Image Ranking with Implicit Feedback from Eye Movements

David R. Hardoon*
Data Mining Department
Institute for Infocomm Research (I^2R)

Kitsuchart Pasupa†
School of Electronics & Computer Science
University of Southampton

Abstract

In order to help users navigate an image search system, one could provide explicit information on a small set of images as to which of them are relevant or not to their task. These rankings are learned in order to present a user with a new set of images that are relevant to their task. Requiring such explicit information may not be feasible in a number of cases, we consider the setting where the user provides implicit feedback, eye movements, to assist when performing such a task. This paper explores the idea of implicitly incorporating eye movement features in an image ranking task where only images are available during testing. Previous work had demonstrated that combining eye movement and image features improved on the retrieval accuracy when compared to using each of the sources independently. Despite these encouraging results the proposed approach is unrealistic as no eye movements will be presented a-priori for new images (i.e. only after the ranked images are presented would one be able to measure a users eye movements on them). We propose a novel search methodology which combines image features together with implicit feedback from users eye movements in a tensor ranking Support Vector Machine and show that it is possible to extract the individual source-specific weight vectors. Furthermore, we demonstrate that the decomposed image weight vector is able to construct a new image-based semantic space that outperforms the retrieval accuracy than when solely using the image-features.

CR Categories: G.3 [Probability and Statistics]: Multivariate Statistics—; H.3.3 [Information Search and Retrieval]: Retrieval models—Relevance feedback

Keywords: Image Retrieval, Implicit Feedback, Tensor, Ranking, Support Vector Machine

1 Introduction

In recent years large digital image collections have been created in numerous areas, examples of these include the commercial, academic, and medical domains. Furthermore, these databases also include the digitisation of analogue photographs, paintings and drawings. Conventionally, the images collected are manually tagged with various descriptors to allow retrieval to be performed over the annotated words. However, the process of manually tagging images is an extremely laborious, time consuming and an expensive procedure. Moreover, it is far from an ideal situation as both formulating an initial query and navigating the large number of retrieved hits is a difficult. One image retrieval methodology which attempts to address these issues, and has been a research topic since the early 1990's, is the so-called "Content-Based Image Retrieval"(CBIR). The search of a CBIR system is analysed from the actual content of the image which may includes colour, shape, and texture rather than using a textual annotation associated (if at all) with the image.

Relevance feedback, which is explicitly provided by the user while performing a search query on the quality of the retrieved images, has shown to be able to improve on the performance of CBIR systems, as it is able to handle the large variability in semantic interpretation of images across users. Relevance feedback will iteratively guide the system to retrieve images the user is genuinely interested in. Many systems rely on an explicit feedback mechanism, where the user explicitly indicates which images are relevant for their search query and which ones are not. One can then use a machine learning algorithm to try and present a new set of images to the users which are more relevant - thus helping them navigate the large number of hits. An example of such systems is PicSOM [Laaksonen et al. 2000]. However, providing explicit feedback is also a laborious process as it requires continues user response. Alternatively, it is possible to use implicit feedback to infer relevance of images. Examples of implicit feedback are eye movements, mouse pointer movements, blood pressure, gestures, etc. In other words, user responses that are implicitly related to the task performed.

In this study we explore the use of eye movements as a particular source of implicit feedback to assist a user when performing such a task (i.e. image retrieval). Eye movements can be treated as an implicit relevance feedback when the user is not consciously aware of their eye movements being tracked. Eye movement as implicit feedback has recently been used in the image retrieval setting [Oyekoya and Stentiford 2007; Klami et al. 2008; Pasupa et al. 2009]. [Oyekoya and Stentiford 2007; Klami et al. 2008] used eye movements to infer a binary judgement of relevance while [Pasupa et al. 2009] makes the task more complex and realistic for search-based task by asking the user to give multiple judgement of relevance. Furthermore, earlier studies of Hardoon et al. [2007] and Ajanki et al. [2009] explored the problem of where an implicit information retrieval query is inferred from eye movements measured during a reading task. The result of their empirical study is that it is possible to learn the implicit query from a small set of read documents, such that relevance predictions for a large set of unseen documents are ranked better than by random guessing. More recently, Pasupa et al. [2009] demonstrated that ranking of images can be inferred from eye movements using Ranking Support Vector Machine (Ranking SVM). Their experiment shows that the performance of the search can be improved when simple images features namely histograms are fused with the eye movement features.

Despite Pasupa et al.'s [2009] encouraging results, their proposed approach is largely unrealistic as they combine image and eye features for both training and testing. Whereas in a real scenario no eye movements will be presented a-priori for new images. In other words, only after the ranked images are presented to a user, would one be able to measure the users eye movements on them. Therefore, we propose a novel search methodology which combines im-

*e-mail: drhardoon@i2r.a-star.edu.sg
†e-mail: kp2@ecs.soton.ac.uk

age features together with implicit feedback from users' eye movements during training, such that we are able to rank new images with only using image features. We believe it is indeed more realistic to have images and eye-movements during the training phase as these could be acquired deliberately to train up such a system.

For this purpose, we propose using tensor kernels in the ranking SVM framework. Tensors have been used in the machine learning literature as a means of predicting edges in a protein interaction or co-complex network by using the tensor product transformation to derive a kernel on protein pairs from a kernel on individual proteins [Ben-Hur and Noble 2005; Martin et al. 2005; Qiu and Noble 2008]. In this study we use the tensor product to constructed a joined semantics space by combining eye movements and image features. Furthermore, we continue to show that the combined learnt semantic space can be efficiently decomposed into its contributing sources (i.e. images and eye movements), which in turn can be used independently.

