Identifying Objects in Images from Analyzing the Users' Gaze Movements for Provided Tags

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

[1]Institute for Web Science and Technology
[2]Institute for Information Systems Research
University of Koblenz-Landau, Germany
{walber,scherp,staab}@uni-koblenz.de
http://west.uni-koblenz.de/

Abstract. Millions of users share, tag, and search for images on social media platforms and social networking sites today. Annotating and searching for specific image regions, however, is still very hard. Assuming that eye tracking will be a common input device in the near future in notebooks equipped with cameras and mobile devices like iPads, it is possible to implicitly gain information about images and image regions from these users' gaze movements. In this paper, we investigate the principle idea of finding specific objects shown in images by looking at the users' gaze path information only. We have analyzed 547 gaze paths from 20 subjects viewing different image-tag-pairs with the task to decide if the tag presented is actually found in the image or not. By analyzing the gaze paths, we are able to correctly identify 67% of the image regions and significantly outperform two baselines. In addition, we have investigated if different regions of the same image can be differentiated by the gaze information. Here, we are able to correctly identify two different regions in the same image with an accuracy of 38%.

1 Introduction

To describe the semantics of images on social media platforms such as Flickr[1] and social networking sites like Facebook[2] users can allocate tags to the images. Nevertheless, tagging describes the semantics of the images in a limited way. One step towards improving the understanding of image semantics is to annotate specific image regions instead of the entire image. Although tagging image regions is in principle possible on these platforms and sites, the annotation is manual and thus quite tedious. In this paper, we are investigating if it is in principle possible to automatically assign tags to objects by analyzing the users' gaze paths. In order to analyze the gaze paths in a controlled manner, we have designed an experiment in which 20 subjects have viewed a sequence of 51 tag-image-pairs each. For each tag shown to the subjects, they had to decide whether or not an object corresponding to that tag can be found in the image. During

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

the experiment, the users' gaze paths with the fixations are recorded. Fixations are the phases of highest visual perception in the movements of the eyes, which are briefly focused on a particular point on the screen. 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. The tag is assigned to the region with the highest fixation measure value. We have applied 13 fixation measures to explore their performance on determining these tag-to-region assignments. The results show a maximum precision of 67% that significantly outperforms two baselines. In addition to finding specific objects in images, we have investigated if it is possible to differentiate different objects shown in the same image by looking at the gaze paths. The results show an accuracy of 38% of two correctly identified objects in the same image and show potential for future improvements.

2 Related Work

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. Other web platforms like LabelMe [9] allow for the 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 [12].

With respect to the automatic *segmentation and labeling of images*, Rowe [8] presents an approach to find the visual focus of an image by applying image processing in terms of segmentation and low-level features. Goal is to link the visual focus with the image caption. This approach is designed for images with a single object only [8]. In addition, it has many limitations concerning the position and characteristics of the shown object.

Usability studies are a standard use case for applying gaze information. For detailed analysis, regions of interest (ROI) are marked on the investigated medium, e.g., a web page or an image showing a commercial. Based on these ROIs, the users' attention is analyzed in order to optimize the object that is under examination [1]. 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.

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, 2, 5]. Jaimes et al. [3] carried out a preliminary analysis of identifying common gaze trajectories in order to classify images into five, predefined semantic categories. They do not consider image regions and the categories are very general. Santella et al. [10] present a method for semi-automatic image cropping using gaze information in combination with image segmentation. Goal is to find the most important image region but not to conduct a general identification of image objects. Klami et al. [4] 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 that identify the regions of interest. The work revealed that the region identified depends on the task given to the subject before viewing the image. However the given task was very general and thus the work does not aim at identifying single objects in the images from the generated heat map. Finally, the work of Ramanathan et al. [7] aims at localizing affective objects and actions in images by using gaze information. Thus, the 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.

The related work shows that it is in principle possible to relate image regions with gaze path information. In contrast to our work, current research does not tackle the identification of objects in images based on the users' gaze information.

