Catching relevance in one glimpse: food or not food?

S. Lopez[1]*, A. Revel[2], D. Lingrand[1], F. Precioso[1], V. Dusaucy[3], A. Giboin[4]
[1] Univ. Nice Sophia Antipolis (UNS) - I3S UMR7271 - UNS CNRS, 06900 Sophia Antipolis - France
[2] Univ. La Rochelle - L3i, 17000 La Rochelle - France
[3] LPC Marseille - UMR 7290, 13003 Marseille - France
[4] INRIA - I3S UMR7271 - UNS CNRS, 06900 Sophia Antipolis - France
{slopez,lingrand,precioso}@i3s.unice.fr, arnaud.revel@univ-lr.fr, vdy@squad.fr, giboin@inria.fr

Keywords

Gaze features; Mental search; Visual Preference Paradigm; Implicit Feedback; Experimentation; Image tag

1. INTRODUCTION

Retrieving specific categories of images among billions of images usually requires an annotation step. Unfortunately, keywords-based techniques suffer from the semantic gap existing between a semantic concept and its digital representation. Content Based Image Retrieval (CBIR) systems tackle this issue simply considering semantic proximities can be mapped to similarities in the image space. Introducing relevance feedbacks involves the user in the task, but extends the annotation step.

To reduce the annotation time, we want to prove that implicit relevance feedback can replace an explicit one. In this study, we will evaluate the robustness of an implicit relevance feedback system only based on eye-tracking features (gaze-based interest estimator, GBIE). In [L$^+$15], we showed that our GBIE was representative for any set of users using "neutral images". Here, we want to prove that it remains valid for more "subjective categories" such as food recipe.

2. METHOD

2.1 Related Work

Image retrieval is a challenging field of research with the increasing amount of available image on the Internet. CBIR systems need a preliminary annotation step, which is often time-consuming and generally require a large amount of images (5011 images are used in PASCAL VOC 2007 database to train the classification system). Exploiting gaze information to ease the processing load in CBIR context has recently aroused research interest. Indeed,[Bus35] has already proved that the gaze contains valuable information. Nowadays, it is well known that fixations depend on the task. It is generally the feature that appears to be the most relevant to predict the user's choice. It is the case of [K$^+$09], with image query task, and [K$^+$08], which is an abstract concept retrieval. In both cases, the gaze is used as an implicit feedback. For [O$^+$06], it is also an implicit control: the display of images is controlled by the gaze, even if the user does not know the target category. For [K$^+$08], the user indicates through an explicit control (the keyboard) if there is one or no image corresponding to the abstract concept among the 4 images displayed. In the Visual Preference Paradigm (VPP) [Fan58] 2 images are displayed to retrieve the one that interests the user the most, reducing even more user attention load.

2.2 Goal

In the presented work, users have an explicit control on the system: they press the space bar whenever they have made a choice within 5 seconds. Meanwhile, the gaze is monitored to be analysed afterwards. To what extend can we reduce the viewing time and get an accurate automatic annotation implicitly with a gaze feature only? To select the most discriminant feature, we perform and analyze a decision tree. Since we target to design a GBIE applicable "on the fly", considering simply the fixations may be not accurate enough in our case. Our GBIE is thus compared with the estimators of [K$^+$08] and [O$^+$06], in term of accuracy of classification and reactivity.

2.3 Procedure

Gaze data are monitored with an eye-tracker TOBII T120, 17", at 60Hz: coordinates of each eye and their validity and pupil diameter. The gaze features are rather common to all similar works and more details can be found in [L$^+$15]. In order to prevent the user from developing a particular strategy to recognize targets, the task consists in identifying images according to 4 general given categories. In *Standard image dataset* experiment, images from PASCAL VOC 2007 database are grouped as follow: *animals*, *persons*, *vehicles* and *furnitures*. In *Food image dataset* experiment, images from [W$^+$15] are organized as follow: *appetizers*, *desserts*, *containing citrus* and *containing red berries*. Each category contains 40 images presented in pairs according to VPP so that only one image per pair belongs to the target category. The users are asked to perform the retrieval task as quickly as possible to prevent gaze distraction. Between each pair of images, a red cross on a gray background is displayed randomly on the median vertical axis.

3. RESULTS

3.1 Gaze-Based Interest Estimator (GBIE)

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

AVI '16 June 7–10, 2016, Barri, Italy

DOI: 10.475/123_4

In order to determine which features among table1 in [L+15] are the most discriminant for predicting the users' selection of the target image, we performed a tree classification. Our hypothesis is that the root of the tree corresponds to the most discriminant feature. We want this feature to be easy to compute and available "on the fly". To do so, we performed experiments on three groups of features.

