Walber, Tina; Scherp, Ansgar; Staab, Steffen

Article

Benefiting from users' gaze: selection of image regions from eye tracking information for provided tags

Multimedia Tools and Applications

Suggested Citation: Walber, Tina; Scherp, Ansgar; Staab, Steffen (2014) : Benefiting from users' gaze: selection of image regions from eye tracking information for provided tags, Multimedia Tools and Applications, ISSN 1380-7501, Springer Science + Business Media B.V, Dordrecht, Vol. 71, Iss. 1, pp. 363-390,
https://doi.org/10.1007/s11042-013-1390-3

This Version is available at:
http://hdl.handle.net/11108/167

Multimed Tools Appl
DOI 10.1007/s11042-013-1390-3

Benefiting from users' gaze: selection of image regions from eye tracking information for provided tags

Tina Walber · Ansgar Scherp · Steffen Staab

Abstract Providing image annotations is a tedious task. This becomes even more cumbersome when objects shall be annotated in the images. Such region-based annotations can be used in various ways like similarity search or as training set in automatic object detection. We investigate the principle idea of finding objects in images by looking at gaze paths from users, viewing images with an interest in a specific object. We have analyzed 799 gaze paths from 30 subjects viewing image-tag-pairs with the task to decide whether a tag could be found in the image or not. We have compared 13 different fixation measures analyzing the gaze paths. The best performing fixation measure is able to correctly assign a tag to a region for 63 % of the image-tag-pairs and significantly outperforms three baselines. We look into details of the image region characteristics such as the position and size for incorrect and correct assignments. The influence of aggregating multiple gaze paths from several subjects with respect to improving the precision of identifying the correct regions is also investigated. In addition, we look into the possibilities of discriminating different regions in the same image. Here, we are able to correctly identify two regions in the same image from different primings with an accuracy of 38 %.

Keywords Region identification · Region labeling · Gaze analysis · Eye tracking · Tagging

T. Walber (✉) · A. Scherp · S. Staab
University of Koblenz-Landau, Universitätsstr. 1, 56070 Koblenz, Germany
e-mail: walber@uni-koblenz.de

A. Scherp
e-mail: scherp@uni-koblenz.de

S. Staab
e-mail: staab@uni-koblenz.de

Published online: 28 February 2013

1 Introduction

Tagging is an approach to describe the semantics of images with simple keywords, called tags. Image tagging is already commonly used on social media platforms such as Flickr[1] as well as on networking sites such as Facebook.[2] Although widely adopted, tagging describes the semantics of images only in a limited way. One step towards improving the understanding of image semantics is to identify and annotate specific image regions instead of tagging the image as a whole. Such annotated image regions can be used for similarity search based on regions [10], for search based on the coherence of individual image regions [13] or as training set in object detection (e.g. [28]). Although the labeling of image regions is implemented on some platforms and sites [23, 31], manual annotation remains tedious and is rarely used. There exist various computer vision approaches that aim at alleviating the users from the manual annotation task [2, 25, 30]. However, these approaches have limitations such as the necessity of a training period and the similarity of low-level features between concept and depicted object. Other approaches are based on calculated saliency in the images [4, 7, 16, 17, 19, 22]. Saliency maps calculate the regions in an image attracting the most visual attention. These methods has limitations on the image complexity, the placement of the objects, and the discrimination of different objects in one image.

In this paper, we offer an entirely different approach to the problem of tagging specific image regions. We investigate if the annotation of image regions can be performed by exploiting gaze information provided by an eye tracker. Our goal is to use the eye tracking device as implicit source of information in order to automatically assign tags to image regions by analyzing the users' gaze paths. The long-term objective of our work is to add this eye tracking functionality to common image applications, which offers functionality like the tagging of images or the search for images.

The experiment setup in this paper simulates the situation of a user viewing images while being interested in a specific object. This interest could also be given as a search query in a search task or a tag describing an image. These scenarios may include further challenges such as possible distractions from the surrounding web page or smaller image size in the search result lists. Because of such additional challenges, we break our approach to the overall research question of assigning tags to image regions based on eye tracking data down into a series of distinct steps as presented in Fig. 1. The first step I is the analysis of gaze data gained in a controlled experiment, with given tags, and the usage of predefined high-quality segmentation (manually drawn in polygons in LabelMe). In this work, we concentrate on this first step. The next step II includes the usage of automatic image segmentation techniques instead of high-quality image regions. It has been addressed in [33]. The last step III will be based on data, which was gained in an experiment without predefined tags and a less controlled setup. The experiment setup will include image tagging and image search.

In terms of hardware are professional eye tracking devices becoming more affordable and easier to use [1] and the interest in eye tracking is high. There are also promising results of using eye tracking data from unmodified, common web

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

[2] http://www.facebook.com/

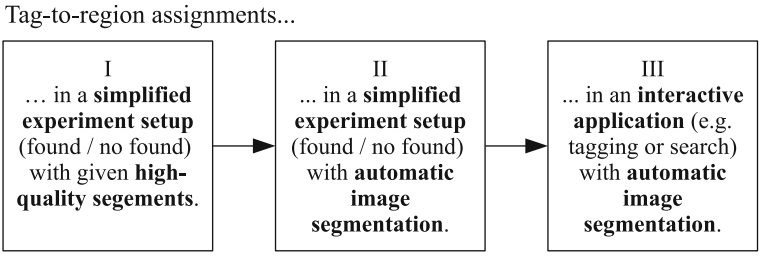


Fig. 1 Embedding of this work in the context of the general goal

cams [26]. Vendors of eye tracking hardware such as Tobii (http://www.tobii.com/) recently released eye trackers which are integrated in the displays of notebooks. Thus, we may assume that eye tracking will become more common in the next years and that it will be possible to make use of gaze information in everyday applications.

Previous research on using eye tracking data for improving the semantics of images such as [1, 3, 8, 11, 20, 34] aimed at finding *some* regions that are relevant to the users' context. We aim at identifying *specific* regions in images by analyzing the users' gaze path information. With this approach we are able to assign a tag to a region and to measure the accuracy of the assignment by comparing it to a ground truth.

In detail, we investigate the following two research questions: (a) To which extend can we predict a region of an image, which is described by the given tag, by applying appropriate fixation measures on the recorded eye tracking data? (b) Can we differentiate different regions (showing different objects) in the images. In order to answer the first research question (a), we conducted a controlled experiment in which we recorded the gaze path information of 30 subjects viewing a sequence of 51 image-tag-pairs. For each tag, the subjects had to decide whether or not an object corresponding to this tag can be found in the image. From the gaze information we calculate tag-to-region assignments, which assigns the given tag to an identified region. Our results show that the best fixation measure reaches a precision of 63 % with taking the potential inaccuracy of the eye tracking data into account and applying a linear weighting function to support smaller image regions. This significantly outperforms three baselines that make use of a random, a saliency-based, and a naive strategy for tag-to-region assignments. Furthermore, we have conducted an in-depth analysis of the obtained results. We have analyzed the size and position of the correctly respectively incorrectly assigned regions to identify typical characteristics that could restrict our approach. The influence of the number of subjects viewing an image on the precision of the tag-to-region assignments is also considered. Finally, we have investigated the impact of the first fixations on an image with respect to identifying the correct image region.

With respect to the second research question (b), we have looked into the possibility of differentiating two objects shown in the same image by analyzing the gaze paths with different primings. Our results show an accuracy of 38 % for two correctly identified objects in the same image. We have analyzed the impact of different user primings such as providing a tag that is different from the region we are interested in or providing a tag that does not correspond to any object in the image.

Our investigation shows that the best fixation measure has a significant impact on identifying the correct image region just by analyzing the users' gaze path.

The main contribution of our work are:

- Assignments of tags to image regions by exploiting the natural human capabilities in object detection
- No limitation on typical visual objects' appearances
- No training set and no training period needed

This article is organized as follows: In the subsequent section we discuss the related work and compare it to our approach. Section 3 presents the 13 fixation measures considered in our experiments and further parameters regarding extending region boundaries and weighting of smaller regions. In Sections 4 to 6 we investigate research question (a): the priming experiment in which we show different image-tag-pairs to participants is presented as well as an in-depth analysis of the results. Finally, in Section 7, we discuss research question (b) and the results of discriminating two objects in the same image, before we conclude the paper.

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

2 Related work

The possibilities to obtain labeled image regions were investigated in different areas of computer science. The work presented in this paper relates to different domains, which need to be discussed. It includes manual labeling of image regions, automatically labeling of images and image regions, and the usage of gaze information for labeling.

2.1 Manual labeling

The simplest approach for annotating image regions is *manual labeling*. For example, the photo sharing platform Flickr allows its users to manually mark image regions by drawing rectangle boxes on it and writing a comment to it. This region labeling is little used in Flickr and mostly for comments. The manual tagging of regions is tedious and the users are pleased with the tagging of images as a whole. Other web platforms like LabelMe [23] allow for a more precise creation of regions by drawing polygons on the images. These regions are annotated with a tag. "Games with a purpose" trigger the human play instinct in order to obtain manually created image regions [31]. For example, in Squigl[3] two randomly selected users team up to mark congruent regions on the same image without seeing the drawing of the partner.

