Exploiting Nonlocal Spatiotemporal Structure for Video Segmentation

Hsien-Ting (Tim) Cheng and Narendra Ahuja
Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign, Urbana, 61801, IL, USA
{hcheng8,n-ahuja}@illinois.edu

Abstract

Unsupervised video segmentation is a challenging problem because it involves a large amount of data, and image segments undergo noisy variations in color, texture and motion with time. However, there are significant redundancies that can help disambiguate the effects of noise. To exploit these redundancies and obtain the most spatio-temporally consistent video segmentation, we formulate the problem as a consistent labeling problem by exploiting higher order image structure. A label stands for a specific moving segment. Each segment (or region) is treated as a random variable which is to be assigned a label. Regions assigned the same label comprise a 3D space-time segment, or a region tube. The labels can also be automatically created or terminated at any frame in the video sequence, to allow objects entering or leaving the scene. To formulate this problem, we use the CRF (conditional random field) model. Unlike conventional CRF which has only unary and binary potentials, we also use higher order potentials to favor label consistency among disconnected spatial and temporal segments. Compared to region tracking based methods, the main advantages of the proposed algorithm are two fold: (1) the label consistency constraints are imposed on multiple regions but in a soft manner, and (2) the labeling decision is postponed until the confidence in the labeling is high. We compare our results with a recent state-of-the-art video segmentation algorithm and show that our results are quantitatively and qualitatively better.

1. Introduction

Analogous to image segmentation, which partitions the image into groups of pixels with photometric similarity and outputs each group as a two-dimensional region, video segmentation partitions the three-dimensional spatio-temporal space into 3D regions (or region tubes) each having photometric coherence formed by the same region moving through consecutive frames. It is an important computer vision problem [16, 13, 2] with applications in areas such activity recognition, video analysis, summarization, surveillance and browsing.

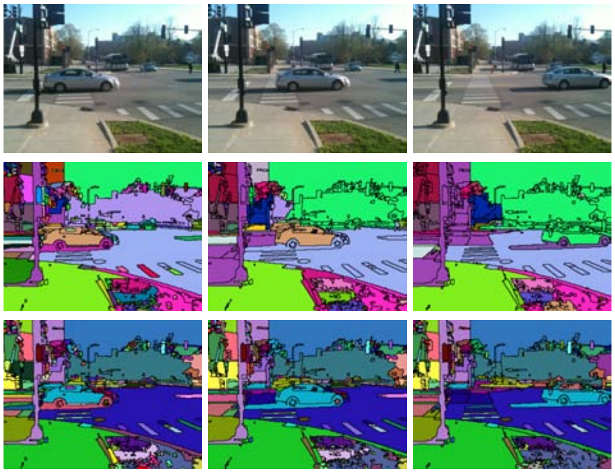


Figure 1. **Comparison between binary potential and higher order potential**. *First row*: three sample frames from the original sequence. *Second row*: results with only unary and binary potentials. *Third row*: results using higher order potentials. As can be seen in the second column, for binary potentials, the background building and tree are merged with the sky, and the zebra crossings are merged with the road. Even the car merges with the surround in the third column. By incorporating higher order potentials, the salient details are retained (*e.g.*, zebra crossing) while the low contrast regions do merge *e.g.*, street light.

One line of research on video segmentation exploits grouping raw pixels across frames [18, 7, 9]. The pixels are represented as a multi-dimensional (space-time) features consisting of photometric, spatial and motion properties. However, this becomes infeasible even for medium-sized videos. Hence methods that use segmentation results of each image followed by tracking these regions have attracted much attention in recent years [12, 21]. Nevertheless, the results are usually less desirable due to the well-known lack of repeatability of image segmentation across frames, and therefore decoupling of the video segmentation problem into two independent subproblems in space

978-1-4673-1228-8/12/$31.00 ©2012 IEEE

and time an unrealistic proposition. For example, regions in one frame across a low-contrast boundary may merge thus be undetected in the next frame. On the other hand, a large region contains a slight variation of brightness may be split into a set of several smaller homogeneous regions in the subsequent frames. Therefore some region tracking efforsts tend to assume (1) region properties are likely be the same in the consecutive frames, and (2) there exists an one-to-one correspondence for regions across frames. In recent years, several methods have been proposed to deal with the above-mentioned problems[4, 11, 10]. [11] assumes that high contrast contours of an *object* are repeated along the video sequence, and then use contour-based partial matching to obtain many-to-many region correspondences. [4] extends the idea of partial contour matching in the video segmentation domain with an efficient matching process using the DTW (dynamic time warping) algorithm. [10] treats video segmentation as a clustering problem in which a "region graph" from a over-segmented video volume is constructed followed by hierarchical merging to generate coherent region boundaries.

Notwithstanding their demonstrated success, there is a major drawback in the tracking based methods, namely, they must make hard decisions about identifying region correspondences and their merging/splitting for each pair of frames. Since we are concerned with bottom up processing, lacking a model of the target being tracked and therefore the nature of the best correspondences, the combinatorial nature of region merging/splitting makes finding optimal region correspondences between two frames an NP-hard problem. This implies that in a straight forward implementation, robust algorithms must take space-time constraints over many frames into consideration, for segmenting regions and determining their correspondences. In the approach we present in this paper, we achieve such robustness while avoiding explicit and hard decisions made from local space and temporal information.

