Semi-Supervised Video Segmentation using Tree Structured Graphical Models

Ignas Budvytis, Vijay Badrinarayanan, Roberto Cipolla
Department of Engineering, University of Cambridge,
Cambridge, UK
ib255,vb292,cipolla@eng.cam.ac.uk

Abstract

We present a novel, implementation friendly and occlusion aware semi-supervised video segmentation algorithm using tree structured graphical models, which delivers pixel labels alongwith their uncertainty estimates. Our motivation to employ superivision is to tackle a task-specific segmentation problem where the semantic objects are pre-defined by the user. The video model we propose for this problem is based on a tree structured approximation of a patch based undirected mixture model, which includes a novel time-series and a soft label Random Forest classifier participating in a feedback mechanism. We demonstrate the efficacy of our model in cutting out foreground objects and multi-class segmentation problems in lengthy and complex road scene sequences. Our results have wide applicability, including harvesting labelled video data for training discriminative models, shape/pose/articulation learning and large scale statistical analysis to develop priors for video segmentation.

1. Introduction

From a Bayesian perspective, unsupervised segmentation must either tackle the issue of model selection (determining the optimal number of segments from data [3]) or marginalize over segmentation hypotheses as in non-parametric Bayesian approaches [19]. The first approach requires determination of model evidence, which for most image models is difficult, and the second requires meaningful priors over segmentations, which would need vast amounts of training data to hypothesize [19]. We choose to avoid semi-heuristic model selection methods [5] and instead tackle the problem of *task-specific segmentation*, where the semantic object labels are initialised by the user (see Fig. 1). In particular, we define our problem as labelling a video into a fixed number of semantic classes, given the labels of the first and last frames of a video sequence [1]. The resulting *pixel soft labels* can be used for harvesting labelled data for maximum likelihood (ML) learning of discriminative models [17], statistical analysis to develop image/video priors [19], and object shape/pose/articulation learning [15].
A video model must capture both short range correlations (within frame and successive frames) and long range correlations (across many frames) in the video to enable occlusion aware segmentation (see Fig. 2). In addition, it must provide a measure of uncertainty which is temporally smooth and helps avoid propagation of erroneous instantaneous decisions. Existing video models [7, 10, 21, 9] do not satisfy one or more of these requirements as discussed in Sec. 2. In contrast, the algorithm we propose is to the best of our knowledge the first of its kind to address all these requirements. Specifically, our contributions in this paper are:
1. A novel rectangular patch based *tree structured graphical model for videos* which capture both short and long range correlations for occlusion aware segmentation.
2. Label *uncertainty estimation* in videos due to *exact* probabilistic inference.
3. An *implementation friendly algorithm* which only requires exact inference and a standard Random Forest classifier based on an entropic information gain criterion [17].
An example object cut-out obtained by our method is shown in Fig. 1.
We present a detailed literature review in Sec. 2. Our proposed algorithm is elaborated in Sec. 3.1. We discuss our experimental setup and results in Sec. 4. We summarise the advantages and drawbacks of our approach in Sec. 5. We conclude in Sec. 6.

2. Literature review

Video Models - Image/video models which learn long range correlations by removing redundancy in video data are the Epitome and Jigsaw models [7], [10]. Their ability to recognize semantic video segments by discovering "semantic clusters" in these compact representations using ad-hoc clustering remains speculative for complex video data. In this paper, we avoid compacting the video and instead use a soft label Random Forest (**slRF**) borrowed from [9] for patch clustering, which is fast and is able to capture long range correlations in the video.

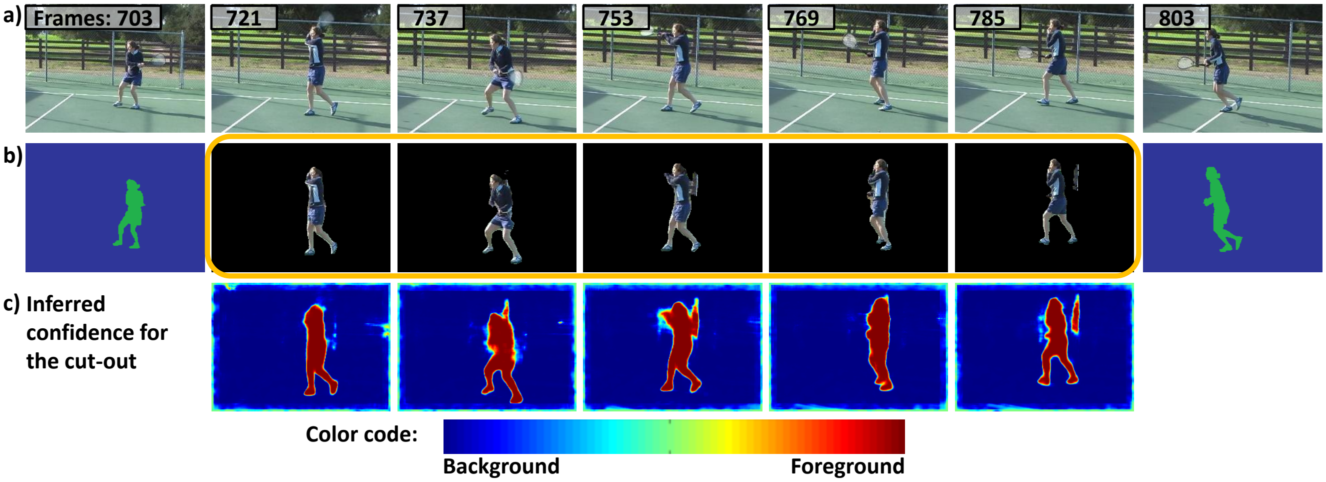


Figure 1. A clean cut-out of an *articulated object* over 100 frames using our method. We used a threshold of 0.90 over the pixel marginal posteriors to produce this cut-out. See supplementary video. Best viewed in colour and zoomed in.

