Multiple Hypothesis Video Segmentation from Superpixel Flows

Amelio Vazquez-Reina[1,2], Shai Avidan[3], Hanspeter Pfister[1], and Eric Miller[2]

[1] School of Engineering and Applied Sciences, Harvard University, MA, USA
[2] Department of Computer Science, Tufts University, MA, USA
[3] Adobe Systems Inc., USA

Abstract. Multiple Hypothesis Video Segmentation (MHVS) is a method for the unsupervised photometric segmentation of video sequences. MHVS segments arbitrarily long video streams by considering only a few frames at a time, and handles the automatic creation, continuation and termination of labels with no user initialization or supervision. The process begins by generating several pre-segmentations per frame and enumerating multiple possible trajectories of pixel regions within a short time window. After assigning each trajectory a score, we let the trajectories compete with each other to segment the sequence. We determine the solution of this segmentation problem as the MAP labeling of a higher-order random field. This framework allows MHVS to achieve spatial and temporal long-range label consistency while operating in an on-line manner. We test MHVS on several videos of natural scenes with arbitrary camera and object motion.

1 Introduction

Unsupervised photometric video segmentation, namely the automatic labeling of a video based on texture, color and/or motion, is an important computer vision problem with applications in areas such as activity recognition, video analytics, summarization, surveillance and browsing [1,2]. However, despite its significance, the problem remains largely open for several reasons.

First, the unsupervised segmentation of arbitrarily long videos requires the automatic creation, continuation and termination of labels to handle the free flow of objects entering and leaving the scene. Due to occlusions, objects often merge and split in multiple 2D regions throughout a video. Such events are common when dealing with natural videos with arbitrary camera and object motion. A complete solution to the problem of multiple-object video segmentation requires tracking object fragments and handling splitting or merging events.

Second, robust unsupervised video segmentation must take into account spatial and temporal long-range relationships between pixels that can be several frames apart. Segmentation methods that track objects by propagating solutions frame-to-frame [3,4] are prone to overlook pixel relationships that span several frames.

K. Daniilidis, P. Maragos, N. Paragios (Eds.): ECCV 2010, Part V, LNCS 6315, pp. 268–281, 2010.

Fig. 1. Results from the on-line, unsupervised, photometric segmentation of a video sequence with MHVS. **Top**: original frames. **Bottom**: segmented frames. MHVS keeps track of multiple possible segmentations, collecting evidence across several frames before assigning a label to every pixel in the sequence. It also automatically creates and terminates labels depending on the scene complexity and as the video is processed.

Finally, without knowledge about the number of objects to extract from an image sequence, the problem of unsupervised video segmentation becomes strongly ill-posed [5]. Determining the optimal number of clusters is a fundamental problem in unsupervised data clustering [5].

Contributions. MHVS is, to the best of our knowledge, the first solution to the problem of fully unsupervised on-line video segmentation that can effectively handle arbitrarily long sequences, create and terminate labels as the video is processed, and still preserve the photometric consistency of the segmentation across several frames.

Although the connections between tracking and video segmentation are well discussed in *e.g.* [6,3,7,4,8], we present the first extension of the idea of deferred inference from Multiple Hypothesis Tracking (MHT) [9,10] to the problem of unsupervised, multi-label, on-line video segmentation. MHVS relies on the use of space-time segmentation hypotheses, corresponding to alternative ways of grouping pixels in the video. This allows MHVS to postpone segmentation decisions until evidence has been collected across several frames, and to therefore operate in an on-line manner while still considering pixel relationships that span multiple frames. This extension offers other important advantages. Most notably, MHVS can dynamically handle the automatic creation, continuation and termination of labels depending on the scene complexity, and as the video is processed.

We also show how higher-order conditional random fields (CRFs), which we use to solve the hypothesis competition problem, can be applied to the problem of unsupervised on-line video segmentation. Here, we address two important challenges. First, the fact that only a subset of the data is available at any time

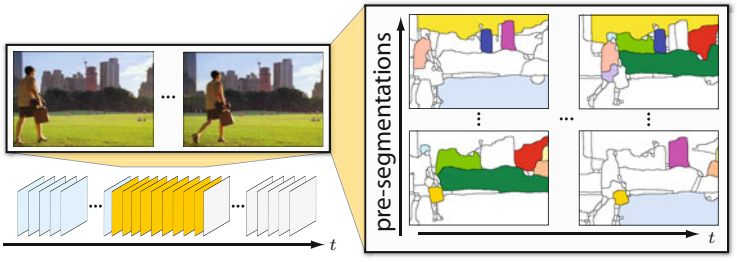

Fig. 2. Left: MHVS labels a video stream in an on-line manner considering several frames at a time. **Right**: For each processing window, MHVS generates multiple pre-segmentations per frame, and finds sequences of superpixels (shown as colored regions) that match consistently in time. Each of these sequences, called a superpixel flow, is ranked depending on its photometric consistency and considered as a possible label for segmentation. The processing windows overlap one or more frames to allow labels to propagate from one temporal window to the next.

during the processing, and second, that the labels themselves must be inferred from the data. A working example of MHVS is illustrated on Fig. 1.

