ONE GAZE IS WORTH TEN THOUSAND (KEY-)WORDS

Stephanie LOPEZ[1]*, *Arnaud REVEL*[2]

Diane LINGRAND[1], *Frederic PRECIOSO*[1]

[1]Univ. Nice Sophia Antipolis
CNRS, I3S, UMR 7271,
06900 Sophia Antipolis - France

[2]University of La Rochelle
L3i Laboratory
17042 La Rochelle Cedex 1 - France

ABSTRACT

With the maturity of machine learning methods to provide satisfying Content-Based Image Retrieval systems (CBIR), research focus has recently turned back towards visual saliency analysis. The goal in these works is to extract even more efficient visual features than the existing ones. However, analyzing visual saliency is critically dependent on the task to be accomplished from the extracted visual features. A significant number of CBIR systems consider image retrieval as a binary classification problem: what is relevant for the user against what is irrelevant. In this paper, we focus on extracting relevant gaze features within the paradigm of visual preference in order to support the design of an eye-tracker CBIR system. We thus define a gaze acquisition protocol, design a benchmark from a subset of Pascal VOC database and present an in depth analysis of eye-tracking data for visual preference paradigm. Our paper provides new informations on relevant gaze features for image binary classification.

Index Terms— Eye-tracking, visual preference, gaze features, experimentation, implicit feedback.

1. INTRODUCTION

It is now rather admitted that even if image search based on key-words keeps being the most prominent technique owing to the computational efficiency of its implementation, it does not always provide us with relevant results. The gap between the description provided by the user through key-words and the expected images leading to irrelevant content is defined as the well-known "semantic gap". In order to overcome this issue, content-based image retrieval approaches have arisen exploiting directly visual features as representation of the query content. This representation is then based on detecting saliency parts in images and on describing these saliency parts. Several generic visual saliency detectors and descriptors have been proposed in the last decade, mainly exploiting gradient information such as Histograms of Oriented Gradients (HOG), Scale-Invariant Feature Transform (SIFT) and all their further extensions. Their main properties are to be robust to image variations, repeatable, both representative and discriminant, while efficient that is to say fast and easy to be computed.

If these last properties are computation oriented, recent increasing research activities on saliency detectors and saliency-based descriptors aim at designing detectors and descriptors closer to human vision mechanisms to improve the relevance of extracted visual information and thus image retrieval performances (see [1] for a study on the lack of correlations between classical features detectors and human fixations on images). Indeed, it is well known that the way we look at something expresses a lot more our interest than what few words could. We could even paraphrase Frederick R. Barnard [2], saying "One gaze is worth ten thousand (key-)words".

1.1. Goal

In this article, we are interested in exploring how visual attention cues can be extracted from raw eye-tracker data to implicitly detect a preferred image in a pair simultaneously presented to a user.

This "protocol", commonly known in psychology as "the visual preference paradigm" (introduced by Fantz in 1958 [3]) is both simple and expected to be discriminant: it consists in presenting 2 images side-by-side, and determining which one is preferred by observing the quick saccadic movements that have been proved to be inferred [4].

The goal of this article is twofold: first, we propose to the community a new open database containing raw eye-tracker data corresponding to the observation of more than 80 subjects in front of a series of pairs of images (one image in each pair only being the target) in controlled conditions; second, we describe a simple estimator, computed on-line, which predicts pretty well the selected target using only eye-tracker data.

Finally, our long term goal is to use this mechanism in an active learning system in order to make it converge rapidly (data being acquired "on the fly" it should make the relevance feedback loop shorter).

*This work is partly funded by French National Agency for Research, VISIIR project, under contract number ANR-13-CORD-0009.

1.2. Related work

Buswell [5] was the first to investigate eye movements during scene perception and revealed that fixations are clustered on informative image regions instead of being randomly distributed over the scene. Moreover, Klami *et al* [6] show that the way we look at images depends on the task.

