Multi-Cue Structure Preserving MRF for Unconstrained Video Segmentation

Saehoon Yi and Vladimir Pavlovic
Rutgers, The State University of New Jersey
110 Frelinghuysen Road, Piscataway, NJ 08854, USA
{shyi, vladimir}@cs.rutgers.edu

Abstract

Video segmentation is a stepping stone to understanding video context. Video segmentation enables one to represent a video by decomposing it into coherent regions which comprise whole or parts of objects. However, the challenge originates from the fact that most of the video segmentation algorithms are based on unsupervised learning due to expensive cost of pixelwise video annotation and intra-class variability within similar unconstrained video classes. We propose a Markov Random Field model for unconstrained video segmentation that relies on tight integration of multiple cues: vertices are defined from contour based superpixels, unary potentials from temporally smooth label likelihood and pairwise potentials from global structure of a video. Multi-cue structure is a breakthrough to extracting coherent object regions for unconstrained videos in absence of supervision. Our experiments on VSB100 dataset show that the proposed model significantly outperforms competing state-of-the-art algorithms. Qualitative analysis illustrates that video segmentation result of the proposed model is consistent with human perception of objects.

1. Introduction

Video segmentation is one of the important problems in video understanding. A video may contain a set of objects, from stationary to those undergoing dependent or independent motion. Human understands a video by recognizing objects and infers the video context(i.e. what is happening in the video) by observing their motion. Depending on the video context, parts or whole objects will have structured motion correlation. However, there may be unrelated entities such as background or auxiliary objects which form additional structures as well. Holistic representation of a video cannot effectively decompose and extract meaningful structure and it may increase intra-variability of a video class. The goal of video segmentation is to obtain coherent object regions over frames so that a video can be represented as a set of objects and a meaningful structure can be extracted.

Ideally, the ultimate goal of video segmentation is to obtain pixelwise semantic segmentation of videos, where the objective is not only to partition a video into object regions but to infer object label of each region. Semantic segmentation is actively investigated in urban driving scene understanding [2, 3, 6, 25]. However, the urban scene videos contain rigid objects such as buildings, cars or road with typically smooth motion. In general, it is more challenging to segment and classify object regions in unconstrained videos due to the labor cost and intra-class variability.

Another fundamental challenge in video segmentation is that the inherent video object hierarchy may be highly subjective. Annotations of multiple human annotators may vary significantly. For example, one annotator may assign a single label to the whole human body, whereas another annotator will label torso and leg part separately. Furthermore, some objects may not have strong correlation to one feature alone. For example, an object may have parts that show different color patterns but move consistently. Hence, in practice, one may induce a video segmentation with different levels of granularity from aggregated information of multi-cue feature channels.

In this paper, we propose a novel video segmentation model which integrates temporally smooth labels with global structure consistency with preserving object boundaries. Our contributions are as follows:

- We propose a video segmentation model that integrates temporally smooth but potentially weak video segmentation proposals with strong static object cues and low-level spatio-temporal cues of color, flow, texture and long trajectories.
- A video segmentation at different granularity levels is inferred through the process of graph edge consistency, which is computationally efficient compared to traditional hierarchy induction approaches.
- The proposed method infers precise coarse grained segmentation, where a segment may represent one whole object.

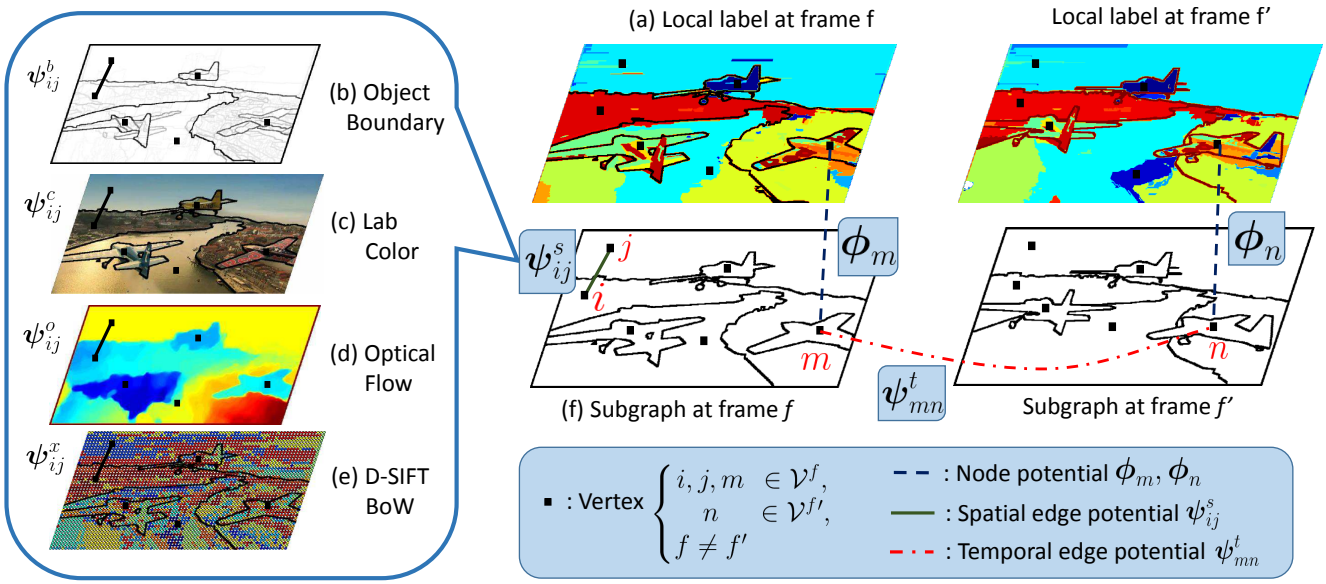


Figure 1: Overview of the framework. (a) temporally smooth pixelwise labels. (b $\sim$ e) Multi-cue spatial edge potentials. (f) Superpixels for corresponding vertices in the frame *f* are illustrated by object contours. Best viewed in color.

