-cut/max-flow [41]. Convergence to a strong local optimum is usually achieved in 3–4 cycles of iterations over our label set. We use Alahari *et al.*'s [42] technique to improve the computation and memory efficiency of each iteration by reusing the flow at each iteration of the min-cut/max-flow algorithm, resulting in a two-fold speed-up.

VI. EXPERIMENTS AND COMPARISONS

We apply our segmentation algorithm to a several video clips exhibiting both slow moving and agile motion, and also a variety of occlusion conditions (no occlusion, self-occlusion, inter-object occlusion)—summarized in Table I. We assess segmentation performance on both a subjective qualitative and objective quantitative basis; the latter using the methodology of the Berkeley Segmentation benchmark [15].

A. Parameter Settings

We first explain the parameter settings in unsupervised segmentation algorithms, i.e., mean shift and *Super-pixel*, that form the basis for the third term (the higher order constraint) in our optimization. There are two key parameters in mean shift algorithm; bandwidth in the spatial domain (h_s), and the range domain (h_r). A set of regions with various granulations are generated by varying h_s and h_r. As segmentations do not change dramatically with varying h_s on our NTSC resolution video frames we obtain 4 over-segmentations with parameters $(h_s, h_r) = \{(6,8),(6,10),(6,12),(6,14)\}$. *Super-pixel* generates a large number of small nearly-uniform regions which has been shown to retain salient structure in real images. The only parameter

(a) (6, 8) (b) (6, 10) (c) (6, 12) (d) (6, 14) (e) 200 Super-Pixels (f) 500 Super-Pixels

Fig. 2. Illustrating the multiple over-segmentations used to promote label consistency via the super-pixel potential in our energy term (4), as governed by parameters documented in Section IV-A. (a)–(d) are generated by mean shift segmentation algorithm with different parameters (h_s, h_r); (e)–(f) are generated by *Super-pixel* with particular number of super-pixels.

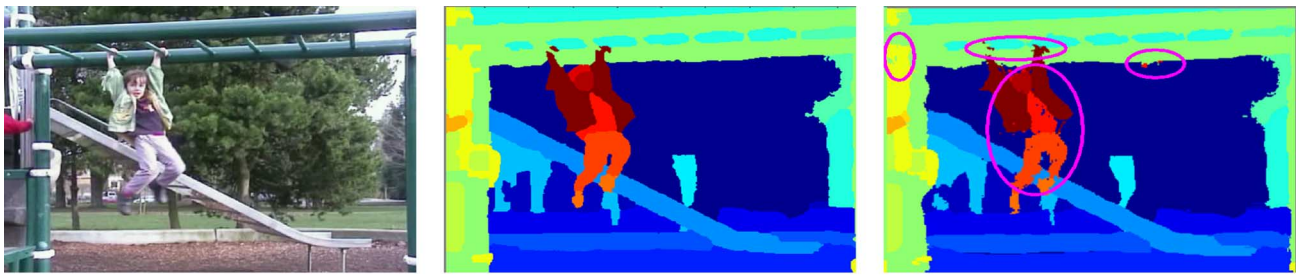
in *Super-pixel* is the number of super-pixels or regions to be generated. We generate two sets of regions using *Super-pixel* with 200 and 500 super-pixels, respectively. An example of multiple over-segmentations is shown in Fig. 2. Weighting parameters θ_{col}, θ_{tex} and θ_{lab} of color potential $\psi_{\text{col}}(x_i)$, texture potential $\psi_{\text{tex}}(x_i)$ and *prior* labeling potential $\psi_{\text{lab}}(x_i)$ are chosen empirically, and we set $\theta_{\text{col}} = 0.31$, $\theta_{\text{tex}} = 0.56$ and $\theta_{\text{lab}} = 0.13$, respectively. θ_λ is set empirically to be 3 to obtain satisfactory segmentation. Other parameter settings are $\theta_s = 0.5, \theta_\gamma = 6, \theta_\mu = 2$.

