An Eye-tracking-based Approach to Facilitate Interactive Video Search

Stefanos Vrochidis
Queen Mary, University of London
Mile End Road, London, UK /
Informatics and Telematics Institute
Thermi, Thessaloniki, Greece
stefanos@iti.gr

Ioannis Patras
Queen Mary, University of London
Mile End Road, London, UK
i.patras@eecs.qmul.ac.uk

Ioannis Kompatsiaris
Informatics and Telematics Institute
Thermi, Thessaloniki, Greece
ikom@iti.gr

ABSTRACT
This paper investigates the role of gaze movements as implicit user feedback during interactive video retrieval tasks. In this context, we use a content-based video search engine to perform an interactive video retrieval experiment, during which, we record the user gaze movements with the aid of an eye-tracking device and generate features for each video shot based on aggregated past user eye fixation and pupil dilation data. Then, we employ support vector machines, in order to train a classifier that could identify shots marked as relevant to a new query topic submitted by new users. The positive results provided by the classifier are used as recommendations for future users, who search for similar topics. The evaluation shows that important information can be extracted from aggregated gaze movements during video retrieval tasks, while the involvement of pupil dilation data improves the performance of the system and facilitates interactive video search.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *relevance feedback, retrieval models*, *search process*.

General Terms
Algorithms, Performance, Experimentation.

Keywords
Implicit feedback, eye-tracking, machine learning, search engine, interactive, video retrieval, support vector machines.

1. INTRODUCTION
During the past decade, the rapid development of digital technologies, the low cost of recording media and the growth of communication networks have led to a great increase in the availability of multimedia content worldwide. The availability of such content, as well as the growing user need for searching into multimedia collections, place the demand for the development of advanced content-based analysis and search techniques. One of the main challenges faced by the state of the art approaches of multimedia indexing and retrieval is to generate efficient representations of the audiovisual source and extract low level features, however due to the well known problem of the semantic gap it is difficult to associate them with human understandable concepts. One of the main methodologies adopted to overcome this problem has been the exploitation of implicit and explicit relevance feedback (RF) provided by the users (i.e. identification of positive and negative examples) of a video search engine that would guide machine learning techniques. Despite the promising results in information retrieval (IR), explicit RF-based functionalities are not very user-popular, as users are usually reluctant to provide explicit information. For that reason, recent works in information retrieval have dedicated efforts focusing on the exploitation of implicit user feedback.

In general case retrieval tasks, the implicit user feedback could be divided into two main categories: the query actions and the physical user reactions. The first category includes the patterns of user interaction with the search engine, as series of mouse movements and clicks, keyboard inputs and key strokes etc, while the second consists of physical user unconscious and affective behavior as heart rate, brain neuron reactions and eye movements that can be gathered with biometric devices such as electroencephalography and electrocardiography sensors, eye trackers, etc. In this work we will focus on exploiting the implicit user feedback that falls into the second category and more specifically the eye movements. Eye-tracking has been used extensively in the psychology literature, and more recently also in tracking users' attention in IR tasks. The promising results on the field of textual IR tasks stimulated the motivation for employing eye-tracking techniques for relevance determination in image [1, 2] and video retrieval tasks.

The objective of this work is to generate recommendations that facilitate video search based on the aggregated gaze movements of past users during interactive video retrieval tasks. The idea is to distill meaningful information from aggregated gaze data, which could be exploited for identifying items that are of interest to a user with respect to his/her query topic. In this context, we propose an approach, in which, the gaze movements of past users are processed, in order to extract fixations (i.e. the eye remains fixed on a specific point for a certain amount of time) and pupil dilations. Then, we propose the extraction of a set of features that describes each video shot based on fixation characteristics and complemented by pupil dilation during fixations. Subsequently, we employ a Support Vector Machine (SVM) approach to train a binary classifier that could predict which of the items viewed by a new user could be classified as interesting for him/her and apparently matches the topic he/she searches for. The positive results of the SVM are provided as recommendations to future users that submit a similar query. To eliminate the searcher effect,

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
ICMR'11, April 17-20, Trento, Italy.
Copyright © 2011 ACM 978-1-4503-0336-1/11/04 $10.00.

we average aggregated fixation information by different users searching for the same topic. We evaluate this approach by conducting a video retrieval experiment, in which, users are recruited to perform video search with an interactive video search engine, while their gaze movements and pupil dilations are captured with the aid of an eye tracker.

The main contribution of this paper is the methodology for processing aggregated gaze data of past users, which combines gaze fixation and pupil dilation information, in order to detect items that are relevant to a given query topic and could be utilized as recommendations for a new user. In addition, the application of eye-tracking techniques in video search experiment, which is conducted in a less controlled environment compared to other approaches (e.g. [2]) can be considered of importance regarding the potential of gaze-based implicit feedback, taking into account the related works in the area have investigated only strictly controlled environments so far.

This paper is structured as follows: section 2 presents the related work, while in section 3 we describe the analysis of gaze movements and we introduce the SVMs employed in this approach. Section 4 presents video search engine we employed for the experiment and the video analysis behind it. In section 5 we describe the experiment conducted, while the results and the evaluation are presented in section 6. Finally, section 7 concludes the paper.

2. RELATED WORK

Studies utilizing eye movements in order to investigate cognitive processes started to appear three decades ago. Based on this research, eye movement data have proven to be very valuable in studying information processing tasks [3]. In information retrieval tasks, eye tracking methods were mostly used for identifying items of interest, as well as to understand the behavior of the user.