The remainder of this paper is organized as follows. Section 2 describes a set of related work and their limitations. Our proposed model is introduced in Section 3. Experiments set up and results are described in Section 4, followed by concluding remarks in Section 5.

2. Related Work

The problem of video segmentation is driven by several underlying goals. Some studies [15, 19, 23, 26] focus on object segmentation, whose aim is to detect, segment out, and track a few objects of interest along the video frames. On the other hand, [7, 8, 9, 10, 12, 13, 18, 24] seek to construct full pixelwise segmentation, where every pixel is assigned one of several labels. Xu and Corso [24] evaluated a set of pixelwise video segmentation algorithms where a video is represented as a graph and partition the graph according to several criteria. In this work, we specifically focus on the latter problem of pixelwise segmentation.

One of the main objectives of video segmentation is to obtain spatio-temporal smoothness of the region labels. Grundmann *et al*. [10] proposed a greedy agglomerative clustering algorithm that merges two adjacent superpixels if their color difference is smaller than internal variance of each superpixel. This results in smooth spatio-temporal segments. In addition, it implicitly detects new objects through the agglomerative clustering process. However, they only focus on color information without spatio-temporal structure. As a consequence, parts of one objects may merge with another object or with background, particularly in the coarse grained segmentation. Furthermore, the approach does not extract object boundaries effectively because the algorithm does not make use of the spatial structure from image gradients or edge detectors.

Object boundary contour defines spatial structure for image data. Arbelaez *et al*. [1] introduced a hierarchical contour detector for image segmentation. The contour strength provides a cue to understand this spatial structure. It is likely that a strong contour separates an object from other objects, while a weak contour only separates two parts inside of an object. However, the approach only works on static images and is not amenable to direct generalization to video data. Hence, applying this approach to videos would require matching of regions across subsequent frames, a nontrivial task.

Galasso *et al*. [9] aim to obtain correspondence of superpixels across video frames by propagating labels from a source frame along the optical flow. However, the quality of propagated labels typically decays due to flow estimation errors as the distance from the source frame increases. In motion segmentation, Elqursh and Elgammal [5] resolve the issue by splitting a group of trajectories if their dissimilarity becomes dominant. However, the robustness of this approach depends highly on the choice of a threshold parameter, which needs to be tuned for each video.

On the other hand, robust temporal structure information can be extracted from long-term trajectories. Ochs *et al*. [18] introduce a video segmentation framework that depends on long-term point trajectories from large displacement optical flow [4]. Although the proposed approach attains robust temporal consistency, it cannot distinguish objects of identical motion patterns because the trajectory

label only depends on motion. Nonetheless, the long trajectories offer a good cue to inferring long range temporal structure in a video. For instance, two superpixels in distant frames can be hypothesized to have common identity if they share sufficiently many trajectories.

Galasso *et al*. [7] aggregate a set of pairwise affinities in color, optical flow direction, long trajectory correspondence and adjacent object boundary. With aggregated pairwise affinity, they adopt spectral clustering to infer segment labels. However, Nadler and Galun [17] illustrate cases where spectral clustering fails when the dataset contains structures at different scales of size and density for different clusters.

We propose a Markov Random Field(MRF) model whose vertices are defined on object contour based regions. The model takes temporally smooth label likelihood as node potentials and global spatio-temporal structure information is incorporated as edge potentials in multi-modal feature channels, such as color, motion, object boundary, texture and long trajectories. Since the proposed model takes contour based superpixels as vertices, the inferred segmentation preserves strong object boundaries. In addition, the model enhances long range temporal consistency over label propagation by incorporating global structure. Moreover, we aggregate multi-modal features in the video so that the model can distinguish objects of identical motion. Finally, MRF inference with unary and pairwise potential results in accurate segmentation compared to spectral clustering which only relies on pairwise relationship. As a result, the proposed model infers video segmentation labels by preserving accurate object boundaries which are locally smooth and consistent to global spatio-temporal structure of the video.

3. Proposed Model

3.1. Multi-Cue Structure Preserving MRF Model

An overview of our framework for video segmentation is depicted in Figure 1. A video is represented as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where a vertex set $\mathcal{V} = \{\mathcal{V}^1, \ldots, \mathcal{V}^F\}$ is defined by object contours from all frames $f \in \{1, \ldots, F\}$ in the video. For each frame, an object contour map is obtained from contour detector [1]. A region enclosed by a contour forms a superpixel. An edge set $\mathcal{E} = \{\mathcal{E}^s, \mathcal{E}^t\}$ describes relationship for each pair of vertices. The edge set consists of spatial edges $e_{ij} \in \mathcal{E}^s$ where $i, j \in \mathcal{V}^f$ and temporal edges $e_{ij} \in \mathcal{E}^t$ where $i \in \mathcal{V}^f, j \in \mathcal{V}^{f'}, f \neq f'$.

