Tagging-by-Search: Automatic Image Region Labeling Using Gaze Information Obtained from Image Search

Tina Walber
Institute WeST
University of Koblenz
Germany
walber@uni-koblenz.de

Chantal Neuhaus
Institute WeST
University of Koblenz
Germany
cneuhaus@uni-koblenz.de

Ansgar Scherp
Kiel University, Germany
Leibniz Information Center for Economics, Kiel, Germany
mail@ansgarscherp.net

ABSTRACT

Labeled image regions provide very valuable information that can be used in different settings such as image search. The manual creation of region labels is a tedious task. Fully automatic approaches lack understanding the image content sufficiently due to the huge variety of depicted objects. Our approach benefits from the expected spread of eye tracking hardware and uses gaze information obtained from users performing image search tasks to automatically label image regions. This allows to exploit the human capabilities regarding the visual perception of image content while performing daily routine tasks. In an experiment with 23 participants, we show that it is possible to assign search terms to photo regions by means of gaze analysis with an average precision of 0.56 and an average F-measure of 0.38 over 361 photos. The participants performed different search tasks while their gaze was recorded. The results of the experiment show that the gaze-based approach performs significantly better than a baseline approach based on saliency maps.

Author Keywords

Region labeling, image search, implicit user feedback, eye tracking

ACM Classification Keywords

H.5.2 User Interfaces: Input devices and strategies

INTRODUCTION

Billions of users are viewing photos on the web. Google published a number of one billion page views per day for their image search service[1]. Photo search can be performed based on simple visual similarity, e. g., on Google by the "Search by image" function. However, such low-level pixel information is often less important to the users than the content actually depicted in the image. Thus, the search is usually conducted based on techniques from text retrieval by using the photo title or the text surrounding an image, e. g., on web pages. To provide better annotations, manually added tags can be used to describe the content of an image. More detailed annotations can be conducted by tagging image regions, instead of the entire image. This information can be used for similarity search based on regions [12] or for search based on the coherence of individual image regions [15]. Additionally, a more detailed labeling can be used to display only the relevant area of a photo in the thumbnails of a search results list. Another potential use of region-based annotation data is its application as training set in object detection algorithms (e. g., [22]).

Manual labeling of image regions is a tedious task and is thus very uncommon. Automatic labeling of image regions as performed by object detection algorithms are limited to a number of trained concepts. They also need a large amount of manually created training data and they depend on the visual similarity of objects. In addition, the high computational efforts of the automatic annotation algorithms restrict their applicability.

The goal of our work is to benefit from users who are viewing photos in the results list of an image search engine to perform automatically the labeling of images at region level. It is intuitive for humans to automatically identify objects depicted in an image. Humans can easily compensate perspective distortions, occlusions, and they can also identify objects with an unusual appearance. The gaze paths of users searching for images are recorded by an eye tracking device. Subsequently, the gaze paths are analyzed and regions of the photos in the search results that caught most attention are identified. The search terms entered by the user is assigned to the most viewed image regions for describing the photo content. The gaze paths of several users are aggregated when they view the same photos with the same search term. The labeled image regions are evaluated by comparing them to ground truth regions, which are part of the experiment data sets. The recent developments of eye tracking hardware[2] supports our approach and the possibility to use eye tracking information in every-day life is expected for the next years.

Our previous research [26] showed that it is possible to annotate image regions by means of gaze information in a controlled priming experiment. In this work, we investigated the possibilities to automatically obtain labeling information

[1] http://www.bbc.co.uk/news/technology-10693439, (last visited Sept. 17, 2013)

[2] http://www.tobii.com/rexvip (last visited Oct. 7, 2013)

IUI'14, February 24–27, 2014, Haifa, Israel.

ACM 978-1-4503-2184-6/14/02..$15.00.
http://dx.doi.org/10.1145/2557500.2557517

for image regions while conducting ordinary routine tasks, namely image search. We asked 23 subjects to perform 23 different search tasks with in total 361 photos from three different data sets of different origin and varying image quality. By comparing the generated region labels to ground truth data, we can show that our approach reaches a maximum average precision of $P = 0.56$ (improvement of 30 % over the best baseline result). The highest F-measure result is $P = 0.38$ (improvement of 14 % over the best baseline result). Two eye tracking approaches are used for analyzing the gaze data, one based on photo segmentation and the other one on eye tracking heat maps. Both are compared to a baseline approach which uses low-level image information to identify the most salient regions in a photo. Additionally, the results for the three photo data sets are investigated in detail.

The related work is discussed below. Subsequently, we describe the experiment design and our methods for image region labeling. The results of our experiment are presented and discussed, before we conclude the paper.

RELATED WORK

Different research has been done in the area of collecting implicit user feedback for improving retrieval quality. Joachims [10] and Jung et al. [11] used click-through data of search engine users as implicit source of information to determine the importance of search results. Other information such as how long a document was displayed were investigated, e. g., by Agichtein et al. [1]. Zhang et al. [30] identified attention times, click-through rates, and mouse movements as implicit feedback measures.

Other work, e. g., by Campbell and Flynn [4] and Viola and Jones [23], has focused on using computer vision techniques for the labeling of image regions. These works require large training data sets as well as extensive computational resources. In addition, the identification of objects is limited to the set of concepts trained on the data and to the visual similarity of the learned concepts. Humans are able to recognize objects based on — but not limited to — their visual appearance. Grabner et al. [7] constitute that objects are identified by human observers based on their function and not only on their visual appearance. This shows the limitations of object detection by visual-similarity-approaches compared to the human capabilities. A very different approach is to offer games for entertaining the users while objects are labeled. In Peekabook [24], users play together for identifying objects or parts of objects on given photos. From the collected data, words are assigned to image regions. We have presented the game EyeGrab [27], with the same goal of image region labeling, but performed in a gaze-controlled game for single users. However, these games follow an approach different from the one presented in this paper, where information is obtained from users performing the routine task of image search. No extra task has to be performed.