B. Objective Evaluation

We first present the comparative objective evaluation of the proposed algorithm against two state-of-the-art video segmentation algorithms: multi-label propagation (MLP) [25], and spatial-temporal mean shift (STMS) [14]. These algorithms, respectively, represent an example of a 2-D + t and 3-D (spatio-temporal) video segmentation algorithm. We additionally compare against a state of the art hierarchical graph based (HGB) approach due to Grundmann *et al.* [21].[1]

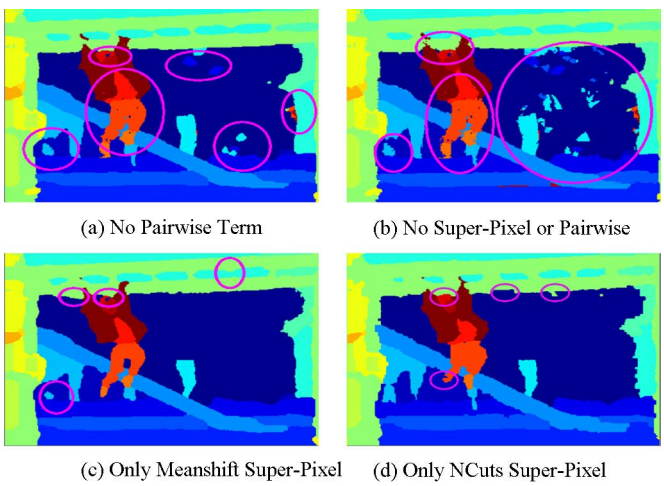
1) Benchmark: For objective evaluation, we adopt the Berkeley Segmentation Benchmark [15] to evaluate segmentation against manual ground-truth. This boundary-based evaluation methodology has become a standard benchmark. This framework considers two aspects of segmentation performance. Precision measures the fraction of true positives in the contours produced by a segmentation algorithm. Recall

[1]Obtained via http://neumann.cc.gt.atl.ga.us/segmentation/



(a) (b) (c)

Fig. 3. Segmentation of the "MONKEYBAR" video with and without the super-pixel consistency term. (a) Original frame. (b) Segmentation obtained with super-pixel potential present in (4) exhibits improved boundary stability when propagated over time, despite computing each frame's super-pixels being computed independently. (c) Segmentation obtained without the super-pixel constraint, differences highlighted in ellipses.



(a) No Pairwise Term (b) No Super-Pixel or Pairwise

(c) Only Meanshift Super-Pixel (d) Only NCuts Super-Pixel

Fig. 4. Illustrating the influence of the unary, pairwise and super-pixel (Spix) terms on segmentation coherence ("MONKEYBAR" sequence). Notable differences to proposed approach [Fig. 3(b)] highlighted in ellipses.

indicates the fraction of ground truth boundaries detected in the segmentation. The global F-measure, defined as the harmonic mean of precision and recall, provides a useful summary score for the segmentation algorithm.

2) Ground Truth: In order to obtain a reliable estimate of segmentation accuracy under [15] we require ground truth region boundaries. We therefore hand labeled individual frames, seeking to preserve fine object boundaries present. Generating manual ground truth segmentations of all the frames of tested videos is very time consuming. Given the frame rate of 24 fps, we opted to hand label the ground truth every 10 frames, and made a second separate manual inspection visually verifying the boundary accuracy.

3) Results: Fig. 6(a)–(c) presents the comparison between the proposed method and the MLP, STMS and HGB algorithms over clips "BOY" (192 frames), "DANCE" (62 frames) and "MONKEYBAR" (300 frames). According to the normalized F-measure with respect to manual ground-truth boundaries, our algorithm consistently outperforms the CRF based MLP algorithm, the graph-based HGB and the spatio-temporal STMS approach across the full duration of the clips. Incorporating labeling prior probability as well as the super-pixel consistency potential in (4) has significantly increased the accuracy and coherence of segmented region boundaries.