Up to now, most of the works that employed gaze analysis in the area of IR are focusing on textual document search. In one of the first attempts to study eye movements during IR tasks [4], the authors investigate how the users interact with the results of a web search engine by employing eye-tracking techniques. A very interesting approach is described in [5], in which proactive information retrieval is proposed by combining implicit relevance feedback and collaborative filtering. More specifically, implicit feedback is inferred from eye movements, with discriminative Hidden Markov Models estimated from data, for which explicit relevance feedback is available. The authors in [6] propose a technique for the restructuring of information that is presented to the user during text retrieval tasks using eye fixation-based features. In a more recent work [7], the identification of indicators and features for eye-tracking in text retrieval is proposed considering viewing time, thorough reading and regressions. The authors in [8] introduce a search strategy, in which a query is inferred from information extracted either from eye movements measured when the user is reading text during an IR task or from a combination of eye movements and explicit relevance feedback. Recently in [9] the authors attempt to consider the reading behavior of the user as implicit feedback during document search for query expansion and reranking. The idea is based on capturing the query context by analyzing what document parts the user looked immediately before issuing the query.

The first applications of eye-tracking in image and video retrieval were in the area of studying the user behavior and evaluating visual interface representations. More specifically, in [10] an eye-tracking study is conducted to investigate whether it is the textual or the visual representation of video that is mostly considered by users in a search engine interface. In another work [11], eye-tracking is applied to evaluate an approach, in which a video timeline is enriched with color information from the video visual data. More recent works in image and video retrieval deal with deriving user interest based on eye movements (focusing mostly on fixation and saccades) and also utilize this technique to develop gaze-based interactive interfaces.

In [12] the idea of an interactive interface for image retrieval is proposed, in which the input is given by the eye movements of the user concluding that eye-trackers could support such an implementation. Furthermore, in [13] the real time interface GaZIR for browsing and searching images is proposed. In this case, the relevance of the viewed images is predicted based fixation and saccade-based features (however pupil dilation is not taken into account), while relevance prediction is performed with classical logic regression.

The authors in [14] conduct experiments to explore the relationship between gaze behavior and a visual attention model that identifies regions of interest in image data. The reported results based on analysis of the fixation duration show that there is a difference in the gaze behavior on images depending on whether they contain a clear region of interest.

In another work [1], the authors propose a nine-feature vector from different forms of fixations and saccades and use a classifier to predict one relevant image from four candidates in two steps: a) first they extract features from the eye trajectory and employ a binary classifier to determine whether a specific page includes images of interest and b) they extract features for each image and use a 4-class classifier to detect which image is of interest. Compared to our approach, this work is evaluated in a very controlled environment following a different two level classifying methodology, while the visual interface is limited to 4 images. Furthermore, it doesn't take into account dilation information and it doesn't study the scenario of predicting the relevance of a new page or image for a new query by a new user.

Recently, an approach for performing RF based on eye features is proposed in [2]. This work employs eye-based features and a decision tree is trained using ground truth provided by the users. However, compared to the proposed approach, this work is evaluated with a rather controlled experiment as the users were told to fix their gaze on a positive image. Furthermore, the exploitation of pupil dilation information, the scenario of aggregating the input of many users for providing recommendations and the usage of a different classifier methodology (we use SVM instead of decision trees), differentiate our approach.

Besides fixations and saccades, pupil dilation has been also studied as an indicator of user interest during visual detection tasks. An interesting work, which falls into the area of visual target detection, is proposed in [15]. The authors investigate whether the pupil response can be considered as a reliable marker of a visual detection event, while viewing complex imagery. After conducting experiments, where viewers were asked to report the presence of a visual target during rapid serial visual presentation (RSVP), the conclusion was that pupil dilation was significantly associated with target detection. In another work [16], pupil information is used to improve the performance of an image classification system based only on EEG signal analysis. More specifically, Pupil responses are proposed as a complementary modality and are utilized for feature-extraction. A two-level linear classifier is then used to obtain cognitive-task-related analysis of EEG and pupil responses.

Finally, the most recent works in image retrieval attempt to combine image features with eye movements, either by using a

ranking SVM approach [17], or by identifying areas of interest in an image to extract local visual features [18], [19].

3. GAZE-BASED APPROACH

This section presents the analysis of implicit user feedback expressed by the gaze movements of the user during interactive video retrieval tasks, the feature extraction process and the SVMs employed in this approach.

3.1 Gaze Movements Analysis

Generally, eye movements can be categorized according to the following ocular behaviors: fixations, saccades, pupil dilation, and scan paths [3]. Fixations are defined as a spatially stable gaze lasting at least 100 milliseconds, during which visual attention is directed to a specific area of the visual interface. The eye fixations could be considered as the most relevant and reliable indicator for evaluating information processing during an online video search. On the other hand, saccades, which are the continuous and rapid movements of eye gazes between fixation points, are believed to occur so quickly across the stable visual stimulus that only a blur would be perceived. On the other hand, pupil dilation is also a metric that is typically used to indicate an individual's arousal or interest in the viewed content, as a larger pupil diameter corresponds to greater arousal [3]. In this paper we will attempt to complement the fixation information with the pupil dilation to identify user interest in the context of a submitted query.