The recently proposed unsupervised segmentation algorithm of Reina et al. [21] aims to link temporally consistent superpixels using a Graph-Cut optimiser. As their segmentation is *task independent*, it is difficult to interpret their over segmented results in terms of semantic object labels. Further, their unsupervised approach does not consider within class appearance variation which is what makes semantic segmentation difficult. We therefore consider the more well-defined yet non-trivial task-specific segmentation problem.

Buchanan et al. [6] store video data in a efficient manner using K-D trees for fast semi-supervised feature point tracking. In contrast to theirs, and optical flow approaches [16], we do not need pixel accurate matching. We rely on intra-class matches in video, as in inpainting problems [7].

The hybrid label propagation (HLP) model of Budvytis et al. [9] employs a directed graphical (DG) model to fuse short and long range correlations for semi-supervised segmentation. Their method suffers from the problem of "explaining away" in directed models [8], which affects the ability to capture long range correlations (see supplementary report). In addition, *multiple connectivity* in their DG restricts them to infer only two extreme states of uncertainty (delta or flat distibutions). To overcome these issues, we propose an undirected graphical (UG) model which retains the essential "inpainting" ability of their model. Our model is able to model the desired correlations and also permits exact inference.

Label Uncertainty Estimation - The energy minimisation method of Sturgess et al. [18] has shown promising results for semantic segmentation, but they require plentiful labelled data for maximum likelihood (ML) parameter learning. Further, there is no attempt to introduce temporal smoothness nor to deliver label uncertainties, both of which are key elements of our work. Indeed, there have been attempts to define label uncertainties through max-marginal probabilities within CRF models [12], but their uncertainty estimates are mis-matched to marginal posteriors. Also, their global uncertainty formulation is unsuitable to prompt temporally smooth local correction [2]. Dynamic MRF models [11] introduce temporal smoothness by using the MAP label estimates in one frame to efficiently drive the MAP computation in the next. As uncertainties of the MAP estimates are not propagated through time, this method is prone to accumulation of errors due to *instantaneous decision making*. In summary, global MAP estimation methods compromise exact inference for a complex prior (MRF) and do not capture long range correlations.

The desire to capture short and long range correlations inherently leads to a "loopy" graphical model, which does not permit exact inference. Variational inference methods such as the popular *mean-field* approximations [9] fail to propagate meaningful uncertainty information in time-series models [20]. Therefore, we avoid such approximations and instead estimate a *tree structured graphical model* for videos, which permits exact inference. This model includes a time-series to capture short-range correlations and slRF as a "black-box" which subsumes the "loopy" long range cliques (see Fig. 2). We believe that this approximation is justified by the quality of the results it produces using exact inference.

Semi-supervised Learning - The semi-supervised Random Forests of [13] make no use of soft labels. They attempt to spread deterministic labels to unlabelled data from labelled ones using a slow iterative annealing scheme designed to minimise a loss function. In contrast, the slRF seamlessly balances the functionality of the Random Forest between the extremes as a classifier (fully certain labels) and a clustering method ("flat" distributions) at almost no extra cost to the training speed of the Random Forest.

Others - Video object cut-out systems like [2] employ ad-hoc fusion of colour, motion and shape cues for interactive segmentation. Since they do not model long range correlations, frequent user input is necessary to deal with occlusions. SIFT-flow [14] based on the optic-flow style optimisation is unsuitable for uncertainty propagation.

3. Proposed Algorithm

We begin by developing our proposed graphical model for video sequences.

3.1. Tree Structured Graphical Model for Videos

We introduce a rectangular patch based undirected graphical (UG) model which is a mixture model similar in topology to the model in [9] (see Fig. 2). This UG model does not suffer from the drawbacks of "explaining away" [8], which makes fusion of the time-series and the slRF difficult (see supplementary report). Our main idea is to perform inference, train the slRF based on the inference, and alternate between these two steps for segmentation in this *feedback* setup. However, when we unravel the dimensions of our UG model, it is clear that exact inference is intractable due to its "loopy" structure (see Fig. 2). We find that training an slRF using *approximate inference* [20] can destabilise this feedback setup. Therefore, we approximate our mixture by a tree structured graphical model using variational analysis and choose it as our video model. As this model permits exact inference, the estimated soft labels for training the slRF are reliable. We also find through empirical studies that this simple model is very effective for semi-supervised segmentation. We explain the components of the UG model below.

Random Variables

1. $I_{0:n}$ are the observed sequence of images.

2. Z_k is a *latent colour image* consisting of "overlapping latent colour image patches", $Z_k = \{Z_{k,j}\}_{j=1}^{\Omega}$, where j is the patch index into the set of patches Ω. As in [7], [1] we first assume these patches (and pixels within them) to be mutually independent, even though they share coordinates, but then enforce agreement in the overlapping parts during inference by using a delta approximation in the variational posterior. This recaptures correlations between latent image patches, but at the cost of only a single point posterior.

3. Z_k^a , C_k and A_k are *latent labelled images* representing the time-series, output of a soft label Random Forest classifier (**slRF**), and their *fused output* respectively. They all consist of "overlapping latent labelled patches". Pixel i in patch j, denoted $Z_{k,j(i)}^a, C_{k,j(i)}, A_{k,j(i)}$, are *multinomial* random variables taking one of L mutually exclusive class labels. Here $j(i)$ denotes coordinate i relative to the top-left corner of patch j. Unlike Z_k, we ignore the correlations between the overlapping parts in order to permit exact inference. We instead average the pixel posteriors at each coordinate of A_k to get the output coordinate-wise distributions. This "post-inference" averaging performs effective video labelling without burdening the inference with intractability issues.