By formulating the video segmentation as a higher order label consistency problem, we propose to solve the above problems via exploiting higher order (instead of local) spatial and temporal structure. Following are the salient features and contributions of our approach. (1) We treat each over-segmented region as a random variable. Random variables in different frames having the same label constitute a region tube. Hence photometric homogeneity within a region tube is achieved by enforcing *neighboring* regions with similar properties to take the same label. (2) Instead of region adjacency, the neighbor definition in common use, we group each region with a larger set of higher order neighbors, by forming its spatial and temporal *cliques*. (3) We do not make hard decisions on region merging or splitting to form the spatial cliques. Similarly, our temporal cliques also include all pixel correspondences across several frames suggested by local motion.. (4) We allow multiple objects entering or leaving the scene, with no assumptions about the number of labels. Label creation and termination are data-driven. (5) We solve the label consistency and competition problem via a conditional random field (CRF) formulation.

The work by Vazquez-Reina *et al.* [17] is one of the few algorithms that consider video segmentation as a labeling problem. They first enumerate multiple trajectories and treat each trajectory as a label and then use CRF to solve the label consistency and competition problem. Our work bears some resemblance to theirs but differs in several aspects. First, they use multiple segmentation as well as superposition while we apply multi-resolution segmentation and build a photometric tree as spatial clique. Second, they prune the label space at beginning by only allowing regions assigned to salient trajectories, while each region in our method can have its own unique label to make automatic label creation possible. Third, they assume corresponding regions in the consecutive frames must overlap with each other while we use optical flow to locate the correspondences.

2. Overview of the Proposed Approach

Unlike approaches that track regions across a pair of images, we simultaneously process a batch of frames to enforce spatial and temporal consistency. This is to reduce the accumulation of image segmentation errors that would be encountered in sequentially forming the region tubes. To formulate video segmentation as a labeling problem, we first construct a photometric segmentation tree for each frame by a multi-resolution segmentation [1]algorithm [1]. The regions having the finest (lowest) contrast are considered to form the leaf level. Each leaf node is assigned a random variable X_i. Together the set of all regions across frames, and their labels, define a random field $\mathbf{X}$. The regions corresponding to variables $\{X_i\}$ having a single label constitute a 3D region (tube.) While we make no assumptions about the number of consistent photometric tubes present in a video, the initial assignment of a unique label to each leaf corresponds to a gross overestimation of the number of tubes, and the process of obtaining the final label set $\mathcal{L}$ must prune the redundancies caused by distinct labels being assigned to connected regions. Objects entering or leaving are handled via automatic label creation and termination.

The consistent labeling problem is formulated as an inference process of a higher order CRF. Conventional CRF formulation uses unary potential and binary potential to ensure label consistency. However,using these two potentials alone tends to *oversmooth* the labeling, resulting in unnecessary region merging. To overcome this problem without sacrificing the tractability of the inference process, we apply the higher order potentials and the robust P^n model proposed by Kohli *et al*. [14] in the context of multi-class image segmentation. Each higher order potential is defined on the set of regions forming a spatial and temporal "*clique*". As indicated earlier, a spatial clique represents a non-leaf node in the photometric tree, which is the union of several regions r_i in the leaf level. A temporal clique is formed by region correspondences across frames. The potential function that penalizes labeling inconsistency in each clique is determined by the photometric property and motion information. As shown in [14], a general submodular higher order function can be transformed to a second order function if the higher order potential is defined using the P^n model. We apply the efficient graph cut based α-expansion and $\alpha\beta$-swap move algorithms of [3] to estimate labels.

3. Higher Order Consistent Labeling

In this section, we describe how to construct cliques and design the unary, binary, and higher order potential to achieve consistent labeling. We begin with the notation and definitions.

3.1. Preliminaries

Given frames $f_1, f_2, ...f_T$, a discrete random field $\mathbf{X}$ is defined over an index system $\mathcal{V} = \{1, 2, ..., M\}$ with a neighborhood system $\mathcal{N}$. Each random variable $X_i \in \mathbf{X}, i \in \mathcal{V}$ is associated with a leaf region in some frame. X_i would take a value from the label set $\mathcal{L} = \{l_1, l_2, ..., l_k\}$. Theoretically we can set the $|\mathcal{L}| = |\mathcal{V}|$ so each region can contribute an unique label l_i. In practice we perform labeling pruning during the construction of temporal clique to reduce redundant labels (detailed in sec3.3). The neighborhood system $\mathcal{N}$ of the random field is defined by the sets $\mathcal{N}_i, \forall i \in \mathcal{V}$, where $\mathcal{N}_i$ denotes the sets of all neighbors of the variable X_i (where, for brevity, we loosely refer to those labels belonging to the neighboring regions as neighboring labels.) Clique c is a set of random variables $\mathbf{X}_c$ which are conditionally dependent on each other. Both neighbors and cliques exist in two forms: spatial and temporal. Any possible assignment of labels to the random variables will be called a *labeling* (denoted by $\mathbf{x}$). The labels take values from the set $\mathbf{L} = \mathcal{L}^T$. A labeling $\mathbf{x}$ is interpreted as the estimated video segmentation. Leaf regions belonging to the same 3D segment are identified by the fact that the random variables associated with them take the same label.

