Can Relevance of Images Be Inferred from Eye Movements?

Arto Klami
Department of Information and Computer Science
Helsinki University of Technology
P.O. Box 5400, 02015 TKK, Finland
arto.klami@tkk.fi

Craig Saunders
School of Electronics and Computer Science
University of Southampton
SO17 1BJ, United Kingdom
cjs@ecs.soton.ac.uk

Teófilo E. de Campos
Xerox Research Centre Europe
6 chemin de Maupertuis
38240 Meylan, France
www.xrce.xerox.com/people/de_campos/

Samuel Kaski
Department of Information and Computer Science
Helsinki University of Technology
P.O. Box 5400, 02015 TKK, Finland
samuel.kaski@tkk.fi

ABSTRACT

Query formulation and efficient navigation through data to reach relevant results are undoubtedly major challenges for image or video retrieval. Queries of good quality are typically not available and the search process needs to rely on relevance feedback given by the user, which makes the search process iterative. Giving explicit relevance feedback is laborious, not always easy, and may even be impossible in ubiquitous computing scenarios. A central question then is: Is it possible to replace or complement scarce explicit feedback with implicit feedback inferred from various sensors not specifically designed for the task? In this paper, we present preliminary results on inferring the relevance of images based on implicit feedback about users' attention, measured using an eye tracking device. It is shown that, in reasonably controlled setups at least, already fairly simple features and classifiers are capable of detecting the relevance based on eye movements alone, without using any explicit feedback.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*relevance feedback*

General Terms

Experimentation, Human Factors, Verification

MIR'08, October 30–31, 2008, Vancouver, British Columbia, Canada.

1. INTRODUCTION

For text retrieval it is relatively easy to define a good set of keywords to be used as query terms. For most other information retrieval (IR) tasks, such as image searches, it is far more difficult to construct efficient queries. Existing solutions typically apply either of two predominant strategies: text-based queries on metadata, or iterative retrieval based on low-level features computed from the content. For the metadata strategy, all the readily available textual IR tools can be used, but the results depend heavily on the quality of the metadata. Constructing good metadata is laborious, time-consuming, expensive and, furthermore, the metadata cannot cover all aspects of the images.

The content-based approach [3, 10] sidesteps the need for metadata, but faces the problem that constructing useful and reliable features is difficult. Even with a good feature representation, it is not straightforward to formulate a query in terms of the features. In an image retrieval task the features may be complex descriptors of local image characteristics, such as color or texture, and a user cannot be expected to specify those manually. The most practical solution is to use pictorial examples for querying, as is done in the QBIC system [5] and in most state-of-the-art methods.

Regardless of the type of query, the modern content-based image retrieval (CBIR) systems, such as PicSOM [9], are typically highly interactive. The user is asked to refine the search by providing explicit feedback on the results, for instance by clicking on the relevant images. The images can then be analyzed and compared with images in a dataset for retrieval. Methods for image comparison have been progressing steadily, both in accuracy and in robustness, as reported in yearly benchmarks such as the PASCAL VOC [4] classification task. In classification, though, the set of classes is known a priori, and there is usually a significant number of training samples from each class.

For CBIR tasks, explicit feedback is accurate, but laborious for the user and limited in complexity. A user may at most be willing to click a few relevant images. In addition to explicit feedback, it is also possible to collect implicit feedback from sources not directly controlled by the user [8].

Various sensors can be used for monitoring the user and the context of the search, and the implicit feedback from them could be used to either complement the explicit feedback to improve search quality, or even to replace it completely. Indeed this is the approach we take in this paper. It is important to note that sensors are typically noisy and they cannot be focused to measure solely relevance. One particular challenge of this approach is to infer the correct relevance feedback from amongst the noise and nuisance signals.

1.1 Using Eye Movements

We consider one particular source of implicit feedback, namely eye movements of the user in an information retrieval task. Eye movements can be collected by relatively inexpensive and small-scale non-invasive equipment, making them a promising source for implicit feedback in practical applications in the near future.