Video segmentation is obtained by MAP inference on a Markov Random Field $\mathbf{Y} = \{y_i | i \in \mathcal{V}, y_i \in \mathcal{L}\}$ on this graph $\mathcal{G}$, where $P(Y) = \frac{1}{Z}\exp(-E(Y))$ and Z is the partition function. Vertex i is labeled as y_i from the label set $\mathcal{L}$ of size L. MAP inference is equivalent to the following energy minimization problem.

$$\min\ E(Y) = \sum_{i \in \mathcal{V}} \boldsymbol{\phi}_i \cdot \mathbf{p}_i + \sum_{(i,j) \in \mathcal{E}} \boldsymbol{\psi}_{ij} : \mathbf{q}_{ij}, \tag{1}$$

$$\text{s.t.} \sum_{l \in \mathcal{L}} p_i(l) = 1, \qquad \forall i \in \mathcal{V} \tag{2}$$

$$\sum_{l' \in \mathcal{L}} q_{ij}(l, l') = p_i(l), \qquad \forall (i,j) \in \mathcal{E}, l \in \mathcal{L} \tag{3}$$

$$\mathbf{p}_i \in \{0,1\}^L, \qquad \forall i \in \mathcal{V} \tag{4}$$

$$\mathbf{q}_{ij} \in \{0,1\}^{L \times L}, \qquad \forall (i,j) \in \mathcal{E} \tag{5}$$

In (1), $\boldsymbol{\phi}_i$ represents node potentials for a vertex $i \in \mathcal{V}$ and $\boldsymbol{\psi}_{ij}$ is edge potentials for an edge $e_{ij} \in \mathcal{E}$. As with the edge set $\mathcal{E}$, edge potentials are decomposed into spatial and temporal edge potentials, $\boldsymbol{\psi} = \{\boldsymbol{\psi}^s, \boldsymbol{\psi}^t\}$. The vector $\mathbf{p}_i$ indicates label y_i and $\mathbf{q}_{ij}$ is the label pair indicator matrix for y_i and y_j. Operators $\cdot$ and : represent inner product and Frobenius product, respectively. Spatial edge potentials are defined for each edge which connects the vertices in the same frame $i, j \in \mathcal{V}^f$. In contrast, temporal edge potentials are defined for each pair of vertices in the different frames $i \in \mathcal{V}^f, j \in \mathcal{V}^{f'}, f \neq f'$. It is worth noting that the proposed model includes spatial edges between two vertices that are not spatially adjacent and, similarly, temporal edges are not limited to consecutive frames.

A set of vertices of the graph is defined from contour based superpixels such that the inferred region labels will preserve accurate object boundaries. Node potential parameters are obtained from temporally smooth label likelihood. Edge potential parameters aggregate appearance and motion features to represent global spatio-temporal structure of the video. MAP inference of the proposed Markov Random Field(MRF) model will infer the region labels which preserve object boundary, attain temporal smoothness and are consistent to global structure. Details are described in the following sections.

3.2. Node Potentials

As described in Section 3.1, a vertex or a node is defined by an object contour proposal. Arbelaez *et al*. [1] extract hierarchical object contours so that taking different threshold values on the contours will produce different granularity levels of the enclosed regions. In our proposed model, we construct a set of vertices $\mathcal{V}^f$ from a video frame f by a single threshold on contours which results in fine-grained(oversegmented) superpixels.

Within each frame f, unary node potential parameters $\boldsymbol{\phi}_i \in \mathbb{R}^L$ represent the cost of labeling vertex $i \in \mathcal{V}$ with labels $l \in \mathcal{L} = \{1, \ldots, L\}$. In a typical video segmentation MRF, node potentials define dependence of l on local appearance features, while the edges impose spatio-temporal smoothness. However, temporal edges depend on motion

estimation that can often be imprecise. Therefore, we redefine the role of node potentials to impose auxiliary smoothness based on weak but temporally smooth oversegmentation, as described below.

Consider a weak video segmentation based on the set of labels $\mathcal{L}$. This segmentation creates a set of superpixels proposals in frame f. Let $h_i(l)$ be the number of pixels with label l within the contour-defined node i, as proposed by the weak superpixels. Then, we define the node potential as:

$$\boldsymbol{\phi}_i = -\frac{[h_i(1), \ldots, h_i(L)]}{\sum_{l=1}^{L} h_i(l)}. \tag{6}$$

In this work, we use [10] to create these weak but temporally smooth proposals. Figure 1 (a) illustrates that a vertex has a mixture of weak pixelwise labels because the unstructured segmentation proposal of [10] is not aligned with object contours. The roles of nodes in our MSP-MRF model will be to enforce this missing contour smoothness.

3.3. Spatial Edge Potentials

Binary edge potential parameters ψ consist of two different types; spatial and temporal edge potentials, ψ^s and ψ^t, respectively . Spatial edge potentials ψ^s_{ij} model pairwise relationship of two vertices i and j within a single video frame f. We define these pairwise potentials as follows:

$$\boldsymbol{\psi}^s_{ij}(l, l') = \begin{cases} \frac{\psi^b_{ij}+\psi^c_{ij}+\psi^o_{ij}+\psi^x_{ij}}{4} & \text{if } l \neq l',\ \psi^s_{ij} \geq \tau, \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

A spatial edge potential parameter $\boldsymbol{\psi}^s_{ij}(l, l')$ is the (l, l') element of $\mathbb{R}^{L\times L}$ matrix which represents the cost of labeling a pair of vertices i and j as l and l', respectively. It takes Potts energy where all different pairs of label take homogeneous cost $\boldsymbol{\psi}^s_{ij}$. Spatial edge potentials $\boldsymbol{\psi}^s$ are decomposed into $\psi^b, \psi^c, \psi^o, \psi^x$, which represent pairwise potentials in the channel of object boundary, color, optical flow direction and texture. Pairwise cost of having different labels is high if the two vertices i and j have high affinity in the corresponding channel. As a result, edge potentials increase the likelihood of assigning the same label to vertices i and j during energy minimization.

