Exploitation of Gaze Data for Photo Region Labeling in an Immersive Environment

Tina Walber[1], Ansgar Scherp[2], and Steffen Staab[1]

[1]University of Koblenz-Landau, Germany
{walber,staab}@uni-koblenz.de
[2]University of Mannheim, Germany
ansgar@informatik.uni-mannheim.de

Abstract. Metadata describing the content of photos are of high importance for applications like image search or as part of training sets for object detection algorithms. In this work, we apply tags to image regions for a more detailed description of the photo semantics. This region labeling is performed without additional effort from the user, just from analyzing eye tracking data, recorded while users are playing a gaze-controlled game. In the game *EyeGrab*, users classify and rate photos falling down the screen. The photos are classified according to a given category under time pressure. The game has been evaluated in a study with 54 subjects. The results show that it is possible to assign the given categories to image regions with a precision of up to 61%. This shows that we can perform an almost equally good region labeling using an immersive environment like in *EyeGrab* compared to a previous classification experiment that was much more controlled.

1 Introduction

The management of digital images is a challenging task, and it is often performed based on metadata. For example, image search makes use of tags manually assigned to images or extracted from surrounding text information on web pages. A more detailed description of photo contents by region labeling can improve the search [4]. Different approaches were investigated for creating region labels. On the one hand, fully automatic approaches are far from delivering results that are on the level of human understanding of visual content [11]. On the other hand, manual labeling is a tedious task for users. The general idea behind our approach is to create image meta information without additional effort from the user. To reach this goal, we exploit the information gained from eye movements, while the user is viewing photos in the context of a specific task. In our first work [15], the data was collected in a controlled experiment. In this first experiment, a tag was first presented to the user, and afterwards, he/she had to decide whether an object described by this tag could be seen on the photo by pressing a key on the keyboard. We obtained a maximum precision of 65% at pixel level from comparing the calculated regions with manually created ground truth regions. In this work, we evaluate if region labeling is also possible in a very different

scenario, while the user is playing a game. As in the first experiment, the task is to decide whether an object, belonging to a given category, can be seen on a photo. While in the first study, the user had no time constraints and the photos were displayed full screen, the game *EyeGrab* [13] was developed to demand fast decision making from the participants and to break up the full concentration on photo viewing by bringing the user into the immersive situation of a game with distractions from the game setup, the gaze control, and the emotional pressure of success and failure. In *EyeGrab*, users classify and rate photos falling down the screen. Photos are selected by fixation them fixated with the eyes. Subsequently, the classification is performed by fixating specific objects on the screen. In addition, the classification comprises a personal rating of the photo. By analyzing the recorded gaze paths, we are able to automatically assign the given category, which describes a specific object like "car" or "tree" to an image region. To evaluate our approach, we have collected gaze data in a user study with 54 participants. All photos used in our evaluation had ground truth information concerning their classification and the depicted objects and image regions, respectively. We can state that the level of difficulty playing *EyeGrab* was not too high as only 7% of the shown images passed without classification and 90% of the classifications were correct with respect to the given category. In order to assign a given category to an image region, we apply two gaze measures and a baseline [15]. The measures predict which region of the photo is assumed to show an object, belonging to the category. This region is compared at pixel level with the ground truth image region from our data set. From the data collected in *EyeGrab*, we obtain a maximum precision of 61% of correctly labeled image region pixels. Thus, region labeling in the immersive environment of a game performs almost equally well as in previous work, where we considered non-moving full-screen images and could predict the regions with precision of about 65% [15]. We have also investigated different falling speed levels for analyzing the influence of speed in the region labeling results. We got a slightly higher number of photos passing without classification or those classified incorrectly for faster speed levels, but only small variations in the precision of the region labeling are observed. Overall, this study shows that one can obtain good region labels in an immersive game environment and that the results are comparable to those where the images were not moving and the experiment was much more restricted.

Please note that we provide the experiment images and gaze data on `http://west.uni-koblenz.de/Research/DataSets/gaze`.

