2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation

Visual Attention-Based Object Detection in Cluttered Environments

Eduardo Machado* §, Ivan Carrillo† *, Miguel Collado *, Liming Chen†
*Department of Innovation and Technology, Ingenieria y Soluciones Informatica, Sevilla, Spain
Email: {Eduardo.machado, Ivan.Carrillo}@isoin.es
†School of Computer Science and Informatics, De Montfort University, Leicester, UK
Email: Liming.chen@dmu.ac.uk

***Abstract*—The study of human visual attention is considered a hot topic in the field of activity recognition, experimental psychology research and human computer interaction. The importance of detecting user objects of interest in real time is critical to provide accurate cues about the user intentions. However, current methods for visual attention extraction and object detection suffer from low performance when moving to ongoing condition. Inherent complexity of cluttered environments is considered the major barrier to achieve good performances. To address this challenge, we present a novel method that includes head-worn eye tracker and egocentric video. Our method exploits sliding window-based time series approach in conjunction with a Heuristic probabilistic function to analyse user fixations around potential object of interest in an egocentric video. We evaluate the proposed method using a new dataset annotated with user gaze data and object within a frame image. Our experimental results show that our approach can outperforms several state-of-the-art commonality visual attention-based object detection methods.**

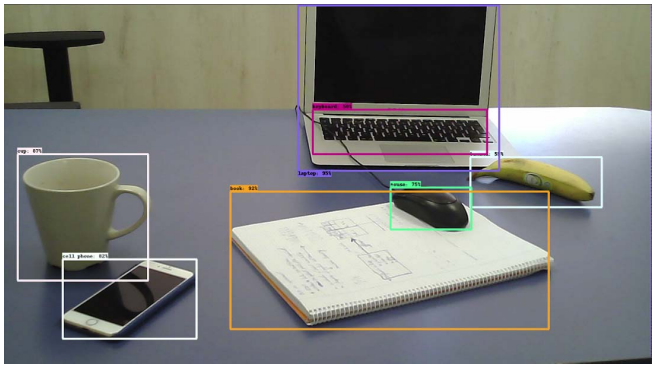

Fig. 1. Cluttered environments as characterized by the variability of overlapped objects regions. Our method exploit visual gaze analysis over overlapped regions to distinguish object of interest.

I. Introduction

Over the last decade eye movements have been intensively explored for the understanding of the nature of the human attention. In this regard, eye tracker emerged as an important enabling technology to extract where a user's visual attention allocation is at any given point in time. The attention information as been been represented long sequences of eye positions, saccades and fixations. Among them, fixations has been the core metric for psychological experiments on visual attention in the context of reading comprehension [1], memory [2] and visual perception [3]. Fixations also have been used to analysis user visual attention on a wide range of applications such as assessment of on-line learning, typing, multimedia content-aware image resizing and so on [4]. More recently, the investigation has shifted towards improving or creating new models of user context for situations-ware interaction. Here, the ability to recognize objects of interests assumes to be critical for constructing an accurate model. In this regards, egocentric cameras and head worn eye tracker are becoming very popular as they overcome the major limitation of common approaches which usually require sensor-equipped environment. More advanced approaches make use of Fixation Density Maps (FDM) generated from eye tracking experiments to train neural networks to predict objects location of the user visual attention in an image [5] [6]. The result is a saliency map of regions that are then extrapolated discover object of interest. Although existing research prove to be promising, there are also limitations and open issues. Due to the nature of this approach, it is designed to be task specific, because the predicted objects do not necessarily correspond to the ones the user is focused on. Instead, they show potential objects of interest due to its shape contours, size and location. Here the variability of the individual natural gaze exploration is not considered because it depends on the user motivation and the specific task that he is performing [4]. Moreover, in daily living contexts, i.e. driving, where a user's environment is continuously changing, the nature of this approach turns out to be unviable for this type settings. In an attempt to overcome this obstacle, other approaches start from using egocentric cameras based on the assumption that user visual attention is allocated at the centre of the visual field [6]. This assumption is supported by the fact that in centre of the human retina there is much higher resolution compared to its more peripheral regions [4], so human gaze tend to be focus on centre region. Studies have used a pre-defined spatial pixels region of egocentric images uses as feature input to object recognition algorithms. However, results from psychological experiments have shown that humans are able to focus their attention in peripheral regions distanced from the gaze centre. An example is a car driver who is fixated on the road but is still able to monitor road signs or potential pedestrians trying to cross the roads. Evidences show that this redirection of visual attention is preceded by prior fixations in peripheral