From [15], the posterior probablity $\Pr(\mathbf{x}|\mathbf{D})$ of CRF given observed data $\mathbf{D}$ is a Gibbs distribution and can be written in the form: $\Pr(\mathbf{x}|\mathbf{D}) = \frac{1}{Z}\exp(-\sum_{c \in \mathcal{C}} \psi_c(\mathbf{x}_c))$, where Z is the normalizing constant known as the partition function, and $\mathcal{C}$ is the set of all cliques. The term $\psi_c(\mathbf{x}_c)$ is called the potential function of the clique c where $\mathbf{x}_c = \{x_i, i \in c\}$. The corresponding Gibbs energy is given by

$$E(\mathbf{x}) = -\log\Pr(\mathbf{x}|\mathbf{D}) - \log Z = \sum_{c \in \mathcal{C}} \psi_c(\mathbf{x}_c) \qquad (1)$$

The MAP labeling $\mathbf{x}^*$ of the random field is defined as:

$$\mathbf{x}^* = \text{argmax}_{\mathbf{x} \in \mathbf{L}} \Pr(\mathbf{x}|\mathbf{D}) = \text{argmin}_{\mathbf{x} \in \mathbf{L}} E(\mathbf{x}) \qquad (2)$$

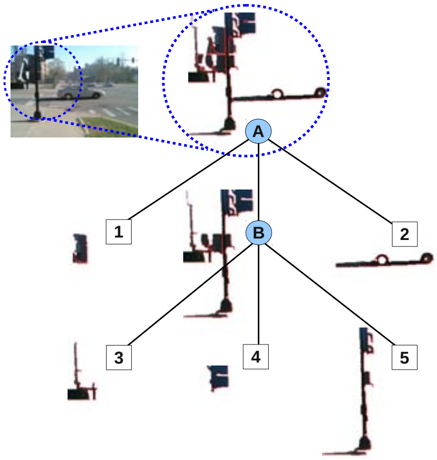


Figure 2. **Spatial cliques and photometric tree**. A and B are two interior regions in a photometric tree, each representing a spatial clique. Clique-A consists of leaf regions {1,2,3,4,5} and clique-B consists of {3,4,5}. It shows that a leaf region (4 here) may belong to multiple spatial cliques.

3.2. Spatial and Temporal Cliques

Conceptually, label consistency is achieved by penalizing variables having similar characteristics but taking different labels. In CRF we approach it by designing penalty potentials. We propose to incorporate higher order potentials to maintain label consistency:

$$E(\mathbf{x}) = \sum_i \psi_i(x_i) + \sum_{i,j \in \mathcal{N}_i} \psi_{i,j}(x_i, x_j) + \sum_{c \in \mathcal{S}} \psi_c(\mathbf{x}_c) \quad (3)$$

where c denotes a *clique* containing multiple regions. In conventional CRF the energy function consists of only the first two terms. If the neighborhood system $\mathcal{N}$ is defined according to region adjacency (both spatial and temporal), we have observed that a lack of higher order terms tends to over-smooth the region tube and lose finer variation (as shown in Figure 1). Therefore ψ_c should be designed such that it allows but penalizes inconsistent labels within the clique. In this section we discuss how to construct the cliques with the higher order potentials, and then describe the detail the formulation details of these potentials in the next section.

[1] The code can be downloaded from http://vision.ai.uiuc.edu/segmentation

To achieve label consistency in both spatial and temporal domain, we propose to construct two kinds of cliques: spatial clique c_s and temporal clique c_t, $c = c_s \cup c_t$. By using the multiple-resolution image segmentation algorithm [1] to obtain the initial image segmentation, we are able to build a (photometric) tree in which: interior regions are the union or merge of the children regions based on photometric similarity. As we traverse up the region hierarchy, the photometric variance within a region gets larger. Since the random field $\mathbf{x}$ is defined on regions at leaf level, for ancestor regions to form a single tube requires label consistency among its descendant leaf regions. Hence each ancestor region in the photometric tree represents a spatial clique as illustrated in Figure 2. A leaf region r_i belongs to multiple spatial cliques c_s, each in general associated with a interior region.

We construct the *temporal cliques* in a similar manner but use motion properties, such as optical flow, for grouping instead of photometric values. For this, we define the overlap of a region r_i with another region r_j , hypothesized as the after-motion correspondence of r_i:

$$h_i(j) = \frac{|(p(i) + u(i)) \wedge p(j)|}{|p(i)|} \in [0, 1] \tag{4}$$

where $p(i)$ denotes the coordinates of pixels in r_i and $u(i)$ is the dense flow vector defined on the pixels in r_i. Note that $h(\cdot)$ is not a symmetric measure. We use the dense flow algorithm proposed in [5]. For a given leaf region r_i in some frame f_t, we find all regions r_j that belong to adjacent frames f_{t+1}, f_{t-1} and group them together if $h_i(j)$ exceeds a threshold T which is set to 0.5. When extended across frames, this grouping defines a temporal clique c_t for r_j. As illustrated in Figure 3, a region may belong to multiple c_t due to noise in image segmentation or optical flow. In both spatial and temporal cliques, this one-to-many scenario leads to a competition between labels. A good design of the potential function would resolve the competition efficiently, assigning the best label to each region.

