Deliverable D2.1.1 + D2.1.2

Prediction of relevance of an image from a scan pattern

Demonstrator for relevance prediction from a scan pattern

Contract number: **FP7–216529** PinView

Personal Information Navigator Adapting Through Viewing

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.

Identification sheet

Project ref. no.	**FP7–216529**
Project acronym	PinView
Status and version	Final, Revision: 1.00
Contractual date of delivery	31.12.2008
Actual date of delivery	31.12.2008
Deliverable number	D2.1.1 + D2.1.2
Deliverable title	Prediction of relevance of an image from a scan pattern Demonstrator for relevance prediction from a scan pattern
Nature	report + prototype
Dissemination level	Report: PU – Public Prototype: PP – Restricted to other programme participants
WP contributing to the deliverable	WP2 Learning relevance feedback from eye tracking
Task contributing to the deliverable	Task 2.1 Relevance of an image from a scan pattern
WP responsible	Teknillinen korkeakoulu
Task responsible	Teknillinen korkeakoulu
Editor	Arto Klami, <arto.klami@tkk.fi>
Editor address	P.O.BOX 5400, FI-02015 TKK, Finland
Authors in alphabetical order	Teófilo de Campos, Samuel Kaski, Arto Klami, Kitsuchart Pasupa, Craig Saunders
EC Project Officer	Pierre-Paul Sondag
Keywords	gaze trajectory, image retrieval, implicit relevance feedback
Abstract	This report considers the task of inferring implicit relevance feedback from eye movements in image retrieval settings. The feasibility of solving the problem without using any image-level features is demonstrated on two different search settings, and the accuracy of inferring the relevance feedback is shown to be relatively high, clearly better than random. In addition, the report provides a list of image-level features that are good cues for relevance.

List of annexes

`klami08mir.pdf - Publication ‘‘Can relevance be inferred from eye movements?’’`
`relevance_predictor.zip - Classification platform for predicting image relevance`

Contents

1 Overview

This deliverable that constitutes the output of Task 2.1 *Relevance of an image from a scan pattern* of the *Personal Information Navigator Adapting Through Viewing*, PinView, project, funded by the European Community's Seventh Framework Programme under Grant Agreement n° 216529, consists of two parts. The first part is this report, Deliverable 2.1.1 *Prediction of relevance of an image from a scan pattern*, describing the task of inferring relevance feedback from eye movements. The second part is a prototype, Deliverable 2.1.2 *Demonstrator for relevance prediction from a scan pattern*, that demonstrates the process of inferring image relevance in practice.

The primary content of Deliverable 2.1.1 is the publication *Can relevance of images be inferred from eye movements?* [8] published in ACM Multimedia Information Retrieval conference. The publication, included as the first annex, shows how relevance feedback can be inferred solely from eye movements in a certain kind of image retrieval setting. The remainder of the report, the results presented in the publication are complemented with additional experiments performed on data collected in Task 8.3 *Eye movement data collection campaign*. A list of eye movement features that are good cues of relevance is produced as the output of the additional experiments.

The Deliverable 2.1.2, demonstrator for relevance prediction, is described as a part of this report, in Section 7. It consists of a platform for testing different classifiers in the task of predicting image relevance from scan patterns.

The work presented in this report will be continued mainly in Task 2.2 *Relevance of parts of an image from the viewing pattern* and Task 2.3 *Data fusion*. In particular, the output of this task will determine how the work in WP 2 is continued. As described in Section 8, the relevance predictor is clearly accurate enough to enable continuing with Task 2.2 without needing to consider external information sources at this stage.

This report describes contributions of three project partners, TKK, SOTON-ECS, and XEROX. The contributions of UCL are ongoing work, and will be reported in the deliverables of the later tasks.

2 Introduction

This report considers the problem of inferring implicit relevance feedback from eye movements in image retrieval setting, to be eventually used as a component in a content-based image retrieval system (CBIR). Given a set of images and a gaze trajectory measured when the user was viewing the images, the system should infer which of the images were relevant for the user in the search task he was performing. Earlier it has been shown that inferring implicit relevance feedback from eye movements is feasible, to an extent, for texts [13, 14]. The main purpose of Task 2.1 was to verify to which degree the results generalize to images.

There is a fundamental difference between texts and images in terms of eye trajectories and their use in machine learning. For texts, the typical task is to infer the relevance of a sentence or a paragraph, and those can be naturally split into smaller elements, typically words. This division is readily available by segmenting the text, and it corresponds to how humans read. Hence, the features used for predicting relevance should be based on how individual words within the sentence or paragraph are viewed. Most state-of-art methods using eye movements in textual information retrieval use this kind of approach [1, 14].