During a video retrieval session the user interacts with a visual interface illustrating several videos. Apparently, the user focuses his/her gaze on the items that are of interest with respect to what he/she searches for. However, it seems that despite the fact that many parts of the graphical interface are viewed constantly for a specific amount of time (i.e. a fixation point was identified) during a video search session, not all of them could be considered as items of interest. This is made clear in the example of Figure 1, in which a user is searching for video scenes that depict books. After the analysis of the gaze movements, many fixations are identified, pointing at different parts of the interface. It is obvious that many fixations on relevant items are reported (e.g. shots on the top left corner of the interface), however it is also clear that some of the video shots that draw the attention of the user (as shown by the fixations) are not relevant to the query (e.g. the shot on the top right corner of the interface). This means that in order to be able to discriminate between relevant and irrelevant items to a query topic, we need to analyze the characteristics of the fixations and identify correlations between the fixation frequency, duration and the user interest depicted in our case by the search topic.

Figure 1. Fixations during video search sessions.

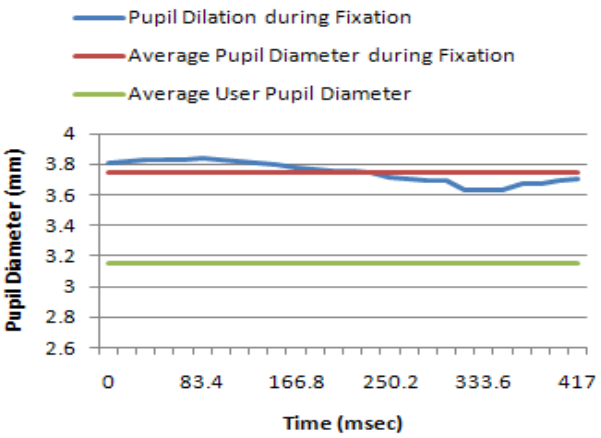

Figure 2. Pupil dilation of a user during a fixation.

Based on previous studies [1], [2], eye fixation-based features have shown discrimination power over items of interest for a user in controlled image retrieval environments.

In order to complement the fixation information, we take also into account the pupil dilation of the user. Many research studies have already documented that emotional and sensory events elicit a pupillary reflex dilation [20], [21], [22]. More specifically, recent experiments in [15] showed that a significant pupil response is reported for visual target detection events. This means that a strong correlation can be assumed between a visual target of interest and the pupil dilation. In Figure 2 we can see how the pupil dilation fluctuates during a fixation of around 400ms. In this case we can observe that the average pupil diameter reported during the fixation has been increased in comparison to the average pupil diameter of the user during the whole search session. Based on this fact, we propose to consider the pupil dilation of a user during a fixation and generate a descriptor that consists of fixation and pupil dilation-based features.

In this work, we extract the aforementioned features for each video shot to train a binary SVM classifier that could discriminate relevant from non relevant shots to a query. In order to minimize the searcher effect (i.e. different behaviors of each searcher) we propose to extract features from aggregated gaze data generated by many past users.

3.2 Feature Extraction

In this section we propose a feature vector that describes each video shot with respect to the relevance to a specific user query based on eye movement information. The fixation based features are based on [1], [2], from which we adopt the fixation total duration, number of fixations and average duration time, enhanced by time relative fixation features (i.e. with respect to the search session duration). On the other hand, the pupil dilation features are inspired by [16], based on which, we consider pupil dilation information in terms of normalized diameter and speed during the "critical time" that is in our case the fixation time window. Specifically, we assume that the pupil diameter and the speed of pupil dilation (i.e. the rate of change of pupil diameter) during a fixation can be considered as indicators of interest. In the sequel, we will present in more detail the features used in this approach.

In order to formally declare the eye-movement based features, we introduce some basic definitions. First, we define as search session $S_{j,k}$ the time period, during which, user j is searching for a specific topic k. We assume that each search session $S_{j,k}$ lasts $t_{S_{j,k}}$ time. We declare as $F_{\alpha,S_{j,k}}$ the total number of fixations and $T_{\alpha,S_{j,k}}$

the total fixation duration time that were reported for a shot a during a search session $S_{j,k}$. During each fixation, a fluctuation of the pupil data diameter takes place, which is represented by a series of pupil diameter values, sampled with a specific frequency. Averaging the pupil diameters reported by long search sessions for each user, we estimate the overall average pupil diameter value for each eye. We declare as $D_{R,j}$ and $D_{L,j}$ the overall average pupil diameter values for the right and the left eye for user j. For each fixation over a shot a we consider the normalized average pupil diameter values (i.e the average pupil diameter value reported during this fixation divided by the overall average value of the same user) $D_{R,a,S_{j,k}}$ and $D_{L,a,S_{j,k}}$ for the right and the left eyes of user j respectively. In parallel, we consider the speed (i.e. rate of change in time) of the pupil dilation. More specifically, we calculate the average speeds $U_{R,a,S_{j,k}}$ and $U_{L,a,S_{j,k}}$ for each fixation reported for shot a by user j for the right and left eyes respectively.

Table 1. Gaze-based features.