2 Related Work

One approach for gaining information from gaze data is relevance feedback in image search. Kozma et al. [5] compared image selection by implicit gaze feedback with explicit user feedback by clicking on relevant images. Gaze information in combination with image segmentation also provides valuable information for photo cropping [9]. Klami et al. [3] identified heat-map-like image regions relevant in a specific task using gaze information. The given task is very general,

and thus, the work does not aim at identifying single objects in the images from the generated heat map. Our previous work [15] showed that it is possible to assign given tags to image regions for describing depicted objects. However, the data was collected in a controlled experiment with static photos and without gaze control. Smith and Graham [10] described the advantages of gaze control in video games. They state that the use of gaze control can improve the game play experience. An example is *EyeAsteroids*[1], an eye-controlled arcade game presented by Tobii. The game is entertaining but does not have the goal to exploit the users' activities while playing. Games with a purpose (GWAPs) are computer games that have the goal to obtain information from humans in an entertaining way. The information is usually easy to create for humans, but challenging or impossible to be created by fully automatic approaches. An example of a GWAP is the game *Peekaboom* [12], presented by von Ahn et al. Two users playing together try to label the same image regions for a given tag. Ni et al. [7] introduced a game for explicitly labeling image regions. The users look for specific objects in photos taken from Flickr and mark them by drawing bounding boxes. The development of eye tracking hardware in the recent years supports the usage of gaze control in everyday device like laptops in the near future. Systems that can detect the eyes and can calculate the viewing direction from cameras integrated in common devices like tablet PCs are already on the market (e. g., Natural User Interface Technology, OKAO Vision[2]). Lin et al. [6] presented an eye tracking system using a web cam that is even working in real-time. Thus, the role of eye tracking as input device for controlling software and for collecting information from it's data is increasing.

3 Gaze-Based Measures for Labeling Image Regions

In this work, we apply two gaze-based measures for labeling image regions and one baseline measure, all introduced in previous work [15]. The two gaze-based measures are the segmentation measure (I) and the heat map measure (II). By means of these measures, we assign a given category to an image region for labeling this region. An overview of both measures is depicted in Figure 1. For all photos belonging to the given category, the input for the gaze analysis are (i) the given category and (ii) the gaze paths of all users who correctly classified the photo. The segmentation measure additionally takes (iii) automatically obtained (hierarchical) photo segments as input data. These photo segments are obtained from applying the *bPb*-owt-ucm algorithm [1]. The different hierarchy levels describe different levels of detail and are controlled by the parameter $k = 0, 0.1, \ldots 0.5$.

In the segmentation approach, the fixations on every region of the segmented photo are counted, which corresponds to the fixation measure *fixationCount*. The segment with the highest outcome is assumed to show the object for the given category. In order to take the inaccuracies in the eye tracking data into

[1] http://www.tobii.com/en/gaze-interaction/global/demo-room/tobii-eyeasteroids/
[2] http://www.omron.com

