

©ACM, 2009. This is the author's version of the work. It is posted here by permission of ACM for your personal use. Not for redistribution. The definitive version was published in ICMI-MLMI'09: Proceedings of the 11th International Conference on Multimodal Interfaces and the 6th Workshop on Machine Learning for Multimodal Interfaces, <http://doi.acm.org/10.1145/1647314.1647379>.

GaZIR: Gaze-based Zooming Interface for Image Retrieval

László Kozma
laszlo.kozma@tkk.fi

Arto Klami
arto.klami@tkk.fi

Samuel Kaski
samuel.kaski@tkk.fi

Helsinki Institute for Information Technology HIIT,
Department of Information and Computer Science,
Helsinki University of Technology
Finland

ABSTRACT

We introduce GaZIR, a gaze-based interface for browsing and searching for images. The system computes on-line predictions of relevance of images based on implicit feedback, and when the user zooms in, the images predicted to be the most relevant are brought out. The key novelty is that the relevance feedback is inferred from implicit cues obtained in real-time from the gaze pattern, using an estimator learned during a separate training phase. The natural zooming interface can be connected to any content-based information retrieval engine operating on user feedback. We show with experiments on one engine that there is sufficient amount of information in the gaze patterns to make the estimated relevance feedback a viable choice to complement or even replace explicit feedback by pointing-and-clicking.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—*Relevance feedback, Search process*; H.5.2 [Information Interfaces and Representation]: User interfaces—*Input devices and strategies (e.g., mouse, touchscreen)*

General Terms

Algorithms, Experimentation, Performance

Keywords

Gaze-based interface, image retrieval, implicit feedback, zooming interface

1. INTRODUCTION

In recent years image retrieval techniques operating on meta-data, such as textual annotations or user-specified tags, have become the industry standard for retrieval from large image collections. They work well with sufficiently high-quality meta-data, but the need for more content-based approaches operating on low-level features extracted from the

image content is still apparent. Content-based techniques are useful for refining results of keyword searches and, moreover, the available meta-data may not be sufficiently rich for all queries.

In content-based image retrieval (CBIR) there has been a lot of research on the retrieval accuracy, developing better feature descriptions, improving the actual retrieval engines, and refining evaluation metrics, resulting in search engines [3]. To focus the search, the engines typically collect explicit feedback from the user, about which of the shown images are relevant.

We study whether it would be possible to make the interface between the user and the search engine more fluent and natural, by collecting the feedback implicitly from what the user would do in any case. We will separate explicit control and implicit feedback, and make the former intuitive to exercise and the latter as informative as possible. In brief, the user will explicitly request for more (better) images by zooming in the interface, and the implicit feedback is inferred from gaze tracking data while the user looks at the images. This paper is a feasibility study on whether it is possible to construct such an interface, and whether it works in practice with an existing CBIR engine.

The main research question is to what extent the explicit relevance feedback can be augmented or eventually replaced by implicit relevance feedback inferred from the actions the user would perform in any case, the idea being that removing a separate relevance feedback phase will make the interface more natural and faster to use [7]. As a practical information source, we use cues obtained by measuring the eye movements of the user, following the success of earlier attempts in inferring relevance from eye movements in text retrieval [2, 4, 11]. As far as we know, there have so far only been preliminary studies related to use of implicit gaze information in image retrieval [8, 10]. Oyekoya et al. present a simple retrieval system that infers relevance from straightforward viewing time [10], whereas Klami et al. introduce a more complex relevance predictor but only measure isolated prediction performance in a simplified artificial setup [8]. We combine the two approaches, developing an even more sophisticated relevance predictor and integrating it with a real retrieval engine. Furthermore, we design the user interface specifically for gaze-based interaction.

Eye movements as a source of implicit relevance feedback have three major advantages. First, the user will by definition need to look at the images in order to make the decision on relevance, and hence if relevance feedback can be inferred from eye movements it will be completely effortless for the

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

ICMI-MLMI'09, November 2–4, 2009, Cambridge, MA, USA.

Copyright 2009 ACM 978-1-60558-772-1/09/11 ...\$10.00.

user. The user just “looks at the images” as he normally would. Secondly, the rich implicit feedback from eye movements may help in the extremely hard problem of solving “I will know it when I see it”-type of search tasks, where the goal is ill-defined at best. Such tasks cannot be solved even with meta-data if the user is not able to formulate explicit queries. The third main advantage of using gaze tracking is that with suitable hardware it is usable in mobile settings when the hands cannot be used, and for users with motor disabilities [1, 16]. While commercial gaze trackers are not yet wide-spread, recent developments [14] suggest that low-cost, robust eye tracking will be possible in the near future also in standard desktop and mobile devices.