#	Feature description	Mathematical Formula
1	Total number of Fixations for shot a	$F_a = \frac{\sum_{S_{j,k} \in Y} F_{a,S_{j,k}}}{L \cdot K}$
2	Total fixation time for shot a	$T_a = \frac{\sum_{S_{j,k} \in Y} T_{a,S_{j,k}}}{L \cdot K}$
3	Average fixation time for shot a	$A_a = \frac{T_a}{F_a} = \frac{\sum_{S_{j,k} \in Y} T_{a,S_{j,k}}}{\sum_{S_{j,k} \in Y} N_{a,S_{j,k}}}$
4	Average fixations for shot a per search session	$V_a = \frac{F_a}{\sum_{S_{j,k} \in Y} t_{S_{j,k}}} = \frac{\sum_{S_{j,k} \in Y} F_{a,S_{j,k}}}{\sum_{S_{j,k} \in Y} t_{S_{j,k}}}$
5	Average fixation time for shot a per search session	$M_a = \frac{T_a}{\sum_{S_{j,k} \in Y} t_{S_{j,k}}} = \frac{\sum_{S_{j,k} \in Y} T_{a,S_{j,k}}}{\sum_{S_{j,k} \in Y} t_{S_{j,k}}}$
6	Average Normalized Right Pupil diameter	$D_{R,a} = \frac{\sum_{S_{j,k} \in Y} \frac{D_{R,a,S_{j,k}}}{D_{R,j}}}{\sum_{S_{j,k} \in Y} F_{a,S_{j,k}}}$
7	Average Normalized Left Pupil diameter	$D_{L,a} = \frac{\sum_{S_{j,k} \in Y} \frac{D_{L,a,S_{j,k}}}{D_{L,j}}}{\sum_{S_{j,k} \in Y} F_{a,S_{j,k}}}$
8	Average Right pupil dilation speed	$U_{R,a} = \frac{\sum_{S_{j,k} \in Y} U_{R,a,S_{j,k}}}{\sum_{S_{j,k} \in Y} F_{a,S_{j,k}}}$
9	Average Left pupil dilation speed	$U_{L,a} = \frac{\sum_{S_{j,k} \in Y} U_{L,a,S_{j,k}}}{\sum_{S_{j,k} \in Y} F_{a,S_{j,k}}}$

We assume that we want to describe a shot a with information retrieved during a set of sessions $Y = [S_{j,k}]$, where $j, k \in \mathbb{N}, 0 < j \leq L, 0 < k \leq K$, where L is the number of different users and K the number of topics involved in these sessions. As in this work we consider that the gaze input could be a result either from one user or aggregated information by many users, the proposed features are normalized against the number of search sessions $L \times K$. The features and the corresponding mathematical formulas are described in Table 1. Hence, the final feature vector for shot a would be:

$$f_a = [F_a, T_a, A_a, V_a, M_a, D_{R,a}, D_{L,a}, U_{R,a}, U_{L,a}] \quad (1)$$

3.3 Support Vector Machines

Support vector machines constitute a set of supervised learning methods, which are employed to solve classification and regression problems. In this work we propose to use a binary SVM in order to classify the viewed items according to the user interest exploiting the fixation-based feature vector. More specifically, we make use of the LIBSVM library [23] and we consider a binary C-Support Vector Classification. In this implementation we used as kernel the radial basis function:

$$K(f_i, f_j) = e^{-\gamma |f_i - f_j|^2} \quad (2)$$

In order to apply a SVM classification we need to have a ground truth for all the video shots viewed by the users (i.e. relevance metric with respect to the query topic).

4. VIDEO SEARCH ENGINE

In this section we present the LELANTUS[1] interactive video search engine, which we used for our experiments, by describing the user interface and the supported functionalities.

4.1 Interface

The search engine interface (Figure 3) is composed of two main parts: the left column, which offers text-based search options and a storage structure, and the main container, where the results are presented offering at the same time video shot-based query options. Four different functionalities are available for each shot: (i) to perform a query by visual example, (ii) to mark a shot as relevant to the topic (i.e. submit a shot), (iii) to view all the shots of the same video, (iv) to view the temporally adjacent shots of a selected video shot with the associated textual transcription.

4.2 Video Analysis and Indexing

In order to support the aforementioned video retrieval functionalities, we first perform offline temporal, textual and visual-based indexing operations as follows. First, in order to index the initial video source according to temporal information, we perform shot boundaries detection and shot segmentation operations that split the video into smaller segments.

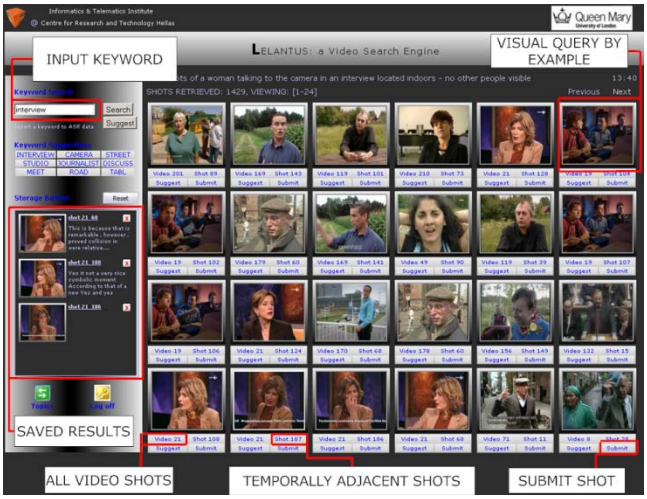

Figure 3. Video Search Engine Interface.

[1] Available at: http://mklab-services.iti.gr/lelantus/

Then, the middle keyframe for each shot, which is considered as the representative one, is extracted. The indexing of video shots according to the associated textual information is performed following the approach of [24]. The audio information is processed off-line with the application of Automatic Speech Recognition (ASR) on the initial video source, so that sets of keywords are extracted for each shot. Indexing and query functions are implemented using the KinoSearch full-text search engine library [25]. Finally, the visual similarity shot indexing is performed with the extraction of MPEG-7 low level visual descriptors, while an r-tree structure is employed for improving the performance of the system in terms of time response [24].

