Gaze-based Relevance Feedback for Realizing Region-based Image Retrieval

Georgios Th. Papadopoulos, *Member, IEEE,* Konstantinos C. Apostolakis and Petros Daras, *Senior Member, IEEE*

***Abstract*—In this paper, a gaze-based Relevance Feedback (RF) approach to region-based image retrieval is presented. Fundamental idea of the proposed method comprises the iterative estimation of the real-world objects (or their constituent parts) that are of interest to the user and the subsequent exploitation of this information for refining the image retrieval results. Primary novelties of this work are: a) the introduction of a new set of gaze features for realizing user's relevance assessment prediction at region-level, and b) the design of a time-efficient and effective object-based RF framework for image retrieval. Regarding the interpretation of the gaze signal, a novel set of features is introduced by formalizing the problem under a mathematical perspective, contrary to the exclusive use of explicitly defined features that are in principle derived from the psychology domain. Apart from the temporal attributes, the proposed features also represent the spatial characteristics of the gaze signal, which have not been extensively studied in the literature so far. On the other hand, the developed object-based RF mechanism aims at overcoming the main limitation of region-based RF approaches, i.e. the frequently inaccurate estimation of the regions of interest in the retrieved images. Moreover, the incorporation of a single-camera image processing-based gaze tracker makes the overall system cost efficient and portable. As it is shown by the experimental evaluation, the proposed method outperforms representative global- and region-based explicit RF approaches, using a challenging general-purpose image dataset.**

***Index Terms*—Relevance feedback, gaze-tracking, gaze analysis, image retrieval.**

I. Introduction

The development and extensive proliferation of advanced image capturing devices (e.g. smart-phones, portable multimedia devices, etc.), as well as the great plurality of the available means for sharing and distributing the generated content (especially over the Internet and through the social networks), have resulted in the formation of literally vast image databases. As a consequence, the semantic analysis of the image content has emerged as a crucial and challenging issue [1]. To this end, extensive research efforts have been invested for developing systems that will enable the understanding of the actual image content. The dominant approach consists of automatically extracting a set of discriminative visual features directly from the images and subsequently estimating their semantic content by applying some formal mathematical models. However, the performance of such methods in realistic environments remains insufficient and prevents the respective systems from being used in real-world applications [2].

One of the most efficient and widely adopted methodologies for facilitating semantic image analysis relies on the fundamental principle of incorporating the human user in the analysis process. In particular, the user is consecutively providing to the system feedback information that is used for refining the retrieved results and the overall procedure unfolds until the user regards the returned results as satisfactory. This category of methods is usually entitled Relevance Feedback (RF) [3]. RF is used for: a) providing the system with the appropriate amount of information that is missing for handling a particular semantic image manipulation task, and b) for identifying and modeling the specific information needs raised by a particular user. RF approaches can be divided, among the use of other possible criteria, into explicit and implicit methods, based on the kind of information that is provided to the system. Explicit methods require from the user to provide explicit statements regarding the relevance of the returned results. On the other hand, implicit RF approaches utilize information that is obtained by the user in an unobscured/non-invasive way (e.g. gaze-tracking, Electroencephalography (EEG), heart beat rate, etc.) and aim at predicting the relevance of the returned results.

In the context of image retrieval, explicit RF methods initially considered global-level feedback information. In [4], an asymmetric bagging and random subspace Support Vector Machine (SVM) efficiently handles the problems caused by the usually small number of positively labeled feedback samples. In [5], the user's labeling effort is reduced following a structural information-based sample selection strategy. Additionally, several query point movement methods that aim at reducing the number of the required iterations and improving the overall retrieval performance are presented in [6]. A direct kernel Biased Discriminant Analysis (BDA)-based approach to RF is introduced for overcoming limitations of the traditional BDA method in [7]. Moreover, Tian et. al [8] efficiently encode the user's labeling information, using a so called 'sparse transfer learning' dimension reduction tool. More recently, the analysis of explicit RF methods has shifted to a finer level of detail and region-based approaches have been proposed. In particular, region-based RF approaches, which however still receive feedback information at global-level, estimate the local-level objects that are considered to be relevant to the query and subsequently using these estimates they refine the retrieved results. The latter fact, i.e. the inaccurate relevant region identification, constitutes the main drawback of this category of methods. Jiang et. al [9] propose an online feature selection approach for improving the image retrieval

Georgios Th. Papadopoulos, Konstantinos C. Apostolakis and Petros Daras are with the Centre for Research and Technology Hellas (CERTH), Information Technologies Institute, Thessaloniki GR-57001, Greece (e-mail: {papad, kapostol, daras}@iti.gr).

performance. Additionally, an image retrieval framework that is based on a graph-theoretic region correspondence estimation is presented in [10]. A SVM-based approach makes use of an adaptive convolution kernel for realizing object-based indexing and retrieval of images in [11]. It must be noted that some region-based RF methods that allow the manual selection of the relevant image regions have also been presented, e.g. [12][13]. Nevertheless, such methods require extensive effort from the user for providing feedback in every iteration.

Over the past few years, implicit RF has received particular attention in the image retrieval community. The main advantages of these approaches are [14]: a) the user's feedback is captured in a non-intrusive time-efficient way, and b) the implicit response is more expressive than the explicitly provided feedback. On the contrary, the main drawback of these methods is the presence of large amounts of noise in the feedback data. Typical types of implicit feedback data that have been used for image RF are click-through [15], EEG signal [16] and gaze-tracking [17], to name a few. More recently, the exploitation of the information that is available in the social media, which can be considered as a particular type of implicit RF data, has gained particular attention. Sang et. al [18] propose a personalized image search framework, taking into account image tags. In [19], image ranking is realized using both social and visual data for improving the relevance between the returned images and the users' intentions. Among the different types of implicit feedback data, gaze-tracking (or eye-tracking) is of particular importance to image retrieval applications, since it can provide valuable information with respect to which parts of the image the user has observed as well as cues regarding the relevance of the latter to the query at hand. The great majority of gaze-tracking approaches related to image retrieval have focused on predicting the user's relevance assessment at the image level, which certainly is not a trivial task, and have little been considered for the development of a complete RF system. In [20], implicit feedback about the users' attention is measured using an eye-tracking device for inferring the relevance of images. Faro et. al [21] propose an implicit RF method for re-ranking the retrieved images according to users' eye gaze data. Additionally, the work of [22] explores possible solutions for image annotation and retrieval, by implicitly monitoring the user's attention via eye-tracking. In [23], the idea of implicitly incorporating eye movement features in an image ranking task is investigated. Klami [24] infers possible target regions in the examined images from gaze data and estimates the relevance of those regions using a simple classifier. Key characteristic of all the above methods is that they rely on the use of explicitly defined gaze features, which are derived from the psychology domain and highlight particular attributes of the gaze signal, for predicting the user's relevance assessment.

In this paper, a gaze-based relevance feedback approach to region-based image retrieval is presented. The fundamental novelty of the proposed approach is the use of the gaze signal for addressing the main challenge in region-based image RF, i.e. the accurate and time-efficient identification of the objects of interest. For that purpose, a novel gaze signal interpretation method is introduced, which iteratively estimates the real-world objects (or their constituent parts) that satisfy the user's information needs and subsequently uses this information for refining the image retrieval results. On the contrary, region-based image RF approaches of the literature incorporate computationally expensive processes (e.g. inexact graph matching [10], region clustering [25], fuzzy codebook creation [9], etc.) that often lead to inaccurate detection of the regions of interest; hence, also resulting in decreased image retrieval performance. Fundamental contributions of this work constitute: a) the introduction of a novel set of gaze features for performing the prediction of the user's relevance assessment at the image region level, and b) the design of a time-efficient and effective object-based RF framework for image retrieval. Regarding the proposed features, they are computed following a mathematical formalization. This constitutes a sharp contradistinction to the entire relevant literature, which is only limited to the use of explicitly defined gaze features that are in principle derived from the psychology domain and highlight only a small set of specific attributes of the gaze signal. In particular, for efficiently describing the temporal characteristics of the gaze signal, a frequency domain analysis is proposed that results into a significantly more detailed and complex representation, compared to typical approaches of the literature that employ simple features, like the number of fixations, the visit length, the number of visits, etc. Additionally, for effectively representing the spatial-related characteristics of the gaze signal, a translation, rotation and scale invariant approach is proposed that models the distribution of the user's attention on the different areas within a given image region, as opposed to the significantly simpler features of the literature (e.g. the difference between the largest and the smallest x(y)-coordinate, the elongation of the x- and y- coordinate spreads, etc.). Moreover, a thorough feature evaluation procedure is applied, which aims at achieving an optimal balance between the needs for selecting the most discriminative features and also increasing the time efficiency of the proposed approach. Regarding the design of the proposed object-based RF framework, this also involves the development of: i) a single-camera image processing-based gaze-tracker for capturing the user's implicit response, and ii) an appropriate dynamic interface for efficiently and accurately capturing gaze-related information at the image region-level. Particular attention is given to maintain low computational complexity, which is significantly facilitated by the proposed gaze signal interpretation procedure that enables the extraction of valuable and detailed feedback information, i.e. a set of local image regions each accompanied with a relevance degree, from the raw gaze data. As it will be shown by the experimental evaluation, the proposed gaze features compare favorably with similar state-of-art ones, while the overall RF framework outperforms representative global- and region-based explicit RF approaches of the literature.

The remainder of the paper is organized as follows: Section II describes the designed system for gaze-based RF. Section III outlines the developed gaze-tracking framework. The gaze-based relevance assessment prediction procedure is detailed in Section IV. Section V describes the developed RF mechanism. Extensive experimental results regarding the evaluation of the proposed gaze features and the designed RF framework, using a challenging dataset of 9933 *Flickr* images, are given in Section VI, and conclusions are drawn in Section VII.

II. System Overview

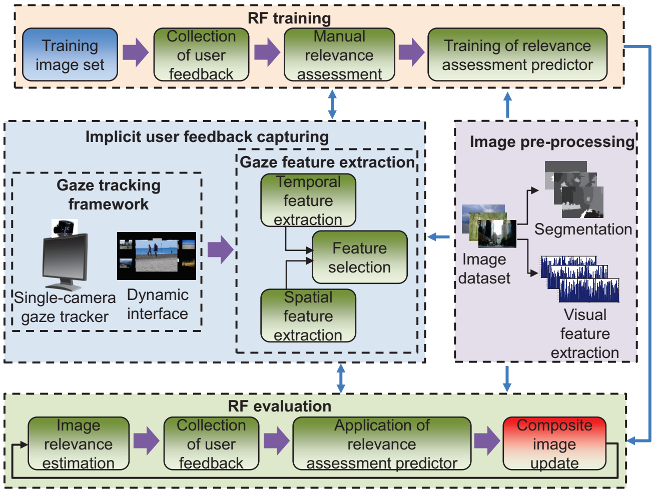

RF training
Training image set
Collection of user feedback
Manual relevance assessment
Training of relevance assessment predictor
Implicit user feedback capturing
Gaze feature extraction
Gaze tracking framework
Single-camera gaze tracker
Dynamic interface
Temporal feature extraction
Feature selection
Spatial feature extraction
Image pre-processing
Image dataset
Segmentation
Visual feature extraction
RF evaluation
Image relevance estimation
Collection of user feedback
Application of relevance assessment predictor
Composite image update

Fig. 1. General architecture of the proposed region-based RF approach for image retrieval.