The paper is organised as follows. In Section 2 we give a brief introduction to the ranking SVM methodology and continue to develop in Section 3 our proposed tensor ranking SVM and the efficient decomposition of the joint semantic space into the individual sources. In Section 5 we give our experimental set up whereas in Section 6 we discusses the feature extraction and representation of the images and eye movements. In section 7 we bring forward our experiments on page ranking for individual users as well as a feasibility study on user generalisation. Finally, we conclude our study with discussion on our present methodology and results in Section 8.

2 Ranking SVM

The Ranking Support Vector Machine (SVM) was proposed by Joachims [2002] which was adapted from ordinal regression [Herbrich et al. 2000]. It is a pair-wise approach where the solution is a binary classification problem. Let $\mathbf{x}_i$ denote some feature vector and let r_i denote the ranking assigned to $\mathbf{x}_i$. If $r_1 \succ r_2$, it means that $\mathbf{x}_1$ is more relevance than $\mathbf{x}_2$. Consider a linear ranking function,

$$\mathbf{x}_i \succ \mathbf{x}_j \iff \langle \boldsymbol{w}, \mathbf{x}_i \rangle - \langle \boldsymbol{w}, \mathbf{x}_j \rangle > 0,$$

where $\boldsymbol{w}$ is a weight vector and $\langle \cdot, \cdot \rangle$ denotes dot product between vectors. This can be placed in a binary SVM classification framework where let c_k be the new label indicating the quality of k^{th} rank pair,

$$\langle \boldsymbol{w}, \mathbf{x}_i - \mathbf{x}_j \rangle = \begin{cases} c_k = +1 & \text{if } r_i \succ r_j \\ c_k = -1 & \text{if } r_j \succ r_i \end{cases}, \qquad (1)$$

which can be solved by the following optimization problem,

$$\min \frac{1}{2} \langle \boldsymbol{w}, \boldsymbol{w} \rangle + C \sum_k \xi_k \qquad (2)$$

subject to the following constrains:

$$\begin{aligned} \forall (i,j) \in \boldsymbol{r}^{(k)} \quad &: \quad c_k(\langle \boldsymbol{w}, \mathbf{x}_i - \mathbf{x}_j \rangle + b) \geq 1 - \xi_k \\ \forall (k) \quad &: \quad \xi_k \geq 0 \end{aligned}$$

where $\boldsymbol{r}^{(k)} = [r_1, r_2, \ldots, r_t]$ for t rank values, furthermore C is a hyper-parameter which allows trade-off between margin size and training error, and ξ_k is training error. Alternatively, we are represent the ranking SVM as a vanilla SVM where we re-represent our samples as

$$\phi(\mathbf{x})_k = \mathbf{x}_i - \mathbf{x}_j$$

with label c_k and m being the total number of new samples. Finally, we quote from Cristianini and Shawe-Taylor [2000] the general dual SVM optimisation as

$$\max_{\boldsymbol{\alpha}} W(\boldsymbol{\alpha}) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j c_i c_j \kappa(\mathbf{x}_i, \mathbf{x}_j) \qquad (3)$$

$$\text{subject to } \textstyle\sum_{i=1}^{m} \alpha_i c_i = 0 \text{ and } \alpha_i \geq 0 \;\; i = 1, \ldots, m,$$

where we again use c_i to represent the label and $\kappa(\mathbf{x}_i, \mathbf{x}_j)$ to be the kernel function between $\phi(\mathbf{x})_i$ and $\phi(\mathbf{x})_j$, where $\phi(\cdot)$ is a mapping from X (or Y) to an (inner product) feature space $\mathcal{F}$.

3 Tensor Ranking SVM

In the following section we propose to construct a tensor kernel on the ranked image and eye movements features, i.e. following equation (2), to then to train an SVM. Therefore, let $\mathbf{X} \in \mathbb{R}^{n \times m}$ and $\mathbf{Y} \in \mathbb{R}^{\ell \times m}$ be the matrix of sample vectors, $\mathbf{x}$ and $\boldsymbol{y}$, for the image and eye movements respectively, where n is the number of image features and ℓ is the number of eye movement features and m are the total number of samples. We continue to define K^x, K^y as the kernel matrices for the ranked images and eye movements respectively. In our experiments we use linear kernels, i.e. $K^x = X'X$ and $K^y = Y'Y$. The resulting kernel matrix of the tensor $T = X \circ Y$ can be expressed as pair-wise product (see [Pulmannová 2004] for more details)

$$\bar{K}_{ij} = (T'T)_{ij} = K^x_{ij} K^y_{ij}.$$

We use $\bar{K}$ in conjunction with the vanilla SVM formulation as given in equation (3). Whereas the set up and training are straight forward, the underlying problem is that for testing we do not have the eye movements. Therefore we propose to decompose the resulting weight matrix from its corresponding image and eye components such that each can be used independently.

