A tracking-based method for video and volume annotation with sparse point supervision

Laurent Lejeune, Jan Grossrieder, and Raphael Sznitman

Ophthalmic Technology Laboratory, University of Bern, Switzerland

1 Introduction

With the goal of strongly reducing the burden of producing pixel-wise annotations for a variety of objects, this work presents a novel framework to produce segmentations in video and volumetric image data using 2D point supervision. In this context, only a single 2D location within an object of interest is provided on each image of the sequence while we make no strict assumption on the object of interest (e.g. shape, color, motion, etc.).

2 Method

The first component of our method estimates a probability map of the object of interest by using the provided image sequence and associated 2D locations. This is achieved by learning a classifier on superpixels in a transductive fashion [1]. To do this, we compute a_t^n, the appearance descriptor of superpixel s_t^n, by means of a U-Net configured in an auto-encoder fashion and use these in a bagging classifier. In the second component, we hypothesise that by tracking superpixels over the entire volume, a complete segmentation of the object can be inferred. That is, we consider each region specified by a provided 2D location to be an individual target, that could potentially depict different parts of the same object. Given the above local object model, we provide a global strategy to infer an accurate segmentation of the object across all frames. In particular, we define $\boldsymbol{Y} = \{Y_t^n | \forall (t,n)\}$ as the set of all Y labels, with y_t^n the binary label of superpixel s_t^n. We then define our segmentation problem as finding the optimal labeling y^* through the Maximum a Posteriori (MAP) optimization:

$$y^* = \arg\max_{y \in \mathcal{Y}} P(\boldsymbol{Y} = \boldsymbol{y} | \boldsymbol{a}, \boldsymbol{g}) = \arg\max_{y \in \mathcal{Y}} \prod_{m,n,t} P(Y_t^n | \boldsymbol{a}, \boldsymbol{g}) P(Y_t^n | a_{t-1}^m) P(Y_t^n | a_t, g_t). \tag{1}$$

With $\boldsymbol{a}$, the grouping variable that corresponds to the a_t^n, and $\boldsymbol{g}$ the set of provided 2D locations. The three terms of Eq. 1 correspond respectively to (1) the transductive object appearance model mentioned above, (2) a similarity model between two superpixels in successive frames, so to describe how frame-to-frame probabilities propagate, and (3) the likelihood that a given superpixel s_t^n is visually similar to the one selected by the 2D location g_t [2]. In practice, we solve problem 1 using the efficient and optimal K-shortest paths algorithm[3].

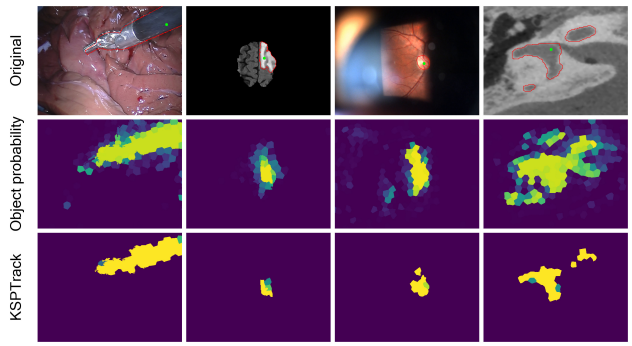

Fig. 1: (Top row): Original images from different datasets. Ground truth contour of the structure of interest is depicted in red and the supervised 2D locations are shown in green. (Middle row): Probability estimates of the object given by our classifier. (Bottom row): Pixel-wise binary segmentations after optimization.

3 Results and Discussion

Fig. 1 depicts example segmentations on four datasets of different nature. From left to right: Video sequences showing a surgical instrument, T2-weighted MRI volumes showing a brain tumor, slitlamp videos showing the optic disk, and CT scans of the inner ear focused on the cochlea. Quantitatively, we report average F1 scores of 0.81, 0.76, 0.78, and 0.66 for the four described types of datasets, each consisting of four sequences. Further experiments showed that our segmentations, when used as ground truth to train supervised machine learning algorithms, induce decreases in performance of about 15% compared to mouse-based segmentations.

Overall our approach appears to effectively produces segmentations for a wide range of image modalities using minimal supervision. By combining more sophisticated user input methods, such as gaze-trackers, our method could provide a reliable way to extract annotations passively from experts or from crowds. Future works will focus on extending this approach to problems with multiple objects of interest in the field of view.

References

1. Mordelet, F., Vert, J.P.: A bagging svm to learn from positive and unlabeled examples. Pattern Recogn. Lett. **37** (February 2014) 201–209
2. Sugiyama, M.: Local fisher discriminant analysis for supervised dimensionality reduction. In: International Conference on Machine Learning. (2006) 905–912
3. Suurballe, J.W.: Disjoint paths in a network. Networks **4**(2) (1974) 125–145