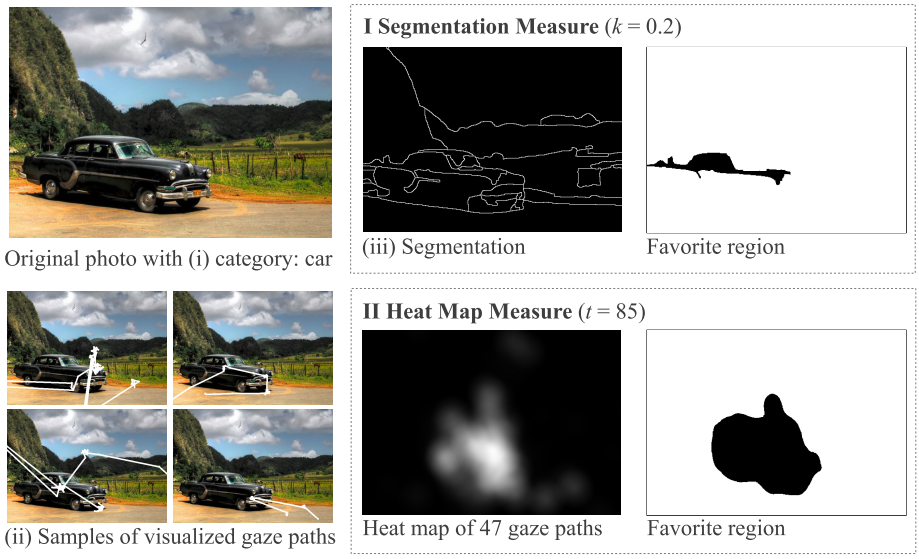


Fig. 1. Gaze-based region labeling with two measures I and II. Input data is (i) the given category, (ii) the users' gaze paths, and (iii) the segmented image (only for I).

account, we apply region extension and weighting introduced earlier [15]. The region extension considers fixations in the surrounding of up to 13 pixels of an segment as being on the segment. Due to the weighting results for segments that are smaller than 5% of the photo are multiplied by a factor up to 4. Different segmentation levels $k = 0, 0.1, \ldots 0.5$ are considered in our analysis. The heat map approach identifies intensively viewed photo regions by summing up the fixations of all gaze paths at pixel level. A value of 100 is applied to the center of each fixation. In a radius of 50 pixels, linear decreasing values are applied to the surrounding pixels. From the created heat map, the object region is calculated by applying a threshold to the data, identifying the mostly viewed pixels. The parameter t indicates the percentage of viewing intensity (e.g. $t = 5$ indicates the 5% of all pixels with the highest values). After the thresholding, the biggest area of connected pixels is assumed to depict the object. The concrete parameter values for both approaches are determined based on the findings in our previous work [15]. The center baseline approach from earlier work [15] is also applied to the data. The element in the center of the segmented photo is considered as a depiction of the object.

By means of ground truth data for the image regions and labels (cf. Section 5), we are able to evaluate the computed object regions. For every pixel, we compare the ground truth with the label obtained from our measures by calculating precision, recall, and F-measure, with F-measure $= 2 \cdot \frac{precision \cdot recall}{precision + recall}$. An example photo with two object regions and their evaluation can be found in Figure 2.

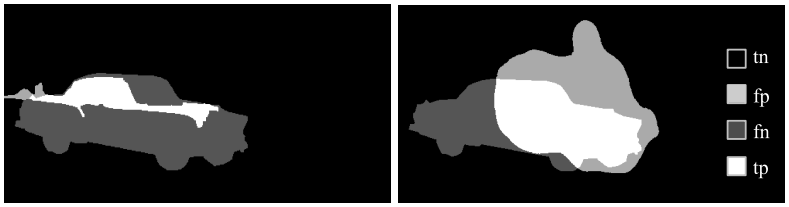
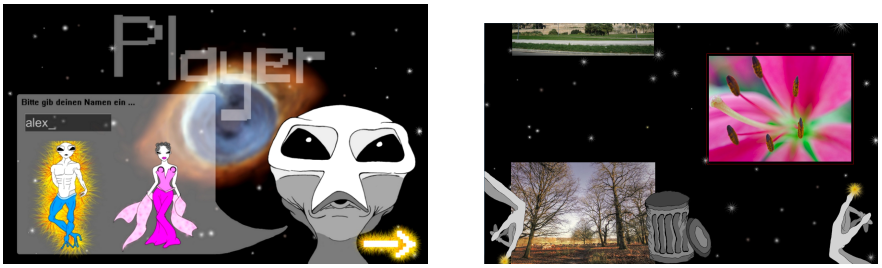


Fig. 2. Comparing labeled image regions and ground truth regions at pixel level.

