MoT - Mixture of Trees Probabilistic Graphical Model for Video Segmentation

Ignas Budvytis
ib255@cam.ac.uk
Vijay Badrinarayanan
vb292@cam.ac.uk
Roberto Cipolla
cipolla@eng.cam.ac.uk

Department of Engineering
University of Cambridge
Cambridge, UK

Abstract

We present a novel mixture of trees (MoT) graphical model for 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. Our time-series model explicitly captures the uncertainty in temporal linkage between adjacent frames which improves segmentation accuracy. We provide a variational inference scheme for this model to estimate super-pixel labels and their confidences in nearly realtime. The efficacy of our approach is demonstrated via quantitative comparisons on the challenging SegTrack joint segmentation and tracking dataset [23].

1 Introduction

Modelling frame to frame correlations is one of the most important components in a video model. It is a common practice in computer vision problems to establish mappings between frames via frame to frame optic flow algorithms [14],[15],[18] or long term point trajectories as in [19],[8]. However for tasks requiring semantic label propagation in video sequences, satisfactory results are not achieved: [13], [10], [2]. Poor performance can be attributed to a lack of robust occlusion handling, label drift caused by round-off errors, high cost of multi-label MAP inference or sparsity of robust mappings. These issues have led to the use of label inference over short overlapping time windows ([23]) as opposed to a full length video volume.

To address these issues, we have developed a novel super-pixel based mixture of trees (**MoT**) video model, motivated by the work of Budvytis et. al [9]. 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 also 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 [23] where only a MAP estimate is obtained.

The pixel label posteriors (confidences) we infer can be used to train a Random Forest clas-

http://dx.doi.org/10.5244/C.26.72

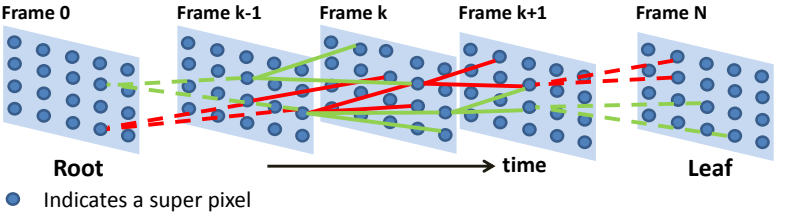

Figure 1: An illustration of the MoT model with two component temporal trees in the mixture. Each node is a super-pixel. Moving from root to leaf two temporal tree structures are visible. During inference, super-pixel label predictions from each tree is weighted by the posterior probability of the corresponding tree in the mixture.

sifier in a semi-supervised setting, as in [9]. 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 [21] .
To summarise, the contributions we make in this paper are as follows:

1. A new mixture of trees (MoT) probabilistic graphical model for video segmentation.
2. An efficient structured variational inference strategy for obtaining super-pixel labels and their confidences.

The remainder of this paper is organised as follows.
We present a comparative literature review in Sec. 2. Our proposed video model is explained in detail in Sec. 3. The inference strategy, we propose for segmentation is elaborated in Sec. 4. We then discuss our experiments and their results on the competitive SegTrack joint tracking and segmentation dataset [12] in Sec. 4. We bring out the advantages and the drawbacks of our approach in Sec. 5. We conclude in Sec. 6 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 [25],[19],[18],[15], especially as extensions of image super-pixellization to space-time super-pixellization. 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 distinctive feature of our algorithm is that it performs probabilistic inference, as opposed to the more commonly used MAP inference. This is with a view towards performing active learning for video segmentation (not addressed in this work).
Other unsupervised techniques are Video Epitomes [11] and Image Jigsaws [16]. 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,