2.2 Automatic labeling

Much work was done on the automatic assignment of *tags to images*. Li et al. [15] makes use of visual similarity for tag recommendation. They present an approach

[3] http://www.gwap.com/squigl-a/

that recommends tags for an unlabeled image by using low-level similarity with already tagged images and by obtaining relevant tags from these images. Tsai et al. [29] performed large-scale annotation of web images by considering images that are visually similar and corresponds to the same semantic concept. They show that their approach facilitates a better prediction performance, compared to competing methods. Tang et al. [27] exploit the extraction of semantic concepts from community-contributed images and tags. They succeed in providing a more robust and discriminative approach compared to other semi-supervised learning approaches. Additionally they propose a label refinement strategy that removes tag noise. However, these approaches do not address the problem of assigning the tags to image regions but to the image as a whole. The problem of labeling image regions is not solved.

Identifying concrete objects and their position in the images is still a challenging task. There are different approaches, based on computer vision or saliency calculation. One approach is the object detection with computer vision techniques. A large amount of training data—consisting of images and labeled image regions—is needed for such a purpose (e.g. [2, 30]). The identification of objects is limited to the learned concepts and to the visual similarity to the learned concepts (e.g.[25]). Different approaches were investigated to make use of *salient image regions* in region labeling. Rowe [22] presents an approach to find the visual focus of an image by applying image processing in terms of segmentation and low-level features. The goal is to link the visual focus with the image caption. This approach is designed for images with a single object only [22]. In addition, it has many limitations concerning the position and characteristics of the shown object. Duygulu et al. [4] perform a mapping between region types and keywords supplied with the images by learning a fixed image vocabulary. Liu et al. [16] propose a method to automatically assign labels at image level to image regions. The method is based on local image patches, gained from image over-segmentation, each of which may partially characterize one image label. They exploit the fact that two images with the same labels are likely to contain some similar patches. The images used in their experiment are simple, with an average of 2 to 3.5 labels per image. Itti et al. [7] present a visual attention model based on multi-scale image features (colors, intensity, and orientation), which outputs salient points in order of decreasing saliency. Their system is offered in a toolkit, which is used in our work as saliency-based baseline.[4] Navalpakkam and Itti [17] present a model, which takes also the influence of tasks into account. Besides the usage of low-level features their system considers an initialization by the user in form of keywords and their relevance. The prediction of visual saliency is then biased by visual information relevant to the given keywords. For their approach, a hand-coded ontology as well as manually created groups of images showing the same objects is needed. Privitera and Stark [19] compares the identification of ROIs by gaze information and by image processing algorithms. They show, that the algorithms cannot predict the sequential ordering of the loci of human fixations.

[4] http://ilab.usc.edu/toolkit/

2.3 Usage of gaze information

Already in 1967 Yarbus [34] has shown the influence of a specific task on the eye movements while viewing an image. Nowadays, *Usability studies* are a standard use case for exploiting gaze information. For detailed analysis, regions of interest (ROI) are marked on the investigated medium, e.g., a web page or a commercial. Based on these ROIs, the users' attention is analyzed in order to optimize the object that is under examination [3]. These ROIs are manually created, have usually simplified shapes like rectangles, and do not aim at correlating image regions with tags for the purpose of region annotation. These approaches show however that the eye tracking data delivers reliable information about the human perception of specific image regions [1].

Jaimes et al. [8] carried out a preliminary analysis of identifying common gaze trajectories in order to classify images into five predefined semantic categories. These semantic categories are handshake, crowd, landscape, main object in uncluttered background, and miscellaneous. The general assumption is that similar viewing patterns occur when different subjects view different images in the same category. To this end, a generic object-definition model is provided that allows the users to specify the relation of objects in the images, such as persons and hands, in an image showing a handshake situation. The results are encouraging and they determine that it may be possible to construct an automatic image category classifier from the approach. However, constructing the object-definition model is tedious. In addition, an object classifier needs to be provided for each object category in the definition model in order to actually be able to classify new images.

In *information retrieval*, several approaches use eye tracking to identify images in a search result as attractive or important and use this information as implicit user feedback to improve the image search, e.g., [6, 12, 14]. Kozma et al. [14] show that a comparison of the implicit gaze feedback with explicit user feedback by clicking on relevant images and a random baseline are promising. Pasupa et al. [18] apply a support vector machine (SVM) algorithm using eye tracking information together with content-based features to rank images. These approaches do not consider image regions.

Santella et al. [24] present a method for semi-automatic image cropping using gaze information in combination with image segmentation. Their goal is to find the most important image region, independent of the objects in the image. Klami et al. [11] present an approach to identify image regions relevant in a specific task using gaze information. Based on several users' gaze paths, heat maps are created, which identify the regions of interest. This work reveals that the regions identified depend on the task given to the subject before viewing the image. However, 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. Ramanathan et al. [21] make use of gaze information to improve the segmentation of digital images. The idea is to analyze the fixation data to identify seeds for the segmentation algorithm. In contrary to our work, the gaze information was collected from users free-viewing images, i. e., without a concrete task or interest in a specific object. The images used in their analysis show only one salient object against the image background. Their gaze-base approach performs the segmentation 10 % better, compared to the segmentation without gaze information. In their work, they do not consider the order of the

fixations. Finally, the work of Ramanathan et al. [20] aims at localizing affective objects and actions in images by using gaze information. Thus, image regions that are affecting the users are identified and correlated with given concepts from an affection model. The affective image regions are identified using segmentation and recursive clustering of the gaze fixations. General identification of image regions showing specific objects is not conducted.

2.4 Summary related work

Manual labeling of image regions is very uncommon in online applications and it is tedious to perform. Automatic assignment of tags to regions is based on the visual similarity, a given training set, and a number of learned concepts. As Grabner et al. [5] constitutes, many objects are identified by human observers based on their function, not on their visual appearance. This shows the limitation of the visual-similarity-approaches. The human is able to identify objects based on but not limited to the visual appearance. The related work on eye tracking shows that it is in principle possible to relate image regions with gaze path information. In contrast to our work, in previous research eye-trackers have been used to identify visual foci in images, find task-related image regions, or localize affective regions in images. However, they have not been used for identifying specific objects in images. This article significantly extends our previous work [32] by taking 10 more subjects into account as well as a new baseline. In addition, we provide crucial in depth analyses of the results that have not been conducted and reported before.

3 Fixation measures, region extension, and weighting

Gaze paths consist of fixations and saccades. Fixations are the phases when the eye is briefly focused on a particular point on the screen. These are the moments of highest visual perception. Saccades are the fast eye movements between the fixations. A fixation measure is a function on the users' gaze path. It is calculated for each image region over all users viewing the same image-tag-pair. In our approach the given tag is assigned to the region with the highest fixation measure value. We call this region the favorite region. In our analysis we investigate which measure provides the highest number of correct aggregations between tag and image region.

In this section we present the considered measures and the additionally investigated parameters for region extension and region weighting.

3.1 Considered fixation measures

We investigate 13 fixation measures concerning their performance to identify the region described by the given tag. The measures including their units are presented below. The way the favorite region is calculated through the particular measure is stated in parentheses behind the measures. It can be, e.g., the minimum of fixation counts on image regions (min count), the maximum distance between two fixations in centimeters (max centimeter), or the maximum fixation duration on the regions in milliseconds (max millisecond). An overview of the measures is presented in Table 1.

Table 1 Eye tracking measures for a region *r*

No	Name	Description	Favorite	Origin
1	firstFixation	Number of times the subject fixates on the image before fixating on region *r* for the first time	min count	Tobii
2	secondFixation	Number of times the subject fixates on the image before fixating on region *r* for the first time without the first fixation on the image	min count	New
3	lastFixation	Number of times the subject fixates on the image after last fixation on region *r*	min count	[11]
4	fixationsBeforeDecision	Number of times the subject fixates on the image after the last fixation on *r* and before the decision	min count	New
5	fixationsAfterDecision	Number of times the subject fixates on the image after the decision and before the fixation on region *r*	min count	New
6	fixationDuration	Sum of the duration of all fixations on *r*	max seconds	Tobii
7	firstFixationDuration	Duration of the first fixation on *r*	max seconds	Tobii
8	lastFixationDuration	Duration of the last fixation on *r*	max seconds	New
9	fixationCount	Number of times the subject fixates on *r*	max count	Tobii
10	maxVisitDuration	Maximum visit length on *r*	max seconds	Tobii
11	meanVisitDuration	Mean visit length on *r*	max seconds	Tobii
12	visitCount	Number of visits within *r*	max count	Tobii
13	saccLength	Length of saccade before fixation on *r*	max centimeter	[14]

The standard measure (1) firstFixation (min count) computes the number of fixations on the image before fixating on a region *r*. The favorite is the region that was fixated first, that means the region with no previous fixations on the image. As a variation of (1) firstFixation, the measure (2) secondFixation (min count) ignores the very first fixation. We also use another modification of the (1) firstFixation measure called (3) lastFixation [11] (min count) to count also the fixations on the image after the last fixation on the examined region. The favorite is the region with no fixations after the last examination. Gaze paths could contain fixations after the making of the decision by pressing the button on the keyboard, due to the inherent reaction time of the experiment application. We have investigated the fixations around the moment of decision with the new measures (4) fixationsBeforeDecision (min count) and (5) fixationsAfterDecision (min count). The measure (6) fixationDuration (max millisecond) describes the sum of the duration of all fixations on a region *r*. The measure (7) firstFixationDuration (max millisecond) considers the order of the fixations and

describes the duration of only the first fixation on a region r. Also the measure (8) lastFixationDuration (max millisecond) was investigated. It provides the duration of the last fixation on the region. The standard measure (9) fixationCount (max count) counts the fixations on a region r. The three measures (10) maxVisitDuration (max millisecond), (11) meanVisitDuration (max millisecond) and (12) visitCount (max count) are based on visits. A visit describes the time between the first fixation on a region and the next fixation outside. The last measure (13) saccLength (max centimeter) [14] provided good results for the relevance feedback in image search. The assumption is that moving the gaze focus over a long distance (i.e., long saccade) to reach an image region r shows high interest in a region.

