Towards Gaze-based Video Annotation

Mohamed Soliman, Hamed R.-Tavakoli and Jorma Laaksonen
Aalto University
e-mail: mohamed.soliman@aalto.fi, h.rtavakoli@yahoo.com, jorma.laaksonen@aalto.fi

***Abstract*— This paper presents our efforts towards a framework for video annotation using gaze. In computer vision, video annotation (VA) is an essential step in providing a ground-truth for the evaluation of object detection and tracking techniques. VA is a demanding element in the development of video processing algorithms, where each object of interest should be manually labelled. Although the community has handled VA for a long time, the size of new data sets and the complexity of the new tasks pushes us to revisit it. A barrier towards automated video annotation is the recognition of the object of interest and tracking it over image sequences. To tackle this problem, we employ the concept of visual attention for enhancing video annotation. Human attention naturally grasps semantic objects that provide valuable information for extracting the objects of interest, which can be exploited to annotate videos. Under task-based gaze recording, we utilize an observer's gaze to filter seed object detector responses for a video sequence. The filtered boxes are then passed to an appearance-based tracking algorithm. We evaluate the gaze usefulness by comparing the algorithm with gaze and without it. We show that eye gaze is an influential cue for enhancing the automated video annotation, improving the annotation significantly.**

***Keywords*— Video annotation, Video processing, Visual attention, Eye gaze, Object detection and tracking.**

I. Introduction

Manual annotation of videos has always been the way to produce the ground-truth to track objects of interest in image sequences. Manual video annotation is performed by asking some one to watch a video and manually draw bounding boxes over objects of interest on each video frame [1], [2]. The massive increase in the amount of multimedia content has made it very costly, in terms of human effort and time, to proceed with manual annotation.

Knowing the detection and tracking of the object of interest is the bottleneck, an alternative to manual annotation is employing a robust automated object tracker. For example, we can employ Direct Acyclic Graphs (DAGs) for detection and segmentation of the primary objects in a video to build an object model and tracking the model over sequences [3]. Another approach can be the use of optical flow and motion boundaries for background-foreground segmentation [4] employed in automated video annotation. There is, however, a severe defect with such an approach, that is there is no robust tracking algorithm to outperform or achieve the human performance. In other words, it is necessary to have kind of an explicit feedback, which provides object bounding boxes, in the loop.

In this paper, instead of explicit feedback, we rely on implicit feedback by eye gaze. The eye gaze can be obtained unobtrusively, eliminating any explicit feedback, while an observer is watching a video clip and looking at a specific target, e.g. a movie character. Furthermore, it can be scaled-up easily by crowd sourcing [5], making it a unique input for fast and accurate annotation of a large-scale video corpus.

II. Related work

Visual attention modeling and fixation prediction in images and videos [6], [7], [8] is a well-researched area with the purpose of saliency modeling, e.g. [9], [10], [11] and applications in various areas such as motion detection [12], object tracking [13], [14], [15], object segmentation [16], interest point detection and recognition [17], etc.

To date, the computational models of attention have been used to replicate human fixations due to the expensiveness of eye-tracking technologies. For instance, in object segmentation, a saliency map is initially computed. The saliency map is then exploited to generate artificial fixations that produce segmentation in conjunction with the edge-boundary constraint as in [18], or it is treated as a feature vector in a conditional random field setup to perform segmentation like [19].

Recently, with the advent of affordable eye-tracking systems, real human fixations can be exploited for the human-in-the-loop applications where eye gaze applied. For example, [20] exploited eye gaze to obtain text and face priors to enhance object detectors performance. A recent study [21] investigated the training of object detectors with weakly annotated data whereby the eye gaze provides the annotation ground-truth rather than object bounding boxes. The annotating process is faster albeit the performance of detectors is not optimal.

The most relevant to our work is [22], which employs gaze as a cue to validate localized objects for salient object detection from videos. Several observers are asked to freely watch a video, meanwhile their eye-gaze data is constantly acquired using an eye tracker. Simultaneously, the object detectors are producing object bounding boxes. The bounding boxes are then merged with the help of gaze data to refine the tracking result. The eventual output is object segments from image sequences. While [22] is designed for online purposes useful in ego-centric vision systems, our approach is planned with an offline pipeline with forward and backward processing passes in mind. We also use an affordable eye tracking device with a low sampling rate (0.03 of the sampling rate used in [22]).