The first step in the developed RF framework, whose general architecture is illustrated in Fig. 1, concerns the pre-processing of the images used. In particular, every image is spatially segmented, in order to identify the real-world objects that it contains. Subsequently, region-level visual features are extracted that will be used for assessing the visual similarity between the respective objects.

Key part of the overall architecture constitutes the implicit user feedback capturing process. For that purpose, an appropriate gaze-tracking framework has been designed. This involves the development of a single-camera image processing-based gaze-tracking sensor and a dynamic interface for efficiently capturing gaze-related information at region-level. Following the estimation of the gaze point trajectory, two sets of features, namely a temporal- and a spatial-related one, are extracted only for those image regions that the user has observed. Then, a feature selection procedure is applied for maintaining the most discriminative features.

As can be seen in Fig. 1, the developed RF mechanism has two distinct modes, namely a training and an evaluation one. During the training mode, the gaze behaviour of the user(s) is modeled. For that purpose, different sets of images, each accompanied with a predefined query, are formed. Then, the user(s) is(are) asked to examine every set of images, using the developed gaze-tracking framework and taking into account the respective query. This results in the generation of a set of gaze features for the regions that have been seen by the user(s). Subsequently, the user(s) is(are) asked to manually annotate the latter image regions as relevant or irrelevant. The computed region-level gaze features, along with their associated manual annotations, are then used for training the user relevance assessment predictor. On the other hand, the RF evaluation mode is initialized with an image relevance estimation step that provides a ranking of the images with respect to the specific query used during the evaluation. Then, similarly to the training mode, the implicit user feedback is captured for the K top-ranked images. The difference is that the developed predictor is now used for estimating the user's relevance assessment, i.e. a relevance degree for every observed region. All regions that have been seen by the user during the whole RF session, along with their associated degrees of relevance, are collected and form a *composite image*, which is used during the computation of the image relevance ranking.

III. Gaze-tracking Framework

A. Gaze-tracker

Sophisticated gaze-tracking systems with increased accuracy and time efficiency, which typically use infrared illumination, are commercially available, e.g. Tobii[1], SMI[2], Eye-Tech[3] and Mirametrix[4], to name the most representative ones. However, a prohibitive factor for the extensive use of such specialized equipment constitutes their significantly high cost. For satisfying the requirements for low-cost and portability, an image processing-based approach that makes use of a single camera is followed for performing gaze-tracking in this work.

For developing the proposed gaze-tracking framework, the method of [26] is followed, which relies on the use of the Candide-3 face model [27]. More specifically, the user needs to adjust the face model to his/her head in an off-line step. For every subsequent frame, the head pose is estimated according to the POSIT algorithm [28], using the face model vertices corresponding to the ones of the off-line step. The latter is carried out by calculating the Pyramidal Lucas-Kanade optical flow [29] on two data streams. The first stream is the camera frame capturing one, where optical flow for frame t is calculated using also the previous frame $t-1$. The second stream is formed by the frame captured at time t and a keyframe. The latter is computed by estimating the head pose at $t-1$, where all face model vertices are known since they correspond to the rigid vertices of the Candide-3 model, and considering a textured instance of the face model at this pose. POSIT then extracts the face model's rotation and translation for the current frame t. This head pose is used to obtain the face model's landmark positions. Out of the available landmarks in the Candide-3 model, a set of points around the user's eyes were selected; hence, creating regions of interest surrounding the area of the eyes. An adaptive thresholding technique is subsequently applied to these regions of interest, similarly to the well-known Otsu histogram shape-based image thresholding algorithm [30], to segment the pupils from the background (e.g. sclera). In particular, for every eye region the image is transformed to the gray scale colour space and eroded. Then, increasing threshold values are consecutively applied until the darker pupil region is clearly segmented from the more lightly toned areas of the eye. More specifically, continuously increasing illumination values, ranging from 0 to 255, are used as thresholds in each step. For every threshold

[1] http://www.tobii.com/
[2] http://www.smivision.com/
[3] http://www.eyetechds.com/
[4] http://mirametrix.com/

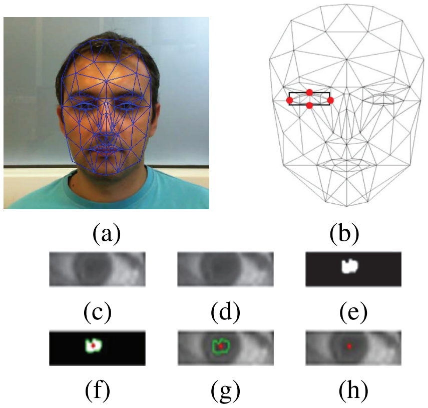

(a) (b)
(c) (d) (e)
(f) (g) (h)

Fig. 2. Pupil center localization process: (a) Adjustment of the Candide-3 face model. (b) Definition of the region of interest. (c) Conversion to gray scale. (d) Erosion of gray scale image. (e) Segmentation of pupil region. (f) Estimation of pupil region contour. (g) Computation of contour's medium. (h) Definition of the pupil center.

value, the largest formed connected area is considered to correspond to the pupil. The respective pupil center is located by identifying points on the pupil region's contour and calculating the medium point $P(\chi, \psi)$, where the coordinates (χ, ψ) refer to the image plain captured by the camera. The above process is considered to converge (and hence terminated) if for 10 successive illumination threshold values, the position of the medium point does not exceed a maximum displacement of 3 pixels in vertical or horizontal direction; the medium point of these 10 successive thresholding steps is considered the final estimated pupil point $P(\chi, \psi)$. In the very unlikely event that the aforementioned procedure does not converge, the first estimated pupil region that is bigger than a predefined value (which is experimentally defined equal to 15 pixels) is assumed to correspond to the final pupil area and hence the pupil point $P(\chi, \psi)$ is estimated as the medium point of its contour, as described above. The overall pupil center tracking process is illustrated in Fig. 2.

For identifying the user's gaze point on the screen, a calibration process is performed off-line. This aims at storing information regarding the pupil center $P(\chi, \psi)$ and one of the eye corners $E(\chi, \psi)$ for each eye. The calibration process comprises the successive display of eight points diametrically placed on the screen, as depicted in Fig. 3. For estimating the user's gaze point, the methodology described in [31] is adopted. In particular, for each eye an eye corner to pupil center vector $\mathbf{U}_i(\chi_i, \psi_i)$, $i \in [1, 8]$, is stored for each of the eight calibration points that correspond to known points $\Delta_i(x_i, y_i)$ on the monitor, where the coordinates (x, y) refer to the image plain depicted on the screen. The method of [31] originally requires the coordinates of only two calibration points, namely the values of variables $(\chi_{right}, \chi_{left}, \psi_{top}, \psi_{bottom})$, which correspond to the respective known values $(x_{right}, x_{left}, y_{top}, y_{bottom})$ defined in Fig. 3, for performing gaze-tracking. On the contrary, the proposed approach exploits information from eight calibration points for estimating these values, according to the following equations: $\chi_{left} = \frac{\chi_1+\chi_4+\chi_6}{3}$, $\chi_{right} = \frac{\chi_2+\chi_5+\chi_8}{3}$, $\psi_{top} = \frac{\psi_1+\psi_3+\psi_5}{3}$ and $\psi_{bottom} = \frac{\psi_2+\psi_4+\psi_7}{3}$. In this way, insertion of noise during the calibration phase is significantly reduced. According to [31], every eye corner to pupil center vector $\mathbf{U}(\chi, \psi)$ is linearly mapped to a gaze point on the screen $\Delta(x, y)$ using the following equations: $x = x_{left} + \frac{\chi-\chi_{left}}{\chi_{right}-\chi_{left}}(x_{right} - x_{left})$ and $y = y_{top} + \frac{\psi-\psi_{top}}{\psi_{bottom}-\psi_{top}}(y_{bottom} - y_{top})$. Having estimated the coordinates of the gaze point $\Delta(x, y)$ from the separate processing of each eye, the final gaze point position on the screen is computed by calculating the average of the coordinates of these two points.

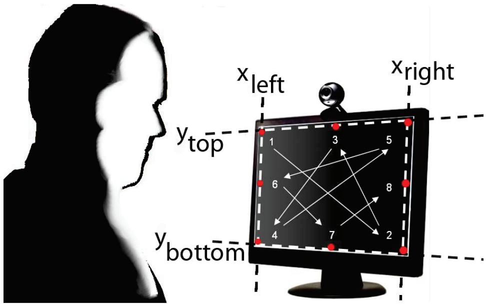
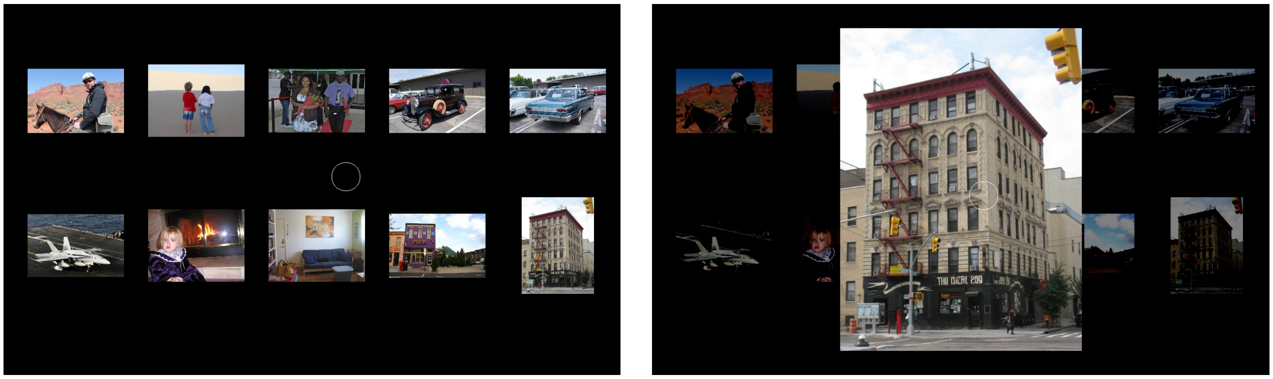


Fig. 3. Gaze-tracking framework setup and calibration procedure.

