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