The goal is to decompose the weight matrix W given by a dual representation

$$W = \sum_{i}^{m} \alpha_i c_i \phi_x(\mathbf{x}_i) \circ \phi_y(\mathbf{y}_i)$$

without accessing the feature space. Given the paired samples $\mathbf{x}, \mathbf{y}$ the decision function in equation is

$$\begin{aligned} f(\mathbf{x}, \mathbf{y}) &= W \circ \phi_x(\mathbf{x}) \phi_y(\mathbf{y})' \qquad (4) \\ &= \sum_{i=1}^{m} \alpha_i c_i \kappa_x(\mathbf{x}_i, \mathbf{x}) \kappa_y(\mathbf{y}_i, \mathbf{y}). \end{aligned}$$

4 Decomposition

The resulting decision function in equation (4) requires both image and eye movement $(\mathbf{x}_i, \mathbf{y}_i)$ data for training and testing. We want to be able to test our model only using the image data. Therefore, we want to decompose the weight matrix (again without accessing the feature space) into a sum of tensor products of corresponding weight components for the images and eye movements (this procedure is given in detail in [Hardoon and Shawe-Taylor 2010])

$$W \approx W^T = \sum_{t=1}^{T} \mathbf{w}_x^t \mathbf{w}_y^{t\prime}, \qquad (5)$$

such that the weights are a linear combination of the data, i.e. $\mathbf{w}_x^t = \sum_{i=1}^{m} \beta_i^t \phi_x(\mathbf{x}_i)$ and $\mathbf{w}_y^t = \sum_{i=1}^{m} \gamma_i^t \phi_y(\mathbf{y}_i)$ where $\boldsymbol{\beta}^t, \boldsymbol{\gamma}^t$ are the

dual variables of $\mathbf{w}_x^t, \mathbf{w}_y^t$. We proceed to define our decomposition procedure such that we do not need to compute the (potentially non-linear) feature projection ϕ. We compute

$$WW' = \sum_{i,j}^{m} \alpha_i \alpha_j c_i c_j \kappa_y(\mathbf{y}_i, \mathbf{y}_j) \phi_x(\mathbf{x}_i) \phi_x(\mathbf{x}_j)' \qquad (6)$$

and are able to express $K^y = (\kappa_y(\mathbf{y}_i, \mathbf{y}_j))_{i,j=1}^m = \sum_{k=1}^K \lambda_k \mathbf{u}^k \mathbf{u}^{k\prime} = U\Lambda U'$, where $U = (\mathbf{u}_1, \ldots, \mathbf{u}_K)$ by performing an eigenvalue decomposition of the kernel matrix K^y with entries $K_{ij}^y = \kappa_y(\mathbf{y}_i, \mathbf{y}_j)$. Substituting back into equation (6) gives

$$WW' = \sum_{k}^{K} \lambda_k \sum_{i,j}^{m} \alpha_i \alpha_j c_i c_j \mathbf{u}_i^k \mathbf{u}_j^{k\prime} \phi_x(\mathbf{x}_i) \phi_x(\mathbf{x}_j)'.$$

Letting $\mathbf{h}_k = \sum_{i=1}^m \alpha_i c_i \mathbf{u}_i^k \phi_x(\mathbf{x}_i)$ we have $WW' = \sum_k^K \lambda_k \mathbf{h}_k \mathbf{h}_k' = HH'$ where $H = \left(\sqrt{\lambda_1}\mathbf{h}_1, \ldots, \sqrt{\lambda_K}\mathbf{h}_K\right)$. We would like to find the singular value decomposition of $H = V\Upsilon Z'$. Consider for $A = diag(\boldsymbol{\alpha})$ and $C = diag(\mathbf{c})$ we have

$$\begin{aligned} \left[H'H\right]_{k\ell} &= \sqrt{\lambda_k \lambda_\ell} \sum_{ij} \alpha_i \alpha_j c_i c_j \mathbf{u}_i^k \mathbf{u}_j^\ell \kappa_x(\mathbf{x}_i, \mathbf{x}_j) \\ &= \left[\left(CAU\Lambda^{\frac{1}{2}}\right)' K^x \left(CAU\Lambda^{\frac{1}{2}}\right)\right]_{k\ell}, \end{aligned}$$

which is computable without accessing the feature space. Performing an eigenvalue decomposition on $H'H$ we have

$$H'H = Z\Upsilon V' V \Upsilon Z' = Z\Upsilon^2 Z' \qquad (7)$$

with Υ a matrix with υ_t on the diagonal truncated after the J'th eigenvalue, which gives the dual representation of $\mathbf{v}_t = \frac{1}{\upsilon_t} H\mathbf{z}_t$ for $t = 1, \ldots, T$, and since $H'H\mathbf{z}_t = \upsilon_t^2 \mathbf{z_t}$ we are able to verify that

$$WW'\mathbf{v}_t = HH'\mathbf{v}_t = \frac{1}{\upsilon_t} HH'H\mathbf{z}_t = \upsilon_t H\mathbf{z}_t = \upsilon_t^2 \mathbf{v}_t.$$

Restricting to the first T singular vectors allows us to express $W \approx W^T = \sum_{t=1}^T \mathbf{v}_t \left(W'\mathbf{v}_t\right)'$, which in turn results in

$$\mathbf{w}_x^t = \mathbf{v}_t = \frac{1}{\upsilon_t} H\mathbf{z}_t = \sum_{i=1}^{m} \beta_i^t \phi_x(\mathbf{x}_i),$$

where $\beta_i^t = \frac{1}{\upsilon_t} \alpha_i c_i \sum_{k=1}^T \sqrt{\lambda_k} \mathbf{z}_k^t u_i^k$. We can now also express

$$\mathbf{w}_y^t = W'\mathbf{v}_t = \frac{1}{\upsilon_t} W'H\mathbf{z}_t = \sum_{i=1}^{m} \gamma_i^t \phi_y(\mathbf{y}_i),$$

where $\gamma_i^t = \sum_{j=1}^m \alpha_i c_i \beta_j^t \kappa_x(\mathbf{x}_i, \mathbf{x}_j)$ are the dual variables of $\mathbf{w}_y^t$. We are therefore now able to decompose W into W_x, W_y without accessing the feature space giving us the desired result.

