Can you see it? Two Novel Eye-Tracking-Based Measures for Assigning Tags to Image Regions

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

[1]Institute for Web Science and Technology, University of Koblenz-Landau, Germany
http://west.uni-koblenz.de
[2]Research Group on Data and Web Science, University of Mannheim, Germany
http://dws.informatik.uni-mannheim.de
{walber,scherp,staab}@uni-koblenz.de

Abstract. Eye tracking information can be used to assign given tags to image regions in order to describe the depicted scene in more details. We introduce and compare two novel eye-tracking-based measures for conducting such assignments: The segmentation measure uses automatically computed image segments and selects the one segment the user fixates for the longest time. The heat map measure is based on traditional gaze heat maps and sums up the users' fixation durations per pixel. Both measures are applied on gaze data obtained for a set of social media images, which have manually labeled objects as ground truth. We have determined a maximum average precision of 65% at which the segmentation measure points to the correct region in the image. The best coverage of the segments is obtained for the segmentation measure with a F-measure of 35%. Overall, both newly introduced gaze-based measures deliver better results than baseline measures that selects a segment based on the golden ratio of photography or the center position in the image. The eye-tracking-based segmentation measure significantly outperforms the baselines for precision and F-measure.

Keywords: Fixation measures, automatic segmentation, heat maps

1 Introduction

The understanding of image content is still a challenge in automatic image processing. Often, tags are used to manually describe images. Another approach is to analyze the text surrounding an image, e.g., on web pages, to draw conclusions about the depicted scene. A better understanding of the objects depicted in an image can improve the handling of images in many ways, e.g., by allowing similarity search based on regions [8] or by serving as ground truth for computer vision algorithms [11]. It is intuitive for humans to identify objects depicted in an image. The human perception system can compensate perspective distortions, occlusions and can also identify objects with an unusual appearance. These adaptions of the perception system are hard tasks for algorithms and have not yet been solved.

The idea of our work is to benefit from human abilities to perceive visual information in order to obtain a better understanding of depicted scenes. We notice a rapid development of sensor hardware (cameras) in devices like laptops and a decreasing of cost for hardware. Extrapolating this development into the future, eye tracking will be more widely available and can be performed using standard sensors like web cameras [12]. In this work, we investigate two new eye-tracking-based measures with regard to their capability of assigning a given tag to a region in an image such that a depicted object is correctly labeled. For this purpose, we have investigated how efficient measures applied on eye fixations may serve the region labeling task. Fixations are the phases in the gaze trajectories when the eyes are fixating a single location. The first measure is the *eye-tracking-based segmentation measure*. It is based on a standard image segmentation algorithm [2] and selects the image segment as most relevant for the given tag which the user fixates on for the longest time interval. The second measure is the *eye-tracking-based heat map measure*. It is based on a traditional heat map and sums up the duration of the fixations.

We compare the two new eye-tracking-based measures with two baseline measures. The baselines also make use of automatically computed segments, but not of additional information. The eye tracking data for our investigations is taken from a controlled experiment conducted with 30 subjects each viewing 51 social media images with given tags. The experiment is presented in [14]. First, the subjects where shown a specific tag. Subsequently, we have recorded their gaze path while they viewed the image and while they had to decide whether an object referring to that tag was depicted or not. The social media images have as ground truth manually labeled objects. We have used this experimental data to tackle the following core research questions:

- To which extent may the two new eye-tracking-based measures identify the correct position in the image for a given tag (maximum precision)?
- To which extent does the area determined by the two new measures cover the actual object depicted in the image (maximum F-measure)?

We show that the segmentation measure performs better for both questions, although the difference to the heat map measure is not significant. The segmentation measure delivers significantly better results for precision and F-measure than the baseline approaches.

In the subsequent section, we discuss the related work. In Section 3, we describe our two novel eye-tracking-based measures and the baselines. In Section 4, the experiment is described from which we have obtained the eye tracking data. The examination of the best parameters determined on a subset of the images is presented in Section 5 followed by the results obtained from our experiments in Section 6.

