Eye tracking assisted extraction of attentionally important objects from videos

S. Karthikeyan[1], Thuyen Ngo[1], Miguel Eckstein[2], B.S. Manjunath[1]
[1]Department of Electrical and Computer Engineering, UC Santa Barbara. [2]Department of Psychology and Brain Sciences, UC Santa Barbara.

Eye tracking data in a free-viewing task is biased towards high level semantics in static and dynamic scenes [2]. Therefore, visual attention can provide a robust prior to assist multiple object segmentation problem in video sequences. Recent advancements in eye tracking technology has opened up avenues for large-scale collection of eye tracking data from multiple subjects without affecting the experience of the viewer. Multimedia content is typically viewed by a large number of people and collecting eye tracking data from a small fraction of the viewers can provide weak supervision to guide object segmentation.

The aim of the proposed approach is to utilize eye tracking data in conjunction with visual information from video frames to extract objects which attract visual attention. We propose a two-step approach. First we process raw eye tracking data and obtain dominant visual tracks which are consistent across multiple subjects. These visual tracks help localize object search in video frames. Next, these localized object proposals are connected using a novel multiple object extraction framework which is designed to simultaneously ensure temporally consistent and spatially distinct objects. An overview of the proposed approach is shown in Fig. 1.

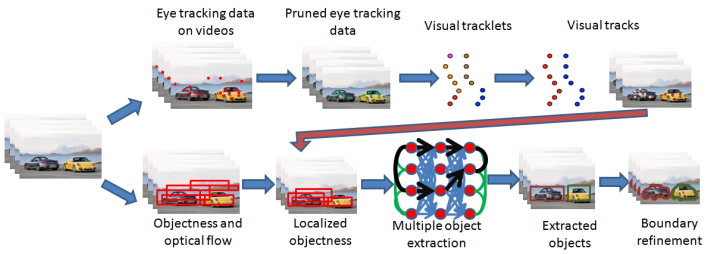

Figure 1: Block diagram of the proposed approach to extract multiple objects from videos using eye tracking prior. The top row indicates the eye tracking processing stage. The bottom row is the multiple object extraction framework guided by the visual tracks.

Eye movement data in dynamic scenes consists of fixations, saccades and smooth pursuit. The fixations and smooth pursuit represent the information gathering stage and saccades represent transitions between fixations. Typically, fixations are present in video regions representing static objects and smooth pursuit is observed when a subject tracks a moving object. Saccades typically do not lie on objects in a video sequence as they indicate transitions between fixations. In our work, the fixation and smooth pursuit samples are utilized in a two step hierarchical association process. First, the eye tracking samples from all the subjects over an entire video sequence are associated in a conservative manner using 3-D mean shift clustering. This gives us visual tracklets representing eye tracking data over small potential temporal object segments through the video sequence. In the next step these tracklets are associated using Hungarian algorithm to eventually represent dominant visual tracks. An example of dominant visual tracks in a video sequence is shown in Fig. 2

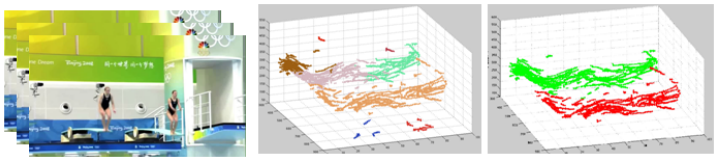

Figure 2: An example of visual tracklets (center) and visual tracks (right) on a video sequence (left) shown in 3-D. The visual tracklets are associated using Hungarian algorithm to obtain the tracks. The horizontal axis in the visual tracks and tracklets represents image frames.

The visual tracks coarsely localize attentionally important objects in a video sequence and thereby reduce the search space for these objects in the scene. Specifically, visual tracks provide the following two critical pieces of information, a) Number of visually salient objects in the scene and b) Coarse spatial localization of the objects of interest. We propose a novel principled framework to extract important objects of interest guided by the visual tracks. As visual tracks provide coarse priors on the object locations, we extract visual track localized bounding box based objectness proposals [1]. Each bounding box is assigned a unary score using objectness which indicates the probability of the bounding box enclosing an object. We refine this unary score to reflect motion information by adding an additional term which measures optical flow magnitude contrast within and outside the bounding box.

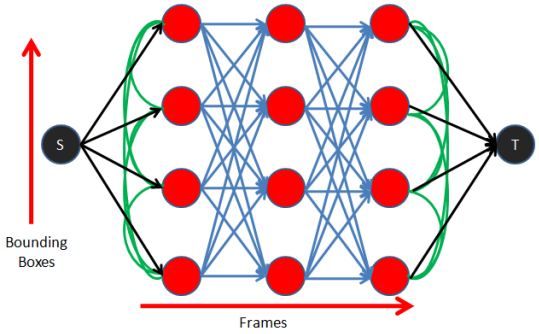

Figure 3: The spatio-temporal graph to extract multiple objects is highlighted here. The bounding boxes in each frame are shown as red circles. The temporal costs shown as blue directed edges indicate inter-frame costs to connect a path through two bounding boxes in successive frames. The intra-frame spatial costs are indicated as green undirected edges. They penalize extraction of the same object in multiple paths. Best viewed in color.

In addition we also define pairwise costs across bounding box pairs in successive frames. This score is determined from spatial overlap distance and color histogram distance between the bounding boxes in the two frames. The overall temporal score across two bounding boxes in successive frames is a weighted combination of the unary and pairwise scores. Now given a set of bounding boxes in every frame, and the number of visual tracks k, we want to extract k distinct objects from the video sequence. For this purpose, we construct a mixed graph as shown in Fig. 3. The nodes of the graph represents the bounding boxes. The directed edges (across successive frames) shown in blue has weights denoting the temporal cost. As the objectness metric extracts multiple bounding boxes around an object of interest, it is possible to extract the same object in several paths. In order to mitigate this we introduce spatial costs quantifying spatial overlap across candidate boxes within a frame. The spatial cost denoted by green undirected edges in Fig. 3 penalizes the extraction of overlapping objects in multiple paths through the graph. The optimal paths through the graph which minimize the overall spatio-temporal cost are obtained by solving a binary integer linear program.

This approach is evaluated using a eye tracking dataset which we collected using Eyelink 1000 eye tracker. The eye tracking data was obtained from 20 subjects over 20 standard sequences for video object segmentation. On this dataset, the proposed approach outperforms single and multiple object segmentation algorithms in videos which do not utilize eye tracking data, as well as state-of-the-art eye fixation based object segmentation algorithm.

[1] Bogdan Alexe, Thomas Deselaers, and Vittorio Ferrari. Measuring the objectness of image windows. *TPAMI*, 2012.

[2] S Karthikeyan, Vignesh Jagadeesh, Renuka Shenoy, Miguel Eckstein, and B.S. Manjunath. From where and how to what we see. In *IEEE ICCV*, 2013.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage. This work was supported in part by the following awards/grants: Institute for Collaborative Biotechnologies through grant W911NF-09-0001 from the U.S. Army Research Office and US Office of Naval Research N00014-12-1-0503