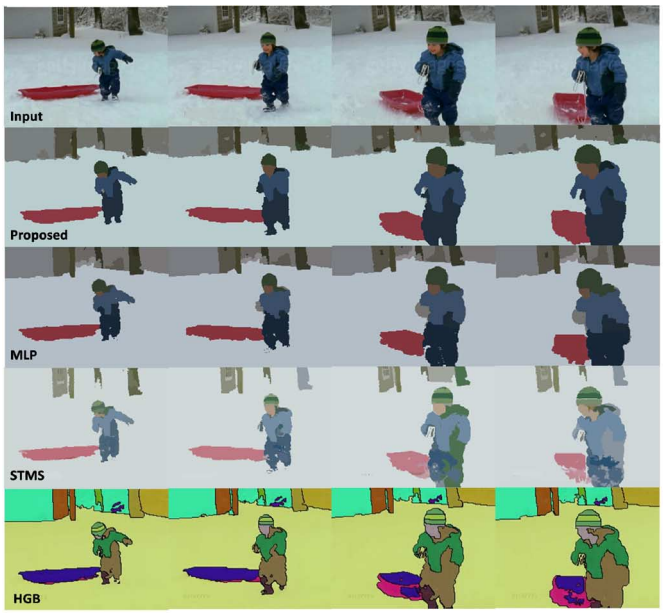
Fig. 6(a) and (b) compares our proposed approach to MLP on clips "BOY" and "DANCE" [25]. We observe the region boundaries from our proposed method to exhibit improved stability and accuracy over time over STMS, HGB, and MLP according to the F-measure with respect to manual ground-truth boundaries. Adopting motion cues as a hard constraint in the CRF framework of the MLP algorithm cumulatively leads to mis-labelings close to boundaries; the non-discriminative color model in MLP further deteriorates the segmentation quality in areas with low contrast or similar color but different texture properties.

In Fig. 6(c) there is significant agile motion in "MONKEYBAR"—the girl twists and suffers frequent inter-occlusion over duration of the clip. Despite the adoption of a forward propagation (2-D+t) strategy over several hundred frames of video there is no significant degradation of F-measure over time; the degradation is comparable to STMS (a spatio-temporal approach). The hard assignment propagation strategy of MLP leads to merging of regions, especially in the wake of moving limbs such as the legs (cf. Fig. 8) resulting in a lower F-measure. We observe the HGB algorithm (also based on a form of hard assignment dense flow propagation) to fragment regions significantly as the sequence progresses, whereas our approach does not, leading that method to produce consistently lower F-measure scores 6(c).

C. Subjective Evaluation

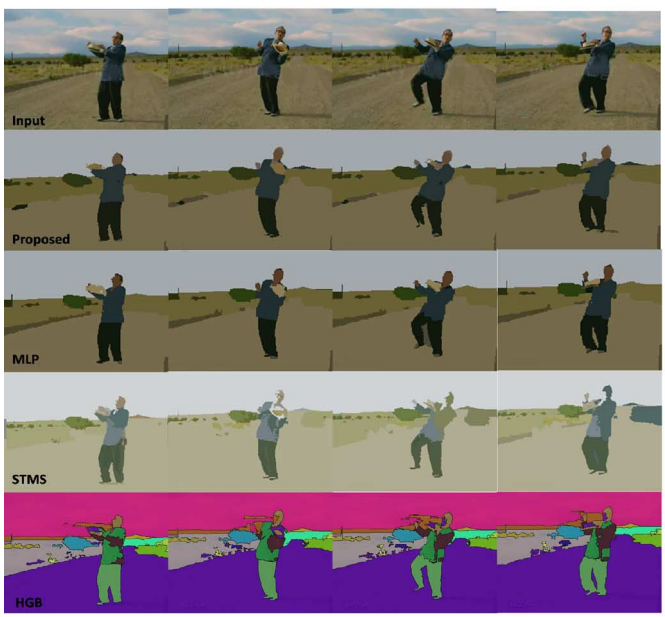
We also demonstrate segmentation results on eight video clips. Each region is shaded with the mean color of pixels in each labeled region on the starting key frame to evaluate longterm coherence and boundary consistency. Fig. 5 makes qualitative comparison of the segmentation results of our proposed algorithm, MLP, HGB and STMS on clips "BOY", "DANCE" and "MONKEYBAR". We observe that the relative coherence and boundary accuracy match the objective evaluations in Section VI-B; for example see the zoomed inset (d). The ability to cope with fast motion and occlusions are significantly improved in the proposed segmentation algorithm over the state-of-the-art. A couple of failure cases are also indicated in Fig. 5(d), in particular the body of the child ("MONKEYBAR") and the hand/hat of the dancer ("DANCE") are shown to deform unnaturally when undergoing erratic motion over background of similar color and or texture.