Previous work. Some of the common features and limitations found in previous work on video segmentation include:

1. The requirement that all frames are available at processing time and can be segmented together [6,11,12,13]. While this assumption holds for certain applications, the segmentation of arbitrarily long video sequences requires the ability to segment and track results in a continuous, sequential manner (we refer to this as *on-line* video segmentation). Unfortunately, those methods that can segment video in an on-line manner usually track labels from frame to frame [3,7,4] (*i.e.*, they only consider two frames at a time), which makes them sensitive to segmentation errors that gradually accumulate over time.
2. The user is often required to provide graphical input in the form of scribbles, seeds, or even accurate boundary descriptions in one or multiple frames to initiate or facilitate the segmentation [14,11]. This can be helpful or even necessary for the high level grouping of segments or pixels, but we aim for an automatic method.
3. The assumption that the number of labels is known *a priori* or is constant across frames [15,16,17,18,12,14] is useful in some cases such as foreground-background video segmentation [18,12,14], but only a few methods can adaptively and dynamically determine the number of labels required to photometrically segment the video. Such ability to adjust is especially important in on-line video segmentation, since the composition of the scene tends to change over time.

Recently, Brendel and Todorovic [6] presented a method for unsupervised photometric video segmentation based on mean-shift and graph relaxation. The main difference between their work and MHVS is that our method can operate in an on-line manner and consider multiple segmentation hypotheses before segmenting the video stream.

2 An Overview of MHVS

The three main steps in MHVS are: hypotheses enumeration, hypotheses scoring, and hypotheses competition.

A *hypothesis* refers to one possible way of grouping several pixels in a video, *i.e.*, a correspondence of pixels across multiple frames. More specifically, we define a hypothesis as a grouping or flow of *superpixels*, where a superpixel refers to a contiguous region of pixels obtained from a tessellation of the image plane without overlaps or gaps. This way, each hypothesis can be viewed as a possible label that can be assigned to a group of pixels in a video (see Fig. 2).

Since different hypotheses represent alternative trajectories of superpixels, hypotheses will be said to be *incompatible* when they overlap; that is, when one or more pixels are contained in more than one hypothesis. In order to obtain a consistent labeling of the sequence, we aim for the exclusive selection of only one hypothesis for every set of overlapping hypotheses (see an example in Fig. 3).

Depending on the photometric consistency of each hypothesis, we assign them a score (a likelihood). This allows us to rank hypotheses and compare them in probabilistic terms. The problem of enumeration and scoring of hypotheses is discussed in Section 3. Once hypotheses have been enumerated and assigned a score, we make them compete with each other to label the video sequence. This competition penalizes the non-exclusive selection between hypotheses that are incompatible in the labeling. In order to resolve the hypotheses competition problem, MHVS relies on MAP estimation on a higher-order conditional random field (CRF). In this probabilistic formulation, hypotheses will be considered as labels or classes that can be assigned to superpixels on a video. Details about this step are covered in Section 4.

For the segmentation of arbitrarily long video sequences, the above process of hypotheses enumeration, scoring and competition is repeated every few frames using a sliding window. By enumerating hypotheses that include the labels from the segmentation of preceeding windows, solutions can be propagated sequentially throughout an arbitrarily long video stream.

3 Enumeration and Scoring of Hypotheses

The enumeration of hypotheses is a crucial step in MHVS. Since the number of all possible space-time hypotheses grows factorially with frame resolution and video length, this enumeration must be selective. The pruning or selective sampling of

Fig. 3. Two hypotheses that are incompatible. The hypotheses (shown in green and red) overlap on the first two frames. The segmentation of the sequence should ensure their exclusive selection. MHVS ranks hypotheses photometrically and penalizes the non-consistent selection of the most coherent ones over time.

hypotheses is a common step in the MHT literature, and it is usually solved via a "gating" procedure [19].