3 Experiment Design

The setup of our experiment was designed such that the users' gaze paths are obtained in a controlled manner. In our experiment application, we show tags to the subjects instead of asking them to enter own tags. In addition, the experiment application is designed such that first a tag and subsequently an image is shown to the subjects. The subjects were asked to decide whether or not an object described by the tag is shown on the image. 20 subjects (4 female) have participated in our experiment. The age of the subjects is between 23 to 40 years (average: 29.6 years). Their professions are undergraduate students (6), PhD students (12), and office clerks (2).

As data set, we use LabelMe[3] with 182.657 user contributed images (download August 2010). The LabelMe community has manually created image regions by drawing polygons into the images and tagging them. These manually created and annotated regions are used as ground truth in our experiment. The labels are used as tags and the regions as a manual, thus high quality image segmentation. For our experiment, we have randomly selected 51 images from the LabelMe data set. The images selected for our experiment have a minimum resolution of 1000x700 pixels and contain at least two labeled regions. We have created two sets of 51 tags and assigned one tag of each set to one image. Thus, each image has two tags. The two sets of tags are needed for the second part of our experiment aimed at discriminating two different objects shown on the same image. For every tag selected and assigned to the images, we have randomly decided if it should be a "true" or "false" tag. Here, "true" means that an object described by the tag can actually be seen on the image. The true tags are obtained from the labeled regions belonging to an image. The other tags were "false" and cannot be seen on the image. They were randomly selected from other LabelMe images. We had to manually replace images from the selected ones when a) the

[3] `http://labelme.csail.mit.edu/`

randomly selected false tags by coincidence correlate to some actually visible parts of the image and thus were true tags. We also replaced images where b) the tags where incomprehensible or expert knowledge is required and nonsense tags. In some cases there is c) a tag associated to a region like bicycle but multiple bicycles are depicted on the image and not all regions are explicitly marked as such. Thus, not all instances of the object the tag is referring to are actually labeled in the image. Finally, we have also removed images, where d) the object of interest is obstructed by other objects like a bicycle behind a car. Please note that the purpose of creating true and false image-tag-pairs is to keep the subjects concentrated during the experiment.

The experiment was performed on a screen with a resolution of 1680x1050 pixels. The subjects' gaze was recorded with a Tobii X60 eye-tracker at a data rate of 60Hz and an accuracy of 0.5 degree. The experiment was running as a simple web page in Microsoft's Internet Explorer. For each image-tag-pair, the following three steps are conducted as illustrated in Figure 1.

1. First, the tag with the question "Can you see the following thing on the image?" is presented to the subjects (see Figure 1, left). After pressing the "space" button, the application continuous with the next screen.
2. In this screen, a small blinking dot in the upper middle is displayed for one second (see Figure 1, middle). The subjects were asked to look at that point in order to let all subjects start viewing the images from the same position. The red dot let all subjects start viewing the image (which is shown next) from the same gaze position. The dot is placed above the actual image that is shown in the third screen.
3. Finally, the image is shown to the subjects (see Figure 1, right). Viewing the image, the subjects had to judge whether the thing shown in the first screen can be seen on the image or not. The decision is made by pressing the "y" (yes) or "n" (no) key.

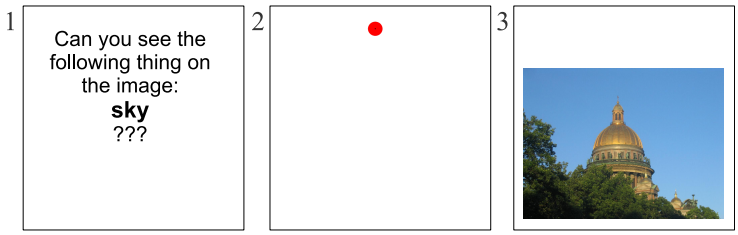

Fig. 1. Steps Conducted for Identifying Image Objects

The first image-tag-pair is used to introduce the application to the subjects and is not used for the analysis. Each subject did evaluate one of the two sets consisting of 51 image-tag-pairs from the data set described above. The subjects

were told that the goal of the experiment is not to measure their efficiency in conducting the experiment task. They could take as much time as they like to make the decision.