The use of eye tracking technology as an explicit input device was investigated in numerous studies. By gaze-control, the users explicitly control software by moving their eyes as presented, e. g., in the evaluations of gaze interaction by Sibert and Jacob [21]. Another area of usage for gaze data is to better understand the users' behavior. For example, Chen et al. [5] analyzed gaze information to classify user behavior while performing tasks. The authors were able to identify transitions between tasks in multi-tasking situations. The research presented in this paper differs insofar as we unobtrusively observe the users' gaze paths for gaining information on the viewed objects.

Several approaches used eye tracking to obtain implicit relevance feedback in image search, e. g., [8, 14, 16]. From these works, we know that it is possible to use gaze information to detect images relevant to a given search task. Xu et al. [29] presented a recommender system based on eye tracking information for online documents, images, and videos. Buscher et al. [3] investigated the annotation of texts by means of gaze data and the usage of this information in retrieval tasks. However, their approach is limited to text documents. Putze et al. [17] combined eye tracking information with EEG data to identify events in video streams. The gaze data was used to identify the location of the perceived event (with an accuracy of 86.3 %) while EEG identified the temporal occurrence of an event. The study was performed in a controlled setting with simulated video sequences. Santella et al. [20] presented a method for semi-automatic image cropping using gaze information in combination with image segmentation. Their work showed that users preferred the gaze-based croppings over baseline croppings. Klami et al. [13] introduced an approach to identify image regions relevant in a specific task by using gaze information. Based on several gaze paths, heat maps were created, which identify regions of interest. This work revealed that these regions depended on the task, given to the subject before viewing the image. The work of Ramanathan et al. [18] aimed at localizing salient objects and actions in images by using gaze information. Image regions that were affecting the users were identified and correlated with concepts taken from a model for affection. The affective image regions were identified using segmentation and recursive clustering of the gaze fixations. The identification of image regions showing specific objects was not conducted in their analysis.

In earlier work, we investigated the potential of labeling image regions by means of gaze data [26]. The eye tracking information was collected in a controlled experiment, where the participants made decisions about the presence of a specific object on a photo. We obtained precision values of up to 65 % at pixel level for the region labeling. In this paper, we go a significant step further and investigate if it is possible to automatically obtain image region labels while asking the participants to do nothing more than performing image search tasks. To the best of our knowledge, this is the first time that the feasibility of automatic labeling of image regions by means of using eye tracking information in a real-world scenario like image search is analyzed.

EXPERIMENTAL DESIGN

We conducted an experiment to investigate the potential of photo region labeling during image search. Therefore, participants used a simulated search page for performing different search tasks.

Figure 1. Sample search tasks and images not fulfilling and fulfilling the task.

Subjects

23 volunteers participated in our experiment, 11 of them were female. Their average age was 23.3 (SD: 2.09) with the youngest person being 20 and the oldest 29. Most of the participants were computer science students, but there were also students of other subjects, like mechanical engineering, biology, geology, and educational science.

Photo Sets

Photographs of natural scenes were presented to the users. These photos were taken from three data sets. All sets provided ground truth region labeling data. The VOC2012 data set [6] was made available for the Visual Object Classes Challenge. The segmentation set, which contains ground truth region labels at pixel level, contains 2913 photos and 20 classes of objects like "aeroplane", "sofa", and "dog". MSRC [28] published by Microsoft Research consists of 592 photos and 23 labeled object classes. The objects belong to simple concepts like in the VOC2012 set, e. g., "bird", "sky", and "sheep". The LabelMe [19] set with 182,657 user contributed images and 291,841 labels (download August 2010) provides images of complex indoor and outdoor scenes. The LabelMe community has manually created region labels by drawing polygons into the images and by tagging them.

The photos for the experiment data set were selected by their labels. The labels were taken from the "All time most popular tags" of the online photo sharing page Flickr[3]. Among the most frequently used tags, 23 occur in at least two of the three data sets. These labels were selected for the use in our experiment application. For each label, a random number of photos between 9 and 24 was chosen from the two resp. three data sets. 10 labels occur in all three data sets, whereas 13 labels are present in only two sets. The label-sets were composed in equal parts of the data sets.

[3]http://www.flickr.com/photos/tags/ (last visited Sept. 29, 2013)

In total, our experiment data set consists of 361 photos, with 103 photos taken from MSRC, 112 from VOC2012, and 146 from LabelMe.

Tasks

For each search set, consisting of a label and a set of photos, a search task was defined with the goal to simulate an online image search and to motivate the users to scan the image search result lists. The tasks request the participants to find an object with specific characteristics. For example for the label "bus", the search task was *"Search for a green bus"*. The tasks were created in a way that at least one photo fulfills the task. Often, even more than one photo could be selected. Also, there exist tasks where the answer depends on the subjective impression of the user. For example, a subject might chose an image showing a bird with an orange bill for the task *"Search for a bird with a red bill"*. Some more examples of search tasks can be found in Figure 1. This figure also shows examples of photos fulfilling and not fulfilling the given search task. 10 of the search tasks ask for a specific color as characteristic (e.g., *Search for a green bus*), 4 for animals with a specific coat color or pattern (e.g., *Search for a dog with black spots*), 5 tasks concentrate on other characteristics (e.g., *Search for a building with balcony*), and 4 ask for objects in specific situations (e.g., *Search for a horse with bridle*). In our analysis, we assign the named object to an image region in all photos of the search result list that were fixated, ignoring the specific characteristics. We investigate possible differences in region labeling results for photos fulfilling the search task (the photos with the *green* bus) and photos not fulfilling the task (photos depicting a bus, but not a green one).