Eye tracking has been used extensively in the psychology literature, and more recently also in tracking users' attention in information retrieval settings. Some examples include the human-computer interaction aspects of how users perform searches [2], analysis of user behavior in web search [6], and using eye movements as implicit relevance feedback in textual IR [7, 13]. In particular, [13] showed that in controlled settings it is possible, to a degree, to infer the relevance of document titles based on the gaze trajectory of the user. The promising results on the textual IR task suggest that using eye movements for relevance determination could be possible also in image retrieval tasks, where they would be even more severely needed.

In this paper we report on the first feasibility studies on using eye movements to improve content-based image retrieval. The task of the users is to search for images matching a literal query, and we try to infer the relevance of the images using the eye movements alone. No image features or metadata are used, but instead the relevance is inferred solely based on the trajectory of eye saccades and fixations on a collage of images. The focus is on inferring the relevance, not in using the inferred relevance as a feedback source in a real CBIR system. However, we also provide a demonstration that the image collection used in the experiments would provide sufficient content-level information for searching similar images based on the feedback.

The test setup of the first experiments is highly controlled, and even though it is simplified (small number of images, relevance determination rather than full search) it does start to reflect the kinds of activities occurring in a complete CBIR system. Even though the idea of using eye movements to detect the attention of the user has been discussed earlier (see e.g. [6, 11]), there has been very little work on inferring the relevance of images based on gaze patterns. Hence, we considered it worthwhile to start with a controlled setting that can then be expanded later. The preliminary results are very encouraging and show that in the chosen setup it is indeed possible to infer the relevance of the images with reasonably high accuracy, and hence eye movements can be used as a source of implicit relevance feedback. This justifies moving to more advanced test setups, using also content-based features instead of just the eye movements, applying more advanced learning algorithms, and using the inferred relevance as a feedback source in a real CBIR system.

Eye tracking provides a non-invasive way to obtain accurate information from the users, without asking them to perform any additional tasks that might interfere with their main task. Only initial calibration is required. As the devices continue to reduce in size and cost, they may become one of the most informative and natural sensor mechanisms for gathering useful user data at low cost. Ubiquitous use of such devices would facilitate personalization and adaptivity of user interfaces, and application-driven scenarios such as information retrieval. One of the goals of this study is to give evidence about the potential of such systems given the current hardware.

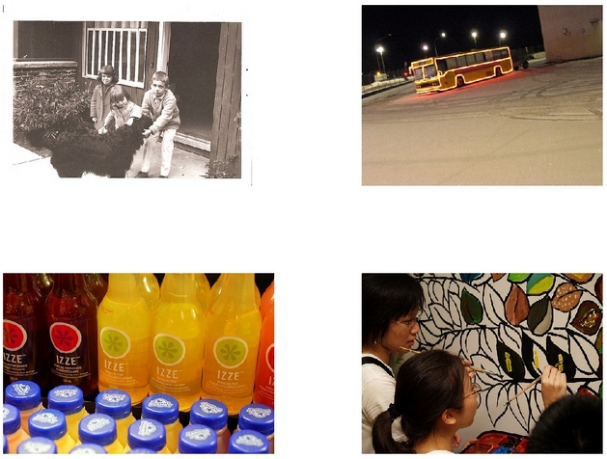

Figure 1: Example of a page with no relevant images.

2. TEST SETUP AND LEARNING TASKS

We study a simple search task, where the user is searching for images related to sports. The rationale behind choosing sports as a search target is that it is a broad and abstract concept, making it challenging for content-based image retrieval methods. Use of implicit feedback will be very valuable for this type of query. The task is intentionally left slightly vague, and some images (such as motorbikes or static shots of a sports celebrity) are open to user-specific interpretation.

The user is searching for the relevant images in a collection of 400 images, which are displayed on 100 pages of 4 images each. Each page has either zero or one sports-related image. Figures 1 and 2 show sample pages, including both simple and challenging examples.

In a complete feedback-driven CBIR system the images would probably be shown as larger collages. Instead of just 4 images, the collage might contain tens of thumbnail images. We start with the small collage to be able to present the images in a larger size, and to avoid any issues related to matching the gaze to the thumbnails. The restriction to at most one relevant image per page was done to ease collecting the data; in a collection phase the users gave explicit feedback, and only two possible values (the page is relevant or it is not) allowed using keyboard as a feedback source. That avoids the problem of the explicit feedback interfering with the eye movements.