2 Related Work

Yarbus [15] has already shown in 1967 that image content strongly influences eye movements. The tendency of humans to fixate faces in images is well known and also the identification of parts of the faces from gaze paths can be performed [4]. Klami [9] investigates which parts of images are relevant for a user in a given task. In his work, relevance is calculated only from the gaze information and it is represented in a Gaussian mixture model, which resembles heat maps. The work reveals that the visual attention depends on the task given to the subject before viewing an image. The work of Ramanathan et al. [10] aims at localizing affective objects and actions in images by using gaze information. Areas that are affecting the users' attention 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 like it is aimed at in this work is not conducted. In a previous work [14], we have investigated the possibilities to assign tags to image regions, where these regions were manually labeled with hand-drawn polygons. Gaze paths of users looking at the images were analyzed by 13 different fixation measures to calculate the assignment. A tag was assigned to a correct image region for 63% of the image-tag-pairs.

Essig [6] takes user-relevance feedback, gained from gaze information, into account to improve the content-based image search. The feedback is calculated on the basis of image regions. He showed that the retrieval results of his approach received significantly higher similarity values than those of the standard approach, which is based only on automatically derived image features. Bartelma [3] investigated the combination of gaze control and image segmentation. He has implemented a system that is controlled by gaze to manually segment images. The gaze is exclusively used as a mouse replacement. The subjects were instructed to outline a given object with their gaze. Santella et al. [13] 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, independent of the objects in the image. Their work shows that the image cropping approach based on gaze information is preferred by the users to fully automatic cropping in 58.4% of the cases.

The related work shows that eye tracking information is exact enough to be used on the level of image regions and that this information can be of value in several use cases. To the best of our knowledge, no work is done on assigning given tags to image regions by using gaze information without a given ground truth segmentation.

3 Identifying Objects in Images

We suggest two methods for assigning tags to image regions, thus identifying objects that correspond to a predefined tag. Both methods, as well as the baseline methods, proceed using the following input:

- An image I is a set of pixels $P(x, y)$, $0 \leq x < width$, $0 \leq y < height$
- A tag t, describing an object depicted in I
- A set of users U that have viewed the images during the experiment

- Set of gaze paths provided by users $u \in U$, to which the tag t was shown and who had to decide whether an object described by t can be seen in the image or not

Gaze paths consist of fixations and saccades. Fixations F are short stops that constitute the phases of the highest visual perception, while saccades are quick movements between the fixations. Every gaze path G_t consists of a set of fixations F, provided by user $u \in U$. Every fixation $f = (x_f, y_f, d)$ is described by a fixated point in the image (x_f, y_f) and a duration d. To measure the human visual attention, the fixations are analyzed by so called fixation measures. From these measures, a value ν is calculated for given regions R of an image I. Example eye tracking measures are the fixationCount, a standard measure which counts the number of fixations on a region and the lastFixationDuration, which sums up the duration of the last fixation on an image region. We have compared 13 fixation measures with respect to their ability to identify a concrete image region for a tag t given to the users [14]. Derived from the results of this work, we use the measure lastFixationDuration, which has delivered the best results.

Subsequently, we present the two novel eye-tracking-based measures and the baseline measures; we also describe the method for evaluating the proposed eye-tracking-based measures.

Eyetracking-based Segmentation Measure: The idea of this approach is to calculate ν for the fixation measure lastFixationDuration for all regions $r \in R$ gained from an automatically segmented algorithm. $\nu(r, u)$ is calculated for every user $u \in U$ viewing the image. The values ν are summed up for every region over all users and the favorite region r_{fav} is determined by the highest value:

$$r_{fav} = \arg\max_{r \in R} \sum\nolimits_{u \in U} \nu(r, u) \tag{1}$$

Eyetracking-based Heat Map Approach: Heat maps are two-dimensional graphical representations of a number of gaze information. They visualize the frequency of fixations for every pixel $P = (x, y)$ in an image. Different colors symbolize how many times or how long a pixel was fixated. The advantage of heat maps is that they can summarize a large quantity of data and are easy to comprehend by humans. Thus, they are often used in usability experiments to visualize users' attention. Different kinds of heat maps can be created based on different measures, e.g., a fixationCount or an absoluteDuration heat map [5]. As the lastFixationDuration was the best measurement for the region identification in our previous work [14], we use this measure as basis for our approach. A radius rd has to be defined for the creation of a heat map. We use a default value of 50 pixels, taken from Tobii Studio [1]. A maximum value of $h_{max} = 100$ is assigned to the pixel fixated by a fixation $f = (x_f, y_f, d)$. Starting from this point, values are added to the pixel in the surrounding of the fixation, based on a linear interpolation between h_{max} and 0. The result is multiplied by the fixation duration d. An example is visualized in Figure 1. For a single fixation, we calculate the heat map values h of all pixels $P = (x, y)$ in the surrounding of the fixation:

$$h(P,f) = \begin{cases} d * (h_{max} - \big(dist(P,f) * \frac{h_{max}}{rd}\big)) & \text{, if } dist(P,f) \leq s \\ 0 & \text{, otherwise} \end{cases} \tag{2}$$

All last fixations f_{last} of all gaze paths provided by the users $u \in U$ are summed up in the final heat map H:

$$H(P) = \sum_{u \in U} h(P, f_{last}) \tag{3}$$

From all heat map values H, the highest value $max(H)$ is determined. To obtain the favorite region from the heat map, we set a threshold $0 < t \leq 100\%$. For example a t = 5% means that only heat map values are considered that belong to the highest 5% of all values. This procedure can be described by an analogy of a flooded region with valleys and elevations. The threshold t symbolizes the water level. With a level of $t = 5\%$, only the highest 5% of the landscape are visible above the water level or in our case all pixels with $H(P) > 0.95 * max(H)$ are determined as possible favorite regions. The biggest area of connected pixels is selected as favorite region r_{fav}. An illustration of this thresholding is presented in Figure 7.

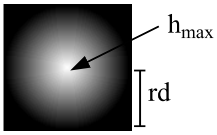


Fig. 1. Heat Map Values

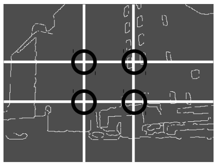
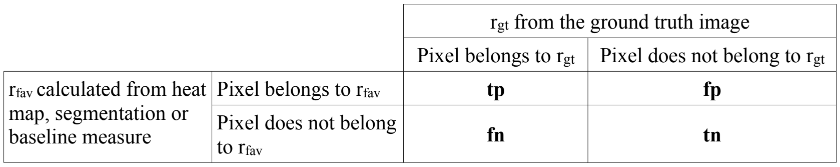


Fig. 2. Golden Sections

Baseline Approach: Initially, we had investigated a random baseline approach as used in our previous work [14], which is randomly selecting one segment of an automatically segmented image as favorite region. As the results of this baseline were very weak, we decided to improve the baseline approach by taking into account the position of the segments in the image in two different ways. As the pictures used in our analysis are taken by humans, we can suppose an inherent photographic bias. The golden ratio rule is a very basic rule in photography [7]. Taking images based on this rule can improve the aesthetics of a photograph and it is often met instinctively to achieve aesthetically appealing pictures. According to the golden ratio, width and height of an image are divided into two parts in the ratio 1 to 1.618. This results into four intersections, at which important objects in the images are often placed. In Figure 2, the golden sections are highlighted by black circles. Another typical bias is to position the important object in the center of the image. For each picture, the golden ratio and the center baselines are calculated. The segment placed at the golden section respectively the center point is selected as favorite region r_{fav}.