In our analysis, only fixations on the images are considered. Fixations on the experiment screen but outside the evaluated image are ignored.

3.2 Extending object boundaries and weighting small objects

We further investigate two parameters for identifying correlations between tags and image regions. The first parameter is an extension of the region boundaries to deal with the inaccuracy of eye tracking data. One obstacle in the identification of image regions from gaze information is the inaccuracy of the eye-tracker. For the Tobii device the accuracy is 0.5°. With a distance of 60 centimeters from the eye to the screen, this inaccuracy equates one centimeter on the screen or about 35 pixels. We investigate if this measurement uncertainty can be diminished by extending the region boundaries. By this, fixations near to a region are also considered belonging to the region. Values for the region extension $d = 1 \ldots 35$ pixels are analyzed.

The second parameter deals with the fact that larger image regions have the advantage to be more likely fixated by coincidence than smaller regions, e.g., while the subject is scanning the image on the search for an object. We analyze if the tag-to-region assignment can be improved by adding a linear weighting function to support smaller regions. The weighting depends on the image region size in relation to the total image size. $f_m(r)$ with $m = 1 \ldots 13$ is a measure functions applied on region r as described in Table 1.

In the following, we consider the linear weighting function *weighted-*f_m on an image region r:

$$weighted\text{-}f_m(r) = \begin{cases} f_m(r) \cdot weight(s_r) & \text{if } s_r \leq T \\ f_m(r) & \text{else} \end{cases} \tag{1}$$

with

$$weight(s_r) = \frac{1 - M}{T} s_r + M$$

The relative region size s_r is calculated from the size of the region in pixels divided by the image size in pixels. The measure is weighted with a factor only when $s_r \leq T$, where T is a predefined threshold. Thus, only image regions up to a specific size gain from the weighting function. The weighting factor itself is calculated depending on the threshold T and the maximum weighting value M. In our analysis, we have investigated the parameters M and T of the weighting function by calculating the precision values of all images for $T = 0 \ldots 1$ and $M = 1 \ldots 50$. An example of applying the *weighted-*f_m for $T = 0.05$ and $M = 4$ is shown in Fig. 2. Here regions of

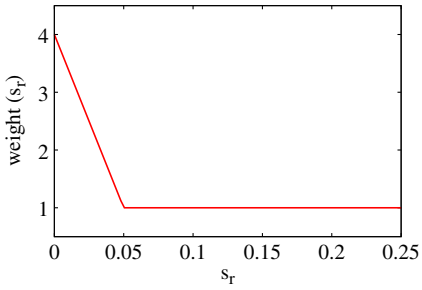


Fig. 2 Example weighting for $T = 0.05$ and $M = 4$

size between 0 % and 5 % of the actual image size are weighted with a factor between 1 and 4.

3.3 Baselines

In the following, we apply three baselines to compare our gaze-based approach to other approaches that are not based on using eye-tracking information. These baselines are (a) a "random" baseline [12], (b) a baseline based on the calculation of the most salient points on the image [17, 22], and (c) a "naive" baseline [14]. The random baseline (a) randomly selects one of the labeled regions of the image as favorite. The saliency baseline (b) assumes the depicted object at the most salient points on the images. The salient points were calculated by the toolbox offered by Itti et al. [7]. The favorite region is selected by using the salient points and their ordering as computed by Itti et al. and interpreting them as simulated gaze paths for our gaze analysis method. We use the measures from Section 3.1 to compute the favorite region from the saliency map. The naive baseline (c) makes the assumption that the area in the center of an image should be the favorite one. It was chosen because it is common that photographers position important motives in the middle of the image.

In this work, we compare with baselines that use the same input data as our approach. This is—besides the eye-tracking data—the set of images, their tags, and the manually created image regions from LabelMe. The use of methods based on a training set, methods requiring a training period, or methods that support a limited number of pre-defined concepts only (like typical object detection algorithms) are hard to compare to our approach as they require additional input data or a bigger data set. Thus, they have limitations our approach does not have. Leveraging such additional information and making use of machine learning approaches are beyond the scope of this work and are part of future work.

4 Experiment design

In the experiment application image-tag-pairs were presented to the subjects with the task to decide whether an object, described by the tag, is depicted on the image or not. The experiment application has been designed such that first a tag and subsequently an image was shown to the subjects.

4.1 Subjects

30 subjects (9 of them female) participated in our experiment. The age of the subjects was between 22 and 45 years (average: 28.7, SD: 6.78). They were undergraduate students (10), PhD students (17), or work in other professions (3). The subjects received a small present for participating.

4.2 Data set

As data set we used LabelMe[5] with 182.657 user contributed images (download August 2010). It provides images of complex indoor and outdoor scenes. The LabelMe community has manually created image regions by drawing polygons into the images and tagging them. The labels were used as tags and the regions as a manual, thus high quality image segmentation. The annotated regions are used as ground truth in our analysis.

For our experiment, we randomly selected images from the LabelMe data set with a minimum resolution of 1000×700 pixels and at least two labeled regions. In average, every image in our selection is labeled with 18.4 tagged regions (SD: 22.4, min: 3, max: 152). 72 % of the image areas are covered by the manually drawn polygons in average (SD: 32 %, min: 1 %, max: 100 %). From these images we created three sets of 51 images each with an assigned tag. For every image selected we randomly chose a "true" or "false" tag. "True" means that an object described by the tag was labeled on the image. "False" means that no label with the tag was given to the image. These "false" tags had been randomly selected from other LabelMe images. The purpose of creating true and false image-tag-pairs was to keep the subjects attentive during the experiment. We had to manually remove images from the selected ones when a) the randomly selected false tags by coincidence correlated to some actually visible parts of the image and thus were true tags. We also removed images where b) the tags where incomprehensible or expert knowledge was required. In some cases there were c) not all instances of an object labeled on the image.

4.3 Experiment setup

The experiment was performed on a screen with a resolution of 1680×1050 pixels. The experiment application was implemented as a simple web application running in Microsoft's Internet Explorer. The subjects' gazes were recorded with a Tobii X60 eye-tracker at a data rate of 60 Hz and an accuracy of 0.5 degree. For each image-tag-pair, the following three steps were conducted:

1. First, the tag with the question "Can you see the following thing on the image?" was presented to the subjects (see Fig. 3, left). After pressing the "space" button, the application continued with the next screen.
2. In this screen, a small blinking dot in the upper middle was displayed for one second (see Fig. 3, middle). The subjects were asked to look at that point. The red dot animated all subjects to start viewing the image (which has been shown

[5] http://labelme.csail.mit.edu/

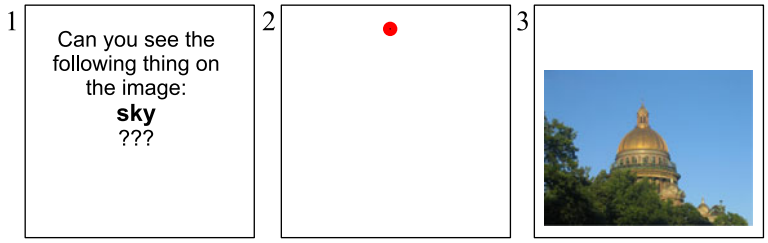


Fig. 3 Steps conducted for identifying image objects

next) from the same gaze position. It was placed above the actual image that is shown in the third screen.

3. Finally, the image was shown to the subjects (see Fig. 3, right). Viewing the image, the subjects had to judge whether the tag shown in the first screen would have an object counterpart in the image or not. The decision was made by pressing the "y" (yes) or "n" (no) key.

The first image-tag-pair was used to introduce the application to the subjects. It has not been used in the analysis. Each subject evaluated one of the three sets consisting of 51 image-tag-pairs from the data set described above. The subjects were told that the goal of the experiment was not to measure their efficiency in conducting the experiment task. No time constraints were given.