4 The *EyeGrab* Game

The task in *EyeGrab* to "clean up an aliens' universe" by categorizing and rating photos. Before starting the game, users have to calibrate the eye tracking device by fixating several points on the screen. Subsequently, a small introduction to the game's rules is given to the gamer. In addition, he/she has to choose a user name and to indicate his gender as depicted in Figure 3(a). Besides entering the gamer's nickname, the game is solely controlled by eye movements. Gaze-based interactions are triggered after a dwell time of 450 m. The ocular dwell time of fixations lies between 200 and 400 ms [2]. Hence, the selection dwell time lies above this value to avoid random selections. For example, the selection of the gender is done by focusing on a male or female character as shown in Figure 3(a). The gender information is used only for adapting the gaming environment, e.g., by changing some colors.

A game consists of several rounds. In each round, a set of photos has to be classified concerning a given category like "car", "person", and "sky". First, the category is presented to the user for 6 s. Subsequently, the photos fall down the screen as depicted in Figure 3(b) and are classified by the gamers. Each round has a different speed level at which the photos move. Several photos can be shown on the screen at the same time. The player selects an image by fixating it for longer than the dwell time of 450 ms. As soon as a photo is selected, it is highlighted by a thin frame, and the user can classify it into one of three categories. The classification takes place by fixating symbols on the screen as shown in Figure 3 (b). The categories are "not relevant" (symbolized by a trash can), "relevant & like" (symbolized by a hand pointing upward), and "relevant & dislike" (symbolized by a hand pointing downward). Playing *EyeGrab*, the gamer scores for each correctly categorized image, receives negative points for each wrong one, and no points for images that fell off the screen without classification. No scores are obtained for the ratings of "like" and "dislike". An acoustic feedback is given for each classification. An applause is played for correct classifications, while a booing sound signals incorrect classifications and missed photos. A high score list is presented to the user at the end of the game.



(a) Entering of personal information (b) Playing screen

Fig. 3. Screen shots from *EyeGrab*.

5 Experiment Description

EyeGrab has been evaluated with 54 subjects (with 19 female). The subjects' ages were between 17 and 56 years (avg = 30 years, SD = 7.7). The majority of the participants were students or research fellows in computer science (70%), but students from other fields of study or members of other professional groups like restorers or psychotherapists participated in the experiment as well. Most subjects enjoyed playing the game. In a questionnaire subsequent to the experiment, 49 of the 54 subjects rated the statement "The game is fun." with a 4 or a 5 on a standard 5-point Lickert scale (avg = 4.22, SD = 0.72). The level of difficulty playing *EyeGrab* seems to have been adequate, as most of the participants did not agree with the statement "The game overexerts me." (M = 2.54, SD = 1). Most of the participants did not feel uncomfortable using the eye tracking device as shown by the low average agreement of 2.24 (SD = 1.15) to the statement "The eye tracker has a negative impact on my well-being."

Procedure. Every participant played four rounds of *EyeGrab*. The first round was a short test round consisting of only 12 photos. This test round with the category "tree" served as an introduction to the game. The data collected during this round was not used in the later analysis. The other three rounds with the categories "car", "person", and "sky" consisted of 24 photos each. The photos of each round were displayed in a randomized order. Different falling speeds were applied to each round. In the slowest pace (speed 1) the photos were falling with 3.6 pixels/ms, and they were visible on the screen for 5200 ms. In the medium pace (speed 2), the photos were visible for 4500 ms (pace = 4.3 pixels/ms). In the most challenging speed (speed 3) the photos were falling down within only 3800 ms (5 pixels/ms). A complete round took between 64,4 s (speed 1) and 50 s (speed 3). A Latin Square design was used in order to randomize the order of the three categories with the three different speed levels. The participants were asked to express their agreement to several statements on a 5-point Likert scale between 1 (strongly disagree) and 5 (strongly agree) in a questionnaire at the end of the experiment. The experiment was performed on a screen with a

resolution of 1680 × 1050 pixels. The subjects' eye movements were recorded with a Tobii X60 eye tracker at a data rate of 60 Hz.

