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

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 [6].

Motivation. It is a common practice in computer vision problems to establish mappings between frames via optic flow algorithms [4] or long term point trajectories. However for tasks requiring semantic label propagation in video sequences, satisfactory results are not achieved: [1]. 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 ([6]) 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 [3]. 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. The uncertainties in the temporal correlations are also inferred, unlike the joint label and motion optimisation method of [6] where only a MAP estimate is obtained.

Model. 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 $S_{i,j}$. 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:

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

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. 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:

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

Note that more sophisticated super-pixel match scores can also be substituted here as in [4]. The *label factor* $\Psi_l(.)$ is defined between the multinomial super-pixel label random variables as follows:

$$\Psi_l\left(Z_{i,j} = l, Z_{i-1,T_{i,j}} = m|\mu\right) \triangleq \mu\,(\text{if } l = m) \text{ or } 1-\mu\,(\text{if } l \neq m), \tag{3}$$

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

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

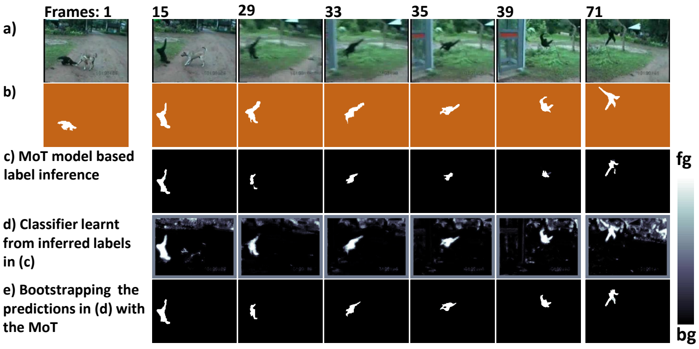
Frames: 1
15
29
33
35
39
71
a)
b)
c) MoT model based label inference
d) Classifier learnt from inferred labels in (c)
e) Bootstrapping the predictions in (d) with the MoT
fg
bg

Figure 1: The first two rows show the image sequence Monkey-dog from the SegTrack dataset [6] 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). Bright white and dark black correspond to confident foreground and background respectively.

The super-pixel label unary factors $\Psi_u\left(Z_{i,j}\right)$ are defined as output of Random Decision Forest Classifier (see Fig. 1 above and Sec. 4 in the paper). 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.

Inference. We use structured variational inference scheme [5] where we assume the following form for the *approximate variational posterior* of the latent variables.

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

The temporal mappings are assumed independent in the approximate posterior, however, the super-pixel latent labels do not factorise into independent terms, thereby maintaining *structure* in the posterior. The observed data log likelihood $\log\left(S_{0:n}|\mu\right)$ is lower bounded using the approximate posterior in Eqn. 5. To maximise the above lower bound we employ calculus of variations [2]. Finally, to compute the approximate super-pixel label and required pair-wise marginals we use variational message passing [2].

Evaluation. We evaluated the performance of our approach in a tracking and segmentation setting using the challenging SegTrack [6] dataset. Fig. 1 illustrates qualitative results of different stages of our algorithm on a Monkey-dog sequence from SegTrack. A detailed qualitavive and quantitative comparisons with some of the recent state of the art approaches are provided in the paper.

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

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

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

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

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

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