Indeed, eye-tracking study may be free viewing or task dependent. In [7], images are displayed one by one during 4 seconds. The purpose is to determine what corresponds to a face and to a text according to the localization of fixations in an image. People do not have any specific task: it is free viewing. However, many studies are task dependent [6, 8, 9, 10, 11, 12].

By the way, the tasks involving images are multiple: retrieve a query image [8, 9] or determine if the image corresponds to a concept or class of images (classification task) [10, 11] for example.

During those tasks, the control of when the decision is taken can be explicitly done by the user. In [9], the user clicks on the image he estimates as the most similar to the target image. In [10], the explicit control is done by the eyes. The images are displayed in concentric circles. If the user wants to look at images in the circles inside, he fixated them in order to zoom them. In [11], the purpose is to identify if there is one or no image corresponding to the concept of Sport. Images are displayed four by four. When the choice is made, the user presses one of the 1/0 buttons to indicate if there is a sport image or not. In [12], images are displayed one by one and the user presses the YES/NO button if the image belongs to the target category or not. However, the control may also be implicit as for example in [8]: the user does not know the criterion that determines the next step. New images are displayed when the sum of duration of fixations on the images is above a threshold.

The decision of the user may be explicit or not to the algorithm. Clicking on an image [9] or selecting one image with a YES/NO button [12] lead to explicit decision of the user. On the opposite side, in [11], the user does not tell which image corresponds to the target, only that there is one among four. As for [8] and [10], the decision is inferred from eye-tracking measurements based on eye positions, fixations and saccades.

In our study, we focus on the classification of images task with implicit decision and explicit control in the particular case of visual preference paradigm. We want to extract the gaze data that enables the decision. Since previous works have used eye tracking data to improve CBIR retrieval systems as input added to image features, we will build a method for studying eye-tracking data regardless of image data.

2. METHOD

Our experimental method aims at determining which gaze features are relevant enough to select in real-time the image of interest for the user among a pair of images.

As aforementioned, our approach is based on the "visual preference paradigm" which is furthermore fully compliant with classic annotation strategies for binary-classification CBIR systems: the user has to retrieve a query concept in a serie of pairs of images, sequentially presented, but for each pair, only one of the two images contains the concept to be recognized. Such a strategy allows us to easily discriminate gaze features corresponding either to user interest or user disinterest.

Two experiments were performed. The first one in Nice and the second one in La Rochelle, with different materials and participants.

2.1. Participants

For the first experiment, there were 46 volunteers (16 females and 30 males) participating. Six were removed: two children, one who did not respect the instructions, one who had problems with glasses and two who had already done the test during the experiment setting development phase. The data reported below are based on the remaining 40 participants.

In the second experiment, there were 48 volunteers (24 women, 24 men). Two were removed: one for a too high eye misdetection rate and the other one because of autism which lead to another strategy of vision (see [13] for a review on peculiarities of perception in autism).

2.2. Material

For the first experiment, we have used an eye-tracker TOBII T120. For this particular equipment, two modes are available: 120 Hz and 60 Hz eye-tracking rates. We conducted the experiment at 60 Hz which corresponds to data recorded every 16ms approximately. The second experiment has been processed with an eye-tracker TOBII X2-30 (only 30 Hz eye-tracking acquisition rate), which provides data about every 32ms.

2.3. Dataset

The data set is composed of 156 images selected from the Pascal VOC 2007 dataset [14] according to four general categories (Animal, Vehicle, Person, Furniture).

Images are rescaled with dimensions calculated relatively to the application window dimensions. Image width is at most set to 27% of the window width (this amount has been experimentally validated). Then, image heights are equalized with respect to the highest image of the pair. All these transformations are done while preserving image ratio.