Given the eye movements, we want to predict the relevance of the target, training the predictor with explicit feed-

Figure 2: Two examples of relevant pages. The left page is unambiguous; the image of a boxer on the top right corner is clearly about sports. The right page is an example of a more challenging page. The image in the top right corner is about sports, but if the user does not recognize the athlete it may well be considered as a non-sports image. Note that correctly detecting it as a sports image (especially e.g. from a dataset of images of people) based on manually dictated image features would be practically impossible.

back as an indicator of the true relevance. Two different relevance determination tasks can be considered. The first is to detect pages with a relevant image. That is, we try to utilize the gaze pattern on the set of four images to infer whether any one of them was about sports. The second possible task is to find out which of the images was relevant on the pages that had a relevant image. We present preliminary results for both of these tasks.

The tasks can be further categorized based on what kind of data is available for training. The simplest learning task is to predict the relevance of a new page or image based on training data from the same user. This corresponds to learning to predict the future behavior of a single user. Alternatively, we might be interested in learning to predict the behavior of a new user, training the classifier with data collected from other users solving the same tasks. That is, this second task is to predict the behavior of a new user in a known task.

The most difficult and interesting task would be to predict the relevance of a new page or image for a new query, possibly performed by a new user. This can only be done if the gaze patterns are sufficiently universal over users and queries. Content-based methods typically do not generalize well to completely new scenarios, such as databases of different kinds of images or queries of clearly different type, but gaze patterns can potentially be partly invariant to the actual content. Alternatively, we could consider a kind of collaborative filtering task, where we would have a few example pages for the new query and user, and a larger collection of training data for users and queries. Here, however, we study only the two more straightforward types of learning, generalization to new pages with a given query and generalization to new users. The question of generalization to new queries is left for future research.

3. DATA COLLECTION AND FEATURE EXTRACTION

The measurements were done with a Tobii X120 eye tracker. The machine has a set of infra-red leds and an infra-red stereo camera, and the tracking is based on detection of pupil centers and corneal reflection. The machine has an accuracy of 0.5 degrees, has a sample rate of 120Hz, and it allows relatively free head movement. The tracker was connected to a PC with a 19-inch flat-panel monitor, using a resolution of 1280x1024. The experiment was done using a standard web browser. In short, the setup mimics the way people typically would use a workstation in a search task.

Out of a total of 400 images, 70 were relevant (i.e., about sports, collected from local databases and internet using simple keyword searches) and 330 were not (taken mainly from the Pascal Visual Objects Challenge VOC2007 [4]). This gives us 70 relevant pages, each including one relevant image and 3 non-relevant images, and 30 pages where all 4 images are non-relevant. The database intentionally contains images that may lead to bias different users (through different personal definitions of 'sport' and a wide variety of subject matter in the non-relevant image set).

A total of 27 test users performed the search task. Each user searched through the same 100 pages, and pressed a key for relevant/non-relevant accordingly. The users were asked to perform the task as quickly as possible, to avoid eye movements not related to the task. Summary statistics gathered from the data are shown in Table 1.

The accuracy of the users was good; on average people got more than 95% of the relevance judgments correct. The best user only made one mistake, whereas the worst made 12. The range of timings for the task was quite variable, and does not seem to correlate well with the accuracy. In the analysis, we ignored the first 10 pages for each user; it typically took a few pages for people to adapt to the task.

We preprocessed the raw eye movement data by finding fixations with the built-in fixation filter provided by Tobii Technology. The filter judges a series of raw coordinates to be a single fixation if the coordinates stay sufficiently long within a sphere of a given radius. We used a threshold of 100ms and a radius of 30px. Example illustrations of the

Table 1: User accuracy (number of correct judgments) and timing summaries for each participant.

