Gaze-guided Object Classification using Deep Neural Networks for Attention-based Computing

Michael Barz
German Research Center for Artificial Intelligence (DFKI)
Stuhlsatzenhausweg 3, 66123 Saarbruecken
michael.barz@dfki.de

Daniel Sonntag
German Research Center for Artificial Intelligence (DFKI)
Stuhlsatzenhausweg 3, 66123 Saarbruecken
sonntag@dfki.de

Ubicomp/ISWC'16 Adjunct, September 12-16, 2016, Heidelberg, Germany
ACM 978-1-4503-4462-3/16/09.
http://dx.doi.org/10.1145/2968219.2971389

Abstract
Recent advances in eye tracking technologies opened the way to design novel attention-based user interfaces. This is promising for pro-active and assistive technologies for cyber-physical systems in the domains of, e.g., healthcare and industry 4.0. Prior approaches to recognize a user's attention are usually limited to the raw gaze signal or sensors in instrumented environments. We propose a system that (1) incorporates the gaze signal and the egocentric camera of the eye tracker to identify the objects the user focuses at; (2) employs object classification based on deep learning which we recompiled for our purposes on a GPU-based image classification server; (3) detects whether the user actually draws attention to that object; and (4) combines these modules for constructing episodic memories of egocentric events in real-time.

Author Keywords
eye tracking; gaze-based interaction; object classification; visual attention

ACM Classification Keywords
H.5.m. [Information Interfaces and Presentation (e.g. HCI)]: Miscellaneous

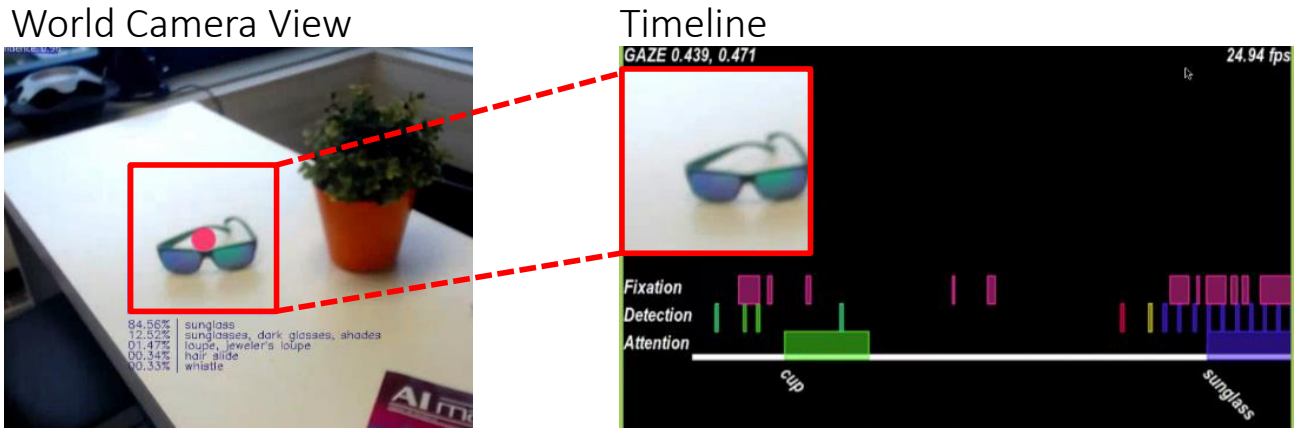

Figure 1: Illustration of the sample application comprising a part of the world camera view (left) and the timeline application (right). The world view shows the current gaze estimate (red dot), the image patch used for classification (red rectangle) and the actual top-5 results. The sample application shows visual attention events augmented with the corresponding label.

Introduction

Human gaze naturally indicates visual attention and thus the interest of a user [10]. Further, recent advances in head-worn eye tracking equipment renewed interest in pervasive eye tracking and mobile measurement of attention [2]. To this end, eye tracking shows great promise for improving or creating models of the user's context for situation-aware interaction. However, existing methods are often limited; for example, they only use high level contextual cues [3] or a sensor-equipped environment [10]. More advanced approaches to determine a user's visual attention include gaze-guided object recognition on video images of egocentric cameras [8]. Most recently, gaze-guided object recognition has been used to create digital episodic memories to support people that suffer from mental disorders, for example dementia patients [9, 6].

Limitations of those state-of-the-art systems concern scalability issues; they allow only for moderate class sizes and little discriminative power (i.e., 10-20 object classes) and must be run on a single computer.

We propose a method that enables gaze-guided object classification by a scalable and powerful decentralized object recognizer based on a state-of-the-art deep learning framework and a light-weight head-mounted eye tracker. We analyse image patches around the user's gaze position from an egocentric camera to determine the object of interest. Finally, we create sequences of *attention events* that provide information about the user's intentions. The history of such events, which is built up continuously, is suitable for assistive technologies, e.g., in the context of activity recognition or everyday memory support [7].

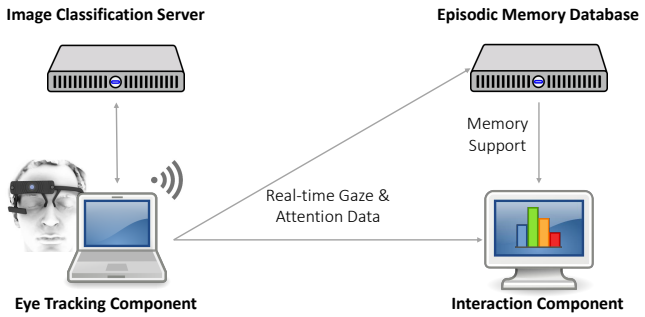

Figure 2: Hardware overview: the system includes an eye tracking component, an image classification server, an episodic memory database, and a generic interaction component.

