Int J Comput Vis
DOI 10.1007/s11263-013-0673-5

Mixture of Trees Probabilistic Graphical Model for Video Segmentation

Vijay Badrinarayanan · Ignas Budvytis · Roberto Cipolla

Received: 31 January 2013 / Accepted: 31 October 2013

Abstract We present a novel mixture of trees probabilistic graphical model for semi-supervised video segmentation. Each component in this mixture represents a tree structured temporal linkage between super-pixels from the first to the last frame of a video sequence. We provide a variational inference scheme for this model to estimate super-pixel labels, their corresponding confidences, as well as the confidences in the temporal linkages. Our algorithm performs inference over full video volume which helps to avoid erroneous label propagation caused by using short time-window processing. In addition, our proposed inference scheme is very efficient both in terms of computational speed and use of RAM and so can be applied in real-time video segmentation scenarios. We bring out the pros and cons of our approach using extensive quantitative comparisons on challenging binary and multi-class video segmentation datasets.

Keywords Video Segmentation · Semi-supervised learning · Mixture of trees probabilistic graphical model · Structured variational inference

Electronic supplementary material The online version of this article (doi:10.1007/s11263-013-0673-5) contains supplementary material, which is available to authorized users.

V. Badrinarayanan (✉) · I. Budvytis · R. Cipolla
Department of Engineering, University of Cambridge, Cambridge, UK
e-mail: vb292@cam.ac.uk

I. Budvytis
e-mail: ib255@cam.ac.uk

R. Cipolla
e-mail: rc10001@cam.ac.uk

1 Introduction

Modelling frame to frame correlations is one of the most important components in a video model. These correlations help propagate semantic labels through the video sequence for joint tracking and segmentation approaches. The standard approach is to use frame to frame optic flow (Fathi et al. 2011; Grundmann et al. 2010; Lee et al. 2003) to build the temporal structure of the video. Some also use long term point trajectories (Brox and Malik 2010; Lezama et al. 2011) to build a sparse temporal structure.

It is well recognised that the use of optical flow is inefficient for temporal propagation of semantic labels (Chuang et al. 2002; Chen and Corso 2010; Badrinarayanan et al. 2010) due to ineffective occlusion handling and label drift caused by round-off errors. To some extent these problems can be overcome by using long term point trajectories, but robust trajectories are sparse and often an additional grouping step is required for segmentation (Lezama et al. 2011; Brox and Malik 2010). These problems combined with costly multi-label MAP inference in video volumes has led to the use of short overlapping time window based segmentation methods (Tsai et al. 2010). To address these issues, we have developed a new super-pixel based mixture of trees (MoT) video model. Our model alleviates the need to use short time window processing and can deal with occlusions effectively. It requires no external optic flow computation, and instead, infers the temporal correlation from the video data automatically. We provide an efficient structured variational inference scheme for our model, which estimates super-pixel labels and their confidences. Furthermore, the uncertainties in the temporal correlations are also inferred (which reduces label drift), unlike the joint label and motion optimisation method of Tsai et al. (2010) where only a MAP estimate is obtained. Our work is partly motivated by the segmentation frame-

Published online: 13 December 2013

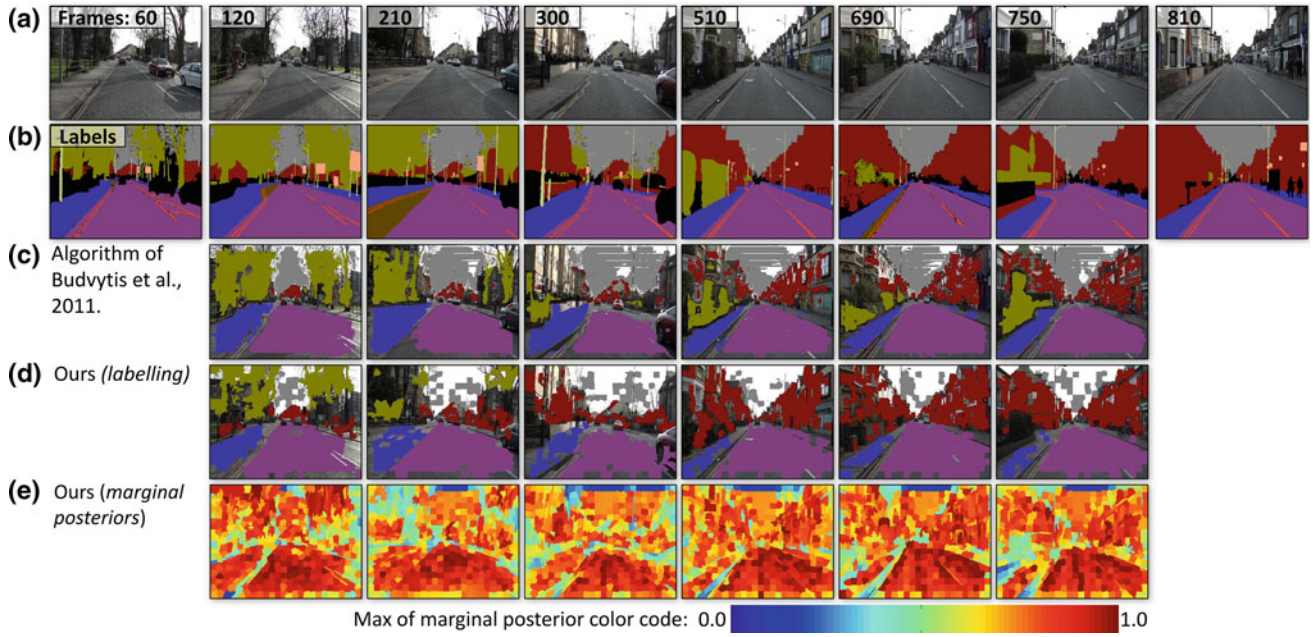

Fig. 1 An illustration of multi-class semi-supervised video segmentation results on a challenging sequence from CamVid (Brostow et al. 2009) dataset. (**a**) shows some sample frames and (**b**) the corresponding ground truth labels. Our proposed algorithm achieves similar accuracy (see Table 2) to the algorithm of Budvytis et al. (2011), yet is about two orders of magnitude faster and requires two orders of magnitude less RAM. Note that pixels which are uncertain are replaced by their original RGB values in (**c, d**). The marginal posteriors of super-pixels are shown in (**e**) (Color figure online)

work presented in Budvytis et al. (2011), Badrinarayanan et al. (2013) which relies on a tree structured video time-series model for semi-supervised video segmentation. The key difference with our proposed algorithm lies in the use of super-pixels to significantly reduce the number of latent variables in the model and the incorporation of a mixture of tree structured temporal components during inference. Figure 1 shows a sample result of multi-class segmentation using our proposed MoT video model and variational inference scheme. A shorter version of this work appeared in Budvytis et al. (2012).

The pixel label posteriors (confidences) we infer can be used to train a Random Forest classifier in a semi-supervised setting as in Budvytis et al. (2011). The predictions from this classifier (super-pixel unaries) can be fused with the MoT time-series to improve the segmentation quality in some sequences. The pixel posteriors also provide an opportunity to perform active learning for video segmentation (Settles 2012).

To summarise, the contributions we make in this paper are as follows:

1. A new MoT video sequence probabilistic graphical model.
2. An efficient structured variational inference strategy for obtaining super-pixel labels and their confidences for multi-class semi-supervised segmentation problems.

The remainder of this paper is organised as follows. We present a comparative literature review in Sect. 2. Our proposed video model is explained in detail in Sect. 3. The inference strategy for segmentation is then elaborated in Sect. 4. We then discuss our experimental results on both binary and multi-class problems using several quantitative and qualitative comparisons in Sect. 5. We discuss the computational requirements of our algorithm in Sect. 6 and bring out the advantages and the drawbacks of our approach in Sect. 7. This is followed by comments in Sect. 8. We conclude in Sect. 9 with pointers to future work.

2 Literature Review

We can broadly divide video segmentation approaches into the following three categories.

Unsupervised segmentation: In recent times, unsupervised video segmentation has gained a lot of attention (Vazquez-Reina et al. 2010; Lezama et al. 2011; Lee et al. 2003; Grundmann et al. 2010; Xu et al. 2012; Cheng and Ahuja 2012) especially as extensions of image super-pixellization to space–time super-pixels. The aim of these methods is to group pixels which are photometrically and motion-wise consistent. In simple cases, where there is a clear distinction between foreground and the background, the grouping may appear to be semantically meaningful. However, in more

complex videos, the result in general is an over-segmentation, and requires additional knowledge (through user interaction for example) to achieve any object level segmentation. In contrast, in this work, we develop a probabilistic framework which jointly models both appearance and semantic labels with a view to perform semi-supervised video segmentation. A second distinction of our algorithm is that it performs probabilistic inference, as opposed to the more commonly used MAP inference. We demonstrate that inference of labels and their marginal posteriors enables bootstrapped semi-supervised learning of classifiers which facilitates segmentation of long and complex sequences.

Other unsupervised techniques are Video Epitomes (Cheung et al. 2005) and Image Jigsaws (Kannan et al. 2006). The common factor underlying these latent variable models is the idea of removing redundancy in a set of images by discovering a compact latent representation (semantic clusters, Epitomes, Jigsaws). For a video sequence, these models can learn correlations between pixels in non-successive frames via the latent representation. However, there is a model selection step (number of clusters, size of Epitomes or Jigsaws) which is usually handcrafted. The main drawback however is the computational complexity in learning these representations. In this work, we train a Random Forest (Breiman 2001) in a semi-supervised setting to establish correlations between non-successive video frames.