We address the enumeration and scoring of hypotheses in two steps. First, we generate multiple pre-segmentations for each frame within the processing window using segmentation methods from the literature, *e.g.*, [20], [21]. Then, we match the resulting segments across the sequence based on their photometric similarity. Those segments that match consistently within the sequence will be considered as hypotheses (possible labels) for segmentation.

The above approach can be modeled with a Markov chain of length equal to that of the processing window. This allows us to look at hypotheses as time sequences of superpixels that are generated by the chain, with the score of each hypothesis given by the probability of having the sequence generated by the chain.

We formalize this approach as follows. Given a window of F consecutive frames from a video stream, we build a weighted, directed acyclic graph $G = (V, E)$ that we denote as a *superpixel adjacency graph*. In this graph, a node represents a superpixel from one of the pre-segmentations on some frame within the processing window, and an edge captures the similarity between two temporally adjacent superpixels (superpixels that overlap spatially but belong to two different and consecutive frames). Edges are defined to point from a superpixel from one of the pre-segmentations on time t to a superpixel from one of the pre-segmentations on $t + 1$. Fig. 4 shows an illustration of how this graph is built.

The above graph can be thought as the transition diagram of a Markov chain of length F [22]. In this model, each frame is associated with a variable that represents the selection of one superpixel in the frame, and the transition probabilities between two variables are given by the photometric similarity between two temporally adjacent superpixels. By sampling from the chain, for example, via ancestral sampling [22] or by computing shortest paths in the transition diagram, we can generate hypotheses with strong spatio-temporal coherency.

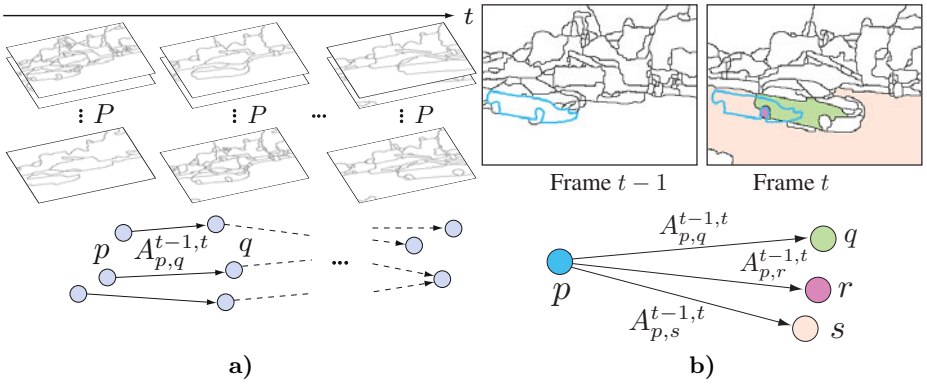

Fig. 4. Construction of the superpixel adjacency graph for the enumeration of hypotheses (flows of superpixels). (a) For each processing window, MHVS generates P pre-segmentations on each frame. Each of them groups pixels at different scales and according to different photometric criteria. The nodes in the graph represent superpixels from some of the pre-segmentations on each frame, and the edges capture the photometric similarity between two temporally adjacency superpixels. (b) Two superpixels are considered to be temporally adjacent if they overlap spatially but belong to two different and consecutive frames.

More specifically, for a given window of F frames, and the set of all superpixels $\mathcal{V} = \{V_1, \ldots, V_F\}$ generated from P pre-segmentations on each frame, we can estimate the joint distribution of a sequence of superpixels $(\mathbf{z}_1, \ldots, \mathbf{z}_F)$ as

$$p(\mathbf{z}_1, \ldots, \mathbf{z}_F) = p(\mathbf{z}_1) \cdot \prod_{t=2}^{t=F} A^{t-1,t}_{j,k}, \tag{1}$$

where the transition matrices $A^{t-1,t}_{j,k}$ capture the photometric similarity between two temporally adjacent superpixels $\mathbf{z}_{t-1} = j$ and $\mathbf{z}_t = k$, and are computed from the color difference between two superpixels in LUV colorspace, as suggested in [23]. In order to generate hypotheses that can equally start from any superpixel on the first frame, we model the marginal distribution of the node $\mathbf{z}_1$ as a uniform distribution. Further details about the generation of pre-segmentations and the sampling from the Markov chain are discussed in Section 5.