Participant ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Sport (70)	70	69	65	68	57	69	69	64	62	69	64	70	70	68
Non-Sport (30)	29	29	26	28	30	28	28	28	26	28	28	29	28	29
Time	4:06	2:51	4:00	3:00	3:50	3:26	3:54	3:41	3:41	4:20	3:21	2:37	4:20	3:32

Participant ID	15	16	17	18	19	20	21	22	23	24	25	26	27	
Sport (70)	65	68	70	69	69	61	69	58	66	68	69	69	67	
Non-Sport (30)	29	29	29	29	28	29	29	29	28	29	29	29	28	
Time	3:20	3:22	2:38	2:59	4:27	2:55	3:29	3:59	4:53	3:43	2:57	2:42	3:30	

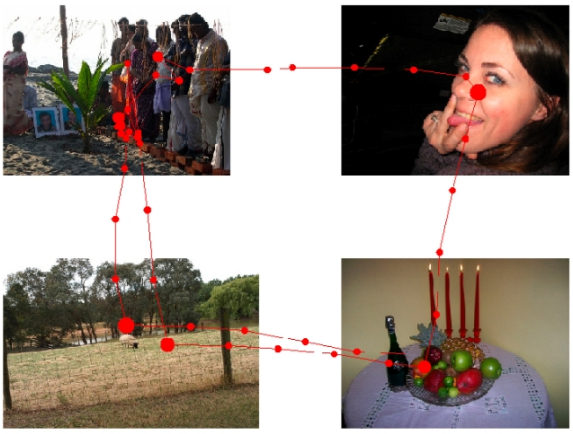

Figure 3: Illustration of the eye movements in the search task. The line connects the raw coordinates of the left eye of a single user, and the bigger spheres mark fixations detected by the fixation filter. The trajectory begins from the top left corner, moves in a clockwise direction and visits all four images, then double checks the other images moving anti-clockwise. Most attention is focused on the top-left image.

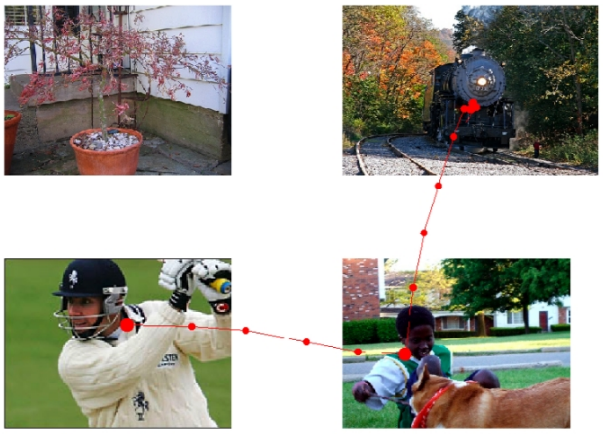

Figure 4: Eye movements for a page with a relevant image. Notice how the user does not even look at one of the images, as soon as a relevant image is found the user can tag the page as relevant and then move on.

measured data and fixations for a page with nonrelevant and relevant images are shown in Figures 3 and 4 respectively.

Based on the fixations, we then computed a simple feature vector for the trajectory of each page. The features are listed in Table 2. Finally, we discarded all user/page pairs for which we had no fixations, typically due to the tracker being unable to detect the pupils during the page.

Table 2: Features used to represent the gaze patterns. The total length of the feature vector is nine.

#	Feature
1	Total length of fixations
2	Number of fixations
3	Average length of fixations
4	Num. of transitions from an image to another
5	Num. of images with at least one fixation
6-9	Number of fixations within each image

4. EVALUATING THE IMAGE SET

Although the primary focus of this paper is in evaluating the possibility of inferring the relevance of images based on eye movements, relevance feedback alone is however of little use unless we can find new images that are similar to the ones deemed relevant. Next, we briefly demonstrate that given a set of images labeled as relevant and non-relevant, it is possible to find new images that are predicted as more likely to be relevant in the image database used for the experiments. Based on these predictions one could then use exploration/exploitation strategies to return a subset of images to the user (from possibly a very large database) in order to get further relevance information and narrow the search criteria.