5. EXPERIMENTAL SETUP

To apply the proposed methodology, we conducted an interactive video retrieval experiment, in which different users searched with LELANTUS video search engine. In this task, we made use of the TRECVID 2008 test video set by NIST[2], which includes about 100 hours of Dutch video segmented into about 30,000 shots. The following query topics were used in our experiments:

A. Find shots of one or more people with one or more horses
B. Find shots of a map
C. Find shots of one or more people with one or more books
D. Find shots of food and/or drinks on a table

In this experiment, 8 (4 male and 4 female) subjects were recruited to search for the topics A-D. The task for each user was to search during a time window of 10 minutes per topic and find as many results that satisfy the given topic. During this task the users were free to make use of all the functionalities of the search engine as described in section 4. In order to imitate as much as possible a real world video retrieval task we instructed the users to search as they normally do, that is without making extra effort to focus their gaze on the shots of interest as they were instructed to do in [2]. In addition, a tutorial session preceded the retrieval task, in order to familiarize the users with the search engine and make them feel comfortably with the eye-tracker existence.

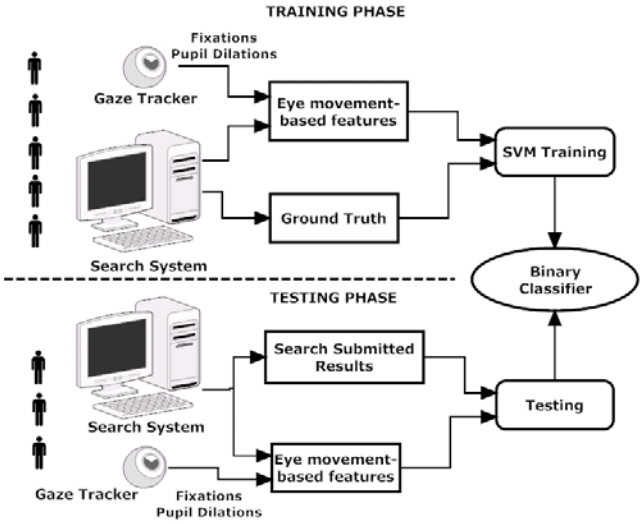

Figure 4. A schematic view of the experiment.

[2] National Institute of Standards and Technology (NIST): http://www.nist.gov/

The whole experiment was divided into the training and the testing phase, as it is depicted in the schematic view of Figure 4.

To record the gaze movement of the users, we employed the faceLAB 5 gaze tracker. The specifications of this device were: typical static accuracy of head measurement within +/- 1mm of translational error and +/- 1° of rotational error; typical static accuracy of gaze direction measurement within 0.5-1° rotational error. The eye-tracker was recording the gaze position and pupil dilation of the user every 16 or 17msecs. We processed the output of the eye-tracker, in order to identify eye fixations on the video shots. We considered as minimum time of 100ms to define a fixation, during which the gaze was stable.

5.1 Training Phase

During the training phase, the first 5 users (1-5) searched for the topics A-D. The results submitted by these users constitute an explicit relevance metric with respect to the query topics for all the viewed items. Considering that very high precisions are reported for interactive systems, as users select a shot only when it is of relevance to the query topic, the submitted shots comprise a very reliable ground truth set for this task. As it is shown in Figure 4, we train the SVM models using the feature vectors produced by the fixation and pupil dilation data and the ground truth. In order to evaluate the approach, we provide a variety of training cases, in which different combinations of training features and topics were used. More specifically, the following four training cases, which are shown in Table 2 are considered: in the first, we train the classifier (model 0 in Table 3) by using the features 1-5 (Table 4) and the 4 topics (A-D), in the second case we train recursively 4 different classifiers (models 2-5 in Table 5) by selecting each time a different combination of the three topics (i.e. (A, B, C), (A, B, D), etc.) and using as vector the 1-5 fixation-based features, while in the third (models 5-8 in Table 5) and forth (models 9-12 in Table 6) training cases we repeat the scenario of the second training case, but we made use of the features 1-7 and 1-9 respectively. In all the aforementioned training cases the gaze data from the same five users were used.

As the ground truth data were not balanced (the positive samples were in average about 10% of the total judged samples), we trained the models introducing a corresponding weight $w_n = 1$ for negative and $w_p = 10$ positive classes. More specifically, we set the cost parameter C to $w_p \cdot C$ and $w_n \cdot C$ for positive and negative samples respectively.

Table 2. Training cases

Training Case	Model No	Features (Table 1)
1	0	1-5
2	1-4	1-5
3	5-8	1-7
4	9-12	1-9

5.2 Testing Phase

In the testing phase, the other 3 users (different from the ones employed in training phase) were recruited to search for the 4 same topics A-D. In a similar way with the ground truth collection we capture the video shots that these users identify as relevant to each topic. Then, we utilize this information, in order to test the classifier against the actual selections of the users. Based on the four aforementioned training cases, we test the first classifier (i.e. model 0 of the first training case) by considering all the topics A-D, while we test the other 12 models (i.e. the ones trained with the 3 topics combinations), by using gaze data captured only during

the retrieval sessions for the remaining topic (e.g. in the case the training was done with topics A,B,C, we test with topic D).

6. RESULTS AND EVALUATION

In this section we will present the evaluation of the results, as well as a visual view of the recommendations provided by the system.