Procedure and Experiment Application

Before starting the experiment application, the participants were introduced to the experiment tasks and the eye tracking device. A calibration of the eye tracker was performed by fixating five dots on the computer screen.

Figure 2. Cropped and scaled screen shots of the three experiment steps: A Search task and start search, B Search results, C Photo selection. The arrows show interaction options.

The experiment application was designed to resemble online image search pages. It consists of three pages. Screen shots of the application can be found in Figure 2. On the first page of the experiment application, page A in Figure 2, the search task was presented to the user. The user had to enter a search term as free text into the search input field. By pressing the OK button the simulated search was started. It was not allowed to start the search with an empty text field, but no further checks with regard to its meaning were performed on the given search query. On the second page B, the photos of the experiment data set were displayed in rows of three photos each. The photos were scaled to a maximum width and height of 450 pixels. The page was scrollable as not all photos could be shown on a static page. The user could go back to the search page by pressing the "Back" button. By clicking on the photos, page C opened. On this page, the user could select a photo by pressing the "Select" button for completing the search task. It was possible to go back to the search result page by clicking on the "Back" button.

Eye tracking data was recorded while the user performed the tasks. No time limitations were given for the 23 search tasks. The order of the tasks was randomly alternated for each participant. Also the order of the photos on the search result pages was randomized. At the end of the experiment, each user filled out a questionnaire. It comprised questions about demographic information (age, profession) and some ratings about the experiment application and tasks.

Apparatus

The experiment was performed on a 22-inch monitor. The participants' gaze paths were recorded with a Tobii X60 eye tracker at a data rate of 60 Hz and an accuracy of 0.5 degrees.

ANALYSIS

In this section, the two approaches for analyzing the gaze data as well as the baseline approaches, introduced in our previous work [26], are briefly presented. We extended the approach in a way that allows to assign a given search term to several image regions in one photo.

Assigning Labels to Image Regions by Gaze Analysis

We applied two gaze-based predictors for labeling image regions and one baseline predictor [26]. The two gaze-based predictors were the I Segmentation Gaze and the II Heat Map Gaze approach. By means of these approaches, we assigned a given search term to an image region for labeling it. An overview of the calculation of both measures with one sample image is depicted in Figure 3. For all photos belonging to a search set, the input for the gaze analysis was (i) the given search term and (ii) the gaze paths of all users who fixated the photo. The I Segmentation Gaze measure additionally took (iii) (hierarchical) photo segments as input data. The photo segments for measures I Segmentation Gaze were obtained from applying the gPb-owt-ucm algorithm [2]. The different hierarchy levels describe different levels of detail and are controlled by the parameter $k = 0, 0.1 \ldots 0.7$, with $k = 0$ as highest level of detail. Please refer to the original publication by Arbeláez et al. [2] for details of the gPb-owt-ucm algorithm.

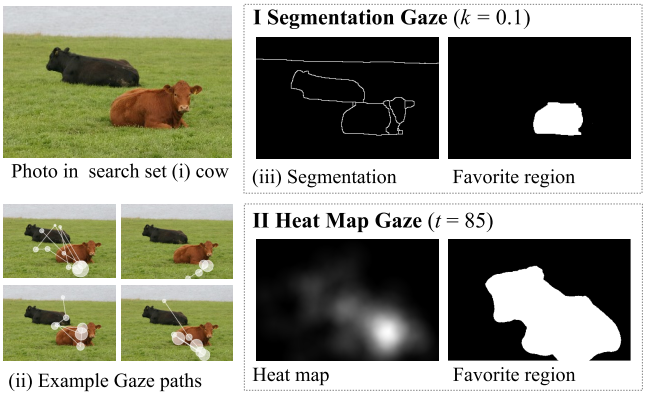

Figure 3. Gaze-based region labeling with predictors I Segmentation Gaze and II Heat Map Gaze. Input data is (i) the given search category, (ii) the users' gaze paths, and (iii) the segmented image (only for I).

The recorded gaze data was analyzed by means of so called eye tracking measures. The segment with the highest measure results was selected and the search term was assigned to it. Different eye tracking measures from literature [26, 16] could be used to perform this selection. The measure (1) fixationCount counted the number of fixations on a segment. (2) fixationDuration calculated the sum of the duration of all fixations on a segment. The measure (3) firstFixationDuration also considered the duration of a fixation, but it only took the very first fixation on a segment into account. Accordingly, (4) lastFixationDuration measured the fixation duration of the very last fixation on a segment. A visit describes the time between the first fixation on a region and the next fixation outside. (5)

visitCount counted the number of visits on a segment and (6) meanVisitDuration calculated the average duration of these visits. The segments with the highest 10 % of the measure values were selected. They were assumed to show an object or several objects described by the search query. The search term was assigned to this region. The measure results for all participants which viewed the same photos are summed up. In order to take the inaccuracies in the eye tracking data into account, we applied the region extension from previous work [26]. The region extension considers fixations in the surrounding of up to 13 pixels of a segment as belonging to the segment.

The II Heat Map Gaze approach identified intensively viewed photo regions by summing up the fixations of all gaze paths at pixel level. A value of 100 was applied to the center of each fixation. In a radius of 50 pixels, linear decreasing values were applied to the surrounding pixels. The value of all fixations were summed up for all pixels of the image for building the so-called heat map. From the created heat map, the assumed object region was calculated by applying a threshold to the data, identifying the mostly viewed pixels. The parameter t indicates the percentage of viewing intensity (e.g. $t = 10$ indicates the 10 % of all pixels with the highest values). The investigated parameters in this work were $t = 1$ and $t = 10 \ldots 100$ in steps of 10.

Baselines

We applied two baseline approaches that were compared with the gaze-based ones. The baseline approaches did not rely on eye tracking data. Furthermore, the baselines did not need training data nor a training period, just like the gaze-based approaches.

