Eye tracking assisted extraction of attentionally important objects from videos

S. Karthikeyan*, Thuyen Ngo*, Miguel Eckstein‡, B.S. Manjunath*
*Department of Electrical and Computer Engineering
‡Department of Psychological and Brain Sciences
University of California Santa Barbara
*{karthikeyan, thuyen, manj}@ece.ucsb.edu , ‡eckstein@psych.ucsb.edu

Abstract

Visual attention is a crucial indicator of the relative importance of objects in visual scenes to human viewers. In this paper, we propose an algorithm to extract objects which attract visual attention from videos. As human attention is naturally biased towards high level semantic objects in visual scenes, this information can be valuable to extract salient objects. The proposed algorithm extracts dominant visual tracks using eye tracking data from multiple subjects on a video sequence by a combination of mean-shift clustering and Hungarian algorithm. These visual tracks guide a generic object search algorithm to get candidate object locations and extents in every frame. Further, we propose a novel multiple object extraction algorithm by constructing a spatio-temporal mixed graph over object candidates. Bounding box based object extraction inference is performed using binary linear integer programming on a cost function defined over the graph. Finally, the object boundaries are refined using grabcut segmentation. The proposed technique outperforms state-of-the-art video segmentation using eye tracking prior and obtains favorable object extraction over algorithms which do not utilize eye tracking data.

1. Introduction

Object extraction in videos is a challenging problem in computer vision. Automated extraction of objects in a video sequence can benefit several applications related to annotation, compression, summarization, search and retrieval. A critical bottleneck in object extraction is defining the importance of objects in a video sequence. Several works in object extraction from videos have focussed on utilizing motion to determine importance of objects. These methods typically aim to extract a single dominant object in the scene, determined by motion. In [27] Lee et al. identify important motion segments representing an object and extrapolate the object of interest throughout the video frames. In Ma et al. [28], important objects are segmented by connecting the extracted object candidates from all video frames using mutual exclusiveness constraints. Also, in [35] Ramakanth et al. utilize video seams to segment moving objects effectively. Recently [47], Zhang et al. proposed a framework to extract objects using objectness [12] and optical flow proposals and segment the key object by dynamic programming on a directed acyclic graph. They also utilize a warping technique to ensure robustness to broken object segments. All the aforementioned methods utilize motion to define the importance of objects and can extract only a single object of interest from a video sequence. However, motion may not be a good metric to determine importance of objects in videos. For example in a video sequence where two subjects are having a conversation, the motion cues might be misleading. We note that salient objects in a scene can be better understood by visual attention regions. Therefore, in this work we investigate the utility of eye tracking to extract multiple interesting objects in a scene.

Figure 1. A simple illustration of the proposed problem. Given a video sequence (left), we collect eye tracking data in the sequence from multiple subjects (center) and utilize this information to extract visually salient objects (right). Best viewed in color.

Eye movements have been shown to reflect a combination of influences of low level image properties, the observer's task, interest, and goals [18, 40, 11, 10, 11, 26]. In a free viewing task, eye movements are biased towards high level semantics [22, 29] in static and dynamic scenes. Therefore, visual attention can provide a robust prior to assist multiple object extraction in video sequences. Recent advancements in eye tracking technology have opened up avenues to collect data without affecting the experience of the viewer. State-of-the-art eye trackers are affordable [2] and this has enabled large-scale collection of eye tracking data from multiple subjects. Multimedia content is typi-

cally 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 extraction of important objects. Therefore, given a video sequence and eye tracking data from multiple subjects in a free viewing task, the objective is to extract relevant objects of interest which attract visual attention. A simple visual illustration of the proposed work is shown in Fig. 1. Recently, there has been active interest in eye tracking assisted computer vision algorithms. An overview of the literature in this area is provided below.

Related Work

Human inspired visual attention modeling [21, 17, 14, 22, 6, 7, 8, 24] has been a well-researched topic in over a decade. Recently there has been significant interest [30, 36, 43, 33, 42, 29, 25, 31, 44, 45, 46, 38] in eye tracking enhanced computer vision. Relevant to our work, Mishra et al. [30] proposed a segmentation using fixation approach which segments objects of interest given a single fixation point. They convert the image to polar coordinate space and graph cut segmentation in this space corresponds to object contour in the image domain. The approach was further extended using optical flow to segment a single object around a fixation point in a video sequence. The primary limitation of [30] is that they use a single fixation point and assume the fixation point is completely inside the object of interest. However, the assumption can be limiting as there is calibration error in real eye tracking data especially when we have to extract small objects. Additionally, [36, 43] have proposed image segmentation algorithms using multiple fixations in order to overcome some of the limitations of [30]. Recently, Papadopoulos et al. [33] explored an interesting problem of weakly annotating objects using eye tracking data to train object class detectors. The eye tracking annotations are used in the training phase to localize object bounding boxes which help train a deformable part model [13] based detector. The final detection performance is considerably lower than perfect ground truth annotations, however these annotations are obtained in about a sixth of the time required to hand annotate the bounding boxes which is encouraging. In [25] we propose a technique to extract face and text semantic priors using eye tracking data from multiple subjects and use this to enhance state-of-the-art object detectors. The algorithm is designed for images and is targeted for only face and text categories. Utility of eye tracking data in action recognition techniques [42, 29] have also shown promise. We note that the works in [33, 25, 36, 43, 45] are designed for object localization in images and in the proposed work we deal with object extraction in videos and therefore are not directly comparable as eye tracking data properties differ in images compared to videos. Additionally, eye tracking provides a platform to quickly annotate a large number of video frames which is of greater practical use compared to image annotation as manually annotating videos is a far more tedious task.

Therefore, in this work we propose an eye tracking assisted object extraction framework which is not restricted to specific object categories. The contributions of the proposed approach are as follows.