Semi-supervised segmentation: The label propagation method of Badrinarayanan et al. (2010) jointly models appearance and semantic labels using a coupled-HMM model. The key idea is to influence the learning of frame to frame patch correlations as a function of both appearance and class labels. This method was extended to include correlations between non-successive frames using a Decision Forest classifier by Budvytis et al. (2011) and Badrinarayanan et al. (2013). In this work, we follow these in jointly modelling appearance and semantic labels. The main difference is that, while these methods employ a patch based tree structured graphical model, we use a super-pixel based mixture of temporal trees. This mixture importantly models the uncertainty in establishing temporal correlation between frames.

Tsai et al. (2010) jointly optimize for temporal motion and semantic labels in an energy minimization framework. In this interesting framework, they use a sliding window approach to process overlapping n-frame grids for the sake of reducing computational burden. The result of one n-frame grid is used as a hard constraint in the next grid and so on. In contrast, we treat the whole video volume at once, inferring both temporal correlations and label uncertainties. Fathi et al. (2011) use semi-supervised and active learning for video segmentation. Each unlabelled pixel is provided a confidence measure based on its distance in a neighbourhood graph to a labelled point. These confidences are used to recommend frames in which more interaction is desired. In our approach, inference directly leads to confidences and active learning can be pursued.

Other recent research has focussed on active frame selection for label propagation in video sequences (Vijayanarasimhan and Grauman 2012). This is another important issue for label propagation in videos, however, this is beyond the scope of this paper. The work of Nagaraja et al. (2012) use local classifiers learnt using the user provided labelled data to correct for errors introduced in optic flow based frame to frame label propagation. In our algorithm, we explicitly model and infer temporal linkage between frames to avoid using erroneous optic flow. The classifier in our algorithm is learnt using all of the video data, as opposed to just the user labelled key frame. Wang and Collomosse (2012) attempt to tackle the problem of erroneous optic flow by propagating labels based on a probabilistic motion model which allows for a distribution over pixel motion vectors. However, they address the problem of streaming label propagation as opposed to considering the full video volume as in our work.

Segmentation by classification: Decision tree architectures such as the popular Randomized Decision Forests (Breiman 2001) have gained popularity in *unstructured classification* tasks. The semantic Texton forests (STF) (Shotton et al. 2008) is an example, where simple pixel intensity comparisons within a patch can be used to transform the image into a descriptor space (such as textons obtained by filtering an image and classifying the filter output descriptors). The authors show that *textonising* an image (Shotton et al. 2006) in this efficient manner can also aid in segmentation when the forest is trained with annotated images. In our MoT model, we employ a decision forest classifier to implicitly establish correlations between temporally non-successive frames (see Fig. 2).

In recent years, *structured classification* models such as conditional random fields (CRF) (Boykov and Jolly 2001) have led the way in image segmentation problems, particularly for foreground object cut-out by querying the user for some supervisory strokes or a bounding box around the object of interest. In practice, their main attraction arises from the ability to perform global optimisation or in finding a strong local minima of a particular class (sub-modular class) of CRF's at interactive speeds (Boykov and Jolly 2001; Boykov et al. 1999). There are one or two notable instances which have tried to extend their image segmentation algorithms directly for videos, either by propagating MAP estimates sequentially (sub-optimal) (Kohli and Torr 2005) or performing MAP inference on N-D sequences (Boykov and Jolly 2001). Another interesting approach for N-D segmentation is the Geodesic Segmentation algorithm (Criminisi et al. 2010) which parametrises the set of possible segmentations and picks the MAP estimate within this limited set

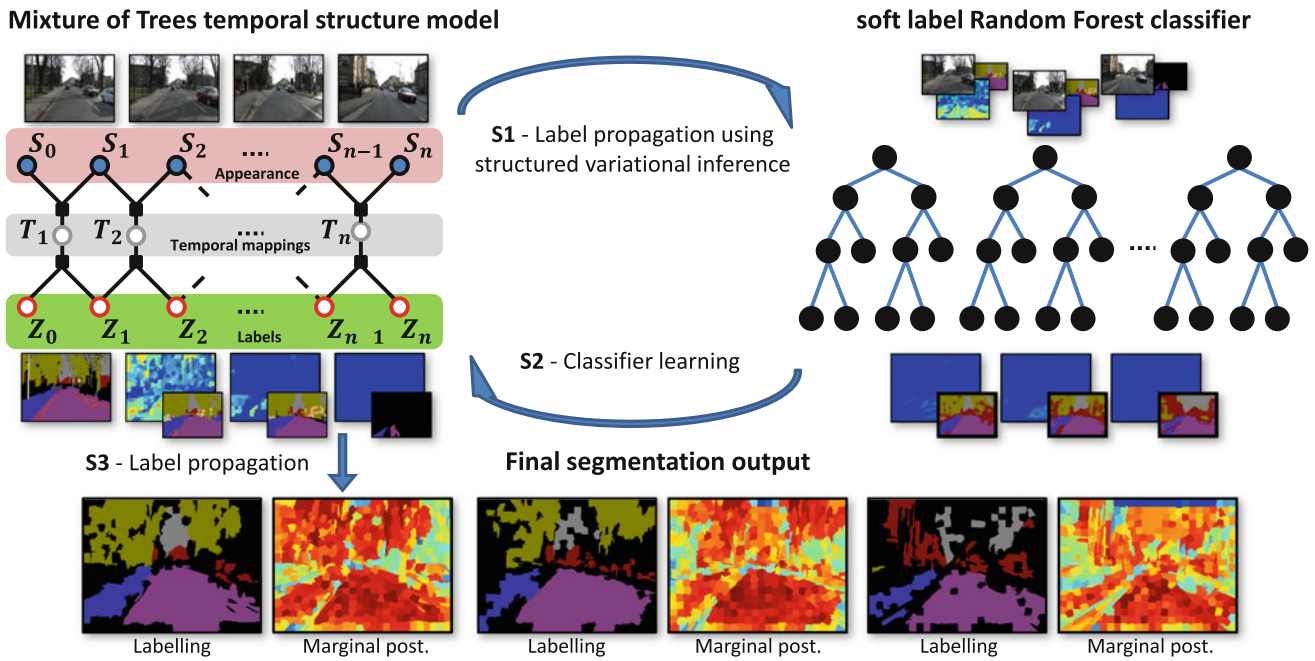

Fig. 2 On the top left is the factor graph of the proposed MoT model. Image appearance, super-pixel labels and their temporal links are jointly modelled. In stage S1, we perform structured variational inference, without any super-pixel unaries, to estimate the super-pixel marginal posteriors. This is followed by stage S2 wherein a soft label Random Forest classifier is trained with pixel marginal posteriors as vector or soft labels. The predictions from the classifier are injected back as super-pixel unaries for a second iteration of label inference (S3) to produce the final super-pixel labels and their uncertainties. Note that pixels which are uncertain are shown in black in the "labelled" images (Color figure online)

by performing energy minimization. As pointed out by Bai et al. (2009), performing MAP inference on large 3D volumes results in an uncontrollable work flow. Finally, multi-label MAP inference is computationally expensive (Tsai et al. 2010), necessitating short overlapping time window based video segmentation.

3 Proposed MoT Video Model

We super-pixelize each image in the video sequence using the SLIC algorithm (Achanta et al. 2010) into about 500 super-pixels. Let $S_{i,j}$ denote super-pixel j at frame i, and $Z_{i,j}$ denote its corresponding missing label. We associate the temporal mapping variable $T_{i,j}$ to super-pixel $S_{i,j}$. $T_{i,j}$ can link to super-pixels in frame $i-1$ which have their centers within a window $W_{i,j}$ (50×50), placed around the center of $S_{i,j}$. Note that this implies that each $T_{i,j}$ can have a different range.

Let $S_i = \{S_{i,j}\}_{j=1}^{\Omega(i)}$, $Z_i = \{Z_{i,j}\}_{j=1}^{\Omega(i)}$ and $T_i = \{T_{i,j}\}_{j=1}^{\Omega(i)}$ denote the set of super-pixels, their labels and the corresponding temporal mapping variables respectively at frame i. $\Omega(i)$ denotes the number of super-pixels in frame i. Our proposed MoT probabilistic graphical model (see Fig. 2) for the video sequence factorises as follows:

$$p\left(S_{0:n}, Z_{0:n}, T_{1:n}|\mu\right) = \frac{1}{\mathcal{Z}\left(\mu\right)} \times \prod_{i=1:n} \prod_{j=1:\Omega(i)} \Psi_a\left(S_{i,j}, S_{i-1,T_{i,j}}\right) \Psi_l\left(Z_{i,j}, Z_{i-1,T_{i,j}}|\mu\right) \times \Psi_u\left(Z_{i,j}\right) \Psi_u\left(Z_{0,j}\right) \Psi_t\left(T_{i,j}\right), \quad (1)$$

where $S_{i-1,T_{i,j}}$ indexes the super-pixel mapped to by $T_{i,j}$ in frame $i-1$ and similarly for $Z_{i-1,T_{i,j}}$.