Evaluation Method: After obtaining favorite regions with one of the two new measures or the baseline measures, the results have to be evaluated by means of comparing them with ground truth object labels. In information retrieval, precision, recall, and F-measure are standard approaches to measure the relevance of search results. We use these measures to evaluate the covering of the ground truth object region r_{gt} by the favorite region r_{fav} at pixel level. The algorithm runs through the image and classifies every pixel as tp (true positive), fp (false positive), fn (false negative), and tn (true negative) as described in Figure 3.



		r_{gt} from the ground truth image	
		Pixel belongs to r_{gt}	Pixel does not belong to r_{gt}
r_{fav} calculated from heat map, segmentation or baseline measure	Pixel belongs to r_{fav}	**tp**	**fp**
	Pixel does not belong to r_{fav}	**fn**	**tn**

Fig. 3. Definition of tp, fp, fn, and tn

4 Experimental Data

This work is based on the data gained in an experiment described in [14]. More details about the experiment setup, the subjects, and the used data set, can be found there. The experiment data was gained in a controlled experiment, performed with 30 subjects organized in three groups. The experiment was performed on a screen with a resolution of 1680x1050 pixels while the subjects' eye movements were recorded with a Tobii X60 eye-tracker at a data rate of 60Hz and an accuracy of 0.5 degree.

The gaze data of the first two groups are used for parameter fitting, while the data from third group is used to verify the results of our measures. The experiment sequence consisted of three steps conducted for each image: First, a tag was presented to the subjects with the experiment task "Can you see the following thing in the image?". After pressing a button, users had to fixate a small blinking dot in the upper middle for one second. In a third step, the image was shown to the subjects. 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 by pressing the "y" (yes) or "n" (no) key. We used images from LabelMe[1] with 182.657 user contributed images (download August 2010) to create three sets of images I, one for each group of subjects. The LabelMe community has manually created image regions by drawing polygons over the images and tagging them. The labels were used as tags t and the polygons as ground truth image segmentation. For every image selected, we randomly chose

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

a “true” (describing an object in the image) or “false” tag. About 50% of the given tags corresponded to an object displayed in the image (“true” tag), while the other half did not. In our analysis of the gaze data, we consider only data belonging to images with a given “true” tag and a correct answer by the user.

5 Determining Best Parameter Settings

The data set is split into two subsets: a training set for the parameter fitting (56 images-tag-pairs each viewed by 10 users) and a test set for the evaluation of the approaches (29 images-tag-pairs each viewed by 10 users). In this section, we investigate different parameters for our approaches and identify the parameters leading to the best results. The outcome is applied to the test data set and used to compare the different measures from Section 6.

Eye-Tracking-Based Segmentation Measure: The segmentation is performed by using the *bPb*-owt-ucm algorithm [2]. Different hierarchy levels for $k = 0 \ldots 1$ are calculated, each representing a different level of detail. An example is presented in Figures 4, showing the segmentation results for different k-values. The first segmentation level $k = 0$ delivers 1831 segments, the segmentation with $k = 0.4$ the least number of segments, namely six.

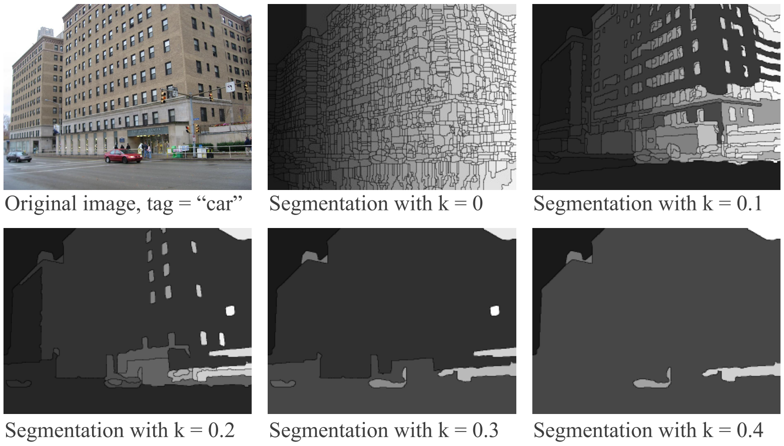
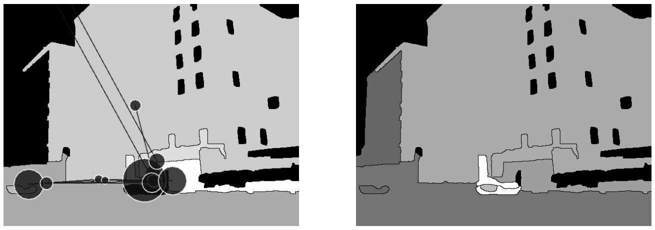