The technique of [12] is used as it was shown to exhibit state-of-the-art performance in the PASCAL VOC 2007 challenge [4] and the image retrieval competition in CLEF 2007 [1]. This method is based on Fisher scores of a bag-or-visual-words (BOV) representation for the images. First, a generative model, here a Gaussian mixture model (GMM), is trained to approximate the distribution of the low-level features in images. It can be seen as a visual vocabulary, where each Gaussian component $\mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ models a visual word. Based on the GMM, a new fixed-dimensional representation for each image is obtained as the normalized Fisher score

of the image. This representation can then be subsequently fed to a discriminate classifier for categorization, or used to compute the similarities between images for retrieval.

Given a maximum likelihood estimate for the parameters of the GMM, $\hat{\Phi} = \{w_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i,\ i = 1, \cdots, K\}$ (where w_i is the mixture's weight, and K is the number of mixture components), the Fisher score $\mathbf{u}(\mathbf{x}, \hat{\Phi})$ of an image represented with low-level features $\mathbf{x}$ is defined as the gradient vector of the log-probability:

$$\mathbf{u}(\mathbf{x}, \hat{\Phi}) = \nabla_{\Phi} \log p(\mathbf{x}|\Phi),$$

evaluated at the maximum likelihood estimate. This gradient describes the direction in which the model parameters should be modified to best fit the image features. Each score is then normalized using the Fisher Information matrix $\mathbf{F}_{\hat{\Phi}}$, giving the final representation of an image as the normalized Fisher score

$$\mathbf{v}(\mathbf{x}, \hat{\Phi}) = \mathbf{F}_{\hat{\Phi}}^{-1/2} \mathbf{u}(\mathbf{x}, \hat{\Phi}) \quad (1)$$

with $\mathbf{F}_{\hat{\Phi}} = E_{\mathbf{x}}[\mathbf{u}(\mathbf{x}, \hat{\Phi})\mathbf{u}(\mathbf{x}, \hat{\Phi})^T]$. Details on how to compute (1) for a GMM are available in [12]. Note that this is a fixed length representation, whose size only depends on the number of parameters in the model.

For the purposes of this paper, the Fisher score representation is used to retrieve images that are more likely to be relevant, using the implicit feedback from eye movements to define the target of the retrieval. The retrieval can be performed either by looking for images similar to the ones marked relevant, or by building a classifier that tries to predict for the rest of the database whether the images are relevant or not. Here, the latter approach is demonstrated, with sparse logistic regression classifiers in the Fisher score representation. We extract locally a set of low-level image features in each image. Two different features are used: a histogram of edge orientations, which describes texture, and the local mean and variances in RGB to describe color. For each feature type, one visual vocabulary was built and one classifier was trained. The two separate classifiers were combined linearly.

To evaluate the sports dataset, we used 10-fold cross validation, obtaining a classification accuracy of 91.2%, with the confusion matrix shown in Table 3. These results show that despite the fact that sport relevance is a broad and abstract concept, it is possible to distinguish between relevant and non-relevant images with a relatively high accuracy, indicating that if enough feedback is gathered, accurate retrieval can be achieved. However, Figure 5 shows some of the images that were miss-classified by this system. Note that such images are not ambiguous for human observers in the sport/not-sport query. This means that if they are presented to the user as intermediate results of the query, the implicit feedback gathered from eye tracking will easily teach the retrieval system that these are not relevant samples, which will improve the chances of reaching the target.

Table 3: Confusion matrix of the sports relevance dataset. The rows indicate true classes, and the columns the predictions given by the categorizer. 43 true positives corresponds to 84.3% precision and 61.4% recall.

	Prediction	
	Relevant	Not-relevant
Relevant	43	27
Not-relevant	8	322
Total	51	349

5. DETERMINING THE RELEVANCE

Having shown that if one has sufficient relevance information it is possible to retrieve images more likely to be relevant from a database, we now return to using the eye-movements alone to infer relevance and obtain the necessary feedback for the retrieval mechanism to operate. In this initial study we use a simple classifier, linear discriminant analysis (LDA), in order to predict the relevance. LDA searches for a linear subspace such that the classes become discriminated as well as possible in that subspace. The method assumes that each class follows a normal distribution with a shared covariance matrix, and measures the separation S in direction $\mathbf{w}$ as

$$S = \frac{\mathbf{w}^T \boldsymbol{\Sigma}_b \mathbf{w}}{\mathbf{w}^T \boldsymbol{\Sigma} \mathbf{w}}.$$

Here $\boldsymbol{\Sigma}_b$ is the sample covariance of the class means, and $\boldsymbol{\Sigma}$ is the covariance of the data.