We are now able to compute, for a given t, the ranking scores in the linear discriminant analysis form $s = \mathbf{w}_x^t \phi(\hat{X}) = \sum_{i=1}^m \beta_i^t \kappa_x(\mathbf{x}_i, \hat{X})$ for new test images $\hat{X}$. These are in turn sorted in order of magnitude (importance). Equally, we can project our data into the new defined semantic space β where we train and test an SVM. i.e. we compute $\tilde{\phi}(\mathbf{x}) = K^x \beta$, for the training samples, and $\tilde{\phi}(\mathbf{x}_t) = K_t^x \beta$ for our test samples. We explore both these approaches in our experiments.

5 Experimental Setup

Our experimental set-up is as follows: Users are shown 10 images on a single page as a five by two (5x2) grid and are asked to rank the top five images in order of relevance to the topic of "Transport". This concept is deliberately slightly ambiguous given the context of images that were displayed. Each displayed page contained 1–3 clearly relevant images (e.g. a freight train, cargo ship or airliner), 2–3 either borderline or marginally relevant images (e.g. bicycle or baby carrier), and the rest are non-relevant images (e.g. images of people sitting at a dining room table, or a picture of a cat).

The experiment had 30 pages in total, each showing 10 images from the PASCAL Visual Objects Challenge 2007 database [Everingham et al.]. The interface consisted of selecting radio buttons (labelled 1^{st} to 5^{th} under each image) then clicking on next to retrieve the next page. This represents data for a ranking task where explicit ranks are given to compliment any implicit information contained in the eye movements. An example of each page is shown in figure 1.

The experiment was performed by six different users, with their eye movements recorded by a Tobii X120 eye tracker which was connected to a PC using a 19-inch monitor (resolution of 1280x1024). The eye tracker has approximately 0.5 degrees of accuracy with a sample rate of 120 Hz and used infrared lens to detect pupil centres and corneal reflection. The final data collected per user is illustrated in table 1. Any pages that contained less than five images with gaze points (for example due to the subject moving and the eye-tracker temporarily losing track of the subject's eyes) were discarded. Hence, only 29 and 20 pages were valid for users 4 and 5, respectively.

Table 1: *The data collected per user. *Pages with less than five images with gaze points were removed. Therefore users 4 and 5 only have 29 and 20 pages viewed respectively.*

User #	Pages Viewed
User 1	30
User 2	30
User 3	30
User 4*	29
User 5*	20
User 6	30

5.1 Performance Measure

We use the Normalised Discount Cumulative Gain (NDCG) [Järvelin and Kekäläinen 2000] as our performance metric, due to our task involving multiple ranks rather than a binary choice. NDCG measures the usefulness, or gain, of a retrieved item based on its position in the result list. NDCG is designed for tasks which have more than two levels of relevance judgement, and is defined as,

$$\mathrm{NDCG}_k(r) = \frac{1}{N_n} \sum_{i=1}^{k} D(r_i) \varphi(g_i)$$

with $D(r) = \frac{1}{\log_2(1+r)}$ and $\varphi(g) = 2^g - 1$, where for a given page r is rank position and k is a truncation level (position), N is a normalising constant which gives the perfect ranking (based on g_i) equal to one, and g_i is the categorical grade; e.g. grade is equal to 5 for the 1^{st} rank and 0 for the 6^{th}.

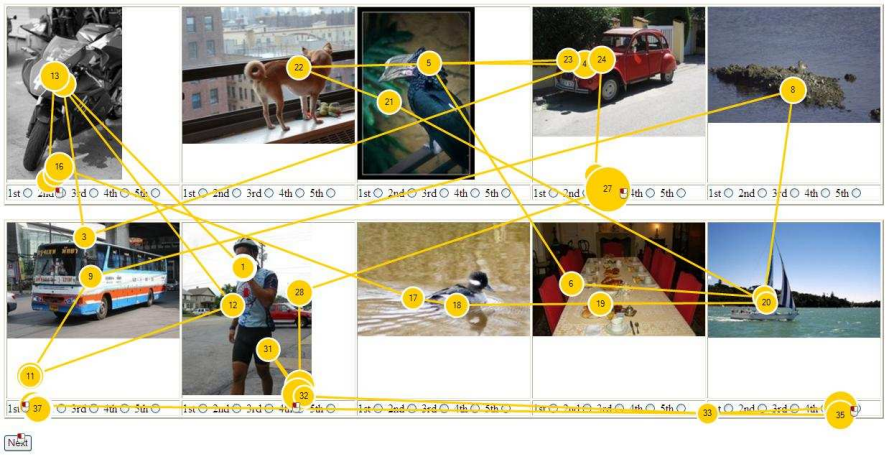


Figure 1: *An example illustrating the outlay of the interface displaying the 10 images with the overlaid eye movement measurements. The circles indicate fixations.*

6 Feature extraction

In the following experiments we use standard image histograms and features collected from eye-tracking. We compute a 256-bin grey scale histogram on the whole image as the feature representation. These features are intentionally kept relatively simple. Although, a possible extension of the current representation is to segment the image and only use regions that have gaze information. We intend to explore this extension in a future study.