Saliency Baselines

The saliency baseline is based on the assumption that the important objects of a photo are the most salient points on an image. These points were calculated by the toolbox offered by Itti et al. [9]. The toolbox calculates salient points by means of multiscale image features. The order of the points depends on decreasing saliency values. The favorite region was selected by using the salient points and their ordering as input data. This saliency paths were interpreted as simulated gaze paths. Subsequently, the same methods as for the gaze analysis approach, described in the previous section, were used to analyze them. Thus, the investigated baseline approaches are called the III Segmentation Saliency approach and the IV Heat Map Saliency approach.

Random Baseline

Finally, for the baseline V Random, the photo was first segmented by the algorithm published by Arbeláez et al. [2]. Subsequently, one of the segments was selected randomly and the search term was assigned to this segment. This very naive baseline serves as measure for how difficult the task of selecting one favorite region was.

Calculating Precision, Recall, and F-measure

By means of ground truth data for all images and assigned labels (cf. Section Photo Sets described above), we were able to evaluate the computed object regions. For every pixel, we compared the ground truth with the labels obtained from our approaches 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 4.

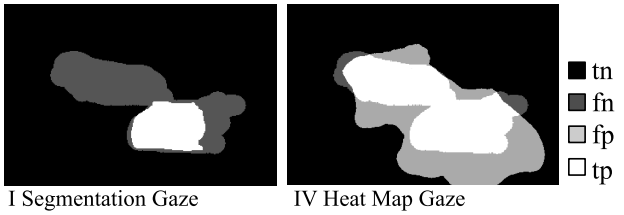

Figure 4. Comparing labeled image regions and ground truth regions at pixel level.

RESULTS

In this section, the labeling results are presented and we compare the gaze-based methods to the saliency methods. Also the results for the three different data sets are compared. Additionally, we investigate the differences for photos fulfilling or not fulfilling the search task.

User Feedback and Behavior

The participants did not feel uncomfortable while their eye movements were recorded by the eye tracking device. Most participants gave an answer of 5 (M: 4.92, SD: 0.28) on a Likert scale from 1 ("I felt uncomfortable while my eye movements were recorded") to 5 ("I did not feel uncomfortable while my eye movements were recorded"). The users' comfort was asked in the questionnaire to check if there was a strong influence of the eye tracker recording on the participants' well-being and thus their gaze. As the users did not feel uncomfortable such an influence is not very likely.

The users did not have problems controlling the application as shown by an average answer of 1.04 (SD: 0.2) on a scale from 1 ("The application was easy to control") to 5 ("It was hard to control the application"). Also the tasks were not to difficult to perform, as the level of difficulty was in average rated with 1.33 (SD: 0.62) on a scale between 1 ("The search for images was easy") to 5 ("The search for images was difficult").

The average time the users spent on a search task was 14.6 s. The longest average search time was obtained for the search task *"Search for a road with median strip."* with 23.3 s. The shortest average time was 8.8 s for task *"Search for a chair with a red seating surface."* The searching behavior of the subjects showed that in 99.98 % of all cases the photo selection page was opened only once, namely for the final selection. Nine times subjects went from photo selection page C back to search page B before they chose an image according to the search query. With regard to the final selections, a percentage of 98.03 % correctly selected images reveals the high quality of the results.

On average, each user fixated 11.63 photos per search query. The average number of fixations over all users per photo is 2.88 (SD: 1.63). The average number of fixations on an image is highest for the search set "bottle" with 6.42 (SD: 1.91). In

contrast, for the search set "car", the number of fixations on an image on average is the lowest with only 1.94 fixations (SD: 0.91).

Comparison of Eye Tracking Measures

First, the six eye tracking measures are compared for the I Segmentation Gaze predictor. As parameter for this approach, the smallest segmentation size $k = 0$ was chosen. Figure 5 depicts the detailed results. For each eye tracking measure the average precision results for each search term are depicted. The box plot diagram shows the first and third quartiles as boxes, the median is displayed inside the boxes as horizontal line, the mean as small circle, and the vertical lines show the range of all values. The measure (5) visitCount clearly performs worse than the other measures. (6) meanVisitDuration and (3) firstFixationDuration have good mean results, but a big spread in the results over the different search terms. The measures (1) fixationCount and (2) fixationDuration perform best. As the measure (1) fixationCount has the best average result ($M = 0.48$, $SD = 0.13$) over all search terms compared to (2) fixationDuration ($M = 0.47$, $SD = 0.13$), (1) fixationCount is used in the following analysis.

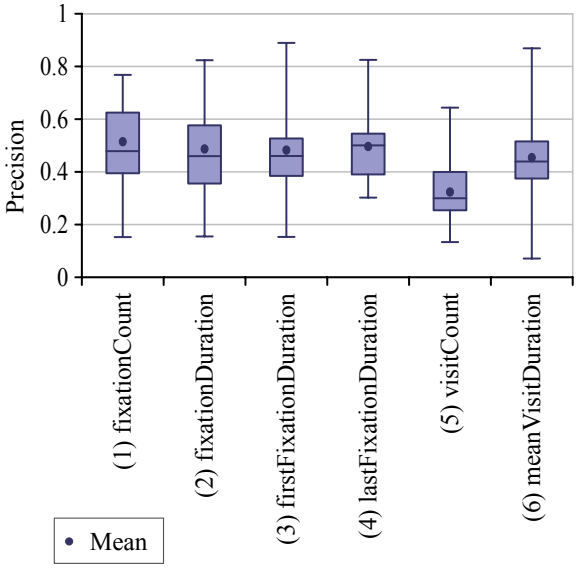

Figure 5. Precision results for I Segmentation Gaze with $k = 0$ for six different eye tracking measures.