3.3. Potential Functions

Unary potential Although each region r_i is made to initially contributes an unique label l_i to $\mathcal{L}$, the eventually surviving labels for all regions is sparse. This selection is initiated by constructing the spatial and temporal cliques c_s and c_t, which constrain the available labels for each region r_i to the set whose corresponding regions belong to the same cliques as r_i. Recall that each label l_i initially corresponds to a leaf region. Let $\tau : \mathcal{L} \rightarrow \mathcal{V}$ denote the function that retrieves region index from label. Then we define the unary potential as:

$$\psi_i(x_i) = \begin{cases} \theta_u d_c(i, \tau(x_i)) & \text{if } i, \tau(x_i) \in \text{some } c \\ \infty & \text{otherwise} \end{cases} \tag{5}$$

where $d_c(i, j)$ is the normalized ℓ_2-distance $\in [0, 1]$ between r_i and r_j in the LUV colors pace, and θ_u is a constant. Note that a region r_i can choose its own unique label l_i, and actually it favors doing so because $d_c(i, i) = 0$. That allows the possibility of assigning new label to objects entering the scene.

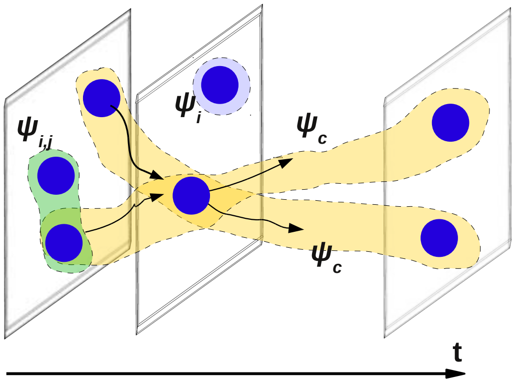


Figure 3. **Potential functions and temporal cliques**. Each leaf region is colored blue and there are three kinds of potentials defined on it: unary ψ_i, binary ψ_{ij} and higher order ψ_c. Here we show only the higher order temporal potential, defined on a temporal clique (yellow). Each temporal clique is a result of inexact region tracking, therefore a region may belong to multiple temporal cliques.

Binary potential We use adjacency matrices A^s and A^t to refer spatial and temporal region adjacencies. We assign the values $A^s_{ij} = 1$ if r_i and r_j are within the same frame and adjacent, and $A^s_{ij} = 0$ otherwise. Using Eq.4, we use $A^t_{ij} = 1$ if $h_i(j) > 0$ or $h_j(i) > 0$, and $A^t_{ij} = 0$ otherwise. We define the following potential for each pair of spatially and temporally adjacent regions r_i, r_j:

$$\psi_{i,j}(x_i, x_j) = \begin{cases} 0 & \text{if } x_i = x_j \\ \theta_b e^{-(w_1 d_c + \bar{w}_1 d_f)} & \text{if } x_i \neq x_j \text{ and } A^s_{ij} = 1 \\ \theta_b e^{-(w_2 d_c + \bar{w}_2 d_o)} & \text{if } x_i \neq x_j \text{ and } A^t_{ij} = 1 \end{cases} \tag{6}$$

where $w_i + \bar{w}_i = 1, i = \{1, 2\}$ and θ_b is a constant. Motion information influences the potential by d_f spatially and by d_o temporally. We define $d_f(i, j)$ as the distance between the flow distributions regions r_i and r_j, computed as $\mathcal{X}^2$-distance. Coupled with the color distance $d_c(i, j)$, it favors assignment of the same label to spatially adjacent regions with similar flow distribution and photometric property. For temporal adjacent regions, $d_o(i, j) = \frac{h_i(j)+h_j(i)}{2}$ the average of two way after-motion overlap, indicates how likely is r_i to move to r_j and vice versa. Coupled with d_c, this enforces consistent labeling.

Higher order potential As mentioned earlier in Section

3.1, each interior region in the photometric tree determines a valid spatial clique c_s. However, not all such interior cliques (e.g., obviously the root node which includes all leaf regions) should be expected to have high consistency among the labels of all the regions comprising them. As shown in Figure 2, although the car, the street light and the background building belong to one interior region, they should not all agree on the same label, e.g., due to their motion differences. The question then arises how to determine which subsets of labels are consistent. For example, how to set a threshold on the value of the variance within an interior region to determine consistency. Likewise, criteria are needed to determine consistency among labels within the temporal cliques. For example, using a loose after-motion overlap threshold T will lead to poor decisions about consistencies.

The above-mentioned problems with label consistency within cliques can be addressed by a potential function like the ones defined in the P^n model [14], where the model has been applied to the problem of supervised multi-class image segmentation. The P^n Potts model is defined in terms of clique c:

$$\psi_c(\mathbf{x}_c) = \begin{cases} \gamma_k & \text{if } x_i = l_k, \forall i \in c \\ \gamma_{\max} & \text{otherwise} \end{cases} \tag{7}$$

where $\gamma_{\max} > \gamma_k, \forall l_k \in \mathcal{L}$. If γ_k are equal, it enforces label consistency rigidly. For instance, if all but one of the regions in a clique take the same label then the penalty incurred is the same as if they were all to take different labels. We extend this to the the P^n model by relaxing the consistency constraint as follows:

$$\psi_c(\mathbf{x}_c) = \min\Big(\min_{l_k \in \mathcal{L}}(|c| - n_k(\mathbf{x}_c))\theta_k + \gamma_k), \gamma_{\max}\Big) \tag{8}$$

where $|c|$ is the number of regions in clique c, $n_k(k)$ denotes the number of regions in c which take the label k in labeling $\mathbf{x}_c$, and $\gamma_k, \theta_k, \gamma_{\max}$ are potential function parameters which satisfy the constraints: $\theta_k = \frac{\gamma_{\max} - \gamma_k}{Q_k}$ and $\gamma_k \leq \gamma_{\max}, \forall l_k \in \mathcal{L}$. Here we design γ_k to be inversely proportional to the "*inseparability*" between the corresponding region r_k and the clique c: $\gamma_k = \exp(-\mathcal{I}_c(r_k))$. $\mathcal{I}(\cdot)$ is the measurement of inseparability and defined as the average contrast along the boundary. In a spatial clique it is the contrast along the region boundary. In a temporal clique,it is the contrast between the overlapping pixels in the consecutive frames; the less the contrast, the harder it is to separate the regions. The truncation parameter Q_k controls the rigidity of the potential and here we set $Q_k = \frac{|c|}{2}$.