In the rest of this paper, we first discuss the proposed

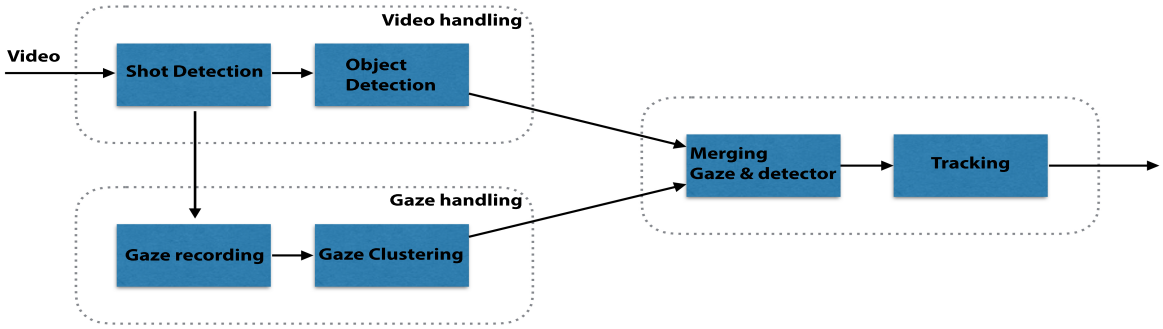

Fig. 1. The overall framework of the gaze-based framework for video annotation.

pipeline and explain its details, where our main focus is on the approach used for inferring the best object bounding box candidate. It is followed by the evaluation and conclusions.

III. Method

Figure 1 visualizes the overall pipeline of the proposed framework. We estimate the object bounding boxes for each video frame. A separate pipeline records the eye movements of observers while they are watching the movie. In order to map eye gaze to bounding boxes, we first cluster the eye gaze data. Afterwards, we associate the eye gaze to the object bounding boxes. The objects with a gazed bounding box are then used to build an appearance model for a tracking algorithm. We explain each part of the pipeline in the rest of this section.

A. Gaze handling

Gaze data is obtained from one observer, watching an episode of the "Scrubs" series. We recorded the eye movements using a Tobii Eye-X controller. The fixation information, i.e. sample time-stamp, fixation location and fixation duration, were obtained using the API of the controller. The video is screened on a 19 inch LCD at the resolution of 1280×1024 and distanced at a bout 65-70 cm from the observer. The observer has normal vision and never reported having eye-sight problems.

Gaze clustering is performed to compensate the few number of samples per frame caused by the low sampling rate of the eye tracking device. To determine the number of clusters for a video shot, we employ self-tuning spectral clustering algorithm [23], that is given a set of gaze points (fixations) $G = \{g_1 \cdots g_n\}$, where $g_i = (x, y, t)$ consist of gaze coordinates and time stamp t, we apply Algorithm 1 to determine the best number of possible clusters.

Once we determine the best number of clusters associated with a video shot, a Gaussian Mixture Model (GMM) clustering is employed to group the eye movements as depicted in Figure 2. The center of each cluster is used to filter out the irrelevant and false positive responses of object detectors.

Algorithm 1 Determine The Number of Clusters

procedure ClusterEstimate(G, C)
  Determine local scale $\sigma_i \leftarrow d(g_i, g_N) \quad \forall g_i \in G$
  Build affinity matrix $A \in \mathcal{R}^{n \times n}$ such that
    $A_{ij} \leftarrow \exp(-d(g_i, g_k)/\sigma_i \sigma_j)$
    $A_{ii} \leftarrow 0$
  $L \leftarrow D^{-1/2} A D^{-1/2}, \quad D_{ii} = \sum_{j=1}^{n} A_{ij}$
  $X \leftarrow [x_1, \cdots, x_C]$, the C largest eigenvectors of L
  Recover rotation matrix R to align columns of X with the canonical coordinate system.
  $Z \leftarrow XR$
  $M_i \leftarrow \max_j Z_{ij}$
  $R_{x_i} \leftarrow \sum_{i=1}^{n} \sum_{j=1}^{C} \frac{Z_{ij}^2}{M_i^2}$ ▷ Rank eigen vectors
  $C_{best} \leftarrow i \quad |R_{x_i} < R_{x_j} \quad \forall j \in [1, \cdots C]$
  return C_{best} ▷ The best number of clusters