Data Set: Categories and Photos. The categories used in *EyeGrab* were taken from the top six of the list with the mostly used tags in LabelMe [8]. The LabelMe data set consists of photos, uploaded by the community, and has manually drawn region labels. The first two categories of this list ("window" and "building") are not taken into account because often not all instances of these objects are labeled on the photos. This could cause problems during the evaluation of our approach, as we need ground truth data with a complete labeling of all occurring objects belonging to the given category. Thus, we have taken the next top categories, which are the above-mentioned categories of "car", "person", and "sky".

In total, 84 photos (24 for each round and 12 for the test round) were selected from the image hosting page Flickr[3] and from LabelMe [8]. To create a challenge for the gamers, only 50% of the selected photos actually belonged to the given category. Thus, half of the photos were randomly chosen from the photos tagged with the given category, the other half from all other photos. An additional criterion for the selected photos was a minimum size of 450 pixels for one of the photo dimensions. All photos were scaled such that the longer edge has a length of 450 pixels. The 46 photos from Flickr belonged to the ones labeled as the most "interesting". For all photos in our experiment, we need ground truth information regarding the region labels. For the LabelMe images, manually drawn polygons describing the shapes of the depicted objects are part of the data set. Some photos had to be replaced after a manual check because not all occurrences of an object were labeled or an object described by the given category was depicted, although the photos were not labeled with it. For the Flickr images, the ground truth region labels were manually created by a volunteer not involved in the research.

6 Photo Classification Results

Excluding the test round, 72 photos in the three rounds were viewed by each subject. This makes a total of 3,888 photo views. In 260 cases (7%), the photo passed without classification, resulting in a total of 3,628 classified photos. 3,279 images (90%) were correctly classified. Overall, we had 1,624 correct classifications for photos belonging to the given category (true-positive), 1,655 correct classifications for photos not belonging to the given category (true-negative). Meanwhile, 241 classifications were false-negative (photo belonged to the category but was classified as not), and 108 classifications were false-positive, which leads to a precision of 94% and a recall of 87% over all users. The number of incorrect assignments per image lies between 2 and 40 with an average of 4. The three photos with the lowest error rate and the three photos with the highest error rate are depicted in Figure 4.

[3] http://www.flickr.com/

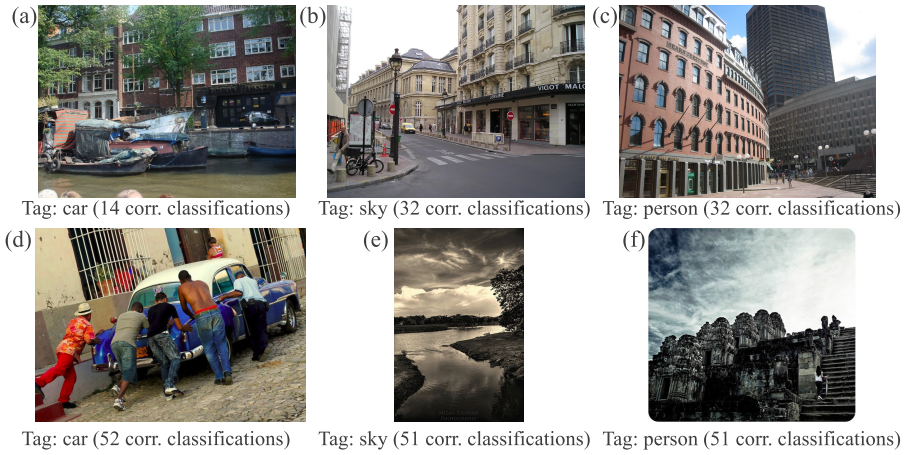


Fig. 4. Upper row: the three photos with the lowest number of correct classifications. Lower row: the photos with the highest number of correct classifications. All photos show an object described by the given category.

When comparing the error rate for different speed levels, we see that the number of unassigned or incorrectly assigned photos is increasing with the falling speed of the photos. The number of not-assigned photos is increasing from 7% to 12%. The number of incorrectly assigned photos is increasing from 4% to 11%. The number of unassigned photos is increasing more strongly than the incorrectly assigned photos. Thus, the subjects are still capable of deciding if an image belongs to a category or not, even with a higher speed level. However, they run out of time to focus each image for classification.