978-1-7281-4034-6/19/$31.00 ©2019 IEEE
DOI 10.1109/SmartWorld-UIC-ATC-SCALCOM-IOP-SCI.2019.00064

regions [4]. This effect is only perceived by analyzing eye movements on eye trackers. Meaning that, fixed centre region approach would fail to analyze important information outside of the centre of the visual field. An alternative approach that seeks to reduce the low accuracy of fixed region models, is to use of head worn eye tracking equipment and egocentric cameras for gaze-based object detection [7][8][9]. Here, object detection algorithms are used in combination with fixation data to discover object of interest. This approach was applied on the creation of digital episodic memories to support people cognitive impaired people, which has where demonstrated good results [10]. Despite that progress has been made in controlled experiments environments, it remains a challenge to apply egocentric visual attention object recognition is when moving these solutions normal daily living environments. In most cases, the existent solutions perform well because they are trained using data exclusively collected for task specific and in controlled environment. The main issue is to extract visual attention in dynamic cluttered environments, where objects are constantly overlapped and partial object occlusion are also frequent. In this paper, we introduce a novel approach to visual attention-based object detection by combining head-worn eye tracker cameras and egocentric video. Specifically, we present a new novel heuristic function that is able to generate a probability estimation of visual attention over objects within a egocentric video. In addition, we produced an egocentric indoor dataset annotated with human fixation during natural exploration in cluttered environment. Therefore, our dataset targets general and task free conditions. Comparing to previous works, our dataset is more realistic because it was recorded in real settings with variations in terms of objects overlapping regions and object sizes. Finally, through experimental evaluations we show that our method outperforms state of the art visual attention methods for object detection.

II. Related work

Predicting visual attention is a task that have been extensively studied in computer vision and human perception. Existing approaches vary in the granularity of predictions, either focusing on predicting saliency regions in image that are likely to attract attention or predicting specific object instances. In both cases human visual gaze is needed. A different alternative includes the use of head-worn eye tracker and egocentric video to combine fixation data and images of user's context to extrapolate object of interest. Others approaches, utilizes multiple egocentric videos to discover objects of joint attention. In the following we summarize previous works on visual attention based on saliency map, egocentric-based object detection and Fixation-based object detection.

A. Visual Attention Detection Based on Saliency Map

Recently in the field of computer vision, researchers use human fixations as input to train neural network algorithms to create saliency maps of an image. The result is a combination of probability of saliency with spatial colour variation of pixels, which represents human visual attention over specific region. Saliency prediction models can be divided in two categories: bottom up and top down [11]. Bottom up category is focused on building saliency maps based on intrinsic features of image like its colour intensity, objects shape, size and location. In addition to bottom up models, top down category takes also in consideration external cues that influence human attention like faces presence, attractive objects and motion. In [12] is presented a bottom up model that combines global contrast with probability of saliency. A different approach is proposed by [13] that combines local and global contract image features to predict saliency in cluttered scenes. In an attempt to increase the accuracy of saliency map, in [11] is included top down cues such face presence and movement combined with global contrast features. Other works exploited cognitive features and scales for a top down visual search model by using multivariate Gaussian distributions [14]. More recently, [15] combined several bottom up and top down saliency models using several combination strategies. Saliency maps reveal to be very useful to extend our understanding over human visual attention and what object features influence it the most. However, in context of our work that envisions the detection of objects under visual attention in runtime, saliency maps would not perform well because it is designed to predict fixations in specific regions of an image which not necessarily correspond to the user attention in particular temporal space.