For images there is no such natural subdivision, which means that the approaches used for text retrieval do not directly generalize to images. This issue can be alleviated with two solutions. The first option is to use only image-level features, building a single feature vector for each image as if it was the smallest scale semantic element. This kind of approach has been used in image retrieval e.g. by [10]. It is also worth noting that recently [3, 11] considered similar approach in text retrieval, predicting word-level relevance and using it for improving document retrieval.

The other solution is to find an implicit subdivision through various interest point detection methods (see e.g. [9] for an overview and empirical comparisons) or image segmentation algorithms, and then build eye-movement trajectories over elements constructed from the output of such algorithms. Alternatively, computational models of visual saliency [5] could be used to provide the interest points, or data measured from users not performing the same task could be used to build data-driven saliency models [7].

In this report we consider the former approach and focus on finding good feature representations for full images, without considering sub-image elements. The latter approach will be studied in conjunction with Task 2.2 *Relevance of parts of an image from a viewing pattern*, where that kind of feature representation level will necessarily be needed. After good sub-image representations are found, we will return to the problem of inferring relevance of full images using the same sub-image elements for building more complex models.

In this report we show that it is possible to infer relevance feedback based on gaze trajectories already with the simplified representation. The accuracy in distinguishing relevant images from non-relevant ones is consistently above random guessing for all users and several search tasks in two different types of search interfaces. We also provide a list of features that are good cues for relevance.

3 Data and setting

All the experiments reported in this deliverable use data measured in Task 8.3. The experimental setup used by [8] is described in detail in the publication itself, and the setup used for the additional results presented in this report is described below. The main difference between the two settings is that [8] shows only four large images at a time, whereas here the display reminds more closely a traditional CBIR system, showing 15 thumbnail images at once.

In total, data from 23 users performing the "Count the relevant images" task described in Section 4.8.3 of Deliverable 8.3 were used. Each user viewed a set of 20 pages, each containing

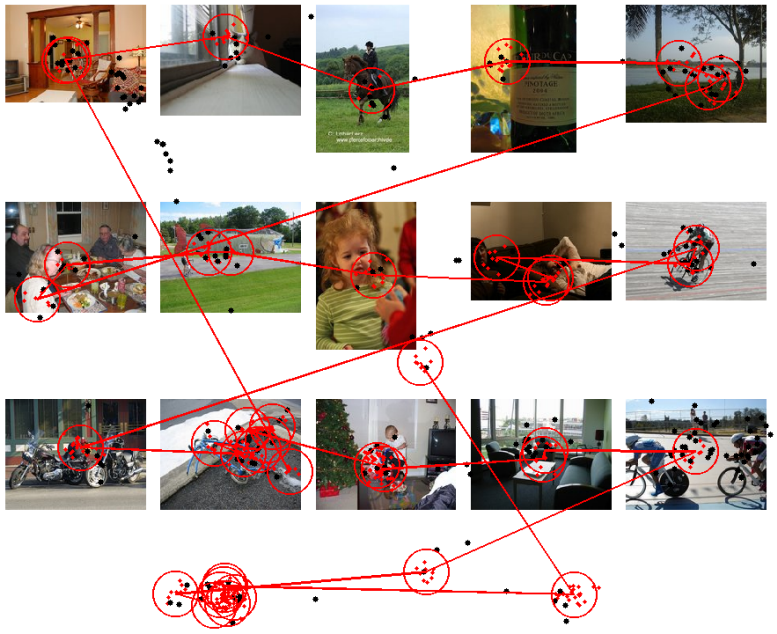

Figure 1: An illustration of the search interface with overlaid eye movement measurements. The red spheres mark fixations and small red dots correspond to raw measurements that belong to those fixations. The black dots mark raw measurements that were not included in any of the fixations. It is worth noting how the scanning pattern closely resembles reading as it proceeds row by row from left to right. However, there are fewer regressions and backtracking than in typical reading experiments. The measurements at the bottom of the screen correspond to control elements used in the experiments, and were ignored in the analysis. The user does not switch between viewing the images and the control elements, demonstrating that the control elements were sufficiently unobtrusive in order to not disturb the eye movement measurements.