The edge potentials take equal weights on all channels. Importance of each channel may depend on video context and different videos have dissimilar contexts. Learning weights of each channel is challenging and it is prone to overfitting due to high variability of video context and limited number of labeled video samples in the dataset. Hence, the propose model equally weights all channels.

The model controls the granularity of segmentation by a threshold τ. Section 3.5 discusses how MSP-MRF obtains segmentation in different granularity in detail. We next discuss each individual potential type in the context of our video segmentation model.

Algorithm 1 Minimum Max-edge Path Weight

procedure MMPW($\mathcal{V}, \mathcal{E}$)
  $d \leftarrow \infty$
  for $v \in \mathcal{V}$ **do**
    $d[v][v] \leftarrow 0$
  for $(u, v) \in \mathcal{E}$ **do**
    $d[u][v] \leftarrow b(e_{uv})$ ▷ assign boundary score
  for $k \in \mathcal{V}$ **do**
    for $i \in \mathcal{V}$ **do**
      for $j \in \mathcal{V}$ **do**
        if $d[i][j] > \max(d[i][k], d[k][j])$ **then**
          $d[i][j] \leftarrow \max(d[i][k], d[k][j])$
  return $d_{\text{MMPW}} \leftarrow d$

Object Boundary Potentials $\boldsymbol{\psi}^b$. Object boundary potentials $\boldsymbol{\psi}^b_{ij}$ evaluate cost of two vertices i and j in the same frame assigned to different labels in terms of object boundary information. The potential parameters are defined as follows:

$$\boldsymbol{\psi}^b_{ij} = \exp(-d_{\text{MMPW}}(i, j)/\gamma_b). \tag{8}$$

where $d_{\text{MMPW}}(i, j)$ represents the minimum boundary path weight among all possible paths from a vertex i to j. The potentials ψ^b are obtained from Gaussian Radial Basis Function(RBF) of $d_{\text{MMPW}}(i, j)$ with γ_b which is the mean of $d_{\text{MMPW}}(i, j)$ as a normalization term.

If the two superpixels i and j are adjacent, their object boundary potentials are decided by the shared object contour strength $b(e_{ij})$, where e_{ij} is the edge connects vertices i and j and the boundary strength is estimated from contour detector [1]. The boundary potentials can be extended to non-adjacent vertices i and j by evaluating a path weight from vertex i to j. The algorithm to calculate $d_{\text{MMPW}}(i, j)$ is described in Algorithm 1, which modifies Floyd-Warshall shortest path algorithm.

Typically, a path in a graph is evaluated by sum of edge weights along the path. However, in case of boundary strength between the two non-adjacent vertices in the graph, total sum of the edge weights along the path is not an effective measurement because the sum of weights is biased toward the number of edges in the path. For example, a path consists edges of weak contour strength may have the higher path weight than another path which consists of smaller number of edges with strong contour. Therefore, we evaluate a path by the maximum edge weight along the path and the path weight is govern by an edge of the strongest contour strength.

Figure 2 illustrates two different path weight models of the max edge weight and the sum edge weight. Figure 2 (a) illustrates contour strength where red color represents

high strength. Two vertices indicated by white arrows are selected in an airplane. In Figure 2 (b), two paths are displayed. *Path 2* consists of less number of edges but it intersects with a strong contour that represents boundary of the airplane. If we evaluate object boundary score between the two vertices, *Path 1* should be considered since it connects vertices within the airplane. Figure 2 (c) shows segmentation result thresholded on edge sum path weight from a vertex at tail to all the other vertices. It displays that the minimum path weight between the two vertices are evaluated on *Path 2*. On the other hand, Figure 2 (d) illustrates that max edge path weight takes *Path 1* as minimum path weight which conveys human perception of object hierarchy.

Color Potentials ψ^c. Color feature for each vertex is represented by a histogram of CIELab color space in the corresponding superpixel. Color potential ψ^c_{ij} between the vertex i and j is evaluated on two color histograms $\mathbf{h}^c_i$ and $\mathbf{h}^c_j$:

$$\boldsymbol{\psi}^c_{ij} = \exp(-d_{\text{EMD}}(\mathbf{h}^c_i, \mathbf{h}^c_j)/\gamma_c). \tag{9}$$

where $d_{\text{EMD}}(\mathbf{h}^c_i, \mathbf{h}^c_j)$ is Earth Mover's Distance(EMD) between $\mathbf{h}^c_i$ and $\mathbf{h}^c_j$ of vertices i and j and γ_c is the normalization parameter.

Earth Mover's Distance [20] is a distance measurement between two probability distributions. EMD is typically more accurate over χ^2 distance in color space of superpixels. An issue with χ^2 distance is that if the two histograms on simplex do not share non-zero color bins, the two histogram are evaluated with the maximum distance of 1. Therefore, distance of vertices i and j is the same as the distance between i and k, if i, j, k do not share any color bins. This occurs often when we compare color feature of superpixels because superpixel is intended to exhibit coherent color especially in the fine grained level. Superpixels on different objects or different parts of an object may have different colors. For example, if we use χ^2 distance to measure color difference of superpixels, distance between superpixels of red and orange will have the same distance of red and blue because they do not share color bins. However, this is not intuitive to human perception. In contrast, EMD considers distance among each color bin, hence it is able to distinguish non overlapping color histograms.

Optical Flow Direction Potentials ψ^o. In each video frame, motion direction feature of ith vertex can be obtained from a histogram of optical flow direction $\mathbf{h}^o_i$. As with the case of color potentials, we use EMD between the two histograms $\mathbf{i}^o_i$ and $\mathbf{i}^o_j$ to accurately estimate difference direction in motion:

$$\boldsymbol{\psi}^o_{ij} = \exp(-d_{\text{EMD}}(\mathbf{h}^o_i, \mathbf{h}^o_j)/\gamma_o) \tag{10}$$

where γ_o is the mean EMD distance on optical flow histogram.