To define the *appearance factor* $\Psi_a(.)$ of the MRF on the R.H.S of (1), we first find the best match pixel in frame $i-1$ for a pixel in frame j by performing patch cross-correlation within a pre-fixed window (3×3 patch size and a window size of 50×50). The appearance factor is then defined using the number of pixels in super-pixel $S_{i,j}$ which have their best matches in $S_{i-1,T_{i,j}}$ as follows,

$$\Psi_a\left(S_{i,j}, S_{i-1,T_{i,j}}\right) \triangleq \#\text{shared pixel matches} \quad (2)$$

Note that more sophisticated super-pixel match scores can also be substituted here, for instance those based on colour histograms, texton histograms, optic flow and SIFT-flow as in Fathi et al. (2011). In our experiments , we demonstrate that the simple measure in (2) already provides us with competitive results (see Table 1).

Table 1 Quantitative evaluation on the SegTrack tracking and segmentation dataset (Tsai et al. 2010)

Video Sequence	Properties		Performance comparison							
	Average object size	No. of frames	Chockalingam et al. (2009)	Tsai et al. (2010)	Fathi et al. (2011)	Budvytis et al. (2011)	Budvytis et al. (2011) tuned	Proposed MoT model		
								Best Tree only	MoT	MoT with Classifier
Parachute	3,683	51	502	**235**	251	404	258	429	301	**296**
Girl	8,160	21	1,755	1,304	1,206	1,705	**820**	3,795	2,387	**1,200**
Monkey-dog	1,440	71	683	563	598	736	**387**	2,054	509	**412**
Penguin	20,028	42	6,627	1,705	1,367	19,310	**1,212**	3,218	**1,736**	29,461
Bird-fall	495	30	454	**252**	342	468	259	570	**434**	508
Cheetah	1,584	29	1,217	1,142	**711**	1,501	923	**851**	870	855

Bold values indicate the best results. In all these experiments only the start frame of the video sequence is user labelled. The score is the average label mismatch per frame computed using the ground truth. Our proposed method with the use of the learnt classifier, and with manually tuned parameters, out performs most methods in two sequences (girl, monkey-dog) and shows comparable performance in another two (parachute, cheetah). We perform poorly in the birdfall sequence due to the very small size of the foreground object and due to severe foreground/background overlap in the penguin sequence. The best overall performance is the manually tuned tree structured model of Budvytis et al. (2011) which uses rectangular patches instead of super-pixels for segmentation. We also provide results when only the best tree among the mixture is chosen for segmentation and compare it alongside the full mixture of trees based segmentation. On this dataset, our MoT model shows more competitive results

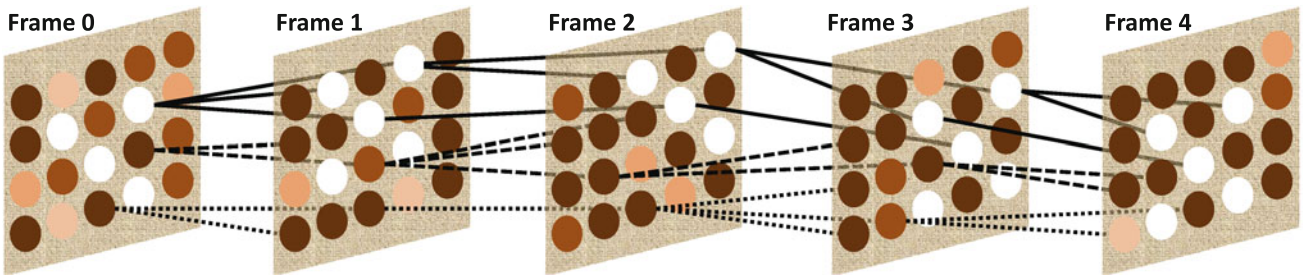

Fig. 3 An illustration of the MoT model with three component temporal trees in the mixture. Each node is a super-pixel. Moving from root to leaf three temporal tree structures are visible. During inference, super-pixel label predictions from each tree is weighted by the posterior probability of the corresponding tree in the mixture (Color figure online)

The *label factor* $\Psi_l(.)$ is defined between the multinomial super-pixel label random variables as follows.

$$\Psi_l\left(Z_{i,j} = l, Z_{i-1,T_{i,j}} = m|\mu\right) \triangleq \begin{cases} \mu & \text{if } l = m, \\ 1-\mu & \text{if } l \neq m. \end{cases} \quad (3)$$

where l, m take values in the label set $\mathcal{L}$. μ is a parameter which controls label affinity. The single node potential for the temporal mapping variables $\Psi_t(.)$ is similar to a box prior and is defined as follows.

$$\Psi_t\left(T_{i,j}\right) \triangleq \begin{cases} 1.0 & \text{if } T_{i,j} \in W_{i,j}, \\ 0.0 & \text{if outside.} \end{cases} \quad (4)$$

The super-pixel label unary factors $\Psi_u\left(Z_{i,j}\right)$ are defined in Sect. 4.2.

From (1) we note that the temporal mapping variable is present both in the appearance and label factor. Thus these variables are *jointly* influenced by both object appearance and semantic labels, a property which is desirable for interactive video segmentation systems.

As shown in Fig. 3, for a given instance of each of the mapping variables, a temporal tree structure is formed linking the super-pixels from the first (root) to the last frame (leaves). Therefore, the probabilistic model in (1) is a *mixture of temporal trees* (MoT model). Below we present our structured variational inference strategy for the MoT model.

4 Inference

It is clear from (1) that computing the partition function $\mathcal{Z}(\mu, \sigma)$ is computationally intractable due to the combinatorial number of states of the temporal mapping variable and the super-pixel labels. Therefore, we resort to structured variational inference, which is an approximate inference scheme (Saul and Jordan 1996) which allows us to incorporate suit-

able types of independency relationships in the approximate posterior. The key idea in this inference scheme is to retain as much structure in the approximate posterior as computationally tractable. This provides more robust inference as opposed to the more basic mean-field variational inference wherein the joint posterior of the variables factorises into independent terms (Turner et al. 2008).

In our work, we assume the following form for the *approximate variational posterior* of the latent variables.

$$Q\left(Z_{0:n}, T_{1:n}\right) \triangleq Q\left(Z_{0:n}\right) \prod_{i=1:n} \prod_{j=1:\Omega(i)} Q\left(T_{i,j}\right), \tag{5}$$

where the temporal mappings are assumed independent in the approximate posterior (as in mean-field approximations). However, the super-pixel latent labels do not factorise into independent terms, thereby maintaining *structure* in the posterior. This form is chosen as a compromise between performing tractable inference and retaining as much structure as possible in the posterior.

The observed data log likelihood can now be lower bounded using the approximate posterior in (5) as follows.

$$\log\left(S_{0:n}|\mu\right) \geq \sum_{Z_{0:n}, T_{1:n}} Q\left(Z_{0:n}, T_{1:n}\right) \times \log\left(\frac{p\left(S_{0:n}, Z_{0:n}, T_{1:n}|\mu\right)}{Q\left(Z_{0:n}, T_{1:n}\right)}\right) \tag{6}$$

To maximise the above lower bound, which is a functional of the variational posterior and the model parameters, we employ calculus of variations (Bishop 2006) and obtain the following fixed point equations for the approximate posteriors.

$$Q\left(T_{i,j}\right) \propto \Psi_t\left(T_{i,j}\right) \exp\Big[\sum_{Z_{i,j}, Z_{i,T_{i,j}}} Q\left(Z_{i,j}, Z_{i,T_{i,j}}\right) \times \log\Big(\Psi_a\left(S_{i,j}, S_{i-1,T_{i,j}}\right) \Psi_l\left(Z_{i,j}, Z_{i-1,T_{i,j}}|\mu\right)\Big)\Big], \tag{7}$$

$$Q\left(Z_{1:n}\right) \propto \prod_{i=1:n} \Psi_u\left(Z_i\right) \exp\Big[\sum_{T_{i,j}} Q\left(T_{i,j}\right) \times \log\Big(\Psi_l\left(Z_{i,j}, Z_{i-1,T_{i,j}}|\mu\right)\Big)\Big]. \tag{8}$$

To compute the approximate super-pixel label marginals and pair-wise marginals required for the above equation we use variational message passing (Bishop 2006). The variational message (vm) which super-pixel $S_{i,j}$ sends to its temporal neighbour $S_{i-1,T_{i,j}}$ is as follows.

$$vm_{Z_{i,j} \to Z_{i-1,T_{i,j}}}\left(Z_{i-1,T_{i,j}}\right) = \sum_{Z_{i,j}} \exp\Big[Q\left(T_{i,j}\right) \times \log\Big(\Psi_l\left(Z_{i,j}, Z_{i-1,T_{i,j}}|\mu\right)\Big)\Big] \prod_{n \in Ne\left(Z_{i,j}\right) \setminus Z_{i-1,T_{i,j}}} vm_{n \to Z_{i,j}}. \tag{9}$$

Algorithm 1: Mixture of Trees (MoT) model for Video Segmentation

Input: Super-pixels $S_{0:n}$ (video), User labelled frames.
Output: Pixel label probabilities.
Initialisation
Set the initial values of μ to those given in Sect. 5.
Set all unaries to uniform distributions.
Set all variational posteriors to uniform distributions.
Set max_iter = 50;

Infer temporal mapping posteriors $Q\left(T_{i,j}\right)$ using (7) for both forward and time reversed sequences.
S1. for $i = 1$ **to** *max_iter* **do**
  Infer $Q\left(Z_{i,j}\right)$ using (10).

