Human-centered Region Selection and Weighting for Image Retrieval

Jean Martinet
LIFL/CNRS-UMR 8022, Université Lille 1, Lille, France

Keywords: Gaze Tracking, Image Indexing and Retrieval, Weighting Scheme, Human-centered Computing.

Abstract: We present an application of gaze tracking to image and video indexing, in the form of a model for selecting and weighting Regions of Interest (RoIs). Image/video indexing refers to the process of creating a synthetic representation of the media, for instance for retrieval purposes. It usually consists in labeling the media with semantic keywords describing its content. When automatized, this process is based on the analysis of visual features, which can be extracted either from the whole image or keyframe, or locally from regions. Since most of the times the whole image is not relevant for indexing (e.g. large flat regions with no specific semantic interpretation, blur regions, background regions that may not be relevant for retrieval purposes, and that should be filtered out), it would be preferable to concentrate the labeling process on specific RoIs that are considered representative of the scene, like the main subjects. The objective of the work presented here is to take advantage of natural human gaze information in order to define a human-centered Region of Interest selection and weighting technique in the context of media retrieval.

1 INTRODUCTION

Automatic image annotation refers to the process of labeling image content – e.g. with semantic keywords, generally for retrieval purposes. This process is based on the analysis of image features, which can be extracted either from the whole image or from image regions. Since most of the times the whole image is not relevant for indexing, it would be preferable to concentrate the analysis on so-called Regions of Interest (RoIs), which represent the most important – or representative – parts of the picture, like the subject.

With the recent development of low-cost gaze tracker devices, the possibility of taking advantage of the information conveyed in gaze has opened many research directions, namely in image compression – where users' gaze is used to set variable compression ratios at different places in an image, in marketing – for detecting products of interest for customers, civil security – for detecting drowsiness or lack of concentration of persons operating machinery such as motor vehicles or air traffic control systems, and in human-computer interaction. In the latter for instance, the user's gaze is used as a complementary input device to traditional ones such as a mouse and a keyboard, namely for disabled users. The goal of our work is to use gaze information as an input for developing a human-centered Region of Interest selection and weighting technique. We present in the remainder of this paper some preliminary works in this direction.

The paper is organized as follows. Section 2 gives an overview of region selection and weighting techniques. Section 3 presents a brief history of research related to gaze tracking. We describe in Section 4 how RoIs are defined by gathering gaze information from several persons. The next step is to define an importance measure for RoIs in an image, so that corresponding labels can be weighted (Section 5). Finally, we describe how to apply and integrate this model in a retrieval system in Section 6. We report and discuss experimental results in Sections 7 and 8, and Section 9 gives a conclusion of this contribution, including some insights for further works.

2 IMAGE REGION SELECTION/WEIGHTING

While keyword selection and weighting techniques are widely used in text retrieval systems, such techniques are seldom applied for image retrieval. The importance of image regions has been considered in some publications. In (Osberger and Maeder, 1998), Osberger and Maeder have defined a way to identify perceptually important regions in an image based on human visual attention and eye movement characteristics. In a similar way, Itti and Koch (Itti et al.,

Martinet J..
Human-centered Region Selection and Weighting for Image Retrieval.
DOI: 10.5220/0004348707290734
In *Proceedings of the International Conference on Computer Vision Theory and Applications* (VISAPP-2013), pages 729-734
ISBN: 978-989-8565-47-1

1998) have developed a visual attention system based on the early primate vision system for scene analysis. Later, Stentiford (Stentiford, 2003) applied the visual attention to similarity matching. Wang (Wang et al., 2001; Wang and Du, 2001) introduced in SIMPLYcity the *region frequency* and *inverse picture frequency* (*rf·ipf*), a region-based measure for image retrieval purposes, inspired from the *tf·idf* weighting scheme for text retrieval. Jing et al. (Jing et al., 2002) have introduced a region weighting scheme that is based on the users' relevance feedback information. The work presented here is aimed at using the natural users' gaze information for image region selection and weighting.