Approach

The proposed system consists of four parts, an eye tracking component to connect a head-mounted eye tracker, an image classification server powered by two graphics processing units (GPU), an episodic memory database to store event sequences, and a generic interaction manager component that receives all generated data (see Figure 2).

Eye Tracking Component

We use a Pupil eye tracker with an egocentric camera and an eye camera in a lightweight package [5]. Calibration and gaze estimation features are provided in the vendor's standard distribution. We extended it by fixation detection, visual attention detection of a certain object, and in addition, broadcast resulting data via network capabilities in real-time. For fixation detection, we use a dispersion based algorithm similar to Barz et al. [1]. To detect visual attention to a certain object, we adopted the threshold-based algorithm proposed by Toyama et al. [8]. In contrast to their SIFT-based object recognition, we use an image classification model based on deep learning running on an additional high performance computer connected to the system.

Image Classification Server

The image classification module is based on a pre-trained model[1] using deep neural networks comprising more than 1000 classes (nodes in output layer). For classification, we use the caffe framework [4] as a background service and offer a REST API for remote function calls. As input, it takes images of size 256×256 pixels and reports the probability of each class as output, with a top-1 accuracy of 68.7% and a top-5 accuracy of 89%. For our system we defined a probability threshold for attention detection of 20%.

Episodic Memory Database

This component stores data about the user's attention similar to the actual human episodic memory [9]. The sequence of events can be used to compensate loss of memory, e.g., caused by mental disorders such as dementia. Utilizing this information can be the target of future work.

Interaction Component

The interaction component is a generic node that offers gaze as input modality for, e.g., ambient gaze-enabled systems [1]. We implemented a sample application that visualises all incoming data in real-time in terms of a timeline. In particular, we show fixations, object detection results such as the user's visual attention including the image patch and its label (see Figure 1). The top-1 result as shown in the world camera view ('sunglass') corresponds to the most recent object detection event above the attention

[1] BVLC GoogLeNet
http://caffe.berkeleyvision.org/model_zoo.html

event box. Previously, the system detected attention to a cup (most recent events occur on the right).

Conclusion

In this work we presented a system for real-time gaze-guided object classification based on a state-of-the-art deep learning framework. Our system enables assistive technologies based on a digital episodic memory with manifold fields of application. Interesting scenarios are to support people in recalling information [7]. A limitation of our system is that we use the most probable result of the image classification only. Incorporating the best five results might increase performance of our system (*top-5* accuracy is higher by 20.3% compared to *top-1* performance).

REFERENCES

1. Michael Barz, Florian Daiber, and Andreas Bulling. 2016. Prediction of gaze estimation error for error-aware gaze-based interfaces. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications (ETRA '16)*. 275–278. DOI: http://dx.doi.org/10.1145/2857491.2857493
2. Andreas Bulling. 2016. Pervasive Attentive User Interfaces. *IEEE Computer* 49, 1 (jan 2016), 94–98. DOI:http://dx.doi.org/10.1109/MC.2016.32
3. Andreas Bulling, Christian Weichel, and Hans Gellersen. 2013. EyeContext: Recognition of High-level Contextual Cues from Human Visual Behaviour. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 305–308. DOI: http://dx.doi.org/10.1145/2470654.2470697
4. Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional Architecture for Fast Feature Embedding. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 675–678. DOI: http://dx.doi.org/10.1145/2647868.2654889
5. Moritz Kassner, William Patera, and Andreas Bulling. 2014. Pupil: An Open Source Platform for Pervasive Eye Tracking and Mobile Gaze-based Interaction. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication (UbiComp '14 Adjunct)*. 1151–1160. DOI:http://dx.doi.org/10.1145/2638728.2641695
6. Alexander Prange, Takumi Toyama, and Daniel Sonntag. 2015. Towards Gaze and Gesture Based Human-Robot Interaction for Dementia Patients. In *2015 AAAI Fall Symposium Series*. 111–113.
7. Daniel Sonntag. 2014. ERmed – Towards Medical Multimodal Cyber-Physical Environments. In *Foundations of Augmented Cognition. Advancing Human Performance and Decision-Making through Adaptive Systems*. Springer, 359–370. DOI: http://dx.doi.org/10.1007/978-3-319-07527-3_34
8. Takumi Toyama, Thomas Kieninger, Faisal Shafait, and Andreas Dengel. 2012. Gaze guided object recognition using a head-mounted eye tracker. In *Proceedings of the Symposium on Eye Tracking Research and Applications*. 91–98. DOI: http://dx.doi.org/10.1145/2168556.2168570
9. Takumi Toyama and Daniel Sonntag. 2015. *Towards Episodic Memory Support for Dementia Patients by Recognizing Objects, Faces and Text in Eye Gaze*. Springer, 316–323. DOI: http://dx.doi.org/10.1007/978-3-319-24489-1_29
10. Roel Vertegaal. 2003. Attentive User Interfaces. *Commun. ACM* 46, 3 (2003), 30–33. DOI: http://dx.doi.org/10.1145/636772.636794