Original image, tag = “car”
Segmentation with k = 0
Segmentation with k = 0.1
Segmentation with k = 0.2
Segmentation with k = 0.3
Segmentation with k = 0.4

Fig. 4. Segmentations with Different Parameters k

Applying eye-tracking-based segmentation measures to those segmentations provides the favorite region r_{fav} from all segments, as described in Section 3. In Figure 5(a), an example for a gaze path of a single user is shown. The fixations are displayed as circles, the fixation duration is presented by the diameter of the circles. The saccades are depicted as lines between the fixations. The brightness of the image segments encodes the fixation measure values $\nu(R)$. The order of the viewed regions is encoded from the favorite region in white to the segments with few fixations in dark gray. The black segments have not been fixated at all. Figure 5(b) shows the results for one image aggregating the gaze paths of all users. To determine the best hierarchy level k, we have compared the results for different levels $k = 0 \dots 1$ by calculating precision, recall, and F-measure. For $k > 0.4$, the number of segments is too low to obtain a reasonable favorite region r_{fav}. Basically the result is a very large segment, covering almost the entire image plus a few very small segments.

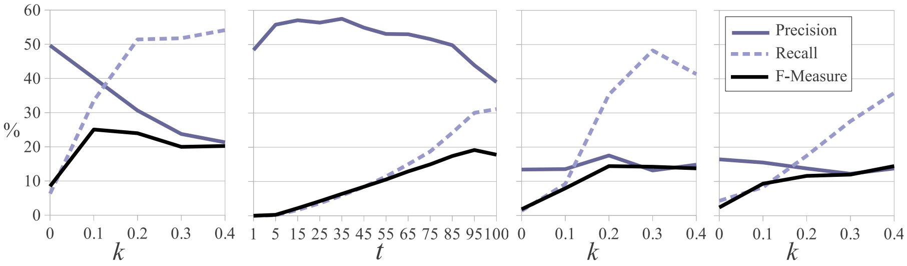


(a) Gaze path with intersected regions for $k = 0.2$

(b) Favorite regions over all users for $k = 0.2$

Fig. 5. Identification of r_{fav} for one user (a) and aggregated for 10 users (b)

The results for all investigated k values are depicted in Figure 6(a). The best precision with 50% is obtained for the smallest sizes of segments for $k = 0$. The best recall with 54% for $k = 0.4$. The maximum F-measure of 25% is reached with $k = 0.1$. It is calculated from a precision of 4% and a recall of 34%. One can see that the F-measure is relatively stable between the $k = 0.1$ and $k = 0.4$, because of the rising recall and the falling precision values.



(a) Segmentation measure with k

(b) Heat map measure with parameter t

(c) BL Golden Section with k

(d) BL Center with k

Fig. 6. Precision, recall, and F-measure for the two gaze-based and the two baseline measures (BL)

Eye-Tracking-Based Heat Map Measure: For the heat map measure, described in Section 3, we have investigated different thresholds $t = 1 \dots 100\%$.

Some examples are depicted in Figure 7. It shows the original image, next to a classical heat map visualization of gaze information from all 10 users. The next four images show different potential favorite areas after applying the threshold t to the heat map. If several areas appear, the biggest one (i.e., the one with the most pixels) is supposed to be the favorite region r_{fav}.