Gaze is used to explicitly guide the interface in Dasher, a system for gaze-based text entry [16] which has been one source of inspiration for our work. In Dasher a language model will offer choices for the next letters to be typed, with size of the letters on the display being proportional to their predicted likelihood of being selected. Then the user will look at the next letter in a zooming interface where new letters will appear with speed controlled by gaze as well. In our case the letters correspond to images, of which the ones predicted to be most relevant are shown, and the user scrolls to get more images. Most of the other features of the systems are different, however; most notably the explicit vs implicit feedback by the gaze. In explicit gaze-driven setups, care has to be exercised to avoid the “midas touch” effect, that it is tiring to use the eyes explicitly as control devices for long because everything you look at will be selected [6]. Implicit feedback should not suffer from the same problem—the intent is not that the user controls the system with eyes, but instead that information is extracted from natural eye movements.

Several techniques have been proposed for visualization and navigation of large image collections, including methods like zooming and other distortions for displaying the contents, and tree- or cone-like structures for organizing the image collection. A comprehensive review of visual interfaces can be found in [17]. Our interface borrows elements from this body of research, the main goal and novelty being in facilitating the interaction with gaze. The remaining visualization decisions were made to create a simple and intuitive interface.

In the remainder of the paper, we first describe the interface and how it interacts with the gaze tracker and the retrieval engine. Then we explain how the relevance of images is predicted from the gaze tracking measurements, and demonstrate with user experiments that the accuracy of the relevance predictions is relatively high. Finally, we perform preliminary experiments on actual image retrieval accuracy using the learned relevance predictor. The approach is shown to have promising performance with high accuracy in certain kinds of search tasks. Work still remains, however; in particular, the performance is not high for all users and search tasks.

2. GAZIR

2.1 Interface

The browsing interface is designed to elicit and collect maximal amount of information from gaze while still being a natural interface for browsing the image collection. Figure 1 illustrates the interface, showing three concentric rings

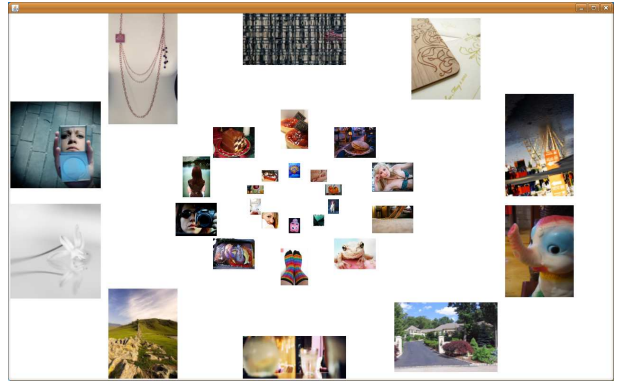


Figure 1: Screenshot of the GaZIR interface. Relevance feedback gathered from outer rings influences the images retrieved for the inner rings, and the user can zoom in to reveal more rings.

of images. The outermost ring contains the first ten images shown to the user, the second ring shows images retrieved given the relevance feedback collected from the outermost ring, and the innermost ring takes into account feedback from the two previous rings. The user can zoom the interface inwards and outwards. When zooming inwards the system retrieves another set of images, using all the previous images and their estimated relevancies as feedback, and eventually the older rings will disappear from the display. They can, however, be recalled by zooming out, and the retrieval process can be restarted from any stage by erasing the rings inside the current main ring.

The concentric rings of images were chosen instead of the standard grid-based thumbnail display of most image retrieval interfaces, in order to avoid imposing gaze trajectories based on the structure of the display instead of the content. On a standard grid the users are likely to go through the images in a row-by-row manner, considerably lowering the amount of relevance information the eye movements contain. Completely random placement of images would break this pattern optimally, but a user is likely to find such an interface unpleasant to use. A circle of images provides a compromise between these two goals. It does not lead to scanning patterns as strongly fixed as a grid would, allowing image content to play bigger role in determining where to look, yet it is sufficiently close to standard user interfaces to feel intuitive.

For the purpose of learning the relevance predictor and studying the interface, we perform the experiments in this paper with two simplifications. First, the user is only expected to zoom inwards and not to reset the retrieval process at any stage. Second, the retrieval engine is set to operate in a sequential manner: A new set of images is fetched only when the user zooms in and they are not updated afterwards. An alternative would be to continuously update the set of images on inner rings when the relevance estimates on the outer rings change. These simplifications were made so that we could collect reliable ground truth for learning the relevance predictor. Finally, in the experiments we used mouse wheel for zooming in and out to make the gaze based interaction completely implicit. The interface can alternatively be zoomed with explicit eye control (looking at the center zooms in and looking at the borders zooms out).