- A method to localize visual tracks from eye tracking data by solving a linear assignment problem, which coarsely corresponds to object locations in video sequences
- A novel object extraction framework guided by visual tracks, which extracts multiple objects in a spatio-temporal mixed graph by solving a binary integer linear program
- A novel eye tracking dataset on standard video sequences

This work is organized as follows. In Section 2 we introduce the eye tracking dataset. Our algorithm to extract visual tracks from eye tracking data is described in Section 3. The novel multiple object extraction framework is also presented in this section following which in Section 4 we demonstrate the results of the proposed approach. Finally the discussion, conclusions and future work are discussed in Section 5.

2. Eye tracking dataset on videos

We collected an eye tracking dataset on videos using Eyelink 1000 eye tracking device [1]. The videos were viewed by 30 subjects (between ages 21 and 35). The viewers sat 3 feet away from a 27 inch screen. The subjects were informed that it was a free viewing experiment and the data was collected without any apriori bias. The users rest their head on a chin rest and the eye position is sampled by the eye tracker at 500 samples per second. Eye tracking data in videos typically consists of fixations, saccades and smooth pursuit. The information gathering stage (while observing objects) is encoded in the fixations and smooth pursuit. The saccades represent attention shift from one fixation to another. The video dataset consists of 20 videos collected from SegTrack [41], GaTech [15] and Xiph.org [3] datasets. Each video in the dataset consists of 1 to 4 dominant objects. The depicted scenes are obtained from static and moving cameras with static and moving objects of interest. Fig. 2 highlights example video frames and the corresponding eye tracking data (excluding saccades) over multiple subjects from the dataset. We notice that the eye tracking data in individual frames is biased towards important semantic objects and coarsely localizes them, including objects without significant motion (faces in the last example in Fig. 2). Therefore, this information can be a useful prior to extract objects from video sequences. This dataset can be downloaded from http://vision.ece.ucsb.edu/~karthikeyan/videoObjectEyeTrackingDataset/.

Figure 2. Shows frames from four example videos in our dataset followed by the corresponding eye tracking data in the bottom row. We note that the dataset consists of single and multiple stationary and moving objects with moving and stationary backgrounds. The eye tracking data in individual frames typically lies on high level semantic objects and coarsely localizes it. Especially in the last (bottom) video sequence, the eye movement regions corresponds to faces where motion information cannot identify important objects in the scene. Therefore, this information is extremely useful to extract salient objects from the video sequences.

3. Proposed approach to extract objects from videos using eye tracking prior

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. As eye tracking data is primarily biased towards objects, we utilize it as a prior to guide object search. 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. 3. The top row indicates the eye tracking processing steps to extract dominant visual tracks from eye movement data from multiple subjects. The bottom row describes the visual track guided object localization and the multiple object extraction framework on a mixed graph. Finally, the object boundaries are refined by segmentation using bounding box prior. The following sections provide a detailed description of the different modules which comprise the proposed framework.

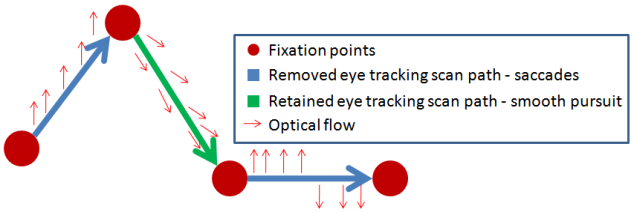

Figure 4. Illustrates that the scanpaths in the direction of optical flow, shown in green, represent smooth pursuit and should be utilized along with the fixations (red circles) for object localization. Scanpaths not in the direction of optical flow indicate attention shift from one object to another (saccades) and can be removed. Best viewed in color.

3.1. Eye tracking data processing to obtain dominant visual tracks

In order to extract dominant visual tracks from eye tracking data, we first introduce a simple pruning step to remove eye tracking data which has low probability of lying on objects. Eye movement data in dynamic scenes is biased towards high level semantic objects as described in Section 1 and 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. The eye tracker localizes fixations accurately, but does not directly distinguish between smooth pursuit and saccades and labels them as a scan path. As smooth pursuit eye tracking samples lie on objects, we first propose a simple technique to identify them. We utilize optical flow [39] to determine the nature of the scan paths. If the scan path lies in the direction of optical flow, it is labelled as smooth pursuit, otherwise it is classified as a saccade. This is illustrated in Fig. 4. We utilize fixations and smooth pursuit for further processing.

These pruned eye tracking samples have a higher probability of lying on objects in the videos than raw eye tracking samples. In the next stage, we associate these eye tracking samples to extract dominant visual tracks which coarsely correspond to objects of interest in a video sequence. This

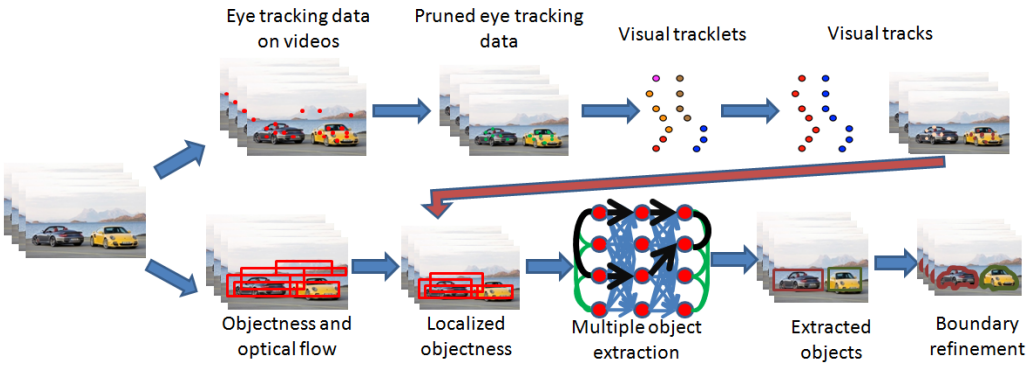

Figure 3. Block diagram of the proposed approach to extract multiple objects from videos using eye tracking prior. The top row indicates the eye tracking processing stage. The bottom row is the multiple object extraction framework guided by the visual tracks. Best viewed in color.

is achieved by a two step hierarchical association process similar to [19, 23]. 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. Eye tracking samples are normalized to have standard deviation one in all dimensions and clustered using a flat kernel with unity bandwidth ensuring invariance to video resolution. 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 to eventually represent dominant visual tracks.