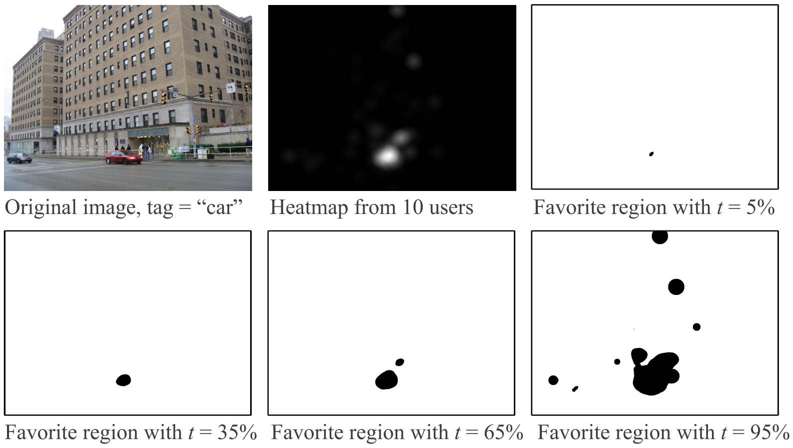

Original image, tag = "car"
Heatmap from 10 users
Favorite region with t = 5%
Favorite region with t = 35%
Favorite region with t = 65%
Favorite region with t = 95%

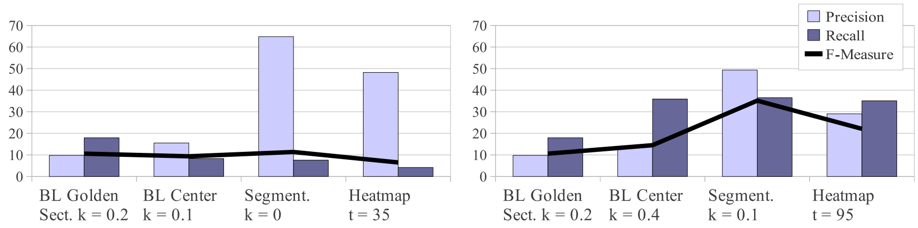
Fig. 7. Visualization of the Heat Map Measure

Precision and F-measure are calculated, comparing the computed favorite region r_{fav} with the ground truth object region r_{gt}. An overview of the results is presented in Figure 6(d). The highest precision value is obtained for $t = 35\%$ with 57%. Even with constantly high precision values of more than 44% the F-measure values cannot get very high because of the poor recall results (maximum: 31%). The best F-measure result is 19% with $t = 95\%$.

Baseline Measures: For the baseline measures, we also compute the segmentation using the bPb-owt-ucm algorithm [2]. For both baselines, we investigate the best parameters $k = 0 \ldots 0.4$. For the golden section baseline, we obtain the highest precision value over all images with 18% for $k = 0.2$ and the highest the F-Measure with 14% for $k = 0.2$. The best results for the center baseline are a precision of 16% for $k = 0.1$ and a F-Measure of 13 % for $k = 0.4$.

6 Evaluation Results

The best performing parameters from the training data set for each of the measures are applied to the test data set. For each measure, we obtain values for precision and F-measure for each image. For comparing the different measures, we have conducted a Kolmogorov-Smirnov test to determine if the precision values and F-measure values exhibit a normal distribution. As most of our computed values do not exhibit a normal distribution (details of the test results omitted for brevity), we have conducted a Friedman test to investigate for a statistical significance in the difference of the obtained precision values and F-measure values. We found that the differences between the four assignment measures (segmentation, heat map, and two baselines) are significant ($\alpha < .05$) for precision ($\chi^2(3) = 32.668$, $p = .000$) and F-measure ($\chi^2(3) = 15.891$, $p = .001$). Thus, post-hoc analyses with pairwise Wilcoxon Tests are conducted with a Bonferroni correction for the significance level (now: $\alpha < .017$). The values used in the pairwise Wilcoxon Tests are presented in Figure 8. We obtain the best precision with 65% for the segmentation measure and the second best with 48% for the heat map measure. These results significantly outperform the two baselines with $Z = -4,059, p = .000$ for the segmentation measure compared to the golden section baseline, respectively $Z = -4,090, p = .000$ for the center baseline. The results for the heatmap measure are $Z = -3,438, p = .001$ and $Z = -3,286, p = .001$, respectively. There is a weakly significant difference between the two eye-tracking-based measures ($Z = -1.905, p = .057$). For 12 of 29 images, r_{fav} lies completely inside r_{gt}. For 20 images at least 1% of r_{fav} intersects the ground truth object region r_{gt}. The highest F-measure is obtained again by the segmentation measure with 35%. The result for the heat map measure is 22% and for the baselines 11% (golden section) and 14% (center). A significant difference is recognized between the segmentation measure and the baselines with $Z = -2,943, p = .003$ for both baseline. The other results do not differ significantly (segmentation - heat map: $Z = -.934, p = .350$, heat map - golden section: $Z = -2,345, p = .019$, heat map - center: $Z = -2,186, p = .029$).