4.4 Evaluation results for effectiveness, efficiency, and satisfaction

Besides recording the raw gaze data, we also measured the time the subjects took to make a decision per image and the correctness of the answers. Additionally, we asked the subjects to express their emotions during the experiment on a 5-point-Lickert scale where a value of 1 means strong disagreement and a value of 5 stands for strong agreement.

Effectiveness We measured how many image-tag-pairs have been correctly classified by the subjects. Correctly classified means that a true tag is confirmed with "yes" and that a false tag is decided with "no" in our experiment application. In total, we received 1500 answers, 10 answers per image-tag-pair. 5.4 % of the given answers of all subjects are incorrect. The proportion of wrong answers for true (5.8 %) tags is close to the value for false tags (4.8 %). The highest number of wrong answers for one image-tag-pair is 8, i.e., most of the users did not correctly identify whether the tag given was true or false. In our work we only analyze the gaze paths of subjects having successfully identified a tag as true or false. We only consider image-tag-pairs with a true tag and a given the correct answer.

Efficiency The average answer time over all images is 3,003 ms (shortest answer time is 204 ms and the longest is 25,163 ms). 50 % of the answers are given in a time between 1,413 ms and 3,920 ms. For true tags, the average answer time over all subjects and all images is 2,818 ms, for false tags it is almost twice as long with 3,854 ms. Also the number of fixations on the image is higher for false tags (13 fixations in average) than for true tags (9.6 fixations). In an independent-samples

Mann-Whitney U Test we compared the answer times and number of fixations measured for true and false tags. For both tests we have obtained a significant difference with $p < 0.0001$. This means that the subjects look longer and more precisely on images when there is no object related to the tag provided.

Satisfaction Concerning the statement "It was easy to decide on an answer.", the subjects answered on average with a score of 3.85 (SD: 0.59). 15 subjects agreed or strongly agreed with the statement. Most of the subjects felt comfortable during the evaluation (average: 4.4, SD: 0.75). 11 strongly agreed and 6 agreed to the statement. Thus, we assume that the results obtained from the experiment application are not influenced by side effects like users feeling discomforted in front of the eye-tracker.

5 Results of finding objects in images

In total 1500 gaze paths were recorded (30 users, each viewing 50 images) during our experiment. Each of them contains fixations on the presented image. An average number of fixations per image over all images and all users is 10.9 (SD: 9.2, min: 1, max: 112). The gaze information was also recorded during and after the decision making by pressing of the "y"- or "n"-button on the keyboard. 88 % of the records contain fixations after the decision before the next page of the experiment application is shown.

We have analyzed only the gaze paths from images with a true tag and a correct answer given by the user (see Section 4.3). In cases where the subjects gave incorrect answers, we cannot conclude if the subjects did not took enough time to examine the image, did not understand the given tag, or if they had other problems. 799 gaze paths were collected during the experiment that fulfill our requirement. 656 (82 %) of these gaze paths have at least one fixation inside or near (10 pixels) a correct region.

The preprocessing of the raw eye tracking data for identifying fixations is performed with the fixation filter offered by Tobii Studio with the default velocity threshold of 35 pixels and a distance threshold of 35 pixels. The algorithm identifies fixations and saccades from detecting quick changes between the gaze points. The velocity threshold defines the maximum speed within one fixation. The distance threshold defines from which distance two fixations are merged into one single fixation.

5.1 Calculating the precision of tag-to-region assignments

The procedure for calculating the tag-to-region assignments is illustrated in Fig. 4. The single steps for every fixation measure are:

1. For every LabelMe region in an image b) a value for a fixation measure is calculated for every gaze path c).
2. For every region, the fixation measure results for every gaze path are summed up. From this, we obtain an ordered list of image regions for a fixation measure that determines the favorite region d) as described in Section 3.1.
3. The label of the favorite region is compared with the tag a) that was given to the subject in the experiment. If label and tag match, the assignment is true positive tp, otherwise it is a false positive fp. We sum up the total number of correct and

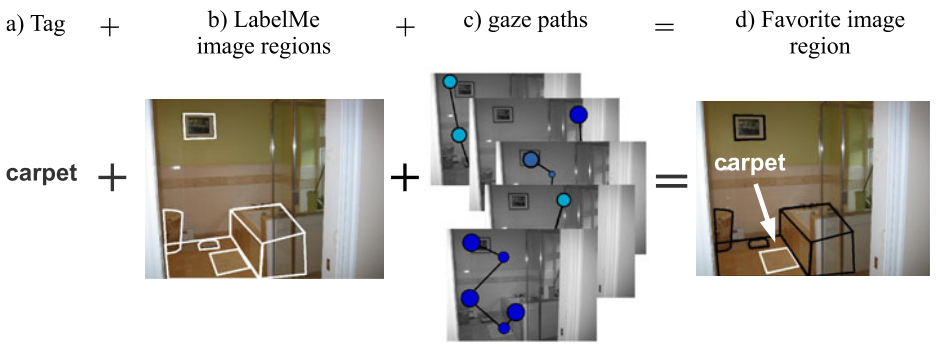


Fig. 4 Overview of calculating the tag-to-region assignments

incorrect assignments over all images and calculate the precision P for the whole image set using the following formula:

$$P = \frac{tp}{tp + fp} \tag{2}$$

5.2 Best fixation measures

The results for all measures are presented in Fig. 5. For each measure the tp and fp results and the precision P, calculated as described in Section 5.1, are depicted. We have received the best results for the measure (8) lastFixationDuration with precision $P = 0.55$. That means, 55 % of the image regions selected by the gaze analysis are described by the tag shown to the subjects. The second best value with $P = 0.54$ is (11) meanVisitDuration, followed by (6) fixationDuration with precision $P = 0.52$. The fourth best result is $P = 0.52$ for (4) fixationsBeforeDecision. We notice that among the best four measures, two measures take the moment of decision into account: (8) lastFixationDuration, (4) fixationsBeforeDecision. The top four measures are the same as in our work [32]. In the following investigations we look into the details of the three

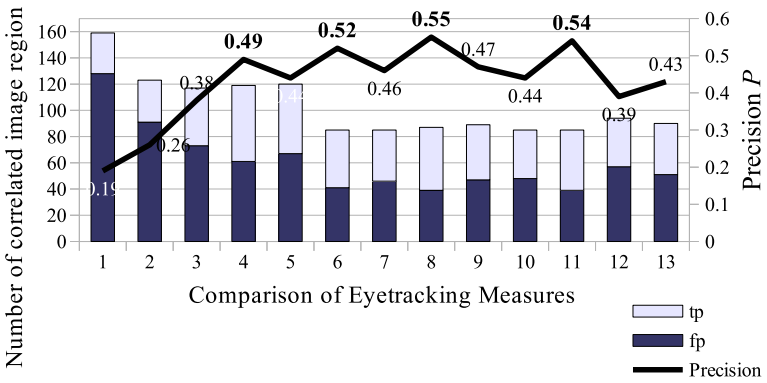


Fig. 5 Precision values for the fixation measures from Section 3 calculated from tp (true positive) and fp (false positive)

best measures. The lowest precision values are 0.19, and 0.26 for (1) firstFixation and (2) secondFixation. These measures are using the first fixations on an image and the fp values are very high. We further examine this problem in Section 6.4.

Figure 7 shows some examples of successfully identified tag-to-region assignments. A closer look at the image region characteristics and a qualitative description of the incorrect correlations can be found in the detailed analysis presented in Section 6.

We have investigated how the complexity of a scene influences the respective fixations measures. As measure for the complexity of a scene, we use the number of tagged regions per image. The number of tagged regions nt is clustered according to the three quartiles ($Q_1 = 6.25$, $Q_2 = 11.5$, $Q_3 = 21$). The maximum difference $diff$ between the precision results for different quartiles for one measure is also calculated. The results are depicted in Fig. 6. For each measure from (1) firstFixation to (13) saccLength the precision values P are calculated separately for images with a number of tagged regions between 0 and Q_1, Q_1 and Q_2, etc . In general, more correct assignments were performed for images with less tagged regions. This finding is not surprising as it is easier to perform a correct assignment by chance for less complex scenes with less regions. The influence of the scene complexity is varying between the measures. The three best performing measures have an average $diff$ value between 0.33 and 0.35. The measure (5) fixationsAfterDecision with the smallest result $diff = 0.21$ shows an average precision performance (Fig. 7). Q1

5.3 Extension of region boundaries

We have investigated the influence of the extension on the precision for the three best performing measure (8) lastFixationDuration, (11) meanVisitDuration and (6) fixationDuration (see Section 3). The precision increases when applying the extension parameter. The best result is precision $P = 0.6$ for (8) lastFixationDuration with $d = 18$ as shown in Fig. 8. This equates to an improvement of about 9 %, compared to the result of $P = 0.55$ without extension. A baseline is added to each diagram, displaying the precision value without extension. The precision is even below (> 1 %) the threshold for $d < 6$ for (8) lastFixationDuration, $d > 32$ for (6) fixationDuration, and $d > 29$ for (11) meanVisitDuration.