Let the set of all tracklets be denoted by $\mathcal{T} = \{T_1, T_2 ... T_N\}$, where T_i denotes an individual tracklet. Similarly, the set of all tracks is denoted as $\mathcal{S} = \{S_1, S_2 ... S_M\}$. A linkage probablity $P_{link}(T_q|T_p)$ is defined between every pair of tracklets p and q. We encourage linkages between tracklets within close spatio-temporal proximity and visual characteristics by defining it as a product of two terms $P_{link}(T_q|T_p) = P_m(T_q|T_p)P_{app}(T_q|T_p)$. Here, we use a linear model to predict the position of the head of one tracklet from the tail of another. The error e in predicting the head of tracklet T_p from the tail of T_q is used to calculate the motion affinity using gaussian distribution, $P_m(T_q|T_p) = \mathcal{N}(e; 0, \sigma_e^2 \mathbf{I}_{2\times 2})$. The appearance affinity is calculated using histogram distance over the video frame pixels corresponding to the eye tracking samples. The histogram distance (d) between the head of T_p and tail of T_q is computed using χ^2 metric. The appearance affinity is calculated by an exponential distribution evaluated at d, $P_{app}(T_q|T_p) = \lambda \exp(-\lambda d)$.

Here we assume the likelihoods of the input tracklets are conditionally independent given $\mathcal{S}$ and the tracks $\{S_l\}$ are independent of each other. Now, the association term is decomposed as

$$\begin{aligned} \mathcal{P}(\mathcal{S}|\mathcal{T}) &\propto \mathcal{P}(\mathcal{T}|\mathcal{S})\mathcal{P}(\mathcal{S}) \\ &= \prod_{T_k \in \mathcal{T}} \mathcal{P}(T_k|\mathcal{S}) \prod_{S_l \in \mathcal{S}} \mathcal{P}(S_l) \end{aligned} \quad (1)$$

We also assume that a larger tracklet by size has a higher probability of being a true positive, i.e., corresponding to observing an object of interest. Therefore, the true positive likelihood of a tracklet is defined as $\mathcal{P}(T_k|\mathcal{S}) = \frac{\phi^{|T_k|}}{\sum_k \phi^{|T_k|}}$, where $|T_k| \in (0, 1]$ is the fraction of eye tracking samples in T_k.

The tracklet association priors in (1) are modeled as Markov Chains.

$$\mathcal{P}(S_l) = \mathcal{P}_{link}(T_{k_1}|T_{k_0}) ... \mathcal{P}_{link}(T_{k_{p_l}}|T_{k_{p_l-1}}) \quad (2)$$

where p_l refers to the number of tracklets associated to form the track S_l. Thus, the association prior is a product of transition terms representing linkage probabilities between tracklets.

As we need to maximize (1), first we convert it into a cost function by taking negative logarithms. The cost described in (1) can be optimized by the Hungarian algorithm similar to the one proposed in [19]. In brief, to associate n tracklets an $n \times n$ cost matrix C_{link} is built with the non-diagonal entries denoting the tracklet association costs. Another $n \times n$ cost matrix obtained from the true positive probabilities in the diagonal entries is concatenated with C_{link} to generate a $n \times 2n$ matrix. The optimal tracklet associations which minimize the cost globally is obtained by the Hungarian assignment on this cost matrix. Therefore, the joint cost matrix $\mathbf{C}_J$ of dimensions $n \times 2n$ is expressed as

$$\mathbf{C}_J(p,q) = \begin{cases} -\ln \mathcal{P}_{link}(T_q|T_p) - \frac{1}{2}[\ln \mathcal{P}(T_p|\mathcal{S}) + \ln \mathcal{P}(T_q|\mathcal{S})] \\ \qquad \text{if } p, q \leq n \text{ and } p \neq q \\ -\ln \mathcal{P}(T_p|\mathcal{S}) \qquad \text{if } p + n = q \\ +\infty \qquad \text{otherwise} \end{cases} \quad (3)$$

The optimal tracks are obtained by the Hungarian algorithm on $\mathbf{C}_J$ which assigns every row to a unique column. Finally, tracks containing above $> 20\%$ of the eye tracking samples are classified as dominant visual tracks. An example of visual tracklets and tracks on a video sequence is shown in Fig. 5. We note that the parameters were chosen empirically ($\lambda = 1, \sigma_e = 0.2$) to give good visual results, as we do not have ground truth for visual tracks.

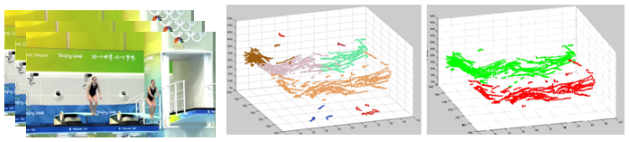

Figure 5. An example of visual tracklets (center) and visual tracks (right) on a video sequence (left) shown in 3-D. The visual tracklets are associated using Hungarian algorithm to obtain the tracks. The horizontal axis in the visual tracks and tracklets represents image frames. Best viewed in color.