An additional qualitative comparison on the "GARDEN" sequence is provided in Fig. 7, comparing against a further



(a) Comparative evaluation over "BOY" sequence

(b) Comparative evaluation over "MONKEYBAR" sequence [17]



(c) Comparative evaluation over "DANCE" sequence

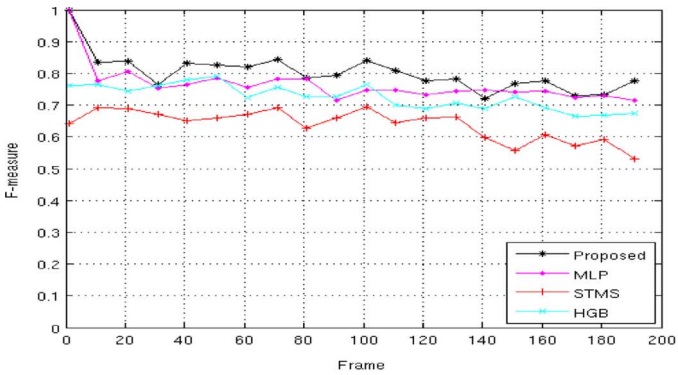
(d) Close-ups of successes vs. inaccuracies

Fig. 5. Comparing the accuracy and coherence of the proposed approach to MLP, STMS and HGB. Boundaries are less prone to variation in shape and topology. Sequences presented as follows: source (1st row); proposed approach (2nd row); MLP (3rd row); STMS (4th row); HGB (5th row). Inset (d). 4×4: Improved performance of the proposed method vs. state of the art on face and hands in "MONKEYBAR". 2×1 Failures cases of the proposed approach, although outperforming compared methods some mislabeling of the hair in "MONKEYBAR" and loss of spatial coherence on hat in "DANCE" can be observed. In both cases these can be attributed to color texture similarity in the presence of erratic motion.

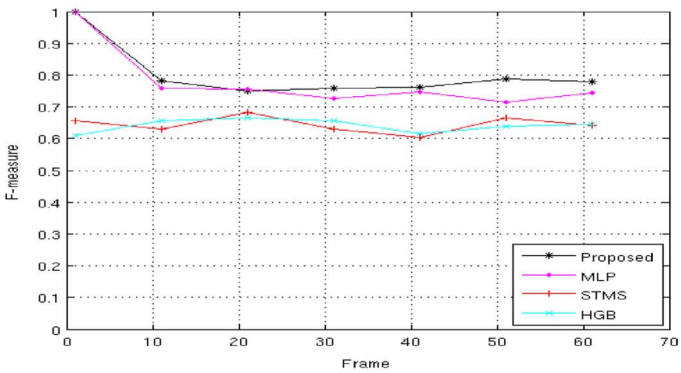
MRF/CRF based method [8], HBG [21] and another recently proposed video segmentation algorithm due to Brendel *et al.* [24]. Our method performs comparably to HBG on this sequence (though see other qualitative comparisons, Fig. 5) and retains a smaller number of coherent regions versus [24], [8].

Fig. 8 directly compares our probabilistic diffusion ("soft") approach to motion propagation, with the hard-assignment strategy of [25]. The experiment is facilitated by temporarily modifying our approach to work with color appearance only (no textons) and omitting the super-pixel term during optimization. The benefits of the probabilistic approach are observed on the feet of the child; hard assignment causes incorrect pixel assignments to cumulatively trail the feet over time. Although soft assignment alone causes minor loss of spatial coherence, this is avoided in our proposed system through incorporation of the super-pixel constraint to produce results such as those of Fig. 5(b).