Texture Potentials ψ^x. Dense SIFT features are extracted for each superpixel and Bag-of-Words(BoW) model is obtained from K-means clustering on D-SIFT features. We evaluate SIFT feature on multiple dictionaries of different K. Texture potentials $\boldsymbol{\psi}^x$ are calculated from RBF on χ^2 distance of two BoW histograms $\mathbf{h}^x_i$ and $\mathbf{h}^x_j$, which is a typical choice of distance measurement for BoW model:

$$\boldsymbol{\psi}^x_{ij} = \exp(-d_{\chi^2}(\mathbf{h}^x_i, \mathbf{h}^x_j)/\gamma_x) \tag{11}$$

where parameter γ_x is the mean χ^2 distance on D-SIFT word histogram.

3.4. Temporal Edge Potentials

Temporal edge potentials define correspondence of vertices at different frames. It relies on long trajectories which convey long range temporal dependencies and more robust than optical flow.

$$\boldsymbol{\psi}^t_{ij}(l, l') = \begin{cases} \frac{\psi^r_{ij} + \psi^c_{ij}}{2} & \text{if } l \neq l',\ \psi^t_{ij} \geq \tau, \\ 0 & \text{otherwise} \end{cases}, \tag{12}$$

$$\psi^r_{ij} = \frac{|T_i \cap T_j|}{|T_i \cup T_j|}, \tag{13}$$

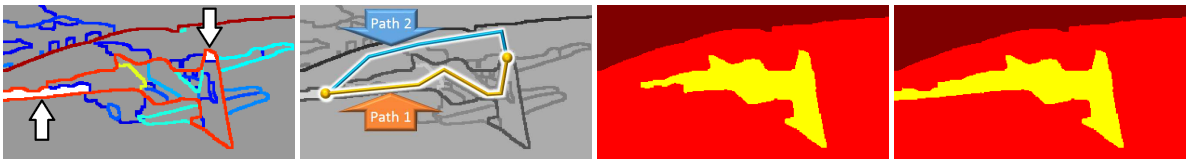
$$\psi^c_{ij} = \exp(-d_{\text{EMD}}(\mathbf{h}^c_i, \mathbf{h}^c_j)/\gamma_c). \tag{14}$$

where T_i is a set of long trajectories which pass through vertex i. Pairwise potential ψ^r_{ij} represents temporal correspondence of two vertices from overlapping ratio of long trajectories that vertices i and j shares, where $i \in \mathcal{V}^f, j \in \mathcal{V}^{f'}$ and $f \neq f'$. In order to distinguish two different objects of the same motion, we integrate color potentials ψ^c between two vertices. Long trajectories are extracted from [22].

3.5. Segmentation Label Inference with Variable Granularity

The proposed model initially forms a complete graph. Before the inference stage, the threshold τ in (7), (12) controls the number of edges in the graph. An edge potential $\boldsymbol{\psi_{ij}}$ in (1) penalizes assigning different labels to two vertices, with the inference algorithm assigning the same label if the overall energy decreases. Intuitively, lowering τ adds edges, which then strongly enforces similar vertices to be labeled identically in the MRF energy minimization. Hence, the model infers coarse-grained segmentation. This process resembles hierarchical segmentation. Although our approach does not guarantee hierarchical segmentation, it typically results in one that is controlled by τ, as evidenced by the smooth PR curve in Figure 5.

A conventional approach that enables hierarchical segmentation is to define a hierarchical vertex set in a graph as in the Hierarchical MRF(HMRF) [11, 27]. It introduces another set of vertices at different levels of hierarchy and



(a) Contour strength (b) Two contour paths (c) Edge sum path weight (d) Max edge path weight

Figure 2: Comparison of two types of path weight models.

edges which connect them. The large size of HMRFs typically makes the inference computationally infeasible.

Our proposed approach to obtain inference on different granularity labels takes computational advantages over graph representation with a hierarchical vertex set. The time complexity of computing node and edge potentials for a complete graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ are $O(|\mathcal{V}||\mathcal{L}|)$ and $O(|\mathcal{V}|^2)$, respectively. However, in practice, edges in the initial graph can be pruned away in the following way. For spatial edge potentials, one may compute the potential if the graph path length of two vertices in a frame is shorter than a threshold. For temporal edge potentials, one computes the potential of two vertices in different frames if they are overlapping by trajectories or optical flows. Once we construct the initial graph, we further threshold on different τ to get the segmentation inference of different granularities.

This paper focuses on constructing a graph that integrates temporally smooth labels with global structures in order to obtain precise video segmentation. However, one can further improve computational complexity for a very long video sequence by running inference on video clips of sliding windows with frame overlaps as in [10, 16, 24]. There will be a trade-off between preserving long term dependency and computational efficiency.

4. Experimental Evaluation

4.1. Dataset

We evaluate the proposed model on VSB100 video segmentation benchmark data provided by Galasso *et al*. [9]. There are a few additional video datasets which have pixelwise annotation. FBMS-59 dataset [18] consists of 59 video sequences and SegTrack v2 dataset [15] consists of 14 sequences. However, the both datasets annotate on a few major objects leaving whole background area as one label. It is more appropriate for object tracking or background subtraction task. On the other hand, VSB100 consists of 60 test video sequences of maximum 121 frames. For each video, every 20 frame is annotated with pixelwise segmentation labels by four annotators. The dataset contains the largest number of video sequences annotated with pixelwise label, which allows quantitative analysis. The dataset provides evaluation measurements in Boundary Precision-Recall(BPR) and Volume Precision-Recall(VPR), which evaluate overlap ratio of the object boundary and volume.