Besides recording the raw gaze data, we have also measured the time the subjects took to make a decision per image and the correctness of the answers.The average answer time over all images and users is about 3,003 ms. 5.7% of the given answers of all subjects were incorrect. The proportion of wrong answers is the same for given true and false tags. Subsequently to the experiment, the subjects were asked to provide subjective feedback in a questionnaire. The eye tracker and the experiment situation did not much influence the users' comfort. 85% of the subjects strongly agreed or agreed on the statement that they felt comfortable during the evaluation.

4 Analysis of Gaze Fixations on the Images

The preprocessing of the raw eye-tracking data was 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 extracted fixations are the base for our measure analysis. We have analyzed the gaze paths for images with a true tag and where the subjects gave a correct answer. In cases where the subjects gave incorrect answers, we do not know if the subjects did not took enough time to examine the image, did not understand the given tag, or if they had other problems. 547 gaze paths have been collected during the experiment that fulfill our requirement. 476 (87 %) of these gaze paths have at least one fixation inside or near a correct region. With this data, we are able to investigate the best fixation measure to identify the correct region in the image, i.e., finding the region of the image the tag shown in the experiment refers to. Please note that we do not use the images with the false tags, as the false image-tag-pairs have only been created in order to keep the subjects concentrated during the experiment (see Section 3). Investigating if it is possible to detect from the gaze path whether a subject had looked at a true image-tag-pair or false image-tag-pair is part of future work.

4.1 Calculating the Precision of Tag-to-Region-Assignments

The procedure for calculating the precision of the tag-to-region assignments is illustrated in Figure 2. The single steps performed for this calculation 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 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).
3. The label of the favorite region is compared with the tag (a) that was given to the subject in the experiment. If the label and tag match, the assignment

is true positive (tp) otherwise it is a false positive (fp). We have summed up the total number of correct and incorrect assignments over all images and calculate the precision P for the whole image set using the following formula:

$$P = \frac{tp}{tp + fp} \tag{1}$$

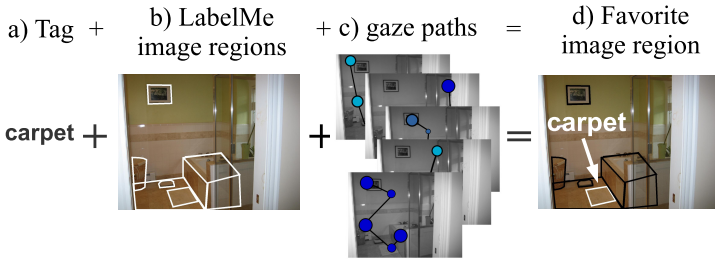

Fig. 2. Overview of Calculating the Tag-to-Region-Assignments

4.2 Considered Fixation Measures

We have selected 13 fixation measures and compared their performance to identify the correct favorite region. The measures including their units are presented below. The way the favorite region is calculated using the measure is summarized in brackets after the measure. It can be, e.g., the minimum of fixation counts on the different 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).

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. The measure (2) secondFixation (min count) ignores the first fixation, because this fixation is influenced by the first visual orientation on an image [13]. We have also used a modification of the secondFixation measure called (3) fixationsAfter [4] (min count) to examine also the fixations on the image after the subjects made their decision, i.e., have pressed the "n" or "y" key. 96% of the gaze paths contain fixations after making the decision by pressing the button on the keyboard. This is due to the inherent reaction time of the experiment setup. The average duration of the recording after making the decision is 834 milliseconds. We have investigated the fixations around the moment of decision with the new measures (4) fixationsBeforeDecision (min count) and (5) fixationsAfterDecision (min count). The last measure includes also fixations at the moment of decision. The (6) fixationDuration (max millisecond) describes the sum of the duration of all fixations on a region r. The Tobii 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 last fixations were taken into consideration in [11]. 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) [6] provided good results for the relevance feedback in image search. Thus, we have also considered it in our experiments. 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.