3.4. Inference and Complexity

Once we have defined unary, binary and higher order potentials for the objective function in Eq.3, the optimal labeling $\mathbf{x}^*$ is obtained by the CRF energy minimization inference process Eq.2.For this we use the algorithms presented in [14] which shows that the *move energy* function of higher order potentials in the robust P^n model can be transformed to submodular quadratic, hence the energy minimization is achieved by a series of α-expansion moves which can be solved efficiently by st-mincut algorithm [3]. The total inference time is: num(cycles)×num(iterations)×T(st-mincut), *i.e.*

$O(|\mathcal{V}|^2|\mathcal{L}|^2\log(\frac{|\mathcal{V}|^2}{|\mathcal{L}|}))$ in worst case where $|\mathcal{V}|$ is the number of variables and $|\mathcal{L}|$ is the number of labels. However in practice num(cycles) can be considered as a constant and dropped from $O(|\mathcal{V}|)$ yielding

$O(|\mathcal{V}||\mathcal{L}|^2\log(\frac{|\mathcal{V}|^2}{|\mathcal{L}|}))$.

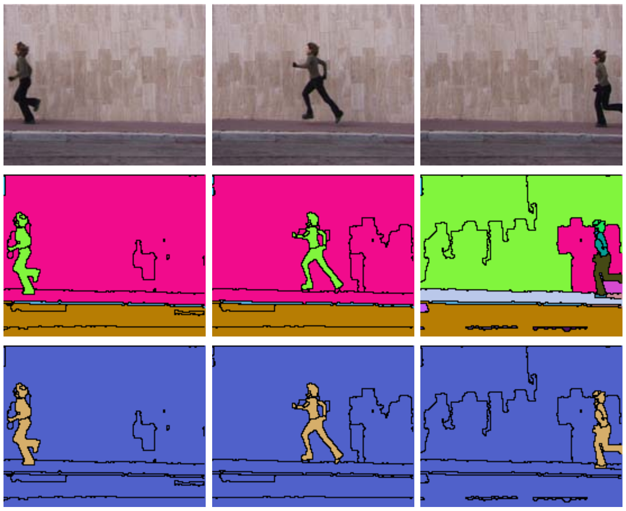


Figure 4. Video segmentation results for the *run* sequence from the Weizmann activity dataset. **Top row:** three consecutive frames (right to left) from *lena run2* activity; **middle row:** results using the mean shift segmentation algorithm; **bottom row:** results of the proposed method. The boundaries of the leaf regions are shown to demonstrate that the proposed method can overcome the instability arising from image segmentation. The background is separated into several, non-repetitive, irregular shape segments, and so is the foreground. Although the mean shift algorithm segments the left two frames successfully, it mislabels the left part of the wall as foreground in the right frame.

4. Implementation Details

To speedup the segmentation process we could reduce the leaf region number $|\mathcal{V}|$ or label size $|\mathcal{L}|$. For the first, we divide the video into overlapping chunks in which segmentation can be performed in parallel followed by merging of the result. The overlapping frames between chunks are used to propagate the labeling result across chunks.The resulting loss in lobal optimality may not be significant since mutual influence among frames decreases with their distance. we use chunks of 10∼15 frames with 1 frame overlap. Second, given the segmentation within each chunk, we observe the CRF inference process can be further decomposed into multiple inference sub-processes as long as their label spaces

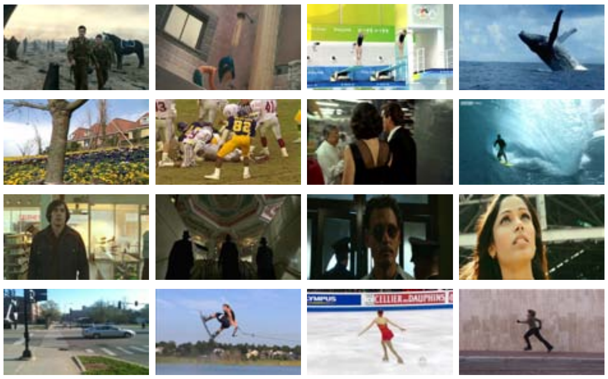


Figure 5. Example frames from video sequences. Top to bottom, left to right, the original sequences are: {atonement, coraline, diving, earth, flower garden, football, goodfellas, sufer, foroldmen, publicenemies1, publicenemies2, slumdog, UI traffic, waterski, ice skate and run example from Wiezeman}.

are disjoint. For example, a temporal clique c_i at upper-left corner is disjoint with another temporal clique c_j at lower-right corner, *i.e.* $c_i \wedge c_j = \emptyset$. Therefore the original label space $\mathcal{L}$ is partitioned into disjoint label subspace $\{\mathcal{L}_i\}$, the inference is performed in each $\mathcal{L}_i$ and the results are combined afterwards. With these speedup operations, it takes 5 seconds to segment each frame on an average (exluding the optical flow computation).