3.2. Visual track guided object extraction from videos

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

- Number of visually salient objects in the scene
- Coarse spatial localization of the objects of interest

In this section 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 [4]. For every visual track, we compute the spatial standard deviation $\sigma = (\sigma_x, \sigma_y)$, which is indicative of the size of the object. Therefore, in every frame, we compute the mean position of visual track and search for objects within 5σ around the mean. Additionally, we only retain object proposals which contain more than 50% of the visual tracks samples in the frame. We notice that objectness [4] provides several overlapping bounding boxes around an object of interest. Each bounding box is assigned a score (objectness) which indicates the probability of the bounding box enclosing an object. We refine this score to reflect motion information by adding an additional term which measures optical flow magnitude contrast [39] within and outside the bounding box. Let the optical flow magnitude average inside a bounding box i and frame f be O^{if}_{in} and outside it be O^{if}_{out}. Then, the optical flow score is measured as $S^{if}_{opt} = 1 - e^{-\frac{(O^{if}_{in} - O^{if}_{out})^2}{\tau_{opt}}}$. The overall unary scores which combines objectness and optical flow score for bounding box i in frame f is a linear combination of individual scores and is given by $S^{if}_{unary} = S^{if}_{obj} + \alpha S^{if}_{opt}$. In this work, we select 35 object proposals within every frame which have the highest S_{unary} as our candidate objects.

Now given a set of bounding boxes in every frame, and the number of objects k (number of visual tracks), we want to extract k distinct objects from the video sequence. Each box has a unary score indicated by S_{unary}. 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. Let b^i_f and b^j_{f+1} represent two bounding boxes in successive frames f and $f+1$, then the pairwise score is represented as $S^{ijf}_{pair} = S^{ijf}_{overlap} + \beta S^{ijf}_{color}$, where $S^{ijf}_{overlap} = \frac{\text{Area}(b^i_f \cap b^j_{f+1})}{\text{Area}(b^i_f \cup b^j_{f+1})}$ and $S^{ijf}_{color} = e^{-\frac{\chi^2(h^i_f, h^j_{f+1})}{\tau_{color}}}$, where h^i_f and h^j_{f+1} are the color histograms of the bounding boxes in frames f and $f+1$. Therefore, the overall combined temporal (unary and pairwise) score is represented as $S^{ijf}_{temp} = S^{if}_{unary} + S^{j(f+1)}_{unary} + \gamma S^{ijf}_{pair}$. The overall temporal cost between bounding boxes b^i_f and b^j_{f+1} is $C^{ijf}_{temp} = 1 - S^{ijf}_{temp}$.

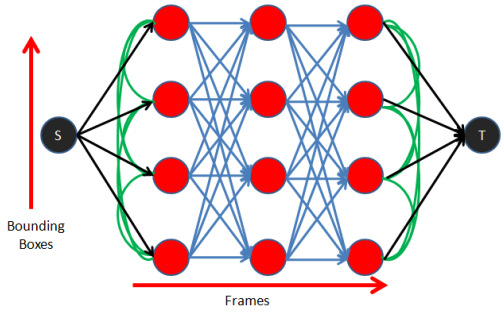

Figure 6. The spatio-temporal graph to extract multiple objects is highlighted here. The bounding boxes in each frame are shown as red circles. The temporal costs shown as blue directed edges indicate inter-frame costs to connect a path through two bounding boxes in successive frames. The intra-frame spatial costs are indicated as green undirected edges. They penalize extraction of the same object in multiple paths. Best viewed in color.

We construct a graph as shown in Fig. 6. The nodes of the graph represents the bounding boxes. The directed edges (across successive frames) shown in blue has weights denoting the temporal cost C^{ijf}_{temp}. We also include additional source (s) and terminal (t) nodes. We want to extract k paths throughout this graph from the source to the terminal node which will represent k extracted objects. However, 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 within a frame. The spatial cost penalizes the extraction of overlapping objects in multiple paths through the graph. The spatial cost associated with two bounding boxes b^i_f and b^j_f in frame f is $C^{ijf}_{spatial} = \frac{\text{Area}(b^i_f \cap b^j_f)}{\text{Area}(b^i_f \cup b^j_f)}$. This spatial cost $C^{ijf}_{spatial}$ is denoted by the undirected green edges (intra-frame) in the graph in Fig. 6.

The aim is to select paths through the graph which minimize the overall spatio-temporal cost. Let the decision variables for the temporal costs be denoted by x^{ij}_f between bounding boxes i in frame f and j in $f+1$. Also, let the decision variables for the spatial costs be denoted by y^{ij}_f between bounding boxes i and j in frame f. Assuming the total number of frames is F, the optimization problem can be formulated as

$$\underset{x_f^{ij}, y_f^{ij}}{\text{minimize}} \sum_{i,j,f} C_{temp}^{ijf} x_f^{ij} + \sum_{i,j,f} C_{spatial}^{ijf} y_f^{ij} \quad (4)$$

subject to:

$$\sum_j x_0^{sj} = k \text{ (Flow from source node = } k) \quad (5)$$

$$\sum_i x_F^{it} = k \text{ (Flow to terminal node = } k) \quad (6)$$

$$\sum_i x_f^{ij} = \sum_k x_{f+1}^{jk} \; \forall j, f \text{ (Conservation of flow)} \quad (7)$$

$$\sum_i x_f^{ij} \leq 1 \; \forall j, f \text{ (At most one active temporal path)} \quad (8)$$

$$y_f^{ij} = (\sum_k x_f^{ik})(\sum_k x_f^{jk}) \; \forall i, j \;\; \text{(Spatial Constraint)} \quad (9)$$

$$x_f^{ij} \text{ and } y_f^{ij} \in \{0, 1\} \quad (10)$$

The edges in the spatio-temporal graphs are either selected (active = 1) or not selected (inactive = 0). This formulation aims to identify the appropriate active x_f^{ij} and y_f^{ij} such that the overall spatio-temporal cost represented in (4) is minimized. The constraints (5) and (6) ensure exactly k paths are selected by our algorithm. The conservation of flow constraint (7) enforces the number of incoming active temporal edges is equal to the number of outgoing active temporal edges at each node. The constraint (8) implies that there can be at most one active temporal path through a node. Finally, the spatial constraint (9) ensures that if there are two temporal paths across nodes p and q in the same frame (f), the corresponding spatial cost (undirected green edge in Fig. 6) connecting them is activated or $y_f^{pq} = 1$.

