

Semi-automatic training of an object recognition system in scene camera data using gaze tracking and accelerometers

M. Cognolato^{1,2*}, M. Graziani^{3*}, F. Giordaniello³, G. Saetta⁴, F. Bassetto⁵,
P. Brugger⁴, B. Caputo³, H. Müller¹ and M. Atzori¹

¹ University of Applied Sciences Western Switzerland (HES-SO), Sierre, Switzerland.
`matteo.cognolato@hevs.ch`.

² Rehabilitation Engineering Laboratory, ETH Zurich, Zurich, Switzerland.

³ University of Rome “La Sapienza”, Rome, Italy.

⁴ Department of Neurology, University Hospital of Zurich, Zurich, Switzerland.

⁵ Clinic of Plastic Surgery, Padova University Hospital, Padova, Italy.

* The first two authors contributed equally to this work.

Abstract. Object detection and recognition algorithms usually require large, annotated training sets. The creation of such datasets requires expensive manual annotation. Eye tracking can help in the annotation procedure. Humans use vision constantly to explore the environment and plan motor actions, such as grasping an object.

In this paper we investigate the possibility to semi-automatically train object recognition with eye tracking, accelerometer in scene camera data, learning from the natural hand-eye coordination of humans. Our approach involves three steps. First, sensor data are recorded using eye tracking glasses that are used in combination with accelerometers and surface electromyography that are usually applied when controlling prosthetic hands. Second, a set of patches are extracted automatically from the scene camera data while grasping an object. Third, a convolutional neural network is trained and tested using the automatically extracted patches.

Results show that the parameters of eye-hand coordination can be used to train object recognition automatically. These can be exploited with proper sensors to **fine-tune** a convolutional neural network for object detection and recognition automatically. This approach opens interesting options to train computer vision and multi-modal data integration systems **and lays the foundations for future applications in robotics**. In particular, this work targets the improvement of prosthetic hands by recognizing the objects that a person may wish to use. However, the approach can easily be generalized.

Keywords: semi-automatic training; object recognition; eye tracking

1 Introduction

Grasping an object is a complex task with several senses being involved. Vision provides the information needed to precisely control the hand and perform the

task. Vision precedes the grasping action according to precise schemes called “hand-eye coordination” [4, 8]. Thus, the parameters related to eye-hand coordination can be used to provide information about the object that the subject is aiming to grasp. Object recognition and detection have been strongly improved during the last years, also thanks to the application of deep learning and Convolutional Neural Networks (CNNs) [20] that require much training data for optimal results. The creation of annotated datasets is most often done manually, which is expensive and time consuming. It was recently shown that eye tracking can simplify the annotation procedure [18, 22].

Human gaze interacts with the surrounding reality in several ways. Gaze alternates between fixations (when the subject’s gaze is stable on a portion of the scene) and saccades (when eyes and/or the body are moved to look somewhere else). The advancement of eye/gaze tracking devices allowed to identify where a subject is looking in real time and in many scenarios. Thanks to these devices, it was shown that the gaze precedes and guides the hand also when performing grasps in routine tasks [16, 12]. In the robotic literature the challenge of grasping various objects while regulating the control with visual information has been covered extensively. The object to be grasped is looked at 40-100 ms before the movement [9, 2, 19] and the fixation lasts for 350-450 ms [14, 5]. During the fixation, the subjects attempts to detect the affordance of an object (the physical possibility of an action on the object). As Bohme and Heinke described in [3], the gaze naturally converges to the grasping point of tools [6]. Eye tracking and gaze information have already been introduced successfully in the control of manual prehension and object detection/recognition [22, 18]. In Mishra et al. [17], the authors propose active segmentation methods using the fixation points in the image. Papadopoulos et al. [18] present a gaze-based method to annotate training sets for object detection. In Toyama et al. [22] the gaze information provided by a head-mounted eye tracker is used to perform real time object recognition.