(a) Best precision

(a) Best F-measure

Fig. 8. Comparison of the two gaze-based measures and the baseline

7 Conclusion

For 63% of the images, we were able to identify the correct image region, described by a given tag. The assignment of tags to regions becomes much harder without the given, manually created regions like the automatically computed segments considered in this work. The reason lies in the inaccuracies involved with the automatic image segmentation. However the results obtained from the eyetracking-based measures are still very good, with an average precision of 65%

over all images for the segmentation-based measure and 48% for the heat-map-based measure. The eyetracking data in this work was gained in a controlled experiment where users had to identify regions for predefined, given tags (see Section 4). To relax this constrain, we will conduct in a next step an experiment with users tagging images using an application in the style of a real online image annotation tool like Flickr. We have introduced a segmentation measure and heat map measure that both use gaze information as source of information. The results show that the new measures perform better in the assignment of tags to image regions than a baseline approach without gaze information. The segmentation measure performs best for both evaluations: precision and F-measure. The segmentation measure significantly outperforms the baseline. The segmentation measure can easily be adapted to different needs by modifying the parameter k to maximize the precision or F-measure.

References

1. Tobii studio 2.x - user manual, 2010. http://www.tobii.com.
2. P. Arbeláez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *IEEE TPAMI*, 33(5):898–916, May 2011.
3. J.M. Bartelma. *Flycatcher: Fusion of gaze with hierarchical image segmentation for robust object detection.* PhD thesis, Massachusetts Institute of Technology, 2004.
4. W.V. Belle, B. Laeng, T. Brennen, et al. Anchoring gaze when categorizing faces' sex: Evidence from eye-tracking data. *Vision research*, 49(23):2870–2880, 2009.
5. A. Bojko. Informative or misleading? heatmaps deconstructed. *Human-Computer Interaction. New Trends*, pages 30–39, 2009.
6. K. Essig. *Vision-Based Image Retrieval (VBIR)-A New Approach for Natural and Intuitive Image Retrieval.* PhD thesis, 2008.
7. M. Freeman. *The Photographer's Eye: Composition and Design for Better Digital Photos.* Focal Press, 2007.
8. D.H. Kim and S.H. Yu. A new region filtering and region weighting approach to relevance feedback in content-based image retrieval. *Journal of Systems and Software*, 81(9):1525–1538, 2008.
9. A. Klami. Inferring task-relevant image regions from gaze data. In *Workshop on Machine Learning for Signal Processing. IEEE*, 2010.
10. S. Ramanathan, H. Katti, R. Huang, T. Chua, and M. Kankanhalli. Automated localization of affective objects and actions in images via caption text-cum-eye gaze analysis. In *Multimedia*, New York, New York, USA, 2009. ACM.
11. B. C Russell, A. Torralba, K. P Murphy, and W. T Freeman. LabelMe: a database and web-based tool for image annotation. *J. of Comp. Vision*, 77(1):157–173, 2008.
12. J. San Agustin, H. Skovsgaard, J.P. Hansen, and D.W. Hansen. Low-cost gaze interaction: ready to deliver the promises. In *CHI*, pages 4453–4458. ACM, 2009.
13. 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.
14. T. Walber, A. Scherp, and S. Staab. Identifying objects in images from analyzing the users' gaze movements for provided tags. *Advances in Multimedia Modeling*, pages 138–148, 2012.
15. A.L. Yarbus. *Eye movements and vision.* Plenum, 1967.