15 thumbnail images, while searching for images belonging to a certain category. The users were performing one of three different search tasks, which were: Bicycle (8 users), Horse (7 users), and Transport (8 users). Explicit relevance feedback was collected as a separate step after viewing each page of images. This explicit relevance feedback was used for learning the relevance predictor. Example illustration of a user viewing a page is shown in Figure 1. The image shows a typical scanning pattern which largely resembles reading; the user scans the images row by row, mostly travelling from left to right within each row. This pattern was shared by many users, but not all of them.

In this report we consider only user- and task-specific models, since there is sufficient amount of training data for each case. However, [8] presents promising results on feasibility of using training data measured from other users, obtaining high accuracy without using any user-specific training data. More generally, it would be possible to use a small training set collected for the specific user and task, and utilize training data from similar users or tasks with multi-task learning methods such as [6].

4 Features

As described in the introduction, we consider only features computed for full images. Each feature can be computed based on only the eye trajectory and locations of the images in the page. This kind of features are general-purpose and easily applicable in all application scenarios.

The features are divided into two categories. The first category uses directly the raw measurements obtained from the eye-tracker, whereas the second category is based on fixations estimated from the raw data. A fixation means a period of maintaining the gaze around a given point, and most of the visual processing happens during fixations, due to blur and saccadic suppression during the rapid saccades between fixations (see, e.g., [2]). Often visual attention features are hence based solely on fixations and relations between them [12]. Here we include additionally raw-measurement features to provide data also for images that contain no fixations, and to enable verifying which kind of features work better. Raw measurement data might also be able to overcome possible problems caused by imperfect fixation detection.

Table 1 shows the list of candidate features considered. Most of the features are motivated by features considered earlier for text retrieval studies [15]. The features cover the three main types of information typically considered in reading studies: fixations, regressions (fixations to previously seen images), and refixations (multiple fixations within the same image). However, the actual forms of the features have been tailored towards being more suitable for images, trying to include measures for things that are not relevant for texts, such as how big a portion of the image was covered. Fixations were detected using the fixation detector provided by Tobii, the manufacturer of the measurement devices, with settings "radius 50pixels, minimum duration 100ms".

Some of the features are not invariant of the location of the image on the screen. For example, the typical pattern of moving from left to right means that the horizontal coordinate of first fixation for the left-most image of each row typically differs from the corresponding measure on the other images. Features that were observed to be position-dependent were normalized by removing the mean of all observations sharing the same position, and are marked in Table 1. Finally, each feature was normalized to have unit variance and zero mean.

5 Feature importance

We study the contribution of features in predicting the relevance of images using linear discriminant analysis (LDA). LDA searches for a linear subspace such that the relevant and non-relevant images become discriminated as well as possible, and hence the found subspace is indicative of the importance of the features in separating the classes. The method implicitly assumes that the classes follow normal distribution, and finds the projection vector $\mathbf{w}$ maximizing the separability criterion

$$S(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{\Sigma}_b \mathbf{w}}{\mathbf{w}^T \mathbf{\Sigma} \mathbf{w}}, \tag{1}$$

where $\mathbf{\Sigma}_b$ is the sample covariance of the class means and $\mathbf{\Sigma}$ is the covariance of the data. A similar approach has earlier been used for finding good features for textual IR studies [15].

We trained a separate model for each user, and hence got 23 projection vectors $\{\mathbf{w}_i\}_{i=1}^{23}$. These are summarised using two different methods. First, simple averaging of the absolute values of the elements of $\mathbf{w}_i$ is used to roughly characterise the overall importance of the features for all users. Absolute values are used because (1) is sign-invariant, i.e. $S(\mathbf{w}) = S(-\mathbf{w})$. Second, we rank the features in order of importance for each user separately, and

Number	Name	Description
		Raw data features
1	numMeasurements	total time of viewing the image
2	numOutsideFix	total time for measurements outside fixations
3	ratioInsideOutside	percentage of measurements inside/outside fixations
4	xSpread	difference between largest and smallest x-coordinate
5	ySpread	difference between largest and smallest y-coordinate
6	elongation	ySpread/xSpread
7	speed	average distance between two consecutive measurements
8	coverage	number of subimages[1] covered by measurements
9	normCoverage	coverage normalized by numMeasurements
10*	landX	x-coordinate of the first measurement
11*	landY	y-coordinate of the first measurement
12*	exitX	x-coordinate of the last measurement
13*	exitY	y-coordinate of the last measurement
14	pupil	maximal pupil diameter during viewing
15*	nJumps1	number of breaks[2] longer than 60ms
16*	nJumps2	number of breaks[2] longer than 600ms
		Fixation features
17	numFix	total number of fixations
18	meanFixLen	mean length of fixations
19	totalFixLen	total length of fixations
20	fixPrct	percentage of time spent in fixations
21*	nJumpsFix	number of re-visits to the image
22	maxAngle	maximal angle between two consecutive saccades[3]
23*	landXFix	x-coordinate of the first fixation
24*	landYFix	y-coordinate of the first fixation
25*	exitXFix	x-coordinate of the last fixation
26*	exitYFix	y-coordinate of the last fixation
27	xSpreadFix	difference between largest and smallest x-coordinate
28	ySpreadFix	difference between largest and smallest y-coordinate
29	elongationFix	ySpreadFix/xSpreadFix
30	firstFixLen	length of the first fixation
31	firstFixNum	number of fixations during the first visit
32	distPrev	distance to the fixation before the first
33	durPrev	duration of the fixation before the first