However, the parameters related to the natural hand-eye coordination have never been used to train an object recognition system automatically. In this paper we investigate if this approach can be used to create a training dataset for object classification automatically. The approach is tested on data acquired to investigate both vision and kinematic aspects of grasping, with the aim of improving the control of a myoelectric hand prosthesis. **The test consisted of intact subjects grasping various object with several grasps. [11]** The data are acquired using head-mounted eye tracking with a scene camera and sEMG electrodes containing accelerometers. The dataset is then used to **fine-tune** and test a CNN to perform object recognition. Currently, the approach is designed and tested to be used to improve the control of a myoelectric hand prosthesis. **Object recognition can make the prosthesis capable to autonomously understand the required grasp. Thus, it can improve control robustness.** On the other hand, the same approach can easily be extended to create a dataset for object detection by annotating the position of the objects in the scene with a bounding-box.

2 Data Acquisition

2.1 Acquisition Setup

The acquisition setup is designed to support the experiment investigating the grasp of several objects. The setup is composed of acquisition hardware and software. The acquisition hardware is composed of a set of surface electromyography electrodes (Delsys Trigno Wireless EMG); a pair of eye tracking glasses with scene camera (Tobii Pro Glasses 2) and a laptop (Dell Latitude E5520). The acquisition software simultaneously records and stores all the data provided by the sensors.

The Delsys Trigno Wireless System consists of 14 electrodes, each equipped with a triaxial accelerometer. It records the surface EMG signal at 2 kHz and the accelerometers at 148 Hz with a baseline noise lower than 0.5 mV RMS (Root Mean Square). The electrodes are placed around the forearm of the subject using a dense sampling approach to record the activity of the muscles controlling the hand. Eight electrodes are placed around the forearm starting from the radio-humeral joint, forming a circular array. The other six electrodes are placed after the first eight, creating a second array in a more distal position, as shown in (Figure 1). These electrodes are placed in correspondence to the gap between the electrodes of the first array, starting from the gap between the first and second electrodes. A latex-free elastic band is used to maintain the electrodes in contact with the skin. The Tobii Pro Glasses 2 are a mobile lightweight gaze tracker recording both point-of-regard and scene in front of the subject. The gaze point data are sampled at 100 Hz with a theoretical accuracy and RMS precision of 0.5 and 0.3 degrees respectively. The scene in front of the subject is recorded with a scene camera embedded in the frame in full HD resolution at 25 fps (frames per second). Finally, the laptop manages hardware connections and runs the acquisition software that guides the subjects during the exercise, while recording and storing the data from all the devices.

The acquisition software is a multithreaded application based on the producer-consumer pattern and developed in C++. To synchronize the data from the various sensors, a high resolution timestamp is assigned to each sampled datum. A Graphical User Interface (GUI) developed with Qt is used to guide the subject with vocal instructions.

2.2 Acquisition Protocol

The test was designed to investigate both kinematic and visual aspects of grasping. **The aim is to robustly identify the grasp that the subject wants to perform. Each grasp is performed on more objects and, if applicable, more grasps are performed on the same object.** Therefore, it mainly consists of grasping 30 objects with 15 types of grasps. Grasps and objects are selected based on a grasp taxonomy [10] and their importance in Activities of Daily Living (ADL). The considered grasps together with the used objects are reported in the columns of Figure 2.

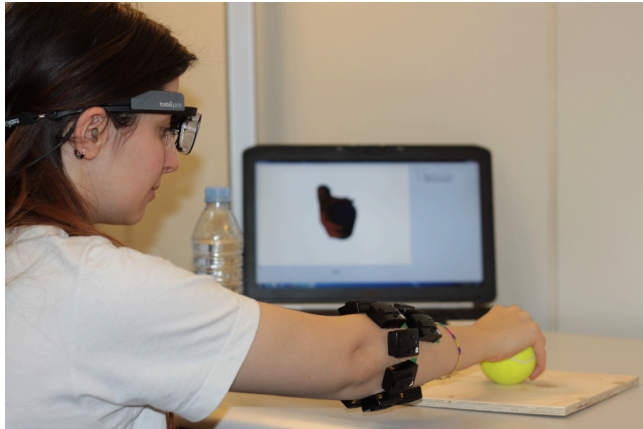


Fig. 1: Overview of the acquisition setup.