B. Egocentric-based object detection

The recent proliferation of wearable cameras prompted the interest for exploring visual analysis in wide range of research fields such as activity recognition, handled object recognition, and object prediction of joint visual attention. Independently of the application domain, there is a common goal among these research topics which is producing a model capable of infer object of interest. Here, the discovery of users visual attention assumes to be a critical step for the models accuracy. In the absence of visual data, common approaches define a specific fixed region in first person video, for which it is extracted features to describe objects being viewed. In [16], the center of the image frame is assumed to be the region under the user visual attention. Then, a Euclidean function is applied to measure the distance from object detected to center frame in order to estimate how likely an object is being focused. In the same direction, [17] computes the distance between of the centroid of an object and the center of image as input features for saliency detection on egocentric video. In the context of our work, fixed-size regions approach would not work well because due to the variability in the size of objects and the different angles perspective, only a limited part of object features is often described in the center region. Moreover, findings suggest that despite human visual attention is mostly focused at the spatial center region of the visual field, attention can be also observed in peripheral regions [18].

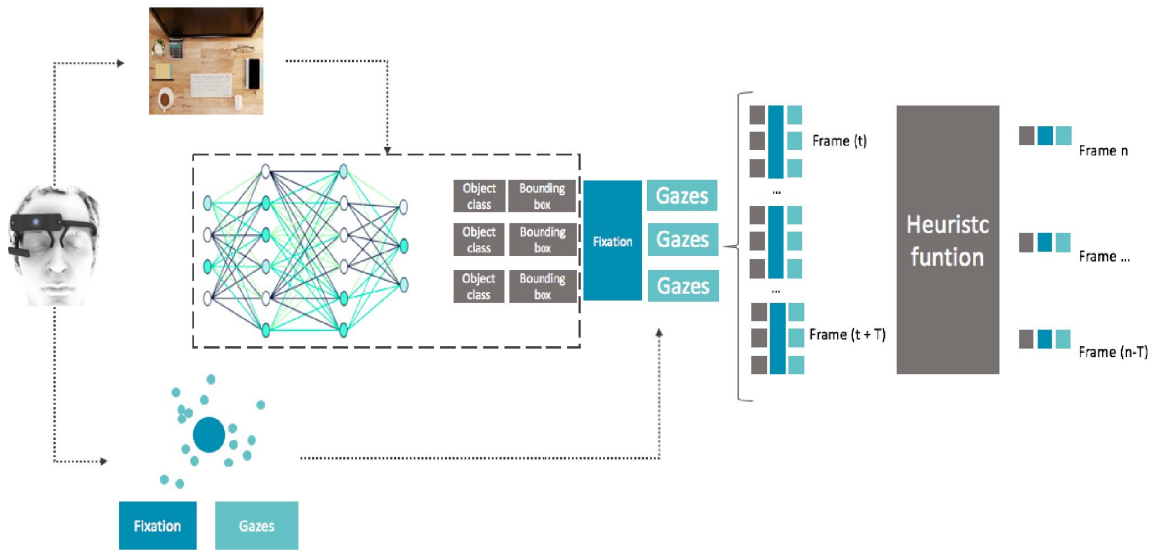

Fig. 2. Detailed scheme exposing the different modules involved in the object of interest detection task.

C. Fixation Based Object Detection

Recent technological advances have made head-worn eye tracker more affordable and widely accessible. This effect sparked the interest for the development of interactive system that incorporates user visual data as an indicator of user visual attention for object of interest detection. A comprehensive list of different applications domains for head-worn eye trackers is provided by [19]. However, one of the issues regarding the eye tracker cameras is the lack of accuracy of the gaze data points resulted by inefficient calibrations. Most of the approaches interpolates data gaze with output from state of art object detection models, for summarizing of object of interest. As result, the accuracy of gaze data becomes very important for the performance of those models. Few studies have tackled this challenge. In [10], it is cropped a rectangle region from image centered on the fixation point and then extracted SIFT features from that region for object detection. The rectangle region is composed based on fixed number of fixations occurred around and after to the first fixation. A different approach is proposed by where fixations into fixation density maps that capture spatial density of fixation over the an image. Further, a Gaze Polling layer combines visual feature and fixation density maps for prediction of object categories [5]. However, prior solutions are designed exclusively on physically manipulated objects., that are specific to certain tasks. None of them, fully exploit common cluttered environments that can be found in free living settings. If we consider a scenario where fixations points disposed on the center of an object that is overlapping a second object (i.e keyboard and laptop or book and pencil) both approaches would consider a region that is concurrent to both objects. As result, the input features provided to the object detection model would classify both objects instead of the target one.