B. Developed interface

The increased needs for combining effective image browsing capabilities and sufficient accuracy in capturing region-level gaze-related information have led to the design of a dynamic interface. In particular, the images are presented to the user in tens; this was selected as a good trade-off between the need for accurate browsing using the gaze sensor (i.e. sufficient size of the images for their details to be adequately visible) and an adequate number of images that will not delay the overall RF procedure. The transition between successive tens of images is performed with the press of a keyboard button. Additionally, the interface has two functional modes: a) the 'browsing' and b) the 'zoom-in-image' modes, which are depicted in Fig. 4. As can be seen, in the 'browsing' mode the images are arranged in two rows of five images each, where the centers of all images are aligned with respect to each row and column. The resolution of the interface image is 1280x768 (horizontal x vertical dimension). The maximum dimension of a depicted image is set to 200 pixels, while the horizontal and the vertical margins between the images are set equal to 50 and 100 pixels, respectively. It must be noted that if an image's bigger dimension is greater than 200, then the image is scaled so that its bigger dimension to be equal to the limit value of 200 pixels, using linear interpolation. On the other hand, the 'zoom-in-image' mode is introduced for handling the problem of capturing accurate gaze-related data at region-level. This mode is entered if the user's gaze point stays on the area of a particular image for a minimum time interval of 800 msec, while in 'browsing' mode. Once the 'zoom-in-image' mode is entered, the image that the user focuses on is enlarged to its original size or (if the bigger image dimension exceeds the limit value of 668 pixels) is linearly interpolated so that its bigger dimension to be equal to 668 pixels. Additionally, the enlarged version of the image is placed in the center of the interface and the background is faded, as it is illustrated in Fig. 4(b). It must be highlighted that gaze-related information for any image region is only captured


(a)

(b)

Fig. 4. Functional modes of the developed interface: (a) ‘browsing’ and (b) ‘zoom-in-image’. The white circle denotes the current position of the gaze point.

when the respective image is in ‘zoom-in-image’ mode. The ‘zoom-in-image’ mode is exited and the interface returns to the ‘browsing’ mode, if the user focuses outside the limits of the zoomed image for a minimum period of 1200 msec. In order to avoid any unintended successive enterings of the ‘zoom-in-image’ mode, the ‘zoom-in-image’ mode is set to be always succeeded by a browsing mode and vice versa.

IV. Gaze-based Relevance Assessment Prediction

A. Image pre-processing

Prior to any user implicit feedback interpretation procedure, every image is segmented to regions and suitable low-level descriptors are extracted for every resulting segment. In this work, the segmentation algorithm of [32] is used, which was experimentally shown to provide satisfactory results in a wide variety of general-purpose image datasets [33]. Output of this segmentation algorithm is a segmentation mask, where the created spatial regions s_n, $n \in [1, N]$, are likely to represent meaningful real-world objects. Every generated image segment s_n is subsequently represented with the use of a visual feature vector $\mathbf{v}_n$. For that purpose, the OpponentSIFT descriptor [34] is extracted at a set of keypoints of a predetermined image grid. Then, adopting the ‘Bag-of-Words’ (BoW) methodology [35], each region is represented by a histogram of 1000 visual words. The latter histogram, which is L1 normalized, constitutes the region feature vector $\mathbf{v}_n$. In the current implementation, the above pre-processing step is performed for all images prior to the application of any gaze interpretation procedure, i.e. it is an off-line process. On the contrary, the user’s relevance assessment prediction and the application of the relevance feedback mechanism, which are described in the following sections, are performed on-line.

B. Extraction of region-level gaze features

In this section, the proposed features for performing user’s relevance assessment prediction at region-level are described. Their definition is based on the fundamental concept of fixation. A fixation is considered to occur when the eyes remain relatively still for a minimum time interval [36], typically a few hundred msec, and takes place when the individual attempts to identify the details of an object of interest. All other continuous and more extensive eye movements are called saccades [37]. In this work, the proposed gaze features are defined considering only the fixations, since they contain more valuable information and less noise than saccades. Among the different definitions of fixation [38], the Dispersion-Threshold Identification (I-DT) method [39] constitutes the most widely used one and is also adopted in this work. According to the latter definition, a fixation is considered to occur if the gaze point remains in a circular area of radius R pixels for a minimum of Θ msec (typically between 100 and 200 msec). For the employed gaze-tracker (Section III-A), the following values, which are also commonly used, were selected based on experimentation: $R = 30$ pixels and $\Theta = 200$ msec. In the sequel, a fixation will be denoted as $F_k(x_k, y_k, t_s, t_e)$, $k \in [1, K]$, where point (x_k, y_k) will correspond to the center of the aforementioned circular area and t_s, t_e to the start, end time of the fixation, respectively. It must be noted that before the fixation identification process the gaze trajectory is low-passed for noise removal. This is performed by applying a simple mean filter of length 5, separately to the horizontal and vertical gaze coordinate signals, as follows: $\tilde{x}(t) = \frac{\sum_{\zeta=-2}^{2} x(t+\zeta)}{5}$ and $\tilde{y}(t) = \frac{\sum_{\zeta=-2}^{2} y(t+\zeta)}{5}$, where $\tilde{x}(t)$, $\tilde{y}(t)$ and $x(t)$, $y(t)$ are the low-passed and original horizontal, vertical coordinates of the gaze signal at time t.

Literature approaches dealing with the definition of gaze features have so far concentrated on guidelines provided solely by the psychological domain, where the proposed features only highlight a small set of specific attributes of the gaze signal, while the spatial dimension of the gaze point trajectory is also naively investigated. A thorough review of the most recent and distinguished gaze features is given in Table I.

For efficiently describing the temporal characteristics of the gaze signal, an analysis in the frequency domain is proposed. In particular, a time sequence, namely a *fixation sequence*, is constructed for denoting when the focus of the user’s gaze lies in every image region s_n^m, where I_m, $m \in [1, M]$, denotes an image of the employed dataset. Region s_n^m is considered to be seen by the user if image I_m is zoomed and a fixation $F_k(x_k, y_k, t_s, t_e)$ occurs in the area that corresponds to s_n^m. The aforementioned fixation sequence for region s_n^m is binary and is estimated according to the following equation:

$$FS_n^m(t) = \begin{cases} 1, & \text{if } \exists\, k : F_k(x_k, y_k, t_s, t_e) \in \mathbf{S}_n^m(t), t_s \leq t \leq t_e \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

where $\mathbf{S}_n^m(t)$ is the area of the interface captured by region

TABLE I

GAZE FEATURES PROPOSED IN THE LITERATURE. FEATURES MARKED WITH AN ASTERISK (*) WERE NOT USED DURING THE EVALUATION.

Approach	Gaze features
[22]	Times an image was visited, Times an image was skipped between fixations, Total time the user spent on image, Average visit duration on image, Maximum visit duration on image, Duration of first visit on image, Rank of images with respect to features 3-6, Proportion of values of features 3-6 to the total time for each image, Pixel number between consecutive images, Duration visit of the image at the end of the transition, The entrance gaze angle to an image, Distance from previously visited image divided by visit duration of current image, Distance from next visited image divided by visit time of current image, Distance of previous and next visited images over visit time of current image, Visit time of the image at start over visit time of the image at end of transition, Feature equal to 1 for the local maximum in the formed time vector and 0 for the rest, Feature equal to 1 for the local minimum in the formed time vector and 0 for the rest
[23]	Number of raw data measurements, Number of measurements outside fixations, Percentage of measurements inside/outside fixations, Difference between largest and smallest x-coordinate, Difference between largest and smallest y-coordinate, Elongation of x- and y- coordinate spreads, Average distance between two consecutive measurements, Number of sub-images covered by measurements, Coverage normalized by number of measurements, x-coordinate of first measurement, y-coordinate of first measurement, x-coordinate of last measurement, y-coordinate of last measurement, Maximum pupil diameter, Number of breaks longer than 60 ms^2, Number of breaks longer than 600 ms^2, Number of fixations, Mean length of fixations, Total length of fixations, Percentage of time spent in fixations, Number of re-visits to the image, Maximum angle between two consecutive saccades, x-coordinate of first fixation, y-coordinate of first fixation, x-coordinate of last fixation, y-coordinate of last fixation, Difference between largest and smallest x-coordinate of fixations, Difference between largest and smallest y-coordinate of fixations, Elongation of x- and y- coordinate spreads of fixations, Length of the first fixation, Number of fixations during first visit, Distance to the fixation before the first*, Duration of the fixation before the first*
[17]	Image shown on onset, Time to first image visit, Mean length of fixations, Standard deviation of fixation occurrence times, Total length of fixations, Maximum continuous image viewing time, Mean length of continuous image viewing sessions, Maximum continuous image viewing time, Proportion of overall image viewing time over total viewing time, Proportion of overall image viewing time over total viewing time in the same ring*, Mean length of saccade before fixation, Proportion of times when previous fixation over the same image, Proportion of times when previous fixation over empty space, Proportion of times when previous fixation over the same ring*, Number of images viewed on the ring before the first fixation on the image*, Number of image revisits, Average distance from previously viewed image on the same ring*
[24]	Number of fixations, Total fixation time, Length of the first fixation, Number of region visits, Time of first fixation since onset, Time of last fixation since onset, Index of first fixation, Index of last fixation, Standard deviation of fixation indices, Whether first fixation occurred in region
[40]	Fixation duration, Fixation count, Fixation length, Number of revisits
[20]	Total fixation length, Number of fixations, Average fixation length, Number of transitions from an image to another, Number of images with at least one fixation, Number of fixations within each image

s_n^m if image I_m is zoomed at time t ($\mathbf{S}_n^m(t) = \emptyset$ if image I_m is not zoomed at time t). Having computed the $FS_n^m(t)$ sequence for all image regions s_n^m seen by the user, a corresponding normalized fixation sequence $\widehat{FS}_n^m(w)$, $w \in [1, W]$, of predetermined length is estimated using linear interpolation. This normalization step is performed for maintaining that the gaze features estimated from fixation sequences derived from different sessions, which are likely not to be of equal duration, to be directly comparable. Then, the Discrete Cosine Transform (DCT) is applied to $\widehat{FS}_n^m(w)$, as follows:

$$fc_n^m(q) = \sum_{w=1}^{W} \widehat{FS}_n^m(w) \cos \frac{\pi}{W}[(w-1) + \frac{1}{2}(q-1)], \tag{2}$$

where $fc_n^m(q)$ are the estimated DCT coefficients and $q \in [1, W]$. The reason for using the DCT transform is twofold: a) its simple form requires relatively reduced calculations, and b) it is a frequency domain transform that receives as input a real sequence and its output is also a real set of values. It must be noted that other common frequency analysis methods (e.g. Fourier transform) were also evaluated; however, they did not lead to increased performance compared to DCT. Out of the W $fc_n^m(q)$ coefficients, only the first L are considered, since the remaining ones: a) were experimentally shown to correspond mainly to noise, and b) their removal will alleviate the subsequent processing steps. These L selected coefficients constitute the temporal gaze features for region s_n^m.

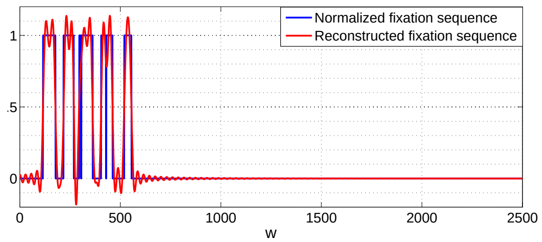


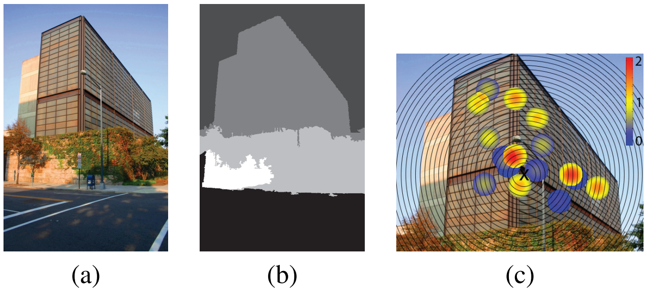
Fig. 5. Example estimation of the normalized fixation sequence and the respective reconstructed one.