Region Labeling Results

The results for the five region labeling approaches are compared in Figure 6. The precision and F-measure results are depicted for different parameters $k = 0 \ldots 0.7$ and $t = 1 \ldots 100$ (see Section Analysis above). Both gaze-based approaches I Segmentation Gaze and II Heat Map Gaze perform better than the baseline approaches. The saliency approach already shows better results than the random baseline. The II Heat Map Gaze approach clearly delivers the best precision and recall results over all parameters. The best F-measure was obtained for II Heat Map Gaze with 0.38 (marked as black circle in Figure 6) with $t = 90$. The overall best precision was obtained for the same measure and parameter with 0.56.

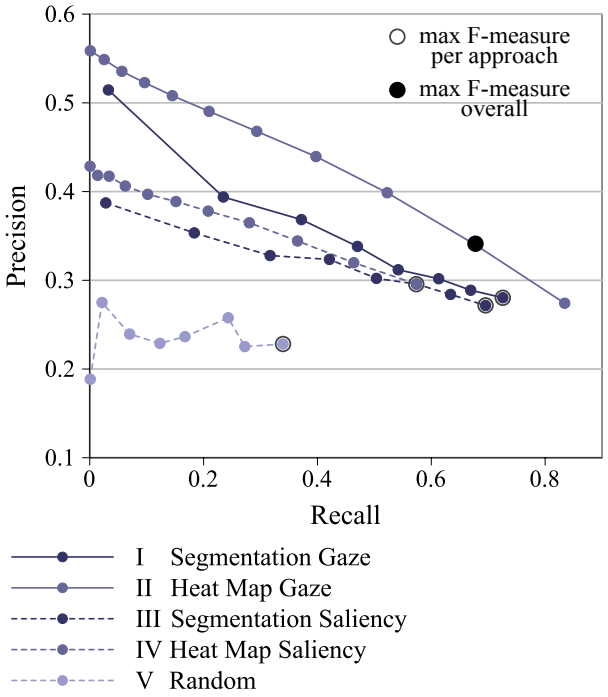

Figure 6. Precision and recall results for the two gaze-based measures I and II, the two saliency-based measures III and IV, and the V Baseline measure.

The best performing baseline approach with a F-measure result of 0.33 is IV Heat Map Saliency with $t = 100$.

A Wilcoxon signed-rank test showed a statistically significant difference with $\alpha < 0.05$ when comparing the average precision and F-measure results per search category for the best performing predictor II Heat Map Gaze with $t = 90$ and the best performing baseline predictor IV Heat Map Saliency with $t = 100$ (precision: $N = 23, Z = -3.194, p < .001$, F-measure: $N = 23, Z = -3.346, p < .001$).

Example Photos

The F-measure results for all photos are depicted in Figure 7, sorted by the F-measure values. We did not find any correlations between the number of fixations on a photo and the precision nor F-measure results. Only 9 of the 361 photos had a precision result of 0, i.e., not a single pixel of the labeled area covered a correct object.

The three photos with the best F-measure results and two photos with the lowest F-measure results are depicted in Figure 8. Besides the original photo, also the region the search tag was assigned to, as well as the ground truth regions for the given object, are depicted. Regarding the average number of fixations for the best labeling predictions one can observe that 1.47 fixations on that image were obtained by 15 subjects (the other ones did not fixate the image). In contrast, the second ranked image was fixated 9.90 times on average by 20 participants. The image placed on rank three was fixated 2.15 times by 13 subjects.

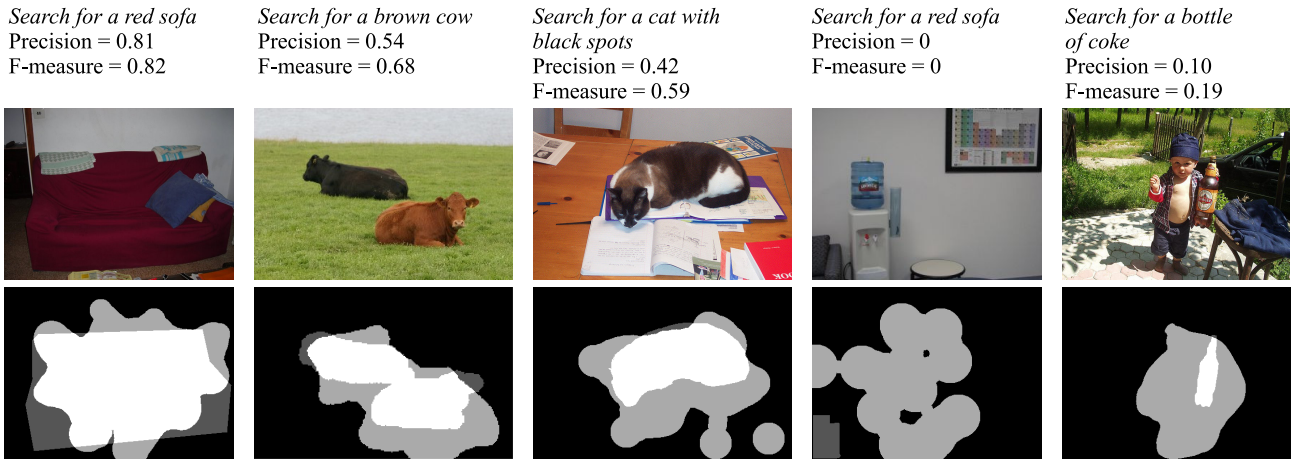

Figure 8. Example image with results for II Heat Map Gaze with $t = 90$ with evaluation of the labeled image regions.