4. $T_k = \{T_{k,j}\}_{j=1}^{\Omega}$ is the set of "patch mapping" variables which *couple* the top and bottom Markov chains. An *instance* of $T_{k,j}$ maps latent image patch $Z_{k,j}$ to an observed patch $I_{k-1,T_{k,j}}$ of the same size in I_{k-1}. The same instance of $T_{k,j}$ also maps latent labelled patch $Z_{k,j}^a$ to a patch $Z_{k-1,T_{k,j}}^a$ of the same size in the image Z_{k-1}. In our experiments, each variable $T_{k,j}$ takes on 1200 instances within a 30×40 window at frame $k-1$ centered on patch j. $T_{k,j}(i)$ denotes pixel i in patch $T_{k,j}$.

Cliques

Top Markov chain cliques:

The two kinds of cliques in this chain involving real-valued images are defined below.

$$\Psi_{top,1}(Z_k, I_{k-1,T_k}; \phi) \triangleq \prod_{j=1}^{\Omega} \prod_{i \in j} \mathcal{N}\left(Z_{k,j(i)}; I_{k-1,T_{k,j}(i)}, \phi\right), \tag{1}$$

where, index j runs over all the (overlapping) latent patches $Z_k = \{Z_{k,j}\}_{j=1}^{\Omega}$. $Z_{k,j(i)}$ is pixel i inside patch j at time k. $T_{k,j}(i)$ indexes the pixel $I_{k-1,T_{k,j}(i)}$ in I_{k-1}. $\mathcal{N}(.)$ is a normalized Gaussian distribution over $Z_{k,j(i)}$, with mean $I_{k-1,T_{k,j}(i)}$ and variance ϕ set to 1.0.

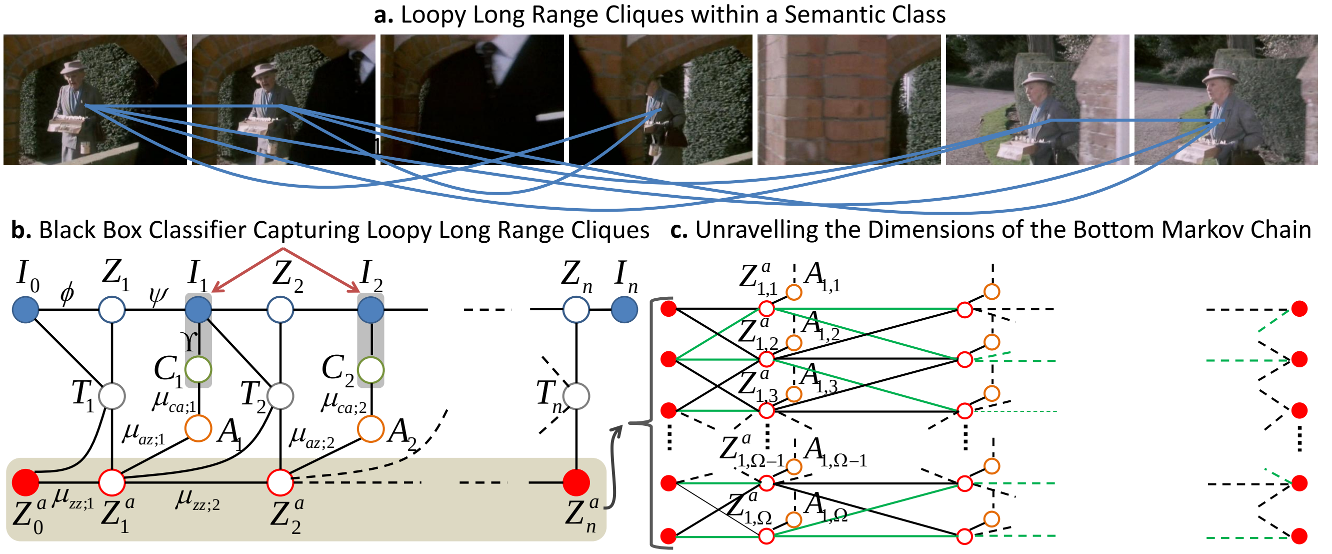
$$\Psi_{top,2}(I_k, Z_k; \psi) \triangleq \prod_{v \in V} \mathcal{N}(I_{k,v}; \frac{1}{N_v} \sum_{\substack{j=1 \\ s.t. v \in j}}^{\Omega} Z_{k,j(v)}, \psi), \tag{2}$$

where $I_{k,v}$ denotes the intensity of *global pixel coordinate* v in the image grid V. j indexes patches in Z_k and the sum is over the patches which overlap v. Note that $j(v) = j(i^{'})$, where $i^{'}$ is a local coordinate in patch j which overlaps global coordinate v. ψ is the variance of the normalized Gaussian which is set to 1.0. Note that in Eqns. 1 and 2 the R,G and B channels are treated independently.

Bottom Markov chain cliques:

$$\Psi_{bot}(Z_k^a, Z_{k-1,T_k}^a; \mu_{zz;k}) \triangleq \prod_{j=1}^{\Omega} \prod_{i \in j} \prod_{l=1}^{L} \prod_{m=1}^{L} \mu_{zz;k,j(i),T_{k,j}(i),l,m}^{\delta\left(Z_{k,j(i)}^a = l, Z_{k-1,T_{k,j}(i)}^a = m\right)}, \tag{3}$$

where the indices on the first two products are the same as in Eqn.1. The last term comprises the joint probability table $\mu_{zz;k,j(i),T_{k,j}(i)}$ for $Z_{k,j(i)}^a, Z_{k-1,T_{k,j}(i)}^a$. l, m are indices


a. Loopy Long Range Cliques within a Semantic Class

b. Black Box Classifier Capturing Loopy Long Range Cliques

c. Unravelling the Dimensions of the Bottom Markov Chain

Figure 2. Illustrates, (a) the increased burden on inference algorithms due to "loops", if both short and long range correlations need to be modelled; (b)the subsuming of loopy cliques into a BlackBox Classifier (in grey) in our mixture model; (c) an expanded view of the nodes in the bottom Markov chain of the mixture and the MAP *tree structured approximation to the mixture* in green. $A_{1:n-1}$ are the **output** nodes, see Alg. 1. See supplementary report for more details. Best viewed in colour and zoomed in.

into this table.