D. Our Contributions to Knowledge

The works of [5] and [10] are the most related to the context of our work. Nevertheless, both works only considered synthesized natural scenes in a closed-word setting. In this setting, all objects of interest are carefully place with the right spatial distance between each other to avoid noise during the classification task. In contrast and in the best our knowledge, our work is the first to address the challenge of object detection based on visual attention in cluttered environments.

III. The Proposed Method

In this work we propose to use information from image, fixation location and gaze location for object detection in run time. The features extracted from each domain along with the proposed heuristic function is described below. A schema diagram of the proposed methodology is shown in fig.1.

A. Fixation Guided Object Recognition

The major advantage of our object detection system compared to approaches based only in egocentric cameras, is that in addition to images we also obtain useful information of the user visual gaze provided by the head-worn eyetracker. Common practices to discover object upon user visual attention needs to perform image analysis to locate where the object of interest is. Thus, for instance, when an image is capture in a highly cluttered environment it becomes quite hard to obtain good performance on detection tasks. Unlike such system, we can take the advantage of having fixation data to infer the location of object of interest and attenuate the noise of objects with close spatial distance.

Typical gaze-based object detection system extracts only a region of the image centered on a fixation point, then neural networks models use this region for feature extraction and

object detection. Nevertheless, this approach is not designed to tackle the challenges inherent to cluttered environments. Firstly, because objects have their different sizes and that can also vary with different camera perspectives, meaning that the region extracted might be too small to capture sufficient features of an object. On the other hand, in case of the size of the fixation region is too large it is likely to capture features related to other objects as well. Finally, for instance, in a situation that the fixation point take place on two overlapped objects or between borders of two objects it would not be capable to distinguish which object have the user attention upon.

Therefore, we simply use a completed image for feature extraction in order to not lose any relevant information. In detail, in compute the total image from the egocentric video with the resolution of 1280x720 pixels at 30 FPS. Regarding the technique, we have used convolutional neural networks due to its robustness and high representation power. A pre-trained MobileNet model (trained on ImageNet dataset) is employed for this purpose. The basis of this choice was its efficient trade-off between latency and accuracy.

As described in figure 2, once we obtain a fixation, we match it to the frame image most close in time. Then, for each of the resulting bounding boxes of the image, we count the corresponding gazes points that occurred during the fixation duration period. The decision of representing gaze points instead of just fixation point in a frame, it is due to the nature of cluttered environments. A fixation point is located at the center of gazes region, meaning that in cluttered situation these gazes can math multiple bounding boxes and not only one like if we consider to use only a fixation point. By this, we can have a better representation of user visual attention during a fixation event.

B. Time Series Sliding Window Approach

As it is possible to observe in figure 1, in cluttered environments the boundaries of bounding boxes between objects are often overlapped or very close to each other. In that sense, the task of detecting objects of interest become very hard because if a fixation occurs over an overlapping region, we need to decide in which object does it belongs to. Also, minimal dispersion of users gaze direction due to the subconscious visual attention effect, can provoke a fixation in a region out of the user visual attention. As result, in a cluttered environment it is possible to observe a significant number of false positive errors. For those reasons, we consider critical to analyze not only each of fixation points but all their gaze points in a time series sliding window before and after a fixation event. By this, we can have an detailed overall picture of the user visual action and determinate in which object was the user visual attention upon. To solve the aforementioned problem, we use sliding window time series approach and a Heuristic probabilistic function. Giving an array of N frames, we define a vector $f =< f_(t_0), f_(t_1).., f_(t_N) >$ where $f_(t_i)$ corresponds to the frame captured in timestamp t_i. Then, as can be observed in the figure 3, for each $f_(t_i)$ we define a sequence of $f_(t_i)$ as