The subject is asked to sit comfortably in front of a desk where the acquisition laptop and the objects to be grasped are positioned. Before each set of grasp repetitions two videos from **lateral and first person** points of view explain to the subject how to perform the grasp. During this phase, the subject can try the grasp on the objects in order to get confident with them. Afterwards, a fixed image of the grasp is shown on screen in order to minimize the distractions during the exercise. The subject is guided by two vocal instructions: the first asking to grasp the object and the second to release it and to return to the rest position. The instructions have the same duration (four seconds) for both grasping and resting and they are recorded as stimulus signal. The exercise can be viewed as composed of five phases: 1) rest (arm comfortably leaning on the desk); 2) movement to reach an object; 3) grasp of an object; 4) release; 5) return to rest. Each grasp is repeated 12 times, while the number of repetitions performed on each object depends on the number of objects used for the specific grasp.

In order to avoid staring at objects between grasps, the subjects were asked to look at the eye tracker calibration target during the resting phase. The acquisition protocol was tested on 6 healthy subjects: 4 males, 2 females, average age 26.6 ± 5.3 , all right handed.

3 Data Analysis

In this section we present the approach used to train a computer vision system automatically with eye tracking and accelerometers using the scene camera data. Learning is based on the natural parameters of hand-eye coordination. The section is divided into two parts: the first describes the automated creation of the training dataset for object classification; the second describes the **fine-tuning** of the CNN using the automatically created training dataset and the evaluation on the test data.





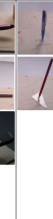







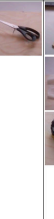
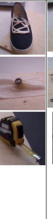
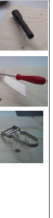
Grasp	1	2	3	4	5&6	7	8	9	10	11	12	13	14	15	
Description	large diameter	small diameter	index finger extension	medium wrap	prismatic 4 fingers & writing tripod	power sphere	precision sphere	lateral	parallel extension	tripod grasp	power disk	using scissors	palmar pinch	adhesive thumb	
Objects															

Fig. 2: Overview of the objects and the grasps performed.

3.1 Automatic Creation of the Training Dataset

Several papers have already investigated the parameters of the coordination between gaze and grasping within eye-hand coordination [14, 2, 5, 16]. They provide information that can be used to localize objects in the scene by detecting the beginning of the movement and the gaze fixations. Thus, we investigated and evaluated the feasibility of using these parameters to automatically extract video patches containing the object that the subject is aiming to grasp. These patches are then used to **fine-tune** a CNN and tested on the object recognition task.

The video patches containing the object that the subject is aiming to grasp are identified by first detecting the beginning of the movement that the subject performs to reach the object. Based on this, the gaze fixations related to the object to be grasped are used. We investigated two approaches to identify the beginning of the movement: the movement of the arm towards the object, identified with the accelerometer or when the hand started to be pre-shaped for grasping, identified as finger movement with the sEMG signals. However, preliminary analysis showed that the accelerometer performed better than the sEMG. Thus, it was decided to use the movement of the arm towards the object using the accelerometer signals. First, the data acquired are preprocessed and synchronized following the procedure described in [1].

In order to detect the beginning of the movement to reach the object, the forearm was considered as a rigid body to which all the sensors are firmly attached. The magnitude of the 3-axis acceleration vector is calculated for each sensor. The signals obtained are averaged and the results denoised using a db2 wavelet of level 9. The beginning of the movement is identified as the maximum value of the signal obtained right after the resting of each repetition (identified thanks to the stimulus signal).

The eye fixation on the object to be grasped can be identified based on the parameters describing the eye-hand coordination. Following the results presented

in the literature, the fixations are identified between 50 ms before and 450 ms after the beginning of the movement using the EyeMMV Toolbox [15]. A fixation cluster is identified when the gaze of the subject is remaining within a radius of 250 px for more than 70 ms. The fixation coordinates are calculated taking into account only the points with a fixed distance regarding the mean point of the cluster lower than $3s$, with $s = (s_x + s_y)^{1/2}$, s_x and s_y standard deviation of horizontal and vertical coordinates.

The identification of the fixation allows the extraction of patches, centered around the fixation point and with a size of 512 by 512 pixels. The patches are extracted from the first two frames of the scene camera video after the beginning of the identified fixation. The labels of the patches, indicating which object the subject was asked to grasp, are automatically assigned using the information provided by the stimulus signal. These patches are then used to **fine-tune** a CNN to recognize the object that the subject is aiming to grasp. A flow diagram of the proposed method is showed on Figure 3.