Given a data with C classes, the optimal set of $C-1$ directions $\mathbf{w}_c$ can be found as eigenvectors of $\boldsymbol{\Sigma}_b \boldsymbol{\Sigma}^{-1}$. The actual classification task in the subspace is solved by assigning the test samples to classes with the highest likelihood, weighted by the class sizes.

6. RESULTS

6.1 User-specific Model for Relevance

The ability to infer user-specific models of relevance is studied with a leave-one-out procedure. Each user has a separate model, and for each page the classifier is trained using all other pages. We considered both the binary classification task of detecting the pages with at least one relevant image, and the task of detecting which of the images was relevant (a four-class classification task, applied only to pages that had a relevant image).

For the binary classification task we use the traditional information retrieval quality measure of area under the ROC curve. That is, the pages are ordered according to the predicted relevance, and we study the curve of true positives versus false positives when lowering the relevance threshold. A perfect retrieval algorithm would obtain a score of 1, whereas random ordering gives 0.5. For the image detection task, we compute the percentage of correct classifications. As there are 4 images on each page, the baseline accuracy obtained by random guessing would be 25%.

The scores for each of the test subjects are collected in Table 4 (columns labeled as "User-specific"). For all subjects, the scores for both tasks are clearly above the random baseline, showing that both the relevant pages and the relevant images on the pages can be detected reasonably well solely based on the gaze pattern. Note that the quality measures of the two tasks are different, and hence are not directly comparable.

6.2 Learning Relevance from Other Users

To test how well the gaze patterns generalize over the users, we use a different kind of leave-one-out procedure. This time each user is left out at a time, and the classifier

Figure 5: Top: examples of images miss-classified as sports. Bottom: examples of images miss-classified as not-sport.

is trained using all the data measured from the other users. The identity of neither the pages nor the users is used in the learning. Instead, all of the training data is pooled together as if it were a collection of independent identically distributed samples. The learning tasks and error measures are shared with the previous case.

The results can be found in Table 4 (columns labeled as "Global"). Again, all classification results are clearly above the random baselines. For both tasks and most users, the accuracies are slightly better compared to the user-specific models. This is most likely caused by having considerably larger amount of training data; in the user-specific model the relevance of a single page is predicted based on at most 89 training examples, whereas here all pages of the 26 other users are used for training. The good performance without using measurements from the user at all suggests that the gaze patterns are relatively universal in this kind of a task. For some subjects the scores for a user-specific model are higher even with a small training set, which hints towards user adaptation still being useful. Combining both approaches should further improve the accuracy.

7. DISCUSSION

In this paper, we made preliminary experiments on inferring relevance of images from the eye movement trajectory of the users. It was shown that at least in a simplified test setup it is possible to detect the relevant images reasonably accurately, even with a simple classifier that uses a set of relatively simple eye movement features. This classifier is able to detect the relevance even when not using any training data from the particular user in question, which suggests that the gaze patterns of different users are similar enough for tasks of this kind.

The visual information contained in the image dataset was evaluated with the image categorization method of [12], showing that it is possible to obtain accurate search results if enough feedback is provided, even if the querry is not straightforwardly described by image features. A natural continuation of this work is to integrate the implicit feedback obtained from eye tracking with a CBIR system based on [12].

Relevance of texts has earlier been estimated based on eye movements. In the most closely related works [13], word-specific features were computed from the gaze trajectory, averaged over the text, and classified with a simple classifier. The features were derived from reading studies, and include total lengths of fixations, lengths of transitions between words etc. We used similar features, modified to be suitable for images. Features more optimal for images need to be developed in future studies, although even the current set performed very well. Another direction forward is to model the sequence of browsing interleaved with inspection of the images, instead of simply averaging all features over the images. For the texts the browsing was modeled with a Markov Chain, in which each state is a Hidden Markov Model which models the reading pattern.