An example of estimating the normalized fixation sequence $\widehat{FS}_n^m(w)$ and the reconstructed one, $\widetilde{FS}_n^m(w)$, using only the first L $fc_n^m(q)$ coefficients is illustrated in Fig. 5. As can be observed, the $fc_n^m(q)$ coefficients encompass significantly more detailed and complex information regarding the characteristics of the gaze signal than typical approaches of the literature. The latter rely on the use of simple features (Table I), like the number of fixations, the visit length, the number of visits, the average visit duration, the time of the first fixation, etc.

For describing the spatial-related characteristics of the gaze signal, an approach that satisfies the needs for translation, rotation and scale invariance is proposed in this work. Key idea for the subsequent analysis constitutes the introduced *gaze energy field*. This field is estimated for every image region s_n^m seen by the user. In particular, it is considered that every fixation $F_k(x_k, y_k, t_s, t_e)$ carries a certain amount of energy, which is proportional to its duration and spreads within the fixation area. More specifically, a fixation $F_k(x_k, y_k, t_s, t_e)$ that occurs on region s_n^m (as defined in (1)) is modelled as a normalized 2D Gaussian distribution of the following form:

$$G_k(x, y) = B_k e^{-\frac{1}{2}(\frac{\varrho}{24})^2}[1 - u(\varrho - R)], \tag{3}$$

where $B_k = t_e - t_s$ is the duration of the fixation measured in sec, $u(\cdot)$ is the unit step function and $\varrho = \sqrt{(x - x_k)^2 + (y - y_k)^2}$. The above definition considers that the gaze energy distribution $G_k(x, y)$ has a peek value at the center point (x_k, y_k) of the fixation area equal to the fixation duration B_k in sec and receives non-zero values in a circular area of radius equal to $R = 30$ pixels (i.e. the same radius length used in the definition of the fixation). Additionally, the standard deviation of the distribution was selected so that the minimum non-zero value of the distribution, which is observed at the margins of the aforementioned circular area, to be approximately half of the respective maximum value of the distribution that is measured at the central point (x_k, y_k). The latter selection was made based on experimentation and was shown to lead to increased performance. Having defined the energy distribution that corresponds to fixation $F_k(x_k, y_k, t_s, t_e)$, the *gaze energy field* for region s_n^m seen by the user is defined, taking into account only the fixations



(a) (b) (c)

Fig. 6. Extraction of spatial-related gaze features: (a) original image, (b) segmentation mask, and (c) spatial-related gaze feature extraction procedure for the region corresponding to the depicted building.

that occur in the region, according to the following equation:

$$EF_n^m(x,y) = \sum_k G_k(x,y) \bigcap \mathbf{S}_n^m(t)$$
$$\forall\ k:\ \ F_k(x_k, y_k, t_s, t_e) \in \mathbf{S}_n^m(t),\ \ t_s \leq t \leq t_e \qquad (4)$$

The estimated energy field provides detailed information, regarding the distribution of the user's attention on the different areas within a given image region.

For extracting a spatial-related description of the gaze signal for region s_n^m, the center of gravity $CG_n^m(x_g, y_g)$ of the gaze energy field $EF_n^m(x,y)$ is initially computed. Then, a set of L concentric ring-shaped areas are estimated, using point $CG_n^m(x_g, y_g)$ as center. These ring-shaped areas, denoted as $RA_{n,l}^m(x,y)$, $l \in [1, L]$, are defined as follows:

$$RA_{n,l}^m(x,y) = \{(x,y):\ (l-1)\cdot\beta \leq \delta < l\cdot\beta\}, \qquad (5)$$

where $\delta = \sqrt{(x-x_g)^2 + (y-y_g)^2}$, $\beta = D_{max}/L$ and D_{max} denotes the maximum Euclidean distance of a point belonging to region s_n^m from the computed center $CG_n^m(x_g, y_g)$. Subsequently, the energy included in each of the formed $RA_{n,l}^m(x,y)$ areas is calculated, according to the following equation:

$$ae_{n,l}^m = \sum_{(x,y)} EF_n^m(x,y):\ \ (x,y) \in RA_{n,l}^m(x,y) \bigcap \mathbf{S}_n^m \qquad (6)$$

where $ae_{n,l}^m$, $l \in [1, L]$, is the computed energy value for the ring-shaped area $RA_{n,l}^m(x,y)$ and $\mathbf{S}_n^m$ denotes the area captured by region s_n^m when image I_m is zoomed. These L calculated values constitute the proposed spatial features. It must be noted that L values were selected for describing the spatial characteristics of the gaze signal, in order for the temporal and the spatial gaze features to receive equal importance in the subsequent analysis.

An example of extracting the spatial gaze features is given in Fig. 6. In this example, a user was asked to search for buildings in a set of presented images during a gaze-tracking session. In the figure, an indicative image is illustrated (Fig. 6(a)) along with the corresponding segmentation mask (Fig. 6(b)). In Fig. 6(c), the bounding box of the region that corresponds to the depicted building is presented, along with the estimated gaze energy field $EF_n^m(x,y)$, the computed center of gravity $CG_n^m(x_g, y_g)$ (highlighted with X) and the corresponding ring-shaped areas $RA_{n,l}^m(x,y)$ defined according to (5).

The estimated spatial features $ae_{n,l}^m$ and the temporal-related $fc_n^m(q)$ coefficients computed for region s_n^m are concatenated and form a region feature vector, denoted by $\mathbf{gf}(s_n^m)$. This vector describes the way that the user has seen region s_n^m.

C. Feature selection

The estimated gaze vector $\mathbf{gf}(s_n^m)$ is of high dimensionality, while it also contains significant amounts of redundant information. In order to improve the prediction performance and time efficiency, a feature selection procedure is performed, aiming at achieving an optimal balance between the needs for selecting the most discriminative features and reducing the dimensionality of the gaze vector. In this work, the following feature selection techniques are evaluated:

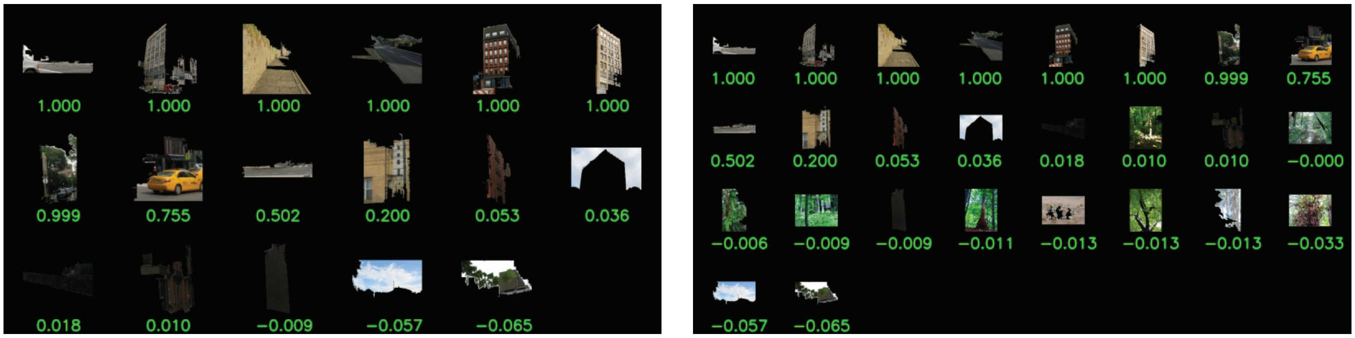
- Principal Component Analysis (PCA): It is a standard technique that considers the linear dependencies among the features [41].
- Correlation-based Feature Selection (CFS): It is based on the fundamental hypothesis that good feature subsets contain features highly correlated with the class, yet uncorrelated with each other [42].
- Chi-Square attribute Selection (CSS): Similarly to CFS, CSS [43] also selects an optimal subset of the features, using the chi-square statistic.
- Information Gain (IG): IG quantifies the effectiveness of a feature by measuring the expected reduction in the entropy caused by partitioning the samples with respect to the examined feature [44].
- Gain Ratio (GR): It extends the IG technique by introducing an additional term that takes into account how each feature splits the samples [44].

Output of the feature selection procedure, regardless of the particular technique used, is a 'reduced' feature vector. This feature vector is denoted by $\widetilde{\mathbf{gf}}(s_n^m)$ and its dimensionality is equal to Z, where $Z < 2L$. $\widetilde{\mathbf{gf}}(s_n^m)$ is the final gaze vector estimated for region s_n^m. It must be noted that additional feature selection techniques were also evaluated (namely Consistency-based, Relief Attribute Selection, SVM-based, Adaboost and Mutual Information-based); however, they led to inferior performance compared to the ones described above.

D. User relevance assessment prediction

The task of predicting the user's relevance assessment, i.e. identifying if the image regions that the user has observed are of interest to him/her or not based on the captured gaze data, is formalized in this work as a binary classification problem. In particular, the observed regions are considered to belong to two distinct classes, namely the relevant and the irrelevant one, with respect to the posed query. For tackling this challenging task, SVMs are selected due to their reported generalization ability [45]. Under the proposed approach, a SVM is introduced for classifying each gaze vector $\widetilde{\mathbf{gf}}(s_n^m)$ as relevant or not. More specifically, the SVM receives as input the vector $\widetilde{\mathbf{gf}}(s_n^m)$ and estimates a degree of relevance, denoted $rd(s_n^m) \in [-1, 1]$, for region s_n^m, where $rd(s_n^m) = 1$ represents a relevant sample and $rd(s_n^m) = -1$ an irrelevant one. A sigmoid function is employed [46] for estimating the value of $rd(s_n^m)$, according to the following equation:

$$rd(s_n^m) = \frac{2}{1 + e^{\eta\cdot\xi_{nm}}} - 1\ , \qquad (7)$$



(a) (b)

Fig. 7. Examples of composite image formation for the posed query term 'street': (a) $CI(0)$ and (b) $CI(1)$.

where ξ_{nm} is the distance of the particular vector $\widetilde{\mathbf{gf}}(s_n^m)$ from the corresponding SVM's separating hyperplane and η is a slope parameter set experimentally. Distance ξ_{nm} is positive in case of a positively classified sample and negative otherwise.

V. Relevance Feedback Mechanism

In this section, the proposed region-based RF mechanism for updating the image retrieval results, by taking into account the user's gaze signal, is presented. Key part with great impact on the final performance constitutes the user's relevance assessment prediction procedure, as detailed in Section IV. Particular attention has been paid during the design of the proposed RF mechanism to maintain low computational complexity, characteristic that is essential for using the overall system in large-scale applications.