6.1 Evaluation

We evaluate our system by reporting the classification accuracy, as well as the precision and recall over the items returned by the system as positive results. During testing the submitted results by the 3 users formed the golden set that was used for the evaluation. Formally, assuming that the classifier returns TP true positives, TN true negatives, FP false positives and FN false negatives for a topic calculated against the V positive and the N negative user selections, the accuracy is computed as $A = \frac{TP+TN}{V+N}$, the precision as: $P = \frac{TP}{TP+FP}$, and the recall as: $R = \frac{TP}{V}$.

The results for the first and second aforementioned training/test cases are reported in Tables 3 and 4. Starting by observing the results of the first case, it can be concluded that the results for model 0 are of good quality, however they have a strong dependence on the query topics, as the same queries were considered both for training and testing.

In the more realistic second case (Table 4, models 1-4), the results are still satisfactory, which shows that this method can provide quality results without depending on the topic. However, it is clear that although the recall values are satisfying, the precision remains rather low (especially in models 1 and 4). In the case that we want to increase the precision at the cost of reducing the recall, we can adjust the weighting w_n and w_p parameters during the training accordingly. In Figure 5, the Precision-Recall curve for model 1 is illustrated, in the case that the ratio $\frac{w_p}{w_n}$ (points on the P-R carve) takes values from 0.2 to 10.

In order to show how the results are improved when considering aggregated user gaze data, we report the fluctuation of the EER in the case of one, two and three test users searching for topic B (i.e. model 2). As it is observed in Figure 6, it is clear that the performance of the classifier is improved, as lower values of EER are reported for the 3 users. More specifically, the EER calculated in the case of one user (i.e. $EER = 32.1\%$), is decreased by 19% when data from a second user are considered, and it is reduced by a further 50% to get a final value of $EER = 13.2\%$, in the case of three users. In a similar way, we present in Figure 7 how precision and recall are changing in the evaluation of model 2, when one, two and three test users are involved.

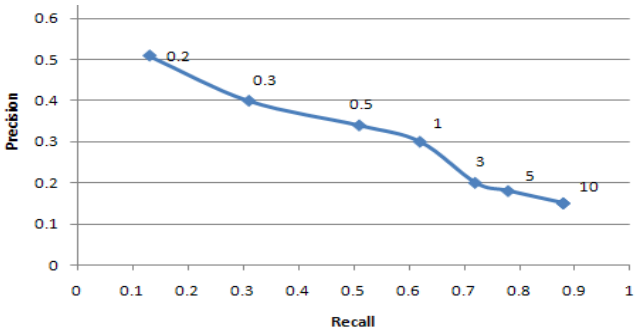

Figure 5. Precision-Recall curve for model 1 when the ratio $\frac{w_p}{w_n}$ (points on the P-R carve) takes values from 0.2 to 10.

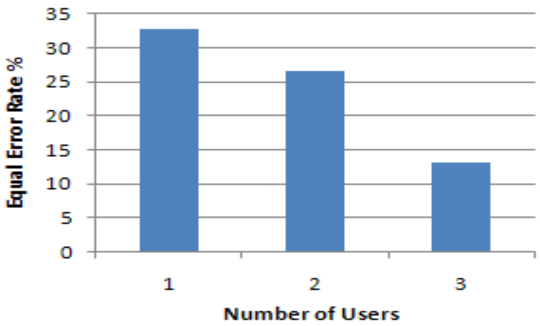

Figure 6. The EER for model 2 is reported, when we use aggregated data from 1,2 and 3 test users respectively.

As far as the precision is concerned, a rise of 33.3% of the initial precision reported for the one user is achieved with the involvement of a second user, followed by a further increase of 46.6% in the case that three users are considered. Finally, we report a slight drop of 3.6% in the initial recall (i.e. in the case of the one user) when aggregated results of two users are considered, however the recall is increased by 34.87% when the third user is involved. This analysis shows that when aggregated gaze information is taken into account the unique gaze behaviors of each user seem to be smoothed to an average gaze behavior, for which the classifier yields better results. The fact that the classifier's performance improves, when the number of the users involved is increased, is an indicator that such an approach could be applied for generating recommendations based on past user aggregated gaze data.

Table 3. First case (features 1-5)

Model No	Train Topics	Test Topics	Classifier Performance	Precision	Recall
0	A,B, C,D	A, B, C, D	94.1748% 5723/6077	49.2%	61.95%

Table 4. Second case (features 1-5)

Model No	Train Topics	Test Topics	Classifier Accuracy	Precision	Recall
1	B,C,D	A	85.70% 1253/1462	15.04%	88.1%
2	A,C,D	B	93.5% 1338/1431	37.16%	61.11%
3	A,B,D	C	70.43% 805/1143	19.52%	89.13%
4	A,B,C	D	72.17% 900/1247	14.32%	73.41%
Average			80.45%	21.51%	77.94%

Table 5. Third case (features 1-7)

Model No	Train Topics	Test Topics	Classifier Accuracy	Precision	Recall
5	B,C,D	A	93.5% 1367/1462	24.6%	73.8%
6	A,C,D	B	94.34% 1350/1431	33.6%	45.56%
7	A,B,D	C	70.69% 808/1143	19.86%	90.17%
8	A,B,C	D	81.48% 1016/1247	19.51%	70.89%
Average			85%	24.39%	70.1%

Table 6. Forth case (features 1-9)