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

4.3 Extending Object Boundaries and Weighting Small Objects

When comparing the fixation measures, we have investigated two further parameters: The first parameter is an extension of region boundaries to deal with the inaccuracy of eye-tracking data. Based on our prior investigations [13], we use an extension of 13 pixels. The second parameter deals with the fact that larger image regions have the advantage of being more likely fixated than smaller images. To support smaller regions, we investigate a linear weighting function with the highest weighting factor 4 [13]. The weighting depends on the image region size in relation to the total image size. All image regions smaller than 5% of the image size are weighted. The detailed analysis of the region extension and weighting parameters can be found in [13].

5 Results of Finding Objects in Images

Comparing the different fixation measures, we have received the best results for the measure (11) meanVisitDuration with precision $P = 0.54$ (cf. Figure 3). That means, 54% of the image regions selected by the gaze analysis belonged to the tag that was shown to the subjects. Two measures reach the second best value ($P = 0.53$): (4) fixationsBeforeDecision and (8) lastFixationDuration. With $P = 0.50$, the measure (6) fixationDuration provides the third best result. The lowest precision values are 0.21 and 0.26 for (1) firstFixation and (2) secondFixation.

Taking the image region extension and the weighting from Section 4.3 into account, we receive for meanVisitDuration the best precision value $P = 0.67$. The following analysis and computations are based on this measure and parameters. Figure 4 shows some positive and negative examples. As we have investigated, the size or the position of an object in the image does not have in principle an influence of the correctness of the assignments (see [13] for details). However, we have identified some characteristics of the images with incorrect assignments.

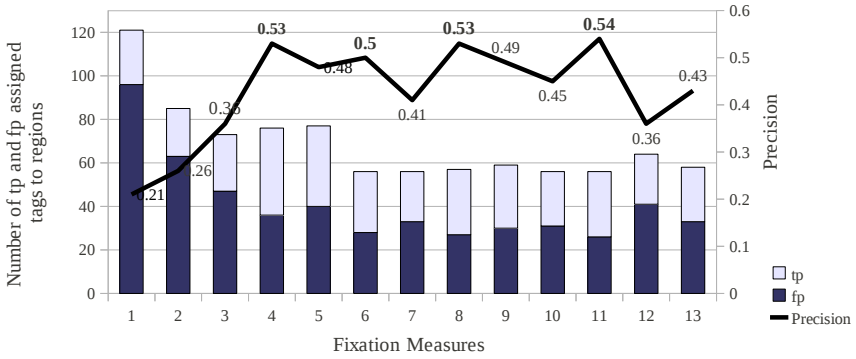

Fig. 3. Precision Values for the Fixation Measures from Section 4.2

First, in some scenes with a small given object the wrongly selected favorite object is also small and located next to the correct object. This problem can be based on the accuracy of the eye-tracker (5 of 19 wrong assignments belong to this category). Second, the object is sometimes located within another object (cf. Figure 4, image 5). In these cases, the outer region is identified as favorite (5 of 19 wrong assignments). Finally, further images show scenes with an object that seems to be very easy to identify. For example large objects like *road* (cf. Figure 4, image 6) or *sky* might be perceived even in the corner of the human eye or based on context knowledge (7 of 19 wrong assignments).