The spatial constraint in (9) is quadratic, however can be linearized as $y_f^{ij} \geq \sum_k x_f^{ik} + \sum_k x_f^{jk} - 1$ as $\sum_k x_f^{ik}$ and $\sum_k x_f^{jk} \in \{0, 1\}$ due to constraint (8). This results in a binary integer linear program. We utilize the GUROBI [16] solver to get the optimal solution to the problem which eventually extracts k distinct paths from the graph. Finally, the bounding box tracks extracted from the spatio-temporal graph are iteratively refined in every frame individually using grabcut segmentation [37]. We note that in our experiment we set the cost-function weights as $\alpha = 0.2, \beta = 0.5, \gamma = 1$, which gave the best results. We also set $r_{color} =$ and $r_{opt} = 5$.

4. Experimental results

In this section we evaluate the performance of object extraction using the proposed approach. We first discuss the procedure to obtain the ground truth for important objects in the video sequence which attract visual attention followed by the evaluation metric. Subsequently, we show quantitative comparison with state-of-the-art.

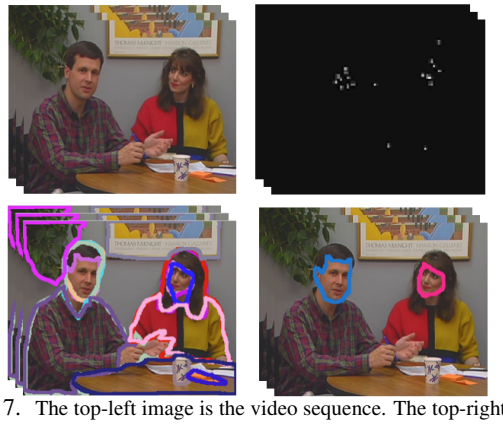

Figure 7. The top-left image is the video sequence. The top-right shows eye tracking samples in a frame from the sequence. The bottom-left figure highlights the exhaustive manual object annotations in the scene. Finally, the bottom-right illustrates the important ground truth objects which attract visual attention obtained according to Section 4.1. Best viewed in color.

4.1. Ground truth creation

In this section, we provide a concrete definition for selecting important objects in a video sequence from eye tracking data. Our eye tracking dataset is obtained from 30 subjects and we randomly select 10 subjects to create the ground truth and the rest for the proposed eye tracking based object extraction algorithm to avoid confirmation bias. In order to obtain ground truth objects from eye tracking data, we first manually annotated (every fifth frame) several meaningful objects in a video sequence. The ground truth objects are assumed to be a subset of these annotated objects. An object in a video sequence is considered important if it captures more than a threshold $(x\%)$ of the visual attention from all the observers. Previous studies have indicated that humans can track upto four objects reliably [5, 20] in video sequence. Therefore, in order to account for slightly uneven distribution of attention among important objects and some attention loss due to calibration error, the threshold is set at $x = 20\%$. Also, an attention region should correspond to a unique object. Therefore, the ground truth is obtained by first sorting the annotations by number of pixels and selecting the annotation which has more than 20% of the attention from the smallest to the largest annotation. Once an object is identified ($> 20\%$ attention), we remove its attention region from the pool and repeat the process and localize subsequent objects. In total we obtain 30 ground truth objects after annotating 136 objects manually. Fig. 7 shows an example of a video sequence, eye tracking data, multiple object annotations and extracted ground truth objects which corresponds to faces. It is well known that faces attract visual attention [9]. This is corroborated by the proposed ground truth extraction scheme.

4.2. Evaluation metric

The output of the object extraction algorithm is a set of object tracks in every frame. These binary masks are matched to the ground truth contour regions obtained ac-

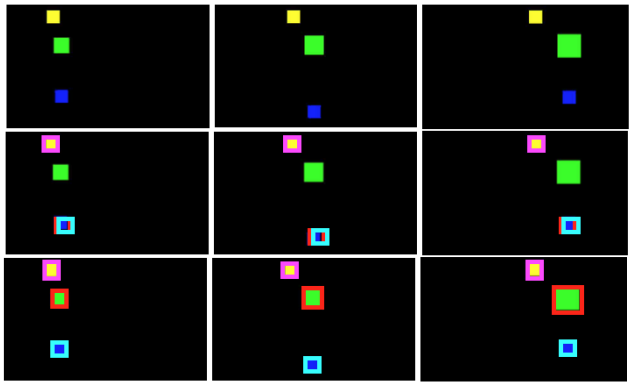

Figure 8. An example highlighting the importance of the spatial constraint in the proposed framework in Section 3.2. The top row indicates three frames from a video sequence where the yellow square moves horizontally, the green square expands and moves horizontally and the blue square undergoes linear slant motion. The second row shows the three objects extracted by the proposed algorithm without the spatial constraint. We notice two overlapping bounding boxes on the blue square. This is avoided by the proposed approach with the spatial constraint in row 3, where all three objects are extracted accurately. Best viewed in color.

cording to Section 4.1. A match score m, between two object tracks is determined as the intersection area divided by the union area between the ground truth and obtained tracks. For a given object track t the best matching track m_b in a set of tracks $\mathcal{T}$ is defined by $m_b(t, \mathcal{T}) = \max\{m(t, t')|t' \in \mathcal{T}\}$. This leads us to the definitions of Precision and Recall as Precision $= \frac{\sum_{t_e \in \mathcal{E}} m_b(t_e, \mathcal{G})}{|\mathcal{E}|}$ and Recall $= \frac{\sum_{t_g \in \mathcal{G}} m_b(t_g, \mathcal{E})}{|\mathcal{G}|}$. Here, $\mathcal{G}$ and $\mathcal{E}$ are the sets of ground truth and estimated object tracks respectively. The precision and recall are combined to a single quantity called F-measure which is defined as $F = \frac{2.\text{Precision}.\text{Recall}}{\text{Precision}+\text{Recall}}$.