The eye movement features are computed using only on the eye trajectory and locations of the images in the page. This type of features are general-purpose and are easily applicable to all application scenarios. The features are divided into two categories; the first category uses 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 in which a user maintains their gaze around a given point. These are important as most visual processing happens during fixations, due to blur and saccadic suppression during the rapid saccades between fixations (see, e.g. [Hammoud 2008]). Often visual attention features are based solely on fixations and the relation between them [Rayner 1998]. However, raw measurement data might be able to overcome possible problems caused by imperfect fixation detection.

In table 2 we list the candidate features we have considered. Most of the listed features are motivated by earlier studies in text retrieval [Salojärvi et al. 2005]. The features cover the three main types of information typically considered in reading studies: fixations, regressions (fixations to previously seen images), and re-fixations (multiple fixations within the same image). However, the features have been tailored to be more suitable for images, trying to include measures for things that are not relevant for text, such as the cover of the image. Similarly to the image features, the eye movement features are intentionally kept relatively simple with the intent that they are more likely to generalise over different users. Fixations were detected using the standard ClearView fixation filter provided with the Tobii eye-tracking software, with settings "radius 30 pixels, minimum duration 100 ms". These are also the settings recommended for media with mixed content[1].

[1]Tobii Technology, Ltd. Tobii Studio Help. url: http://studiohelp.tobii.com/StudioHelp_1.2/

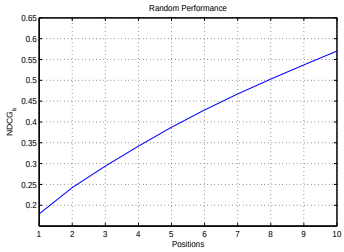


Figure 2: *NDCG performance for predicting random rankings.*

Some of the features are not invariant to the location of the image on the screen. For example, the typical pattern of moving from left to right means that the horizontal co-ordinate of the 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 normalised by removing the mean of all observations sharing the same position, and are marked in Table 2. Finally, each feature was normalised to have unit variance and zero mean.

7 Experiments

We evaluate two different scenarios for learning the ranking of image based on image and eye features; 1. Predicting rankings on a page given only other data from a single specific user. 2. A global model using data from other users to predict rankings for a new unseen user.

We compare our proposed tensor Ranking SVM algorithm which combines both information from eye movements and image histogram features to a Ranking SVM using histogram features and to a Ranking SVM using eye movements alone. We emphasis that training and testing a model using only eye movements is *not realistic* as there are no eye movements presented a-priori for new images, i.e. one can not test. This comparison provides us with

Table 2: *We list the eye movement features considered in this study. The first 16 features are computed from the raw data, whereas the remainder are based on pre-detected fixations. We point out to the reader that features number 2 and 3 use both types of data since they are based on raw measurements not belonging to fixations. All the features are computed separately for each image. Features marked with * are normalised for each image location.*

Number	Name	Description
		Raw data features
1	numMeasurements	total number of measurements
2	numOutsideFix	total number of 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 covered by measurements[1]
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 longer than 60 ms[2]
16*	nJumps2	number of breaks longer than 600 ms[2]
		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.
[2] A sequence of measurements outside the image occurring between two consecutive measurements within the image.
[3] A transition from one fixation to another.

a baseline as to how much it may be possible to improve on the performance using eye movements. Furthermore, we are unable to make direct comparison to [2009] as they had used an online learning algorithm with different image features.

In the experiments we use a linear kernel function. Although, it is possible to use a non-linear kernel on the eye movement features as this would not effect the decomposition for the image weights (assuming that $\phi_x(\mathbf{x}_i)$ are taken as the image features in equation (6)). In figure 2 we give the NDCG performance for predicting random ranking.

7.1 Page Generalisation

In the following section we focus on predicting rankings on a page given only other data from a single specific user (we repeat this for all users). We employ a leave-page-out routine where at each iteration a page, from a given user, is withheld for testing and the remaining pages, from the same user, are used for training.

We evaluate the proposed approach with the following four setting:

- $T1$: using the largest component of tensor decomposition in the form of a linear discriminator. We use the weight vector corresponding to the largest eigenvalue (as we have a t weights).
- $T2$: we project the image features into the learnt semantic space (i.e. the decomposition on the image source) and train and test within the projected space a secondary Ranking SVM (Ranking SVM).
- $T1^{all}$: similar to $T1$ although here we use all t weight vectors and take the mean value across as the final score.
- $T1^{opt}$: similar to $T1$ although here we use the n-largest components of the decomposition. i.e. we select n weight vectors to use and take the mean value across as the final score.

We use a leave-one-out cross-validation for $T1^{opt}$ to obtain the optimal model for the later case which are selected based on maximum average NDCG across 10 positions.

We plot the user specific leave-page-out NDCG performances in figure 3 where we are able to observe that $T2$ consistently outper-

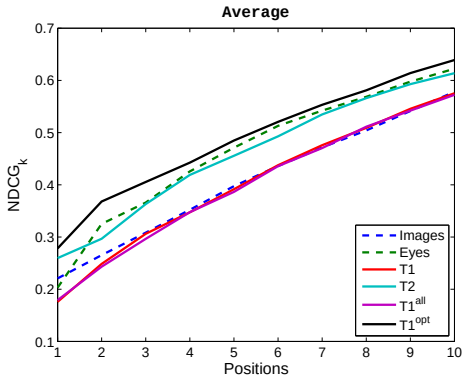


Figure 4: *Average NDCG performance across all users for predicting rankings on a page given only other data from a single specific user.*

forms the image feature Ranking SVM across all users, demonstrating that it is indeed possible to improve on the image ranking with the incorporation of eye movement features during training. Furthermore, it is interesting to observe that for certain users $T1^{opt}$ improves on the ranking performance, suggesting that there is an optimal combination of the decomposed features that may further improve on the results.