We compare the classification results of *EyeGrab* with results from the photo classifications in our previous experiment [15]. In the previous experiment, a specific tag was first presented to the subjects. Subsequently, a photo was presented to the user who had to decide whether an object described by the given tag is depicted. The decision was made by pressing a key on the keyboard. Of all classification, 5.4% were incorrect. In this work, 10% of all classifications are incorrect over all speeds. The slowest speed level with an error rate of 7% is close to the results observed in previous work [15].

The subjects were asked in the questionnaire how much effort they put into the subjective classification of the photos into "like" and "dislike". They answered this question with a mean value of 3.43 (SD = 1.35), which shows that their effort was not very high. Of the photos, 62% were rated as "like", the rest as "dislike". Of the Flickr images, 70%, were liked in comparison with 56% of the LabelMe images. As the Flickr photos were selected from the most interesting, we can assume that they are more attractive to most viewers than the LabelMe photos. This assumption is only reflected slightly in the rating results.

In summary, the user gave a rating, but it does not seem to be of high quality. Thus, the rating information is not further considered in the remainder of the paper.

7 Photo Labeling Results

We evaluate our approach by analyzing the region labeling for all photos using the aggregated data of all users who correctly classified a photo. In Figure 5, the results for the region labeling using the different eye tracking-based measures are depicted by comparing precision and recall, as well as precision and F-measure. The best precision with 61% is obtained for the segmentation measure with parameter $k = 0$, which corresponds to very small segments. The highest precision for the heat map measure is obtained for $t = 1$ with 59%; for the baseline approach is only 19% ($k = 0$). The best recall results are 96% for the heat map measure with $t = 100$, 70% for segmentation measure with $k = 0.5$, and 53% for the baseline with also $k = 0.5$. We also look into the F-measure results to consider both, precision and recall. The overall best F-measure is obtained by the segmentation approach with 32% ($k = 85$), followed by the heat map approach with 31% ($k = 0.5$). The baseline approach clearly performs weaker, with a maximum result of 21% ($k = 0.5$). We applied a Friedman test to compare the results for the best performing parameters. We found that the differences are significant ($\alpha < .05$) for precision ($\chi^2(2) = 15.436$, $p = .000$) and F-measure ($\chi^2(2) = 18.048$, $p = .000$). A post-hoc analysis with pairwise Wilcoxon tests with a Bonferroni correction ($\alpha < .017$) showed two significant results for precision between heat map and baseline ($Z = -3.527, p = .000$) and segmentation and baseline ($Z = -3.704, p = .000$). No significance was measured in the post-hoc test for F-measures.

The results vary for the three categories "car", "person", and "sky". For example, the precision values for $k = 0$ are $p_{car} = 0.79$, $p_{person} = 0.28$, and $p_{sky} = 0.76$. This range of results seems to be caused by the sizes of the objects. The average size of the ground truth objects of the different categories are (compared with the whole image size) as follows: $size_{car} = 11.5\%$ (SD = 8.3%) , $size_{person} = 11.7\%$ (SD = 19.9%), and $size_{sky} = 42.8\%$ (SD = 23.1%). Although the $size_{car}$ and $size_{person}$ are similar, the high standard derivation for "person" shows that the object sizes vary strongly. Very small objects are known to complicate the region labeling [14].

In addition, we analyzed the region labeling results for the different falling speeds. A faster falling speed increases the pressure on the user to perform the classification. A summary of the results for the different speed levels can be found in Figure 6. It shows that the falling speed does not have a high impact on the precision and F-measure. For both eye tracking measures, the medium speed level delivers the best results. However, only minor differences can be noticed. Please note that the results for all speeds are not the average of all speed levels as the region labeling for the different speed levels is done with only one-third of the data. This is caused by the fact that every user played the game in three