Black-box clique:

$$\Psi_{bb}(C_k, I_k; \Upsilon) \triangleq \prod_{j=1}^{\Omega} \prod_{i\in j} \prod_{l=1}^{L} \pi_{k,j(i),l}\left(I_{k,j(i)}; \Upsilon\right)^{\delta\left(C^{k,j(i)}=l\right)}, \tag{4}$$

where, class probabilities obey $\sum_{l=1}^{L} \pi_{k,j(i),l} = 1.0$. Υ represents the internal parameters specific to the chosen classifier, a Random Forest in our case. The tree structure and split node functions are the internal parameters ([17]). Notice that, this definition masks the internal "loopy" structure of the Random Forest as a compromise for exact inference.

Fusion cliques:

$$\Psi_{fus,1}(C_k, A_k; \mu_{ca;k}) \triangleq \prod_{j=1}^{\Omega} \prod_{i\in j} \prod_{l=1}^{L} \prod_{m=1}^{L} \mu_{ca;k,j(i),l,m}^{\delta\left(C_{k,j(i)}=l, A_{k,j(i)}=m\right)}, \tag{5}$$

where the indices on the first two products are the same as in Eqn.1. The last term comprises the joint probability table between $C_{k,j(i)}, A_{k,j(i)}$ in corresponding patches. l, m are indices into this table. Similarly,

$$\Psi_{fus,2}(A_k, Z^a_k; \mu_{az;k}) \triangleq \prod_{j=1}^{\Omega} \prod_{i\in j} \prod_{l=1}^{L} \prod_{m=1}^{L} \mu_{az;k,j(i),l,m}^{\delta\left(A_{k,j(i)}=l, Z^a_{k,j(i)}=m\right)}. \tag{6}$$

Given the random variables and cliques, the joint posterior distribution of the latent variables $H = \{Z_{1:n}, Z^a_{1:n-1}, C_{1:n-1}, A_{1:n-1}, T_{1:n}\}$, given the visible data $V = \{I_{0:n}, Z^a_0, Z^a_n\}$ and model parameter set $\Xi = \{\psi, \phi, \mu_{zz}, \mu_{az}, \mu_{ca}, \Upsilon\}$ is as follows:

$$\begin{aligned} p\left(H|V,\Xi\right) \propto \prod_{k=1}^{n} & \Psi_{top,1}(Z_k, I_{k-1,T_k}; \phi)\Psi_{top,2}(I_k, Z_k; \psi)\times \\ & \Psi_{bot}(Z^a_k, Z^a_{k-1,T_k}; \mu_{zz;k})\Psi_{bb}(C_k, I_k; \Upsilon)\times \\ & \Psi_{fus,1}(C_k, A_k; \mu_{ca;k})\Psi_{fus,2}(A_k, Z^a_k; \mu_{az;k}), \end{aligned} \tag{7}$$

where the proportionality constant is computationally intractable.

3.2. Inference

The log probability of the visible data V can be lower bounded as follows:

$$\log p(V|\Xi) \geq \int_H q(H) \log \frac{p(V,H|\Xi)}{q(H)}, \tag{8}$$

where $q(H)$ is a variational posterior. We choose,

$$q(H) = q_1(\mathcal{T})q_2(\Theta), \tag{9}$$

where $\Theta = \{Z_{1:n}, Z^a_{1:n-1}, C_{1:n-1}, A_{1:n-1}\}$, $\mathcal{T} = T_{1:n}$, and,

$$q_1(\mathcal{T}) \triangleq \prod_{k=1}^{n} \prod_{j=1}^{\Omega} q_1(\mathcal{T}_{k,j})$$

$$q_2(\Theta) \triangleq \prod_{k=1}^{n} \prod_{j=1}^{\Omega} \prod_{i\in j} \delta_{Z^*_{k,j(i)}}(Z_{k,j(i)})\tilde{q}_2(\Theta_{/Z_{1:n}}). \tag{10}$$

We then apply the calculus of variations to maximise the lower bound w.r.t q_1, q_2 and arrive at,

$$q_1(\mathcal{T}_{k,j}) \propto \exp\left\{\int_{Z_{k,j},Z^a_{k,j},Z^a_{k-1,T_{k,j}}} \tilde{q}_2(Z^a_{k,j}, Z^a_{k-1,T_{k,j}}) \times \log\left[\Psi(Z^*_{k,j}, I_{k-1,T_{k,j}}; \phi)\Psi(Z^a_{k,j}, Z^a_{k-1,T_{k,j}}; \mu_{zz;k})\right]\right\}, \quad (11)$$

$$\tilde{q}_2(\Theta_{\setminus Z_{1:n}}) = \exp\int_{\mathcal{T}} q_1(\mathcal{T}) \log p(\Theta_{/Z_{1:n}}|V, \mathcal{T}; \Xi). \quad (12)$$

The second of the above fixed point equations is still computationally intractable as it involves marginalising over all the mapping variables. For this reason we approximate it as,

$$\tilde{q}_2(\Theta_{\setminus Z_{1:n}}) \approx \exp\int_{\mathcal{T}} \delta_{\mathcal{T}^*}(\mathcal{T}) \log p(\Theta_{\setminus Z_{1:n}}|V, \mathcal{T}; \Xi),$$
$$= p(\Theta_{\setminus Z_{1:n}}|V, \mathcal{T}^*; \Xi) \quad (13)$$

where $\mathcal{T}^* = \text{argmax}_{\mathcal{T}}\, q_1(\mathcal{T})$. A second motivation for this approximation is that $p(\Theta_{\setminus Z_{1:n}}|V, \mathcal{T}^*; \Xi)$ is tree structured or in other words $\mathcal{T}^*$ represents the best (MAP) tree structured approximation of the mixture model from a variational inference viewpoint (see Fig. 2). In consequence, it is now straightforward to evaluate the exact marginals of the variables in $\Theta_{\setminus Z_{1:n}}$ and the pairwise marginals to evaluate $q_1(\mathcal{T})$.
In practice, we start by setting the $\tilde{q}_2(Z^a_{k,j}, Z^a_{k-1,T_{k,j}})$ to uniform and $Z^*_{1:n} = I_{1:n}$ to evaluate $q_1(\mathcal{T})$ (Eqn. 11(a)). With this initialisation, this step is similar to *patch cross-correlation*. Although simple to implement this step is computationally demanding (Fig. 7) and therefore, we only evaluate $q_1(\mathcal{T})$ once for each sequence to derive the corresponding tree structured model for that video.