4.2. MSP-MRF Setup

In this section, we present the detailed setup of our Multi-Cue Structure Preserving Markov Random Field (MSP-MRF) model for unconstrained video segmentation problem. As described in Section 3.2, we take a single threshold on image contour, so that each frame contains approximately 100 superpixels. We assume that this granularity level is fine enough such that no superpixel at this level will overlay on multiple ground truth regions. Node potential (6) is evaluated for each superpixel with temporally smooth label obtained with agglomerative clustering [10]. Although we chose the 11th fine grained level of hierarchy, Section 4.4 illustrates that the proposed method shows stable performance over different label set size $|\mathcal{L}|$ for node potential. The average $|\mathcal{L}|$ on VSB100 video sequence is 355 labels for the chosen hierarchy. We assume the label size is large enough to encompass true segment labels.

The proposed model requires node potential to be evaluated on a temporally smooth segmentation label set $\mathcal{L}$. We chose [10] for $\mathcal{L}$ because their agglomerative clustering algorithm ensures temporal smoothness. Furthermore, our study shows that it achieves the-state-of-the-art performance among publicly available video segmentation implementations.

Finally, edge potential is estimated as in (7), (12). For color histograms, we used 50 bins for each CIELab color channel. In addition, 50 bins were set for horizontal and

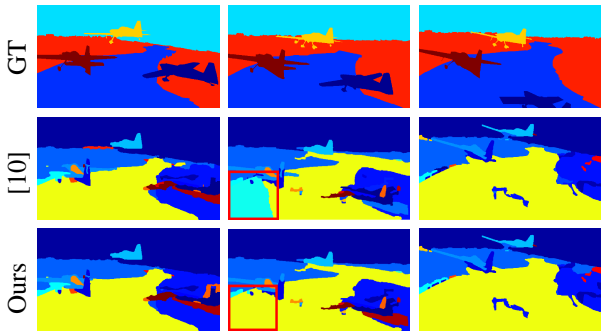


Figure 3: Temporal consistency recovered by MSP-MRF.

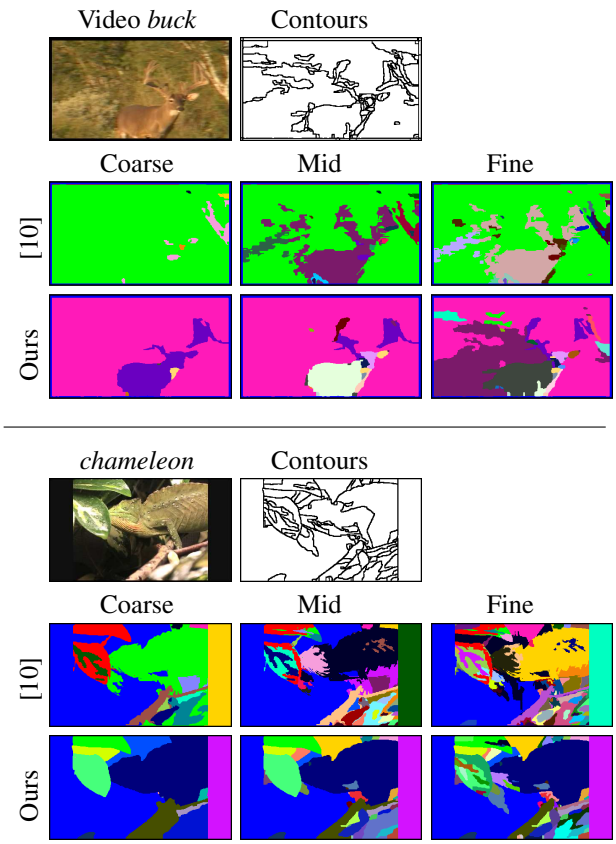


Figure 4: Segmentation comparison on two videos.

vertical motion of optical flow. For D-SIFT Bag-of-Words model, we used $K = 1000$ words. Energy minimization problem in (1) for MRF inference is optimized using FastPD algorithm [14].

4.3. Qualitative Analysis

Figure 3 illustrates a segmentation result on an *airplane* video sequence. MSP-MRF rectifies temporally inconsistent segmentation result of [10]. For example, in the fourth column of Figure 3, the red bounding boxes show MSP-MRF rectified label from Grundmann's result such that labels across frames become spatio-temporally consistent.

In addition, control parameter τ successfully obtains different granularity level of segmentation. For MSP-MRF, the number of region labels is decreased as τ decreases. Figure 4 compares video segmentation results of MSP-MRF with Grundmann's by displaying segmentation on the same granularity levels, where the two methods have the same number of segments in the video frame. MSP-MRF infers spatial smooth object regions, which illustrates the fact that the proposed model successfully captures spatial structure of objects. Furthermore, in the *buck* video sequence of Figure 4, the coarse segmentation result of [10] fails to identify the foreground object, whereas the proposed MSP-MRF retains a separate segmentation on the buck object. In addition, the fine level segmentation of [10] propagates the foreground segment label to a fragment of the background, which illustrates the lack of structural information in [10].

4.4. PR Curve on High recall regions

We specifically consider high recall regions of segmentation since we are typically interested in videos with relatively few objects. Our proposed method improves and rectifies state-of-the-art video segmentation of greedy agglomerative clustering [10], because we make use of structural information of object boundary, color, optical flow, texture and temporal correspondence from long trajectories. Figure 5 shows that the proposed method achieves significant improvement over state-of-the-art algorithms. MSP-MRF improves in both BPR and VPR scores such that it is close to *Oracle* which evaluates contour based superpixels on ground truth. Hence, it is worth noting that *oracle* is the best accuracy that MSP-MRF could possibly achieve because MSP-MRF takes contour based superpixels from [1] as well.