4.3. Importance of spatial constraint

This section shows a synthetic example which illustrates the importance of the spatial constraint in our proposed framework in Section 3.2. Fig. 8 shows an example with three squares moving over multiple frames. In the proposed algorithm, we extracted 35 candidate bounding boxes which have multiple overlapping bounding boxes over each of the squares. When we set number of objects, $k = 3$ and extracted three paths without the spatial constraint, two paths overlapped on the same object. However, the spatial constraint which penalizes overlap extracted the three objects accurately. Therefore, we note that without the spatial constraint the optimization can be solved by a simpler linear program. However, it does not yield favorable results for multiple object extraction.

4.4. Performance of multiple object extraction

The video dataset presented in this work contained 30 ground truth objects in total. The proposed approach extracted 31 objects from the visual tracks of which two were false positives, and one false negative was obtained. After extracting the visual tracks from 20 subjects, we want to segment the objects of interest from the videos. Our graph based object extraction algorithm extracts bounding boxes representing important objects in the scene which are refined by grab cut segmentation. Our approach is compared with [30] to evaluate the capability of multiple object extraction using eye tracking prior. As [30] requires a unique fixation point per object, we extract the median of the visual track in every frame to represent the fixation point which provides the pivot for the segmentation. A comparison of our multiple object extraction algorithm with [30] is highlighted in Table 1. We notice that the proposed approach outperforms [30], which is state-of-the-art in eye tracking assisted object extraction, by a significant margin. Additionally, we also obtain better performance than state-of-the-art video object segmentation algorithms [32, 34] which do not utilize eye tracking data.

We also want to understand the role of eye tracking and object extraction module individually to localize objects in a video. For this purpose, we selected bounding boxes around every visual track by using $\mu \pm 2\sigma$, where μ and σ are the mean and variance of the visual track in every frame. These eye tracking based bounding boxes ignore visual information from the video sequence. Additionally, the multiple object extraction module is individually run on the video sequence without the eye tracking based localization prior to quantify the performance of the multiple object extraction framework without utilizing the eye tracking data. However, here we utilize the number of visual tracks to extract k objects from the video sequence. We notice in Table 1 that the proposed approach outperforms individual eye tracking and object extraction methods. Some example bounding box based object extraction results using the proposed spatio-temporal graph are shown in Fig. 9. The final multiple object video segmentation results (after applying grabcut) on a few example videos are shown in Fig. 10. We also illustrate some results where our algorithm outperforms that of [30] in Fig. 11. This is attributed to the high sensitivity of [30] to the location of the fixations and noise in fixation localization.

4.5. Performance of single object extraction

We also compare the performance of the proposed object extraction algorithm with state-of-the-art video segmentation algorithm [47]. As [47] extracts a single object of interest, we set number of objects, $k = 1$ in the proposed approach. We compare the extracted object to the ground truth annotations using the precision metric. We do not compute the recall metric as we are limited to single object extraction where recall is not meaningful. The proposed approach obtains a precision of 0.557 while [47] obtains a precision of 0.374. Therefore, the proposed algorithm outperforms [47] in our dataset, which establishes that our eye tracking based object extraction is suitable for single object extraction too.

	Our algorithm without eye tracking data	Visual tracks only	Active segmentation [30]	Ochs et al. [32]	Papazoglou et al. [34]	Our algorithm bounding boxes	Our algorithm with grabcut
Precision	0.263	0.253	0.448	0.255	0.354	0.439	0.513
Recall	0.251	0.271	0.483	0.245	0.264	0.476	0.526
F-Measure	0.257	0.262	0.465	0.250	0.303	0.457	0.519

Table 1. Comparison of the performance of our multiple object extraction algorithm with state-of-the-art. We also selectively compare the performance of different sub-blocks of our model. We notice that both the object extraction module and eye tracking data contribute equally to extract objects which attract visual attention.

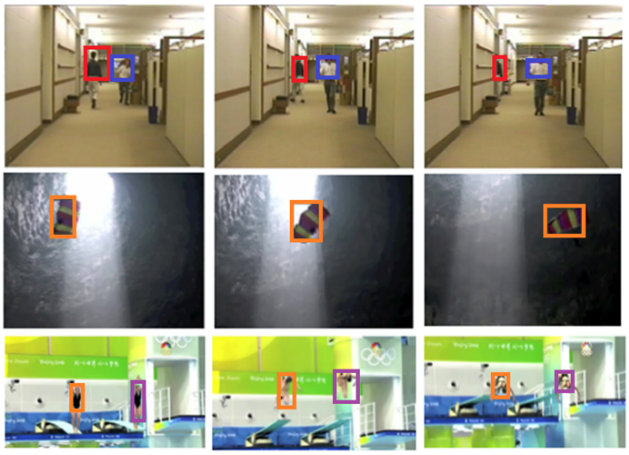

Figure 9. Shows example results using the proposed approach to extract multiple objects represented by bounding boxes. We see the proposed approach is able to localize different visually salient objects in the video sequences with reasonable accuracy. Best viewed in color.