Under the proposed approach, an extension of the traditional 'Query-expansion' method is introduced. In particular, fundamental idea of the overall approach constitutes the so called *composite image*. This image is gradually constructed by continuously adding the image regions s_n^m observed by the user along with their estimated degree of relevance $rd(s_n^m)$. Intuitively, the *composite image* can be considered as a drawing canvas, where the user continuously adds pieces of visual information that satisfy his/her information needs and eventually builds a complex high-level semantic concept that represents his/her perception of the posed query. The RF mechanism is formalized as follows: The *composite image* at iteration τ, denoted as $CI(\tau)$, comprises all image regions s_n^m that have been seen by the user from the beginning of the session, along with their corresponding degree of relevance $rd(s_n^m)$, and is represented according to the following equation:

$$CI(\tau) = \{(s_{n_\gamma}^{m_\gamma}, rd(s_{n_\gamma}^{m_\gamma})),\ \gamma \in [1, \Gamma]\}\ , \tag{8}$$

where index γ is introduced for denoting the number of regions that are present in the composite image. $CI(\tau)$ is updated at the end of iteration τ, as described in Section II and illustrated in Fig. 1. It must be noted that during the initialization of the proposed RF approach the composite image is considered to be empty, while in the very unlikely event of a particular region being observed by the user in more than one iterations then only the highest value of $rd(s_{n_\gamma}^{m_\gamma})$ is included in $CI(\tau)$. For estimating the updated image retrieval results (i.e. an updated image ranking) at iteration $\tau+1$, the following image relevance metric is used:

$$IR_{\tau+1}(I_m, CI(\tau)) = \sum_n rd(s_{n_\phi}^{m_\phi}) \cdot \mu \cdot area(s_n^m)$$
$$\mu = [1 - D(\mathbf{v}_{n_\phi}^{m_\phi}, \mathbf{v}_n^m)],\ \phi = \arg\min_\gamma [D(\mathbf{v}_{n_\gamma}^{m_\gamma}, \mathbf{v}_n^m)] \tag{9}$$

where $D(\mathbf{v}_{n_\phi}^{m_\phi}, \mathbf{v}_n^m) \in [0, 1]$ denotes the normalized Euclidean distance between the feature vectors $\mathbf{v}_{n_\phi}^{m_\phi}$ and $\mathbf{v}_n^m$ that correspond to regions $s_{n_\phi}^{m_\phi}$ and s_n^m, respectively, and $area(s_n^m) \in [0, 1]$ denotes the relative area of region s_n^m in image I_m. From the above definition, it can be seen that every region s_n^m in the examined image I_m is associated with a unique region $s_{n_\phi}^{m_\phi}$ of the composite image $CI(\tau)$ based solely on visual similarity. The importance of the latter assignment, which in principle controls the RF process, is weighted by the $rd(s_{n_\phi}^{m_\phi})$ degree, which is estimated according to (7).

Indicative examples of composite image formation are given in Fig. 7. The composite images at the end of the first and the second iteration (i.e. $CI(0)$ and $CI(1)$, respectively) are depicted from an image retrieval session, where the posed query to the user was the term 'street'. The regions $s_{n_\gamma}^{m_\gamma}$ in Figs. 7(a) and 7(b) are sorted in descending order with respect to their associated $rd(s_{n_\gamma}^{m_\gamma})$ degree, starting from the upper-left corner of each figure and moving from left-to-right and in a line-by-line mode. As can be seen, the proposed approach achieves to distinguish the relevant from the irrelevant objects. Inevitably, some false negative samples are also present, like the last building object in both $CI(0)$ and $CI(1)$ that is identified as irrelevant with a low degree though.

VI. Experimental Results

In this section, experimental results from the application of the proposed gaze-based RF approach to region-based image retrieval are presented. In particular, extensive experiments have been conducted for: a) performing an empirical evaluation of the employed gaze-tracker, b) evaluating the efficiency of the proposed gaze features and comparing them with features presented in the literature, c) examining the image retrieval performance of the overall gaze-based image RF approach and its comparison with representative state-of-art explicit image RF methods, and d) comparatively evaluating the time efficiency of the proposed image RF method. Additionally, a discussion regarding the factors that affect the performance of gaze interpretation algorithms is also given.

The user was initially given a particular query term and asked to examine a ranked set of images. In this work, a random initial ranking of the images was considered. The tracking of the user's gaze was performed using the framework described in Section III. Subsequently, the prediction of the user's relevance assessment was realized using temporal and spatial region-level gaze features, as detailed in Section IV. Then, an updated ranking of the images, with respect to their relevance to the given query, was estimated, using the relevance feedback mechanism outlined in Section V. The aforementioned iterative procedure was performed a predetermined number of times. The value of parameter W in (2), i.e. the length of the normalized fixation sequence $\widehat{FS}_n^m(w)$, was set equal to 2500, since the length of any fixation sequence $FS_n^m(t)$ was not observed to exceed this value during the experiments conducted. Additionally, parameter η in (7) was set equal to 0.7, based on experimentation. Relatively small deviations around this value led to negligible variations in the overall image retrieval performance. The SVM predictor in Section IV-D was implemented using a polynomial kernel function.

For evaluating the efficiency of the proposed approach, a challenging general-purpose image dataset was assembled. For that purpose a set of 10 high-level semantic concepts was initially defined, namely set $C = \{car,\ building,\ person,\ road,\ sea,\ beach,\ street,\ living-room,\ forest,\ desert\}$. Using the concepts in C as keywords, a corresponding set of 9933 images were retrieved from the Flickr[5] online photo management and sharing application, maintaining that approximately 1000 images were collected for every keyword. The dataset, as well as the source code for computing the proposed region-level gaze features, can be downloaded from http://vcl.iti.gr/evaluation-datasets/.

A. Empirical gaze-tracker accuracy evaluation

The efficiency of the gaze-tracker has a significant impact on the performance of the overall RF approach. As a consequence, the accuracy of the employed tracker is examined in this section. Table II briefly outlines some of the approaches that have been introduced for measuring the performance of image processing-based gaze-trackers. For producing directly comparable evaluation results, the performance of the developed gaze-tracker was evaluated using the experimental protocols described in the works of [47][31][48][49][50][51]. The comparative evaluation results given in Table II show that the developed gaze-tracker outperforms most state-of-art approaches.

TABLE II
EXPERIMENTS FOR MEASURING THE ACCURACY OF IMAGE PROCESSING-BASED GAZE-TRACKERS

Approach	Type of tracker	Experiment	Reported error	Error of employed tracker
[47]	Remote	Nine markers	0.23-0.9°	0.37°
[31]	Remote	Zigzag	1.4°	0.85°
[48]	Remote	Markers on wall	3.2°	–
[49]	Stereo, remote	Four points	4.6°	0.34°
[50]	Stereo head-mounted	Nine markers	1.0-1.38°	0.37°
[51]	Stereo head-mounted	Twelve markers	0.88°	0.55°
Employed tracker	Remote	Circular trajectory	–	0.83°

Most of the works reported in Table II rely on the use of static markers in fixed positions for evaluating the gaze-tracking performance. In this way, the behavior of the gaze-tracker is not adequately examined. For that purpose, a significantly more challenging and thoroughly defined experiment is proposed with the following two key characteristics: a) it can be easily reproduced, and b) it takes into account both the spatial accuracy and the temporal coherence of the tracker. More specifically, a red circle was depicted on the screen performing a circular trajectory and the user was asked to follow the center of this circle with his/her gaze. The tracker's accuracy was defined as the mean gaze angle deviation (in degrees) that corresponds to the distance of the estimated gaze point from the center of the circle, where the gaze point trajectory is considered a continuous signal that is low-passed as described in Section IV-B. Five individuals participated in the experiments, each performing the aforementioned task five times. Regarding the specifications of the defined experiment, the monitor plane was vertically aligned, while the perpendicular vector originating from the monitor's center was maintained to approximately target the nose of the user and to also be perpendicular to the user's face plane during the calibration step. Additionally, the camera was placed on top of the monitor and at the center of the respective monitor's side, with the nose of the user to be set to correspond approximately to the center pixel of the captured video sequence. The distance of the user's head from the screen was maintained approximately at 65cm, while the radius of the depicted red circle was set to 0.7cm. The radius and the period of the circular trajectory were set to 13.5cm and 30sec, respectively. The average value of the measured accuracy for the employed gaze-tracker was approximately equal to $0.83°$. The employed gaze-tracker also achieved a sampling frequency of approximately 25Hz, using a PC with Intel i7 processor at 3.5 GHz and a total of 16GB RAM. It must be noted that the resolution of the video captured by the camera for performing gaze tracking was selected to be 800x448. Commercially available gaze-trackers (like Tobii, SMI, EyeTech, Mirametrix, etc.) report accuracy around $0.4-0.5°$, at a sampling rate of $30-300$Hz. However, the employed gaze-tracker constitutes a low-cost alternative (it only requires a single camera) and it is also portable (e.g. most laptops are equipped with a camera above their screen), while the 'zoom-in-image' mode of the designed interface (Section III-B) accounts for the difference in the gaze-tracking accuracy.

At this point, it must be highlighted that the focus of this work does not include the proposal of a new gaze-tracker, whose performance needs to be accurately measured and to be superior compared to other state-of-art methods. On the contrary, aim of this work is the proposal of a novel framework for interpreting the gaze signal and subsequently utilizing this information for realizing RF in the context of image retrieval, irrespectively of the particular gaze-tracker that is used. To this end, the empirical gaze-tracking evaluation study reported above is conducted only for roughly demonstrating that the

[5] http://www.flickr.com/

developed gaze-tracker achieves state-of-art performance and that similar image retrieval results, as it will be discussed in Section VI-C, can be obtained with any state-of-art image processing-based tracker. Further image retrieval performance improvements could potentially be obtained if an infrared illumination-based gaze-tracker is used (Section III-A), i.e. if the relevant resources are available and the possible portability issues are not of particular concern.

B. Relevance assessment prediction results

In this section, extensive experimental results regarding the evaluation of the proposed gaze features are reported. For performing the evaluation, 15 subjects participated in the experiments. Each subject underwent 10 gaze-tracking sessions, where in every session he/she was presented a set of 10 images that were randomly chosen from the assembled image dataset. In each session, the user was provided as a query term one of the semantic concepts of set C. It must be highlighted that out of the 10 randomly selected images in each session, 8 were chosen so as to be relevant to the given query and 2 to be irrelevant. Then, the user was asked to observe the images, taking into account the posed query. At the end of each session, the user was presented the objects that he/she has seen and was asked to manually annotate them as relevant or irrelevant to the query, based on his/her understanding of the query term. In this way, a gaze-tracking dataset with associated ground truth annotations was formulated, consisting of a set of 3027 samples in total, out of which 2277 were annotated as relevant and 750 as irrelevant.

In Table III, experimental results from the application of the different feature selection techniques are given. Classification accuracy and relevant (irrelevant) classification rate were used as performance measures, where the former represents the percentage of all samples that were correctly classified and the latter denotes the percentage of the relevant (irrelevant) samples that were correctly identified. All experiments were performed following the 5-fold cross validation approach [52]. From the presented results, it is shown that the PCA technique exhibits the best overall classification performance. This suggests that considering the linear dependencies among the features and choosing those that present the greatest variance is the most efficient methodology for selecting the most discriminative gaze features. The results reported in Table III are estimated for parameter L (i.e. half of the length of $\mathbf{gf}(s_n^m)$) equal to 175 and Z (i.e. length of $\widetilde{\mathbf{gf}}(s_n^m)$) equal to 30, 248, 100, 80 and 60 for PCA, CFS, CSS, GR and IG, respectively. The detailed results, in terms of classification accuracy, for different combinations of values for parameters Z and L obtained using the PCA technique are illustrated in Fig. 8. From the presented results, it can be seen that the maximum performance is observed for $Z = 30$ for most values of parameter L, while significantly lower or higher values of Z lead to decrease in performance. Additionally, it is shown that when selecting 30 features from a pool of 350 available temporal- and spatial-related ones, i.e. when $L = 175$, leads to the best overall performance. Similar behavior, i.e. achieving a maximum classification performance for $L = 175$ and around a particular value of Z, was also observed for the other feature selection techniques.