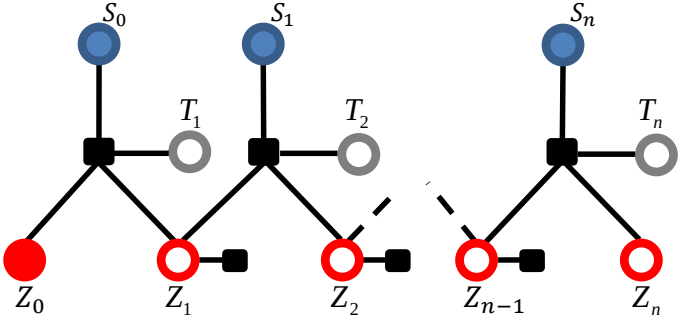

Figure 2: The factor graph for our proposed Mixture of Trees (MoT) model. Observed nodes are fully shaded. For each frame blue and red nodes represent a collection of super-pixels and their labels. The multi-node factors involve super-pixel, their labels and the mapping variable connecting labels in adjacent frames. All the label nodes have a unary factor. View in colour.

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 (no 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 [7] in a semi-supervised setting to establish correlations in non-successive video frames.

Semi-Supervised Segmentation: The label propagation method of Badrinarayanan et. al. [2] 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 via 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. [9]. 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 and improves segmentation quality.

Tsai et. al [23] 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 a single 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 [14] 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.

Segmentation by Classification: Decision tree architectures such as the popular Randomized Decision Forests [7] have gained popularity in *unstructured classification* tasks. The Semantic Texton Forests (STF) [22] is an example. In our MoT model, we employ a decision forest classifier for semi-supervised learning of super-pixel label unaries (likelihood).

In recent years, *structured classification* models such as conditional random field (CRF) models [5] have led the way in image segmentation problems. In practice, their main attraction arises from the ability to perform global optimisation or in finding a strong local minima

[5],[6]. 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) [17] or for N-D sequences [5]. Also, as noted by [3], performing MAP inference on large 3D volumes results in an uncontrollable work flow. Finally, multi-label MAP inference is computationally expensive [23], necessitating short overlapping time window based video segmentation.

3 Mixture of Trees (MoT) Video Model

We super-pixelize each image in the video sequence using the SLIC algorithm [1] into approximately 1000 super-pixels. Let $S_{i,j}$ denote super-pixel j at frame i, and $Z_{i,j}$ denote the 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 lying within a window $W_{i,j}$, placed around the center of $S_{i,j}$. Note that this implies 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 correponding temporal mapping variables respectively at frame i. $\Omega(i)$ denotes the number of super-pixels in frame i. Our proposed mixture of trees (MoT) probabilistic model for the video sequence factorises as follows:

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

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 Eqn. 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 [12]. In our experiments, we demonstrate that the simple measure in Eqn. 2 already provides us with competitive results (see Table 1).

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. We set it to a value of 0.95 in our experiments. The single node potential for the temporal

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

Input: Super-pixels $S_{0:n}$ (video), User labelled frame Z_0.
Output: Pixel label probabilities.
Intialisation
Set the initial value of μ to the value given in Sec. 4.
Set all unaries to uniform distributions.
Set all variational posteriors to uniform distributions.
Set max_iter = 50;
1a. Infer temporal mapping posteriors $Q\left(T_{i,j}\right)$ using Eqn. 7.
1b. for *i = 1* **to** *max_iter* **do**
  Infer $Q\left(Z_{i,j}\right)$ using Eqn. 9.
  Infer $Q\left(Z_{i,j}, Z_{i-1,T_{i,j}}\right)$ using Eqn. 10.
1c. Train the Random Forest with posteriors $Q\left(Z_{i,j}\right)$ as labels.
Set the super-pixel unaries to the predictions from the Random Forest. See Sec. 4.
1d. Perform step **1b** again.