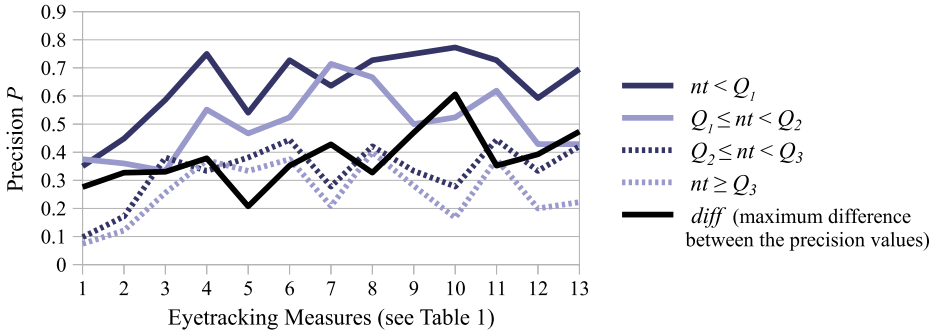


Fig. 6 The precision P compared for different levels of scene complexity (measured by the number of tagged regions nt)

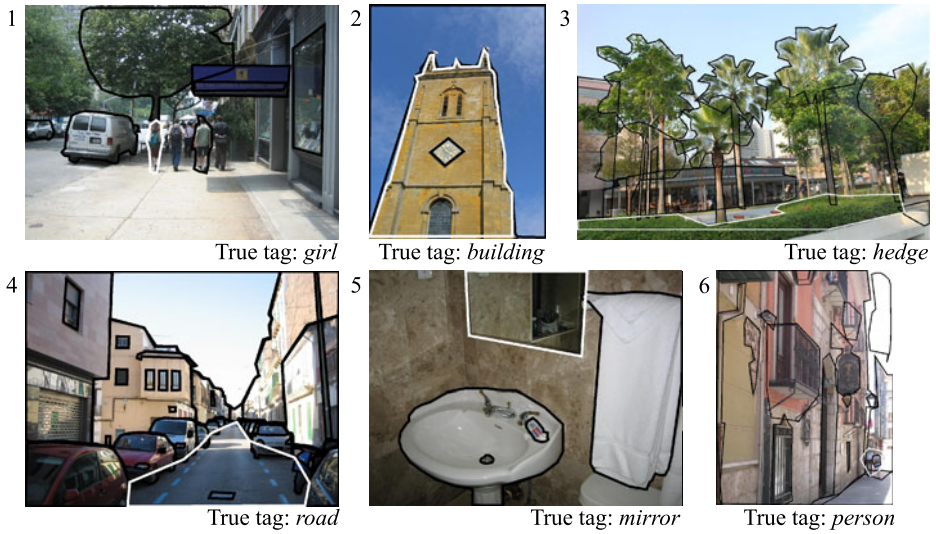


Fig. 7 All labeled regions (black shapes) and correctly identified favorite objects (white shapes)

The results suggest that it is reasonable to include the extension of region boundaries in the calculation of tag-to-region assignments. The precision is fluctuating depending on the chosen extension value d. In our investigations best results are obtained for $6 \leq d \leq 29$.

5.4 Weighting function

The best precision value applying the weighting function on the fixation measure (8) lastFixationDuration is $P = 0.56$, the worst result is $P = 0.47$. (for (11) meanVisitDuration best $P = 0.6$ and worst $P = 0.53$, for (6) fixationDuration: best $P = 0.54$ and worst

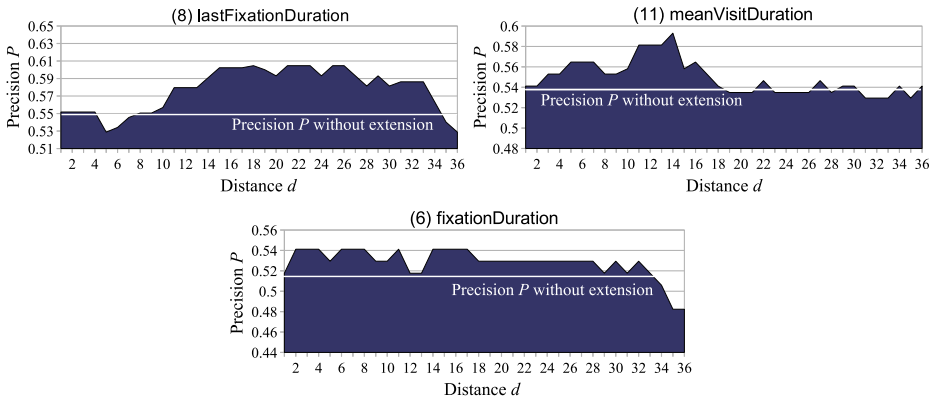


Fig. 8 Influence of extension parameter

$P = 0.48$). These values are provided by different combination of M and T. In Fig. 9, the results for the two weighting parameters are displayed. As baseline precision, we consider the precision values obtained without extension and weighting (see Section 3). Values equal to this baseline are marked in white. Values higher than and lower than the baseline precision are highlighted in the figure in red respectively blue. From the results depicted in Fig. 9, one can see that the influence of parameter T is higher than the influence of M. The precision values are strongly varying. Every chart (a) to (c) shows an area of highest values for $0.04 < T < 0.1$. The precision decreases for every measure from $T > 0.13$, but also here good precision values can appear for higher T.

The usage of the weighting function can improve the results. However, the precision can also decrease. Further investigations are necessary to better explain the fluctuation of the graph.

5.5 Combination of region extension and weighting function

Finally, we use the three best performing fixation measures and combine both parameters of the region extension and the weighting function. The best precision value for fixation measure (8) lastFixationDuration is $P = 0.62$, the worst result is $P = 0.49$. The best result is delivered with $P = 0.63$ by (11) meanVisitDuration, including extension $d = 10$ and weighting (e.g., $T = 0.05$, $M = 4$). For (6) fixationDuration the best result is $P = 0.61$, the worst $P = 0.51$.

5.6 Comparison of the eye tracking approach with three baselines

We compare the precision values obtained with our approach to the three baselines described in Section 3.3. The results in Fig. 10 show that the random baseline has an average precision of 0.17 over 30 samples (SD: 0.04, min: 0.1, max: 0.26). The saliency approach has a best precision of 0.21 for the measure (11) meanVisitDuration, followed by a precision of 0.20 for (1) firstFixation. The worst result was obtained with a precision of 0.15 for the measure (2) secondFixation. The naive approach achieves a precision of also 0.21. These baseline results are compared to the gaze-based approach with precisions between 0.52 and 0.55 for the measures (6), (8), and (11), and between 0.61 and 0.63 for the measures with extension and weighting. The identification of assignments based on gaze or on gaze including extension and weighting performs better than the baseline approaches. We have performed 18 Chi-square tests to investigate significant differences between the approaches. They all show a statistical significance of $p < 0.0015$. We received the least significant result with $X^2(1, N = 124) = 10.723$, $p < 0.0015$, $\phi = 0.162$ for the Naive Baseline and measure (6) fixationDuration without extension and weighting. Further details of the X^2-tests are omitted for reasons of brevity.

6 Detailed analysis of image region characteristics and gaze paths

The best precision value $P = 0.63$ for measure (11) meanVisitDuration (including extension and weighting) is calculated from 54 tp and 32 fp assignments. In this section, we first present a qualitative analysis of the fp assignments. Subsequently,

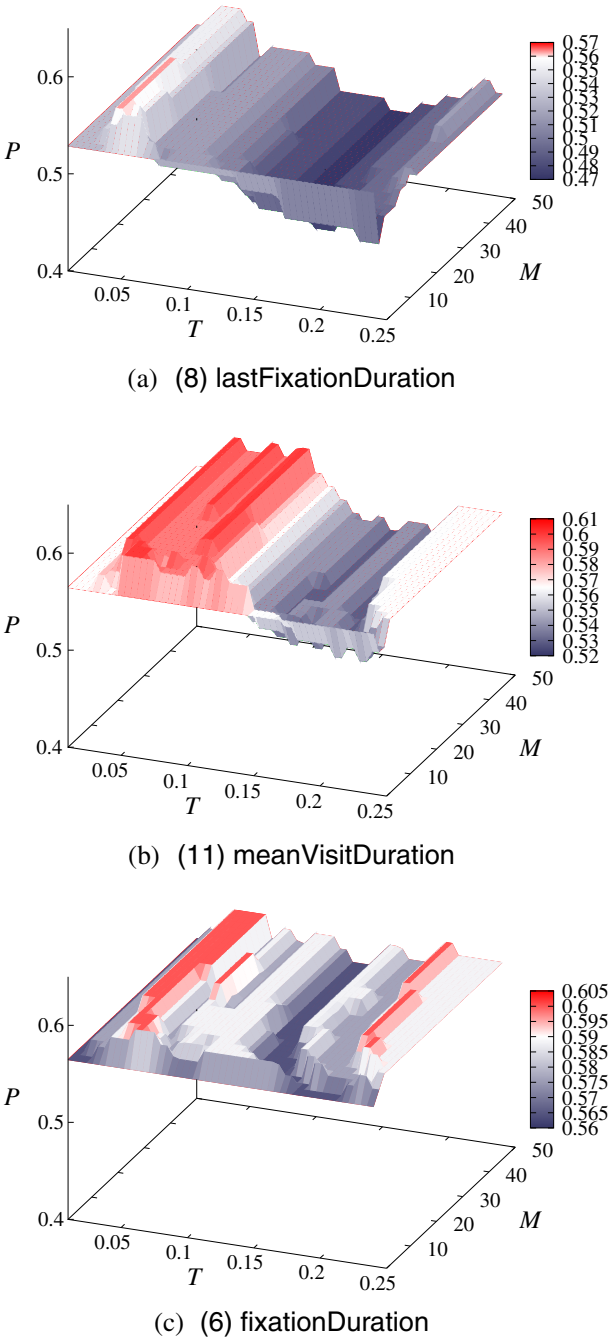


Fig. 9 Influence of the weighting function on precision P for different measures (white: baseline without weighting)

we investigate if there are typical characteristics concerning region sizes or positions of image regions for correct and incorrect tag-to-region assignments, followed by a