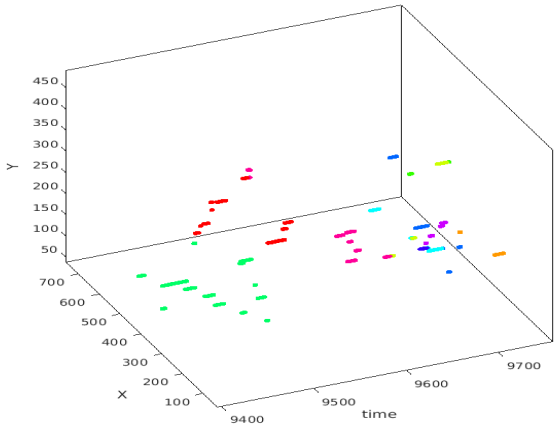

Fig. 2. The spatio-temporal clustering of eye gaze data of a movie shot. Each cluster is colored differently.

B. Video handling

Shot detection is necessary in order to cope with the dynamics of scenes and simplifying the gaze to object association by limiting the maximum number of possible objects and gaze clusters. There exist various shot boundary detection algorithms [24] including, color-histogram, edge change ratio,

contrast, etc. In this work, we, however, use manually tagged video shots in order to reduce the noise caused by inaccurate shot detection. Furthermore, the manual annotation can be used to help choosing a suitable shot-boundary detection algorithm.

Obeject detection is based on deformable part models [25] and is chosen in regard to the task assigned to the observer. In other words, if an observer is instructed for spotting cars, a car detector is used. In our implementation, we utilize the upper-body detection [26] and Viola-Jones face detector [27] to detect people faces and upper-bodies in each video frame. The non-maximum suppression is employed to fuse the detector responses. As depicted in Figure 3, there exists a high number of false positive detection.

Fig. 3. The result of upper-body and face detection after non-maximum suppression. A high number of false positive and irrelevant responses exist.

C. Merging gaze and object detectors

The eye gaze data is used to filter out the false positive responses. For each video frame of a shot, the distance of detector's bounding box to gaze cluster centers is used for this purpose. A detection is valid, if it has a distance to a gaze cluster less then a threshold value. We evaluated two approaches to determine the threshold value, 1) The mean/median distance of all object bounding boxes to the gaze clusters for the shot, 2) The mean/median distance of object bounding boxes using eye gaze cluster active time (time-activation), i.e., the duration of the gaze cluster validity. The filtering step reduces the number of false positives significantly, as shown in Figure 4. The resulted set of valid bounding boxes is used as the input to the tracking algorithm.

D. Tracking

The valid bounding boxes are used as seed object detectors in each video frame of the tracking algorithm. We employ an appearance-based multi-object tracking algorithm by [28] in order to keep track of the detected objects and fine tune the final results. That is for a valid detected object o in frame t, denoted by a binary function $v^o(t) = 1$, the state of the object

Fig. 4. The valid upper-body and face detection after filtering detections using gaze. The number of false positive detections decreases significantly.

is represented by $\mathbf{x}_t^o = (\mathbf{p}_t^o, \mathbf{s}_t^o, \mathbf{v}_t^o)$, where $\mathbf{p}_t^o$, $\mathbf{s}_t^o$ and $\mathbf{v}_t^o$ are the position, size, and velocity, respectively. The algorithm then relies on tracklets for tracking an object. A tracklet is defined as $T^o = \{\mathbf{x}_k^o | \mathbf{v}^o(k) = 1, 1 \leq t_s^o \leq k \leq t_e^o \leq t\}$, where the start- and end-frame of the tracklet is defined by t_s^o and t_e^o, respectively.