3.3. Parameter updates and slRF training

The tree model parameters $\mu_{zz}, \mu_{az}, \mu_{ca}$ are updated in the standard maximum likelihood (ML) style using the inferred pairwise marginals [3]. Optimising the lower bound in Eqn. 8 w.r.t Υ we get the following ML update equation;

$$\hat{\Upsilon} = \underset{\Upsilon}{\text{argmax}} \sum_{k=1}^{n-1} \sum_{j=1:\Omega} \sum_{i \in j} \sum_{l=1}^{L} \tilde{q}_2(C_{k,j(i),l}) \times \log \pi_{k,j(i),l}\left(I_{k,j(i)}; \Upsilon\right)^{\delta\left(C^{k,j(i)}=l\right)}. \quad (14)$$

Updating Υ is simply equivalent to minimising the KL-divergence between the inferred soft label (marginal posterior of $\tilde{q}_2(C_{k,j(i)})$ and the predicted soft label (prior $\pi_{k,j(i)}$). Therefore, we approximate this parameter update step as a training of the slRF using soft-labels. In practice, we adopt the "information gain" evaluation criterion of [17] to train our slRF, as it is directly suited to training by taking into account the entropy of soft labels. We summarise the discussions of this section in a psuedo-code shown in Algorithm 1.

```
Algorithm 1: Semi-supervised Video Segmentation
  Input: $I_{0:n}$ (video),$Z^a_0$, $Z^a_n$ (hand labelled end frames).
  Output: Pixel label probabilities.
  Intialisation
  Set the initial values of $\mu_{zz}, \mu_{az}, \mu_{ca}, \psi, \phi$ to the values
  given in Sec. 4.
  Set $\pi_{k,j(i),l} = \frac{1}{L}$ , $l = 1 : L$ and $\forall k = 1 : n-1$, which is
  equivalent to an untrained slRF.

  Building the tree model
  Compute the MAP tree structured approximation to the
  mixture model by evaluating Eqn. 11 (Sec. 3.2) .

  Segmentation
  1. Infer marginals of $C_{1:n-1}$ [3].
  /* See S1 in Fig.3.3 & Sec.  3.2.        */
  2. Update $\Upsilon$ by learning a soft label Random Forest (slRF)
  using the marginals of $C_{1:n-1}$ as soft pixel labels .
  /* See S2 in Fig.3.3 & Sec.  3.3.        */
  3. Infer marginals of $A_{1:n-1}$ using the updated $\Upsilon$.
  /* See S3 in Fig.3.3 & Sec.  3.2.        */
  Compute probability of pixel $v$ taking label $l$ at frame $k$ as
  $\frac{1}{N_v}\sum_{j=1, s.t. j \ni v}^{\Omega} A_{k,j(v),l}$
  /* An example of typical results at each
     step is in a supplementary report to
     encourage repeatability.              */
```

4. Experiments and Results

Our colour video sequences are 320×240 resolutions. Our test sequences are taken from the CamVid road scene dataset [4] and Berkeley Motion Segmentation (BMS) dataset [5]. For qualitative studies, we use the tennis and Miss Marple sequences from BMS. We use the CamVid dataset for a quantitative study. Each sequence in CamVid is 750 frames in length, but we down sample to every 5^{th} frame to have a length of 150 frames. Ground-truth is available every 30 frames. We study 9 static classes like sky, road, etc. and treat cars, pedestrians as outliers as they are not permanent in a road scene. We assign a "flat" distribution to these outlier classes in the start and end frames and examine their false positive rate to gain insight into outlier rejection performance.
Each channel in all the images are scaled to lie between $[0.0, 1.0]$. We use patches of size is set to 7×7 and patches overlap except for 1 pixel.
In our tree model, we set the entries in the joint probability tables $\mu_{zz}, \mu_{az}, \mu_{ca}$ to 0.9 along the diagonals and equal values along the non-diagonal elements such that the sum of all entries is unity. We choose the 1^{st} stage Random Forest (RF) classifier, as in [17], with 16 trees, each of depth 10. Input LAB patches of 21×21 are extracted around every 5^{th} pixel on both axis. We leave out border pixels in a 12 pixel