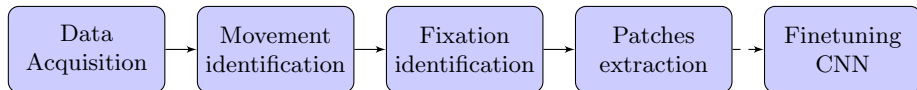


Fig. 3: Flow diagram of the proposed method.

3.2 Evaluation of the Training Dataset on Object Recognition

As reported in section 1, CNNs are currently obtaining good results in many recognition tasks, including object recognition. The training data automatically extracted are tested using one of the CNN architectures that has shown the best results on both image classification and single object localization: the GoogLeNet network [20]. The 2158 patches extracted (subsection 3.1) were divided into 30 classes representing the objects. The object recognition was performed using the Berkeley Vision and Learning Center (BVLC) GoogleNet model [21] pre-trained on ImageNet Large Scale Visual recognition Competition (ILSVRC) 2012 [7] using the deep learning framework caffe 1.0.0-rc3 [13] **together with the framework NVIDIA Deep Learning GPU Training System (DIGITS)**. To evaluate the performance, a leave-one-person-out approach was used. The **fine-tuning** of the network is performed with the patches of 5 subjects and tested on the patches of the subject left out. The training and validation datasets are composed of 1438 and 360 (20%) patches on average resized to 256 px respectively. The test datasets were containing the patches extracted from a single subject (360 patches on average).

4 Results

The object top 1 classification accuracy is $71.97\% \pm 2.67\%$, considering the 30 classes (range between 68.33% and 75.56%). The probability of having the correct

class in the first 5 results (indicated as top 5 accuracy in Table 1) is, $82.53\% \pm 1.77\%$, with a range of 3.33% ($81.39\% - 84.72\%$). The detailed results of the classification for each subject are reported in Table 1.

The average per-class accuracy (accuracy achieved on each object class) highlights how various objects can be more easily and robustly recognized by the detector. As shown in Table 2, the *remote control* is the object recognized with the best accuracy ($91.67\% \pm 12.91\%$), while *sunglasses* is the one with the lowest per-class accuracy ($43.61\% \pm 15.00\%$). The *button* is the object with the highest classification uncertainty ($\pm 49.16\%$ of standard deviation) while the *belt* is the one classified with highest precision ($\pm 4.19\%$ of standard deviation).

The per-class accuracies for all subjects (30 objects for 6 subjects) range from 0% to 100%. In 4 cases ($\sim 2\%$) the object is recognized in none of the tested images, so with a per-class accuracy of 0%. This happened for classes: *button* for Subject 1, *cardboard cup* for Subject 2, and *button* and *peeler* for Subject 3. On the other hand, in 44 cases ($\sim 24\%$) the object is recognized in all tested patches, so a per-class accuracy of 100%.

5 Discussion and Conclusion

The results show that the parameters of hand-eye coordination can be used to semi-automatically create a training dataset for object recognition from eye tracking, accelerometer and scene camera data. Several approaches can be applied to obtain further improvements.

The application of the eye-hand coordination parameters to gaze and accelerometer data allows the identification of the time interval during which the subject is looking at the object to grasp. This information allows to automatically identify the object within a video or image of the scene. The described procedure can be applied to automatically create a training dataset, **in particular in application involving the eye-hand coordination, such as myoelectric hand prosthesis**. A trained system can then be used to perform object recognition, for example to automatically identify the object that the subject is aiming to grasp. This can be useful in several applications and fields such as in robotics and pros-

Table 1: Top 1 and Top 5 accuracy on object identification for each subject.

	Top 1 Accuracy [%]	Top 5 Accuracy [%]
Sbj 1	70.00	80.56
Sbj 2	72.78	81.94
Sbj 3	68.33	81.39
Sbj 4	74.02	81.84
Sbj 5	75.56	84.72
Sbj 6	71.11	84.72
Average Accuracy	71.97	82.53

thetics where movements can be adapted to objects to be grasped. The same approach can also be used to train or fine-tune object detectors by automatically drawing a bounding-box around the objects to annotate their position in the scene. It can be used to create a training set for object detection or recognition in a real world scenario, with the objects represented in various orientations and positions. Moreover, it can be also extended in an unsupervised scenario in which the ground truth is given by saying the object class and then transcribed by speech recognition.