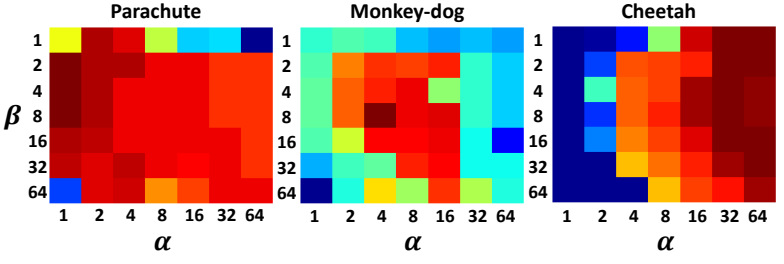

Figure 3: Sensitivity of pixel labelling accuracy to parameters α, β shown for three of the SegTrack [12] sequences (also see Sec. 4). In each heat map, deep blue hues represent very low accuracy and deep red represents high accuracy. It is clear from these maps that there is no common parameter value(s) which provides good performance for all sequences. In the parachute sequence, high accuracy is obtained for low values of α which corresponds to several contributing components in the mixture (highly *loopy* model). In the Monkey-dog and Cheetah sequences, the optimal α values are higher. This reduces the number of influential components in the mixture, making the model less *loopy*. Zoom-in and view in colour.

mapping variables $\Psi_t(.)$ is similar to a box prior and is defined as follows.

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

The super-pixel label unary factors $\Psi_u\left(Z_{i,j}\right)$ are defined in Sec. 4.
From Eqn.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. 1, 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 Eqn. 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 Eqn. 1 that computing the partition function $\mathcal{Z}(\mu)$ 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 [20]. The key idea in this inference scheme is to retain as much model structure in the approximate posterior as computationally tractable. This provides more robust inference as opposed to mean-field variational inference, where the joint posterior of the variables factorises into independent terms [24].
In our work, we assume the following form for the *approximate variational posterior* of the latent variables.

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

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

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

To maximise the above lower bound, which is a functional of the variational posterior and the model parameters, we employ calculus of variations [4]. This leads to the following fixed point equations for the approximate posteriors.

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

To compute the approximate super-pixel label marginals and pair-wise marginals required for the above equation we use variational message passing [4]. The variational message (vm) which super-pixel $S_{i,j}$ sends to its temporal neighbour $S_{i-1,T_{i,j}}$ is as follows.

$$vm_{Z_{i,j} \to Z_{i-1,T_{i,j}}}(Z_{i-1,T_{i,j}}) = \sum_{Z_{i,j}} \exp\Big[Q(T_{i,j}) \log\Big(\Psi_l(Z_{i,j}, Z_{i-1,T_{i,j}}|\mu)\Big)\Big] \prod_{n \in Ne(Z_{i,j}) \setminus Z_{i-1,T_{i,j}}} vm_{n \to Z_{i,j}}. \tag{8}$$

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

$$Q(Z_{i,j}) \propto \prod_{n \in Ne(Z_{i,j})} vm_{n \to Z_{i,j}} \tag{9}$$

$$Q(Z_{i,j}, Z_{i-1,T_{i,j}}) \propto \prod_{n \in Ne(Z_{i,j}) \setminus Z_{i-1,T_{i,j}}} vm_{n \to Z_{i,j}} \prod_{n \in Ne(Z_{i-1,T_{i,j}}) \setminus Z_{i,j}} vm_{n \to Z_{i-1,T_{i,j}}} \tag{10}$$

$$\times \exp\Big[Q(T_{i,j}) \log\Big(\Psi_l\Big(Z_{i,j}, Z_{i-1,T_{i,j}}|\mu\Big)\Big)\Big]. \tag{11}$$

In our experiments, we first set all the variational posteriors to a uniform distribution and compute $Q(T_{i,j})$ only once at the first iteration. Then Eqns. 9, 10 are iterated for a fixed