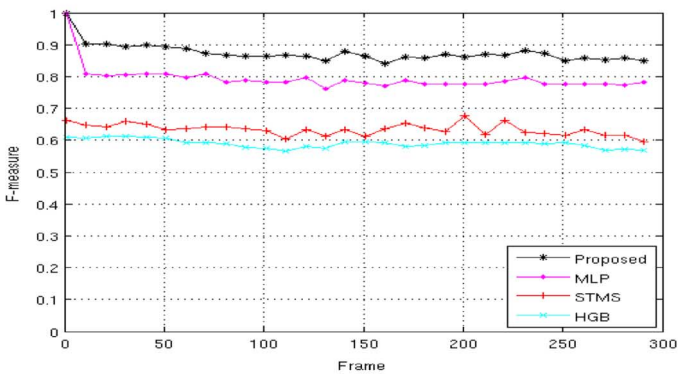
Fig. 9 shows the remaining five segmentation results on clips "YUNAKIM" (560 frames), "COWGIRL" (224 frames), "BASEBALL" (171 frames), "SKATEBOARD" (146 frames), and "WALKDOG" (300 frames). Our segmentation algorithm exhibits consistent region identity and stable boundaries under conditions such as fast motion, low contrast, ambiguous color,



(a) Comparative F-measure for "BOY" over time



(b) Comparative F-measure for "DANCE" over time



(c) Comparative F-measure for "MONKEYBAR" over time

Fig. 6. Evaluation of video segmentation algorithms against manual ground-truth on the Berkeley Segmentation Benchmark. Our proposed algorithm outperforms Multi-label Propagation (MLP) [25], Grundmann *et al.* [21], and spatial-temporal mean shift (STMS) [14] according to their F-measure (harmonic mean of precision and recall) with respect to manual ground-truth.

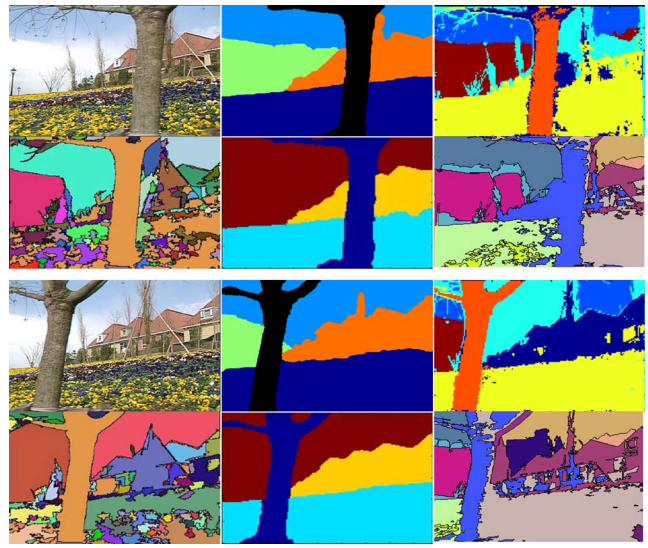


Fig. 7. Additional qualitative comparison of performance on frames 1 and 30 of the "GARDEN" sequence. Order left-right, then top-bottom: Original; Proposed approach; [43]; [24]; [8]; [21].

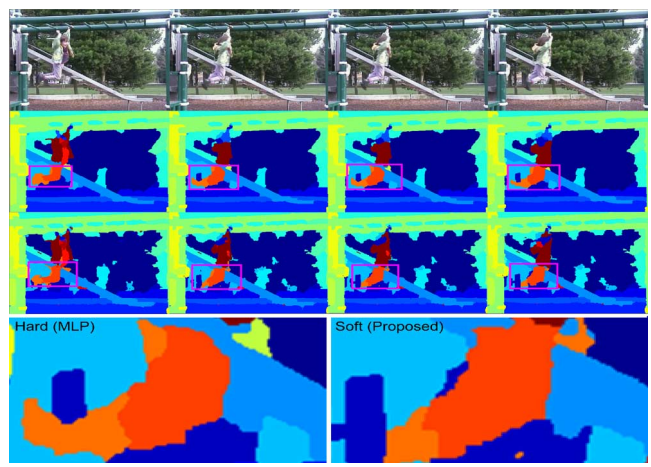


Fig. 8. Comparison of motion propagation strategy; soft (proposed) vs. hard ([25]) assignment. Textons and the super-pixel term are disabled to enable comparison between motion propagation strategies of [25] (row 2) and our approach (row 3); zoomed in sections of third frame sample (row 4). Note the cumulative errors of hard assignment incorrectly smear the feet (orange label) into elongated region over time (highlighted), where the region is correctly segmented using our proposed motion diffusion approach.

non-rigid shape, occlusions. Object boundaries are accurately preserved with color and texture homogeneous regions grouped to ensure temporal and spatial coherence.