3 GAZE TRACKING

The analysis of gaze has been studied for over a century in several disciplines, including physiology, psychology, psychoanalysis, and cognitive sciences. The purpose is to analyze eye saccades and fixations of persons watching a given scene, in order to extract several kinds of information.

3.1 Techniques and Systems

In order to detect and track users' gaze, it is necessary to employ a gaze tracking device which is able to determine the fixation point of a user on a screen from the position of their eye. Earliest gaze trackers were very intrusive. In addition to constrain the user to be totally static, they were in direct contact with him by sticking a reflective white dot directly onto the eye or attaching a number of electrodes around the eye.

Nowadays, the most accurate gaze tracking systems generally consist of head mounted devices which allow detecting the direction of the gaze without having to cope with the pose of the user's head. These trackers are also intrusive; they consist of devices generally composed with 3 cameras mounted on a padded headband (2 eye cameras to allow binocular eye tracking with built-in light sources, and 1 camera to allow accurate tracking of the user's point of gaze).

Non-intrusive gaze tracking systems usually require a static camera capturing the face of the user and detecting the direction of their gaze with respect to a known position. A basic gaze tracker system is composed with a static camera, a display device and software to provide an interface between them. The precision of the system can be increased by different ways, like adding a specific light source such as an infrared beam – in order to create reflections on the eye and produce more accurate tracking.

3.2 Visual Attention

During visual perception, human eyes move and successively fixate at the most informative parts of the image (Yarbus, 1967). While RoIs can be defined in various ways according to the requirement, the presented indexing model is based on visual attention. Attention is the cognitive process of selectively concentrating on one aspect of the environment while ignoring other things. For images and videos, the visual attention is at the core of the visual perception, because it drives the gaze to salient points in the scene.

4 IDENTIFICATION OF REGIONS OF INTEREST

Gaze trackers provide the horizontal and vertical coordinates of the point of regard relative to the display device. Thus, one can easily obtain a sequence of points corresponding to the sampled positions of the eye direction. These points correspond to triplets of the form (x,y,t) and reflect the scan path. The scan path consists of a sequence of eye fixations (gaze is kept still in a location, yielding regions with important density of points), separated by eye saccades (fast movement, yielding large spaces with only few isolated points).

4.1 Fixations and Saccades

An essential part in scan path analysis and processing is the identification of fixations and saccades. Indeed, fixations and saccades are often used as basic source for the various metrics that are used for interpreting eye movements (number of fixations, saccades, duration of the first fixation, average amplitude of saccades, etc.) (Jacob and Karn, 2004; Poole et al., 2004).

The most widespread identification technique is by computing the velocity of each point (defined as the angular speed of the eye in degrees per second). The velocity of a point corresponds to the distance that separates it from its predecessor or successor. Separation of points into fixations and saccades is achieved by using a velocity threshold. Successive points labelled fixations are then grouped into what can be considered as an eye fixation. Another threshold involving a minimal duration of a fixation allows eliminating insignificant groups.

4.2 Point Clustering

After combining collected data from several participants by merging all their fixations into a single set, a clustering process allows reducing the spatial characteristics of fixations into a limited subset of clusters K_i, which define the RoIs (see Figure 1).

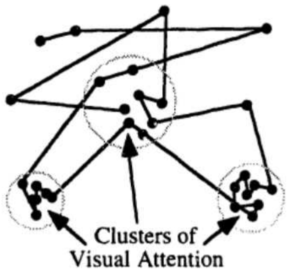

Figure 1: Clustered fixation points highlight conspicuous locations in the scene, allowing to define the Regions of Interest.

The clustering is to be achieved by unsupervised techniques, such as K-means, because the spatial distribution of points from all scan paths is unknown, and their number is finite.