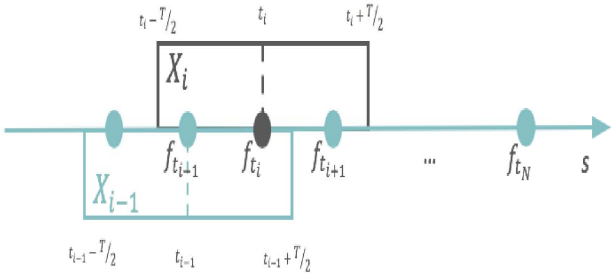

Fig. 3. Example of our time series sliding window approach.

a vector $X_i =< .., f_k, .. >, t_k[t_i - T2, t_i + T2)$ during the period of T. In which i is the number of each frame and T corresponds to 500 milliseconds. Then, we compute the total gazes occurred in sequence as $R = _(k = 0)(X_k)$.

C. Probabilistic Heuristic Function

At this time, for each element f of the vector X we compute the percentage of the interception area between element f and the following elements of X as $P(A \cap I) = \{I \in A \wedge I \in B\}$ for each, A corresponds the target object area, I to the interception area between both objects and B to the area of secondary object .

After this step, we initiate a process that aims to assign the gazes points of overlapped areas to objects by the following rules:

- In case of the $P(A \cap I) \geq T$ and $P(B \cap I) \geq T$: we consider that all the gazes points in the interception area belong to the object with smallest area.
- In case of the $P(A \cap I) \geq T$ and $P(B \cap I) < T$: we determine that the all the gazes points in the interception belong to Object A.
- In case of the $P(A \cap I) < T$ and $P(B \cap I) < T$: the gazes points of the interception area are assigned to both objects.

We define T as threshold T = 70. Since within a sequence X it is possible to find duplicated object classes we create a new vector O, with unique object class and their corresponding gazes within a sequence X as $O =< o_(i_0), o_(i_1).., o_(i_N) >$, $o \in X$. Then, we calculate the percentage of gazes bellowing each object class o during a sequence x as $i \in O.P(o_i) = o_i/(R)$, $O \in X$ with O and R representing respectively the total number of gazes during a sequence. Finally, for each element of the vector o we compare percentage of gazes with a predefined threshold G= 50. The result of this comparison is the following:

- In case $P(o_i) \geq G$: we define that in this particular frame of the sequence, the user visual attention was focused on the object $P(o_i)$.

- In case $P(o_i) < G$: we conclude that there is not enough confidence to attribute user visual attention to a specific object, so we define it as *Null*.

It is also worth to mention that in case of more than one element in the vector O shows percentage greater than G, we define the object class upon user visual attention as *Null*.

IV. Experiment and Results

To evaluate the effectiveness of our approach and given the lack of an appropriate dataset, we designed a new visual attention study to collect visual data over typical cluttered environment. The experiment demonstrate that our approach can outperform commonly state-of-the-art visual attention object detection approaches.

A. Data Collection

In contrast to previous approaches that strategically well-spaced out objects [8][20], our goal was to simulate a cluttered environment characterized by overlapped objects with different sizes and disposed in different positions. In that sense, we set an experiment composed by six different objects and arranged by groups of two objects. For each group, we intentionally disposed objects in way that they get overlapped in different regions and in different magnitudes of space. Our aim was test our approach in a scenario with great variability of overlapping conditions. The experiment settings are shown in 1. Each participant was equipped with a head worn camera and eye trackers to record first person videos and gaze data collectively. To the best of our knowledge, this data set is the first to use point-of-gaze sources in video vision task targeting cluttered environments. During each recording, eight participants were asked to focus their attention upon various objects such as banana, book, laptop, mouse, mobile and cup in the same manner as they do in their live, instead of forcing them to fixate in the center of the objects. In addition, we asked them to pay attention to the contours of each object. In this way, we can explore the impact of the subconscious visual attention effect as well as make sure that some of the gaze points reach the overlapped regions. We used the Pupil Lab eye trackers to record HD resolution first person video at 30 FPS with points-gaze-data at 120Hz. Parameters of fixation detection were left at their defaults where fixation duration was between 100-200 milliseconds. Eye trackers were calibrated before each recording session, and validated by the eye tracker accuracy. As soon as the participants started to paying attention to an object, they pressed a key. The same procedure was repeated when they finished the contour of object visualization. This aimed to obtain ground truth labels of the time intervals where participant looked at a specific object. In detail, for each frame that a fixation occurred we annotated a label based on the object that the user was looking at that specific time interval.