2.4. Protocol

The experiment begins with oral explanations about the eye-tracking equipment and how the experiment will be processed

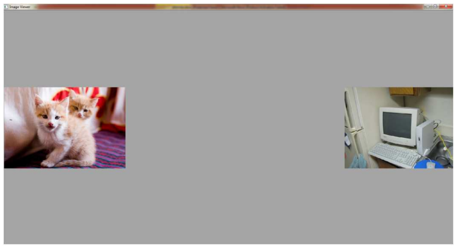

Fig. 1. A pair of images is displayed on the screen

(following recommendations from [15]). First, we have to check that the participant's eyes are well detected. Then, the participant has to operate calibration, in order to be sure that the device gets the true positions of the participant's eyes during the experiment. Calibration consists in fixating five points on the screen ([16]) and is done again if errors are too important.

The experiment consists in four retrieval tasks randomly ordered. Each task corresponds to finding one of the four categories (Animal, Vehicle, Person, Furniture): 20 pairs of images are then displayed respecting the visual preference paradigm (i.e. one and only one image corresponds to the target category).

The participants are asked to perform the retrieval task as quickly as possible. For each of the four task, the following three steps are performed:

1. instructions are displayed (until the user presses the space bar)
2. the image pair is displayed at most 5 seconds (if the user founds the target image, he presses the space bar and goes immediately to step 3)
3. gray screen with a cross on the median vertical axis is displayed during 2 seconds.

Steps 2 and 3 are repeated 20 times.

The pairs of images and their positions (left or right) are randomly selected. The images are displayed at both extreme sides of the screen (see figure 1) in order to avoid any participant using the peripheral vision instead of moving the eyes to find the target.

3. RESULTS

3.1. Database

We have collected raw data about the position of the eyes and pupilla's size for each eye [16]. To the best of our knowledge, this is the first database of measurements from gaze data for a visual preference protocol on various categories (not only for faces as in several previous works).

In order not to perform a user specific analysis, but to determine features that would work for any user, all the data have been aggregated in a single data file available at `http://visiir.univ-lr.fr/VISIIR-data`.

3.2. Gaze features

First, we want to check if the gaze features provided in related articles [11, 10] are appropriate for the visual preference paradigm and which among these gaze features would hold more information on the visual content. We have reported in table 1 all the gaze features we have extracted from raw data. Data are computed only when both eyes are well-detected and values are given for left eye as an arbitrary choice since no relevant difference has been observed between the two eyes.

Table 1. Gaze features calculated from raw data

1	line number
2,3	max. size of pupilla on left, right image
4	total fixation number (F)
5	F/(F + number of saccades)
6,7	left image: spread in x, y
8,9	right image: spread in x, y
10,11	both images: spread in x, y
12	average distance between fixations
13,14	left image: spread in x, y for fixations
15,16	right image: spread in x, y for fixations
17,18	both images: spread in x, y for fixations
19,20	first and last seen image
21	image label with maximal pupilla size
22,23	first and last fixated image
24,25	duration of first and last fixation
26,27	number of fixations during 1st and last visit
28	total fixation duration
29,30	number of fixations on left (right) image

x values are normalised between 0 and 1. Thus, a value of x below 0.5 (resp. above) corresponds to the left (resp. right) image. Fixations and saccades are computed according to the same parameters as GaZIR [10]: a fixation is detected if the points have a dispersion less than 30 pixels, (0.6 visual degrees for our monitor of 17" screen with resolution 1280*1024) in a period of at least 120 ms.

3.3. Features to predict the preference

We look for the simplest estimator of the image implicitly chosen by the user, that is to say the gaze feature the most compliant with real-time decision by analysing subsets of these different gaze features. For that purpose, we use a CART decision tree (R implementation), which is both a predictive and explicative model due to its capacity to generate a set of explicit rules based on parameters analysis to classify data.