Do this for both forward and backward MoT models.
S2. Train the Random Forest with average of posteriors $Q\left(Z_{i,j}\right)$ from both forward and backward MoT models as labels.
S3. Set the super-pixel unaries to the predictions from the Random Forest. See Sect. 4.2.
Perform label inference with unaries on both forward and backward MoT models.
For each super-pixel, average its label posteriors inferred from both the models. See Sect. 4.3.
Assign all pixels of a super-pixel the same marginal posterior as the super-pixel itself.

Using the above messages, the approximate variational posteriors can be obtained as shown below.

$$Q\left(Z_{i,j}\right) \propto \prod_{n \in Ne\left(Z_{i,j}\right)} vm_{n \to Z_{i,j}} \tag{10}$$

$$Q\left(Z_{i,j}, Z_{i-1,T_{i,j}}\right) \propto \prod_{n \in Ne\left(Z_{i,j}\right) \setminus Z_{i-1,T_{i,j}}} vm_{n \to Z_{i,j}} \times \prod_{n \in Ne\left(Z_{i-1,T_{i,j}}\right) \setminus Z_{i,j}} vm_{n \to Z_{i-1,T_{i,j}}} \times \exp\Big[Q\left(T_{i,j}\right) \log\Big(\Psi_l\left(Z_{i,j}, Z_{i-1,T_{i,j}}|\mu\right)\Big)\Big]. \tag{11}$$

In our experiments, we first set all the variational single and pairwise posteriors to uniform distributions. Then we compute $Q\left(T_{i,j}\right)$ once at the first iteration. Then after a fixed number of message passing iterations, (10) is used to compute the approximate super-pixel marginal posterior. We also tried to alternate between inferring the temporal mappings and the labels but this is computationally expensive and did not improve the results. Therefore, we only performed a single round of inference of the mapping variables. A summary of the inference technique with a view to encourage implementation is given in Algorithm 1.

4.1 Influential Parameters

We introduce two parameters α, β which control the effect of mixing from different trees for label inference and the strength of variational messages respectively. We compute

$Q\left(T_{i,j}\right)^{\alpha}$ and re-normalize to obtain an α controlled posterior over the temporal mapping variables. Larger values of α imply the label inference is influenced by fewer components of the temporal mixture of trees. This reduces the effective number of loopy cliques in the model.

The MoT model is loopy by construction and the features used to create temporal linkages [see (2)] can result in very flat $Q(T_{i,j})$ distributions. Therefore, in practice, variational messages reduce to near uniform distribution after a few frames only. To tackle this problem, at each iteration of variational message passing, we raise the messages to a power β and re-normalize. This step helps propagate messages over longer durations. In Sect. 5 we discuss the effect of varying these parameters on the accuracy of segmentation.

4.2 Semi-Supervised Learning of Unaries

In the first iteration of inference, we set the super-pixel unaries to uniform distributions and use our proposed inference technique to estimate the super-pixel marginal posteriors. We assign each member pixel of a super-pixel the same marginal posterior as the super-pixel itself. We then train a Random Decision Forest (Breiman 2001) using these posteriors as soft pixel labels, i.e each pixel has a vector label instead of a scalar class label. At training time, we compute a histogram of soft labels at each node of a tree in the Forest by element wise addition of the vector labels and use the "entropic information gain" criterion used by Shotton et al. (2008) to evaluate the split function. We use simple but computationally efficient pixel intensity difference features at each split node as in Shotton et al. (2008). We term this Random Forest, trained in a *semi-supervised* manner, the soft label Random Forest (slRF) as in Budvytis et al. (2011).

We bootstrap the predictions from the learnt slRF into the MoT time-series model. We assign each super-pixel the average of the slRF predicted label distributions of its pixels. The *averaged distribution* is the unary $\Psi_u\left(Z_{i,j}\right)$ for each super-pixel which is used in the second iteration of super-pixel label inference. These unaries, learnt in a semi-supervised manner, can help improve segmentation accuracy as shown in Fig. 8. Unlike traditional tracking algorithms were the unaries are learnt using the first frame labels, we use the entire video data and the corresponding inferred labels to learn the unaries. In some approaches to segmentation (Fathi et al. 2011), labels are propagated to the adjacent frame and their MAP estimate is used to update the unary parameters. This is sub-optimal, given that the entire video volume is not used to update the unary. In contrast, our efficient inference method allow us to pool in the entire video data and the label posteriors to learn the unaries.

Our semi-supervised training of the Random Forest is different from the transductive forest described in Criminisi and Shotton (2013). In the transductive forests, labelled and unlabelled data are treated separately and a new information gain criterion is introduced to combine label and appearance based entropies. In contrast, we first assign each unlabelled data point a soft label obtained from the label inference step. At training time, we compute a histogram of soft labels at each node and use the information gain criterion of Shotton et al. (2008) to evaluate the split function.

4.3 Forward and Backward MoT

The component trees in the MoT model have their root in the first frame and leaves in the last frame. This introduces a temporal bias. We correct for this bias by performing inference on a time reversed video sequence and averaging the super-pixel label posteriors from the forward and reversed time-series. This averaged posterior of super-pixel labels is used in the semi-supervised training of the slRF discussed above.

4.4 Best Tree Versus MoTs

If we approximate the variational posterior $Q\left(T_{i,j}\right)$ by a single point posterior at its MAP location then the variational message passing (9) reduces to standard message passing on a best tree structured MoT model (exact inference). Inference on this best tree structure can be performed very efficiently at low memory and computational cost (see Table 3). This best tree structure quite often demonstrates even better segmentation performance than the full mixture of trees model. This is particularly when the feature used to measure super-pixel similarities is weak and thus unable to provide a good quantitative ranking of the possible matches for a super-pixel in an adjacent frame. This poor ranking is directly reflected in the $Q\left(T_{i,j}\right)$ distribution where the probabilities of each state of $T_{i,j}$ are unreliable. In our experiments, we have found instances where the mixture model performs better than the best tree model and some others where its performance is inferior to the best tree model (see Sect. 5).

5 Experiments and Results

We evaluated the performance of our proposed MoT model based algorithm on binary and multi-class semi-supervised video segmentation problems. The experiments are explained in more detail below.

5.1 Binary Segmentation

We evaluated the performance of our approach in a tracking and segmentation setting using the challenging SegTrack (Tsai et al. 2010) dataset. This dataset with each frame ground truth consists of six sequences with clutter, self-occlusion, small sized objects and deformable shape filmed with a moving camera. In Table 1 we report our scores (number of pixel

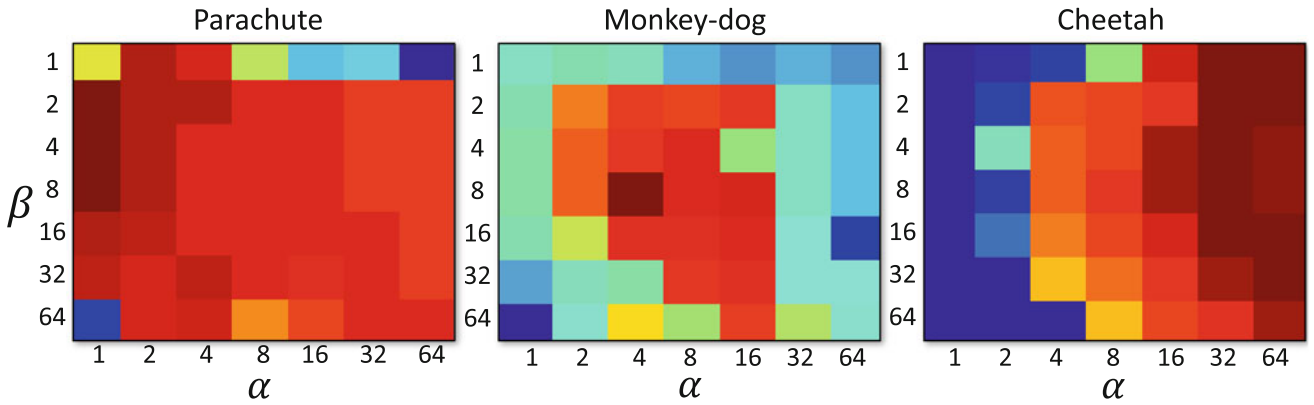

Fig. 4 Sensitivity of pixel labelling accuracy to parameters α, β (see Sect. 4.1) shown for three of the SegTrack (Chockalingam et al. 2009) sequences. In each heat map, deep blue hues represent very low accuracy and deep red represents high accuracy. It is clear from these maps that there is no common parameter value(s) which provide high performance for all the sequences. In the parachute sequence, high accuracy is obtained for low values of α which corresponds to many mixture components in the MoT model. In the Monkey-dog and Cheetah sequences, the α values progressively increase for higher accuracy. This reduces the number of influential components in the mixture, making the model less *loopy* (Color figure online)