In figure 4 we plot the average performance across all users. The figure shows that $T1$ and $T1^{all}$ are slightly worse than using image histogram alone. However, when selecting using cross-validation the number of largest components in tensor decomposition, the performance of the classifier is improved and outperforms the Ranking SVM with eye movements. Furthermore, we are able to observe that we perform better than random (figure 2). Using classifier $T2$, the performance is improved above the Ranking SVM with image features and it is competitive with Ranking SVM with eye movements features.

7.2 User Generalisation

In the following section we focus on learning a global model using data from other users to predict rankings for a new unseen user. Although, as the experiment is set up such that all users view the same pages, we employ a leave-user-leave-page-out routine, i.e;

```
For all users
   Withhold data from user i
   For all pages
      Withhold page j from all users
      Train on all pages-j from all users - i

      Test on page j from user i
   Endfor
Endfor
```

Therefore we only use the users from table 1 who viewed the same number of pages, i.e. users 1, 2, 3 and 6, which we refer to henceforth as users 1-4.

We evaluate the proposed approach with the following two setting:

- $T1$: using the largest component of tensor decomposition in the form of a linear discriminator. We use the weight vector corresponding to the largest eigenvalue (as we have a t weights).
- $T2$: we project the image features into the learnt semantic space (i.e. the decomposition on the image source) and train and test within the projected space a secondary Ranking SVM.

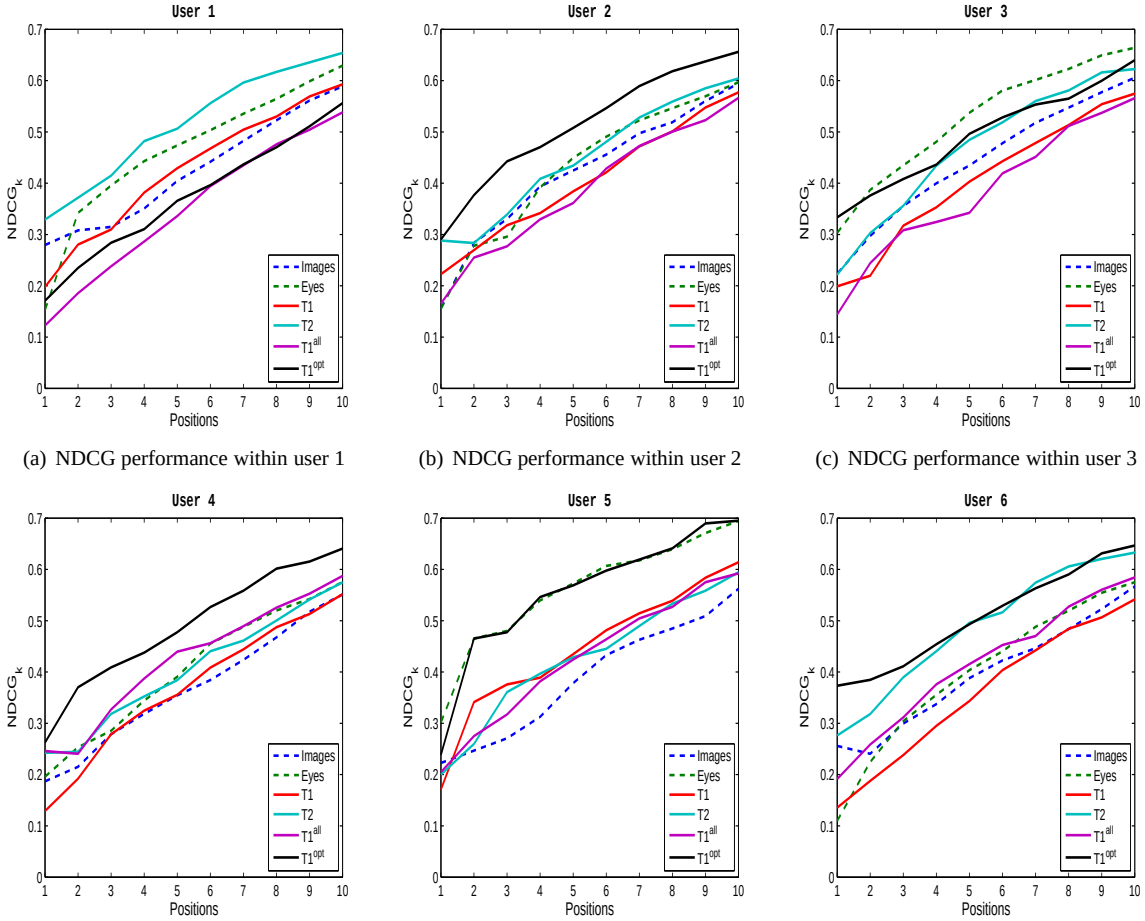
We plot in figure 5 the resulting NDCG performance for the leave-user-out routine. We are able to observe, with the exclusion of user 2 in figure 5(b), that $T2$ is able to outperform the Ranking SVM on image features. Indicating that it is possible to generalise our proposed approach across new unseen users. Furthermore, it is interesting to observe that $T2$ achieves a similar performance to that of a Ranking SVM trained and tested on the eye features. Finally, even though we do not improve when testing on data from user 2, we are able to observe that we perform as-good-as the baselines. In figure 5(e) we plot the average NDCG performance on the leave-user-out routine, demonstrating that on average we improve on the ranking of new images for new users and that we perform better than random (figure 2).