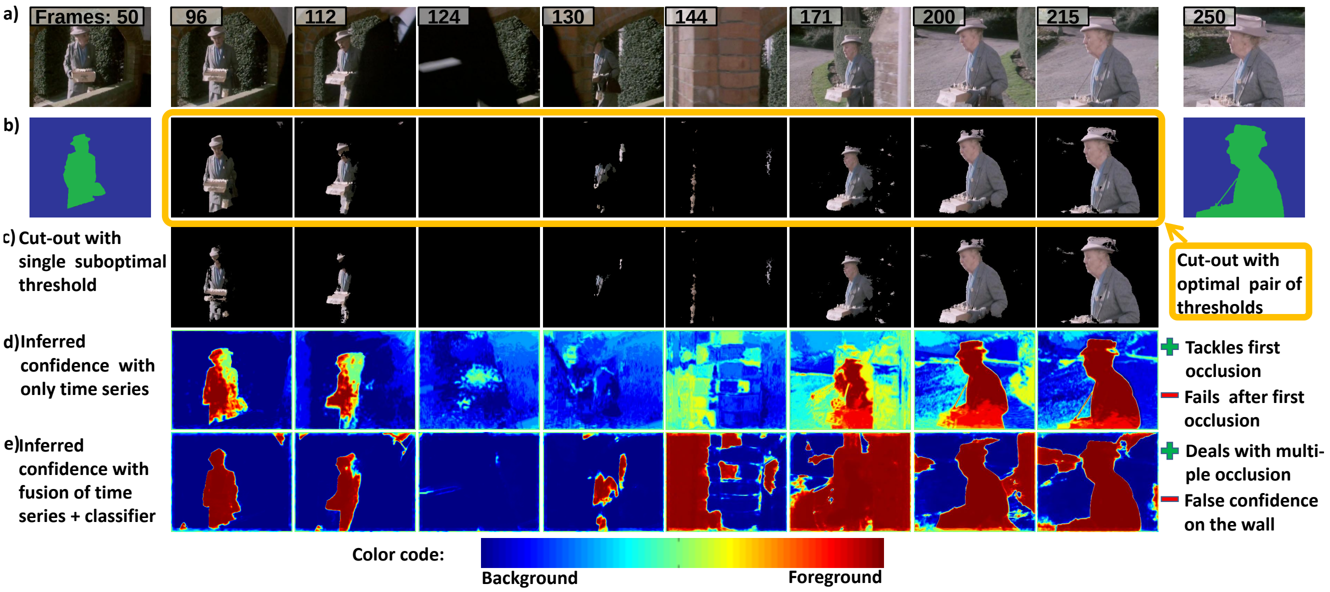

a)
Frames: 50
96
112
124
130
144
171
200
215
250
b)
c) Cut-out with single suboptimal threshold
Cut-out with optimal pair of thresholds
d) Inferred confidence with only time series
Tackles first occlusion
Fails after first occlusion
e) Inferred confidence with fusion of time series + classifier
Deals with multiple occlusion
False confidence on the wall
Color code:
Background
Foreground

Figure 3. (a) Miss Marple sequence from the Berkeley Dataset [5] with multiple complete occlusions. The occluding wall has similar colour composition as the skin colour, but our method handles this difficult occlusion effectively as seen from the cut-out in row (b). We obtained this *optimal cut-out* by thresholding the marginal posteriors with a value of 0.97 (frames $51 - 130$), and 0.99 (frames $131 - 250$). This pair of thresholds could only be obtained due to the uncertainties in (e). In comparison row (c) shows a poorer cut-out using a sole threshold of 0.99. (d) are the uncertainties without including the slRF, and in which inter-class separation is clear (frame 171). This clarity is sometimes lost in row (e) after fusing the predictions of the slRF. Best viewed in colour and zoomed in. See supplementary video.

band to fit all rectangular patches. We use the same kind and number of features as in [17]. The key difference is that we use the inferred *soft labels* to train the slRF. We compute the split function information gain and the leaf node distributions (normalized histograms) by treating the data point label as a *vector* whose elements sum to unity.

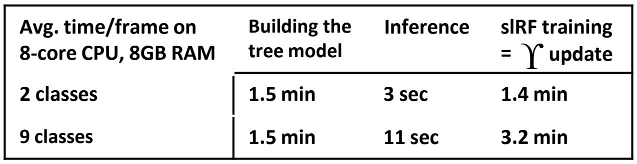
As the tree model expands from root to leaf from instance 1 to n (see Fig. 2) the model is unsymmetric in time, which can result in unnecessary biases in the labels. We rectify this by repeating the segmentation for a time-reversed video and perform pixel-wise averaging of the two inferences to get our final results as suggested in [9].

Figs. 3, 4 are our results on the foreground/background problem and the multi-class problem. For the convenience of the reader, we have provided the highlights of these results along side the images. We report our quantitative studies in Table 5 and study ROC curves for the sequence in Fig. 4, both underlined by appropriate comments to ease understanding. We also present the typical computational load for our method in Fig. 7.

5. Advantages and Drawbacks

The key **advantages** of our proposed approach are:

1. Using exact inference we avoid sequential propagation of erroneous instantaneous decisions and therefore reduce false positives.



Avg. time/frame on 8-core CPU, 8GB RAM	Building the tree model	Inference	slRF training = Υ update
2 classes	1.5 min	3 sec	1.4 min
9 classes	1.5 min	11 sec	3.2 min

Figure 7. Typical computational load of our method with an unoptimised C++ program for 320×240 sized images. We have assumed the arrays holding the marginals, incoming messages, and model parameters have been loaded into RAM in order to generate these numbers.