TABLE III
COMPARISON OF FEATURE SELECTION TECHNIQUES

	PCA	CFS	CSS	GR	IG
Clas. accuracy	**69.71**%	64.19%	63.33%	65.37%	64.50%
Relevant clas. rate	**65.66**%	52.69%	48.33%	49.78%	49.29%
Irrelevant clas. rate	75.75%	80.45%	84.50%	**87.76**%	86.54%

TABLE IV
COMPARISON OF PROPOSED GAZE FEATURES (USING THE PCA TECHNIQUE FOR FEATURE SELECTION)

	Combination of features	Temporal features	Spatial features
Clas. accuracy	**69.71**%	67.04%	66.28%
Relevant clas. rate	65.66%	**80.47**%	53.98%
Irrelevant clas. rate	75.75%	48.38%	**83.94**%

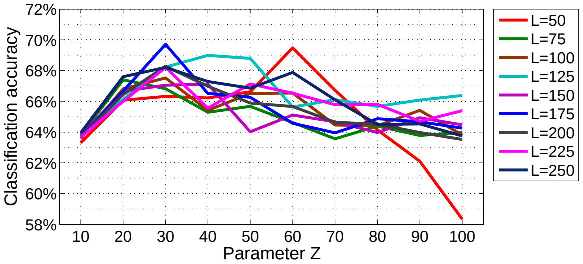


Fig. 8. Prediction of user's relevance assessment using the PCA technique.

The different types of the proposed gaze features are comparatively evaluated in Table IV, where the PCA technique was used for feature selection with $L = 175$ and $Z = 15$. From the presented results, it can be seen that the temporal gaze features lead to increased classification performance, while the combination of the temporal and the spatial features outperforms the performance accomplished when each type of features is used alone. The latter observation demonstrates the usefulness of incorporating the spatial characteristics of the gaze signal, when attempting to predict the user's relevance assessment based on gaze data.

The proposed gaze features are also comparatively evaluated with the features presented in the works of [22], [23], [17], [24], [40], [20] and described in Table I, using the proposed user's relevance assessment predictor (Section IV-D). It must be noted that the definitions of the image-level features described in Table I are appropriately modified so that the features eventually extracted to refer to image regions (e.g. the feature 'times an image was visited' was modified to the 'times a region was visited'). Features, for which a counterpart at the region-level could not be defined, are marked with an asterisk (*) in Table I and were not used in the evaluation. The obtained experimental results are given in Table V. From the presented results, it can be seen that the proposed features significantly outperform all other features of the literature. This demonstrates that the mathematical formulation of the gaze data analysis problem is advantageous compared to the explicit definition of a set of features (as the works of [22], [23], [17], [24], [40] and [20] do). In order to investigate whether the performance difference between the proposed approach and the methods of [22], [23], [17], [24], [40] and [20] is sufficiently large to be also statistically significant, the 'paired t-test' [53] statistical significance test is used. In particular, the following null hypothesis, i.e. the hypothesis that is to be

TABLE V
COMPARATIVE EVALUATION WITH GAZE FEATURES OF THE LITERATURE

Method	Classification accuracy	Relevant clas. rate	Irrelevant clas. rate
Proposed	**69.71**%	**65.66**%	75.75%
[22]	61.60%	47.45%	**82.30**%
[23]	60.10%	47.74%	78.59%
[17]	59.04%	44.57%	80.38%
[24]	58.02%	43.49%	79.45%
[40]	56.07%	50.53%	63.46%
[20]	55.24%	43.29%	72.99%

rejected if the respective test is passed, is defined: "there is no significant difference in the obtained user relevance assessment prediction performance (classification accuracy) between the proposed approach and another similar method of the literature in a gaze-tracking session". For performing the statistical significance test, 150 sessions (i.e. an individual session is considered for each of the 10 supported semantic concepts and each of the 15 performing subjects) are taken into account, resulting in 149 degrees of freedom (df) in the defined t-test. The test revealed that the performance difference between the proposed approach and all the aforementioned state-of-art methods is statistically significant. More specifically, the lowest t-value calculated according to the aforementioned pair-wise method comparisons is ($t-value = 54.7801$, $df = 149$, $P < 0.01$), which corresponds to the performance comparison with the method of [22], i.e. the best performing state-of-art method.

C. Image retrieval results

In this section, image retrieval results from the application of the proposed gaze-based RF approach are presented. For computing the results, the same 15 individuals and the image dataset described in Section VI-B were involved. In particular, 10 image retrieval sessions were performed by each individual, where one of the concepts of C was given again as query term to the user in each session. Every user was initially presented a set of randomly ranked images and asked to observe them, taking into account the query term. Subsequently, the user underwent 5 successive gaze-tracking iterations, where at the beginning of every iteration a new ranking of the images was presented to the user based on the collected gaze data of the previous iteration. In this work, top-20 image retrieval experiments were performed, i.e. the 20 most relevant images were presented to the user in each iteration. The latter choice was considered to be representative of the actual behavior of an average user, who is usually interested in only examining the first very few images that are retrieved based on a query that he/she has posed [54].

Performance is measured using the precision metric, i.e. the percentage of the retrieved images that are relevant to the query. Additionally, the average performance of the proposed approach for each feedback iteration is estimated by calculating the mean precision value taking into account all supported concepts and all users involved, as is typically the case in the literature [55][56][57][58]. Apart from the precision, metrics that take into account the rank of the images in the retrieved results have also been proposed for measuring the image retrieval performance, like Average Precision (AP), Mean Average Precision (MAP) and Normalized Discounted Cumulative Gain (NDCG). The reason that these metrics are not used in this work is twofold: a) Examining the ranking of the images in the adopted top-20 experimental setting could potentially lead to inaccurate performance estimation. b) Secondly, and most importantly, the use of AV, MAP or NDCG would inevitably require the user to be aware of the ranking of the images in the retrieved results, in order to ensure a fair performance measurement with this particular metric. However, the nature of the considered feedback information (gaze signal) requires the user to be unaware of the image ranking, i.e. maintaining that the user's implicit response is captured in a non-intrusive/unbiased way. For that purpose, the images are only presented in tens by the developed interface (Section III-B) and no other information is provided to the user. Including any kind of information regarding the ranking of the retrieved images would very likely violate the fundamental principle of implicit RF methods, i.e. that the user's implicit response should be captured in a non-invasive way.

The average performance of the proposed approach is depicted in Fig. 9(a), while the detailed retrieval results for every individual concept are given in Fig. 9(b). From the presented results, it can be seen that the proposed approach achieves an increase in the mean precision value from 12.47% (random initialization at iteration 0) to 80.43% after five gaze-tracking iterations, i.e. at iteration 5. Additionally, there are concepts whose detection is particularly favored by the proposed approach, i.e. concepts forest, desert and sea. These concepts are composed of real-world objects (or their constituent parts) with characteristic visual appearance. However, there are also concepts that do not exhibit that increased improvement in their retrieval performance, such as the concepts car, living-room and road. For these concepts the significant variance of their constituent objects, in terms of low-level visual features, hinders further performance improvement. The above results justify the fact that the region-based analysis of the gaze signal can lead to increased image retrieval performance.

The proposed approach is also comparatively evaluated with two representative explicit RF methods of the literature for image retrieval, namely the Local Neighboring Movement (LNM) method presented in [6] and the approach of [10]. In particular, the LNM method constitutes a variant of the traditional global-level Query Point Movement approach [59], where previously checked images are not re-examined in subsequent iterations. On the other hand, the method of [10] is representative of the so called region-level image RF category and it estimates a correspondence between the image regions, following an inexact graph matching methodology for estimating the degree of similarity between the respective images. It must be noted that both implemented methods used the same features with the proposed approach, i.e. SIFT-based BoWs at global- and local-level, instead of the simpler ones originally proposed in the works of [6] and [10], respectively. Additionally, for the method of [10], which requires an initialization of the 'query image', the first image annotated by the user as relevant was used as the initial query image.

The average retrieval results from the application of the LNM approach [6] and the method of [10] are depicted in Fig. 9(a), while the respective per concept results are shown

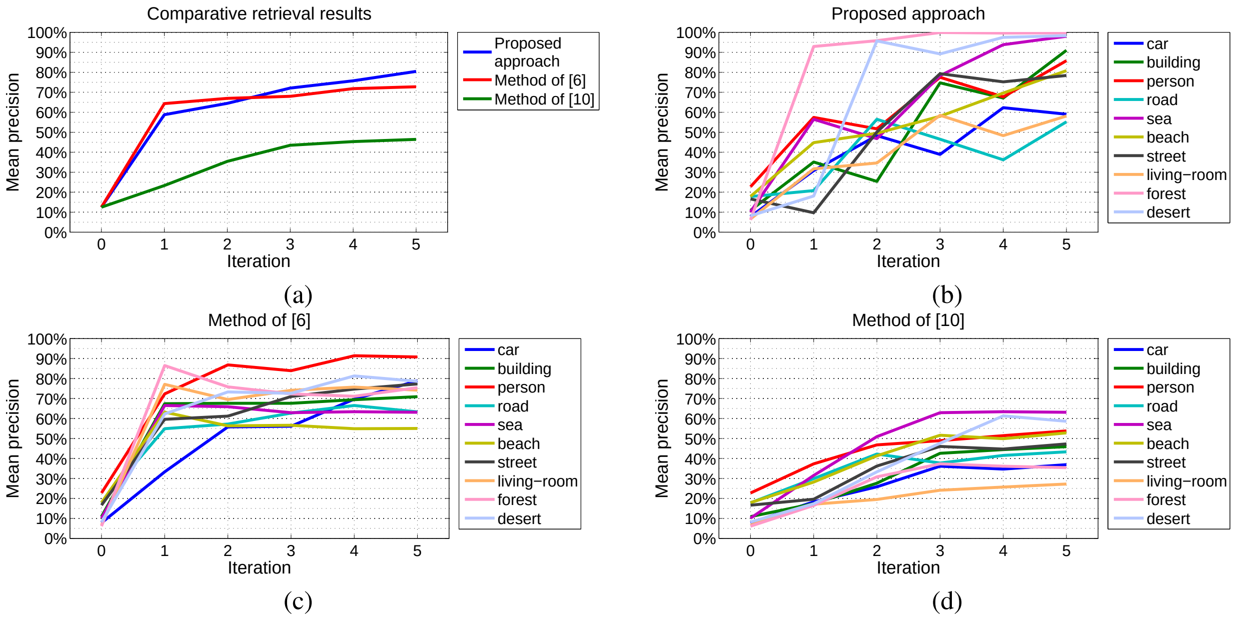


Fig. 9. Image retrieval results: (a) average performance of all methods, (b) per concept performance of the proposed approach, (c) per concept performance of LNM [6] method, and (d) per concept performance of method of [10].

in Figs. 9(c) and 9(d). From the presented results, it can be seen that the proposed approach outperforms both state-of-art methods, exhibiting 7.68% higher precision performance than the LNM approach [6] and 34.01% compared to the method of [10]. These observations suggest that: a) receiving feedback information at region-level can overcome the limitations of global-level RF methodologies, and b) exploiting implicit (i.e. gaze-tracking) RF information can lead to improved image retrieval performance, compared to explicit RF approaches.