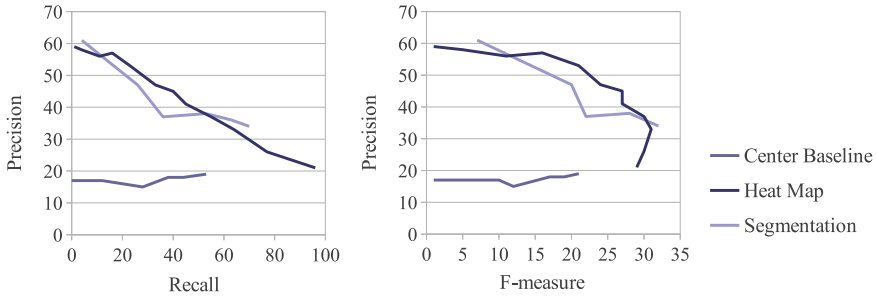


Fig. 5. Precision, recall, and F-measure results for the three labeling approaches. The curves are limited by the investigated parameters (e.g., the Center Baseline by the number of segmentation levels).

different speed levels (cf. Section 5). We conclude that the influence of the speed on the region labeling results is, at the least, not strong.

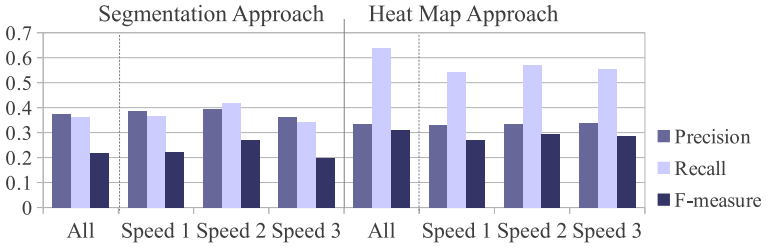


Fig. 6. Region labeling results for different falling speeds.

8 Comparison with Previous Results

We have compared the results in terms of precision and F-measure from our *EyeGrab* experiment with the results obtained from our previous work [15]. The best performing parameters were determined in the previous work by means of a training set and applied to the test set of [15] and to the *EyeGrab* data. The parameters are $k = 0.1$ for the segmentation measure, $t = 95$ for the heat map measures, and $k = 0.4$ for the baseline. The results are depicted in Figure 7.

The segmentation measure performs best, while the baseline approach delivers clearly weaker results than both eye tracking methods. The F-measure results are more diverse. The differences between the two gaze-based measures and the baseline are less distinct for the *EyeGrab* data than for the data from

the previous experiment [15] (i. e., the results between the measures and baseline in our earlier experiment differ more). Using the parameters from earlier work [15] for the *EyeGrab* analysis delivers only slightly better results for the segmentation approach than the baseline, whereas the heat map approach performs clearly better. We compare the center baseline results for photos of the first experiment [15] and *EyeGrab* data in a Mann-Whitney U test and do not obtain a significant difference, neither for precision ($U = 467, p = .291$) nor for F-measure ($U = 446, p = .302$). Thus, we conclude that the photo sets are comparable concerning the center baseline results and infer that the region labeling results can be compared. No statistically significant differences can be found comparing the results from *EyeGrab* and our previous work [15] with regard to the segmentation measure and the heat map measure, neither for precision (segmentation: $U = 528, p = .909$; heat map: $U = 480, p = .467$), nor for F-measure (segmentation: $U = 436, p = .19$; heat map: $U = 468, p = .376$). Thus, we conclude that we can obtain similar results in region labeling in *EyeGrab* and the previous, simplified experiment [15].

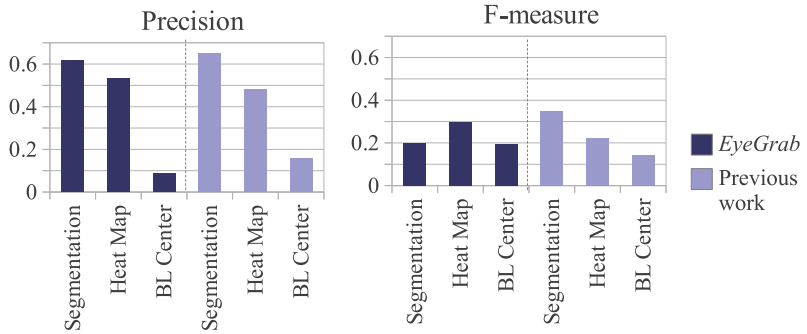


Fig. 7. Region labeling results for *EyeGrab* and previous work [15].