By defining a tracklet T^o for an object as the set of states up to frame t, a set of tracklets consisting of all the objects up to frame t is represented by $\mathbb{T}_{1:t}$. Similarly, the valid detected object o at frame t is defined as $\mathbf{z}_t^o$ and the set of all detections up to frame t as $\mathbb{Z}_{1:t}$. The multi-object tracking is then defined as the estimation of the maximum a posteriori (MAP) of $\mathbb{T}_{1:t}$, given $\mathbb{Z}_{1:t}$:

$$\hat{\mathbb{T}}_{1:t}^{\text{MAP}} = \arg\max_{\mathbb{T}_{1:t}} p(\mathbb{T}_{1:t} | \mathbb{Z}_{1:t}). \tag{1}$$

The above equation is solved using the tracklet confidence approach [28] due to the infeasibility of the solution caused by the innumerability of the combinations of $\mathbb{T}_{1:t}$ and $\mathbb{Z}_{1:t}$. The tracking process can be repeated via a backward pass and merged with the forward pass to increase the performance. Figure 5 depicts the tracked objects using gaze validated bounding boxes.

IV. Experiments and Results

We evaluate the contribution of gaze in the proposed framework on a sequence of "Scrubs" TV series with the gaze data of one observer. We manually annotated the regions of interest for the sequence. To evaluate, we employ precision, recall, multiple-object tracking accuracy and the average overlap. The average overlap at frame t is defined as the Jaccard similarity coefficient:

$$a_t = \frac{|PB_t \cap GT_t|}{|PB_t \cup GT_t|}, \tag{2}$$

Fig. 5. The tracked characters using the gaze-validated detector responses.

where PB is the detected bounding box, GT is the ground-truth bounding box and $|\cdot|$ indicates area. Any bounding box with $a_t \geq 0.3$ is a true positive. For a sequence, the mean a_t of true positive detections over all frames is reported. We measure the multiple-object tracking accuracy (MOTA) [29] defined as follows:

$$MOTA = 1 - \frac{\sum_t (M_t + FB_t)}{\sum_t GT_t}, \tag{3}$$

where the M_t is the number of misses and FB_t is the number of false detected bounding boxes (false positives).

A. Suitable object detector

In the experiments, we are focused on the movie characters as the objects of interest. To understand the suitability of object detectors, we evaluate two object detectors, including the upper-body detector [26] and the Viola-Jones face detector [27], and their combination. We feed the detector response to the tracking algorithm without gaze-based validation. Therefore, the results can also be used as a baseline for the assessment of gaze contribution. The results are summarized in Table 1.

Table 1. Baseline: the evaluation of object detectors without gaze validation.

Name	Detector	Average overlap	Precision	Recall	MOTA
B1	face	39.9%	36.2%	17.9%	6.8%
B2	upper-body	48.7%	39.1%	37.2%	23.2%
B3	face+upper-body	49.3%	41.1%	39.2%	15.1%

The results reveal that the face detector by itself is not a good detector for movie character detection and tracking. It, however, shows that the upper-body detector is not that different from the combination of the face and upper-body detectors. We, thus, adopt upper-body detector and face+upper-body detector in the further experiments.

B. Gaze-based detection and tracking

In this section, we evaluate the parameters affecting the gaze-based filtering of the object detectors. We investigate the thresholding method, mean distance or median distance, and clustering with gaze time-activation for the upper-body detector and the face+upper-body detector. The results are reported in Table 2. We also report the F1-score in order to interpret the relation between the precision and recall. This analysis shows that the M1 model, using the upper-body detector and mean distance threshold with no time-activation achieves the highest MOTA score, followed by the M6 model which employs face+upper-body detector and has time-activation. The difference between the two models is that in the latter one emphasis is on recall improvement, while the first one is more precise. Nonetheless, taking the F1-score into account the M6 model is a clear winner by having a better overall precision-recall performance and the second highest MOTA score.

The overall result comparison shows that the models with mean distance value as threshold are on average performing better in terms of recall, F1 score, and MOTA. The models with the combination of detectors, i.e. face+upper-body, has on the average the highest recall and F1 score.