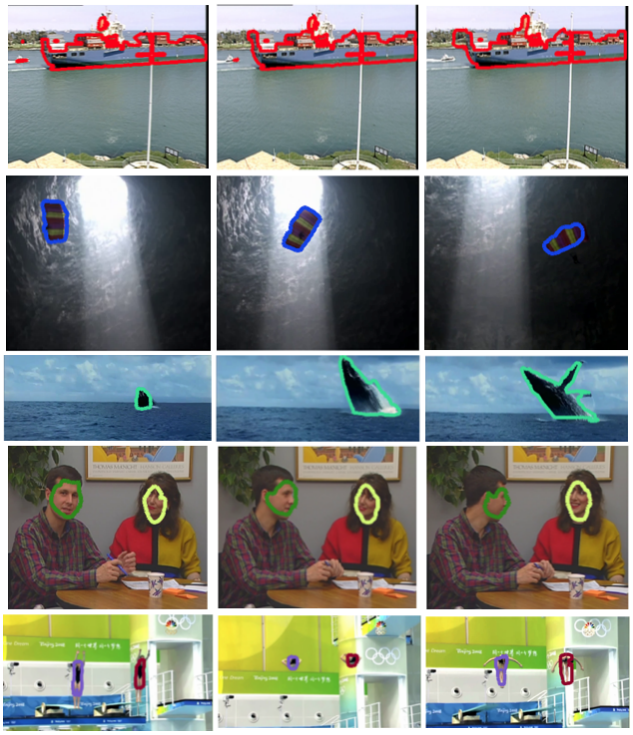

Figure 10. Shows example results from the proposed approach after applying grabcut based video segmentation to the extracted multiple object bounding boxes. We see the proposed approach is able to segment multiple objects in the video sequences with good accuracy. Best viewed in color.

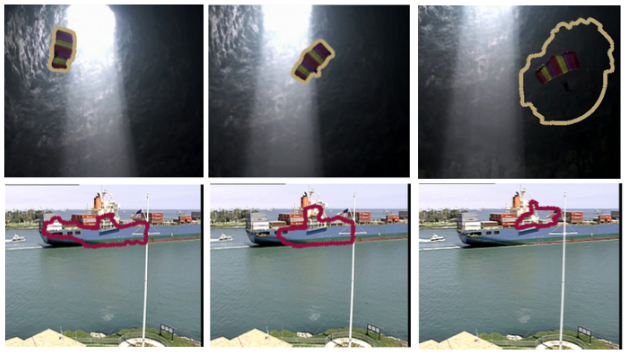

Figure 11. Shows some segmentation results using [30]. We notice that in the top row, possibly due to fixation localization error, the segmentation [30] breaks down. In addition, the algorithm suffers from similar issues in the bottom row as well, where it is not robust to the occluding pole.

5. Summary, Discussion and Future Work

Recent advances in eye tracking technology has enabled collection of eye movement data on a large scale. Multimedia applications can significantly benefit from the availability of such technology. Towards this end, we utilize human eye movements to extract important objects in videos. The algorithm first clusters the eye tracking data using 3-D mean shift to obtain visual tracklets, which are in turn associated to get visual tracks. The visual track guided object search provides object proposals in every frame. Further, this information is used to build a spatio-temporal mixed graph and we extract paths representing objects from this graph by inference using binary integer linear programming. The extracted bounding box based objects are refined using grabcut segmentation to get object contour based segmentation.

To the best of our knowledge, the proposed work is the first attempt to tackle multiple object extraction from videos guided by eye tracking data in a free viewing task. This work was implemented in MATLAB on a 8-core 2.26 GHz machine, with computationally intensive functions in C++. The spatio-temporal multiple object inference is fast, takes less than 10 seconds on 100 frames. We note that the optical flow and objectness computations can be parallelized. We choose bounding box based objectness in [4] as it was significantly faster than contour based approaches [12].

In the future, we want to investigate how the number of subjects affects object extraction performance. It would also be interesting to explore utility of eye tracking in other problems such as image and video retrieval. Finally, we believe single subject eye tracking guided algorithms need further research as they will enable applications beyond multimedia where it can be combined with wearable technology.

6. Acknowledgements

This work was supported in part by the following awards/grants: US Office of Naval Research N00014-12-1-0503 and Institute for Collaborative Biotechnologies through grant W911NF-09-0001 from the U.S. Army Research Office. The content of the information does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

References

[1] Eyelink 1000. *http://www.sr-research.com/EL_1000.html/*. Accessed: 2014-11. 2

[2] Eyetribe eye tracker. *https://theeyetribe.com/*. Accessed: 2014-11. 1

[3] Xiph dataset. *https://media.xiph.org/video/derf/*. Accessed: 2014-11. 2

[4] B. Alexe, T. Deselaers, and V. Ferrari. Measuring the objectness of image windows. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(11):2189–2202, 2012. 5, 8

[5] G. A. Alvarez and S. L. Franconeri. How many objects can you track?: Evidence for a resource-limited attentive tracking mechanism. *Journal of Vision*, 7(13):14, 2007. 6

[6] A. Borji and L. Itti. State-of-the-art in visual attention modeling. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(1):185–207, 2013. 2

[7] A. Borji, D. N. Sihite, and L. Itti. Salient object detection: A benchmark. In *Computer Vision–ECCV 2012*, pages 414–429. Springer, 2012. 2

[8] A. Borji, D. N. Sihite, and L. Itti. Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *Image Processing, IEEE Transactions on*, 22(1):55–69, 2013. 2

[9] M. Cerf, E. P. Frady, and C. Koch. Faces and text attract gaze independent of the task: Experimental data and computer model. *Journal of vision*, 9(12), 2009. 6

[10] M. P. Eckstein. Visual search: A retrospective. *Journal of Vision*, 11(5):14, 2011. 1

[11] K. A. Ehinger, B. Hidalgo-Sotelo, A. Torralba, and A. Oliva. Modelling search for people in 900 scenes: A combined source model of eye guidance. *Visual cognition*, 17(6-7):945–978, 2009. 1

[12] I. Endres and D. Hoiem. Category independent object proposals. In *Computer Vision–ECCV 2010*, pages 575–588. Springer, 2010. 1, 8

[13] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008. 2

[14] S. Goferman, L. Zelnik-Manor, and A. Tal. Context-aware saliency detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(10):1915–1926, 2012. 2