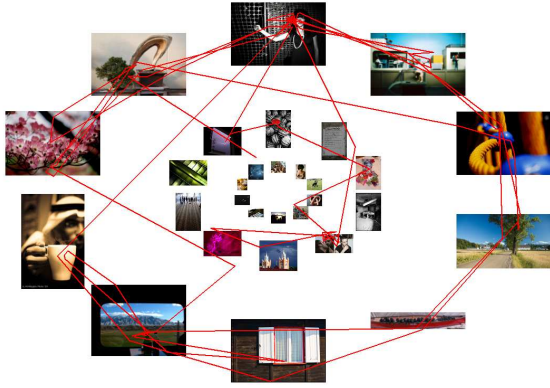


Figure 2: A sample gaze pattern of a user interacting with GaZIR. The line segments indicate saccades, and the joints of two adjoining saccades correspond to the fixation locations. Here the user is primarily looking at images on the outermost ring, only occasionally visiting the inner rings.

2.2 Eye movements

We measured the eye movements with Tobii 1750 eye tracker with 50Hz sample rate. The tracker has a set of infra-red lights and an infra-red stereo camera attached to a standard flat-screen monitor, and the tracking is based on detection of pupil centers and measurement of corneal reflection. See Figure 2 for an example gaze trajectory.

Raw eye measurements are preprocessed by first extracting fixations and saccades with the algorithm described in [13], judging a set of consecutive raw measurements to be a single fixation if they occur within a dispersion of 30 pixels, which at normal viewing distance is equivalent to roughly 0.6 visual degrees (17" screen with resolution of 1280*1024). A fixation is defined to be a period of at least 120 milliseconds of looking at a single location on screen, and processing of the visual stimulus takes place primarily during fixations [12]. Hence we adopt the common approach of basing our gaze trajectory representation in fixations.

The eye movements are measured for predicting which of the viewed images were relevant for the search. To achieve this, we convert the eye movements into a 17-dimensional feature representation for each image and learn a classifier from these representations to binary relevance labels available in a training phase. The feature representations characterize primarily eye movements between the different images, whereas the information within the images is summarized through averages over all fixations landing on the image. This is because the images are small enough to fall within the parafoveal vision and hence the viewer can typically extract sufficient information on the content by fixating anywhere within the image. Hence, movements within the image are not expected to provide much information in this specific case, even though in general the task influences the scan pattern within the images (see e.g. [15]).

Table 1 lists the features used in the experiments. In brief, the features characterize aspects like how long the image was viewed, how often it was viewed, and how soon after the onset it was viewed for the first time. Some of the features also take into account the overall pattern over the different

circles of the interface. Each raw feature was further pre-processed by the z-transform, by removing the mean and dividing the values by the standard deviation.

2.3 Relevance prediction

The system needs relevance feedback to be able to choose the set of images displayed on the next circle when zooming in. It is possible to use the interface by explicitly indicating the relevant images, but the main focus of this paper is in investigating whether the explicit feedback can be replaced by implicit feedback inferred from eye movements. The relevance is predicted in real time for all of the images shown within the main circle, using a classifier operating on the features extracted from the gaze trajectory (see Table 1).

To predict the relevancies we use classical logistic regression. If we denote the feature vector of the i th image by \mathbf{x}_i , then the model estimates the probability of that image being relevant as

$$p(r_i|\mathbf{x}_i) = \frac{1}{1 + \exp(-\mathbf{w}^T \mathbf{x}_i - \alpha)}.$$

Here \mathbf{w} is a projection vector that weights the different features, and α is a bias term. Both are learned to maximize the likelihood of the true relevancies in labeled training data. The final relevance predictions are made by thresholding the probabilities with a pre-specified threshold t , so that all images with relevance probability above the threshold are deemed relevant. In absence of further information the threshold is set to $t = 0.5$, but depending on whether the retrieval engine is more sensitive to false positives or negatives the threshold can be tuned higher or lower.

We chose logistic regression as the predictor for three main reasons. First, it is very light-weight and hence enables computing the predictions in real-time. It would also be efficient enough for on-line adaptation. Second, the weight vector \mathbf{w} reveals information about importance of the different features, enabling us to characterize what kinds of eye movements are predictive of relevance. Finally, it provides probabilities of relevance as output, instead of just discrete predictions of relevant/non-relevant, and could hence be applied in retrieval engines capable of utilizing non-binary feedback.