Once a set of hypotheses has been enumerated, we measure their temporal coherency using the joint distribution of the Markov chain. Given a set of L hypotheses $\mathcal{H} = \{H_1, \ldots, H_L\}$, we define the score function $s : \mathcal{H} \rightarrow [0, 1]$ as:

$$s(H_k) = N_1 \cdot p(\mathbf{z}_1 = v_1, \ldots, \mathbf{z}_F = v_F) = \prod_{t=2}^{F} A^{t-1,t}_{v_{t-1},v_t}, \tag{2}$$

where $(v_1, \ldots, v_F)$ is a sequence of superpixels comprising a hypothesis H_k and N_1 is the number of superpixels on the first frame.

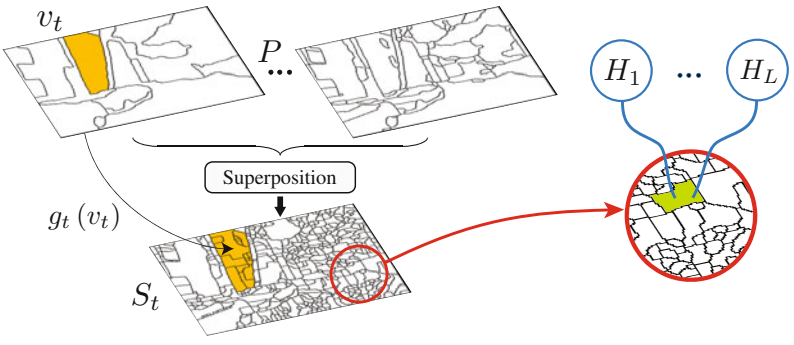

Fig. 5. We define our higher-order conditional random field on a sequence of fine grids of superpixels $\mathcal{S} = \{S_1, \dots S_F\}$. Each grid S_t is obtained as the superposition of the P tessellations that were generated for the enumeration of hypotheses. The mapping g_t takes superpixels v_t from one of the pre-segmentations to the superposition S_t. Each superpixel in S_t is represented in our CRF with a random variable that can be labeled with one of the hypotheses $\{H_1, \dots, H_L\}$.

Propagation of solutions. The above approach needs to be extended to also enumerate hypotheses that propagate the segmentation results from preceding processing windows. We address this problem by allowing our processing windows to overlap one or more frames. The overlap can be used to consider the superpixels resulting from the segmentation of each window when enumerating hypothesis in the next window. That is, the set of pre-segmented superpixels $\mathcal{V} = \{V_1, \dots, V_F\}$ in a window w, $w > 1$, is extended to include the superpixels that result from the segmentation of the window $w - 1$.

4 Hypotheses Competition

Once hypotheses have been enumerated and scored for a particular window of frames, we make them compete with each other to label the sequence. We determine the solution to this segmentation problem as the MAP labeling of a random field defined on a sequence of fine grids of superpixels. This framework allows us to look at hypotheses as labels that can be assigned to random variables, each one representing a different superpixel in the sequence (see Fig. 5).

Our objective function consists of three terms. A unary term that measures how much a superpixel within the CRF grid agrees with a given hypothesis, a binary term that encourages photometrically similar and spatially neighboring superpixels to select the same hypothesis, and a higher-order term that forces the consistent labeling of the sequence with the most photometrically coherent hypotheses over time (See Fig. 6 for an illustration).

We formalize this as follows. For each processing window of F frames, we define a random field of N variables X_i defined on a sequence of grids of superpixels $\mathcal{S} = \{S_1, \dots S_F\}$, one for each frame. Each grid S_t is obtained as the superposition of the P pre-segmentations used for the enumeration of hypotheses, and

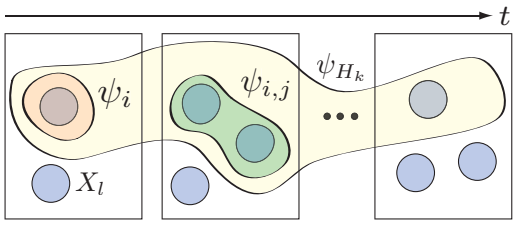

Fig. 6. The unary, pairwise and higher-order potentials, $\psi_i, \psi_{i,j}$ and ψ_{H_k}, respectively, control the statistical dependency between random variables X_l, each one representing a different superpixel within the processing window.

yields a mapping g_t that takes every superpixel from the pre-segmentations to the set S_t (see Fig. 5). The random variables X_i are associated with superpixels from $\mathcal{S}$, and take values from the label set $\mathcal{H} = \{H_1, \ldots, H_L\}$, where each hypothesis H_k is sampled from the Markov chain described in the previous section.