B. Experimental results

In this section, we present the evaluation of our experiments. We calculate the precision based on the results obtained

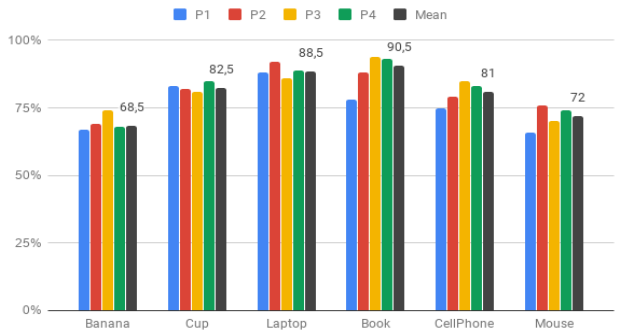

Fig. 4. Results of real-time simulation for each object class. Although precision drops significantly both classes banana and mouse, it remains at an acceptable level.

from each methodology and the ground truth labels. Firstly, we present a comparison of our method with some baseline methods for visual attention object detection. We implement the following two methods for baseline:

1) Central Biases Object Detection: The authors [16]introduced the Central biases object detection methodology which uses a central fixed region in image to extract object features for object detection algorithms. In the experiment, we defined a central region for each captured frame and used Convolutional Neural networks (CNN) object detection algorithm to find object of interest.

2) Fixation Based Object Detection: To provide evidences of the effectiveness of our methodology we implement a simplified version of ours that is also used by the authors in [10][7]. This methodology combines fixations with CNN object detection algorithms to determine object of interest. In the experiment, we manually configure fixations as the unique parameter for object detection, excluding the time series gaze analysis along with the heuristic probabilistic function.

C. Evaluation of Our Method by Participant and Objects

Figure 4 show the results our method accuracy regarding each object class and participants. We can observe that our method in overall performed better on the objects class laptop and book with 88,5% and 90,5% respectively. This significant improvement can be explained by the fact that those objects had major bonding boxes areas over the rest of the objects. So, the majority of user visual gaze tend to be situated outside of overlapped regions. In the opposite direction, can be observed a decrease of the average accuracy among the eight participants on the objects banana and mouse with 68,5% and 72% respectively. One of the reasons that can justify these poor results is the lack of precision of the eye trackers. Small deviations of user gazes can have a negative impact when dealing with reduced bounding boxes areas. Even so, our method shows to be robust against this type of adversities.

D. Evaluation of key parameters on performance

In addition to the previous results, this section shows the influence of different thresholds on the overall accuracy. We

Fig. 5. Results of the influence of the varying the values of threshold T in the performance.

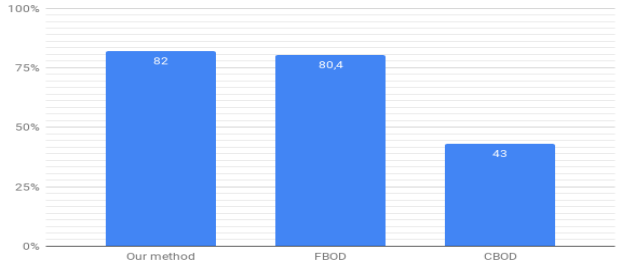

Fig. 6. Comparison of overall accuracy between baseline methods.

tested with different parameters the T threshold of our heuristic function that is responsible to decide within a time window the object of interest based on its percentage of user gazes. Figure 5 shows the average accuracy of our method by changing T for $T1 = 40$, $T2 = 50$, $T3 = 60$, $T4 = 70$, $T5 = 80$. We can observe that $T1$ threshold has achieves the higher accuracy with 84%. On the other, when decreasing this optimal parameter the accuracy drops drastically as can be seen by $T1$.