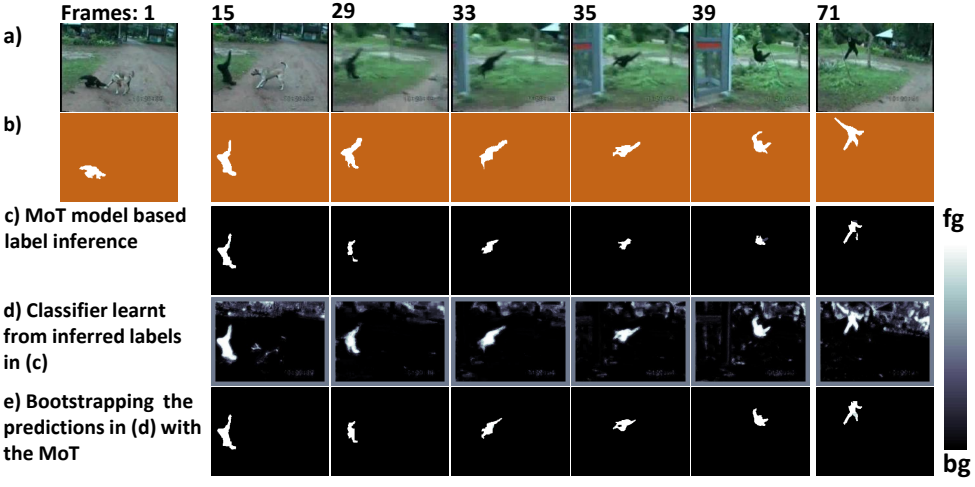

Figure 4: The first two rows show the image sequence Monkey-dog from the SegTrack dataset [23] and the corresponding ground truth. The segmentation algorithm in this sequence has to cope with fast shape changes, motion blur and overlap between foreground and background appearances. Row (c) is the inferred labels using the MoT time-series and with flat unaries. Row (d) are the Random Forest predictions when trained using the posteriors in row (c). Fusing these predictions with the MoT time-series results in an improved segmentation in row (e). In heat maps above bright white and dark black correspond to confident foreground and background labels respectively.

number of iterations (See Sec.4). A summary of the inference technique with a view to encourage implementation is given in Alg. 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 $Q(T_{i,j})^{\alpha}$ 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 in the temporal mixture of trees. This reduces the number of loopy cliques in the model.

As the MoT model is loopy by construction and the features used to make temporal linkages (Eqn. 2) can be arbitrarily weak, 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 Sec. 4, we discuss the effect of varying these parameters versus the accuracy of segmentation.

Semi-supervised Learning of Unaries. We follow the approach of [9] and learn a Random Forest Classifier using the inferred pixel labels. Note that all pixels in the super-pixel share the same label distribution after inference. We use the entropic information gain criterion (as in [22]) to learn the classifier. As the information entropy is evaluated on probability distributions, this criterion is directly suitable to incorporate marginal posteriors (soft labels) of pixels to learn the tree classifier.

To bootstrap the predictions from the learnt classifier into the MoT time-series model, we first compute the average of the classifier predicted label distributions for pixels within a super pixel. This *averaged distribution* acts as the unary $\Psi_u(Z_{i,j})$ for each super-pixel.

Video sequence	Properties		Performance comparison				
	Average object size	No of frames	Chockalingham et. al [12]	Tsai et. al [24]	Fathi et. al [14]	Budvytis et. al [9]	Ours
Parachute	3683	51	502	**235**	251	404	296
Girl	8160	21	1755	1304	1206	1705	**1200**
Monkey-dog	1440	71	683	563	598	736	**412**
Penguin	20028	42	6627	1705	**1367**	19310	1736
Bird-fall	495	30	454	**252**	342	468	508
Cheetah	1584	29	1217	1142	**711**	1501	855

Table 1: Quantitative evaluation on the SegTrack tracking and segmentation dataset [23]. 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 manually tuned parameters, outperforms other 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.

5 Experiments and Results

We evaluated the performance of our approach in a tracking and segmentation setting using the challenging SegTrack [23] dataset. This dataset with each frame ground truth consists of 6 sequences with fast motion, self-occlusion, small sized objects and deformable shape. In Table 1 we report our scores (number of pixel label errors per frame) using manually selected parameters and settings. Scores of some of the recent state of the art approaches are also reported.

To demonstrate that a single value of influential parameters α, β is insufficient to cover this entire dataset, we plot the accuracy for sample sequences over different values of the parameters. Note that we do not include any unaries while computing this score. Therefore, these plots gives us a indication of the performance of the MoT time-series alone. Using this plot, we select the parameter settings for each sequence which result in the highest accuracy. In comparison, Fathi et. al [14] 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. As seen in Table 1, this method works for a few sequences but fails for other sequences, perhaps when self-training collapses. Therefore, in this paper we do not adopt this approach to parameter setting.