VII. Conclusion

In this paper, we presented a novel algorithm for video segmentation driven by multi-label graph-cut. Our core contribution was a multi-frame probabilistic motion diffusion model to incorporate labeling priors from previous frames to influence the segmentation in new frame. Uniquely this diffusion model propagated a *per-pixel distribution of labeling priors* forward based on the probability distribution of motion vectors for that pixel. Motion flow estimation remains a challenging open problem in Computer Vision, and our approach mitigates against inaccuracy in such estimates via this "soft" propagation strategy. This was shown to improve temporal coherence over similar hard-assignment strategies [25], graph based schemes based on flow propagation [21] and spatio-temporal segmentation [14]. We combined this motion framework with a spatially "higher order" constraint additionally imposing the soft label consistency constraint across image regions (super-pixels) obtained via various unsupervised segmentations—as is now common in image segmentation. By enforcing labeling consistence, both the spatial coherence and boundary accuracy of the segmentation was enhanced (demonstrated via comparison to a manually labeled ground truth). We demonstrated our algorithm on a variety of sequences exhibit both simple and challenging motion and occlusion conditions.

(a) Representative frames from "YUNAKIM" segmentation

(b) Representative frames from "COWGIRL" segmentation

(c) Representative frames from "BASEBALL" segmentation

(d) Representative frames from "SKATEBOARD" segmentation

(e) Representative frames from "WALKDOG" segmentation

Fig. 9. Additional segmentation results applying our approach to NTSC video sequences (source in top row, our result in bottom row). Please refer to http://personal.ee.surrey.ac.uk/Personal/Tinghuai.Wang/TMM2011 for these and further results.

A current bottleneck in our system is the SIFT-flow estimation, which can take around 10 seconds in total to compute the flow between historic frames at the currently processed frame. Were our algorithm to be used for real-time segmentation, an alternative and perhaps less accurate optical flow method could be trivially substituted.

One interesting direction for future work would be to explore the possibility of propagation labeling priors both forward and backward in the sequence. This could provide an additional temporal constraint with the potential to further enhance temporal coherence. Currently our motion diffusion is Gaussian, and possibly some form of anisotropic diffusion in the direction of motion could further enhance motion coherence. However we do not believe such extensions are necessary to show the value of our motion diffusion model and segmentation framework which in their current form already exhibit improved accuracy on state of the art approaches under the Berkeley F-measure. Furthermore, the dependency on data from only previous time-steps preserves the future possibility of applying an optimized version of our algorithm to online (incrementally streamed) video data. Although our run-time complexity is currently tens of seconds per frame, GPU implementations of the bottle-neck in our system (α-expansion) are available. These future applications are in line with our original project motivations which are to to develop a coherent video object segmentation algorithm for multimedia graphics applications such as video stylization [4].

ACKNOWLEDGMENT

The authors would like to thank D. Slatter, P. Cheattle, and D. Greig at Hewlett Packard Labs Bristol for their valuable discussions and input to this work.