The discriminant gaze feature among all the possible features, but removing the average gaze position in x (i.e. horizontal position) and adding our GBIE, (this feature set is called α) depends on the viewing time. Thus, we defined a subset of features (called β), removing all the features based on the last viewed image in the displayed pair (this subset does not contain our GBIE). Then, γ group is similar to β adding our GBIE.

Table 1: Main features for classification until decision for *Standard* (S) and *Food* (F) image datasets

Case	Root feature	rate S	rate F
α	last seen image	96.5%	92.8%
β	right image: spread of gaze in x	93.8%	88.8%
γ	average of gaze position in x	94.1%	92.2%

The root features defined by the decision tree creation are exactly the same in both experiments, which means it is generalizable.

3.2 Detailed analysis of the GBIE

As the average duration of visualization of image pairs is 1840 ms for *Standard image dataset*, we aim to reduce by two the time of prediction. Thus, we study the average position of the gaze in x (i.e. horizontal displacement) before one second. We can correlate the slope with the target image before the time limit (one second). As an approximation of the slope, two values of the cumulative average of x are considered: at T_0 and at $T_1 > T_0$ (table 2).

The later T_1, the better the results. Considering decision time, we selected T_0=800 ms and T_1=960 ms as a good compromise. Indeed, waiting until T_1=992 ms does not affect the performances too much.

Table 2: Cumulative average at T_0 and T_1

T_0	640		704		800	928
T_1	800	960	864	1024	960	1248
rate S %	60.1	66.1	63.3	68.4	67.8	69.4
rate F %	50.9	54.4	53.4	54.5	54.4	61.5

3.3 Comparison to other works

We compare our results to the GBIE defined in [O+06] and [K+08]. In [O+06], the users have to retrieve a query image among different sets of 4 by 4 set of images. They do not know the criterion that displays the next set. This criterion consists in a threshold on the sum of the fixation length on one image (either 400 or 800ms). A fixation lasts at least 80ms in this study. The hypothesis is that the image is more likely to be relevant if the user is looking at it for a long time. We apply this method to both experiment, on the data collected until the decision time (see table 3).

In many cases, the viewing time does not allow to reach the threshold of fixation duration on one image (400 ms or 800 ms), whereas our GBIE allows to get labels for all the images. This method is not well suited to an "on the fly" study. Moreover, this method does not provide good accuracy. It means that the user takes a decision before this threshold. Our GBIE gets a better accuracy in a shorter time than the maximal viewing time.

Table 3: Results with method [O+06]

	Standard		Food	
Threshold (ms)	Accuracy (%)	unpredicted labels (%)	Accuracy (%)	unpredicted labels (%)
400	46.1	13	47.2	11.9
800	35.8	40.4	37.4	34.7

In [K+08], 4 images are displayed. The user has to indicate with the keyboard if there is one or no image corresponding to the abstract concept. The combined gaze features that are used in Linear Discriminant Analysis (LDA with leave-one-out) are: total length of fixations, number of fixations, average length of fixations, number of transitions from an image to another, number of images with at least one fixation, number of fixations within an image.

We apply this method to our 2 image databases until decision while limiting time analysis to 960 ms (see table 4).

Table 4: Results with method [K+08]

experiment	Standard	Standard960	Food	Food960
accuracy (%)	64.8	55.1	61.1	49.5

For the data monitored until the user's decision, the accuracy of the prediction is comparable. Nevertheless, our GBIE provides slightly better accuracy in a shorter time (less than one second, whereas the average viewing time for *Standard image dataset* and *Food image dataset* are respectively: 1840ms and 2208ms). Thus, our GBIE appears to be better suited to an "on the fly" study.

4. CONCLUSION

Gaze tagging images in less than one second appears to be intuitive. Our GBIE provides good classification accuracy with no extra actions required, given that no image features were used. It is also user and category independent. It is not based on traditional features like fixations, which appear to be not adapted to an "on the fly" context. Completing our GBIE with explicit feedback would provide more confidence in the labeling but would be time consuming and gaze distracting.

5. REFERENCES

[Bus35] Guy T. Buswell. How people look at pictures: A study of the psychology of perception in art. *The Art Bulletin*, 18(3):198, 1935.

[Fan58] R. Fantz. Pattern vision in young infants. *The Psychological Record*, 8:43–47, 1958.

[K+08] A. Klami et al. Can relevance of images be inferred from eye movements ? In *ACM MIR*, 2008.

[K+09] L. Kozma et al. Gazir: Gaze-based zooming interface for image retrieval. ICMI-MLMI, 2009.

[L+15] S. Lopez et al. One gaze is worth ten thousand (key-)words. In *IEEE ICIP*, 2015.

[O+06] O. Oyekoya et al. Perceptual image retrieval using eye movements. In *Int. Workshop on Intelligence Comput. in Pattern Analysis/Synthesis*, 2006.

[W+15] X. Wang et al. Recipe recognition with large multimodal food dataset. In *IEEE ICME*, 2015.