[1] The image was divided into a regular grid of 4x4 subimages, and covering a subimage means that at least one measurement falls within it

[2] A sequence of measurements outside the image occurring between two consecutive measurements within the image

[3] A transition from one fixation to another

Table 1: List of features considered in the study. The first 16 features are computed from the raw data, whereas the rest are based on pre-detected fixations. Note that features 2 and 3 use both types of data since they are based on raw measurements not belonging to fixations. All features are computed separately for each image. Features marked with * were normalized for each image location; see text for details.

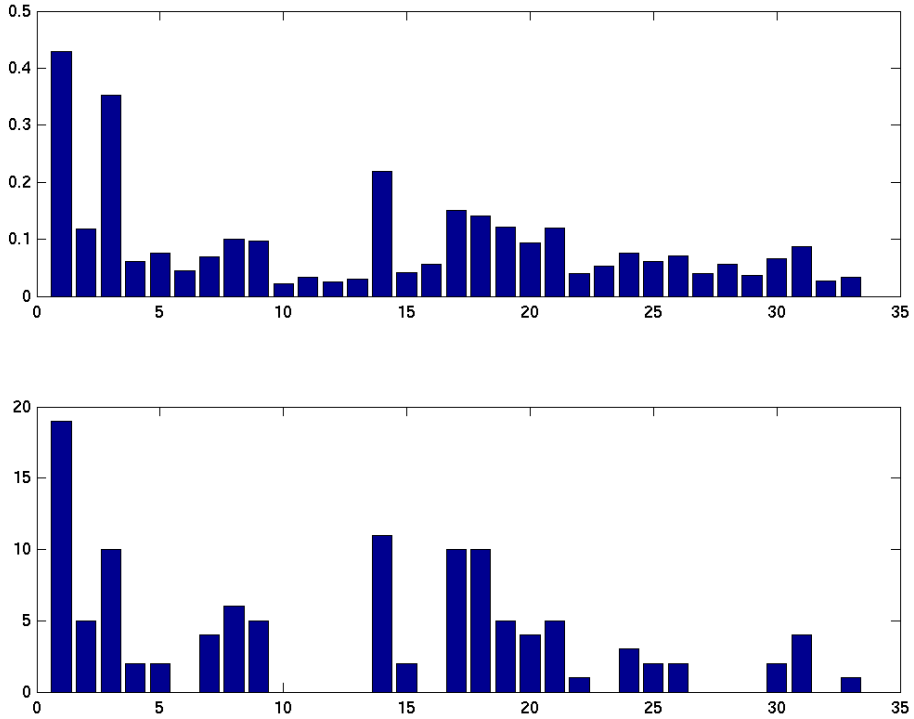

Figure 2: Two illustrations of feature importance. The top sub-figure shows the absolute values for the LDA projection vectors $\mathbf{w}_i$ averaged over users i. The bottom sub-figure shows for how many users the particular feature was within the best 5 features. The numbering of the features corresponds to the list of features in Table 1. In both figures the height of the bar corresponds to the importance of the feature, and both measures seem to be give roughly comparable results. See text for analysis of the results.

then study, for each feature, for how many users it was ranked amongst the top 5 features. This measure tries to extract most reliable features that work for all users and search tasks.

The results are illustrated in Figure 2, and Table 2 collects in a single list all features selected in the top 5 features for at least 5 users. The most important features are closely related to the viewing time of the image. Feature 1 measures exactly the time the image was viewed, whereas feature 2 indicates how many measurements not mapped to fixations were observed. Also features 17 (number of fixations) and 19 (total length of fixations) on the fixation side are tightly connected to viewing time, since most of the time is spent during fixations.