Once a set of RoIs is identified for an image, corresponding objects can be labeled, and the labels can be given more or less importance according to their importance in the image. The final objective is to integrate this generated index into a retrieval system.

5 REGION OF INTEREST IMPORTANCE

In the context of image indexing for retrieval purpose, it is a central problem to assign weights to index items in order to give emphasis to important ones, quantifying how well they describe and represent documents (Martinet et al., 2008). For instance, in text retrieval, popular weighting scheme such as $tf \times idf$ are based on two sources, one representing the local importance of the word in the document, and the other representing the global importance of the word in the whole collection (Baeza-Yates and Ribeiro-Neto, 1999; Salton, 1971). The importance measure of an RoI in an image can be interpreted in the same way as a tf measure of a word in a text document.

5.1 Cluster Importance Criteria

Having the main Regions of Interest defined as described in Section 4, we are interested in identifying criteria to estimate their relative importance, based on the characteristics of the clusters. Inspired from (Nguyen et al., 2006), we identify the following measures:

- **The Cardinal $C(K_i)$ of a Cluster:** the cardinal of a cluster K_i measures the number of gaze points that belong the cluster. This measure is related to the duration of gaze within the cluster when a uniform gaze sampling is recorded. Hence, it represents the total time spent viewing/gazing at the corresponding region. The cluster importance is conjectured to be proportional to the cardinal $C(K_i)$: $importance(K_i) \propto C(K_i)$. The cardinal $C(K_i)$ of a cluster is defined as:

$$C(K_i) = \frac{1}{n_p} \sum_{p \in K_i} 1 \qquad (1)$$

where p denotes the successive recorded gaze points from all participants, and n_p is the total number of recorded points.

- **The Surface $S(K_i)$ of a Cluster:** the surface of a cluster measures the area (in pixels) of the region covered by gaze points in the cluster. The surface $S(K_i)$ is conjectured to be inversely proportional to the cluster importance. Indeed, for a given number of gaze points, a smaller surface indicates a higher concentration of points: $importance(K_i) \propto \frac{1}{S(K_i)}$. The surface $S(K_i)$ of a cluster is given by the surface of its convex hull:

$$S(K_i) = convexHullSurface(\{p | p \in K_i\}) \qquad (2)$$

- **The Variance $V(K_i)$ of a Cluster:** the variance of gaze points from their cluster mean. The variance $V(K_i)$ is conjectured to be inversely proportional to the cluster importance : $importance(K_i) \propto \frac{1}{V(K_i)}$. It is defined by the following equation:

$$V(K_i) = \sum_{p \in K_i} (p - \mu_i)^2 \qquad (3)$$

where μ_i represents the centroid of K_i.

- **The time-weighted Visit Count $W(K_i)$ of a Cluster:** the time-weighted visit count is used to weight the count (or duration) of the k^{th} cluster visited by the inverse of the cluster visit time.

$$W(K_i) = \sum_{p \in K_i} \frac{1}{rank_p} \qquad (4)$$

where $rank_p$ denotes the cluster visit rank of point p in the scan path (i.e. 1 for the first visited cluster, 2 for the second, and so on). This measure gives more importance to the first sampled points.

- **The Revisit Count $R(i)$ of a Cluster:** the revisit count measures the number of saccade revisits to a given cluster during the scan path. The revisiting of the fixation-saccade path to regions in an image has been found to be a fundamental property of eye movements. The cluster importance is conjectured to be proportional to the number of cluster revisits : $importance(K_i) \propto R(K_i)$. It is defined by the following equation:

$$R(K_i) = card(\{p | (p \in K_i) \wedge (rank_p > rank_q) ||\}) \quad (5)$$

where $card(X)$ is the cardinal of the set X, and $rank_q$ is the lowest rank of visit to cluster K_i – this is to count in the cluster only points *after* the first visit to the cluster.