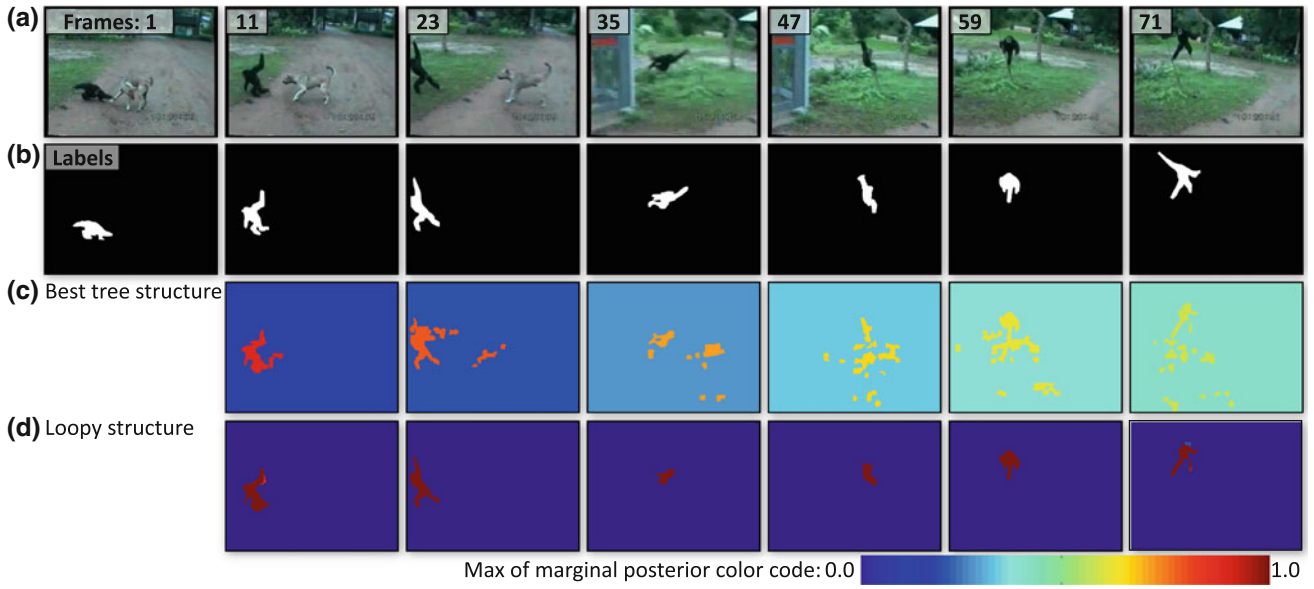

Fig. 5 Qualitative difference in segmentation when only the best tree of the mixture is used for segmentation as compared to when the full mixture of trees is used. The best tree is computed by retaining only the MAP estimate of each of the temporal mapping variables. In this example, the best tree based inference where each super-pixel has only one link to another super-pixel in the previous frame, performs poor quality label propagation. This is because of some erroneous super-pixel linkages in the tree (connections between foreground and background super-pixels). In contrast, in the MoT model each super-pixel is connected to several others in the neighbouring frame. Therefore, even if the best link for a super-pixel is erroneous, still erroneous label propagation can be avoided if the majority of other possible links are correct, i.e connect to the correct class (Color figure online)

label errors per frame) using manually selected parameters and settings. Scores of some of the recent state of the art approaches are also reported alongside results of our method with different settings. The value of μ was empirically set to 0.95 for this experiment.

Effect of the influential model parameters α, β To demonstrate that a single value of influential parameters α, β is insufficient to cover this entire dataset, we plot the score for the sequences over different values of the parameters (see Fig. 4). Note that we do not include any unaries while computing this score. Therefore, this error score gives us an indication of the performance of the MoT time-series alone. Using this plot, we select the parameter settings for each sequence which results in the lowest labelling error. It is clear from this experiment that the optimal temporal structure is different for each sequence. For example, in the monkey-dog sequence from the SegTrack dataset (Tsai et al. 2010) shown

in Fig. 5 the more "loopy" temporal structure MoT model performs better than the single best tree structured MoT model. If any single link in the best tree is incorrect (due to the feature used to measure super-pixel similarities) then the labelling error propagates. However, in the loopy MoT model with several tree components in the mixture, each super-pixel has several links in the adjacent frame. Therefore, even if the MAP link is erroneous (connects to another class), it is still possible to avoid erroneous label propagation if the majority of the other links connect to super-pixels of the correct class. In other sequences, such as the cheetah sequence from the same dataset, a very high value of α provides the best performance. At such as high value of α, the model is closer to a tree structure than a loopy one.

In comparison, Fathi et al. (2011) attempt to automatically learn their model parameters by self-training, that is by fixing some of the inferred labels and using these labels to learn the weights given to temporal predictions versus unary (classifier) predictions. However, in this paper we do not adopt this approach to parameter setting as it relies on instantaneous decision making which often leads to label drift.

We also find from our experiments that in some instances the performance improves when a unary term learnt from only the first frame is included in the starting iteration (Girl sequence in Table 1). In the Penguin sequence, we avoid the use of any unaries as there is significant overlap between foreground object and background. In the remaining sequences, we follow the setting prescribed in Algorithm 1.

We can observe good qualitative performance of our algorithm in Fig. 6. It is important to note that a lower quantitative score does not necessary imply a poor qualitative result. For instance, in the fast motion cheetah sequence, the foreground object is tracked and segmented reasonably well. However, a small part of the background which appears as true positive foreground in a few frames [see Fig. 6 (22–26)] lowers quantitative accuracy.

In all our experiments, each channel in all the images are scaled to lie between [0.0, 1.0]. We choose the 1*st* stage Random Forest (RF) classifier, as in Shotton et al. (2008), with 16 trees, each of depth 8. Input LAB patches of 21×21 are extracted around every 2*nd* pixel on both axis. We leave out border pixels in a 10 pixel band to fit all rectangular patches. We use the same kind and number of features as in Shotton et al. (2008). The key difference is that we use the inferred pixel label posteriors to train the slRF. We compute the entropic information gain and the leaf node distributions (normalized histograms) by treating the data point label as a *vector* whose elements sum to unity. This has the advantage that entropies can be computed directly using label distributions (see Sect. 4.2).

5.2 Multi-Class Segmentation

We used the publicly available CamVid driving video dataset (Brostow et al. 2009) for our multi-class video segmentation experiments on long and challenging video sequences. We chose sequence seq05VD (30 Hz) in this dataset and divided it into six sequences as in Budvytis et al. (2011). We use three of the six sequences, described in Table 2, to perform various quantitative and qualitative comparisons. All the sequences considered are of 750 frames in length and are uniformly down sampled to a length of 150 frames in order to reduce computational and memory requirements. In order to make a fair comparison with the state-of-the-art algorithm of Budvytis et al. (2011) the resolution of video frames are reduced to 320×240. The value of μ was empirically set to 0.7 for this experiment.

The quantitative analysis reported in Table 2 is performed on classes like roads, pavements, road markings and others (10 classes including a void class) which are relevant to driving. The metrics used for this evaluation are global accuracy (percentage of pixels labelled correctly) and class average accuracy measured for all static classes (ASC), small static classes (SSC = {signs, poles, road markings}) and large static classes (LSC) which do not include SSC. These accuracies are computed only over pixels labelled into known classes in the ground truth.

We first discuss a study of the performance of the stages S1, S2, S3 (see Fig. 2) of our proposed semi-supervised segmentation/label propagation algorithm. Note that the experimental set up for label propagation and slRF learning is nearly identical to the binary segmentation experiment described earlier in Sect. 5.1, including the method to choose the optimal values of the α and β parameters. The only difference is that, the patch size is set to 7×7 (see Sect. 3).

Labelling accuracy and density We present evaluations of our algorithm under various metrics for Sequence 1 from the CamVid dataset in Fig. 7. The graph on the left panel of Fig. 7 shows how the segmentation accuracy changes when the number of accepted pixel labels is varied. The number of accepted pixels can be varied by applying a threshold over the uncertainty of the most likely pixel label. Note that all the stages of our algorithm provide uncertainty estimates of each pixel label. From this graph it is clear that, in general, the accuracy improves as we move from stages S1 to S3. This shows the beneficial effect of semi-supervised learning using the pixel label distributions, particularly as the number of accepted points are increased. This point is again emphasized in Fig. 8 which shows qualitative results for each one of the three stages along with their uncertainty maps.