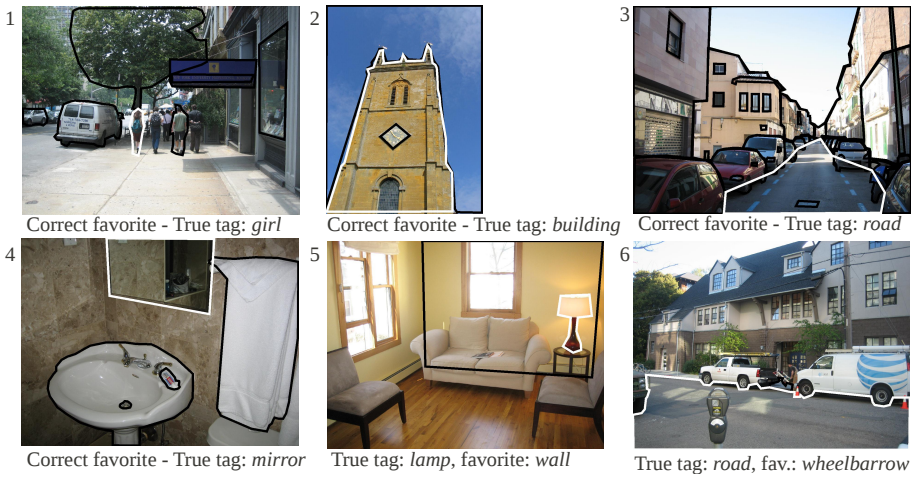

Fig. 4. Correctly (1. - 4.) and incorrectly (5., 6.) identified favorite objects

5.1 Compare with Baselines

We use two baselines that have been applied to evaluate relevance feedback from gaze information in [6] and [5]. We compare the precision P for image-tag-pair assignments calculated from the baseline "naive" (a) and the baseline "random" (b) with the mere measure meanVisitDuration (c) and the meanVisitDuration measure including region extension and weighting (d). The naive baseline makes the assumption that the largest area in an image should be the favorite one. The random baseline randomly chooses one of the labeled regions of the image as favorite. The results in Figure 5 show, the naive approach has a precision of 0.16 and the random baseline of 0.21 compared to the gaze-based approach with a precision of 0.54 and the extended and weighted of 0.67. The identification of assignments based on gaze information or on gaze information including extension and weighting performs better than both baseline approaches. Applying Chi-square tests shows that the gaze assignments are significantly better than the baselines (all with $\alpha < 0.001$).

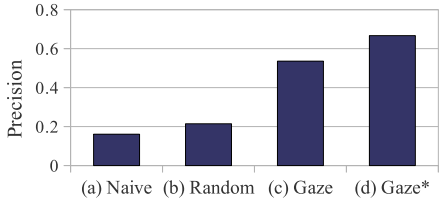

Fig. 5. Precision for two baselines and gaze based analysis

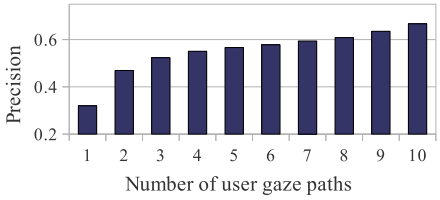

Fig. 6. Effect of aggregation of gaze paths from one up to ten users

5.2 Effect of Aggregation of Gaze Paths on Precision

We have investigated how strong the influence of the aggregation over multiple subjects on the precision. We present precision values for aggregations 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 Figure 6 shows, the influence of the number of users is very high. With the gaze paths of only a single user, we have received an average precision (over all users and all images) of $P = 0.31$. For the aggregated data for all 10 users we got a precision $P = 0.67$. This corresponds to an improvement of 109%. 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%. Between nine users and ten users per group, there is only an improvement of 4%.

The results based on multiple gaze paths are considerably better than the ones calculated from only a few gaze paths. However, the improvement of the

precision gets lower when aggregating more gaze paths. Compared with the two baselines from Section 5.1, the results for single users are still significantly better than the naive or random baseline. The Chi-square test provides for the naive approach $\alpha < 0.001$ and for the random approach $\alpha < 0.002$.

6 Results of Discriminating Objects in Images

As an extension to the experiment described above, we have investigated if it is possible to differentiate objects by analyzing the users' gaze paths given that different tags of the same image are shown to the subjects. For this experiment, we have used the two tag sets assigned to the 51 images as described in Section 3. We use the measure meanVisitDuration, including extension and weighting, to calculate the results. For 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 Figure 7, 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 67% (see Section 5). For two images, the probability of identifying correct assignments for both tags is 44%. 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.