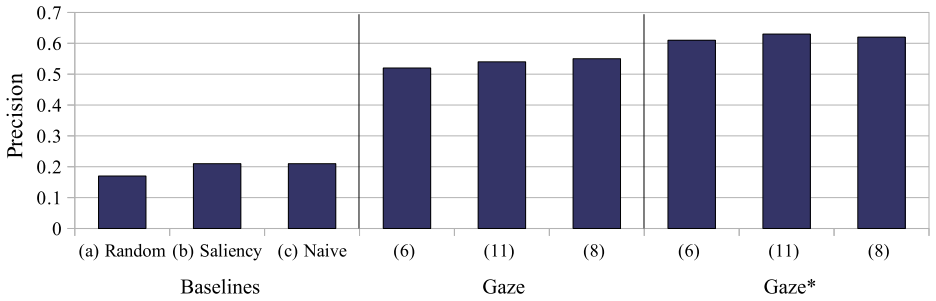


Fig. 10 Precision for three baselines and gaze based analysis

look into typical patterns for the first fixations. Finally, the effect of aggregating gaze paths of several subjects is investigated.

6.1 Qualitative analysis of incorrect assignments

Some examples of incorrect assignments can be seen in Fig. 11. The white boundaries show the object that corresponds to the tag given to user. The black boundaries show the objects determined as favorite from the gaze information. The correlations are calculated with measure (11) meanVisitDuration including extension and weighting.

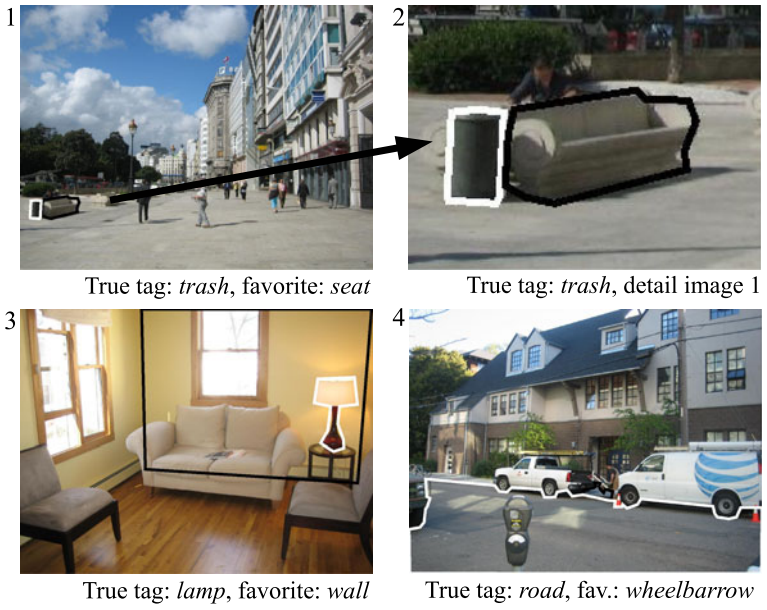


Fig. 11 Examples of image-tag-pairs with given tags (white shape) and incorrectly identified favorites (black shape)

From an qualitative analysis of the 32 wrongly assigned tags to regions, we identified the following characteristics of the images with incorrect assignments:

- Some images show scenes with a small correct object and a wrongly selected favorite object also small and located next to the correct object (cf. images 1 and 2). Six images belonged to this category. This problem can be based on the inaccuracy of the eye-tracker.
- In some images, the correct object is displayed within another object (cf. image 3, *lamp* inside *wall*). In these cases, the outer region is identified as favorite. That means our weighting function does not work for all occurrences of smaller regions. Eight images belong to this category.
- Further images show scenes with an object that seems to be easy to identify. For example larger objects like *road* (cf. image 4), *sky* or *tree* might be perceived even in the corner of the human eye or based on context knowledge (e.g., sky is above sea is above sand in a beach scene). Nine images belong to this category. This is a basic limitation of our approach, but it appears infrequently in comparison to the number of all shown images.

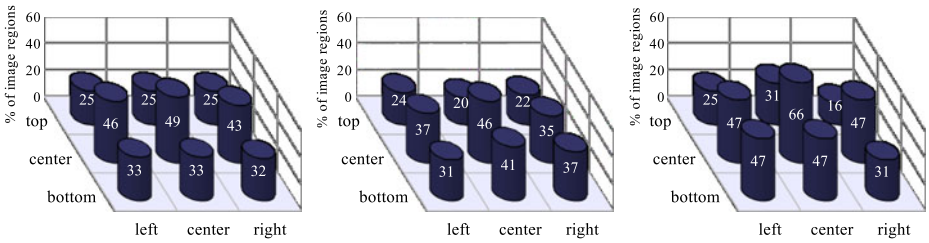
6.2 Comparing the region size for correct vs. incorrect assignments

The average size of the LabelMe regions in the images used in our experiment is 66,3811 pixels. The average region size for correctly assigned regions *tp* is 123,609, for incorrectly assigned regions *fp* 214,704 pixels. The region size of the selected favorite regions (*tp* or *fp*) is clearly larger than the average region size. Thus, larger regions are selected with a higher probability for tag-to-region assignments by our approach. It is also interesting to notice that the average region size of *fp* assignments is about 70 % larger than the region size of *tp* assignments.

6.3 Comparing the region positions for correct vs. incorrect assignments

We have divided the images into nine uniform areas. Based on these areas, we have investigated the positions of the assigned regions. The percentage of image regions having an overlap with the particular area is calculated. In Fig. 12a, the positions of all regions in our data set corresponding to true tags are depicted. 49 % of the regions overlap with the center field of the image. In the upper third of the images is only one fourth of the regions located. In the lower areas it is about one third. This could be explained by how people take images, e.g., with the object in the center of the image and sky or the ceiling in the upper areas. The differences between the left and the right areas are very small. In Fig. 12b and c, the positions of correctly and incorrectly assigned regions are summarized. One can see in Fig. 12b, that the positions of the correctly assigned regions are distributed over the image areas in a similar way as the true-tag image regions (cf. Fig. 12a). For the correct assignments it is not possible to identify a privileged area on the image.

For the incorrect assignments in Fig. 12c, one can notice that the positions of the regions concentrate in the center of the image. One can further observe that in the center top part the value is also increased compared to the true-tag image regions. The total number of touched areas is bigger for (c) compared to (a) and (b). This finding is based on the bigger size of incorrect areas, as described in Section 6.2. This higher percentage of wrongly assigned regions might be caused by a concentration



(a) Positions of all tagged regions (b) Positions of correct favorite regions (c) Positions of incorrect favorite regions

Fig. 12 Percentage of regions located in image areas

of fixations in the center of the images. This concentration has been observed during the first fixations on the images as shown in the next section.

6.4 Influence of the first five fixations to assignment precision

Figure 13 shows an illustration of the first five fixations over all subjects and all images. One can see that the first fixations are concentrated in the center of the images. Later, the fixations are better distributed over the whole image. This effect is called center bias and was described amongst others by Judd et al. [9] and Zhao and Koch [35]. The appearance of such a bias is based on different factors like the experience that photographers place the most important objects in the center of an image or simply the straight-ahead position in front of the screen. In both above-mentioned works, the eye tracking information showing the center bias is collected in free-viewing scenarios (i. e., no specific task was given to the users, they were asked to just view the images). The influence of this bias was not clear in task-driven viewing and a fixated starting point outside the image itself. As can be seen in Fig. 13, the center bias is highly distinct only for the first fixations. This is a valuable finding for our work, as we also consider the fixation order in our analysis. The weak results for the measure (1) firstFixation and (2) secondFixation show this problem. In our experiment setup, the subjects were asked to look at a red dot—placed above the image position—before the image appeared on the screen (see Section 4.3). The influence of this point can be seen in the illustration of the first and second fixations, because of the fixations in the upper center of the images. This also provides an explanation for the high value of incorrect aggregations in the middle of the images in the previous Section 6.3.