Examining the results in more details, it must be highlighted that the LNM method and the approach of [10] are favored due to the fact that the users were asked to provide feedback information regarding the relevance of all retrieved images. In this way, the aforementioned methods received greater amounts of feedback information, compared to the proposed approach. For instance, when the retrieved results were sufficiently good (i.e. precision $> 70\%$), it was noted that the users typically observed only some of the relevant images. However, despite the aforementioned fact, the proposed approach outperforms the best performing LNM method, given that a minimum number of gaze-tracking iterations, which according to the conducted experiments were shown to be equal to 3 (Fig. 9(a)), have taken place. The latter implies that learning from region-level feedback information may be slower during the first few RF iterations compared to the case of using global-level feedback, but region-level information can lead to improved retrieval performance. On the other hand, the method of [10] achieves inferior performance compared to the proposed approach, due to the frequently inaccurate region-correspondence estimation, as it is also mentioned in the original text of this work. The latter suggests that directly selecting the image regions and estimating their degree of relevance, even with the use of an imperfect predictor (i.e. the interpreter of the gaze signal), can lead to significantly increased performance, compared to explicit RF methods that receive as input global-level feedback and attempt to estimate a region correspondence between the examined images. Similarly to the case in Section VI-B, the ‘paired t-test’ [53] statistical significance test is also used in this section, in order to investigate whether the performance difference between the proposed RF approach and the methods of [6] and [10] is sufficiently large to be also statistically significant. The null hypothesis is now defined as follows: “there is no significant difference in the obtained image retrieval performance (retrieval precision) between the proposed approach and another similar method of the literature in a relevance feedback session”. The test, which has 149 degrees of freedom (as in Section VI-B), showed that the performance difference between the proposed approach and the aforementioned state-of-art methods is again statistically significant. More specifically, the lowest calculated t-value is ($t-value = 5.0687$, $df = 149$, $P < 0.01$) and corresponds to the performance comparison with the method of [6], i.e. the best performing one.

D. *Time efficiency evaluation*

Time efficiency in image RF applications concerns two factors: a) time required for user response capturing and b) time needed for estimating an updated set of results. The estimated average time for each of the aforementioned procedures, as well as their summation, for the proposed approach and the methods of [6] and [10] are given in Fig. 10. It must be noted that time performance is measured for every feedback iteration, while considering the experimental framework described in Section VI-C for image retrieval performance evaluation. Additionally, the reported times were obtained using a PC with the same specifications described in Section VI-A. From the presented results, it can be seen that the time required for Response Capturing (RC) is almost half for the proposed gaze-based approach, compared to the respective times needed for the methods of [6] and [10]. This is due to the explicit response, which involves human judgment/manual annotation processes, required by both methods. On the contrary, the proposed approach only requires the tracking of the user's gaze. Regarding the Feedback Interpretation (FI) step, the global-RF method of [6] is the best performing one, while the region-based RF method of [10] is by far the slowest. The latter is mainly due to the time consuming procedures (i.e. inexact graph matching) that the method of [10] adopts for localizing the regions of interest in the observed images. The proposed region-based RF method significantly outperforms the approach of [10], mainly due to the more time-efficient gaze-based methodology for predicting the regions of interest. Examining the overall time required for both RC and FI procedures, it can be seen that the time performance of the proposed approach is competitive to that of the global-RF method of [6], while it significantly outperforms the one of the method of [10]. The latter suggests that the proposed RF approach combines increased time efficiency and improved retrieval performance (Section VI-C). Additionally, the implicit way that user feedback is captured renders the proposed system user-friendly and easy to use, which is a desirable characteristic especially for large-scale applications. It must be highlighted that the aforementioned time performances were measured without applying any particular optimizations to any of the above methods. In case that significantly larger datasets than the utilized one are to be used, significant time performance improvements can be obtained by employing GPU implementations, incorporating indexing structures, performing code optimizations, applying incremental clustering of the estimated regions of interest for constraining the time performance of the proposed approach, etc.

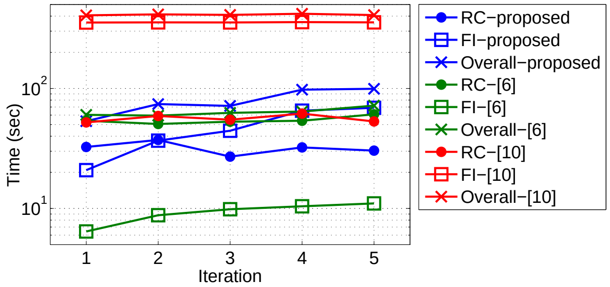


Fig. 10. Time performance evaluation considering the Response Capturing (RC) and Feedback Interpretation (FI) procedures. The reported times in the vertical axis are in logarithmic scale.

E. Discussion on relevance assessment prediction

The performance of gaze interpretation algorithms is affected by a wide series of factors. The most commonly met ones can be roughly categorized in the following classes: a) visual appearance-related (e.g. salient objects in the images, positioning of the images on the screen, image depth, images' aspect ratio, etc.), b) environmental (e.g. acoustic noise, illumination changes, interruptions during the gaze-tracking session, etc.) and c) psychological (e.g. users' personality, mood of the users during the session, etc.). The significant difficulty in efficiently modeling the above factors reveals the great complexity of the gaze interpretation problem and also highlights that significant performance improvements can still be accomplished (the proposed approach achieves approximately 69.71% in user relevance assessment prediction accuracy, as described in Section VI-B). In this work, a two-stage approach is followed for efficiently tackling the factors of the first class, namely at a) the gaze signal capturing/processing stage and b) the response modeling level. In particular, visual appearance-related factors typically cause sudden movements of the gaze, i.e. saccades (Section IV-B). However, the proposed gaze features, which are used for predicting the user's relevance assessment, are computed taking into account only the observed fixations (Section IV-B). Additionally, the zoom-in/zoom-out functionalities of the developed interface further facilitate in eliminating undesirable gaze distortions, as detailed in Section III-B. In case that gaze deteriorations consist of fixations that occurred during the 'zoom-in-image' mode, the employed discriminative learning classifier (Section IV-D) aims among others at discriminating them from the respective user behavior types that correspond to image regions that are truly of interest. It must be reminded that the employed classifier receives as input manual annotations of the captured user gaze responses during the training stage. In order to reduce the detrimental effects caused by the environmental factors, particularly increased attention was given so that the users to remain un-obscured during the conducted gaze-tracking sessions. Moreover, the deviations caused by the third category of factors, i.e. the psychological related ones, are mainly encountered by the significantly increased expressiveness of the proposed gaze features (Section IV-B). However, since this aspect of the problem (i.e. the psychological state of the user during the gaze-tracking session) is identified as the biggest challenge in the interpretation of the gaze signal, it is considered as future work with a great potential to increase the user relevance assessment prediction performance, as it will be discussed in the sequel.

VII. Conclusions

In this paper, a novel gaze-based RF approach to region-based image retrieval was presented. Aim of the overall approach was to iteratively estimate the real-world objects (or their constituent parts) that are of interest to the user and subsequently use this information for refining the image retrieval results. A novel set of region-level gaze features, which represent both the temporal and spatial characteristics of the gaze signal, was presented for performing user's relevance assessment prediction. Extensive experiments demonstrated their efficiency, compared to other features presented in the literature. Additionally, an object-based RF mechanism was developed, which handles the main limitation of region-based RF approaches, i.e. the inaccurate estimation of the regions of interest in the retrieved images, in a satisfactory way. The experimental evaluation proved that the proposed approach outperforms representative global- and region-based explicit RF approaches of the literature, using a challenging general-purpose image dataset. Moreover, the incorporation of a single-camera image processing-based gaze tracker makes the overall system cost efficient and portable. From the reported experimental results, it is shown that there still exists a strong potential for further improving the prediction of the user's relevance assessment based on gaze data. Towards this goal, future work includes the investigation and modeling of the factors that affect the way that users see (e.g. personality, mood, etc.) and their integration to the developed framework.

Acknowledgment

The work presented in this paper was supported by the European Commission under contract FP7-258749 CEEDs.

References

[1] R. Datta, D. Joshi, J. Li, and J. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Computing Surveys (CSUR)*, vol. 40, no. 2, p. 5, 2008.

[2] A. Hanjalic, R. Lienhart, W. Ma, and J. Smith, "The holy grail of multimedia information retrieval: So close or yet so far away?" *Proceedings of the IEEE*, vol. 96, no. 4, pp. 541–547, 2008.

[3] X. Zhou and T. Huang, "Relevance feedback in image retrieval: A comprehensive review," *Multimedia systems*, vol. 8, no. 6, pp. 536–544, 2003.

[4] D. Tao, X. Tang, X. Li, and X. Wu, "Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval," *Pattern Analysis and Machine Intelligence, IEEE Trans. on*, vol. 28, no. 7, pp. 1088–1099, 2006.

[5] X. Tian, D. Tao, X.-S. Hua, and X. Wu, "Active reranking for web image search," *Image Processing, IEEE Trans. on*, vol. 19, no. 3, pp. 805–820, 2010.

[6] D. Liu, K. Hua, K. Vu, and N. Yu, "Fast query point movement techniques for large cbir systems," *Knowledge and Data Engineering, IEEE Trans. on*, vol. 21, no. 5, pp. 729–743, 2009.

[7] D. Tao, X. Tang, X. Li, and Y. Rui, "Direct kernel biased discriminant analysis: a new content-based image retrieval relevance feedback algorithm," *Multimedia, IEEE Trans. on*, vol. 8, no. 4, pp. 716–727, 2006.

[8] X. Tian, D. Tao, and Y. Rui, "Sparse transfer learning for interactive video search reranking," *ACM Trans. on Multimedia Computing, Communications, and Applications (TOMCCAP)*, vol. 8, no. 3, p. 26, 2012.

[9] W. Jiang, G. Er, Q. Dai, and J. Gu, "Similarity-based online feature selection in content-based image retrieval," *Image Processing, IEEE Trans. on*, vol. 15, no. 3, pp. 702–712, 2006.

[10] C. Li and C. Hsu, "Image retrieval with relevance feedback based on graph-theoretic region correspondence estimation," *Multimedia, IEEE Trans. on*, vol. 10, no. 3, pp. 447–456, 2008.

[11] D. Djordjevic and E. Izquierdo, "An object-and user-driven system for semantic-based image annotation and retrieval," *Circuits and Systems for Video Technology, IEEE Trans. on*, vol. 17, no. 3, pp. 313–323, 2007.

[12] G. Aggarwal, T. Ashwin, and S. Ghosal, "An image retrieval system with automatic query modification," *Multimedia, IEEE Trans. on*, vol. 4, no. 2, pp. 201–214, 2002.

[13] M. Kherfi and D. Ziou, "Relevance feedback for cbir: a new approach based on probabilistic feature weighting with positive and negative examples," *Image Processing, IEEE Trans. on*, vol. 15, no. 4, pp. 1017–1030, 2006.