The graph in the middle panel compares global accuracy and percentage of labelled points obtained at various thresh-

Fig. 6 An qualitative results on the SegTrack dataset (Tsai et al. 2010). In all these experiments only the start frame of the video sequence is user labelled. Notice how our algorithm is able to cope with fast motion, motion blur, shape changes (**a, b, c, d, g, h, i, j, k, l**) and small sized objects (**f**). The main failure case is (**e**) due to its small size. Also see the supplementary video (Color figure online)

olds of label uncertainty for the MoT model at different values of the α parameter (this controls the "loopyness" of the model), and including the MoT model which only uses the best tree structure (see Sect. 4.4). The highest accuracy is obtained for the best tree structured MoT model and the MoT model with α set to 32. At such a high value of α

Table 2 This figure illustrates a qualitative comparison with other recent semi-supervised video segmentation algorithms (Budvytis et al. 2010) (HLP) and (Budvytis et al. 2011) (PBM) (Color table online)

Settings			Class accuracies for static classes									All static classes (ASC)			Large static classes(LSC)			Small static classes (SSC)		Outlier Classes
Sequence	Frames	Model	Threshold	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 – 810	HLP	-	75	26	27	0	7	93	91	97	78	79	55	53	82	77	51	57	44	87
		PBM	0.75	100	77	0	0	0	94	99	94	0	90	52	53	92	78	52	69	31	86
		MoT, α=8	0.971	0	91	0	0	0	100	0	20	0	73	23	53	76	35	53	51	49	67
		MoT, α=16	0.911	31	96	0	0	0	98	43	36	0	81	34	53	84	51	53	47	53	60
		MoT, α=32	0.756	92	98	0	0	0	97	85	57	0	88	48	53	90	72	53	60	40	65
		MoT, tree	0.663	98	95	0	0	0	97	88	70	0	91	50	53	93	75	54	64	36	73
2	2310 – 3060	HLP	-	99	87	77	30	65	88	78	35	-	83	70	94	88	80	82	62	39	2
		PBM	0.12	94	94	16	3	66	88	88	59	-	84	63	90	90	85	80	39	61	0
		MoT, α=8	0.464	76	99	0	0	0	96	68	6	-	78	43	91	86	58	90	3	97	4
		MoT, α=16	0.33	87	98	0	0	0	97	69	2	-	78	44	91	87	59	90	5	95	3
		MoT, α=32	0.263	91	98	0	0	0	95	74	3	-	79	45	91	87	60	90	7	93	3
		MoT, tree	0.205	91	98	0	0	0	94	76	3	-	79	45	91	87	60	91	8	92	1
3	3060 – 3810	HLP	-	98	92	16	12	37	93	85	9	-	90	55	45	92	76	44	62	38	38
		PBM	0.77	100	99	0	0	6	93	93	0	-	89	49	47	90	77	46	75	25	39
		MoT, α=8	0.987	0	98	0	0	0	97	81	0	-	75	35	46	77	46	47	63	37	67
		MoT, α=16	0.936	4	100	0	0	0	96	86	0	-	80	37	47	81	48	48	61	39	51
		MoT, α=32	0.797	82	100	0	0	0	95	80	0	-	80	45	47	82	60	48	68	32	49
		MoT, tree	0.714	91	99	0	0	0	95	83	0	-	81	46	47	82	61	48	70	30	51

Video segmentation is performed on long and complex (750 frames) sequences the from CamVid (Brostow et al. 2009) dataset. Our method obtains comparable accuracy over large static classes (road, sky, building, pavements) for a similar label density and a similar false positive rate as the method of Budvytis et al. (2011) for Sequence 1. The average accuracy is lower in Sequences 2 and 3, mainly due to misalignment of super-pixels with class edges and a poor ranking of super-pixel matches across adjacent frames. However our segmentation algorithm is both faster and more memory efficient by about two orders of magnitude

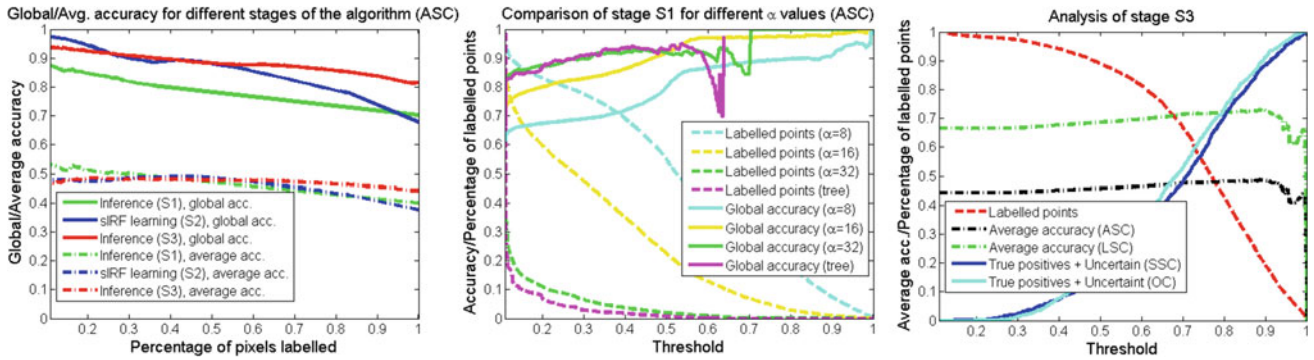

Fig. 7 The three plots illustrate various performance measures for different stages of our algorithm. The plot on the left panel indicates that the segmentation quality increases from stage S1 to stage S3. The middle panel shows that for high values of α (with fewer tree components in the mixture) we obtain a better global accuracy as the threshold over the super-pixel MAP label uncertainty is varied. The plot on the right panel demonstrates how average accuracy and number of potentially correctly labelled points (true positives and uncertain) behaves depending on the uncertainty threshold. This plot is used to choose an optimal threshold over label certainty to obtain the results reported in Table 2 (Color figure online)

there are very few tree components in the MoT model and it begins to resemble the best tree structured MoT model more closely. For lower values of α, which retains many components in the mixture, the model is quite loopy. This has the effect that the labels of pixels towards the middle of the sequence are falsely confident, and thus the accuracy for a particular threshold is poorer even though the label density is higher.

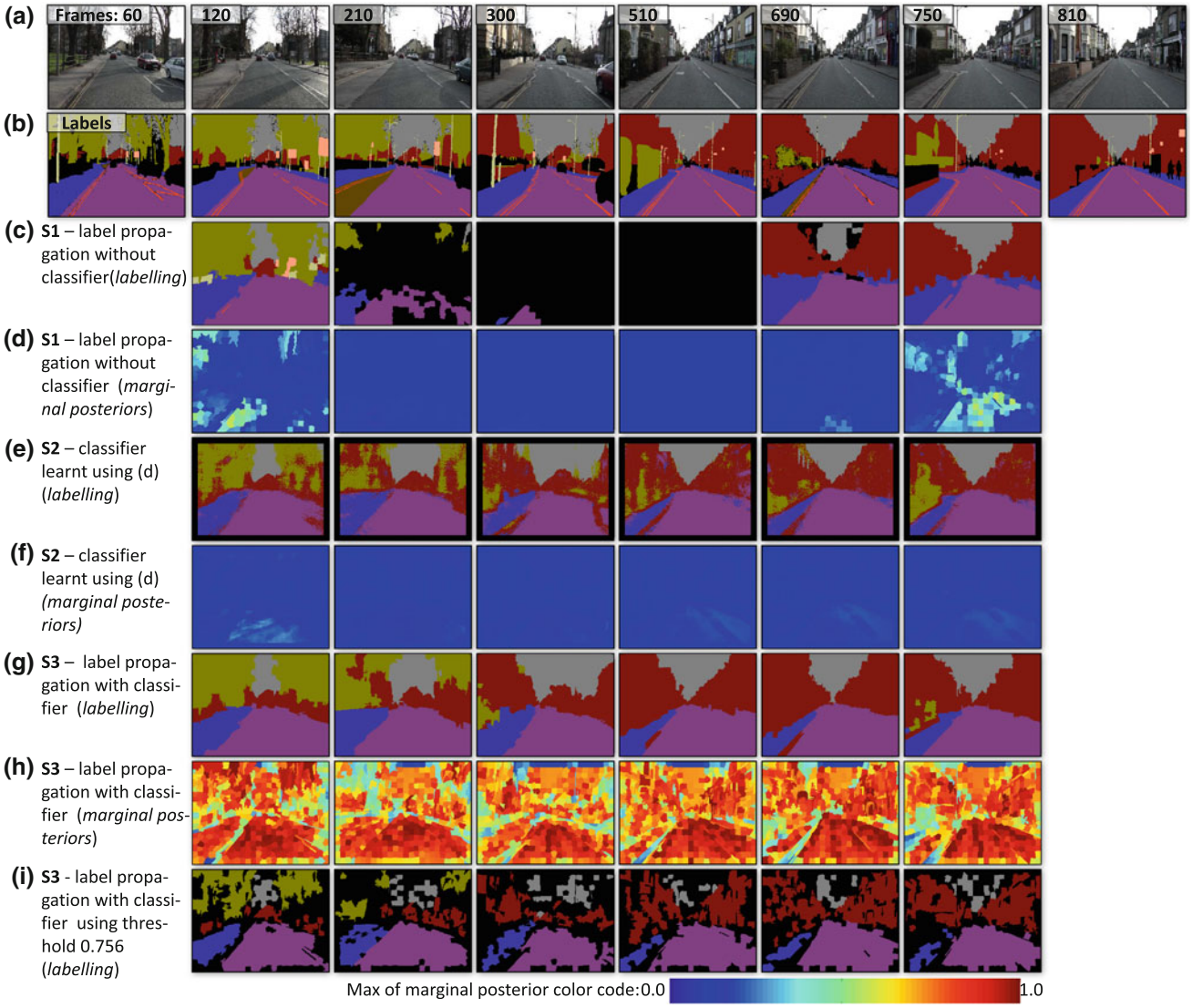

Fig. 8 This figure illustrates the three stages S1 (**c, d**), S2 (**e, f**), S3 (**g, h, i**) of our segmentation algorithm performed on Sequence 1 from the CamVid (Brostow et al. 2009) dataset. Row (**a**) shows the frames and (**b**) the corresponding ground truth. The rows (**c, e, g**) indicate the most likely class with its corresponding marginal posterior or uncertainty map represented respectively in rows (**d, f, h**). Row (**i**) shows the segmentation result when a threshold of 0.756 is applied over the marginal posteriors represented in row (**h**). Note the increasing super-pixel label uncertainty further away from the labelled ends in (**d**). Also note that large static classes such as road, pavement, building are labelled quite well. However, our algorithm performs poorly for small static classes such as signs, road markings or poles, mainly due to their very small resolution. Note that pixels which have completely uncertain label distributions are shown in black color in (**c, e, g, i**). Also see the supplementary video (Color figure online)

Finally, the graph on the right panel shows a more detailed analysis of the class average accuracy of our algorithm on all static classes (ASC), large static classes (LSC), small static classes (SSC) and outlier class (OC). The outlier class consists of all pixels with a void label in the ground truth. As the threshold over the uncertainty of the most likely label at each pixel is increased, intuitively the accuracy of ASC, LSC increases with a corresponding decrease in label density. The true positive labels and uncertain labels over SSC increases, which indicates that these classes are not falsely labelled as other large classes. Similarly, the true positives and uncertain labels over OC also increases, which indicates that the algorithm does not fill in the void labels with other classes as the threshold in increased. Note that for the outlier class accuracy computation, the true positives are those which are due to propagation of void labels provided in the user labelled end frames and uncertain labels correspond to those which are below the threshold. We use this graph to compute an optimal threshold for pixel MAP label uncertainty such that there is a balance between the percentage of labelled points and the class average accuracy.