2.4 Retrieval engine

The interface can interact with any image retrieval engine that operates through relevance feedback. Our current implementation is built on top of the PicSOM [9] engine that uses self-organizing maps for retrieval and MPEG-7 features for describing the content. As PicSOM utilizes binary relevance feedback, taking into account both positive and negative feedback, we restrict our interface to also send only binary feedback. It should, however, be noted that the relevance predictor could directly provide also probabilities of relevance, and could be trivially extended to multiple levels of relevance (e.g., relevant, not-relevant, no decision).

3. EXPERIMENTS

3.1 Relevance prediction

3.1.1 Data collection

We use the MIRFLICKR-25000 [5] database for evaluating the system. To learn the predictor we need training data with known relevance judgments. The data needs to

| Nr. | Name | Description | Type |
|-----|-----------------|---|----------------|
| 1 | IsRandom | Whether the images were shown in the first stage (before any user feedback) | binary |
| 2 | FirstVisit | Time passed between image onset and first visit | continuous |
| 3 | MeanLength | Mean length of fixations | continuous |
| 4 | FixTimeSpread | Standard deviation of fixation occurrence times | continuous |
| 5 | SumLength | Total length of fixations | continuous |
| 6 | MaxContView1 | Maximum continuous viewing time without viewing other images | continuous |
| 7 | MeanContView | Mean length of continuous viewing sessions of image | continuous |
| 8 | MaxContView2 | Maximum continuous viewing time without fixating outside the image | continuous |
| 9 | RatioTotal | Proportion of total viewing time over total viewing times of all other images | continuous |
| 10 | RatioRing | Proportion of total viewing time over total viewing times in the same ring | continuous |
| 11 | MeanSaccLength | Mean Length of saccade before fixation | continuous |
| 12 | MeanPrevImage | Proportion of times when previous fixation also over this image | continuous |
| 13 | MeanPrevEmpty | Proportion of times when previous fixation over empty space | continuous |
| 14 | MeanPrevRing | Proportion of times when previous fixation over the same ring | continuous |
| 15 | FirstVisitIndex | How many images viewed on this ring before the first fixation on image | discrete (0-9) |
| 16 | RevisitCount | How many times image revisited in total | discrete |
| 17 | PrevDist | Average distance from previously viewed image on the same ring | discrete (0-5) |

Table 1: The eye movement features used for predicting the relevance. The distance for PrevDist is measured along the ring, the distance between neighbors being 1.

be collected in a realistic use scenario, to match the kind of eye movements we would expect in real use, yet it needs to be collected before we have a model for making the predictions. The easiest way to do this is to use a stand-in estimator that provides relevance feedback sufficiently close to what the users are expected to think.

In practice, we set up a specific given search task that matches one of the existing high-quality category labels in the MIRFLICKR collection, and asked the users to search for images matching the category description. The true category labels were used as feedback while collecting the data, and after the experiment the users were taken back to the beginning and asked to indicate which of the seen images they had considered relevant by clicking them. This gave us training data with actual user-specific relevancies, while using a substitute relevance feedback during data collection to ensure that the sets of images shown to the user are close to what we should expect in real use.

We collected training data from 6 different users, each performing 6 search tasks. Within each search task the user saw on average around 120 images, so in total we have eye movement measurements over 4300 user-task-image instances. The tasks used in the experiments were chosen randomly from 8 potential classes.

3.1.2 Accuracy

To evaluate the accuracy of relevance prediction, we carried out a number of experiments where a subset of the training data was used for learning the predictor, and the accuracy was measured on the remaining left-out data. The accuracy is measured with standard information retrieval measures, namely area under the ROC-curve (AUC) and mean average precision (MAP).

First we trained user-specific prediction models, using a cross-validation procedure to evaluate the performance. We learned a separate predictor for a subset of 4 search tasks (out of the total 6), and computed the accuracy measures as averages over the corresponding left-out sets. Four tasks were chosen for training since they already provided fairly large amount of data while the training phase still was quick enough for practical user-adaptation in real-world use.

Figure 3 shows the distribution of AUC and MAP scores for each user separately, using 50 random split-ups of the tasks into the training and test sets. The prediction accuracy is above random for all users, indicating that the eye movements provide information on the image relevance. Somewhat surprisingly, the accuracies of different users show substantial differences. This is partly due to the fact that different users had different search tasks, but it may also indicate systematic differences in using the interface. Further investigation is required to find out why exactly the performance is poor for users 3 and 4.