A few aspects negatively influenced the results, in particular the quality of the patches and the small size of the dataset. The main limitation of the automatic extraction of the patches is related to object representation. **Many external factors such as distraction or adaptation to the task (decreasing the level of visual attention) may have an influence on patch extraction.** In some cases, the patches can contain: part of the object; the object being occluded by the subject’s hand; the object to be grasped and parts of the other objects in the scene and background. In the worst case a patch does not contain any object (Figure 4). From visual inspection we noticed that approximately one third of the patches extracted are not containing the object or just a small portion of it. Increasing the quality of the patches can most likely increase the performance of the classification. This can be done by segmenting the objects in the scene (for example, using an objectiveness filter or a Region Proposal Network). The gaze information may then be used to select the object that the subject seems to be interested in, i.e. selecting the region of the object containing the gaze or the closest one. This can lead to a more robust and precise patch extraction, avoiding the mentioned problems. The acquisitions were made in the same location

Table 2: Average per-class object classification accuracies.

Object	Per-class Accuracy [%]	Object	Per-class Accuracy [%]
belt	68.75 ± 4.19	razor	55.56 ± 27.22
book	69.45 ± 24.53	remote control	91.67 ± 12.91
fork	74.82 ± 17.94	scissor	78.46 ± 16.21
cardboard cup	63.89 ± 37.14	screwdriver	78.94 ± 10.74
hairbrush	68.10 ± 37.05	shoe	45.00 ± 20.74
key	65.65 ± 22.22	cell phone	81.25 ± 15.31
knife	79.17 ± 24.58	torch	72.92 ± 16.61
button	58.33 ± 49.16	zip	61.90 ± 26.60
mug	68.06 ± 26.04	measuring tape	87.92 ± 6.74
mouse	75.00 ± 29.35	can	80.56 ± 12.54
peeler	66.67 ± 40.83	cup	60.00 ± 21.91
pen	74.00 ± 16.00	disk	56.25 ± 24.69
bottle	80.00 ± 25.30	handle	70.83 ± 29.23
pencil	77.65 ± 10.89	ball	76.25 ± 33.68
plate	75.00 ± 22.36	sunglasses	43.61 ± 15.00

using the same setup and objects. This aspect that can influence the classification accuracy as well as limit the portability of the approach. On the other hand, it does not reduce the importance of the presented approach. Increasing both size and variability of the dataset can also help to recognize a larger number of objects in various conditions, such as changing light. Moreover this method can easily be applied in a semi-supervised manner, setting a threshold for the classification accuracy below which the system requires human intervention to define the correct object in the scene.

The information provided by the eye-hand coordination was successfully used to automatically create a training dataset for object detection and classification starting from eye tracking, accelerometer and scene video data.

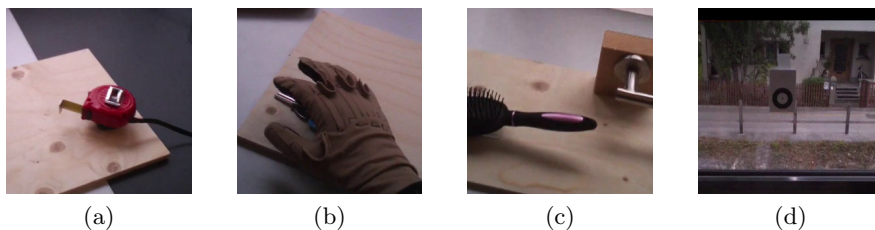


Fig. 4: (a) Object correctly represented in the patch. (b) Object being occluded by the subject’s hand. (c) Two objects partially represented in the patch. (d) The patch does not contain the object.

Acknowledgments The authors would like to thank A. Gigli, A. Gijsberts and V. Gregori from the University of Rome “La Sapienza” for their help on pre-processing the data and their helpful suggestions.