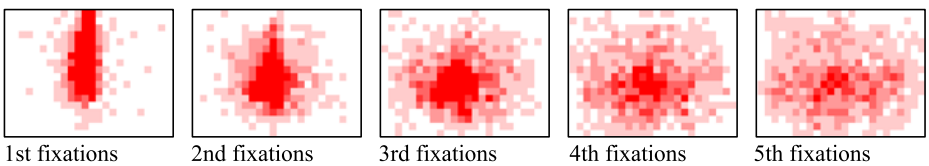


Fig. 13 First five fixations accumulated over all subjects and all images

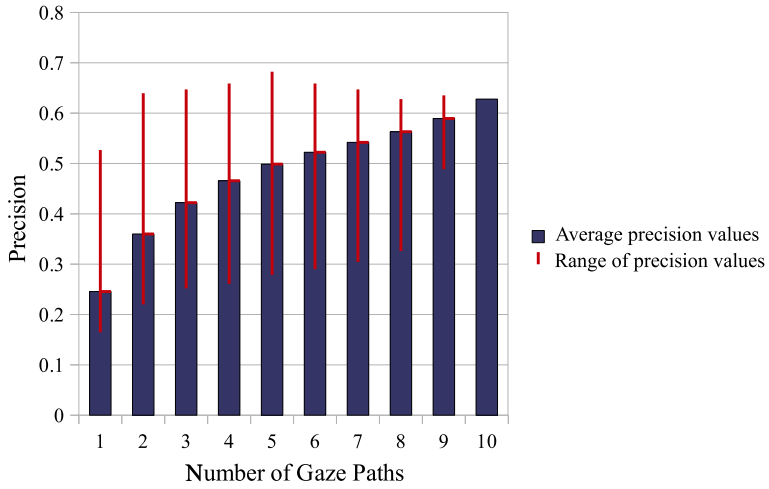


Fig. 14 Effect of aggregation of gaze paths from one up to ten users

6.5 Effect of aggregation of gaze paths on precision

Finally, it is interesting to know how many users are needed to accomplish a certain level of reliability in assigning a tag to the correct region. Thus, we have investigated which precision can be reached when aggregating and using an increasing number of users' gaze paths. We present precision values for aggregating the gaze paths of 1 to 10 subjects for the measure meanVisitDuration, including extension and weighting. Precision P is calculated for every possible subset of subjects and averaged for all subgroups of the same size. As shown with the bars in Fig. 14, the number of users has a high influence. With the gaze paths of only single users, we have received an average precision (over all users and all images) of $P = 0.25$ (SD: 0.1, min: 0.16, max: 0.53). For the aggregated data for all 10 users we got a precision $P = 0.63$. This corresponds to an improvement of 152 %. The biggest improvements take place between the first group sizes. For example between one and two users per group we have an improvement of 46 % in average. Between nine users and ten users per group, there is only an improvement of 7 %.

In addition, the range between minimum and maximum precision values is depicted in Fig. 14. The range decreases from the single user results to the multiple user results. Even for single users and single images a good precision can be achieved for some images and regions, respectively. For 10 users we only have one set, therefore no range can be indicated. The big step for the minimum values between the subgroups of 8 and 9 users could also be caused by the small number of only 10 subsets for 9 users. The results based on multiple gaze paths are considerably better than the ones calculated from only a few gaze paths.

7 Discriminating different objects in one image

With research question (b) (see introduction in Section 1), we investigated the possibility to differentiate objects by analyzing the users' gaze paths. To perform this

analysis, two of the three image data sets from Section 4.2 were composed from the same image subsets. As a result, we got two sets of 51 image-tag-pairs each, sharing the same images but different tags. All combinations of correct and incorrect tags appear: images with a correct tag for both sets, images with one correct, one incorrect tag and images with two incorrect tags. Our data set includes 16 true-true image-tag-pairs (tags for both groups are true), 24 true-false image-tag-pairs (one tag is true, one false), and 10 false-false image-tag-pairs. In this section, we use again the measure (11) meanVisitDuration, including extension and weighting.

7.1 Proportion of correctly discriminating two objects

For the 16 images with two correct tags, the favorite image regions were calculated. In 6 images, two correct image regions were identified. This is a proportion of 38 %. In Fig. 15, some examples with two correctly identified regions are shown. As the figure shows, the two tags *sky* and *sea* could be distinguished in the upper image. Also the tags *water pot* and *teas* in the lower image could be identified using gaze information. The average probability to identify the correct region in one image is 63 % (see Section 5). Therefore, the probability to obtain two correct tag-to-region assignments in two different images is 40 %. With a value of 38 % for two image regions in one image, the probability is close to the probability for two image regions in two different images. Thus, it is possible to identify different image regions in one image with an accuracy close to the accuracy of the single assignments. The 16 images with two correct tags provided to the subjects has in average 15 tagged regions (SD: 16, min: 3, max: 62). The six images with two correctly identified favorite regions have

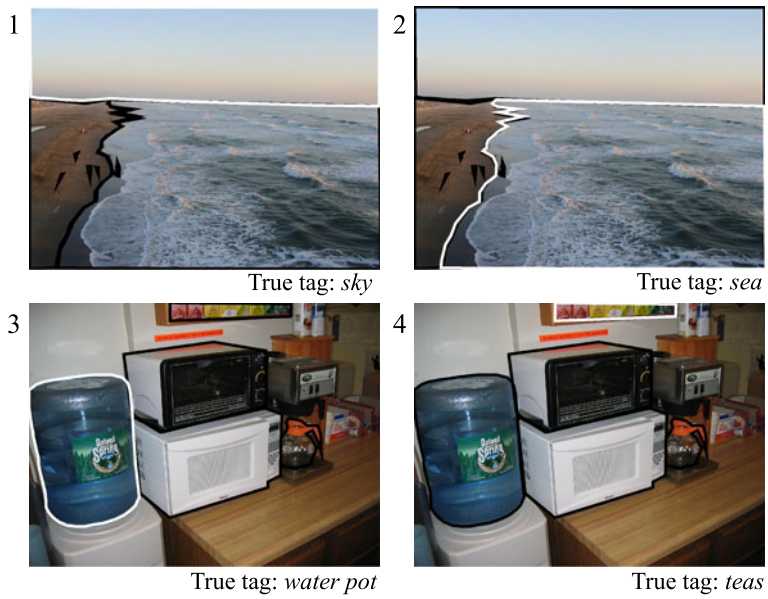


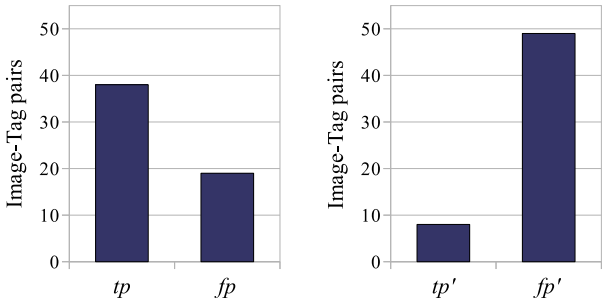
Fig. 15 Example images with two correctly identified regions (white boundaries)

an average of 17 tagged regions (SD: 22, min: 5, max: 62), whereas the 10 images with one or two incorrect favorite regions has in average 14 tagged regions (SD: 11, min: 3, max: 37). These results indicate that the rate of successfully assigned tags is not or only weakly influenced by the complexity of the scene. An accumulation of the error for detecting multiple objects in one image could lead to an overall low precision of the tag-to-region assignments. However, in this work the number of investigated images with their tags assigned is small and more detailed investigations are part of interesting future work.

7.2 Influence of different tag primings on tag-to-region assignments

In this section, we examine the influence of priming by the given tags to the tag-to-region assignments. Every true tag t_r assigned to an image describes one or multiple image regions r. We compare the results of our approach for users with a provided true tag t_r with results from users, viewing the same image, but given a false tag or a tag describing another region on the image. We investigate how often region r is determined as favorite region by chance, i.e., when not the tag t_r was viewed in advance. We use the 16 true-true and 24 true-false image-tag-pairs to perform this calculation.

The tp and fp values in Fig. 16a show the results for the assignment of tag t_r to region r from our analysis in Section 5. A tp assignment means that the favorite region is described by the tag presented to the user. The fp assignments describe the incorrect correlations, i.e., when a favorite region is selected that is not r. The second tag for the same image is used to investigate if (for some reason) the previous region r is selected as favorite. tp' in Fig. 16b shows how often the region r is identified as favorite from gaze paths that did not belong to tag t_r, i.e., where a tag is provided to the users that does not refer to region r. Rather, the tag given to the users could be incorrect (true-false image-tag-pairs) or correct for another region in the image (true-true image-tag-pairs). In case of fp', the region r is not identified as favorite. Thus the fp' assignments mean that the investigated region is not described by the tag presented to the user and the region r was not determined as favorite. We get a low precision value of $P = 0.12$ in our calculations from tp' and fp'. That means that the region r referring Fig. 16a is rarely selected when a tag is shown to the



(a) Results for users given a tag t_r describing region r (see Section 5)

(b) Results for users not given tag t_r

Fig. 16 Comparing the identification of region r as favorite from gaze paths (**a**) corresponding and (**b**) not corresponding to tag t_r

user that does not correlate to the image at all or correlated to a region different from r.