However, there is also evidence of the viewing pattern playing a role. The coverage features 8 and 9 both appear on the list, and while 8 is strongly correlated with the total viewing time the normalized coverage feature 9 is clearly indicative of the pattern within image being relevant for analysis. Also features 3 (ratio of raw measurements within and outside fixations) and 18 (average fixation length) are related to how the image is viewed.

The remaining two features are also promising. Feature 21 essentially measures how many times the image was viewed during scanning of the page, and hence gives evidence on the overall scanning pattern of the whole page being important, even though the search tasks were all so easy that most users were scanning most images just once. Finally, feature 14 measures the maximal pupil diameter while viewing the image. The feature had low weight

Feature	Name	Mean weight	Count
1	numMeasurements	0.43	19
14	pupil	0.22	11
3	ratioInsideOutside	0.35	10
17	numFix	0.15	10
18	meanFixLen	0.14	10
8	coverage	0.10	6
2	numOutsideFix	0.12	5
9	normCoverage	0.10	5
19	totalFixLen	0.12	5
21	nJumpsFix	0.12	5

Table 2: List of all features selected within the top 5 features for at least 5 users, showing both importance scores described in the text and ranked according to the Count column. See Table 1 for descriptions of the features and the text for analysis of the top features.

for some users, but was consistently ranked near the top for others. This suggests that the feature is highly user-specific.

In summary, very simple features essentially measuring how long the user was viewing the page seem to be highly informative of the relevance. Such features are easy to measure and robust, to the degree that some earlier works have used viewing time as the only feature [10]. At the same time, the importance of features that measure how the eye movements progress within and between images indicates that there is more information in the actual trajectory.

6 Classification accuracy

We consider two different methods for inferring relevance feedback from the gaze pattern. First, we use LDA as a low-complexity baseline algorithm readily available for real-time use. The classification task is solved by assuming the classes follow normal distributions in the projection space, and assigning each validation sample to the class with the highest likelihood.

The other algorithm for inferring the image relevance is parsimonious kernel Fisher discriminant analysis (kFDA) [4], a kernel-based extension of LDA. The algorithm utilizes a connection between Fisher discriminant analysis and least-squares problem. The complexity is controlled by L_q penalty function where $0 < q \leq 1$. This penalty function is well-known to have a sparsity-inducing property, and it leads to a non-smooth formulation. The problem is solved by the majorize-minimize principle, which gives a very simple iterative algorithm. See [4] for details of the algorithm.

In this report we consider only user-specific models. For each user separately, we use a leave-one-out validation procedure to measure the accuracy of predicting relevant images. Out of the 20 pages 19 are used for learning the classifier, and the performance is measured on the remaining page. The overall results are averaged over 20 runs, each having a separate page as test data. The kernel and regularization parameters of kFDA were chosen separately based on the selected performance criteria for each user, using further leave-one-out validation within each of the training sets of 19 pages. Here, we use the L_1 penalty function to induce parsimonious solutions. For both methods, images with no raw measurements (5.8% of all pairs of users and images) were excluded from the training data and considered non-relevant in the test data.

Three different measures were used to evaluate the accuracy. First, classification accuracy measures directly the ratio of correct labelings as relevant or non-relevant. Second, a classical information retrieval measure of area under the receiver operating characteristics (ROC) curve

measures how close to the top true relevant images are if the predictions of the classifier are ranked according to their predicted probability of being relevant. Finally, the ratio of how often a relevant image was ranked the highest is presented, to provide information on the accuracy of retrieving a smaller subset of relevant images. The classical retrieval measure of average precision was left out due to small number of images per page; the standard interpolated precision measurements do not work well with just 15 images per page or with just a few relevant images.

Table 3 shows area under curve (AUC) results for LDA using four different features sets (only raw features, only fixation features, all features, and the collection of top features presented in Table 2), and for kFDA using all features. The scores are shown separately for each user. kFDA is usually slightly better than LDA, but the difference is relatively small for most users. For all users the best accuracy is higher than random guessing (random ordering gives AUC score of 0.5 by definition), which implies that it is possible to gain information on the relevance based on the gaze trajectory alone. We also observe that, on average, the results obtained with the set of features chosen in the previous section are the best, but there are also users for which other feature collections provide better results. This indicates that the collection of best features is not a universal property of the users.