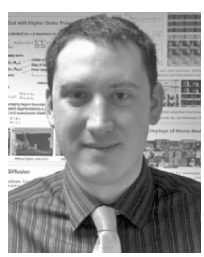
REFERENCES

[1] D. B. Goldman, B. Curless, S. Seitz, and D. Salesin, "Schematic storyboarding for video visualization and editing," in *Proc. ACM SIGGRAPH*, 2006, pp. 862–871.
[2] J. Collomosse, G. McNeill, and Y. Qian, "Storyboard sketches for content based video retrieval," in *Proc. ICCV*, 2009.
[3] M. A. Ahmed, F. Pitie, and A. Kokaram, "Extraction of non-binary blotch mattes," in *Proc. Int. Conf. Image Processing (ICIP)*, 2009, pp. 2757–2760.
[4] T. Wang, C. J. , D. Slatter, P. Cheatle, and D. Greig, "Video stylization for digital ambient displays of home movies," in *Proc. ACM NPAR*, Jun. 2010, pp. 137–146.
[5] X. He, R. S. Zemel, and D. Ray, "Learning and incorporating top-down cues in image segmentation," in *Proc. ECCV*, 2006, pp. 338–351.
[6] A. Rabinovich, S. Belongie, T. Lange, and J. M. Buhmann, "Model order selection and cue combination for image segmentation," in *Proc. CVPR*, 2006, pp. 1130–1137.
[7] P. Arbelaez and L. D. Cohen, "Constrained image segmentation from hierarchical boundaries," in *Proc. CVPR*, 2008, pp. 1–8.
[8] P. Kohli, L. Ladicky, and P. H. S. Torr, "Robust higher order potentials for enforcing label consistency," *Int. J. Comput. Vis. (IJCV)*, 2009.
[9] T. H. Kim, K. M. Lee, and S. U. Lee, "Nonparametric higher-order learning for interactive segmentation," in *Proc. CVPR*, 2010, pp. 3201–3208.
[10] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 603–619, May 2002.
[11] X. Ren and J. Malik, "Learning a classification model for segmentation," in *Proc. ICCV*, 2003, vol. 1, pp. 10–17.
[12] Y. Boykov and M. Jolly, "Interactive graph cuts for optimal boundary and region segmentation of objects in n-d images," in *Proc. ICCV*, 2001, vol. 1, pp. 105–112.
[13] I. Patras, E. A. Hendriks, and R. L. Lagendijk, "Video segmentation by map labeling of watershed segments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1553–1567, Sep. 2001.
[14] S. Paris, "Edge-preserving smoothing and mean-shift segmentation of video streams," in *Proc. ECCV*, 2008, pp. 460–473.
[15] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. 8th Int. Conf. Computer Vision*, 2001, pp. 416–423.
[16] D. Dementhon, "Spatio-temporal segmentation of video by hierarchical mean shift analysis," *J. Image Vis. Comput.*, 2002.
[17] J. Wang, B. Thiesson, Y. Xu, and M. Cohen, "Image and video segmentation by anisotropic kernel mean shift," in *Proc. ECCV*, 2004, pp. 238–249.
[18] J. Shi and J. Malik, "Motion segmentation and tracking using normalized cuts," in *Proc. ICCV*, 1998, pp. 1154–1160.
[19] M. Ristivojevic and J. Konrad, "Space-time image sequence analysis: Object tunnels and occlusion volumes," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 364–376, Feb. 2006.
[20] H. Greenspan, J. Goldberger, and A. Mayer, "A probabilistic framework for spatio-temporal video representation and indexing," in *Proc. ECCV*, 2002, pp. 461–475.
[21] M. Grundmann, V. Kwatra, M. Han, and I. Essa, "Efficient hierarchical graph based video segmentation," in *Proc. CVPR*, 2010.
[22] F. Moscheni, S. Bhattacharjee, and M. Kunt, "Spatiotemporal segmentation based on region merging," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 9, pp. 897–915, Sep. 1998.
[23] J. Collomosse, D. Rowntree, and P. Hall, "Stroke surfaces: Temporally coherent artistic animations from video," *Trans. Vis. Comput. Graph.*, vol. 11, no. 5, pp. 540–549, Sep./Oct. 2005.
[24] W. Brendel and S. Todorovic, "Video object segmentation by tracking regions," in *Proc. ICCV*, 2009.
[25] T. Wang, J.-Y. Guillemaut, and J. Collomosse, "Multi-label propagation for coherent video segmentation and artistic stylization," in *Proc. IEEE Int. Conf. Image Processing (ICIP)*, 2010.
[26] B. Price, B. Morse, and S. Cohen, "Learning-based interactive video segmentation by evaluation of multiple propagated cues," in *Proc. ICCV*, 2009.
[27] X. Bai, J. Wang, and G. Sapiro, "Dynamic color flow: A motion-adaptive color model for object segmentation in video," in *Proc. Eur. Conf. Computer Vision (ECCV)*, 2010, pp. 617–630.
[28] D. Hoiem, A. A. Efros, and M. Hebert, "Geometric context from a single image," in *Proc. ICCV*, Oct. 2005, pp. 654–661.
[29] B. C. Russell, W. T. Freeman, A. A. Efros, J. Sivic, and A. Zisserman, "Using multiple segmentations to discover objects and their extent in image collections," in *Proc. CVPR*, 2006, pp. 1605–1614.
[30] Y. J. Lee and K. Grauman, "Object-graphs for context-aware category discovery," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 2, pp. 346–358, Feb. 2012.
[31] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. ICML*, 2001, pp. 282–289.
[32] C. Liu, J. Yuen, A. B. Torralba, J. Sivic, and W. T. Freeman, "Sift flow: Dense correspondence across different scenes," in *Proc. ECCV*, 2008, pp. 28–42.
[33] T. Leung and J. Malik, "Representing and recognizing the visual appearance of materials using three-dimensional textons," *Int. J. Comput. Vision*, vol. 43, pp. 29–44, Jun. 2001.
[34] M. Varma and A. Zisserman, "A statistical approach to texture classification from single images," *Int. J. Comput. Vision*, vol. 62, pp. 61–81, 2005.
[35] J. Winn, A. Criminisi, and T. Minka, "Object categorization by learned universal visual dictionary," in *Proc. Int. Conf. Computer Vision (ICCV)*, 2005, pp. 1800–1807.
[36] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context," *Int. J. Comput. Vision*, vol. 81, pp. 2–23, Jan. 2009.
[37] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," *ACM Trans. Graph.*, vol. 23, pp. 309–314, 2004.
[38] Y. Boykov and G. Funka-Lea, "Graph cuts and efficient n-d image segmentation," *Int. J. Comput. Vision*, vol. 2, no. 70, pp. 109–131, 2006.
[39] B. C. Russell, W. T. Freeman, A. A. Efros, J. Sivic, and A. Zisserman, "Using multiple segmentations to discover objects and their extent in image collections," in *Proc. CVPR*, 2006, pp. 1605–1614.
[40] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 11, pp. 1222–1239, Nov. 2001.
[41] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 9, pp. 1124–1137, Sep. 2004.
[42] K. Alahari, P. Kohli, and Torr, "Reduce, reuse & recycle: Efficiently solving multi-label mrfs," in *Proc. CVPR*, 2008, pp. 1–8.
[43] S. Khan and M. Shah, "Object based segmentation of video using color, motion and spatial information," in *Proc. Computer Vision and Pattern Recognition (CVPR)*, 2001.
[44] R. Hu, T. Wang, and J. Collomosse, "A bag of regions based approach to sketch based image retrieval," in *Proc. Int. Conf. Image Processing (ICIP)*, 2011.