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 Alg. 1. We can observe good qualitative performance of our algorithm in Fig. 5. 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 in Fig.5, the foreground object is tracked and segmented reasonably well. However, a small part of the background appearing in a few frames (Fig. 5, frames 21-26) lowers quantitative figures. 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 [22], 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

Figure 5: Qualitative results on the SegTrack dataset [23], [13]. 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,c,i) and small sized objects (f). The main failure case is (e) due to its small size. Zoom in and view in colour. See supplementary videos.

and number of features as in [22]. The key difference is that we use the inferred pixel label posteriors to train the RF.

Advantages and Drawbacks: The main advantages of our approach are enumerated 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.
2. We infer pixel-wise labels and their confidences (marginal posteriors). This is useful for semi-supervised and active learning systems.
3. In addition, we model uncertainty in the temporal links explicitly and this helps avoid having to use any off the shelf optical flow methods.
4. Our inference method is both computationally and memory wise efficient. For example, for the entire Penguin sequence our method requires 40MB of RAM and 1 minute of processing time. In contrast the method of Budvytis et. al [9] requires 1.5GB of RAM and 4 minutes when using one core of Intel Xeon 2.5GHZ CPU.

The main drawbacks and pointers to future work are.
1. The influential parameters α, β are manually set using grid-search. This was 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 must be replaced with more sophisticated features as in [14].

6 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. We demonstrated its efficacy on a very challenging video segmentation dataset. We look forward to exploiting this model in both interactive and multi-class video segmentation problems.

References

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. Slic superpixels. Technical report, EPFL Technical Report no. 149300, 2010.

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

[3] 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. ISBN 978-1-60558-726-4.

[4] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[5] Y. Boykov and M. P. Jolly. Interactive graph cuts for optimal boundary and region segmentation of objects in n-d images. In *ICCV*, 2001.

[6] Y. Boykov and O. Veksler & R. Zabih. Fast approximate energy minimization via graph cuts. In *ICCV*, 1999.

[7] L. Breiman. Random forests. In *Machine Learning*, pages 5–32, 2001.

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

[9] I. Budvytis, V. Badrinarayanan, and R. Cipolla. Semi-supervised video segmentation using tree structured graphical models. In *CVPR*, 2011.

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

[11] V. Cheung, B. J. Frey, and N. Jojic. Video epitomes. In *CVPR*, 2005.

[12] P. Chockalingam, N. Pradeep, and S. Birchfield. Adaptive fragments-based tracking of non-rigid objects using level sets. In *ICCV*, 2009.

[13] Y. Chuang, A. Agarwala, B. Curless, D. H. Salesin, and R. Szeliski. Video matting of complex scenes. *ACM SIGGRAPH*, 21, No. 3:243–248, 2002.

[14] A. Fathi, M. Balcan, X. Ren, and J. M. Rehg. Combining self training and active learning for video segmentation. In *BMVC*, 2011.

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

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

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

[18] K.C. Lee, J. Ho, M.H. Yang, and D. Kriegman. Video-based face recognition using probabilistic appearance manifolds. In *CVPR, Madison, Wisconsin*, 2003.

[19] J. Lezama, K. Alahari, J. Sivic, and I. Laptev. Track to the future: Spatio-temporal video segmentation with long-range motion cues. In *CVPR*, 2011.

[20] L. K. Saul and M. I. Jordan. Exploiting tractable substructures in intractable networks. In *NIPS*, 1996.

[21] B. Settles. Active learning literature survey. Technical report, Computer Sciences Technical Report 1648. University of Wisconsin Madison, 2010.

[22] J. Shotton, M. Johnson, and R. Cipolla. Semantic texton forests for image categorization and segmentation. In *CVPR*, 2008.

[23] D. Tsai, M. Flagg, and J. M. Rehg. Motion coherent tracking with multi-label mrf optimization. In *BMVC*, 2010.

[24] 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.

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