Table 4 is a similar table showing for how many pages the image estimated to most likely be relevant was correctly labeled. Again both methods are considerably better than random guessing (around 33% of images were on average judged to be relevant, and hence random guessing gives a bit less than 7 correct predictions), but now kFDA clearly outperforms LDA for all users. This indicates that even though kFDA was not able to improve AUC score dramatically, it is considerably more accurate in predicting the most relevant images.

The classification accuracy results are not shown in detail, but average scores for both methods and the three performance criteria are collected in Table 5, together with results for the baseline of ranking the images randomly.

In conclusion, it is definitely possible to infer information on relevance of the images from eye movements alone, already with fairly simple methods. This result is comparable to the one obtained by [8] with a simpler interface. The prediction accuracy is particularly high for the images ranked very high in relevance order, which hints at a retrieval system that uses only the most reliable estimates for feedback while ignoring (or using explicit feedback for) images with more uncertain estimates.

The relatively modest increase in AUC scores is likely to be a result of too simple feature representations. It seems that simple averaged features for the whole images are not sufficiently good representations of the trajectory, preventing accurate analysis of borderline cases. Instead, models that subdivide the image into smaller elements and build trajectory models for the traversal between those elements are needed. Sub-image level features will be considered in Task 2.2, and we intend to briefly return to verifying their benefit in the full-image relevance prediction task after the new trajectory representations have been developed.

7 Demonstrator for relevance prediction from a scan pattern

The Deliverable 2.1.2 *Demonstrator for relevance prediction from a scan pattern* implements a platform for testing other classification methods in the relevance prediction task. The platform is written in Matlab, and is accompanied with the feature representations for the images. The demonstrator implements the straightforward way of inferring the relevance with the LDA classifier, including code for computing the accuracy measures used in the experiments, and hence provides a readily applicable testbed that enables testing any classification algorithm on the same data. Results obtained with other classifiers can be directly compared to the results shown in Tables 3-5.

		LDA				
User	Search term	Raw	Fixation	All	Top	kFDA
1	Bicycle	0.72	0.74	0.74	0.75	**0.79**
2	Bicycle	0.60	0.62	0.61	0.62	**0.65**
3	Bicycle	0.69	0.63	0.67	0.71	**0.73**
4	Bicycle	0.77	0.74	0.77	0.77	**0.78**
5	Bicycle	0.66	0.60	0.64	0.65	**0.67**
6	Bicycle	0.76	0.77	0.75	**0.80**	0.77
7	Bicycle	0.72	0.68	0.71	0.75	**0.76**
8	Bicycle	0.55	0.54	0.52	0.55	**0.60**
9	Horse	0.69	0.69	0.69	**0.70**	0.70
10	Horse	0.60	0.64	0.64	**0.66**	0.65
11	Horse	0.49	0.51	0.47	0.55	**0.58**
12	Horse	0.72	0.75	0.73	0.75	**0.75**
13	Horse	0.69	0.68	0.67	0.72	**0.75**
14	Horse	**0.74**	0.66	0.69	0.74	0.73
15	Horse	0.71	0.64	0.68	0.70	**0.74**
16	Transport	0.68	0.70	0.72	0.72	**0.72**
17	Transport	**0.78**	0.69	0.73	0.77	0.77
18	Transport	0.49	0.51	0.47	0.49	**0.55**
19	Transport	**0.68**	0.63	0.66	0.66	0.67
20	Transport	0.70	0.67	0.69	**0.73**	0.68
21	Transport	**0.63**	0.50	0.57	0.57	0.61
22	Transport	0.69	0.66	**0.71**	0.69	0.71
23	Transport	0.67	0.69	0.71	0.71	**0.73**
Mean	-	0.67	0.65	0.66	0.69	**0.70**
Std	-	0.08	0.08	0.08	0.08	0.07

Table 3: The table shows user-specific area under the ROC curve scores, averaged over 20 test pages. The first 4 result columns corresponds to LDA classifier with four different feature sets as described in the text, and the last column shows the score for parsimonious kernel Fisher discriminant analysis (kFDA). For most users kFDA obtains the highest score, shown in boldface. A random classifier would obtain a score of 0.5, and all users clearly surpass this threshold, showing that it is possible to infer relevance of images based on the gaze trajectory. The difference between all models and the random baseline is statistically significant (paired Wilcoxon test, p-values below 10^{-5}), and kFDA is significantly better than the best LDA variant (p-value 0.01) despite modest increase in absolute figures. Note that users 18-23 were measured with Tobii X120, whereas the rest were measured with Tobii 1750. The measurement equipment does not seem to have impact on the feasibility of inferring the relevance.