Even though the user-specific accuracies vary a lot, we tried also a user-independent prediction model. If a predictor trained on other users is sufficiently accurate, the interface can be used without requiring a separate user-specific training phase. To evaluate this we split all the available data into training and validation parts across the users, and again computed the results as averages over a number of random splits. Precision-recall curves of one run are used for illustrating the performance, and the accuracy is compared to two baseline approaches (Figure 4). The worse baseline corresponds to ordering the images randomly, whereas the other baseline is the simplest possible approach that utilizes the gaze information: Images with at least one fixation are predicted to be relevant, whereas the rest of the images are not relevant. The purpose of the first baseline is to demonstrate that gaze contains useful information, whereas the latter baseline is included for showing the advantage of the more advanced relevance predictor. The prediction model outperforms the baselines according to both performance measures.

3.1.3 What is informative in gaze?

We first discuss the link between the gaze features and predicted relevance, and then briefly illustrate the typical mistakes of the predictor.

Figure 5 shows the values of the feature weights in the logistic regression model. These were obtained using a subset of two-thirds of all the collected data as a training set. To assess the stability of the weights, the training was done 50 times, each time with a randomly chosen subset of all

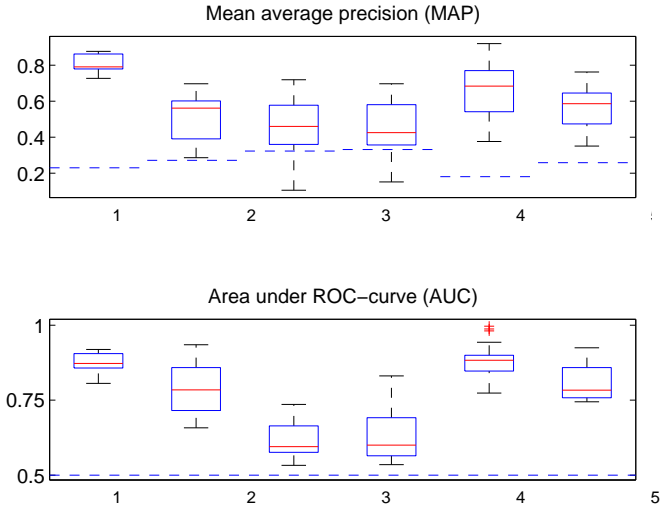


Figure 3: Comparison of retrieval accuracies (MAP and AUC scores) of user-specific relevance prediction models. The box-plots show the distribution of values over different splits to training and validation data. For all users, numbered from one to six on the horizontal axis, the accuracy is above the random baselines shown with dashed lines, but the accuracy depends considerably on the user. For MAP the baselines are user-specific due to different proportions of relevant images.

the data. Most features tend to have a consistently positive or negative effect on the predicted relevance. The first feature has a strong negative weight, indicating that fewer images should be predicted relevant when starting the retrieval process – a random collection of images should show less relevant images on the average.

Example features having strong positive effect are mean length of fixations (3), total viewing time relative to other images (9,10), and the average distance from the previous image (17). All these are fairly intuitive, indicating that the image is more likely to be relevant if the user is looking at it for a long time and breaks the generic viewing pattern to visit the image. At the same time, some observations are less intuitive; for example, the number of images looked at before fixating on the image (15) has a strong positive weight, contradictory to the intuitive expectation of more relevant images visited early on.

Figure 6 shows two false positives (first row) and two false negatives (second row). It can be seen that the first of the “false positive” images in fact contains animals, but this was not noticed by the user when marking the relevant images. The second image is indeed falsely detected as relevant, however, similar images of underwater scenes often show animals. Often the false positives are images that are fairly similar to the relevant ones in terms of content, while missing some detail that would make them relevant. For most retrieval engines these kinds of false positives are not a problem, since the engine would not be able to capture the semantic component that made some of the images relevant

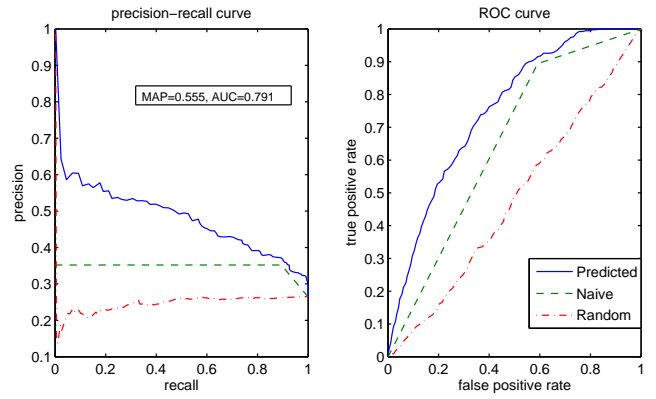


Figure 4: Precision-recall and ROC curves for user-independent relevance prediction model. The predictions (solid line) are clearly above the baseline of random ranking (dash-dotted line), showing that relevance of images can be predicted from eye movements. The retrieval accuracy is also above the baseline provided by a naive model making a binary relevance judgement based on whether the image was viewed or not (dashed line), demonstrating the gain from more advanced gaze modeling.

anyway. Indicating these borderline images as relevant may even improve the retrieval performance, since the engine is guided towards retrieving images that match the conditions where the human user will need to actually process the image in order to determine its relevance.