The proposed MSP-MRF model rectifies agglomerative clustering by merging two different labels of vertices if it reduces overall cost defined in (1). By increasing the number of edges in the graph by lowering threshold value, the model leads to coarser grained segmentation. As a result, MSP-MRF only covers higher recall regions from precision-recall scores of the selected label set size $|\mathcal{L}|$ from [10]. A hybrid model that covers high precision regions is described

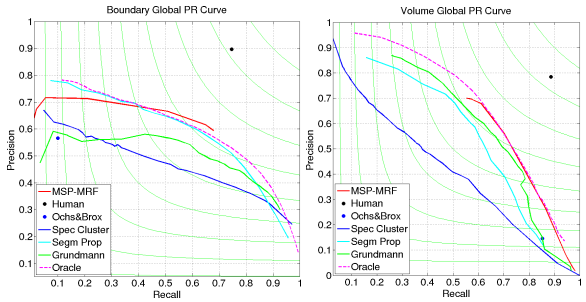


Figure 5: PR curve comparison to other models.

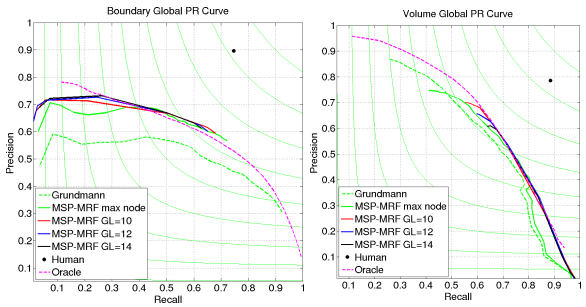


Figure 6: PR curve on different size of label set $\mathcal{L}$.

Table 1: Performance of MSP-MRF model compared with state-of-the-art video segmentation algorithms on VSB100.

	BPR			VPR			Length	NCL
Algorithm	ODS	OSS	AP	ODS	OSS	AP	$\mu(\delta)$	μ
Human	0.81	0.81	0.67	0.83	0.83	0.70	83.24(40.04)	11.90
Ochs and Brox [21]	0.17	0.17	0.06	0.25	0.25	0.12	87.85(38.83)	3.73
Spectral Clustering [7]	0.51	0.56	0.45	0.45	0.51	0.42	80.17(37.56)	8.00
Segmentation Propagation [9]	0.61	0.65	0.59	0.59	0.62	0.56	25.50(36.48)	258.05
$\mathcal{G}^{\mathcal{Q}} \equiv \mathcal{G}'$ SC [8]	0.62	0.66	0.54	0.55	0.59	0.55	61.25(40.87)	80.00
$[\mathcal{M}(\mathcal{G}^{\mathcal{SC}}{}_2)]^{\text{NCut}}$-1SC [12]	0.61	0.64	0.51	0.58	0.61	0.58	60.48(43.19)	50.00
Grundmann *et al*. [10]	0.57	0.62	0.48	0.61	0.65	0.61	51.83(39.91)	117.90
Khoreva *et al*. [13]	**0.64**	**0.70**	**0.61**	0.63	0.66	0.63	83.41(35.27)	50
MSP-MRF	0.63	0.67	0.60	**0.64**	**0.67**	**0.65**	35.83(38.93)	167.27
Oracle [9]	0.62	0.68	0.61	0.65	0.67	0.68	-	118.56

in Section 4.5.

Figure 6 illustrates the PR curve of MSP-MRF on different granularity levels of label set $|\mathcal{L}|$ in node potential (6). Dashed-green line is the result of greedy agglomerative clustering [10]. Solid-green line is the result of MSP-MRF with edge threshold τ set to 1, which leaves no edge in the graph. The figure shows that results of MSP-MRF are stable over different size of $|\mathcal{L}|$, particularly in the high recall regions.

4.5. Quantitative Analysis

The proposed model effectively merges labels of each pair of nodes according to edge set $\mathcal{E}$. As the number of edges increases, the size of the inferred label set will decrease from $|\mathcal{L}|$, which will cover higher recall regions. Although we are interested in high recall regions, the model needs to be evaluated on high precision regions of PR curve. For this purpose, we take a hybrid model that obtains rectified segmentation results from MSP-MRF on the high recall regions but retains segmentation result of [10] on high precision regions as an unrectified baseline.

Table 1 shows performance comparison to state-of-the-art video segmentation algorithms. The proposed MSP-MRF model outperforms state-of-the-art algorithms on most of the evaluation metrics. BPR and VPR is described in Section 4.1. Optimal dataset scale(ODS) aggregates F-scores on a single fixed scale of PR curve across all video sequences, while optimal segmentation scale(OSS) selects the best F-score with different scale for each video sequence. All the evaluation metrics are followed from dataset [9].

Khoreva *et al*. [13] achieves similar performance to MSP-MRF but it relies on model training over a large feature set. Our model uses no supervised training. As described in Section 4.4, *Oracle* is a model that evaluates contour based superpixels on ground truth. MSP-MRF infers segmentation label by integrating object boundary, global structure and temporal smoothness based on [10]. The result shows that incorporating boundary and global structure rectifies [10] by significant margin.

4.6. Contribution of Each Cue in Spatial Edge Potentials

We compare contributions of each cue in the spatial edge potentials. Table 2 shows the performance comparison of MSP-MRF on different combinations of cues. Although most of the combinations result in stable performance, the result without boundary potential achieves low average precision suggesting that its contribution is significant compared to other cues.