		LDA				
User	Search term	Raw	Fixation	All	Top	kFDA
1	Bicycle	10	14	13	12	**17**
2	Bicycle	7	8	9	6	**12**
3	Bicycle	11	10	10	11	**13**
4	Bicycle	9	12	11	10	**15**
5	Bicycle	9	9	10	9	**12**
6	Bicycle	11	**14**	**14**	13	**14**
7	Bicycle	12	9	12	13	**15**
8	Bicycle	7	5	4	4	**11**
9	Horse	13	**15**	14	**15**	**15**
10	Horse	9	12	11	11	**13**
11	Horse	5	8	4	8	**11**
12	Horse	12	11	12	10	**14**
13	Horse	10	10	6	12	**17**
14	Horse	9	10	9	13	**14**
15	Horse	10	9	10	10	**13**
16	Transport	13	12	10	13	**18**
17	Transport	14	13	12	12	**15**
18	Transport	4	4	8	7	**14**
19	Transport	10	11	10	11	**15**
20	Transport	11	11	12	9	**13**
21	Transport	13	10	13	12	**17**
22	Transport	12	13	15	15	**17**
23	Transport	10	12	14	9	**15**
Mean	-	10.04	10.52	10.57	10.65	**14.35**
Std	-	2.53	2.69	2.98	2.72	1.97

Table 4: The table shows for how many pages (out of 20) the image with the highest relevance score was relevant. That is, the results indicate how reliable the best predictions of the models are. The first 4 result columns corresponds to LDA classifier with four different feature sets as described in the text, and the last column shows the score for parsimonious kernel Fisher discriminant analysis (kFDA). Boldface font indicates highest score, which is always obtained with kFDA. kFDA is significantly better than LDA, and all methods are significantly better than random guessing (paired Wilcoxon test, all p-values below 10^{-4}).

	LDA					
Measure	Raw	Fixation	All	Top	kFDA	Random
AUC	0.67	0.65	0.66	0.69	0.70	0.50
Classification	66.96	67.45	67.29	67.70	70.80	58.75
Top prediction	10.04	10.52	10.57	10.65	14.35	6.65

Table 5: The three quality measures used in the experiments, averaged over all users. kFDA uses all features, and the parameters of the model were validated separately for each performance measure, whereas for LDA the results are shown for four different feature sets. kFDA outperforms LDA on all measures, and both methods clearly surpass random baseline. Classification accuracies are shown in percentages, whereas top prediction is the count of pages (out of 20) with correct label on the image predicted most likely to be relevant. The column of random baseline corresponds to a method that orders the images in a random order, but uses the correct ratio of relevant and non-relevant images for classification.

8 Conclusions

The aim of Task 2.1 was to produce information on the feasibility and accuracy of inferring relevance feedback based on eye movement trajectories. Both [8] and the additional experiments reported in this deliverable show that it is possible to infer information about relevance feedback based on feature representations that do not consider the image content at all. The accuracy is consistently above what would be obtained by random guessing, but it remains an open question whether this level of accuracy is sufficient for practical retrieval systems. The accuracy is higher with the simplified search interface used by [8], which suggests that showing a somewhat smaller number of relatively large images on a display makes eye-movement based inference easier.

The results of Task 2.1 will guide the progress of the rest of WP 2. According to Description of Work, the work will be continued by Task 2.2 *Relevance of parts of an image from the viewing pattern* if eye movements prove to be sufficiently accurate for inferring relevance, and otherwise by Task 2.3 *Data fusion*. As shown in this report, it is clearly possible to obtain information on the relevance based on eye movements alone. The classification accuracy of the better method (kFDA) was on average 12% above the random baseline (70.8% vs 58.8%), and the image estimated to be the most relevant was correct twice as often with kFDA than by random guessing (14 vs 7 out of 20 pages). These figures compare favorably to the 10% threshold stated in DoW, and especially the latter show that at least partial relevance feedback (the most relevant image) can be obtained with high accuracy. As a conclusion, we will continue working for Task 2.2, and will move to considering fusion with other data sources (Task 2.3) after that.

During the task it was, however, observed that there is not much to be gained by applying more complex classifiers for feature representations that only consider full images. The accuracy of simple machine learning methods such as LDA is already relatively good, and the representations are too simple to reveal the full strength of more advanced methods like discriminative Hidden Markov models [14]. The kernel method considered in this report was clearly better for inferring the most relevant image, but overall did not improve the results considerably, most likely because the feature representations were too simple to reveal differences between borderline cases. It has become apparent that a method inferring the relevance of full images should already consider sub-image level elements like local interest-points [9] or image saliency [5] for maximal accuracy.