Model No	Train Topics	Test Topics	Classifier Accuracy	Precision	Recall
9	B,C,D	A	94.12% 1376/1462	25.21%	69.05%
10	A,C,D	B	95.11% 1361/1431	35.19%	42.22%
11	A,B,D	C	71.65% 819/1143	20.39%	90.22%
12	A,B,C	D	81.88% 1021/1247	20.14%	72.15%
Average			85.7%	25.23%	68.71%

The results for the third and forth training cases are presented in Tables 5 and 6 respectively. With a view to evaluating the different set of features, we observe that when pupil dilation information is involved, the accuracy of the classifier is slightly improved for all cases. The comparison of the accuracy of the classifiers for the different feature sets is illustrated in Figure 8. The major improvement is reported in the testing of topic D (i.e. models 4, 8, 12). In this case, the accuracy of the classifier is boosted from 72.17% to 81.48%, reporting an improvement of 12.9%, when the second feature set (i.e. features 1-7) is involved and a total increase by 13.45%, when we employ the third feature set (i.e. features 1-9).

Furthermore, as it is shown in Tables 4-6, the involvement of pupil features improves the precision of the system by an average of 13.4% when we employ features 1-7 and by a further 3.45% when the third feature set (i.e. features 1-9) is involved. On the other hand, the recall seems to drop by 10% and a further 1.98% for the two aforementioned cases. The F-score is calculated as 33.7%, 36.2% and 36.9% for feature sets 1, 2 and 3 respectively, showing that the overall performance of the system slightly improves with the employment of pupil dilation information.

6.2 Recommendations

In order to provide recommendations for a specific query submitted by a new user, the system utilizes the output of the classifier after processing the aggregated input by many past users, who searched for the same topic as discussed in section 3. In this case, the output is expressed as a distance from the hyperplane, which discriminates the two different classes (i.e. relevant and irrelevant to a submitted query) and ranks the resulted shots accordingly.

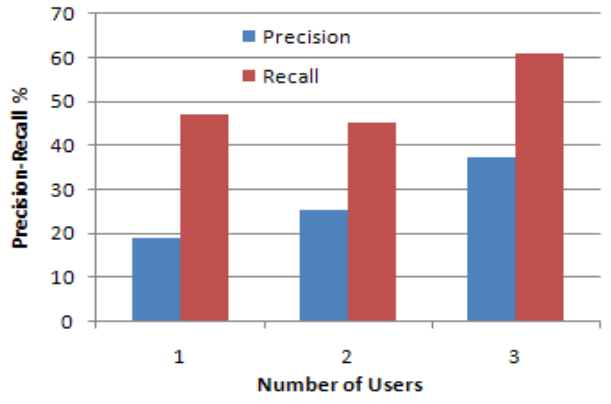

Figure 7. The precision and recall for model 2 are presented.

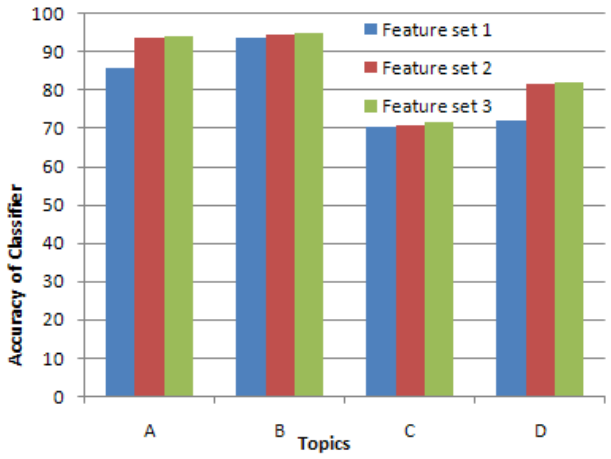

Figure 8. Accuracy of the classifiers for the different feature sets: 1: 1-5, 2: 1-7, 3: 1-9 introduced in Table 1.

A visual example of the recommendations provided for topic B by model 10 (Table 6) is shown in Figure 9. By performing a first assessment of the visual results that are ranked higher, it is clear that a precision of 83.3% (15/18) is achieved in this case.

7. CONCLUSIONS

In this paper we have investigated the potential of utilizing aggregated gaze movement data, of past users during interactive video retrieval tasks, in order to generate recommendations for specific queries. Our results show that exploiting gaze-based implicit feedback, expressed in terms of fixations and pupil dilations, could be of added value, as important information regarding the relevance of a video shot to a query topic can be inferred even in not strictly controlled environments (i.e. when the users are not instructed to focus on interesting items). Such information could be exploited for identifying user interest in the context of a specific query and for generating recommendations with a view to facilitating video retrieval tasks. Future work includes alternative methodologies for complementing fixation with pupil dilation information, as well as combination of gaze-based user feedback with patterns of user interaction (i.e. click throughs and keyboard inputs). Recently the work is conducted towards these goals.

Figure 9. Recommendations for Topic B: "Find shots of a map", based on the results of model 10 (Table 6).

8. ACKNOWLEDGMENTS

This work was supported by the projects PESCaDO (FP7-248594) and PetaMedia (FP7-216444).

9. REFERENCES

[1] Klami, A., Saunders, C., De Campos, T.E., Kaski, S. 2003. Can relevance of images be inferred from eye movements?. In *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval* (Vancouver, Canada 2008), 134-140.

[2] Zhang, Y., Fu, H., Liang, Z., Chi, Z., Feng, D. 2010. Eye movement as an interaction mechanism for relevance feedback in a content-based image retrieval system. In *Proceedings of the Symposium on Eye-Tracking Research & Applications* (Austin, Texas, 2010), 37-40.