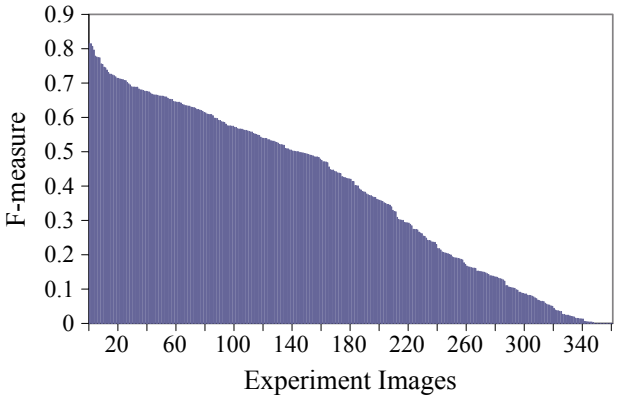

Figure 7. F-measure results for all images of the experiment data set calculated with II Heat Map Gaze with $t = 90$. The images were sorted according to their F-measure value in descending order.

Example Sets

In Figure 9, the precision and F-measure results for approach II Heat Map Gaze with $t = 90$ are split up for the different search tasks. In the diagrams, the results for all photos in each task are displayed (boxes show the area between the first and the third quartile, median as horizontal line, and the range of all photo results as vertical line). One can see that the range in the results is high. This means that the labeling results strongly depend on the given photos. The highest average precision value over all photos of one search task is obtained for "tree" with $P = 0.61$, the worst for "bottle" with $P = 0.09$. The best average F-measure value is obtained for "building" with $P = 0.63$, the worst for "sky" with $P = 0.16$.

Comparison of the Data Sets

Our experiment data was composed of photos from three different data sets, as described in Section Photo Sets. For the best performing approach II Heat Map Gaze, the best performing baseline approach IV Heat Map Saliency, and the V Random Baseline, we split up the precision and recall results for the three data sets VOC2012, MSRC, and LabelMe in Figure 10. Already the random baseline shows differences in the level of difficulty for the segmentation approach. In total, the results are much higher for the MSRC data set containing scenes of low complexity, compared to the most challenging data set LabelMe which includes images showing scenes of high complexity (i. e., many different objects). However it can be observed that the gaze-based approach improves the results for all data sets over the saliency baseline. The results of II Heat Map Gaze always lie above IV Heat Map Saliency.

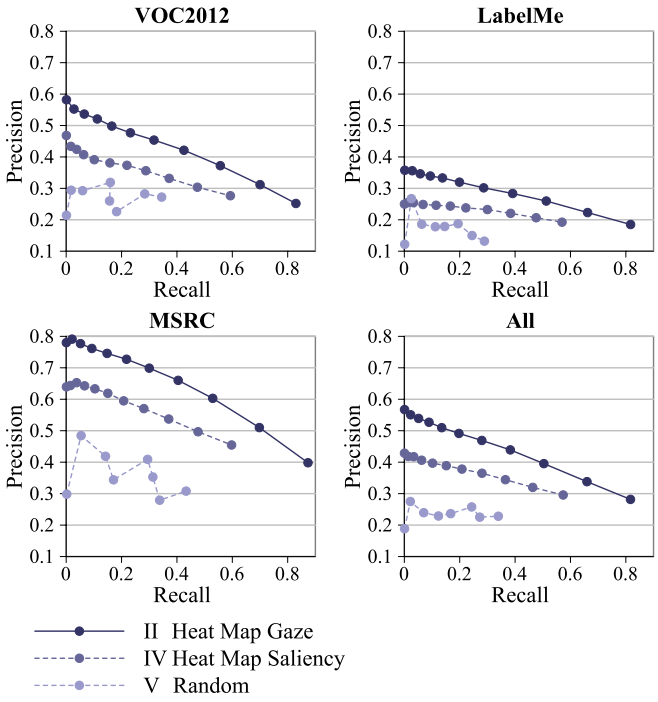

Figure 10. Compare results for the different data sets.

Comparison of True and False Images

In the experiment application, a search task was given to the participants asking for an object with specific characteristics, e. g., *"Search for a green bus"*. In the search result list, all photos depicted an object we asked for (e. g., "bus"). But