C. Gaze contribution

We evaluate the contribution of gaze in object detection and tracking by comparing the model M6 and the baseline models, B2 and B3. The results are summarized in Table 3. While the average overlap of the baseline model B3 is high, it is clearly evident that the M6 model outperforms the baselines in terms of all the other metrics by a large margin. Therefore, we conclude that gaze as an implicit cue is a significant contributing cue for semi-automatic video annotation.

Table 3. Gaze contribution: comparing performance of gaze-based filtered detections and all the detectios.

Name	Average overlap	Precision	Recall	F1 score	MOTA
B2	48.7%	39.1%	37.2%	38.13%	23.2%
B3	**49.3%**	41.1%	39.2%	40.30%	15.1%
M6	48.3%	**71.7%**	**57.8%**	**64.00%**	**33.7%**

V. Conclusion

This paper described a part of our ongoing effort towards a gaze-based video annotation system. We presented a simple framework for video annotation that could be used to evaluate the usefulness of gaze data for the task. We combined gaze and object detection to provide seed object bounding boxes to an appearance-based object tracking algorithm. We evaluated various parameters for the fusion of gaze and object detector responses and then assessed the contribution of gaze.

The results demonstrate that gaze is a strong cue for automated seed detection and tracking. The future work will be focused on more rigorous integration of gaze and object

Table 2. Gaze-based detector filtering: analysis of the parameters.

Name	Detector	Threshold	Time-activation	Average overlap	Precision	Recall	F1	MOTA
M1	upper-body	mean	no	60.2%	77.1%	50.3%	60.88%	**34.6%**
M2	upper-body	mean	yes	58.4%	75.8%	47.5%	58.40%	31.6%
M3	upper-body	median	no	**61.7%**	**78.8** %	39.5%	52.62%	28.1%
M4	upper-body	median	yes	57.9%	74.1%	39.6%	51.62%	25.1%
M5	face+upper-body	mean	no	49.0%	68.5%	56.8%	62.10%	30.6%
M6	face+upper-body	mean	yes	48.3%	71.7%	**57.8%**	**64.00%**	33.7%
M7	face+upper-body	median	no	51.2%	72.4%	47.0%	56.99%	28.9%
M8	face+upper-body	median	yes	48.0%	72.3%	47.3%	57.19%	29.1%

tracking mechanism and scaling-up the proposed framework for multiple object types.

VI. Acknowledgment

This work has been funded by the grant 251170 of the Academy of Finland.

References

[1] Z. Palotai, M. Lang, A. Sarkany, Z. Toser, D. Sonntag, T. Toyama, and A. Lorincz, "Labelmovie: Semi-supervised machine annotation tool with quality assurance and crowd-sourcing options for videos," in *CBMI*, June 2014, pp. 1–4.

[2] C. Vondrick, D. Ramanan, and D. Patterson, "Efficiently scaling up video annotation with crowdsourced marketplaces," in *CVPR*, 2010.

[3] D. Zhang, O. Javed, and M. Shah, "Video object segmentation through spatially accurate and temporally dense extraction of primary object regions," in *CVPR*, 2013.

[4] A. Papazoglou and V. Ferrari, "Fast object segmentation in unconstrained video," in *ICCV*, 2013.

[5] P. Xu, K. A. Ehinger, Y. Zhang, A. Finkelstein, S. R. Kulkarni, and J. Xiao, "Turkergaze: Crowdsourcing saliency with webcam based eye tracking," in *arXiv:1504.06755v1*, 2015.

[6] J. K. Tsotsos, L. Itti, and G. Rees, "A brief and selective history of attention," in *Neurobiology of Attention*, L. Itti, G. Rees, and J. K. Tsotsos, Eds. Burlington: Academic Press, 2005, pp. xxiii – xxxii. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780123757319500033

[7] A. Borji, D. Sihite, and L. Itti, "Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study," *Image Processing, IEEE Transactions on*, vol. 22, no. 1, pp. 55–69, 2013.

[8] A. Borji, H. R.-Tavakoli, D. N. Sihite, and L. Itti, "Analysis of scores, datasets, and models in visual saliency prediction," in *ICCV*, 2013.