5.2 Criteria Combination

The application of gaze tracking to image indexing is done under the hypothesis that the importance of an image region is related to its conspicuity. The notion of importance of the RoI is calculated on the basis of the measures presented above. The total importance of an RoI is obtained by combining these normalized values.

$$\begin{aligned} importance(K_i) = k \times (\quad & \alpha_C \cdot \frac{C(K_i)}{\sum_i C(K_i)} + \\ & \alpha_S \cdot \frac{S(K_i)}{\sum_i S(K_i)} + \\ & \alpha_V \cdot \frac{V(K_i)}{\sum_i V(K_i)} + \\ & \alpha_W \cdot \frac{W(K_i)}{\sum_i W(K_i)} + \\ & \alpha_R \cdot \frac{R(K_i)}{\sum_i R(K_i)}) \end{aligned} \quad (6)$$

where

$$k = \frac{1}{\alpha_C + \alpha_S + \alpha_V + \alpha_W + \alpha_R}$$

is a normalizing constant, and

$$(\alpha_C, \alpha_S, \alpha_V, \alpha_W, \alpha_R)$$

is the vector of parameters allowing to set the relative importance of each criterion.

6 APPLICATION TO IMAGE INDEXING AND RETRIEVAL

Image indexing consists in creating a synthetic representation of an image, e.g. by assigning semantic descriptors (keywords) to images describing their content. There are several ways to automatically generate semantic descriptors for an image. Most popular techniques are based on a learning process, using Support Vector Machines for instance. In this process, several samples of a given concept are used to feed a training algorithm, which is later use to detect the presence of the concept in a test image.

Based on the previous definition of RoIs, the annotation process can be concentrated on RoI locations, since they represent the main areas that people perceive in the media. As a consequence, the indexing tool will focus the process on the most important regions in the image, by extracting visual features (such as color, texture, etc.) preferably from these specific locations, generate labels to be assigned to them together with the estimated importance value. At query time, the search engine will then the query features with region features, considering their importance. This makes the search process more efficient. The importance of a given index term regarding an image is related to the *aboutness* of the image considering the term. The use of RoI importance as weighting method for index terms is driven by the assumption that if most users concentrate their attention on a specific object, then the *aboutness* of the image considering the object is high.

We can extend the indexing and retrieval with semantics. In this case, we integrate a weighting scheme to semantic concepts associated to the most important RoIs. As indicated in (Salton, 1971), weighting schemes are often based on the formula $tf \times idf$. The importance of label (or index term) attached to an RoI is considered the visual counterpart of the tf measure for text. The idf value can be computed traditionally (Baeza-Yates and Ribeiro-Neto, 1999).

7 EXPERIMENTS

We have carried out experiments to validate our approach. We describe in this section the experimental settings, and some preliminary resutls.

7.1 Settings

In our experiments, we used a single-camera gaze tracker setting based the Pupil-Centre/ Corneal-Reflection (PCCR) method to determine the gaze direction. The video camera is located below the computer screen, and monitors the subject's eyes. No attachment to the head is required, but the head still needs to be motionless. A small low power infrared light emitting diode (LED) embedded in the infrared camera and directed towards the eye. The LED generates the corneal reflection and causes the bright pupil effect, which enhances the camera's image of the pupil. The centers of both the pupil and corneal reflection are identified and located, and trigonometric calculations allow projecting the gaze point onto the image.

Most recent systems still suffer from the following problems: (1) the need for calibration for each session, (2) the large restriction on head motion – especially for single-camera systems, (3) the limitation to a single-user tracking only. The main additional difficulty in a single-camera setting is determining the distance of the user from the camera, since a triangulation as in a multi-camera setting cannot be carried out without calibration.

7.2 Preliminary Results

Given this experimental setting, we have recorded the gaze of 10 persons participating in our test session. Participants were shown images successively and asked to watch attentively the presented scenes. All gaze information have been recorded and processed according to the the description of Section 5.