We demonstrate that the assignments to region r by providing a tag referring to a different region or not referring to any region at all are significantly lower compared to the assignments based on true tags of region r. The result of a Chi-square test shows that the difference is significant with $x^2(1, N = 114) = 32.8005$, $p < 0.0001$, $\phi = 0.5364$. The correct assignments do not appear coincidentally but strongly depend on the gaze paths, guided by the given tag.

8 Conclusions

The results of our experiment show that the best performing fixation measure can identify image regions at a precision of 63 %. In addition, we can state that taking extensions of region boundaries into account as well as weighting of smaller regions improves the results. However, it is not possible to clearly determine the best parameters for region extension and weighting from our experiment. Our detailed analysis shows that there is a higher concentration of regions in the center of the images. In addition, more incorrect tag-to-region assignments are made in the center than correct assignments. Investigating the first fixations on the image explains the low precision values of measures like firstFixation and shows the center bias. Additionally, we have shown that two regions can be differentiated in the same image with an accuracy of 38 %. By evaluating the effect of different primings such as providing different tags, we have shown that the identified tag-to-region assignments are not just a matter of chance but are the results of analyzing the users' gaze path.

Acknowledgement We thank the subjects participating in our experiment. The research leading to this article was partially supported by the EU project SocialSensor (FP7-287975).

References

1. Bruneau D, Sasse M, McCarthy J (2002) The eyes never lie: The use of eye tracking data in HCI research. In: Proceedings of the CHI, vol 2
2. Campbell RJ, Flynn PJ (2001) A survey of free-form object representation and recognition techniques. Comput Vis Image Underst 81(2):166–210
3. Castagnos S, Jones N, Pu P (2010) Eye-tracking product recommenders' usage. In: Proceedings of the 4th ACM conference on recommender systems. ACM, pp 29–36
4. Duygulu P, Barnard K, De Freitas J, Forsyth D (2006) Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In: Computer vision, ECCV 2002, pp 349–354
5. Grabner H, Gall J, Van Gool L (2011) What makes a chair a chair? In: 2011 IEEE conference on computer vision and pattern recognition (CVPR). IEEE, pp 1529–1536
6. Hajimirza S, Izquierdo E (2010) Gaze movement inference for implicit image annotation. In: Image analysis for multimedia interactive services. IEEE
7. Itti L, Koch C, Niebur E (1998) A model of saliency-based visual attention for rapid scene analysis. IEEE Trans Pattern Anal Mach Intell 20(11):1254–1259
8. Jaimes A (2001) Using human observer eye movements in automatic image classifiers. In: SPIE. ISSN 0277786X. doi:10.1117/12.429507
9. Judd T, Ehinger K, Durand F, Torralba A (2009) Learning to predict where humans look. In: IEEE international conference on computer vision (ICCV). Citeseer

10. Kim D, Yu S (2008) A new region filtering and region weighting approach to relevance feedback in content-based image retrieval. J Syst Softw 81(9):1525–1538
11. Klami A (2010) Inferring task-relevant image regions from gaze data. In: Workshop on machine learning for signal processing. IEEE
12. Klami A, Saunders C, De Campos T, Kaski S (2008) Can relevance of images be inferred from eye movements? In: Multimedia information retrieval. ACM
13. Kompatsiaris I, Triantafyllou E, Strintzis M (2001) A World Wide Web region-based image search engine. In: Conference on image analysis and processing. doi:10.1109/ICIAP.2001.957041
14. Kozma L, Klami A, Kaski S (2009) GaZIR: gaze-based zooming interface for image retrieval. In: Multimodal interfaces. ACM
15. Li X, Snoek CGM, Worring M (2009) Annotating images by harnessing worldwide user-tagged photos. In: Acoustics, speech, and signal processing. IEEE, pp 3717–3720
16. Liu X, Cheng B, Yan S, Tang J, Chua T, Jin H (2009) Label to region by bi-layer sparsity priors. In: Proceedings of the 17th ACM international conference on multimedia. ACM, pp 115–124
17. Navalpakkam V, Itti L (2005) Modeling the influence of task on attention. Vis Res 45(2): 205–231
18. Pasupa K, Saunders C, Szedmak S, Klami A, Kaski S, Gunn S (2009) Learning to rank images from eye movements. In: IEEE 12th International conference on computer vision workshops, (ICCV Workshops '09)
19. Privitera CM, Stark LW (2000) Algorithms for defining visual regions-of-interest: comparison with eye fixations. IEEE Trans Pattern Anal Mach Intell 22(9):970–982
20. Ramanathan S, Katti H, Huang R, Chua T-S, Kankanhalli M (2009) Automated localization of affective objects and actions in images via caption text-cum-eye gaze analysis. In: Multimedia. ACM. New York, USA. ISBN 9781605586083. doi:10.1145/1631272.1631399
21. Ramanathan S, Katti H, Sebe N, Kankanhalli M, Chua T (2010) An eye fixation database for saliency detection in images. In: Computer vision–ECCV 2010, pp 30–43
22. Rowe N (2002) Finding and labeling the subject of a captioned depictive natural photograph. IEEE Trans Knowl Data Eng 14(1):202–207. ISSN 1041-4347
23. Russell BC, Torralba A, Murphy KP, Freeman WT (2008) LabelMe: a database and web-based tool for image annotation. International Journal of Computer Vision 77(1):157–173
24. Santella A, Agrawala M, DeCarlo D, Salesin D, Cohen M (2006) Gaze-based interaction for semi-automatic photo cropping. In: CHI. ACM, pp 780
25. Schneiderman H, Kanade T (2000) A statistical method for 3d object detection applied to faces and cars. In: Proceedings of the IEEE conference on computer vision and pattern recognition, vol 1. IEEE, pp 746–751
26. Sewell W, Komogortsev O (2010) Real-time eye gaze tracking with an unmodified commodity webcam employing a neural network. In: Proceedings of the 28th of the international conference extended abstracts on human factors in computing systems. ACM, pp 3739–3744
27. Tang J, Yan S, Hong R, Qi G, Chua T (2009) Inferring semantic concepts from community-contributed images and noisy tags. In: Proceedings of the 17th ACM international conference on multimedia. ACM, pp 223–232
28. Torralba A, Murphy K, Freeman W (2007) Sharing visual features for multiclass and multiview object detection. IEEE Trans Pattern Anal Mach Intell 29(5):854–869
29. Tsai D, Jing Y, Liu Y, Rowley H, Ioffe S, Rehg J (2011) Large-scale image annotation using visual synset. In: 2011 IEEE international conference on computer vision (ICCV). IEEE, pp 611–618
30. Viola P, Jones M (2001) Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE Computer Society Conference on computer vision and pattern recognition, CVPR 2001, vol 1. IEEE, pp 511–518
31. von Ahn L, Liu R, Blum M (2006) Peekaboom: a game for locating objects in images. In: CHI. ACM, 2006. ISBN 1-59593-372-7
32. Walber T, Scherp A, Staab S (2012) Identifying objects in images from analyzing the users gaze movements for provided tags. In: Advances in multimedia modeling. Springer, pp 138–148
33. Walber T, Scherp A, Staab S (2013) Can you see it? two novel eye-tracking-based measures for assigning tags to image regions. In: Advances in multimedia modeling. Springer, pp 36–46
34. Yarbus A (1967) Eye movements and vision. Plenum press
35. Zhao Q, Koch C (2011) Learning a saliency map using fixated locations in natural scenes. J Vis 11(3):1–15

Tina Walber is a PhD student at the Web Science and Technologies Institute in Koblenz, Germany since 2010. She studied computer science with focus on computer vision at the University of Koblenz-Landau, Germany and finished her degree in 2007 with the diploma thesis on a 3D face scanner. She is cofounder of a startup company and worked as web-developer and interaction designer. Her current research interests include eye tracking and human computer interaction.

Ansgar Scherp is leader of the focus group on Interactive & Multimedia Web at the Institute for Web Science and Technologies. He has received his PhD at the University of Oldenburg, Germany with distinction in 2006. He has been EU Marie Curie Fellow at the Donald Bren School of Information and Computer Sciences, University of California, Irvine, USA between 2006 and 2008. He received the best paper award for "Paving the Last Mile for Multi- Channel Multimedia Presentation Generation", MMM, 2005, and is winner of the Billion Triple Challenge of the Semantic Web Conference in 2008 and 2011. He has published over 50 peer- reviewed scientific publications.

Steffen Staab is leading the Web Science and Technologies Institute in Koblenz, Germany. His research led to over 150 refereed publications and 7 books. In his previous work, he has originally developed the concept of semantic portals, the notion of ontology learning and was among the first to work on emergent semantics. While his primary expertise is on Semantic Web and data publishing, over the last four years he has investigated data mining on sensor data, in particular social media.