A sample $\mathbf{x} = (x_1, \ldots, x_N) \in \mathcal{H}^N$ from the field, *i.e.* an assignment of labels (hypotheses) to its random variables, is referred to as a *labeling*. From the Markov-Gibbs equivalence, the MAP labeling $\mathbf{x}^*$ of the random field takes the form:

$$\mathbf{x}^* = \underset{\mathbf{x} \in \mathcal{H}^N}{\arg\min} \sum_{c \in \mathcal{C}} \alpha_c \psi_c \left(\mathbf{x_c}\right), \tag{3}$$

where the potential functions ψ_c are defined on cliques of variables c from some set $\mathcal{C}$, and α_c are weighting parameters between the different potentials. The labeling $\mathbf{x}_c$ represents the assignment of the random variables X_i within the clique c to their corresponding values in $\mathbf{x}$.

We next define three different types of potentials ψ_c (representing penalties on the labeling) for our objective function in Eq. 3. The potentials enforce the consistent photometric labeling of the sequence. The unary potentials favor the selection of hypotheses that provide a high detail (fine) labeling of each frame. The pairwise potentials encourage nearby superpixels to get the same label, depending on their photometric similarity. Finally, the higher-order potentials force the exclusive selection of hypotheses that are incompatible with each other.

Unary potentials. The mappings $g = (g_1, \ldots, g_F)$ between the pre-segmentations and the grids S_t (see Fig. 5) are used to define the penalty of assigning a hypothesis x_i to the random variable X_i representing the superpixel s_i as

$$\psi_i \left(x_i\right) = 1 - d\left(s_i, g\left(x_i\right)\right), \tag{4}$$

where $g\left(x_i\right)$ represents the mapping of the superpixels within the hypothesis x_i to the set of superpixels $\mathcal{S}$. The function $d\left(a, b\right)$ measures the Dice coefficient $\in [0, 1]$ on the plane between the sets of pixels a and b (the spatial overlap between a and b), and is defined as $d\left(a, b\right) = 2|a \cap b| / \left(|a| + |b|\right)$. Since the set of superpixels $\{S_1, \ldots S_F\}$ represents an over-segmentation on each frame (it is obtained from a superposition of tessellations), the unary potential favors

labelings of the sequence with spatially thin hypotheses, *i.e.* those with the highest overlap with superpixels on the CRF grid, in the Dice-metric sense.

Pairwise potentials. We define the following potential for every pair of spatially adjacent superpixels s_i, s_j in each frame:

$$\psi_{i,j}\left(x_i, x_j\right) = \begin{cases} 0 & \text{if } x_i = x_j \\ b\left(i,j\right) & \text{otherwise,} \end{cases} \tag{5}$$

where $b\left(i,j\right)$ captures the photometric similarity between adjacent superpixels, and can be obtained by sampling from a boundary map of the image. The above potential guarantees a discontinuity-preserving labeling of the video, and penalizes label disagreement between neighboring superpixels that are photometrically similar [24]. A discussion on the choice of $b\left(i,j\right)$ is given in Section 5.

Higher-order potentials. As mentioned in Section 2, we penalize the non-exclusive selection of hypotheses that are incompatible with each other. To do this, we design a higher-order potential that favors the consistent selection of the most photometrically coherent hypotheses over time. The notion of label consistency was formalized by Kohli *et al.* in [25] and [26] with the introduction of the Robust P^n model, which they applied to the problem of supervised multi-class image segmentation. Here, we use this model to penalize label disagreement between superpixels comprising hypotheses of high photometric coherency. For each hypothesis H_k, we define the following potential:

$$\psi_{H_k}\left(\mathbf{x}_k\right) = \begin{cases} N_k\left(\mathbf{x}_k\right) \frac{1}{Q_k} s(H_k) & \text{if } N_k\left(\mathbf{x}_k\right) \leq Q_k \\ s(H_k) & \text{otherwise,} \end{cases} \tag{6}$$

where $\mathbf{x}_k$ represents the labeling of the superpixels comprising the hypothesis H_k, and $N_k\left(\mathbf{x}_k\right)$ denotes the number of variables not taking the dominant label (*i.e.*, it measures the label disagreement within the hypothesis). The score function $s\left(H_k\right)$ defined in the previous section measures the photometric coherency of the hypothesis H_k (see Eq. 2). The truncation parameter Q_k controls the rigidity of the higher-order potential [25], and we define it as:

$$Q_k = \frac{1 - s\left(H_k\right)}{\max\limits_{m \in [1,L]} \left(1 - s\left(H_m\right)\right)} \cdot \frac{|c|}{2}. \tag{7}$$

The potential ψ_{H_k} with the above truncation parameter gives higher penalties to those labelings where there is strong label disagreement between superpixels that belong to highly photometrically coherent hypotheses (the more photometrically coherent a hypothesis is, the higher the penalty for disagreement between the labels of the CRF superpixels comprising it). See Fig.7(a) for an example.