[14] D. Kelly and J. Teevan, "Implicit feedback for inferring user preference: a bibliography," in *ACM SIGIR Forum*, vol. 37, no. 2. ACM, 2003, pp. 18–28.

[15] E. Cheng, F. Jing, L. Zhang, and H. Jin, "Scalable relevance feedback using click-through data for web image retrieval," in *Proc. of the 14th annual ACM Int. Conf. on Multimedia*. ACM, 2006, pp. 173–176.

[16] J. Wang, E. Pohlmeyer, B. Hanna, Y. Jiang, P. Sajda, and S. Chang, "Brain state decoding for rapid image retrieval," in *Proc. of the 17th ACM Int. Conf. on Multimedia*. ACM, 2009, pp. 945–954.

[17] L. Kozma, A. Klami, and S. Kaski, "Gazir: Gaze-based zooming interface for image retrieval," in *Proc. of the 2009 Int. Conf. on Multimodal interfaces*. ACM, 2009, pp. 305–312.

[18] J. Sang, C. Xu, and D. Lu, "Learn to personalized image search from the photo sharing websites," *Multimedia, IEEE Transactions on*, vol. 14, no. 4, pp. 963–974, 2012.

[19] S. Liu, P. Cui, H. Luan, W. Zhu, S. Yang, and Q. Tian, "Social visual image ranking for web image search," in *Advances in Multimedia Modeling*. Springer, 2013, pp. 239–249.

[20] A. Klami, C. Saunders, T. De Campos, and S. Kaski, "Can relevance of images be inferred from eye movements?" in *Proc. of the 1st ACM Int. Conf. on Multimedia information retrieval*. ACM, 2008, pp. 134–140.

[21] A. Faro, D. Giordano, C. Pino, and C. Spampinato, "Visual attention for implicit relevance feedback in a content based image retrieval," in *Proc. of the 2010 Symposium on Eye-Tracking Research & Applications*. ACM, 2010, pp. 73–76.

[22] S. Hajimirza, M. Proulx, and E. Izquierdo, "Reading users' minds from their eyes: A method for implicit image annotation," *Multimedia, IEEE Trans. on*, vol. 14, no. 3, pp. 805–815, 2012.

[23] D. Hardoon and K. Pasupa, "Image ranking with implicit feedback from eye movements," in *Proc. of the 2010 Symposium on Eye-Tracking Research & Applications*. ACM, 2010, pp. 291–298.

[24] A. Klami, "Inferring task-relevant image regions from gaze data," in *Workshop on Machine Learning for Signal Processing. IEEE*, 2010.

[25] F. Jing, M. Li, H.-J. Zhang, and B. Zhang, "Relevance feedback in region-based image retrieval," *Circuits and Systems for Video Technology, IEEE Trans. on*, vol. 14, no. 5, pp. 672–681, 2004.

[26] L. Terissi and J. Gómez, "3d head pose and facial expression tracking using a single camera," *Journal of Universal Computer Science*, vol. 16, no. 6, pp. 903–920, 2010.

[27] J. Ahlberg, "Candide-3–an updated parameterized face," *Report No. LiTH-ISY*, 2001.

[28] D. DeMenthon and L. Davis, "Model-based object pose in 25 lines of code," *Int. Journal of Computer Vision*, vol. 15, no. 1, pp. 123–141, 1995.

[29] B. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. of the 7th Int. joint Conf. on Artificial intelligence*, 1981.

[30] N. Otsu, "A threshold selection method from gray-level histograms," *Automatica*, vol. 11, no. 285-296, pp. 23–27, 1975.

[31] J. Zhu and J. Yang, "Subpixel eye gaze tracking," in *Automatic Face and Gesture Recognition, Proc. Fifth IEEE Int. Conf. on*. IEEE, 2002, pp. 124–129.

[32] V. Mezaris, I. Kompatsiaris, and M. Strintzis, "Still image segmentation tools for object-based multimedia applications," *Int. Journal of pattern recognition and artificial intelligence*, vol. 18, no. 04, pp. 701–725, 2004.

[33] G. T. Papadopoulos, C. Saathoff, H. J. Escalante, V. Mezaris, I. Kompatsiaris, and M. G. Strintzis, "A comparative study of object-level spatial context techniques for semantic image analysis," *Computer Vision and Image Understanding*, vol. 115, no. 9, pp. 1288–1307, Sept. 2011.

[34] K. Van De Sande, T. Gevers, and C. Snoek, "Evaluating color descriptors for object and scene recognition," *Pattern Analysis and Machine Intelligence, IEEE Trans. on*, vol. 32, no. 9, pp. 1582–1596, 2010.

[35] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Workshop on statistical learning in computer vision, ECCV*, vol. 1, 2004, p. 22.

[36] K. Rayner, "Eye movements in reading and information processing: 20 years of research." *Psychological bulletin*, vol. 124, no. 3, p. 372, 1998.

[37] E. Kowler, "Eye movements: The past 25years," *Vision research*, vol. 51, no. 13, pp. 1457–1483, 2011.

[38] D. Salvucci and J. Goldberg, "Identifying fixations and saccades in eye-tracking protocols," in *Proc. of the 2000 symposium on Eye tracking research & applications*. ACM, 2000, pp. 71–78.

[39] H. Widdel, "Operational problems in analysing eye movements," in *Proc. of The Second European Conf. on Eye Movements*, 1984, vol. 22, pp. 21 – 29.

[40] Y. Zhang, H. Fu, Z. Liang, Z. Chi, and D. Feng, "Eye movement as an interaction mechanism for relevance feedback in a content-based image retrieval system," in *Proc. of the 2010 Symposium on Eye-Tracking Research & Applications*. ACM, 2010, pp. 37–40.

[41] R. Harris, *A primer of multivariate statistics*. Lawrence Erlbaum, 2001.

[42] M. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, The University of Waikato, 1999.

[43] N. Johnson, S. Kotz, and N. Balakrishnan, "Continuous univariate distributions. vol. 1," 1994.

[44] T. Mitchell *et al.*, "Machine learning," 1997.

[45] V. Vapnik, *The nature of statistical learning theory*. springer, 1999.

[46] D. Tax and R. Duin, "Using two-class classifiers for multiclass classification," in *Pattern Recognition, Proc. 16th Int. Conf. on*, vol. 2. IEEE, 2002, pp. 124–127.

[47] T. Ohno, N. Mukawa, and A. Yoshikawa, "Freegaze: a gaze tracking system for everyday gaze interaction," in *Proc. of the 2002 symposium on Eye tracking research & applications*. ACM, 2002, pp. 125–132.

[48] T. Ishikawa, S. Baker, I. Matthews, and T. Kanade, *Passive driver gaze tracking with active appearance models*. Carnegie Mellon University, the Robotics Institute, 2004.

[49] E. Pogalin, A. Redert, I. Patras, and E. Hendriks, "Gaze tracking by using factorized likelihoods particle filtering and stereo vision," in *3D Data Processing, Visualization, and Transmission, Third Int.l Symposium on*. IEEE, 2006, pp. 57–64.

[50] J. Chen, Y. Tong, W. Gray, and Q. Ji, "A robust 3d eye gaze tracking system using noise reduction," in *Proc. of the 2008 symposium on Eye tracking research & applications*. ACM, 2008, pp. 189–196.

[51] E. Lee and K. Park, "A robust eye gaze tracking method based on a virtual eyeball model," *Machine Vision and Applications*, vol. 20, no. 5, pp. 319–337, 2009.

[52] R. Kohavi *et al.*, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Int. joint Conf. on artificial intelligence*, vol. 14. Lawrence Erlbaum Associates Ltd, 1995, pp. 1137–1145.

[53] J. O'Connor and E. Robertson, "Student's t-test," *MacTutor History of Mathematics archive, University of St Andrews*.

[54] N. Vasconcelos and A. Lippman, "Learning from user feedback in image retrieval systems," *Advances in neural information processing systems*, vol. 12, pp. 977–983, 1999.

[55] Y. Yang, F. Nie, D. Xu, J. Luo, Y. Zhuang, and Y. Pan, "A multimedia retrieval framework based on semi-supervised ranking and relevance feedback," *Pattern Analysis and Machine Intelligence, IEEE Trans. on*, vol. 34, no. 4, pp. 723–742, 2012.

[56] L. Zhang, L. Wang, and W. Lin, "Generalized biased discriminant analysis for content-based image retrieval," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Trans. on*, vol. 42, no. 1, pp. 282–290, 2012.

[57] J.-H. Su, W.-J. Huang, P. S. Yu, and V. S. Tseng, "Efficient relevance feedback for content-based image retrieval by mining user navigation patterns," *Knowledge and Data Engineering, IEEE Trans. on*, vol. 23, no. 3, pp. 360–372, 2011.

[58] A. Grigorova, F. De Natale, C. Dagli, and T. Huang, "Content-based image retrieval by feature adaptation and relevance feedback," *Multimedia, IEEE Trans. on*, vol. 9, no. 6, pp. 1183–1192, 2007.

[59] Y. Ishikawa, R. Subramanya, and C. Faloutsos, "Mindreader: Querying databases through multiple examples," *Computer Science Department*, p. 551, 1998.

Georgios Th. Papadopoulos (S'08, M'11) was born in Thessaloniki, Greece, in 1982. He received the Diploma and Ph.D. degrees in Electrical and Computer Engineering from the Aristotle University of Thessaloniki (AUTH), Thessaloniki, Greece, in 2005 and 2011, respectively.

He is currently a Post-doctoral Researcher at the Information Technologies Institute (ITI) of the Centre for Research and Technology Hellas (CERTH), Thessaloniki, Greece. He has published 8 international journal articles and is a coauthor of 21 international conference proceedings. His research interests include computer vision, pattern recognition, semantic multimedia analysis, image and video processing, relevance feedback, context-based analysis and machine learning techniques.

Dr. Papadopoulos is a member of the Technical Chamber of Greece.

Konstantinos C. Apostolakis received his university degree in Computer Science from the Computer Science Department of the Aristotle University of Thessaloniki in September 2009. Since 2007 and until graduating in 2009, he was a member of the "Talos RoboCup Soccer Simulation Team" of the Aristotle University of Thessaloniki, where he worked as a programmer on the implementation of a virtual soccer team of autonomous software agents. As of May 2011, he works as a research associate in the Information Technologies Institute of the Centre for Research and Technology Hellas.

Petros Daras (M'07, SM'13) was born in Athens, Greece, in 1974. He received the Diploma degree in Electrical and Computer Engineering, the M.Sc. degree in Medical Informatics, and the Ph.D. degree in Electrical and Computer Engineering, all from the Aristotle University of Thessaloniki, Greece, in 1999, 2002, and 2005, respectively. He is a Researcher Grade B (Assoc. Professor), at the Information Technologies Institute (ITI) of the Centre for Research and Technology Hellas (CERTH). His main research interests include 3-D object processing & reconstruction, search, retrieval and recognition of 3-D objects, medical informatics, medical image processing, 3-D object watermarking, and bioinformatics. Dr. Daras is the chair of the IEEE MMTC IMVIMEC IG and a key member of the IEEE MMTC 3DRPC IG. He is a senior Member of IEEE.