5. Experimental Results

We compare our approach (abbreviated as CRF in the context) against three other approaches: mean shift, state-of-the-art Grundmann's graph-based grouping approach [10] and our approach without using higher order spatial potentials (abbreviated as CRF − SP). The potential parameters are set the same across all video sequences: $\theta_u = 1000, \theta_b = 50, w_i = \hat{w}_i = 0.5$ for $i = 1, 2$. They are determined by a preliminary human sanity test. Since no benchmark for generic video segmentation is available, we conduct our experiments with standard datasets to evaluate the work both quantitatively and qualitatively. First we report the precision and recall on the Weizmann activity [8] dataset in Table 1. It consists of 90 videos: 10 distinct human activities (bend, jump, run *etc.*) each with 9 human subjects with foreground and background ground truth labeling. The average precision and recall rates over all videos are shown in Table 1. The rates are computed in terms of the total image area across all frames correctly segmented as foreground and background. For a given region tube with the same label (or the regions within the same cluster for the mean shift algorithm) we classify it as foreground if majority the area is covered by foreground, and vice versa. From Table 1, it is clear that the CRF method outperforms the mean shift by significant margin. It also demonstrates the role of higher order spatial structure in achieving label consistency. Examples of the foreground and background segmentations obtained by our algorithm are shown in Figure 4. The boundaries of each leaf region are shown to visualize regions merging effect. We do not display the boundary for the following more textured image sequences to make the figure less noisy.

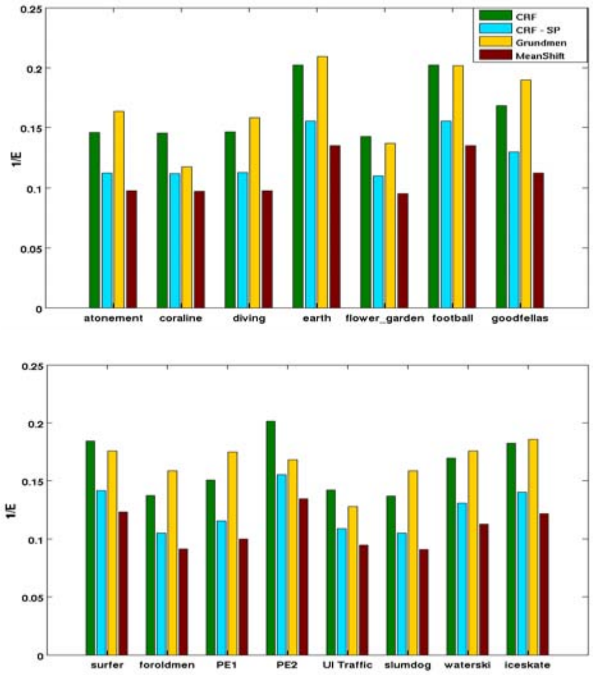


Figure 6. Quantitative measurement $\frac{1}{E}$ among 4 methods (CRF, CRF−SP, Grundmann[10] and mean shift) across 15 videos depicted in Figure 5. Higher value indicated better quality.

We next evaluate the methods on 15 more complicated videos as depicted in Figure 5. Since humans tend to perform object-level segmentation, producing ground truth for the textured image sequence via human annotation may not fully capture the performance. Performance evaluation on generic image/video segmentation has been discussed in [20, 6]. In this work we use entropy as a measure $\frac{1}{E}$, computed as [19]:

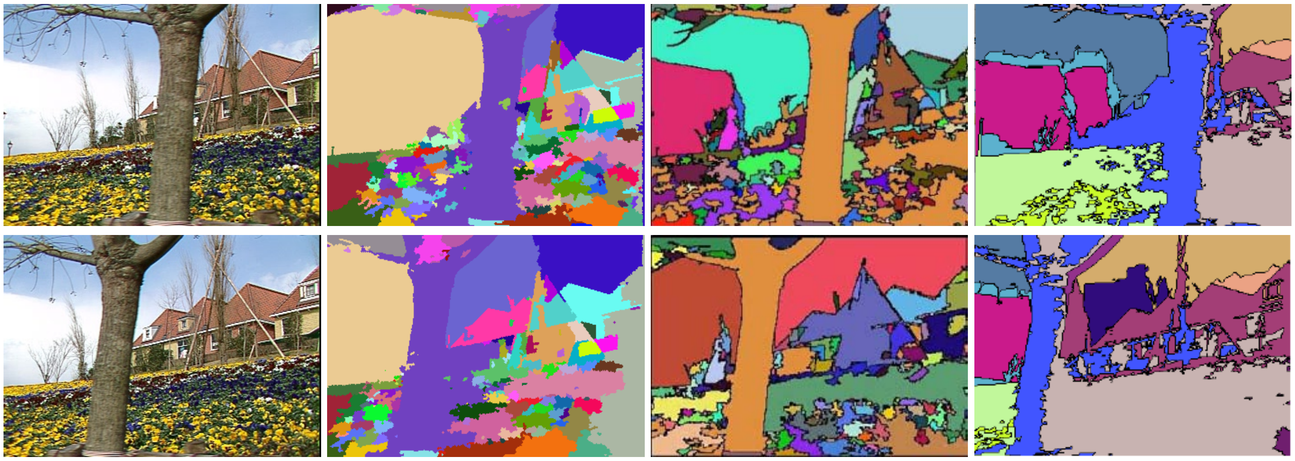
$$E = H_r(I) + H_l(I), \tag{9}$$

where $H_r(I) = \sum_{j=1}^{N}(\frac{V_j}{V_I})H(R_j)$ denotes the expected *intra region tube entropy* as the sum of individual tube entropies (weighted by volume), and $H_l(I) = -\sum_{j=1}^{N}\frac{V_j}{V_I}\log\frac{V_j}{V_I}$ denotes the *layout entropy* which is used to penalize over-segmentation. Higher $\frac{1}{E}$ value indicates better quality. Since [10] provides a multi-resolution (coarse to fine) segmentation result, we select the layer with the most similar number of 3D regions (i.e. labels) to compare with. The values of $\frac{1}{E}$ for each method for the 15