On the other hand, the false negatives are often images that can be very easily recognized as being relevant. In the most extreme cases a user zooming in rapidly might correctly categorize the image without fixating on it even once, since the peripheral vision can capture sufficient visual features. Separating such images based on the gaze data alone is likely to be extremely hard or even impossible. Such behavior may, however, be an artifact of the experimental setting where the user is searching for images of a given category while not having a real interest in the images. In a real search task peripheral detection would more likely lead to closer inspection of the image.

3.2 Image retrieval

3.2.1 Setup

Encouraged by the good accuracy of the user-independent relevance predictor, we made a preliminary experiment with the full retrieval system using a pre-trained relevance predictor operating in real time. The predictor was trained on all of the data from the six users, and relevance threshold of $t = 0.6$ was chosen based on sensitivity analysis of the PicSOM retrieval engine (not shown). Then a set of experiments was performed, using a subset of the same six test subjects to evaluate the actual retrieval accuracy. The subjects were numbers 2, 5 and 6 in Figure 3, chosen to represent average users and excluding the users with the best and worst accuracy.

Each user performed the same six new search tasks. The tasks, shown in Figure 7, were chosen to provide a range of tasks with different complexities and proportions of rel-

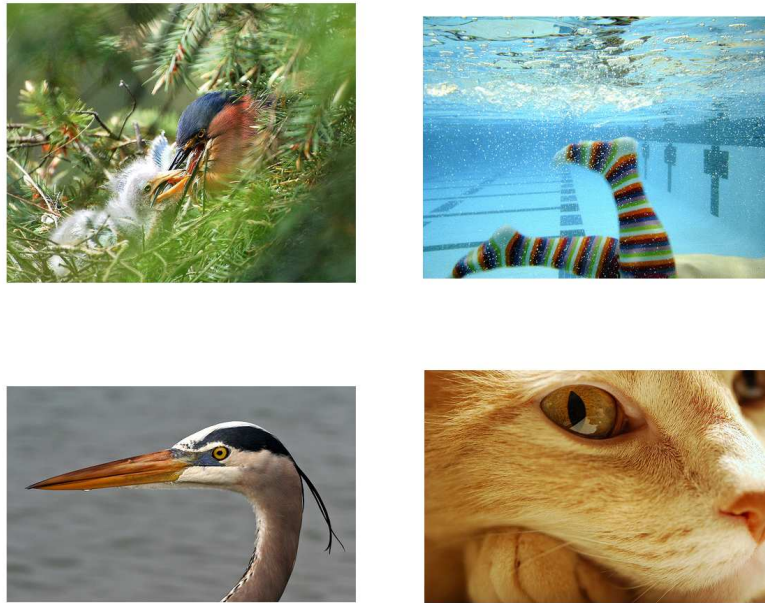


Figure 6: Sample images classified erroneously for the search task *animals*. The first row contains non-relevant images (according to the ground truth given by the users) that were predicted as relevant, whereas the second row contains images that were falsely predicted as non-relevant. Both false positive images show natural backgrounds that might well have animals (in fact, the left one has birds, yet the user did not indicate it as relevant). The false negative images are such that the user can trivially judge them to include animals based on a quick glance, not needing to focus on the image. The images are from the MIRFLICKR database, and released under Creative Commons license by the Flickr usernames jeslu (top left), amanky (top right), Terry Foote (bottom left), and Malingering (bottom right).

evant images. For two of the tasks each subject used the interface with the gaze-based relevance predictor. For two other tasks the predictor was replaced with a dummy one, predicting each image to be relevant with the probability of the images matching the proportion of relevant category labels in the image collection. This worked as a baseline for determining whether the interface can provide more relevant images than mere browsing in random order would. Finally, the remaining two searches were done with the same interface but using explicit relevance feedback from mouse clicks. This provides an estimate of the accuracy achievable with a traditional point-and-click relevance feedback.

Again the users indicated true relevance judgments after performing the search (not needed for the click-based comparison method). We then measured the performance of the different methods by counting the proportion of relevant images during the whole experiment. To take into account the variability caused by the search target, we constructed the experiment such that each of the three users did the same 6 tasks starting from the same initial conditions. However, different tasks were assigned to the three alternative methods, so that each combination of task+method was measured.