E. Evaluation of Methods for Object Detection Under Visual Attention

In this subsection, we show the results of our experiment in real-time processing. We compare all the detection methods in the same experiment setting conditions. We did the average accuracy of each method for all the participants. Based on the results in figure 6, we can conclude that our method shows to outperform both centre biases object detection (CBOD) and fixation-based object detection(FBOD) methodologies. We observe that our methodology achieved a mean average of 82%, outperforming by 2% the FBOD method and by 39% the CBOD method. These results indicate that in cluttered environment our method performs significantly better than CBOD and slightly better than FBOD.

V. Discussion

This study points out an important but overlooked issue of visual attention-based object detection in mobile settings. Since most of existent approaches are applied in offline conditions, we started by addressing the challenge of the performing it in real-time settings. In this regard, we demonstrated that salience object detection approach is not suitable for this type of settings, instead we presented an approach based on head worn eye tracker cameras and egocentric video. The change of offline to *real-time* settings introduces the one of the major challenges in this field of study, that is the lack of performance in detecting objects of interest in cluttered environments. We are the first to address this challenge by exploiting visual gaze data analysis with CNN object detection algorithms. We also provide the first mobile data set of viattention in cluttered environment settings. In addition, we also have conducted an in-depth evaluation of our method against the existing widely used methodologies in mobile settings. It is encouraging to see that our methodology can perform well in cluttered environments. It significantly outperforms centre biases-based object detection methodology. As it was expected, our algorithm also slightly outperforms fixation-based object detection methodology thanks to our time series sliding window approach and heuristic probabilistic function. Our experimental results also reveal that our method is very robust against extreme overlapping regions and small objects, showing small difference in its accuracy comparing to objects with reduced overlapping area and large size. Given the technology advance of head-worn eye tracking and emerging interest to mobile computing, we believe that our method can open numerous opportunities for assistive technology studies as well as follow-up visual behaviour research. Assuming that the goal of this paper is to study the detection of object upon the user visual attention in cluttered environment, our experiment was performed with the participants in controlled settings. In the future work, we will evaluate our approach on a novel data set covering free-living settings.

VI. Conclusion

In this work is introduced a new method for discovering objects upon user visual attention. We tackle one of the major challenges in this field that is the poor of performance of existent methods in both real time settings and cluttered environments. To address this problematic we presented a novel method that combines object detection algorithms with egocentric video and user gaze analysis. Our method uses a heuristic probabilistic function as well as time series windows approach to analyze user gazes data around objects in egocentric images. Our results shown that in cluttered environments our method outperforms commonly used methods for detection of object of interest in real time. To evaluate our method we conducted an experiment and constructed an indoor dataset with data annotated of eight participants during visual attention tasks in cluttered scenario. We considered that this dataset fulfill all requirements needed to evaluated our method.

Acknowledgment

This project has received funding from the European Unions Horizon 2020 research and innovation program under the

Marie Skodowska-Curie grant agreement No. 676157, project ACROSSING.

REFERENCES

[1] Joshua Snell, Sebastiaan Mathôt, Jonathan Mirault, and Jonathan Grainger. Parallel graded attention in reading: A pupillometric study. *Scientific Reports*, 8(1):3743, 2018. ISSN 2045-2322. doi: 10.1038/s41598-018-22138-7. URL https://doi.org/10.1038/s41598-018-22138-7.

[2] Dima Amso and Gaia Scerif. The attentive brain: insights from developmental cognitive neuroscience. *Nature Reviews Neuroscience*, 16:606, 9 2015. URL https://doi.org/10.1038/nrn4025http://10.0.4.14/nrn4025.

[3] Dobromir Rahnev, Brian Maniscalco, Tashina Graves, Elliott Huang, Floris P de Lange, and Hakwan Lau. Attention induces conservative subjective biases in visual perception. *Nature Neuroscience*, 14:1513, 10 2011. URL https://doi.org/10.1038/nn.2948http://10.0.4.14/nn.2948https://www.nature.com/articles/nn.2948#supplementary-information.