In table 2, we present the percentage of good classification of different CART decision trees taking into consideration several features subsets for Nice and La Rochelle. In the first subset α, the CART tree has been trained with all the features listed in table 1. The decision is taken according to

Table 2. Main features to get good decisions grouped in four different sets.

subsets	α	β	γ	δ
Nice	95.94%	95.51%	94.71%	94.91%
La Rochelle	95.26%	95.06%	95.52%	95.29

the label of the target image. The most discriminant feature is ”last seen image”(table 1:20)(similar results can be found in the subset β with the last fixated image (table 1:23) as most discriminant feature).

However, this information can only be available *a posteriori* (the last image can only be determined after the end of the sequence), while in our use case, we are interested in computing predictive features *a priori*. Another CART tree has then been trained removing these inadequate features (results reported in table 2 subset γ). The new main feature is then the spread in x (table 1:6,8,10) ($\Delta(x) = \max(x) - \min(x)$). This feature can be computed at each time step from only a simple raw data, the x value. In our context, it is a good candidate. Deriving this feature, we then proposed to substitute the average position to the ”last seen” and ”last fixated” features (table 1:20 and 23) and report the new results in table 2 in subset δ, which remains pretty satisfying. In the subset δ, it can be seen that results remain very good. Moreover, analyzing the CART tree, the average position is the most discriminant feature. If considering only this average position instead of all those features from table 1, the results are around 83% for La Rochelle and 87% for Nice, which leads to the conclusion that this feature is a good candidate. However, the prediction accuracy of the gaze average is very dependent on the observation time. Thus, it may be not the best candidate for real-time decision. We thus proposed to use the cumulative average ($\overline{x} = \frac{\sum_{1,N} x}{N}$). In figure 2, two typical examples of the cumulative average for two gaze recordings are presented. Analyzing these curves, we noticed that the first 480 ms concern the retinal persistence of the cross (left picture). Then, around 640 ms, we presume the accumulated position of the gaze indicates what is the preferred image (an intermediate saccade can be observed if the irrelevant image has been selected first - see right picture).

Our goal is to capture “on the fly” the interest of the user, to improve results of mouse click and to get a feature relevant enough to get a good decision with minimum thought and manual involvement [8]. We thus proposed to use only $\overline{x}$ as a predictor of the target image. The profiles observed, such as in figure 2, lead us to consider two values of the cumulative average of x: at t_0, noted $\overline{x}_{t_0}$, and at $t_1 > t_0$, noted $\overline{x}_{t_1}$. For the sake of clarity, only few results of quality of decisions with respect to t_0 and t_1 are presented in table 3.

It is interesting to notice that using a single t value for the decision leads to lower the performance (approximately 10% below results in table 3).

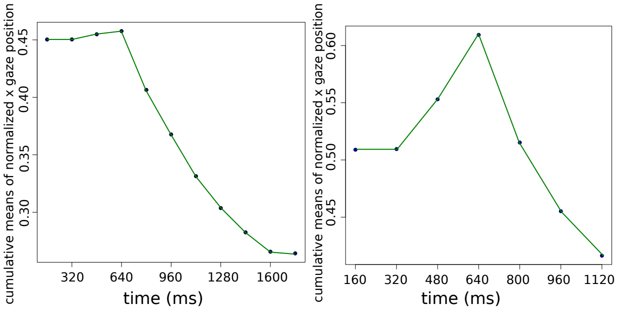

Fig. 2. Cumulative means of gaze position in x for different image pairs. We observe that depending on the profile, different decision strategies could be decided.

Table 3. Good decisions ratio from $\overline{x}$ at t_0 and t_1

t_0	t_1	**Nice**	**La Rochelle**
800 ms	960 ms	71.2%	68.2%
736 ms	960 ms	70.8%	67.7%
768 ms	928 ms	70.2%	67.7%

Increasing t_1 value increases the performance but, in order to satisfy “on-the-fly” decision, we limited t_1 to 960 ms. The optimal difference between t_0 and t_1 is 160ms. All values of table 3 are about 70% showing that lowering t_1 from 960ms to 928ms does not affect too much the perfomances. Based on these promising results for a real-time decision with only average position, we investigate how, considering all features from subset γ in table 2 at t_1=960ms could improve the accuracy. We then reach 88% of good classification which can already be considered as a convincing relevance feedback mechanism in active learning CBIR system.