References

- [1] Atzori, M., Gijsberts, A., Castellini, C., Caputo, B., Hager, A.G.M., Elsig, S., Giatsidis, G., Bassetto, F., Müller, H.: Electromyography data for non-invasive naturally-controlled robotic hand prostheses. *Scientific data* 1 (2014)
- [2] Biguer, B., Jeannerod, M., Prablanc, C.: The coordination of eye, head, and arm movements during reaching at a single visual target. *Experimental Brain Research* 46(2) (1982)
- [3] Böhme, C., Heinke, D.: Where do we grasp objects? - An experimental verification of the Selective Attention for Action Model (SAAM). In: *Lecture Notes in Computer Science*. vol. 5395 LNAI (2009)
- [4] Bulloch, M.C., Prime, S.L., Marotta, J.J.: Anticipatory gaze strategies when grasping moving objects. *Experimental Brain Research* 233(12) (2015)
- [5] Castellini, C., Sandini, G.: Gaze tracking for robotic control in intelligent teleoperation and prosthetics. *Proceedings of {COGAIN} - Communication via Gaze Interaction (November 2014)* (2006)

- [6] Connolly, J.D., Goodale, M.a.: The role of visual feedback of hand position in the control of manual prehension. *Experimental brain research. Experimentelle Hirnforschung. Expérimentation cérébrale* 125(3) (1999)
- [7] Deng, J.D.J., Dong, W.D.W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition* (2009)
- [8] Desanghere, L., Marotta, J.J.: "Graspability" of objects affects gaze patterns during perception and action tasks. *Experimental Brain Research* 212 (2011)
- [9] Desanghere, L., Marotta, J.J.: The influence of object shape and center of mass on grasp and gaze. *Frontiers in Psychology* 6(OCT) (2015)
- [10] Feix, T., Pawlik, R., Schmiedmayer, H.B., Romero, J., Kragi, D.: A comprehensive grasp taxonomy. *Robotics, Science and Systems Conference: Workshop on Understanding the Human Hand for Advancing Robotic Manipulation* (2009)
- [11] Giordaniello, F., Cognolato, M., Graziani, M., Gijbets, A., Gregori, V., Saetta, G., Hager, A.g.M., Tiengo, C., Bassetto, F., Brugger, P., Müller, H., Atzori, M.: Megane Pro: myo-electricity , visual and gaze tracking data acquisitions to improve hand prosthetics. *IEEE 15th International Conference on Rehabilitation Robotics (ICORR)*. Accepted. (2017)
- [12] Hayhoe, M.: Vision Using Routines: A Functional Account of Vision. *Visual Cognition* 7(1997) (2000)
- [13] Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional Architecture for Fast Feature Embedding. *Proceedings of the ACM International Conference on Multimedia* (2014)
- [14] Johansson, R.S., Westling, G., Bäckström, A., Flanagan, J.R.: Eye-hand coordination in object manipulation. *The Journal of neuroscience : the official journal of the Society for Neuroscience* 21(17) (2001)
- [15] Krassanakis, V., Filippakopoulou, V., Nakos, B.: EyeMMV toolbox: An eye movement post-analysis tool based on a two-step spatial dispersion threshold for fixation identification. *Journal of Eye Movement Research* 7(1)(1) (2014)
- [16] Land, M.F.: Eye movements and the control of actions in everyday life. *Progress in Retinal and Eye Research* 25(3) (2006)
- [17] Mishra, A., Aloimonos, Y., Fah, C.L.F.C.L.: Active segmentation with fixation. *IEEE 12th International Conference on Computer Vision* (2009)
- [18] Papadopoulos, D.P., Clarke, A.D.F., Keller, F., Ferrari, V.: Training object class detectors from eye tracking data. In: *European Conference on Computer Vision*. pp. 361–376 (2014)
- [19] Pinpin, L.K., Johansson, R.S., Laschi, C., Dario, P.: Gaze interface: Utilizing human predictive gaze movements for controlling a HBS. *Proceedings of the 2nd Biennial IEEE/RAS-EMBS International Conference on Biomedical Robotics and Biomechatronics* (2008)
- [20] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision* 115(3) (2015)
- [21] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2015)
- [22] Toyama, T., Kieninger, T., Shafait, F., Dengel, A.: Gaze guided object recognition using a head-mounted eye tracker. *ETRA '12 Proceedings of the Symposium on Eye Tracking Research and Applications* 1(212) (2012)