	bend				jack				jump				pjump			
	Foreground		Background		Foreground		Background		Foreground		Background		Foreground		Background	
Methods	PR	RE	PR	RE	PR	RE	PR	RE	PR	RE	PR	RE	PR	RE	PR	RE
Mean shift (%)	62	42	71	55	63	65	66	67	56	41	63	77	73	68	70	65
CRF - SP (%)	82	73	83	85	75	80	83	84	87	67	85	90	89	80	85	78
CRF (%)	**97**	**83**	**96**	**96**	**89**	**94**	**93**	**93**	**99**	**80**	**95**	**96**	**95**	**93**	**94**	**94**

run				side				skip				walk				wave1				wave2			
61	50	67	71	71	62	72	62	75	50	70	72	65	53	71	65	61	58	68	74	64	67	71	68
90	78	88	80	80	76	77	70	87	70	87	80	86	69	80	78	81	77	80	75	85	77	82	81
99	**80**	**92**	**93**	**94**	**90**	**93**	**93**	**99**	**75**	**95**	**96**	**99**	**81**	**92**	**93**	**94**	**88**	**91**	**91**	**99**	**87**	**92**	**92**

Table 1. The precision (PR) and recall (RE) rate of foreground and background on the Weizmann Activities [8] image sequence. The first row is the activity category. Comparison is made between the mean shift segmentation algorithm and the proposed method without using spatial higher order potentials. Note that each activity has 10 different subjects. The reported rate is the average over all subjects.



(a) (b) (c) (d)

Figure 7. Comparison with [4] and [10] on the *garden* sequence. Column (a) shows two frames from the original sequence. Columns (b,c,d) show the results of the proposed method, [4], and [10], respectively. Correspondences are shown using the same color. Performance can be evaluated by checking whether corresponding parts have the same color and different parts have different colors. The *garden* sequence is considered difficult for video segmentation/tracking for two reasons: (1) four major parts (sky, house, garden and tree) are at different depths, thereby causing different motions since camera is moving, and (2) extremely textured regions within the garden produce a huge number of unstable regions which are hard to track. [4] successfully tracks the front tree but misses to track almost all other regions. [10] over-merges the regions, e.g., tree and house in the top frame, as well as the garden becomes a single region tube despite the texture variation within. Our results preserve the local texture without over-smoothing, while correctly tracking each parted. The shift of regions with the same color is highly correlated to the image motion.

videos are shown in Figure 6. Our CRF method and [10] consistently outperform mean shift and CRF−SP. While competing head-to-head against [10], CRF does better in 6 out of 15. Due to space limitation, Figure 8 shows only two qualitative segmentation results. In the *ice skate* sequence, all the major parts (*i.e.*, legs, body, hand, ground and back advertisement panel) are consistently segmented across frames. Note that from the first to the second frame, a new label for the text “OLYMPUS” is created. While in the *foroldmen* sequence, not only the person is consistently segmented, the new label corresponding to car explosion is being created. For the classic garden sequence, we also add comparison with [4]. Note that these are three different approaches: [4] is based on region tracking, [10] is graph-based grouping whereas ours is based on labeling of many frames simultaneously using a conditional random field representation. Our algorithm (column (b)) erroneously segments part of the garden in the bottom frame as tree. However, for the remaining parts, our method outperforms the other two methods: [4] does not track any of the regions except the tree, and [10] simply over-merges the regions (tree and house, all the grass area and flowers). More importantly, the shift between corresponding regions (indicated by the movement of colored regions) in our method is highly correlated with the image motion.