[4] Author Manuscript and Tract Structures. Visual Attention and Applications in Multimedia Technologies Patrick. 6(9):247–253, 2009. ISSN 08966273. doi: 10.1111/j.1743-6109.2008.01122.x.Endothelial.

[5] Hosnieh Sattar, Andreas Bulling, Mario Fritz, and Max Planck. Predicting the Category and Attributes of Visual Search Targets Using Deep Gaze Pooling.

[6] Gedas Bertasius, Hyun Soo Park, Stella X. Yu, and Jianbo Shi. First Person Action-Object Detection with EgoNet. (1), 2016. ISSN 2330765X. URL http://arxiv.org/abs/1603.04908.

[7] Julian Steil, Michael Xuelin Huang, and Andreas Bulling. Fixation detection for head-mounted eye tracking based on visual similarity of gaze targets. *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications - ETRA '18*, pages 1–9, 2018. doi: 10.1145/3204493.3204538. URL http://dl.acm.org/citation.cfm?doid=3204493.3204538.

[8] Takumi Toyama, Thomas Kieninger, Faisal Shafait, and Andreas Dengel. Gaze guided object recognition using a head-mounted eye tracker. *Proceedings of the Symposium on Eye Tracking Research and Applications - ETRA '12*, (March):91, 2012. ISSN 9781450312219. doi: 10.1145/2168556.2168570. URL http://dl.acm.org/citation.cfm?doid=2168556.2168570.

[9] Jeff Klingner. Measuring cognitive load during visual tasks by combining pupillometry and eye tracking. *Perspective*, (May):130, 2010.

[10] Michael Barz and Daniel Sonntag. Gaze-guided object classification using deep neural networks for attention-based computing. *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing Adjunct - UbiComp '16*, pages 253–256, 2016. doi: 10.1145/2968219.2971389. URL http://dl.acm.org/citation.cfm?doid=2968219.2971389.

[11] Anna Rogalska and Piotr Napieralski. The visual attention saliency map for movie retrospection. *Open Physics*, 16(1):188–192, 2018. doi: 10.1515/phys-2018-0027.

[12] Gkhan Yildirim and Sabine Süsstrunk. Fasa: Fast, accurate, and size-aware salient object detection. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9005(d):514–528, 2015. ISSN 16113349. doi: 10.1007/978-3-319-16811-1{_}34.

[13] Jian Li, Martin D Levine, Xiangjing An, and Hangen He. Saliency Detection Based on Frequency and Spatial Domain Analyses. In *BMVC*, 2011.

[14] A Oliva, A Torralba, M S Castelhano, and J M Henderson. Top-down control of visual attention in object detection. In *Proceedings 2003 International Conference on Image Processing (Cat. No.03CH37429)*, volume 1, pages I–253, 2003. ISBN 1522-4880 VO - 1. doi: 10.1109/ICIP.2003.1246946.

[15] J Wang, A Borji, C . Jay Kuo, and L Itti. Learning a Combined Model of Visual Saliency for Fixation Prediction. *IEEE Transactions on Image Processing*, 25 (4):1566–1579, 2016. ISSN 1057-7149 VO - 25. doi: 10.1109/TIP.2016.2522380.

[16] Yong Jae, Lee Kristen, C V May, and Yong Jae Lee. Predicting Important Objects for Egocentric Video Summarization. (January), 2015.

[17] Gedas Bertasius and Hyun Soo Park. Exploiting Egocentric Object Prior for 3D Saliency Detection.

[18] Benjamin J Tamber-Rosenau and Ren Marois. Central attention is serial, but midlevel and peripheral attention are parallel-A hypothesis. *Attention, perception & psychophysics*, 78(7):1874–1888, 10 2016. ISSN 1943-393X. doi: 10.3758/s13414-016-1171-y. URL https://www.ncbi.nlm.nih.gov/pubmed/27388496https://www.ncbi.nlm.nih.gov/pmc/PMC5014686/.

[19] Andrew T Duchowski. A breadth-first survey of eye-tracking applications. *Behavior Research Methods, Instruments, & Computers*, 34(4):455–470, 2002. ISSN 1532-5970. doi: 10.3758/BF03195475. URL https://doi.org/10.3758/BF03195475.