9 Conclusion

We have shown that the labeling of image regions is possible from data collected from subjects playing the immersive game-with-a-purpose *EyeGrab*. For one of two gaze-based measures, the results are comparable to those from a previous experiment [15]. This is quite interesting as the conditions for obtaining the gaze data are more difficult due to factors like time pressure and distraction by the gaming environment in *EyeGrab*. The region labeling results are only slightly influenced by different speed levels, which are forcing the subjects to make decisions on the photo classifications faster. As a broader spread of eye tracking hardware is assumed for the near future, it will become possible to use

eye tracking technology in everyday tasks like image search on the web or for playing games. Thus, the results of our research may be applied for labeling image regions based on the gaze data obtained from users viewing the results of image search engines.

Acknowledgments The research leading to this paper was partially supported by the EU project SocialSensor (FP7-287975). We thank our students D. Arndt, L. Buxel, K. Kramer, C. Neuhaus, C. Saal, H. Swerdlow, and M.-S. Usta.

References

1. P. Arbeláez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *IEEE TPAMI*, 33(5):898–916, May 2011.
2. J.H. Goldberg, M.J. Stimson, M. Lewenstein, N. Scott, and A.M. Wichansky. Eye tracking in web search tasks: design implications. In *Symposium on Eye tracking research & applications*, pages 51–58. ACM, 2002.
3. A. Klami. Inferring task-relevant image regions from gaze data. In *Workshop on Machine Learning for Signal Processing*. IEEE, 2010.
4. I. Kompatsiaris, E. Triantafyllou, and M.G. Strintzis. A World Wide Web region-based image search engine. *Conference on Image Analysis and Processing*, 2001.
5. L. Kozma, A. Klami, and S. Kaski. Gazir: Gaze-based zooming interface for image retrieval. In *Multimodal interfaces*, pages 305–312. ACM, 2009.
6. Yu-Tzu Lin, Ruei-Yan Lin, Yu-Chih Lin, and Greg C Lee. Real-time eye-gaze estimation using a low-resolution webcam. *Multimedia Tools a. Applications*, 2013.
7. Yuzhao Ni, Jian Dong, Jiashi Feng, and Shuicheng Yan. Purposive Hidden-Object-Game: Embedding Human Computation in Popular Game. *IEEE Transactions on Multimedia*, 14(5):1496–1507, October 2012.
8. Bryan C Russell, Antonio Torralba, Kevin P Murphy, and William T Freeman. Labelme: a database and web-based tool for image annotation. *International journal of computer vision*, 77(1-3):157–173, 2008.
9. A. Santella, M. Agrawala, D. DeCarlo, D. Salesin, and M. Cohen. Gaze-based interaction for semi-automatic photo cropping. In *CHI*, page 780. ACM, 2006.
10. J.D. Smith and T.C. Graham. Use of eye movements for video game control. In *Advances in Computer Entertainment Technology*. ACM, 2006.
11. P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001.* IEEE, 2001.
12. L. Von Ahn, R. Liu, and M. Blum. Peekaboom: a game for locating objects in images. In *Human Factors in computing systems*, pages 55–64. ACM, 2006.
13. T. Walber, C. Neuhaus, and A. Scherp. Eyegrab: A gaze-based game with a purpose to enrich image context information. In *EuroHCIR*, 2012.
14. T. Walber, A. Scherp, and S. Staab. Identifying objects in images from analyzing the users's gaze movements for provided tags. pages 138–148. Springer, 2012.
15. T. Walber, A. Scherp, and S. Staab. Can you see it? two novel eye-tracking-based measures for assigning tags to image regions. In *MMM*. Springer, 2013.