8 Discussion

Improving search and content based retrieval systems with implicit feedback is an attractive possibility given that a user is not required to explicitly provide information to then improve, and personalise, their search strategy. This, in turn, can render such a system more user-friendly and simple to use (at least from the users' perspective). Although, achieving such a goal is non-trivial as one needs to be able to combine the implicit feedback information into the search system in a manner that does not then require the implicit information for testing. In our study we focus on implicit feedback in the form of eye movements, as these are easily available and can be measured in a non-intrusive manner.

Previous studies [Hardoon et al. 2007; Ajanki et al. 2009] have shown the feasibility of such systems using eye moments for a textual search task. Demonstrating that it is indeed possible to 'enrich' a textual search with eye features. Their proposed approach is computationally complex since it requires the construction of a regression function on eye measurements on each word. This was not realistic in our setting. Furthermore, Pasupa et al. [2009] had extend the underlying methodology of using eye movement as implicit feedback to an image retrieval system, combining eye movements with image features to improve the ranking of retrieved images. Although, still, the proposed approach required eye features for the test images which would not be practical in a real system.

In this paper we present a novel search strategy for combining eye movements and image features with a tensor product kernel used in a ranking support vector machine framework. We continue to show that the joint learnt semantic space of eye and image features can be efficiently decomposed into its independent sources allowing us to further test or train only using images. We explored two different search scenarios for learning the ranking of images based on image and eye features. The first was predicting ranking on a page given only other data from a single specific user. This experiment was to test the fundamental question of whether eye movement are able to improve ranking for a user. Demonstrating that it was indeed possible to improve in the single subject setting, we then proceeded to our second setting where we constructed a global model across users in attempt to generalise on data from a new user. Again our results demonstrated that we are able to generalise out model to new users. Despite these promising results, it was also clear that using a single direction (weight vector) does not necessarily improve on the baseline result. Motivating the need for a more sophisticated combination of the resulting weights. This, as well as extending our experiment to a much larger number of users, will be addressed in a future study. Finally, we would also explore the notion of image segmentation and the use of more sophisticated image features that



(a) NDCG performance within user 1 (b) NDCG performance within user 2 (c) NDCG performance within user 3

(d) NDCG performance within user 4 (e) NDCG performance within user 5 (f) NDCG performance within user 6

Figure 3: *In the following sub-figures 3(a)-3(f) we illustrate the NDCG performance for each user in a leave-page-out routine, i.e. here we aim to generalise over new pages rather new users. We are able to observe that $T2$ and $T1^{opt}$ routinely outperform the ranking with only using image features. The 'Eyes' plot in all the figures demonstrates how the ranking (only using eye-movements) would perform if eye-features were indeed available a-priori for new images.*

are easily computable.

Acknowledgements

The authors would like to acknowledge financial support from the European Community's Seventh Framework Programme (FP7/2007–2013) under *grant agreement* n° 216529, Personal Information Navigator Adapting Through Viewing (PinView) project (http://www.pinview.eu). The authors would also like to thank Craig Saunders for data collection.

References

AJANKI, A., HARDOON, D. R., KASKI, S., PUOLAMÄKI, K., AND SHAWE-TAYLOR, J. 2009. Can eyes reveal interest? implicit queries from gaze patterns. *User Modeling and User-Adapted Interaction 19*, 4, 307–339.

BEN-HUR, A., AND NOBLE, W. S. 2005. Kernel methods for predicting protein-protein interactions. *Bioinformatics 21*, i38–i46.