Labeling. Once we have defined unary, binary and higher-order potentials for our objective function in Eq. 3, we approximate the MAP estimate of the CRF using a graph cuts solver for the Robust P^n model [25]. This solver relies on a

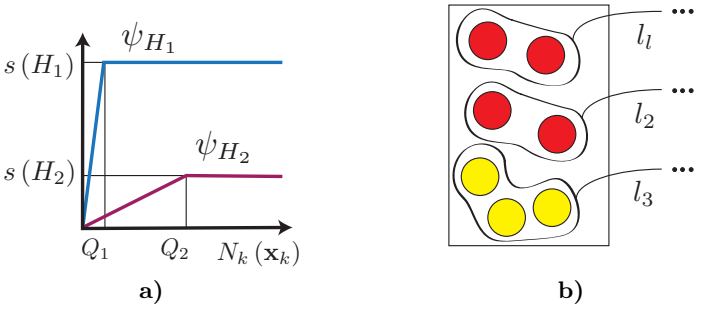

Fig. 7. (a) Higher-order penalty (y-axis) as a function of label disagreement within a hypothesis (x-axis) for two overlapping hypotheses H_1 and H_2, with H_1 being more photometrically coherent than H_2. The potential ψ_{H_1} strongly penalizes any label disagreement within H_1, while ψ_{H_2} tolerates significantly higher label disagreement within H_2. (b) The colored circles represent superpixels that were labeled in the preceding processing window (each color being a different label). The groupings l_1, l_2 and l_3 are the result of the MAP labeling within the current processing window. Depending on the selection of γ_1 and γ_2 (see text), l_1 and l_2 are considered as new labels or mapped to the label depicted in red.

sequence of alpha-expansion moves that are binary, quadratic and submodular, and therefore exactly computable in polynomial time [25]. From the association between variables X_i and the superpixels in S, this MAP estimate also yields the segmentation of all the pixels within the processing window.

Handling mergers and splits. The implicit (non-parametric) object boundary representation provided by the random field [24] allows MHVS to easily handle merging and splitting of labels over time; when an object is split, the MAP labeling of the graph yields disconnected regions that share the same label. Since labels are propagated across processing windows, when the parts come back in contact, the labeling yields a single connected region with the same label. The automatic merging of object parts that were not previously split in the video is also implicitly handled by MHVS. This merging occurs when the parts of an object are included within the same hypothesis (i.e. one of the pre-segmentations groups the parts together).

In order to create new labels for parts of old labels, when the parts become distinguishable enough over time to be tracked, a final mapping of labels is done before moving to the next processing window. We handle this scenario by comparing the spatial overlap between new labels (from the current processing window) and old labels (from the preceding processing window). We check for new labels l that significantly overlap spatially with some old label p, but barely overlap with any other old label q. We can measure such overlaps using their Dice coefficients, and we denote them by γ_p and γ_q. Then, if $\gamma_p > \gamma_1$ and $\gamma_q < \gamma_2$, $\forall q \neq p$, for a pair of fixed parameters $\gamma_1, \gamma_2 \in [0, 1]$, we map the label l to p, otherwise l is considered a new label (see Fig. 7(b) for an example).

5 Experimental Results

Most previous work on unsupervised photometric video segmentation has focused on the segmentation of sequences with relatively static backgrounds and scene complexity [4,6,12,16]. In this paper, however, we show results from applying MHVS to natural videos with arbitrary motion on outdoor scenes. Since existing datasets of manually-labeled video sequences are relatively short (often less than 30 frames), and usually contain a few number of labeled objects (often only foreground and background), we collected five videos of outdoor scenes with 100 frames each, and manually annotated an average of 25 objects per video every three frames. The videos include occlusions, objects that often enter and leave the scene, and dynamic backgrounds (see Figs. 1 and 8 for frame examples).

We compared MHVS with spatio-temporal mean-shift (an *off-line* method, similar to [13]), and pairwise graph propagation (an on-line method with frame-to-frame propagation, similar to [4]). In both methods we included color, texture and motion features. For the test with mean-shift, each video was processed in a single memory-intensive batch. For our MHVS tests, F was set to 5 frames to meet memory constraints, but values between 3 and 10 gave good results in general. The size of the processing window was also observed to balance MHVS's ability to deal with strong motion while preserving long-term label consistency. We used an overlap of one frame between processing windows and generated $P = 30$ pre-segmentations per frame using the gPb boundary detector introduced by Maire *et al.* [21], combined with the OWT-UCM algorithm from [27].