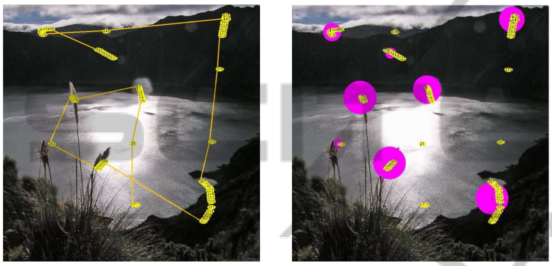

Figure 2: Example of a photograph from our database, displaying a scan path (left), and showing the processed clusters as superimposed disks, with their estimated importance denoted by the size of the disks (right).

Figure 2 shows an example of image from our database. It is a photograph showing a landscape. The upper part displays a scan path, composed with several points of regards, each being linked to the previous one to represent the path. The lower part shows the processed clustdeters – represented in the image with the superimposed disks, with their estimated importance denoted by the size of the disks. Note that after filtering out small clusters, the algorithm kept 8 locations. Figure 3 shows a similar example, with a painting in order to illustrate our approach. The painting represents the *Raft of the Medusa* painted by the French Théodore Géricault. Here, after filtering out small clusters, the algorithm kept 9 locations.

Figure 4 presents the result of our approach for another painting, and also shows how a 3D interest map can be automatically generated from gaze information. These results illustrate how our approach takes advantage of natural human gaze. We have collected gaze information from participants, clustered the points into clusters, and determined locations of interest in images with their respective importance. The next step is to integrate this automatically generated index into a retrieval system on order to demonstrate the impact of our approach, namely for the task of query-by-example retrieval.

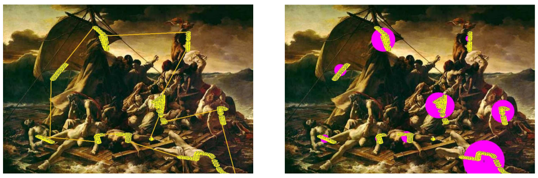

Figure 3: Another example with a painting: scan path (left) and processed clusters with their importance (right).

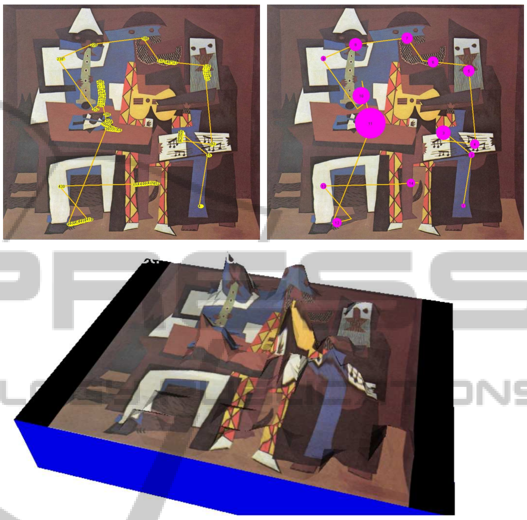

Figure 4: Another example of painting from our database, displaying a scan path (upper left), the processed clusters and their importance (upper right), and also an automatically generated 3D interest map (below).

8 DISCUSSION

For large collections of images, building RoI from users' gaze information requires the availability of gaze data from several users, for each image in the collection. We could imagine a scenario in the near future where the gaze can be estimated from a simple webcam, so that it is possible to collect users' gaze information while they are searching/browsing image collections on the Web. An alternative solution is to use saliency maps (Itti et al., 1998; Itti and Koch, 1999) to simulate this information. Saliency maps have been widely used in the computer vision field to solve the attention selection problem. The purpose of saliency maps is to represent the conspicuity (as a scalar value) of each locations of the image. According to such maps, it is possible to simulate users' gaze, and to generate RoIs together with their importance scores as described in previous sections.