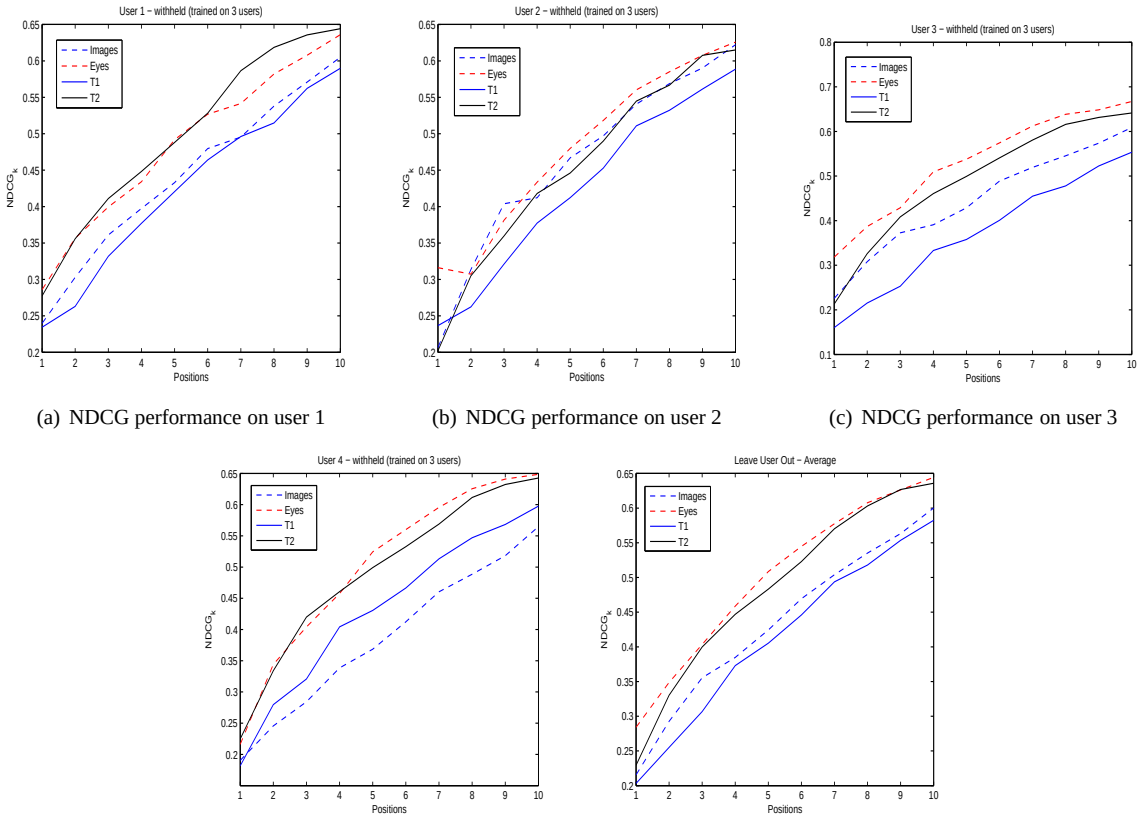
CRISTIANINI, N., AND SHAWE-TAYLOR, J. 2000. *An Introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press.

EVERINGHAM, M., VAN GOOL, L., WILLIAMS, C. K. I., WINN, J., AND ZISSERMAN, A. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

HAMMOUD, R. 2008. *Passive Eye Monitoring: Algorithms, Applications and Experiments*. Springer-Verlag.

HARDOON, D. R., AND SHAWE-TAYLOR, J. 2010. Decomposing the tensor kernel support vector machine for neuroscience data with structure labels. *Machine Learning Journal: Special Issue on Learning From Multiple Sources 79*, 1-2, 29–46.

HARDOON, D. R., AJANKI, A., PUOLAMÄKI, K., SHAWE-TAYLOR, J., AND KASKI, S. 2007. Information retrieval by inferring implicit queries from eye movements. In *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics (AISTATS)*, Electronic proceedings at www.stat.umn.edu/ aistat/proceedings/start.htm.



(a) NDCG performance on user 1

(b) NDCG performance on user 2

(c) NDCG performance on user 3

(d) NDCG performance on user 4

(e) Average NDCG performance

Figure 5: *In the following sub-figures 5(a)-5(d) we illustrate the NDCG performance in a leave-user-out (leave-page-out) routine. The average NDCG performance is given in sub-figure 5(e) where we are able to observe that $T2$ outperforms the ranking of only using image features. The 'Eyes' plot in all the figures demonstrates how the ranking (only using eye-movements) would perform if eye-features were indeed available a-priori for new images.*

HERBRICH, R., GRAEPEL, T., AND OBERMAYER, K. 2000. *Large margin rank boundaries for ordinal regression*. MIT Press, Cambridge, MA.

JÄRVELIN, K., AND KEKÄLÄINEN, J. 2000. IR evaluation methods for retrieving highly relevant documents. In *SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, New York, NY, USA, 41–48.

JOACHIMS, T. 2002. Optimizing search engines using clickthrough data. In *KDD '02: Proceedings of the 8th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM Press, New York, NY, USA, 133–142.

KLAMI, A., SAUNDERS, C., DE CAMPOS, T. E., AND KASKI, S. 2008. Can relevance of images be inferred from eye movements? In *MIR '08: Proceeding of the 1st ACM international conference on Multimedia information retrieval*, ACM, New York, NY, USA, 134–140.

LAAKSONEN, J., KOSKELA, M., LAAKSO, S., AND OJA, E. 2000. Picsom–content-based image retrieval with self-organizing maps. *Pattern Recognition Letter 21*, 13–14, 1199–1207.

MARTIN, S., ROE, D., AND FAULON, J.-L. 2005. Predicting protein-protein interactions using signature products. *Bioinformatics 21*, 218–226.

OYEKOYA, O., AND STENTIFORD, F. 2007. Perceptual image retrieval using eye movements. *International Journal of Computer Mathematics 84*, 9, 1379–1391.

PASUPA, K., SAUNDERS, C., SZEDMAK, S., KLAMI, A., KASKI, S., AND GUNN, S. 2009. Learning to rank images from eye movements. In *HCI '09: Proceeding of the IEEE 12th International Conference on Computer Vision Workshops on Human-Computer Interaction*, 2009–2016.

PULMANNOVÁ, S. 2004. Tensor products of hilbert space effect algebras. *Reports on Mathematical Physics 53(2)*, 301–316.

QIU, J., AND NOBLE, W. S. 2008. Predicting co-complexed protein pairs from heterogeneous data. *PLoS Computational Biology 4(4)*, e1000054.

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

SALOJÄRVI, J., PUOLAMÄKI, K., SIMOLA, J., KOVANEN, L., KOJO, I., AND KASKI, S. 2005. Inferring relevance from eye movements: Feature extraction. Tech. Rep. A82, Computer and Information Science, Helsinki University of Technology.