Tinghuai Wang (S'10) received the M.Sc. degree in electrical engineering from the Royal Institute of Technology (KTH) and RWTH Aachen University in 2007. He is pursuing the Ph.D. degree based in the Centre for Vision Speech and Signal Processing at the University of Surrey, Surrey, U.K., and HP Labs in Bristol, U.K., funded by the HP Labs Innovation Research Programme.

He was with Chinese Academy of Science as a research engineer in 2008. He was a research intern in Sony China Research Lab and HP Labs in Bristol, where he is now a visiting researcher. His research interests lie in the convergence area between computer graphics and vision.

Mr. Wang is a student member of the IEEE Signal Processing Society.



John Collomosse (M'09) received the Ph.D. degree on the topic of non-photorealistic rendering (NPR) and computer vision from the University of Bath, Bath, U.K., in 2004.

He is a Lecturer (Assistant Professor) within the Centre for Vision Speech and Signal Processing (CVSSP) at the University of Surrey, Surrey, U.K. Prior to joining CVSSP, he held a lectureship for 4 years in the Department of Computer Science at the University of Bath. His research explores the relationships between artistic depiction and real-world imagery. Applications focus on NPR and the use of sketches to drive visual search of multimedia collections. He is a visiting researcher at Hewlett Packard (HP) Labs, Bristol, U.K.

Dr. Collomosse is a Chartered Engineer.