[15] M. Grundmann, V. Kwatra, M. Han, and I. Essa. Efficient hierarchical graph based video segmentation. *IEEE CVPR*, 2010. 2

[16] I. Gurobi Optimization. Gurobi optimizer reference manual, 2014. 6

[17] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *Advances in neural information processing systems*, pages 545–552, 2006. 2

[18] M. Hayhoe and D. Ballard. Eye movements in natural behavior. *Trends in cognitive sciences*, 9(4):188–194, 2005. 1

[19] C. Huang, B. Wu, and R. Nevatia. Robust object tracking by hierarchical association of detection responses. In *Computer Vision–ECCV 2008*, pages 788–801. Springer, 2008. 4

[20] J. Intriligator and P. Cavanagh. The spatial resolution of visual attention. *Cognitive psychology*, 43(3):171–216, 2001. 6

[21] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(11):1254–1259, 1998. 2

[22] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *Computer Vision, 2009 IEEE 12th international conference on*, pages 2106–2113. IEEE, 2009. 1, 2

[23] S. Karthikeyan, D. Delibaltov, U. Gaur, M. Jiang, D. Williams, and B. Manjunath. Unified probabilistic framework for simultaneous detection and tracking of multiple objects with application to bio-image sequences. In *Image Processing (ICIP), 2012 19th IEEE International Conference on*, pages 1349–1352. IEEE, 2012. 4

[24] S. Karthikeyan, V. Jagadeesh, and B. Manjunath. Learning top-down scene context for visual attention modeling in natural images. *ICIP, IEEE*, 2013. 2

[25] S. Karthikeyan, V. Jagadeesh, R. Shenoy, M. Eckstein, and B. Manjunath. From where and how to what we see. In *Computer Vision, 2013 IEEE International conference on*. IEEE, 2013. 2

[26] K. Koehler, F. Guo, S. Zhang, and M. P. Eckstein. What do saliency models predict? *Journal of vision*, 14(3):14, 2014. 1

[27] Y. J. Lee, J. Kim, and K. Grauman. Key-segments for video object segmentation. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1995–2002. IEEE, 2011. 1

[28] T. Ma and L. J. Latecki. Maximum weight cliques with mutex constraints for video object segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 670–677. IEEE, 2012. 1

[29] S. Mathe and C. Sminchisescu. Dynamic eye movement datasets and learnt saliency models for visual action recognition. In *Computer Vision–ECCV 2012*, pages 842–856. Springer, 2012. 1, 2

[30] A. Mishra, Y. Aloimonos, and C. L. Fah. Active segmentation with fixation. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 468–475. IEEE, 2009. 2, 7, 8

[31] A. K. Mishra, Y. Aloimonos, L.-F. Cheong, and A. A. Kassim. Active visual segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions On*, 34(4):639–653, 2012. 2

[32] P. Ochs, J. Malik, and T. Brox. Segmentation of moving objects by long term video analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(6):1187–1200, 2014. 7, 8

[33] D. P. Papadopoulos, A. D. Clarke, F. Keller, and V. Ferrari. Training object class detectors from eye tracking data. In *Computer Vision–ECCV 2014*, pages 361–376. Springer, 2014. 2

[34] A. Papazoglou and V. Ferrari. Fast object segmentation in unconstrained video. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 1777–1784. IEEE, 2013. 7, 8

[35] S. A. Ramakanth and R. V. Babu. Seamseg: Video object segmentation using patch seams. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 376–383. IEEE, 2014. 1

[36] S. Ramanathan, H. Katti, N. Sebe, M. Kankanhalli, and T.-S. Chua. An eye fixation database for saliency detection in images. In *Computer Vision–ECCV 2010*, pages 30–43. Springer, 2010. 2

[37] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. In *ACM Transactions on Graphics (TOG)*, volume 23, pages 309–314. ACM, 2004. 6

[38] R. Subramanian, V. Yanulevskaya, and N. Sebe. Can computers learn from humans to see better?: inferring scene semantics from viewers' eye movements. In *Proceedings of the 19th ACM international conference on Multimedia*, pages 33–42. ACM, 2011. 2

[39] D. Sun, S. Roth, and M. J. Black. Secrets of optical flow estimation and their principles. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2432–2439. IEEE, 2010. 3, 5

[40] A. Torralba, A. Oliva, M. S. Castelhano, and J. M. Henderson. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological review*, 113(4):766, 2006. 1

[41] D. Tsai, M. Flagg, and J. M.Rehg. Motion coherent tracking with multi-label mrf optimization. *BMVC*, 2010. 2

[42] E. Vig, M. Dorr, and D. Cox. Space-variant descriptor sampling for action recognition based on saliency and eye movements. In *Computer Vision–ECCV 2012*, pages 84–97. Springer, 2012. 2

[43] T. Walber, A. Scherp, and S. Staab. Can you see it? two novel eye-tracking-based measures for assigning tags to image regions. In *Advances in Multimedia Modeling*, pages 36–46. Springer, 2013. 2

[44] K. Yun, Y. Peng, H. Adeli, T. Berg, D. Samaras, and G. Zelinsky. Specifying the relationships between objects, gaze, and descriptions for scene understanding. *Journal of Vision*, 13(9):1309–1309, 2013. 2

[45] K. Yun, Y. Peng, D. Samaras, G. J. Zelinsky, and T. L. Berg. Exploring the role of gaze behavior and object detection in scene understanding. *Frontiers in psychology*, 4, 2013. 2

[46] K. Yun, Y. Peng, D. Samaras, G. J. Zelinsky, and T. L. Berg. Studying relationships between human gaze, description, and computer vision. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 739–746. IEEE, 2013. 2

[47] D. Zhang, O. Javed, and M. Shah. Video object segmentation through spatially accurate and temporally dense extraction of primary object regions. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 628–635. IEEE, 2013. 1, 7