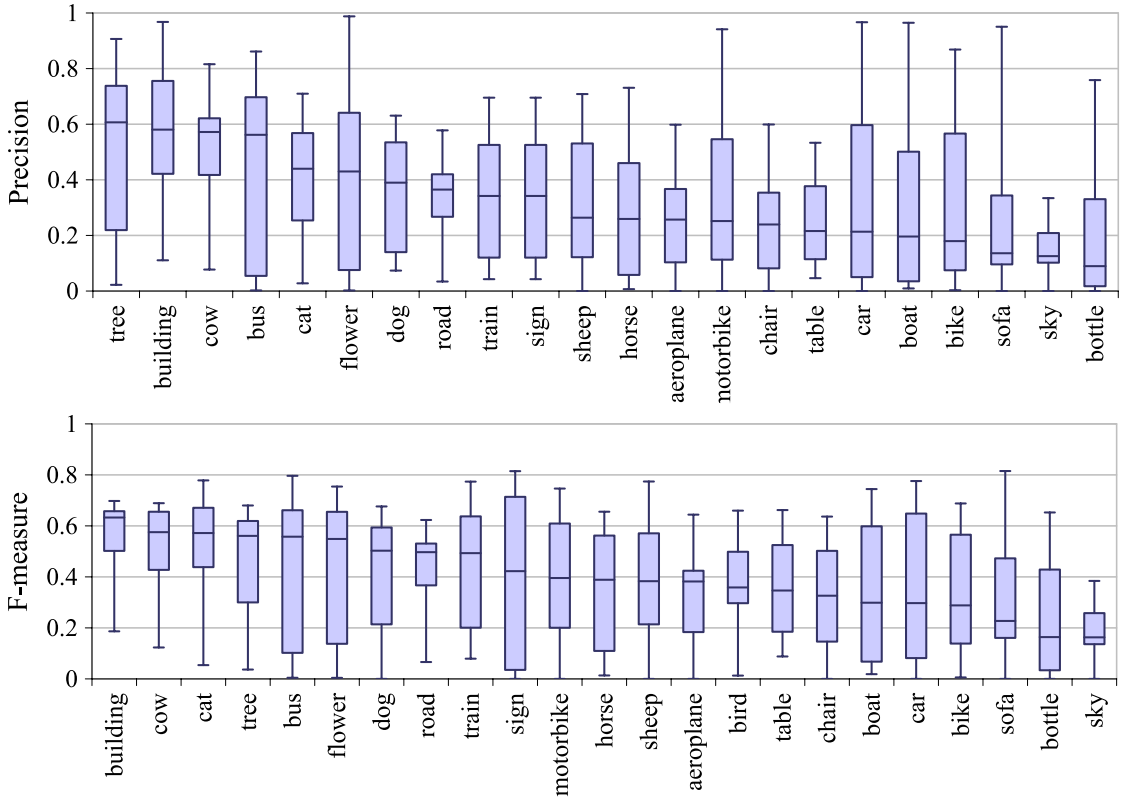

Figure 9. Detailed precision and F-measure region labeling results for each search task for approach II Heat Map Gaze with $t = 90$. The terms are sorted in descending order by their median precision value (above) and F-measure value (below), respectively.

only a few photos showed the object with the specific characteristics (e. g., "green bus") . In total, 97 of the 361 photos fulfilled the search task, 264 did not. For the approaches II Heat Map Gaze and IV Heat Map Saliency, we compared the labeling results for photos fulfilling the search task and not fulfilling the task. Precision and recall results are depicted in Figure 11. As can be seen in the figure, the curves lie close to each other. The results for the photos fulfilling the tasks are slightly higher. We compared the results for photos fulfilling the tasks and not fulfilling the tasks in a Wilcoxon signed-rank test. The results were computed using the values obtained from the approach II Heat Map Gaze with $t = 90$. The differences in the results are not significant with $\alpha < 0.05$ for precision ($N = 23, Z = -.487, p < .626$) and F-measure ($N = 23, Z = -3.346, p < .001$). This suggests that the approach also works for objects that do not exactly fulfil the task, i. e., where the photos show the object asked for but where the object does not match the additional characteristics like the color. With other words, the results imply that the labeling of objects is agnostic to characteristics of the objects the user is looking for.

DISCUSSION

Our experiment results suggest that the labeling of image regions by means of gaze data is possible. Comparing the best precision and F-measure results ($P = 0.56$, F-measure $= 0.33$) of this work shows slightly lower results compared to the ones obtained in previous work [26] ($P = 0.65$, F-measure $= 0.35$). The results strongly vary for different search terms and photos. There are usually two reasons for difficulties in identifying objects in photos: One reason is caused by the characteristics of human visual perception. Big objects and objects that can easily be identified in the corner of ones eyes. Here, the user does not have to fixate it directly. One of the weak categories, "sky", is very likely to belong to this group of objects. Another challenge are very small objects due to inaccuracy of the eye tracking data and the segmentations of the photos. One of our previous works showed difficulties with small objects [25]. A detailed analysis of the factors influencing the results (like how many details are depicted on a photo) could be subject of a future study. More data and different photos might be needed for such a study.

We selected only "correct" photos for the search sets. Correct means that on each photo at least a correct object is depicted, even though the object does not have the specific characteristics. In a real world application, search engines reach a very high quality for simple search queries. Thus, we assume that the results may be transferred to a real search engine. However, when applying our method to real image search, this question has to be handled and wrong photos in the result set have to be considered as well.

From the two approaches for the gaze-based (I, II) and the saliency-based (III, IV) methods, the heat map approach performs better. An additional advantage of this approach is –

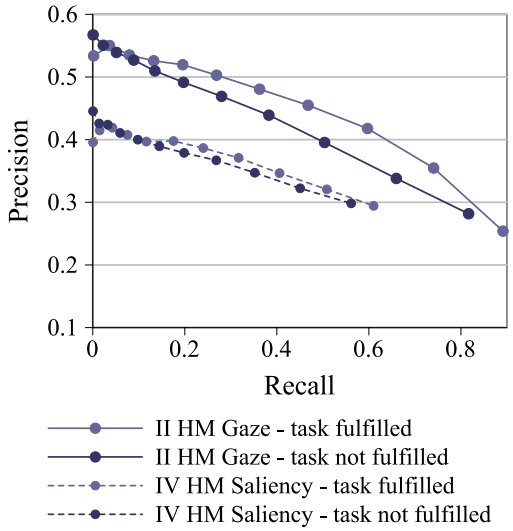

Figure 11. Precision and F-measure results for II Heat Map Gaze with $t = 90$ and IV Heat Map Saliency with $t = 100$ for photos fulfilling the search task versus not fulfilling the search task.

compared to the segmentation-based approach – that no segmentations have to be calculated. The computation of high-quality segmentations could be time-consuming. By varying the parameters of the II Heat Map Measure Gaze approach, the focus can be moved from good F-measures results (a higher parameter t which leads to bigger selected areas) to good precision values (small t values).

A further step could be the combination of I Segmentation Approach and II Heat Map Approach. For example, a segmentation inside the selected heat map areas could deliver interesting results. Until now, the II Heat Map Approach does not use any information that can be obtained from the image content.

The possibility to identify personal preferences from gaze information can also be adapted to other domains. One application could be the recommendation of products based on previous fixations on photos or objects in photos. In social media content, it could be possible to identify persons that are important to the user.

The steps forward in the development of eye tracking hardware are big and state-of-the-art open source solutions are developing rapidly. It could be an interesting next step in our work to test a low-cost device to investigate if there are differences in the accuracy and how this affects our approach.

SUMMARY

Our work shows that it is possible to assign search terms to image regions by means of gaze paths recorded by users searching for images. The usage of gaze data significantly improves the labeling results over a baseline approach using only saliency information. The method works even for photos depicting an object that was asked for, but did not fulfill the specific characteristic mentioned in the search task.