[3] Rayner, K.1998. Eye movements in reading and information processing. *Psychological Bulletin*, 124, (1998), 372-252.

[4] Granka, L. A., Joachims, T., Gay, G. 2004. Eye-tracking analysis of user behavior in WWW search. In *Proceedings of the 27th Annual International ACM SIGIR conference on Research and Development in Information Retrieval* (New York, USA, 2004), 478-479.

[5] Puolamaki, K., Salojarvi, J., Savia, E., Simola, J., Kaski, S. 2005. Combining eye movements and collaborative filtering for proactive information retrieval. In *Proceedings of the 28th Annual International ACM SIGIR conference on Research and Development in Information Retrieval* (Salvador, Brazil, 2005), 46-153.

[6] Brooks, P., Phang, K. Y., Bradley, R., Oard, D., White, R., Guimbretire, F. 2006. Measuring the utility of gaze detection for task modeling: A preliminary study. In *Workshop on Intelligent Interfaces for Intelligent Analysis* (Sydney, Australia, 2006).

[7] Moe, K. K., Jensen, J. M., Larsen, B. 2007. A qualitative look at eye-tracking for implicit relevance feedback. In *Proceedings of the 2nd International Workshop on Context-Based Information Retrieval* (Roskilde, Denmark 2007), 36-47.

[8] Hardoon, D. R., Shawe-Taylor, J., Ajanki, A., Puolamaki, K., Kaski, S. 2007. Information retrieval by inferring implicit queries from eye movements. In *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics* (San Juan, Puerto Rico, 2007).

[9] Buscher, G., Dengel, A., Van Elst, L. 2008. Query expansion using gaze-based feedback on the subdocument level. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval* (Singapore, July 20 - 24, 2008), 387-394.

[10] Hughes, A., Wilkens, T., Wildemuth, B., Marchionini, G. 2003. Text or Pictures? An Eyetracking Study of How People View Digital Video Surrogates. In *Proceedings of the 2nd International Conference on Image and Video Retrieval* (Urbana, IL, USA, 2003), 271-280.

[11] Moraveji, N. 2004. Improving video browsing with an eye-tracking evaluation of feature-based color bars. In *Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries*, (Tuscon, AZ, USA 2004), 49-50.

[12] Oyekoya, O.K., Stentiford, F. W. M. 2004. Eye tracking as a new interface for image retrieval. *BT Technology Journal*, 22, 3 (2004), 161-169.

[13] Kozma, L., Klami, A. and Kaski, S. 2009. GaZIR: gaze-based zooming interface for image retrieval. In *Proceedings of the 2009 International Conference on Multimodal interfaces* (New York, NY, USA). ICMI-MLMI '09. ACM, 305-312.

[14] Oyekoya, O., Stentiford, F. 2004: Exploring Human Eye Behaviour using a Model of Visual Attention. In *Proceedings of the 17th International Conference on Pattern Recognition*, 4 (Washington, DC, USA 2004), 945-948.

[15] Privitera, C., Renninger, L., Carney, T. Klein, S. and Aguilar, M. 2008. Pupil dilation during visual target detection. In *Proceedings of the SPIE 20th Annual Symposium on Electronic Image Science and Technology* (Jan. 2008), 68060T-1-68060T-11.

[16] Qian, M., Aguilar, M., Zachery, K., Privitera, C., Klein, S., Carney, T., Nolte L.W. 2009. Decision-level fusion of EEG and pupil features for single-trial visual detection analysis, *IEEE Trans Biomed Eng*, 56, 7 (2009), 1929-1937.

[17] Hardoon, D. R., Pasupa, K.: Image Ranking with Implicit Feedback from Eye Movements. In *Proceedings of the Symposium on Eye-Tracking Research & Applications* (Austin, Texas, 2010), 291-298.

[18] Liang, Z., Fu H., Zhang, Y., Chi, Z., Feng, D. 2010. Content-based image retrieval using a combination of visual features and eye tracking data. In *Proceedings of the Symposium on Eye-Tracking Research & Applications* (Austin, Texas 2010), 41-44.

[19] Faro, A., Giordano, D., Pino, C., Spampinato, C. 2010. Visual attention for implicit relevance feedback in a content based image retrieval. In *Proceedings of the Symposium on Eye-Tracking Research & Applications* (Austin, Texas 2010), 73—76.

[20] Krenz, W., Robin, M., Barez, S. and Stark, L. W. 1985. Neurology model of the normal and abnormal human pupil. *IEEE Trans. Biomed. Eng.*, 32, 10 (Oct. 1985), 817–825.

[21] Loewenfeld I., Lowenstein, O. 1993. *The Pupil: Anatomy Physiology and Clinical Applications* (Detroit, 1993), MI: Wayne State University Press.

[22] Smith, J. D., Masek, G. A., Ichinose, L. Y., Watanabe, T. and Stark, L. W. 1970. Single neuron activity in the pupillary system, Brain Research, 24 (1970), 219-234.

[23] Chang, C., Lin, C. *LIBSVM: a library for support vector machines*, Software available at: http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[24] Zhang, Q., Tolias, G, Mansencal, B., et. al. 2008. COST292 experimental framework for TRECVID 2008. In *Proceedings of TRECVID 2008 Workshop* (Gaithersburg, MD, USA, 2008).

[25] Kinosearch search engine library. http://www.rectangular.com/kinosearch/