Active research in computer vision is aimed at developing gaze tracking software operating from a simple webcam. From an accurate face detection, it is possible to determine the position of the eyes, and therefore to find the iris. The eye detection can be based on (among other methods) template matching, appearance classification, or feature detection. In the template matching methods, a generic eye model is first created based on the eye shape, and a template matching process is then used to search eyes in the image. The appearance based methods detect eyes based on their appearance using a classifier trained using a large amount of image patches representing the eyes of several users under different orientations and illumination conditions. The feature detection methods explore the visual characteristics of the eyes (such as edge, intensity of iris, or color distributions) to identify some distinctive features around the eyes.

9 CONCLUSIONS

One of the most difficult problems in content-based image indexing and retrieval is the automatic identification of regions of interest from images. The difficulty is related to the subjective semantics associated to the region. This difficulty is the main reason to consider semi-automatic approach as the most realistic approach to extract regions of interest from images for indexing and retrieval. More image regions are semantically labeled, better is the quality of indexing. The semi-automatic approach needs user collaboration and cooperation with algorithms to determine regions of interests. Generally, experts use graphical tools to determine these regions. This task, although popular, is time consuming when considering huge quantities of images.

Exploiting the information carried in natural human gaze is an interesting approach to determine efficiently potential semantic regions with almost no human effort. Our model is based on the use of successive fixations and saccades from people watching the media. It processes these data in order to determine clusters of points, and to extract several metrics for estimating the importance of areas in the image. The metrics are: the cardinal, the variance, the surface, the time-weighted visit count and the revisit count. All these metrics are combined together into a single estimator of the importance of image regions, in a human-centered fashion. The use of this natural information is en effective way of dealing with the high number of images in searched collections, which is a crucial issue in indexing.

Also we have concentrated our study on static images/keyframes, video indexing and retrieval can benefit from this approach, where the quantity of data is very high. Extensions of our work to the specificity of the video will need to face other challenging problems related to the temporal aspect of data.

REFERENCES

Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison-Wesley.

Itti, L. and Koch, C. (1999). Learning to detect salient objects in natural scenes using visual attention. In *In Image Understanding Workshop*.

Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259.

Jacob, R. J. and Karn, K. S. (2004). Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. In Elsevier Science, Oxford, U., editor, *The Mind's Eyes: Cognitive and Applied Aspects of Eye Movements*.

Jing, F., Li, M., jiang Zhang, H., and Zhang, B. (2002). Learning region weighting from relevance feedback in image retrieval. In *in Image Retrieval, Proc. the 27th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP*.

Martinet, J., Satoh, S., Chiaramella, Y., and Mulhem, P. (2008). Media objects for user-centered similarity matching. *Multimedia Tools and Applications, Special Issue on Multimedia Semantics*.

Nguyen, A., Chandran, V., and Sridharan, S. (2006). Gaze tracking for region of interest coding in jpeg 2000. *Signal Processing: Image Communication*, 21(5):359–377.

Osberger, W. and Maeder, A. J. (1998). Automatic identification of perceptually important regions in an image using a model of the human visual system. In *International Conference on Pattern Recognition*, Brisbane, Australia.

Poole, A., Ball, L. J., and Phillips, P. (2004). In search of salience: A response-time and eye-movement analysis of bookmark recognition. In *Conference on Human-Computer Interaction (HCI)*, pages 19–26.

Salton, G. (1971). *The SMART Retrieval System*. Prentice Hall.

Stentiford, F. (2003). An attention based similarity measure with application to content based information retrieval.

Wang, J., J.L., and Wiederhold, G. (2001). SIMPLIcity: Semantics-sensitive Integrated Matching for picture LIbraries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(9):947–963.

Wang, J. Z. and Du, Y. (2001). Rf x ipf: A weighting scheme for multimedia information retrieval. In *ICIAP*, pages 380–385.

Yarbus, A. L. (1967). *Eye Movements and Vision*. Plenum Press, New York.