4. CONCLUSION AND PERSPECTIVES

In this paper, we proposed a protocol for gaze data acquisition in the context of visual preference paradigm. The data acquired from 86 subjects are made available to the research community. We analyzed various gaze features, inspired from previous studies. Our purpose was to determine which ones would be relevant for implicit decision “on the fly” in the visual preference paradigm.

Our preliminary results are very promising and show that the gaze average position, which has never been considered in previous works, is indeed a convincing user feedback in the visual paradigm context. Our on-going deep analysis of gaze features already provided some clues on possible improvements combining several features while remaining compliant with “on the fly” decision. The next step will consist in using these results as annotation process in a CBIR system.

5. REFERENCES

[1] A. Dave, R. Dubey, and B. Ghanem, “Do humans fixate on interest points?,” in *IEEE ICPR*. 2012, pp. 2784–2787.

[2] S. Burton, *The Home Book of Proverbs, Maxims, and Familiar Phrases*, Mac-Millan Publishing Company, April 1987.

[3] R. Fantz, “Pattern vision in young infants,” *The Psychological Record*, vol. 8, pp. 43–47, 1958.

[4] H. Kirchner and S. J. Thorpe, “Ultra-rapid object detection with saccadic eye movements: Visual processing speed revisited,” *Vision Research*, vol. 46, pp. 1762–1776, 2006.

[5] G.T. Buswell, *How people look at pictures: A study of the psychology of perception in art*, University of Chicago Press, 1935.

[6] A. Klami, “Inferring task-relevant image regions from gaze data,” in *Machine Learning for Signal Processing (MLSP), 2010 IEEE International Workshop*. Aug-Sept 2010, pp. 101–106.

[7] S. Karthikeyan, V. Jagadeesh, R. Shenoy, M. Exkstein, and B.S. Manjunath, “From where and how to what we see,” in *IEEE Int. Conf. on Computer Vision (ICCV)*, Sydney, NSW, December 2013, pp. 625 – 632

[8] O. Oyekoya and F. Stentiford, “Perceptual image retrieval using eye movements,” in *Proceedings of the International Workshop on Intelligence Computing in Pattern Analysis/Synthesis*. In N. Zheng, X. Jiang, and X. Lan, editors, 2006, pp. 281–289.

[9] K. Essig, *Visual-Based Image Retrieval (Vbir) - A New Eye-Tracking Based Approach to Efficient and Intuitive Image Retrieval*, Ph.D. thesis, Bielefeld University, 2007.

[10] L. Kozma, A. Klami, and S. Kaski, “Gazir: Gaze-based zooming interface for image retrieval,” in *Proceedings of the 2009 International Conference on Multimodal Interfaces (ICMI-MLMI)*, 2009, pp. 305–312.

[11] A. Klami, C. Saunders, T.E. de Campos, and S. Kaski, “Can relevance of images be inferred from eye movements ?,” in *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval (MIR’08)*, Vancouver, British Columbia, Canada, October 2008, pp. 134–140.

[12] D. Papadopoulos, A. D. F. Clarke, F. Keller, and V. Ferrari, “Training object class detectors from eye tracking data,” in *European Conference on Computer Vision (ECCV 2014)*, Zurich, Switzerland, 2014, pp. 361–376, Springer.

[13] S. Dakin and U. Frith, “Vagaries of visual perception in autism,” *Neuron*, vol. 48, pp. 497–507, November 2005.

[14] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results,” http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[15] K. Pernice and J. Nielsen, “How to conduct eyetracking studies,” http://www.nngroup.com/reports/how-to-conduct-eyetracking-studies/.

[16] Tobii Technology, “Tobii eye tracking, an introduction to eye tracking and tobii eye tackers,” Tech. Rep., Tobii, 2010.