8. ACKNOWLEDGMENTS

We wish to sincerely thank the volunteers from Xerox RCE for participating in the experiment, and F. Perronnin for the useful feedback. 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. We also acknowldedge the support of the PASCAL2 network of excellence. AK and SK belong to the Finnish Center of Excellence in Adaptive Informatics Research and Helsinki Institute for Information Technology.

Most of the images reproduced in the figures of this paper are from the PASCAL VOC 2007 database. For the list of copyright owners, see [4].

9. REFERENCES

[1] S. Clinchant, J. Renders, and G. Csurka. XRCE s participation to imageclefphoto 2007. In *Working Notes of the CLEFWorkshop*, 2007.

Table 4: Results of the relevance determination tasks. The first two columns (Page relevance) are for the experiment on detecting whether the page contains a relevant image. The reported values are area under curve (AUC) scores, and the baseline of random guessing would be 0.5. The last two columns (Image relevance) are for predicting which of the 4 images was relevant, and the reported values are classification accuracies (percentage of correct choices). Random guessing would achieve a score of 25%. In both experiment types, user-specific refers to a model trained with all other pages of the particular user, whereas global refers to a model trained with all other users except the one being tested.

	Page relevance		Image relevance	
ID	User-specific	Global	User-specific	Global
1	0.83	0.87	54.0	58.7
2	0.86	0.78	58.7	57.1
3	0.83	0.80	60.9	51.6
4	0.84	0.88	64.5	62.9
5	0.80	0.83	65.0	68.3
6	0.68	0.76	44.0	58.0
7	0.82	0.89	52.4	68.3
8	0.87	0.81	64.8	70.4
9	0.71	0.81	59.4	60.9
10	0.87	0.90	66.7	77.8
11	0.81	0.86	73.4	71.9
12	0.91	0.93	74.6	74.6
13	0.90	0.94	76.6	79.7
14	0.87	0.91	60.9	70.3
15	0.74	0.84	71.9	81.2
16	0.84	0.88	65.9	68.2
17	0.75	0.78	71.4	76.2
18	0.60	0.64	62.7	72.9
19	0.81	0.83	71.9	75.0
20	0.64	0.79	33.3	66.7
21	0.91	0.94	76.2	73.0
22	0.65	0.70	45.2	54.8
23	0.72	0.76	62.5	70.3
24	0.86	0.89	59.4	64.1
25	0.88	0.90	62.5	62.5
26	0.88	0.92	76.6	65.6
27	0.87	0.92	61.9	68.3
Mean	0.81	0.84	62.9	67.7
Std	0.09	0.08	10.5	7.7

[2] E. Cutrell and Z. Guan. What are you looking for?: An eye-tracking study of information usage in web search. In *CHI '07: Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 407–416, New York, NY, USA, 2007. ACM.

[3] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Transactions on Computing Surveys*, 40(2), 2008.

[4] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. `http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html`, 2007.

[5] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker. Query by image and video content: The QBIC system. *Computer*, 28:23–32, 1995.

[6] L. A. Granka, T. Joachims, and G. Gay. Eye-tracking analysis of user behavior in WWW search. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and Development in Information Retrieval*, pages 478–479, New York, NY, USA, 2004. ACM.

[7] 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*, 2007.

[8] D. Kelly and J. Teevan. Implicit feedback for inferring user preference: A bibliography. *ACM SIGIR Forum*, 37(2):18–28, 2003.

[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] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain. Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2(1):1–19, 2006.

[11] P. P. Maglio and C. S. Campbell. SUITOR: An attentive information system. In *Proceedings of the 5th International Conference on Intelligent User Interfaces*, pages 169–176, 2000.

[12] F. Perronnin and C. Dance. Fisher kernel on visual vocabularies for image categorization. In *Proc. Computer Vision and Pattern Recognition*, Minneapolis, MN, USA, June 18-23 2007.

[13] J. Salojärvi, I. Kojo, J. Simola, and S. Kaski. Can relevance be inferred from eye movements in information retrieval? In *Proceedings of WSOM'03, Workshop on Self-Organizing Maps*, pages 261–266. Kyushu Institute of Technology, Kitakyushu, Japan, 2003.