Table 2: Comparison on different cue combinations.

Multi-cue combination	BPR AP	VPR AP
All cues	0.60	0.65
Color+Flow+Texture	0.58	0.63
Boundary+Flow+Texture	0.59	0.64
Boundary+Color+Texture	0.59	0.64
Boundary+Color+Flow	0.60	0.64

5. Conclusion

In this paper, we have presented a novel video segmentation model that considers three important aspects of video segmentation. The model preserves object boundary by defining vertex set from contour based superpixels. In addition, temporally smooth label is inferred by providing unary node potential from agglomerative clustering label likelihood. Finally, global structure is enforced from pairwise edge potential on object boundary, color, optical flow motion, texture and long trajectory affinities. Experimental evaluation shows that the proposed model outperforms state-of-the-art video segmentation algorithm on most of the metrics.

References

[1] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 33(5):898–916, May 2011.

[2] V. Badrinarayanan, F. Galasso, and R. Cipolla. Label propagation in video sequences. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[3] G. J. Brostow, J. Fauqueur, and R. Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 2008.

[4] T. Brox and J. Malik. Large displacement optical flow: descriptor matching in variational motion estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 33(3):500–513, 2011.

[5] A. Elqursh and A. M. Elgammal. Online motion segmentation using dynamic label propagation. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2008–2015, 2013.

[6] B. Frhlich, E. Rodner, M. Kemmler, and J. Denzler. Large-scale Gaussian process multi-class classification for semantic segmentation and facade recognition. *Machine Vision and Applications*, 24(5):1043–1053, 2013.

[7] F. Galasso, R. Cipolla, and B. Schiele. Video segmentation with superpixels. In *Asian Conference on Computer Vision (ACCV)*, 2012.

[8] F. Galasso, M. Keuper, T. Brox, and B. Schiele. Spectral graph reduction for efficient image and streaming video segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[9] F. Galasso, N. S. Nagaraja, T. J. Cardenas, T. Brox, and B. Schiele. A unified video segmentation benchmark: Annotation, metrics and analysis. In *IEEE International Conference on Computer Vision (ICCV)*, December 2013.

[10] M. Grundmann, V. Kwatra, M. Han, and I. Essa. Efficient hierarchical graph based video segmentation. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[11] M. Keuper, T. Schmidt, M. Rodriguez-Franco, W. Schamel, T. Brox, H. Burkhardt, and O. Ronneberger. Hierarchical markov random fields for mast cell segmentation in electron microscopic recordings. In *International Symposium on Biomedical Imaging (ISBI)*, pages 973 – 978, 2011.

[12] A. Khoreva, F. Galasso, M. Hein, and B. Schiele. Learning must-link constraints for video segmentation based on spectral clustering. In *German Conference on Pattern Recognition (GCPR)*, 2014.

[13] A. Khoreva, F. Galasso, M. Hein, and B. Schiele. Classifier based graph construction for video segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[14] N. Komodakis and G. Tziritas. Approximate labeling via graph cuts based on linear programming. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 29(8):1436–1453, Aug. 2007.

[15] F. Li, T. Kim, A. Humayun, D. Tsai, and J. M. Rehg. Video segmentation by tracking many figure-ground segments. In *IEEE International Conference on Computer Vision (ICCV)*, 2013.

[16] O. Miksik, V. Vineet, P. Perez, and P. H. S. Torr. Distributed non-convex admm-inference in large-scale random fields. In *British Machine Vision Conference (BMVC)*, 2014.

[17] B. Nadler and M. Galun. Fundamental limitations of spectral clustering methods. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems (NIPS)*, Cambridge, MA, 2007. MIT Press.

[18] P. Ochs, J. Malik, and T. Brox. Segmentation of moving objects by long term video analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 36(6):1187 – 1200, Jun 2014.

[19] A. Papazoglou and V. Ferrari. Fast object segmentation in unconstrained video. In *IEEE International Conference on Computer Vision (ICCV)*, December 2013.

[20] O. Pele and M. Werman. Fast and robust earth mover's distances. In *IEEE International Conference on Computer Vision (ICCV)*, 2009.

[21] P.Ochs and T.Brox. Object segmentation in video: a hierarchical variational approach for turning point trajectories into dense regions. In *IEEE International Conference on Computer Vision (ICCV)*, 2011.

[22] T.Brox and J.Malik. Object segmentation by long term analysis of point trajectories. In *European Conference on Computer Vision (ECCV)*, Lecture Notes in Computer Science. Springer, Sept. 2010.

[23] D. Tsai, M. Flagg, A. Nakazawa, and J. Rehg. Motion coherent tracking using multi-label MRF optimization. *International Journal of Computer Vision (IJCV)*, 100(2):190–202, 2012.

[24] C. Xu and J. J. Corso. Evaluation of super-voxel methods for early video processing. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[25] C. Zhang, L. Wang, and R. Yang. Semantic segmentation of urban scenes using dense depth maps. In *European Conference on Computer Vision (ECCV)*, pages 708–721, Berlin, Heidelberg, 2010. Springer-Verlag.

[26] D. Zhang, O. Javed, and M. Shah. Video object segmentation through spatially accurate and temporally dense extraction of primary object regions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 0, pages 628–635, Los Alamitos, CA, USA, 2013. IEEE Computer Society.

[27] J. Zhu, Z. Nie, B. Zhang, and J.-R. Wen. Dynamic hierarchical markov random fields and their application to web data extraction. In *International Conference on Machine learning (ICML)*, pages 1175–1182, New York, NY, USA, 2007. ACM.