[9] H. Rezazadegan Tavakoli, E. Rahtu, and J. Heikkilä, "Fast and efficient saliency detection using sparse sampling and kernel density estimation," in *Image Analysis*, ser. Lecture Notes in Computer Science, A. Heyden and F. Kahl, Eds. Springer Berlin / Heidelberg, 2011, vol. 6688, pp. 666–675.

[10] H. R. Tavakoli, E. Rahtu, and J. Heikkil"a, "Stochastic bottom-up fixation prediction and saccade generation," *Image and Vision Computing*, vol. 31, no. 9, pp. 686 – 693, 2013.

[11] H. Rezazadegan Tavakoli, E. Rahtu, and J. Heikkilä, "Spherical center-surround for video saliency detection using sparse sampling," in *Advanced Concepts for Intelligent Vision Systems*, ser. Lecture Notes in Computer Science, J. Blanc-Talon, A. Kasinski, W. Philips, D. Popescu, and P. Scheunders, Eds. Springer International Publishing, 2013, vol. 8192, pp. 695–704.

[12] ——, "Temporal saliency for fast motion detection," in *Computer Vision - ACCV 2012 Workshops*, ser. Lecture Notes in Computer Science, J.-I. Park and J. Kim, Eds. Springer Berlin Heidelberg, 2013, vol. 7728, pp. 321–326.

[13] S. Frintrop and M. Kessel, "Most salient region tracking," in *ICRA*, 2009.

[14] A. Borji, S. Frintrop, D. Sihite, and L. Itti, "Adaptive object tracking by learning background context," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, 2012, pp. 23–30.

[15] H. Rezazadegan Tavakoli, M. Shahram Moin, and J. Heikkilä, "Local similarity number and its application to object tracking," *International Journal of Advanced Robotic Systems*, vol. 10, no. 184, 2013.

[16] Y. Fu, J. Cheng, Z. Li, and H. Lu, "Saliency cuts: An automatic approach to object segmentation," in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, Dec 2008, pp. 1–4.

[17] D. Gao, S. Han, and N. Vasconcelos, "Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition," *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, vol. 31, no. 6, pp. 989 – 1005, 2009.

[18] A. Mishra, Y. Aloimonos, and C. L. Fah, "Active segmentation with fixation," in *Computer Vision, 2009 IEEE 12th International Conference on*, Sept 2009, pp. 468–475.

[19] E. Rahtu, J. Kannala, M. Salo, and J. Heikkilä, "Segmenting salient objects from images and videos." in *Proc. 11th European Conference on Computer Vision (ECCV 2010), Crete, Greece*, 2010. [Online]. Available: http://www.ee.oulu.fi/mvg/page/saliency

[20] S. Karthikeyan, V. Jagadeesh, R. Shenoy, M. Ecksteinz, and B. S. Manjunath, "From where and how to what we see," in *ICCV*, 2013.

[21] D. P. Papadopoulos, A. D. F. Clarke, F. Keller, and V. Ferrari, "Training object class detectors from eye tracking data," in *ECCV*, 2014.

[22] S. Karthikeyan, T. Ngo, M. Eckstein, and B. Manjunath, "Eye tracking assisted extraction of attentionally important objects from videos," in *CVPR*, 2015.

[23] L. Zelnik-Manor and P. Perona, "Self-tuning spectral clustering," in *NIPS*, 2004.

[24] R. Lienhart, "Comparison of automatic shot boundary detection algorithms," 1999, pp. 290–301.

[25] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 9, pp. 1627–1645, 2010.

[26] V. Ferrari, M. Marin-Jimenez, and A. Zisserman, "Progressive search space reduction for human pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2008.

[27] P. Viola and M. Jones, "Robust real-time object detection," in *International Journal of Computer Vision*, 2001.

[28] S. H. Bae and K. J. Yoon, "Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, June 2014, pp. 1218–1225.

[29] K. Bernardin, E. Elbs, and R. Stiefelhagen, "Multiple object tracking performance metrics and evaluation in a smart room environment," in *VS*, 2006.