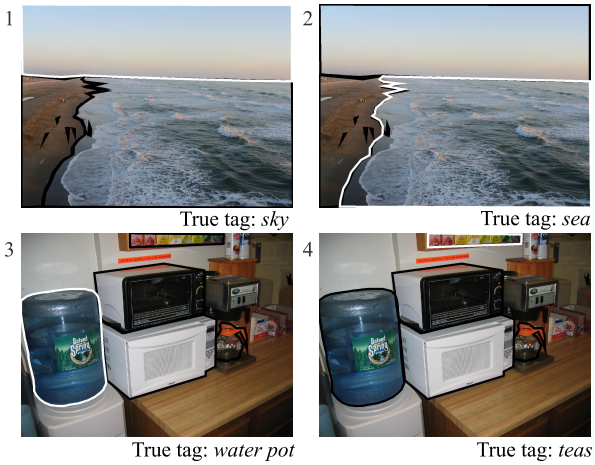

Fig. 7. Example images with two correctly identified regions (white borders)

7 Conclusions

In this paper, we have shown that it is possible to identify image regions by analyzing the gaze paths of users viewing the image with a given tag and given image regions at a precision of 67%. In addition, we have shown that two different regions can be differentiated in the same image with an accuracy of 38%. The results are gained in a controlled experiment with manually segmented images from the LabelMe data set. We have used LabelMe instead of applying automatic segmentation based on low-level features because of the additional error that would have been introduced in the experiment by automatic segmentation. The next step will be to apply the experiment on automatically segmented images. Such automatic segmentation can be improved by using the gaze information [10].

Acknowledgement: The research leading to this paper was partially supported by the EU projects Petamedia (FP7-216444) and SocialSensor (FP7-287975).

References

1. S. Castagnos, N. Jones, and P. Pu. Eye-tracking product recommenders' usage. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 29–36. ACM, 2010.
2. S.N. Hajimirza and E. Izquierdo. Gaze movement inference for implicit image annotation. In *Image Analysis for Multimedia Interactive Services*. IEEE, 2010.
3. Alejandro Jaimes. Using human observer eye movements in automatic image classifiers. *SPIE*, 2001.
4. A. Klami. Inferring task-relevant image regions from gaze data. In *Workshop on Machine Learning for Signal Processing*. IEEE, 2010.
5. A. Klami, C. Saunders, T.E. De Campos, and S. Kaski. Can relevance of images be inferred from eye movements? In *Multimedia information retrieval*. ACM, 2008.
6. L. Kozma, A. Klami, and S. Kaski. GaZIR: gaze-based zooming interface for image retrieval. In *Multimodal interfaces*. ACM, 2009.
7. Subramanian Ramanathan, Harish Katti, Raymond Huang, Tat-Seng Chua, and Mohan Kankanhalli. Automated localization of affective objects and actions in images via caption text-cum-eye gaze analysis. *Multimedia*, 2009.
8. N.C. Rowe. Finding and labeling the subject of a captioned depictive natural photograph. *IEEE Transactions on Knowledge and Data Engineering*, pages 202–207, 2002.
9. B. C Russell, A. Torralba, K. P Murphy, and W. T Freeman. LabelMe: a database and web-based tool for image annotation. *Journal of Computer Vision*, 77(1):157–173, 2008.
10. A. Santella, M. Agrawala, D. DeCarlo, D. Salesin, and M. Cohen. Gaze-based interaction for semi-automatic photo cropping. In *CHI*, page 780. ACM, 2006.
11. S. Shimojo, C. Simion, E. Shimojo, and C. Scheier. Gaze bias both reflects and influences preference. *Nature neuroscience*, 6(12):1317–1322, 2003.
12. Luis von Ahn, Ruoran Liu, and Manuel Blum. Peekaboom: a game for locating objects in images. In *CHI*. ACM, 2006.
13. Tina Walber, Ansgar Scherp, and Steffen Staab. Towards improving the understanding of image semantics by gaze-based tag-to-region assignments. Technical Report 08/2011, Institut WeST, Universität Koblenz-Landau, 2011. `http://www.uni-koblenz.de/~fb4reports/2011/2011_08_Arbeitsberichte.pdf`.