With a performance time of 14.6 s per search query, including the scanning of numerous photos, the labeling of image regions is very fast compared to the manual drawing of polygons. Also, no more effort is needed by the users than viewing search engine results. Another advantage of the suggested method is that the visual appearance of an object is not of importance and even unusual objects could be labeled as long as they are identified by the users. Also the labeling of image regions depicting more abstract concepts like “love” and “speed” could be performed by our approach.

ACKNOWLEDGMENTS

We thank all subjects for participating in our experiment. The research leading to these results has received funding from the European Community’s Seventh Framework Programme (FP7/2007-2013) under grant agreement 287975.

REFERENCES

1. Agichtein, E., Brill, E., and Dumais, S. Improving web search ranking by incorporating user behavior information. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM (2006), 19–26.
2. Arbeláez, P., Maire, M., Fowlkes, C., and Malik, J. Contour detection and hierarchical image segmentation. *IEEE TPAMI 33*, 5 (May 2011), 898–916.
3. Buscher, G., Dengel, A., and van Elst, L. Query expansion using gaze-based feedback on the subdocument level. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, ACM (2008), 387–394.
4. Campbell, R. J., and Flynn, P. J. A survey of free-form object representation and recognition techniques. *CVIU 81*, 2 (2001), 166–210.
5. Chen, S., Epps, J., and Chen, F. Automatic and continuous user task analysis via eye activity. In *Proceedings of the 2013 international conference on Intelligent user interfaces*, ACM (2013), 57–66.
6. Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.
7. Grabner, H., Gall, J., and Van Gool, L. What makes a chair a chair? In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, IEEE (2011), 1529–1536.
8. Hajimirza, S., and Izquierdo, E. Gaze movement inference for implicit image annotation. In *Image Analysis for Multimedia Interactive Services*, IEEE (2010).
9. Itti, L., Koch, C., and Niebur, E. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence 20*, 11 (Nov 1998), 1254–1259.

10. Joachims, T., Granka, L., Pan, B., Hembrooke, H., and Gay, G. Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM (2005), 154–161.
11. Jung, S., Herlocker, J. L., and Webster, J. Click data as implicit relevance feedback in web search. *Information Processing & Management 43*, 3 (2007), 791–807.
12. Kim, D., and Yu, S. A new region filtering and region weighting approach to relevance feedback in content-based image retrieval. *Journal of Systems and Software 81*, 9 (2008), 1525–1538.
13. Klami, A. Inferring task-relevant image regions from gaze data. In *Workshop on Machine Learning for Signal Processing*, IEEE (2010).
14. Klami, A., Saunders, C., De Campos, T., and Kaski, S. Can relevance of images be inferred from eye movements? In *Multimedia information retrieval*, ACM (2008), 134–140.
15. Kompatsiaris, I., Triantafyllou, E., and Strintzis, M. A World Wide Web region-based image search engine. *Conference on Image Analysis and Processing* (2001).
16. Kozma, L., Klami, A., and Kaski, S. GaZIR: gaze-based zooming interface for image retrieval. In *Multimodal interfaces*, ACM (2009).
17. Putze, F., Hild, J., Kärgel, R., Herff, C., Redmann, A., Beyerer, J., and Schultz, T. Locating user attention using eye tracking and eeg for spatio-temporal event selection. In *Proceedings of the 2013 international conference on Intelligent user interfaces*, ACM (2013), 129–136.
18. Ramanathan, S., Katti, H., Huang, R., Chua, T.-S., and Kankanhalli, M. Automated localization of affective objects and actions in images via caption text-cum-eye gaze analysis. In *Multimedia*, ACM (New York, New York, USA, 2009).
19. Russell, B., Torralba, A., Murphy, K., and Freeman, W. Labelme: a database and web-based tool for image annotation. *International journal of computer vision 77*, 1 (2008), 157–173.
20. Santella, A., Agrawala, M., DeCarlo, D., Salesin, D., and Cohen, M. Gaze-based interaction for semi-automatic photo cropping. In *CHI*, ACM (2006), 780.
21. Sibert, L. E., and Jacob, R. J. Evaluation of eye gaze interaction. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM (2000), 281–288.
22. Torralba, A., Murphy, K., and Freeman, W. Sharing visual features for multiclass and multiview object detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 29*, 5 (2007), 854–869.
23. Viola, P., and Jones, M. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1, IEEE (2001), I–511.
24. Von Ahn, L., Liu, R., and Blum, M. Peekaboom: a game for locating objects in images. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, ACM (2006), 55–64.
25. Walber, T., Scherp, A., and Staab, S. Benefiting from users gaze: selection of image regions from eye tracking information for provided tags. *Multimedia Tools and Applications* (2013), 1–28.
26. Walber, T., Scherp, A., and Staab, S. Can you see it? two novel eye-tracking-based measures for assigning tags to image regions. In *Advances in Multimedia Modeling*. Springer, 2013, 36–46.
27. Walber, T., Scherp, A., and Staab, S. Exploitation of gaze data for photo region labeling in an immersive environment. In *Advances in Multimedia Modeling*. Springer, 2014.
28. Winn, J., Criminisi, A., and Minka, T. Object categorization by learned universal visual dictionary. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, vol. 2, IEEE (2005), 1800–1807.
29. Xu, S., Jiang, H., and Lau, F. Personalized online document, image and video recommendation via commodity eye-tracking. In *Proceedings of the 2008 ACM conference on Recommender systems*, ACM (2008), 83–90.
30. Zhang, B., Guan, Y., Sun, H., Liu, Q., and Kong, J. Survey of user behaviors as implicit feedback. In *Computer, Mechatronics, Control and Electronic Engineering (CMCE), 2010 International Conference on*, vol. 6, IEEE (2010), 345–348.