3.2.2 Results

Figure 7 shows, for each method, the proportion of relevant images shown during the experiment. In all six cases both the explicit and predicted relevance feedback result in

more relevant images than random ordering, although for both methods one of the six tasks (*dog* for explicit feedback and *flower* for predicted feedback) are very close to random. The explicit feedback gives on average, as expected, most relevant images, but for three out of six tasks (*people*, *animals*, *dog*) the implicit feedback is either comparable or even better than explicit feedback.

The results are promising, but more extensive testing is required to provide conclusive evidence. Content-based retrieval with relevance feedback is inherently very noisy, and already small differences in early stage feedback can lead to large changes in retrieval accuracy, which could only be smoothed out by averaging over a much larger collection of experiments. Taking into account also the user-specific prediction accuracies presented in Figure 3 we can, however, conclude that the performance of the GaZIR retrieval system is currently sensitive to the personal viewing patterns and the search task.

Further research is needed in particular to figure out why exactly the performance is so poor for *sunset* images, as well as to find out whether the predicted relevancies can actually outperform explicit feedback as shown here for *animals* and *dogs*. We have, however, potential explanations for both. The *sunset* images are easy to spot even with peripheral vision, since they can be characterized with simple color features. As suggested in Section 3.1.3, these kinds of images are often predicted incorrectly to be non-relevant. The ani-

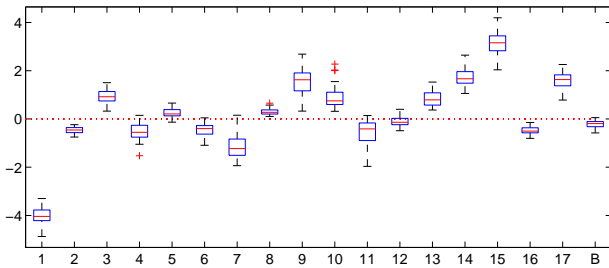


Figure 5: Weights of the user-independent relevance predictor. The numbering of the features corresponds to that in Table 1, and the last item is the bias term. The box-plots show the distribution of the weights in 50 runs, each using a random subset of the data for learning the model. Most of the weights are consistently above/below zero, indicating that the effect of that feature increases/decreases the probability of relevance. See text for further discussion of the features.

mal categories, on the other hand, are such that the user will inspect certain kinds of images to find out whether an animal is present or not, based on prior expectations about natural contexts for animals. Such images are often predicted to be relevant by our model even if there were no animals, which results in the retrieval engine fetching more such images that in turn are likely to contain animals. It may be that the retrieval engine can even more efficiently retrieve potentially relevant images with this kind of feedback as the feature representations used for retrieval consider the whole image, not just the object of interest.

4. DISCUSSION

We introduced an interface for content-based image retrieval. The interface interacts with existing image retrieval engines that utilize relevance feedback, and automates the relevance feedback collection by using eye tracking. The eye movement measurements are fed into a classifier predicting the relevances, and the predictions are in turn given to the engine.

The interface was designed to improve the information content of eye movements while still being simple and intuitive to use. It provides the user with a set of images at a time, but not in a standard grid structure of most retrieval interfaces. Instead, the images are shown as rings that enable easy movement from image to image, and different sets of images are shown as nested rings zooming towards infinity. This allows easy backtracking.

We demonstrated by empirical experiments that the relevance of the images can be inferred from eye movements with a reasonably good accuracy. Some typical cases of mistaken predictions were demonstrated, and high variance over users was pinpointed as the main open question — for some users the prediction accuracy is excellent, whereas for some the accuracy is not much better than random guessing. We also made brief preliminary experiments with the full retrieval system. The results are promising, but more extensive testing is required for conclusive evidence. Again, the model seems to work very well in some conditions (certain kinds of

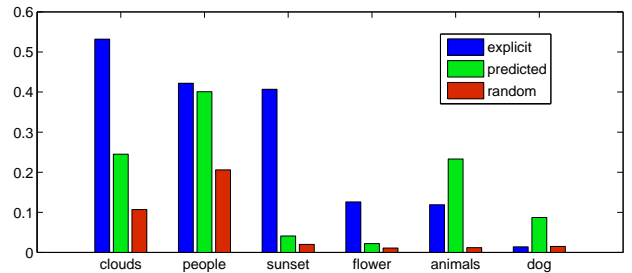


Figure 7: Retrieval performance in real user experiments. The bars indicate the proportion of relevant images shown during the search in six different search tasks for three different feedback methods. Explicit denotes the standard point-and-click feedback, predicted means implicit feedback inferred from gaze, and random is the baseline of providing random feedback. In all cases both actual feedback types outperform the baseline, but the relative performance of explicit and implicit feedback depends on the search task. See text for further analysis of the results.

search tasks, sufficiently large number of relevant images), while underperforming in others.