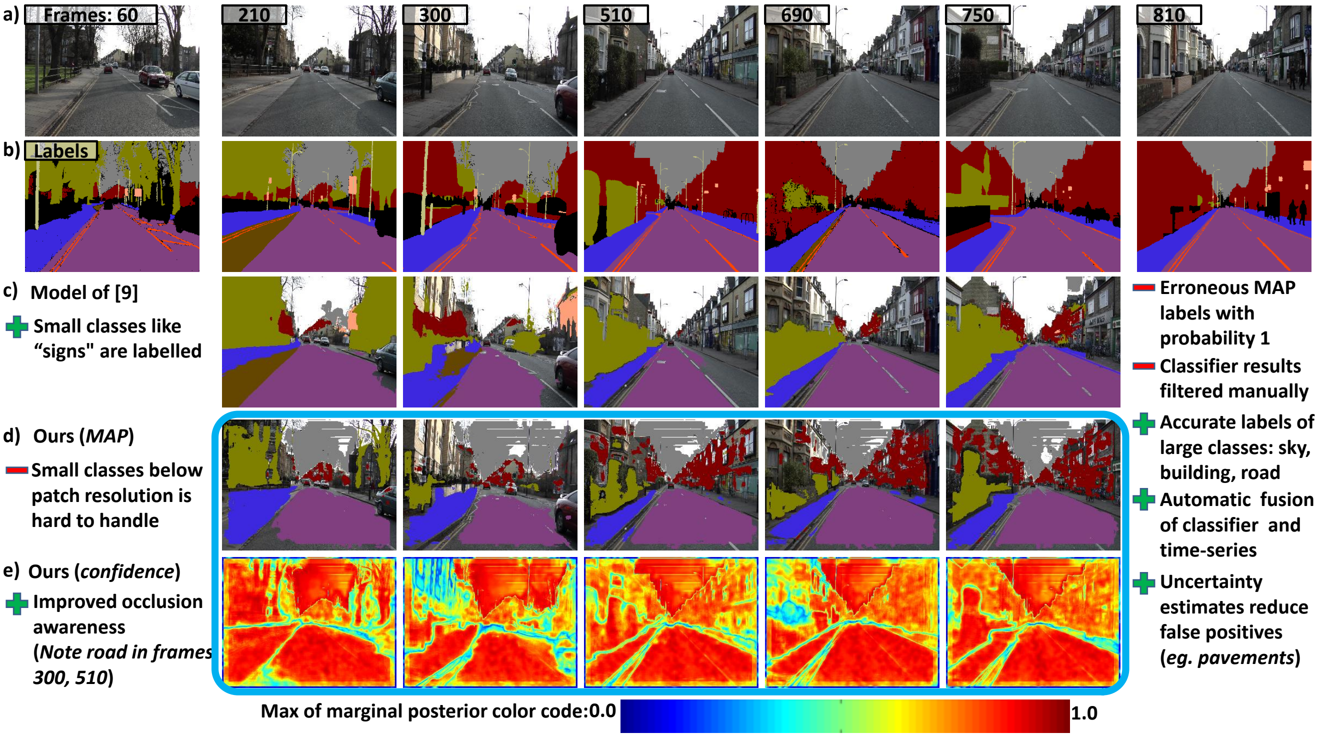
2. We avoid the common and unreliable demands of "generalization" from a classifier, and only use it as a method to setup long range correlations within its training data obtained by inference. Therefore, the classifier operates only in the "closed-world" of a video.

3. Inference and training on our model is *implementation friendly* and free from hacks.

4. The inferred uncertainties lead to better object cut-outs by leading to *optimal thresholds* for local time segments of the video (see Fig. 3).

Our approach suffers from the following **drawbacks**:

1. The uncertainty in the marginal posteriors is based on the number of pairwise cliques a patch is part of (its neigh-



✚ Small classes like "signs" are labelled

▬ Small classes below patch resolution is hard to handle

✚ Improved occlusion awareness (*Note road in frames 300, 510*)

▬ Erroneous MAP labels with probability 1

▬ Classifier results filtered manually

✚ Accurate labels of large classes: sky, building, road

✚ Automatic fusion of classifier and time-series

✚ Uncertainty estimates reduce false positives (*eg. pavements*)

Figure 4. Seq05VD from the CamVid dataset with ground-truth [4]. Black "outlier" labels at the ends have uniform distributions. The labels in row (d) were obtained by thresholding the marginal posteriors at a value of 0.75, selected using the ROC curves in Fig. 3.3. We encourage the reader to view the labels along with the confidence map in row (f) to see that our approach reduces false positive labelling. Best viewed in colour and zoomed in. See video in supplementary material.

Settings			Class accuracies for static classes									All static classes (ASC)			Large static classes(LSC)			Small static classes (SSC)		Outlier classes
Sequence	Frames	Model	Sky	Building	Sign	Pole	Road marking	Road	Pavement	Tree	Concrete	Global class acc	Average class acc	Label density	Global class acc	Average class acc	Label density	True positives + Uncertain	False positives	Uncertain (void)
1	60 –	HLP	75	26	27	0	7	93	91	97	78	79	55	53	82	77	51	57	44	87
	810	Ours	100	77	0	0	0	94	99	94	0	90	52	53	92	78	52	69	31	86
2	2310–	HLP	99	87	77	30	65	88	78	35	-	83	70	94	88	80	82	62	39	2
	3060	Ours	94	94	16	3	66	88	88	59	-	84	63	90	90	85	80	39	61	0
3	3060–	HLP	98	92	16	12	37	93	85	9	-	90	55	45	92	76	44	62	38	38
	3810	Ours	100	99	0	0	6	93	93	0	-	89	49	47	90	77	46	75	25	39

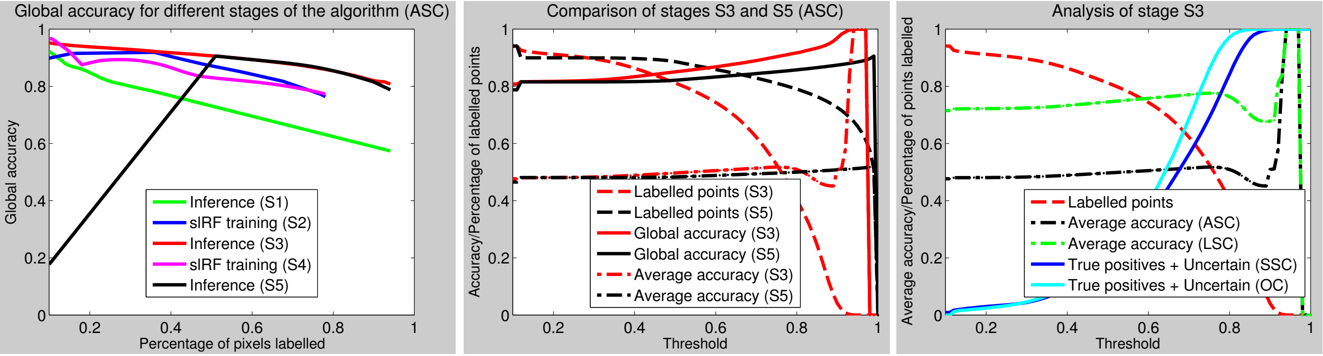
For similar label density as HLP;
✚ better LSC accuracies,
✚ comparable density of uncertain labels for outlier classes,
▬ lower accuracy over SSC due to low image resolution,
✚ reduced false positive rate by remaining uncertain,
✚ **no manual filtering of classifier output** as in HLP **[9].**

Figure 5. Quantitative comparison on complex and lengthy (750 frames) video sequences from CamVid [4] dataset. Unlike our method, HLP [9] uses manual classifier monitoring. We used ROC curves (Fig. 3.3) to get optimal thresholds of 0.75,0.12,0.77 for the three videos.

bourhood connectivity) and does include the uncertainty with which the clique was formed in the tree model. As part of future work, we would like include this information in the model to improve performance.
2. We are currently restricted to segment classes which have sizes above the patch resolution of 7×7. Using higher resolution images should alleviate this problem to a large extent.