Comparison with other state-of-the-art approaches Table 2 compares the accuracy of stage S3 of our algorithm with methods of Budvytis et al. (2010) and (2011) for a similar label density. A trend which is common to all three sequences is the increase in accuracies (both global and class average) as the α value is increased and which culminates with the highest accuracy when only the best tree structured MoT model is used. For Sequence 1, we obtain accuracies comparable to the other competing methods, however the performance is poorer for the remaining sequences. We offer two reasons for this observation. First, the use of an off the shelf super-pixellization algorithm which produces super-pixel boundaries that often does not align with class edges. The second reason is the use of a simple feature (number of shared patch matches) to measure super-pixel similarities. This results in an unreliable ranking of similarities between super-pixels and so the probabilities of different states of the mapping variables are not a true indicator of its strength.

Since the CamVid sequences are long, it is desirable to expect the labels to be more uncertain towards the middle of the sequence, where objects unseen in the labelled frames appear. Unfortunately, this effect does not occur for low values of α when the MoT model is very loopy. The variational messages reinforce themselves after several iterations of message passing resulting in very confident marginal posteriors. Here we would like to clarify that the uncertainty does not decrease in the case where the full mixture of trees is used. In fact, for low values of α, the uncertainty level remains constant throughout the sequence. Another issue with very loopy models is that the marginals are quite sensitive to even small numerical approximations and round-off errors. The more tree structured versions of the MoT model (low α values) are less prone to these issues and result in more uncertain labels toward the middle of the long sequences. However, for short sequences and binary segmentation problems such as those used in Sect. 5.1 good performance can still be obtained for fairly loopy MoT models as indicated in Fig. 5. Figure 9

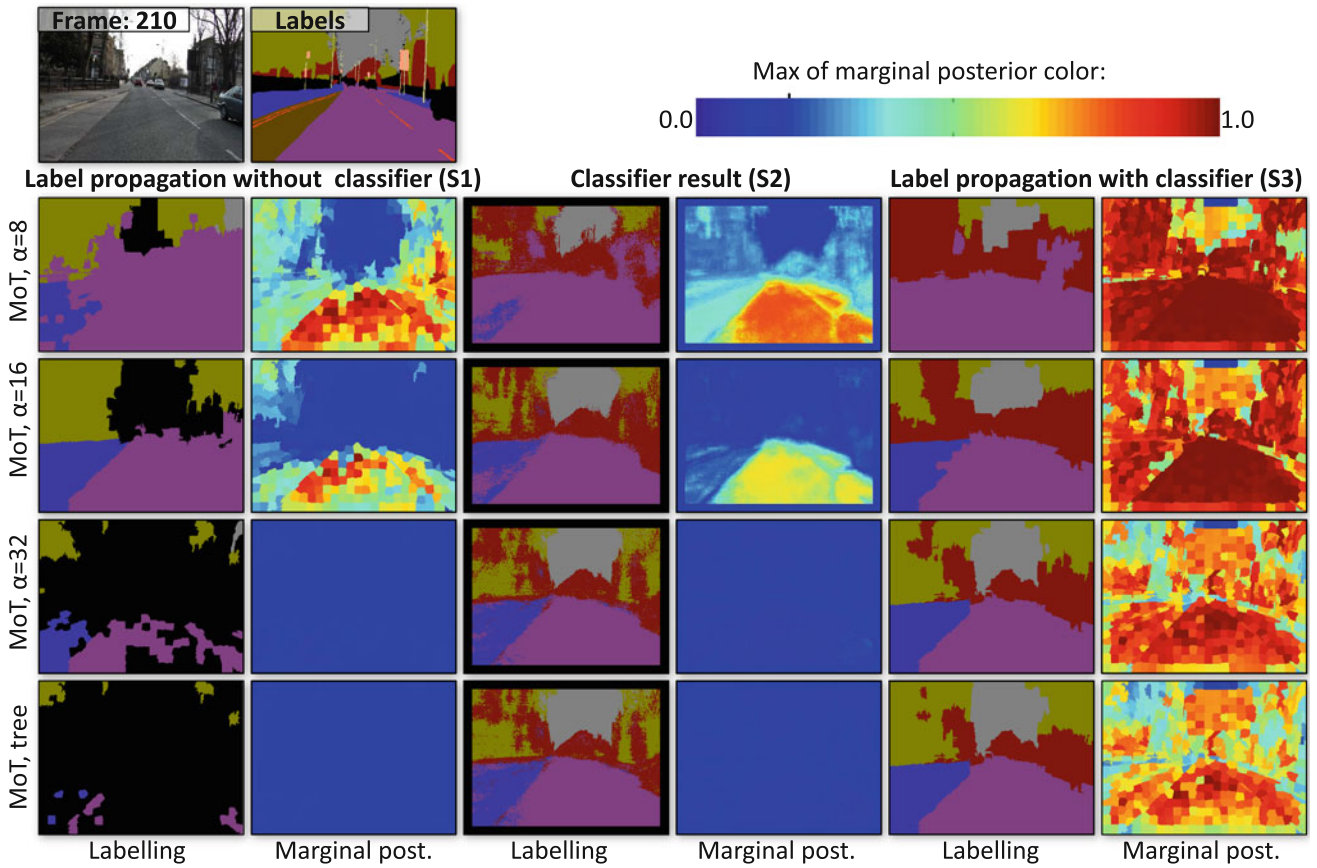

Fig. 9 This figure illustrates our segmentation output (labels and corresponding marginal posteriors for MoT models with different α values. The higher the value of this parameter, the fewer trees contribute to the mixture, thus making the model less loopy. The best result, as also seen from Fig. 7 and Table 2, is achieved when only the best tree in the mixture is used for inference. This is mainly due to misalignment of super-pixels to class edges and poor ranking of super-pixel matches across adjacent frames, which fails to exploit the flexibility of the mixture model as compared to the binary segmentation problem. Note that pixels which have completely uncertain label distributions are shown in black in the "labelling" panels. Also see the supplementary video (Color figure online)

shows a qualitative comparison of labelling results for varying α values for a frame drawn from the middle of Sequence 1. The confident but erroneous marginal posteriors can be observed at all stages of inference for high values of α.

6 Computational Requirements

We performed all comparisons of computational requirements on a machine with 8 core Intel Xeon (2.5 GHz) CPU with 8GB of RAM. Table 3 compares time taken per frame and the maximum RAM requirement for various stages of the algorithm of Budvytis et al. (2011) and our algorithm. The manner in which the temporal structure is inferred and the training of the classifier is the same as in our algorithm, hence identical memory usage and CPU usage numbers are reported for those steps. The main difference in computational effort between the algorithm of Budvytis et al. (2011) and our MoT model based algorithm can be seen in the label inference stages. The model of Budvytis et al. (2011) contains more than ten million random variables in total, compared to about hundred thousand in the MoT model. Therefore, their approach requires large amounts of RAM for both binary and multi-class segmentation. For the same reasons, it is also significantly slower than our best tree structured MoT model. The difference in computational time between the MoT model with many components in the mixture and their method is smaller as performing several iterations of variational message passing is time consuming.

The difference in computational time and memory requirement for the algorithm of Budvytis et al. (2011) and our model is large for the multi-class segmentation of Sequence 1 from the CamVid dataset. This is due to the need of storing and loading large chunks of memory into a limited RAM. However, our more efficient MoT model based algorithm has low computational load and thus can be applied to interactive video segmentation on devices with limited computational power such as smart phones or tablet PC's.

7 Advantages and Drawbacks

The main advantages of our approach are summarised below.

1. Our MoT video time-series model and the accompanying efficient variational inference scheme alleviates the need to perform overlapping time window based video volume processing. This helps avoid instantaneous decision making which often causes label drift.
2. We infer pixel-wise labels and their confidences (marginal posteriors). This is useful for both semi-supervised and active learning systems.
3. In addition, we model uncertainty in the temporal links explicitly which can sometimes correct for errors in the best tree structured MoT model of the video as shown in Fig. 5.
4. Our inference method is both computationally and memory wise efficient as can be seen from Table 3.