Task 2.2, which seeks to infer relevance of parts of images, will necessarily require sub-image level features, and the first part of that task is to find such representations for the eye trajectory. It is likely that similar features will be useful also for predicting the relevance of the whole images, and hence we intend to quickly re-visit the problem studied in Task 2.1 after the feature representations used in Task 2.2 are defined, strengthening the link between the Tasks of WP 2.

Acknowledgements

We wish to thank Prof. Peter Auer of University of Leoben for his valuable contributions in the writing of this report and for comments on its draft versions.

References

[1] Georg Buscher, Andreas Dengel, and Ludger van Elst. Eye movements as implicit relevance feedback. In *CHI '08: CHI '08 extended abstracts on Human Factors in Computing Systems*, pages 2991–2996, New York, NY, USA, 2008. ACM.

[2] R.I. Hammoud, editor. *Passive Eye Monitoring: Algorithms, Applications and Experiments*. Springer, Berlin, 2008.

[3] David R. Hardoon, John Shawe-Taylor, Antti Ajanki, Kai Puolamäki, and Samuel Kaski. Information retrieval by inferring implicit queries from eye movements. In Marina Meila and Xiaotong Shen, editors, *Proceedings of AISTATS 2007, the 11th International Conference on Artificial Intelligence and Statistics*. Omnipress, 2007. Proceedings on CD and at `http://www.stat.umn.edu/~aistat/proceedings/start.htm`.

[4] R.F. Harrison and K. Pasupa. A simple iterative algorithm for parsimonious binary kernel fisher discrimination. *Pattern Analysis & Applications*, 2008. In press.

[5] L. Itti and C. Koch. Computational modelling of visual attention. *Nature Reviews Neuroscience*, 2(3):194–203, 2001.

[6] Samuel Kaski and Jaakko Peltonen. Learning from relevant tasks only. In Joost N. Kok, Jacek Koronacki, Ramon Lopez de Mantaras, Stan Matwin, Dunja Mladenič, and Andrzej Skowron, editors, *Machine Learning: ECML 2007*, pages 608–615. Springer, Berlin Heidelberg, 2007.

[7] W. Kienzle, F.A. Wichmann, B. Schölkopf, and M.O. Franz. A nonparametric approach to bottom-up visual saliency. In *Advances in Neural Information Processing Systems 19*, pages 689–698. MIT Press, 2007.

[8] Arto Klami, Craig Saunders, Teófilo de Campos, and Samuel Kaski. Can relevance of images be inferred from eye movements? In *MIR '08: Proceeding of the 1st ACM International Conference on Multimedia Information Retrieval*, pages 134–140. ACM, New York, NY, USA, 2008.

[9] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J.G. Matas, F. Schaffalitzky, T. Kadir, and L.J. Van Gool. A comparison of affine region detectors. *International Journal of Computer Vision*, 65(1-2):43–72, November 2005.

[10] Oyewole Oyekoya and Fred Stentiford. Perceptual image retrieval using eye movements. In N. Zheng, X. Jiang, and X. Lan, editors, *Proceedings of the International Workshop on Intelligence Computing in Pattern Analysis/Synthesis 2006*, pages 281–289, 2006.

[11] Kai Puolamäki, Antti Ajanki, and Samuel Kaski. Learning to learn implicit queries from gaze patterns. In Andrew McCallum and Sam Roweis, editors, *Proceedings of ICML 2008, Twenty-Fifth International Conference on Machine Learning*, pages 760–767, Madison, WI, 2008.

[12] K. Rayner. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3):372–422, 1998.

[13] Jarkko Salojärvi, Ilpo Kojo, Jaana Simola, and Samuel 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.

[14] Jarkko Salojärvi, Kai Puolamäki, and Samuel Kaski. Implicit relevance feedback from eye movements. In Włodzisław Duch, Janusz Kacprzyk, Erkki Oja, and Sławomir Zadrożny, editors, *Artificial Neural Networks: Biological Inspirations — ICANN 2005*, volume I, pages 513–518, Berlin, 2005. Springer.

[15] Jarkko Salojärvi, Kai Puolamäki, Jaana Simola, Lauri Kovanen, Ilpo Kojo, and Samuel Kaski. Inferring relevance from eye movements: Feature extraction. Publications in Computer and Information Science A82, Helsinki University of Technology, Espoo, Finland, 2005.