As mentioned in Section 3, hypotheses can be obtained via ancestral sampling [22] (*i.e.* sampling from the conditional multinomial distributions in the topological order of the chain), or by computing shortest paths in the transition diagram from each superpixel on the first frame to the last frame in the window (*i.e.* computing the most likely sequences that start with each value of the first variable in the chain). We follow this second approach. Neither guarantees that every CRF superpixel is visited by a hypothesis. In our implementation, such CRF superpixels opt for a dummy (void) label, and those that overlap with the next processing window are later considered as sources for hypotheses. The parameters α_e weighting the relative importance between the unary, pairwise and higher-order potentials in Eq. 3 were set to 10, 2 and 55, respectively, although similar results were obtained within a 25% deviation from these values. The pairwise difference between superpixels $b(i,j)$ was sampled from the boundary map generated by OWT-UCM and the parameters γ_1 and γ_2 that control the mapping of new labels to old labels were set to 0.8 and 0.2, respectively.

We measured the quality of the segmentations using the notion of *segmentation covering* introduced by Arbeláez *et al.* in [27]. The covering of a human segmentation S by a machine segmentation S', can be defined as:

$$C(S' \to S) = \frac{1}{N} \sum_{V \in S} |V| \cdot \max_{V' \in S'} d(V, V') \tag{8}$$

where N denotes the total number of pixels in the video, and $d(V, V')$ is the Dice coefficient in 3D between the labeled spatio-temporal volumes V and V'

Fig. 8. Top to fourth row: Results from the on-line, unsupervised, photometric segmentation of four video sequences of varying degrees of complexity with MHVS. The examples show MHVS's ability to adjust to changes in the scene, creating and terminating labels as objects enter and leave the field of view. **Fourth and fifth row:** Comparison between MHVS (fourth row) and pairwise graph propagation (similar to [4]) (fifth row). The frames displayed are separated by 5-10 frames within the original segmented sequences.

Table 1. Best segmentation covering obtained with MHVS, pairwise graph propagation and mean-shift across five outdoor sequences that were manually annotated. Frame examples from Video 1 are shown in Fig. 1, and from Videos 2 to 5 in Fig. 8, top to bottom. Higher segmentation coverings are better.

Method	Video 1	Video 2	Video 3	Video 4	Video 5
MHVS (multi-frame on-line)	**0.62**	**0.59**	**0.45**	**0.54**	0.42
Graph propagation (pairwise on-line)	0.49	0.37	0.36	0.39	0.34
Mean-shift (off-line)	0.56	0.39	0.34	0.38	**0.44**

within S and S', respectively. These volumes can possibly be made of multiple disconnected space-time regions of pixels. Table 1 shows the values of the best segmentation covering achieved by each method on our five videos.

6 Discussion and Future Work

In our tests, we observed that sometimes labels have a short lifespan. We attribute this to the fact that it is difficult to find matching superpixels in pre-segmentations of consecutive frames. The use of multiple pre-segmentations per frame was introduced to alleviate this problem, and further measures, such as the use of "track stitching" methods (*e.g.* see [28]) could help reduce label flickering in future work.

Running time. The unary, pairwise and higher-order potentials of Eq. 3 are sparse. Each random variable (representing an over-segmented superpixel) overlaps few other hypotheses. No overlap makes the unary and higher-order terms associated with the hypothesis zero. The pre-segmentations, enumeration of hypotheses and measuring of photometric similarities between superpixels can be parallelized, and each processing window must be segmented (Eq. 3 solved) before moving to the next processing window. With this, in our tests, MHVS run on the order of secs/frame using a Matlab-CPU implementation.

Conclusions. MHVS is, to the best of our knowledge, the first solution to the problem of fully unsupervised on-line video segmentation that can segment videos of arbitrary length, with unknown number of objects, and effectively manage object splits and mergers. Our framework is general and can be combined with any image segmentation method for the generation of space-time hypotheses. Alternative scoring functions, to the ones presented here, can also be used for measuring photometric coherency or similarity between superpixels.

We believe our work bridges further the gap between video segmentation and tracking. It also opens the possibility of integrating the problem of on-line video segmentation with problems in other application domains such as event recognition or on-line video editing. Future work could include extensions of MHVS based on on-line learning for dealing with full occlusions and improving overall label consistency.