The main drawbacks and pointers to future work are:

1. The influential parameters α, β are manually set using grid-search. This is made possible due to efficient label inference. In future, we aim to learn these parameters for each video sequence in an interactive setting.
2. We use simple patch cross correlation based features to set the temporal links. This degrades performance and so must be replaced with more robust features which are able to rank the quality of matches between super-pixels.

Table 3 A comparison of computational requirements of our proposed algorithm and the one of Budvytis et al. (2011)

Algorithm step		Computing the trees		Label inference		Classifier learning	
Experiment\Computational requirements		Time *(per frame)*	Memory *(max)*	Time *(per frame)*	Memory *(max)*	Time *(per frame)*	Memory *(max)*
Girl sequence, (SegTrack), 2 classes, 20 frames, image resolution 320x240	Budvytis et. al.,2011	1.5 min	2 MB	3 sec	8 GB	1.4 min	60 MB
	MoT , best tree			0.15 sec	40 MB		
	MoT			1.0 sec			
Sequence 1, (CamVid), 9 classes, 150 frames, image resolution 320x240	Budvytis et. al.,2011	1.5 min	2 MB	2 min*/11 sec	40 GB	3.2 min	250 MB
	MoT , best tree			0.9 sec	80 MB		
	MoT			4.6 sec			

Note that our algorithm is faster and consumes less RAM by two orders of magnitude. This makes it applicable to real time interactive video segmentation applications. *2 min is the estimate when only 8GB of RAM is available, while the time reduces to 11 s when there is no such restriction

8 Discussions

The MoT model can be seen as an extension of the single tree model proposed earlier (Badrinarayanan et al. 2013). However, there are two key changes, the first is the use of super-pixels as the basic labelling unit and second is the use of a loopy temporal structure. The loopy temporal structure is employed to model the uncertainty present in the temporal mappings as opposed to using only the best tree as in Badrinarayanan et al. (2013). This mixture model, in principle, can be used with patches instead of superpixels as in Badrinarayanan et al. (2013). However, this is computationally very intensive and unsuitable for real time video segmentation. Using superpixels can make the mixture model usable in interactive settings as well.

Our proposed MoT model is loopy by construction and so we chose to perform variational inference (see Sect. 4) to estimate the super-pixel labels. However, variational inference is affected by issues such as numerical precision and this restricts us to explore those parameter settings which trade-off label density for accuracy (see Fig. 9). The use of better features for estimating temporal mappings and a more robust inference scheme can reduce this restriction.

From our experiments, we observed that in some sequences where temporal links (computed with simple features) can be established more reliably the MoT model performs better than the single tree model (see Fig. 5). However, in sequences where temporal links cannot be established (ranked) reliably the power of the MoT model diminishes (see Table 2). This effect was observed in complex and lengthy sequences where the MoT model failed to achieve the same level of performance as the single tree model. The performance of the MoT model is also affected by the super-pixellization algorithm used. We believe the use of more sophisticated features for establishing temporal mappings and better super-pixellization can improve the performance of our MoT model.

9 Conclusion

We presented a novel mixture of temporal trees (MoT) model for video segmentation. Each component in the mixture connects super-pixels from the start to the end of a video sequence in a tree structured manner. We provided a computationally and memory wise efficient inference scheme to estimate pixel-wise labels and their confidences, both for binary and multi-class segmentation problems. This inference scheme alleviates the need to perform short sliding time window based video volume processing which often results in erroneous label propagation. We demonstrated the efficacy of our algorithm on challenging binary and multi-class video segmentation datasets. For short sequences, we find that several components of the mixture contribute towards achieving good performance, however, on lengthy and complex sequences fewer mixture components or the best tree structure alone performs better. The use of an off-the-shelf super-pixellization algorithm and weak features to measure super-pixel similarities affect the performance of our algorithm, especially when more components in the mixture model are used. More improved super-pixellization schemes and similarity measures can produce a better performance of the MoT model. We now look forward to exploiting our proposed algorithm in interactive settings where its computational efficiency is desirable.

References

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., & Snsstrunk, S. (2010). *Slic superpixels*. Technical report, EPFL Technical Report no. 149300.

Badrinarayanan, V., Galasso, F., & Cipolla, R. (2010). Label propagation in video sequences. In *CVPR*.

Badrinarayanan, V., Budvytis, I., & Cipolla, R. (2013). Semi-supervised video segmentation using tree structured graphical models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*, 2751–2764.

Bai, X., Wang, J., Simons, D., & Sapiro, G. (2009). Video snapcut: Robust video object cutout using localized classifiers. *ACM Transactions on Graphics*, *28*, 70:1–70:11.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York: Springer.

Boykov, Y., & Jolly, M. P. (2001). Interactive graph cuts for optimal boundary and region segmentation of objects in n–d images. In *ICCV*.

Boykov, Y., Veksler, O., & Zabih, R. (1999). Fast approximate energy minimization via graph cuts. In *ICCV*.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Brostow, G., Fauqueur, J., & Cipolla, R. (2009). Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, *30*(2), 88–97.

Brox, T., & Malik, J. (2010). Object segmentation by long term analysis of point trajectories. In *ECCV*.

Budvytis, I., Badrinarayanan, V., & Cipolla, R. (2010). Label propagation in complex video sequences using semi-supervisedlearning. In *BMVC*.

Budvytis, I., Badrinarayanan, V., & Cipolla, R. (2011). Semi-supervised video segmentation using tree structured graphicalmodels. In *CVPR*.

Budvytis, I., Badrinarayanan, V., & Cipolla, R. (2012) Mot: Mixture of trees probabilistic graphical model for video segmentation. In *BMVC*.

Chen, A. Y. C., & Corso,J. J. (2010). Propagating multi-class pixel labels throughout video frames. In *Proceedings of Western New York Image Processing Workshop*.

Cheng, H. -T., & Ahuja, N. (2012). Exploiting nonlocal spatiotemporal structure for video segmentation. In *CVPR*.

Cheung, V., Frey, B. J., & Jojic, N. (2005). Video epitomes. In *CVPR*.

Chockalingam, P., Pradeep, N., & Birchfield, S. (2009). Adaptive fragments-based tracking of non-rigid objects using level sets. In*ICCV*.

Chuang, Y., Agarwala, A., Curless, B., Salesin, D. H., & Szeliski, R. (2002). Video matting of complex scenes. *ACM Transactions on Graphics*, *21*(3), 243–248.

Criminisi, A., & Shotton, J. (Eds.). (2013). Decision forests in computer vision and medical image analysis. *Advances in computer vision and pattern recognition*. Berlin: Springer.

Criminisi, A., Sharp, T., Rother, C., & Perez, P. (2010). Geodesic image and video editing. *ACM Transactions on Graphics*, *29*(5), 1–15.

Fathi, A., Balcan, M., Ren, X., & Rehg, J. M. (2011). Combining self-training and active learning for video segmentation. In *BMVC*.

Grundmann, M., Kwatra, V., Han, M., & Essa, I. (2010). Efficienthierarchical graph-based video segmentation. In *CVPR*.

Kannan, A., Winn, J., & Rother, C. (2006). Clustering appearance and shape by learning jigsaws. In *NIPS*, (Vol. 19).

Kohli, P., & Torr, P.H.S. (2005). Efficiently solving dynamic markov random fields using graph cuts. In *ICCV*, (Vol. II, pp. 922–929).

Lee, K.C., Ho, J., Yang, M.H., & Kriegman, D. (2003). Video-based face recognition using probabilistic appearance manifolds. In *CVPR, Madison, WI*.

Lezama, J., Alahari, K., Sivic, J., & Laptev, I. (2011). Track to the future: Spatio-temporal video segmentation with long-range motioncues. In *CVPR*.

Nagaraja, N.S., Ochs, P., Liu, K., & Brox, T. (2012). Hierarchy of localized random forests for video annotation. In *Pattern Recognition (Proceedings of DAGM), Springer, LNCS*.

Saul, L. K., & Jordan, M. I. (1996). Exploiting tractable substructures in intractable networks. In *NIPS*.

Settles, B. (2012). Active learning literature survey. Technical report, Computer Sciences Technical Report 1648. University of Wisconsin Madison.

Shotton, J., Johnson, M., & Cipolla, R. (2008). Semantic texton forests for image categorization and segmentation. In *CVPR*.

Shotton, J., Winn, J., Rother, C., & Criminisi, A. (2006). Textonboost: Joint appearance, shape and context modeling formulti-class object recognition and segmentation. In *ECCV*.

Tsai, D., Flagg, M., & Rehg, J. M. (2010). Motion coherent tracking with multi-label mrf optimization. In *BMVC*.

Turner, R. E., Berkes, P., & Sahani, M. (2008). Two problems with variational expectation maximisation for time-series models. In *Workshop on Inference and Estimation in Probabilistic Time-SeriesModels*.

Vazquez-Reina, A., Avidan, S., Pfister, H., & Miller, E. (2010). Multiple hypothesis video segmentation from superpixel flows. In *ECCV*.

Vijayanarasimhan, S., & Grauman, K. (2012). Active frame selection forlabel propagation in videos. In *ECCV*.

Wang, T., & Collomosse, J. (2012). Progressive motion diffusion of labeling priors for coherent video segmentation. *IEEE Transactions on Multimedia*, *14*(2), 389–400.

Xu, C., Xiong, C., & Jason J. C. (2012). Streaminghierarchical video segmentation. In *ECCV*.