6. Conclusions

We presented a novel tree structured graphical model for videos for semi-supervised segmentation. Unlike traditional global MAP inference, our patch based video model permits exact inference of pixel marginal posteriors within an implementation friendly setup. Using simple patch cross-correlation to model temporal correlations among



- □ S3 obtained using S2 (sIRF training) is best
- □ sIRF (S4) trained after S3 overfits. This does not improve performance (S5)
- □ Improved accuracy going from S1 to S3

- □ S5 has higher average accuracy due to overfitting , but has lower global accuracy.
- □ We stop at S3 in our experiments to avoid overfits, obtain high global accuracy and label density.

- □ At threshold 0.75, used in Fig. 4, we balance label density, avoid false positives over small classes and outlier classes.

Figure 6. Stages S1 to S3 correspond to our algorithm in Alg. 1. We update model parameters $\mu_{zz}, \mu_{az}, \mu_{ca}$ using S3 results. S4 corresponds to re-training the sIRF (equivalently Υ update) using S3. S5 corresponds to inference using all the updated parameters and demonstrates model overfitting. In the first plot, the curves fall short of 1.0 label density as we do not count (1) outlier labels in computing accuracies and (2) leave-out border pixels in sIRF predictions. Best viewed in colour and zoomed in.

pixel labels and patch-clustering to model long range label correlations, we have demonstrated that our video model can produce effective soft labels for a wide variety of applications, including object cut-outs and road scene learning. Quantitative tests demonstrate the efficiency of our approach for multi-class segmentation of segmenting lengthy and complex video sequences involving frequent occlusions. Another novelty is that the uncertainty information can be used extract better object cut-outs in complex videos.

References

[1] V. Badrinarayanan, F. Galasso, and R. Cipolla. Label propagation in video sequences. In *CVPR*, 2010. 1, 3

[2] X. Bai, J. Wang, D. Simons, and G. Sapiro. Video snapcut: robust video object cutout using localized classifiers. In *ACM SIGGRAPH*, pages 70:1–70:11, 2009. 2, 3

[3] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006. 1, 5

[4] G. Brostow, J. Fauqueur, and R. Cipolla. Semantic object classes in video: A high-definition ground truth database. *PRL*, 30(2):88–97, 2009. 5, 7

[5] T. Brox and J. Malik. Object segmentation by long term analysis of point trajectories. In *ECCV*, 2010. 1, 5, 6

[6] A. Buchanan and A. Fitzgibbon. Interactive feature tracking using k-d trees and dynamic programming. In *CVPR*, 2006. 2

[7] V. Cheung, B. J. Frey, and N. Jojic. Video epitomes. In *CVPR*, 2005. 1, 2, 3

[8] G. E. Hinton. Learning to represent visual input. *Philosphical Transactions of the Royal Society, B.*, 365:177–184, 2010. 2, 3

[9] I.Budvytis, V. Badrinarayanan, and R.Cipolla. Label propagation in complex video sequences using semi-supervised learning. In *BMVC*, 2010. 1, 2, 3, 6, 7

[10] A. Kannan, J. Winn, and C. Rother. Clustering appearance and shape by learning jigsaws. In *NIPS, Volume 19.*, 2006. 1

[11] P. Kohli and P. Torr. Efficiently solving dynamic markov random fields using graph cuts. In *ICCV*, pages II: 922–929, 2005. 2

[12] P. Kohli and P. H. S. Torr. Measuring uncertainty in graph cut solutions - efficiently computing min-marginal energies using dynamic graph cuts. In *ECCV*, pages 30–43, 2006. 2

[13] C. Leistner, A. Saffari, J. Santner, and H. Bischof. Semi-supervised random forests. In *ICCV*, 2009. 2

[14] C. Liu, J. Yuen, and A. Torralba. Nonparametric scene parsing: Label transfer via dense scene alignment. In *CVPR*, 2009. 3

[15] R. Navaratnam, A. W. Fitzgibbon, and R. Cipolla. The joint manifold model for semi-supervised multi-valued regression. In *ICCV*, 2007. 1

[16] P. Sand and S. Teller. Particle video: Long-range motion estimation using point trajectories. In *CVPR*, 2006. 2

[17] J. Shotton, M. Johnson, and R. Cipolla. Semantic texton forests for image categorization and segmentation. In *CVPR*, 2008. 1, 4, 5, 6

[18] P. Sturgess, K. Alahari, L. Ladicky, and P. H.S.Torr. Combining appearance and structure from motion features for road scene understanding. In *BMVC*, 2009. 2

[19] E. B. Sudderth and M. I. Jordan. Shared segmentation of natural scenes using dependent pitman-yor processes. In *NIPS*, pages 1585–1592, 2008. 1

[20] R. E. Turner, P. Berkes, and M. Sahani. Two problems with variational expectation maximisation for time-series models. In *Workshop on Inference and Estimation in Probabilistic Time-Series Models*, 2008. 2, 3

[21] A. Vazquez-Reina, S. Avidan, H. Pfister, and E. Miller. Multiple hypothesis video segmentation from superpixel flows. In *ECCV*, 2010. 1, 2