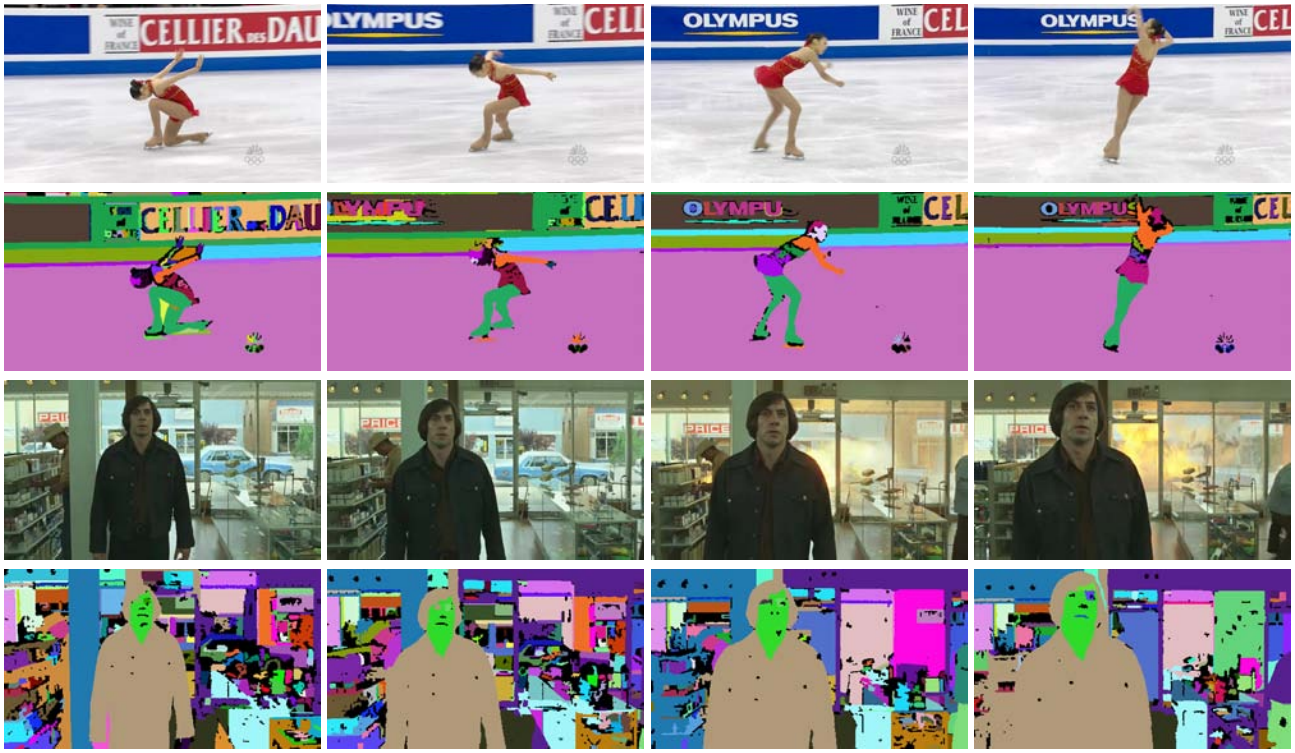


Figure 8. Video segmentation results for *ice skate* and *foroldmen* videos, shown here for qualitative evaluation based on color coded-correspondences.

6. Conclusion

In this paper we formulate video segmentation as a consistent labeling problem in which regions assigned the same label constitute a tube in the spatio-temporal space. By not making any assumption about the number of labels, the objects entering and leaving the scene are modeled by label creation and termination, respectively. To capture the relationships among regions across spatial and temporal domains, we make use of higher order structures and softly enforce label consistency.

7. Acknowledgement

This research was supported by ONR: N00014-06-1-0017 and Motorola: 239 Motorola 2010-06418.

References

[1] E. Akbas and N. Ahuja. From ramp discontinuities to segmentation tree. In *ACCV*, 2009.

[2] J. Y. A.Wang and E. H. Adelson. Representing moving images with layers. *TIP*, 1994.

[3] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *PAMI*, 2001.

[4] W. Brendel and S. Todorovic. Video object segmentation by tracking regions. In *ICCV*, 2009.

[5] T. Brox and J. Malik. Large displacement optical flow: Descriptor matching in variational motion estimation. *PAMI*, 2010.

[6] C. Erdem, B. Sankur, and A. M. Tekalp. Performance measures for video object segmentation and tracking. *TIP*, 13:947–51, 2004.

[7] C. Fowlkes, S. Belongie, F. Chung, and J. Malik. Spectral grouping using the nystrom method. *PAMI*, 26(2):214–225, 2004.

[8] L. Gorelick, M. Blank, E.Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *PAMI*, 29:2247–2253, 2007.

[9] H. Greenspan, J. Goldberger, and A. Mayer. A probabilistic framework for spatio-temporal video representation. In *ECCV*, 2002.

[10] M. Grundmann, V. Kwatra, M. Han, and I. Essa. Efficient hierarchical graph-based video segmentation. In *CVPR*, 2010.

[11] V. Hedau and N. Ahuja. Matching images under unstable segmentations. In *CVPR*, 2008.

[12] S. Khan and M. Shah. Object based segmentation of video using color, motion and spatial information. In *CVPR*, 2001.

[13] J. Kim and J. Woods. Spatiotemporal adaptive 3-d kalman filter for video. *TIP*, 1997.

[14] P. Kohli, L. Ladicky, and P. Torr. Robust higher order potentials for enforcing label consistency. *IJCV*, 2009.

[15] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labelling sequence data. In *ICML*, 2001.

[16] S. Paris. Edge-preserving smoothing and mean-shift segmentation of video streams. In *ECCV*, 2008.

[17] A. Vazquez-Reina, S. Avidan, H. Pfister, and E. Miller. Multiple hypothesis video segmentation from superpixel flows. In *ECCV*, 2010.

[18] J. Wang, B. Thiesson, Y. Xu, and M. Cohen. Image and video segmentation by anisotropic kernel mean shift. In *ECCV*, 2004.

[19] H. Zhang, J. Fritts, and S. Goldman. An entropy-based objective evaluation method for image segmentation. In *SPIE*, 2003.

[20] H. Zhang, J. Fritts, and S. Goldman. Image segmentation: a survey of unsupervised methods. In *CVIU*, 2009.

[21] C. L. Zitnick, N. Jojic, and S. B. Kang. Consistent segmentation for optical flow estimation. In *ICCV*, 2005.