Label propagation in complex video sequences using semi-supervised learning

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

Department of Engineering,
Cambridge University

Motivation. Fast and efficient discriminative classifiers like Random Forests have shown promising results for video segmentation [4]. However, training these classifiers require copious quantities of labelled video data, which unfortunately is extremely strenuous to obtain by hand labelling. To reduce the burden of hand labelling, label propagation methods for semi-supervised learning, like [1],[5], exploit the structure in the distribution of data points to infer unknown labels from the a labelled points. Label propagation in video sequences for training multi-class classifiers designed for video segmentation has been recently proposed in [1]. Given hand-labelled start and end frames of a video sequence, the goal is to propagate labels throughout the rest of the video sequence. To this end, [1] suggested a coupled Bayes net for joint modelling of the image sequence and their pixel-wise labels. A simple variational EM strategy is employed to infer the most probable class label for the pixels in the video. This scheme provides high quality labels for 2-3 second videos. However, their scheme is afflicted by *time-assymetry* in labelling and an inability to tackle occlusions, which degrades performance over greater lengths.

Contribution. We propose a novel directed graphical model for label propagation (see Figure 1) in lengthy ($\approx$ 30 seconds) and complex video sequences. Given hand-labelled start and end frames of a video sequence, a variational EM based inference strategy propagates either one of several class labels or assigns an unknown class (void) label to each pixel in the video. These labels are used to train a multi-class classifier. The pixel labels estimated by this classifier are injected back into the Bayesian network for another iteration of label inference. The novel aspect of this iterative scheme, as compared to a recent approach [1], is its ability to handle occlusions. This is attributed to a hybrid of generative propagation and discriminative classification in a *pseudo time-symmetric* video model. The end result is a conservative labelling of the video (see Figure 2); large parts of the static scene are labelled into known classes, and a void label is assigned to moving objects and remaining parts of the static scene. These labels can be used as ground truth data to learn the static parts of a scene from videos or for semantic video segmentation.

Model, inference and learning. In the model (see Figure 1) $I_{0:n}$ are

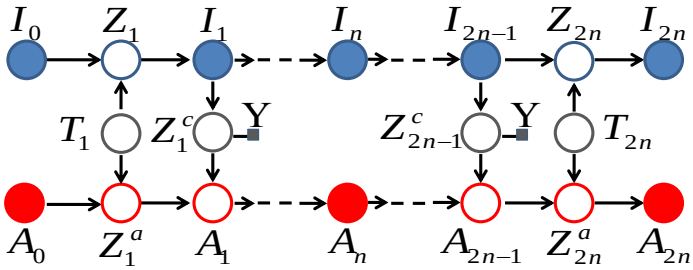

Figure 1: Proposed hybrid model (PHM) for label propagation. Shaded nodes represent observed/clamped variables. The Markov chains have time-reversed frames $I_{n+1:2n}$ placed to get a pseudo time-symmetric video model.

the observed sequence of images and $\{I_k = I_{2n-k}\}_{k=n+1}^{2n}$ is the sequence of images in time reversed order, so arranged to achieve time-symmetric label propagation. Z_k is a *latent colour image* consisting of "overlapping latent colour image patches" are used to "explain" image I_k. Correspondingly, Z_k^a is a *corresponding latent labelled image* consisting of "overlapping latent labelled patches" used to "explain" A_k, which is is an image sized "averaging variable" representing the annotation for image I_k. Each coordinate of A_k captures a "local uncertainty" in the labels for image I_k. Z_k^c is a latent labelled image obtained as a result of feeding the observed image I_k through a "black box" classifier controlled by parameter

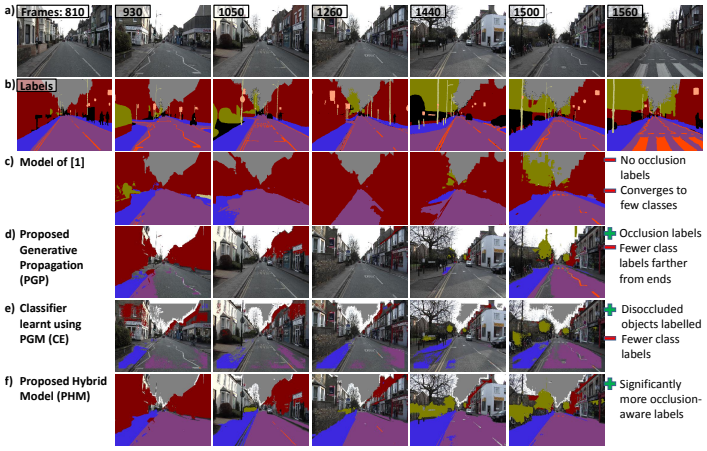

Figure 2: Qualitative comparisons of [1], proposed generative model (PGP) and proposed hybrid model (PHM) on sequence S1 (see paper).

Υ. Finally, T_k is the set of "patch mapping" variables, which in broad terms balances the contribution of the generative and discriminative components towards modelling the observed data [3].

The proposed variational inference proceeds by lower bounding the visible data $\{I_{0:2n}, A_0, A_n, A_{2n}\}$ as shown below.

$$\log p(I_{0:2n}, A_0, A_n, A_{2n}) \geq \int_{\Theta} q(Z_{1:2n}, Z^a_{1:2n}, Z^c_{1:2n}, A_{0:2n}, T_{1:2n}) \times \log \frac{p(Z_{1:2n}, Z^a_{1:2n}, Z^c_{1:2n-1}, A_{0:2n}, T_{1:2n}, I_{0:2n})}{q(Z_{1:2n}, Z^a_{1:2n}, Z^c_{1:2n}, A_{0:2n}, T_{1:2n})}, \quad (1)$$

where $q(.)$ is an auxiliary distribution. The form of $q(.)$ is;

$$q(Z_{1:2n}, Z^a_{1:2n}, A_{1:2n}, Z^c_{1:2n}, T_{1:2n}) = \prod_{k=1}^{2n} q(T_k)\delta(Z_k - Z_k^*)\delta(Z_k^a - Z_k^{a*})\delta(A_k - A_k^*)\delta(Z_k^c - Z_k^{c*}). \quad (2)$$

The above computationally tractable form [2] allows us to alternately approximate a posterior distribution over the mapping variables and a MAP estimate over the remaining ones in the process of maximizing the above bound. Learning implies estimating the the classifier internal parameter Υ. In principle, the inferred values of the $Z^c_{1:2n}$ act as the "desired output" for training a multi-class classifier (random decision forest) on the image sequence $I_{0:2n}$. The pixel labels estimated by this classifier are injected back into the network to prolong occlusion aware label propagation. **Results.** The benefits of our proposed hybrid model (PHM) are shown in the qualitative comparison in Figure 2. PHM provides occlusion-aware labelling (unseen objects are not labelled) which is not possible to achieve by [1] as the latter provides labels to all the pixels. Also the use of classifier allows PHM to get many more pixels labelled that PGP (PHM with classifier turned of) as classifier can recover the labels of an object which has been occluded and then dis-occluded. The quantitative comparison between algorithms is given in the paper.

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

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

[3] J. A. Lasserre, C. M. Bishop, and T. P. Minka. Principled hybrids of generative and discriminative models. In *CVPR, NewYork*, 2006.

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

[5] X. Zhu and Z. Ghahramani. Learning from labeled and unlabeled data with label propagation. Technical report, CMU-CALD-02-107, CMU., 2002.