Acknowledgments. This work was supported in part by the NSF Grant No. PHY-0835713. We also thank the generous support from Microsoft Research, NVIDIA, the Initiative in Innovative Computing (IIC) at Harvard University, and Jeff Lichtman from the Harvard Center for Brain Science.

References

1. Turaga, P., Veeraraghavan, A., Chellappa, R.: From videos to verbs: Mining videos for activities using a cascade of dynamical systems. In: CVPR 2007 (2007)
2. Pritch, Y., Rav-Acha, A., Peleg, S.: Nonchronological video synopsis and indexing. PAMI 30, 1971–1984 (2008)
3. Ren, X., Malik, J.: Tracking as repeated figure/ground segmentation. In: CVPR 2007 (2007)

4. Liu, S., Dong, G., Yan, C., Ong, S.: Video segmentation: Propagation, validation and aggregation of a preceding graph. In: CVPR 2008 (2008)
5. Jain, A.: Data clustering: 50 years beyond k-means. Pattern Recognition Letters (2009)
6. Brendel, W., Todorovic, S.: Video object segmentation by tracking regions. In: ICCV 2009 (2009)
7. Bugeau, A., Pérez, P.: Track and cut: simultaneous tracking and segmentation of multiple objects with graph cuts. JIVP, 1–14 (2008)
8. Yin, Z., Collins, R.: Shape constrained figure-ground segmentation and tracking. In: CVPR 2009 (2009)
9. Yilmaz, A., Javed, O., Shah, M.: Object tracking: A survey. ACM Comput. Surv. 38, 13 (2006)
10. Reid, D.B.: An algorithm for tracking multiple targets, vol. 17, pp. 1202–1211 (1978)
11. Wang, J., Xu, Y., Shum, H.Y., Cohen, M.F.: Video tooning. In: SIGGRAPH 2004 (2004)
12. Huang, Y., Liu, Q., Metaxas, D.: Video object segmentation by hypergraph cut. In: CVPR 2009 (2009)
13. De Menthon, D.: Spatio-temporal segmentation of video by hierarchical mean shift analysis. In: SMVP 2002 (2002)
14. Bai, X., Wang, J., Simons, D., Sapiro, G.: Video snapcut: robust video object cutout using localized classifiers. In: SIGGRAPH 2009 (2009)
15. Chan, A., Vasconcelos, N.: Variational layered dynamic textures. In: CVPR 2009 (2009)
16. Hedau, V., Arora, H., Ahuja, N.: Matching images under unstable segmentations. In: CVPR 2008 (2008)
17. Ayvaci, A., Soatto, S.: Motion segmentation with occlusions on the superpixel graph. In: ICCVW 2009 (2009)
18. Unger, M., Mauthner, T., Pock, T., Bischof, H.: Tracking as segmentation of spatial-temporal volumes by anisotropic weighted tv. In: Cremers, D., Boykov, Y., Blake, A., Schmidt, F.R. (eds.) EMMCVPR 2009. LNCS, vol. 5681, pp. 193–206. Springer, Heidelberg (2009)
19. Blackman, S., Popoli, R.: Design and Analysis of Modern Tracking Systems. Artech House (1999)
20. Comaniciu, D., Meer, P.: Mean shift: A robust approach toward feature space analysis. PAMI 24, 603–619 (2002)
21. Maire, M., Arbelaez, P., Fowlkes, C., Malik, J.: Using contours to detect and localize junctions in natural images. In: CVPR 2008 (2008)
22. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, Heidelberg (2007)
23. Fulkerson, B., Vedaldi, A., Soatto, S.: Class segmentation and object localization with superpixel neighborhoods. In: ICCV 2009 (2009)
24. Boykov, Y., Funka-Lea, G.: Graph cuts and efficient n-d image segmentation. IJCV 70, 109–131 (2006)
25. Kohli, P., Ladický, L., Torr, P.H.S.: Robust higher order potentials for enforcing label consistency. IJCV 82, 302–324 (2009)
26. Kohli, P., Kumar, M.P., Torr, P.H.S.: P3 & beyond: Solving energies with higher order cliques. In: CVPR 2007 (2007)
27. Arbelaez, P., Maire, M., Fowlkes, C., Malik, J.: From contours to regions: An empirical evaluation. In: CVPR 2009 (2009)
28. Ding, T., Sznaier, M., Camps, O.: Fast track matching and event detection. In: CVPR 2008 (2008)