To our knowledge, the current system is the first attempt of building a sophisticated image retrieval interface that utilizes implicit gaze information. As such, it is definitely not a finalized version. Further work will be needed for instance to provide the user the possibility to correct mistakes made by the relevance predictor. As shown by the experiments, implicit cues from eye movements do not work well in all situations, and hence some kind of overriding or other interaction with explicit commands might be beneficial. One way towards that is to use the relevance predictions to alter the saliency of the images, for example by showing images that are less likely to be relevant in smaller size.

5. ACKNOWLEDGMENTS

The authors belong to the Finnish Center of Excellence in Adaptive Informatics. The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under *grant agreement* n° 216529, and from the TKK MIDE Programme (project UI-ART). We also acknowledge the support of the PASCAL2 EU FP7 network of excellence.

All photographs displayed in the paper are from the MIR-FLICKR database consisting of Flickr images released under Creative Commons license. We thank the photographers of the images and have credited the creators of the images reproduced in this paper, excluding the small thumbnail versions visible in the screenshots.

6. REFERENCES

- [1] R. Bates and H. Istance. Towards eye based virtual environment interaction for users with high-level motor disabilities. *Special Issue of International Journal of Disability & Human Development: The International Conference Series on Disability, Virtual Reality and Associated Technologies*, 4(3):275–282, 2005.

- [2] G. Buscher, A. Dengel, and L. van Elst. Eye movements as implicit relevance feedback. In *CHI '08: extended abstracts on Human Factors in Computing Systems*, pages 2991–2996, ACM, New York, NY, USA, 2008.
- [3] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 40(2):1–60, April 2008.
- [4] D. R. Hardoon, J. Shawe-Taylor, A. Ajanki, K. Puolamäki, and S. Kaski. Information retrieval by inferring implicit queries from eye movements. In *11th International Conference on Artificial Intelligence and Statistics*, Omnipress, 2007.
- [5] M. J. Huiskes and M. S. Lew. The MIR Flickr retrieval evaluation. In *MIR '08: Proceedings of the 2008 ACM International Conference on Multimedia Information Retrieval*, pages 39–43. ACM, New York, NY, USA, 2008.
- [6] R. J. K. Jacob. The use of eye movements in human-computer interaction techniques: what you look at is what you get. *ACM Transactions on Information Systems*, 9(2):152–169, 1991.
- [7] D. Kelly and J. Teevan. Implicit feedback for inferring user preference: a bibliography. *SIGIR Forum*, 37(2):18–28, 2003.
- [8] A. Klami, C. Saunders, T. de Campos, and S. Kaski. Can relevance of images be inferred from eye movements? In *MIR '08: Proceeding of the 2008 ACM International Conference on Multimedia Information Retrieval*, pages 134–140. ACM, New York, NY, USA, 2008.
- [9] J. Laaksonen, M. Koskela, and E. Oja. PicSOM – self-organizing image retrieval with MPEG-7 content descriptions. *IEEE Transactions on Neural Networks*, 13(4):841–853, 2002.
- [10] O. Oyekoya and F. Stentiford. Perceptual image retrieval using eye movements. *International Journal of Computer Mathematics, Special Issue on Computer Vision and Pattern Recognition*, 84:9, 2007.
- [11] K. Puolamäki, J. Salojärvi, E. Savia, J. Simola, and S. Kaski. Combining eye movements and collaborative filtering for proactive information retrieval. *Proceedings of the 28th annual international ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 146–153. ACM, New York, NY, USA, 2005.
- [12] K. Rayner. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3):372–422, 1998.
- [13] D. Salvucci and J. Goldberg. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the Eye Tracking Research and Applications Symposium 2000 (ETRA2000)*, pages 71–78. ACM, New York, NY, USA, 2000.
- [14] J. San Agustin, H. Skovsgaard, J. P. Hansen, and D. W. Hansen. Low-cost gaze interaction: ready to deliver the promises. In *CHI EA '09: extended abstracts on Human Factors in Computing Systems*, pages 4453–4458. ACM, New York, NY, USA, 2009.
- [15] A. Torralba, A. Oliva, M. Castelhana, and J. Henderson. Contextual guidance of eye movements and attention in real-world scenes: The role of global features on object search. *Psychological Review*, 113(4):766–786, 2006.
- [16] D. J. Ward and D. J. MacKay. Fast hands-free writing by gaze direction. *Nature*, 418:838–840, 2002.
- [17] C. Ware. *Information Visualization: Perception for Design*. Morgan Kaufmann Publishers, 2nd edition, 2004.