A Multimodal Information Collector for Content-Based Image Retrieval System

He Zhang, Mats Sjöberg, Jorma Laaksonen, and Erkki Oja

Department of Information and Computer Science,
Aalto University School of Science, Espoo, Finland
{he.zhang,mats.sjoberg,erkki.oja,jorma.laaksonen}@aalto.fi

Abstract. Explicit relevance feedback requires the user to explicitly refine the search queries for content-based image retrieval. This may become laborious or even impossible due to the ever-increasing volume of digital databases. We present a multimodal information collector that can unobtrusively record and asynchronously transmit the user's *implicit* relevance feedback on a displayed image to the remote CBIR server for assisting in retrieving relevant images. The modalities of user interaction include eye movements, pointer tracks and clicks, keyboard strokes, and audio including speech. The client-side information collector has been implemented as a browser extension using the JavaScript programming language and has been integrated with an existing CBIR server. We verify its functionality by evaluating the performance of the gaze-enhanced CBIR system in on-line image tagging tasks.

Keywords: Implicit relevance feedback, JavaScript, gaze tracking, content-based image retrieval, image tagging.

1 Introduction

Relevance feedback has been widely utilized in content-based image retrieval (see [1] for an extensive survey). Often, people need to attentively indicate or answer whether or not the retrieved information is relevant, and thus give *explicit relevance feedback*. With large databases and long retrieval sessions, this will inevitably become a laborious task. The interest for using *implicit relevance feedback* [2], although less accurate than explicit, has increased in recent years. By using implicit relevance feedback, an information retrieval system can unobtrusively record the user's behavior, such as gaze direction, facial expressions, body gestures etc., and use this information to infer the user's search preferences [3]. Moreover, a combination of explicit and implicit feedback can even better model the user's potential interests [4].

In the current work, we present a multimodal information collector that can unobtrusively record and transmit the user's implicit feedback, in addition to the explicit feedback, to the remote CBIR server for image retrieval. The feedback modalities include eye movements, pointer tracks and clicks, keyboard strokes, and audio including but not limited to speech. We focus on using gaze as a primary feedback modality since eye movements have earlier been found to have

B.-L. Lu, L. Zhang, and J. Kwok (Eds.): ICONIP 2011, Part III, LNCS 7064, pp. 737–746, 2011.

a strong correlation with human cognitive processes (see [5] for a thorough review). Using eye movements as an implicit feedback source is a relatively new research area. However, eye movements have already demonstrated strong potentials in inferring people's interests in tasks such as image retrieval [6] and image ranking [7].

A similar information collector can be found in [8], where a prototype attentive information system was implemented to track the user's behavior and suggest helpful information to the user. However their system was not evaluated quantitatively. In [9], a Web-Accessible Multimodal Interfaces (WAMI) toolkit was developed at MIT. Their approach was tightly connected to the use of speech input and audio output in web applications and had no directions to eye movement analysis.

The following section describes the implementation and operation principles of the multimodal information collector. Section 3 introduces the gaze-enhanced CBIR system. In Section 4, we verify the functionality of the information collector by evaluating the system performance with real user experiments. Section 5 concludes the paper and discusses our future work.

2 Implementation and Operation Principles

Figure 1 illustrates the overall schematic diagram of a client-server based image retrieval system with four forms of user interaction modalities collected at the client side. The multimodal information collector has been implemented as an extension of Mozilla Firefox, which is a free and open source web browser of great popularity today.

2.1 Client Implementation

The client-side collector is programmed by using the JavaScript language since it is the primary implementation language of Mozilla Firefox extensions and supports prototype-based object construction and object-oriented programming including class inheritance.

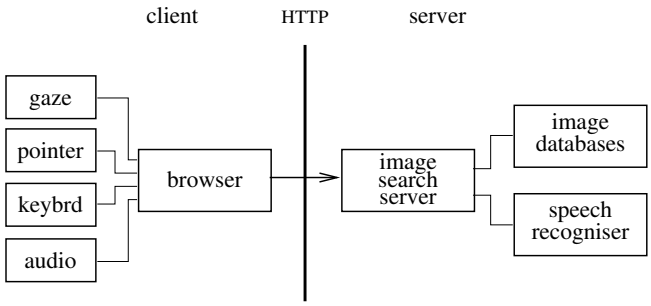

Fig. 1. The block diagram of our system capable of transferring multimodal feedback from a browser client to a content-based image search server

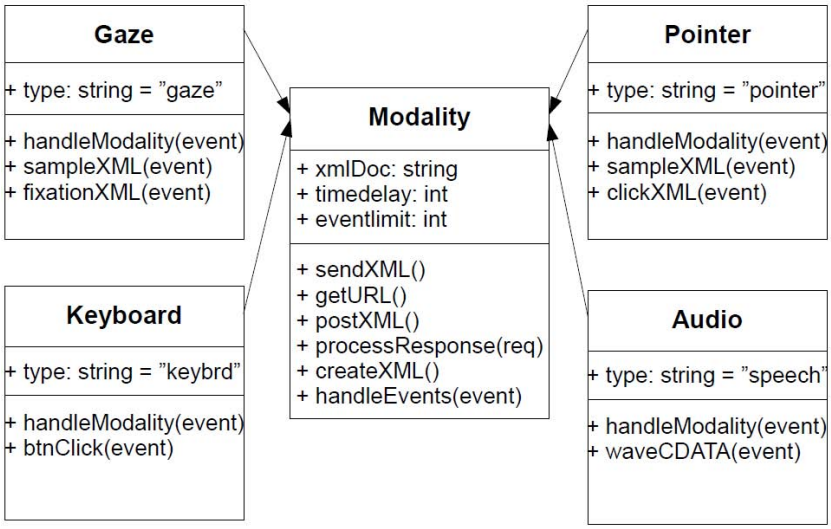

Fig. 2. A class diagram showing the relationships between the **Modality** super-class and its four sub-classes, **Gaze**, **Pointer**, **Keyboard**, and **Audio**

Figure 2 shows a class inheritance diagram of the JavaScript implementation. The roles of the classes are defined as follows:

> **Modality** is defined to be the super-class of the classes for modalities of user interaction. These all collect the input and store them internally in the `xmlDoc` object which is then periodically transmitted to the server in XML-formatted packets.
> **Gaze** is defined to be a sub-class of Modality. This module detects the user's gaze samples and fixations on a displayed image and converts them into 2D coordinates relative to the shown image.
> **Pointer** is defined to be a sub-class of Modality. This module detects the user's pointer movements (and clicks) and converts them into 2D coordinates.
> **Audio** is defined to be a sub-class of Modality. This module detects and converts the user's voice into binary audio data for the speech recognizer.
> **Keyboard** is defined to be a sub-class of Modality. This module detects and records the user's keyboard events (strokes).

Each sub-class inherits all the functions defined in **Modality**, but adds its own methods for that particular functionality. The method `handleModality()` is overridden in each sub-class for a specific modality. It generates the XML data structure for that particular interaction modality. For example, in the class of **Gaze**, the method `handleModality()` calls the `sampleXML()` method to generate XML data for gaze samples, and `fixationXML()` for gaze fixations. Then the general method `handleEvents()` in the super-class periodically sends the collected data to the specified server URL. Similar routines apply to the other three sub-classes.

2.2 Client-Server Interaction

The client-server communications are based on the World Wide Web Consortium (W3C) XMLHttpRequest protocol[1], which has recently been employed extensively for implementing *Asynchronous JavaScript and XML* (AJAX)[2] type of asynchronous content updates in web applications.

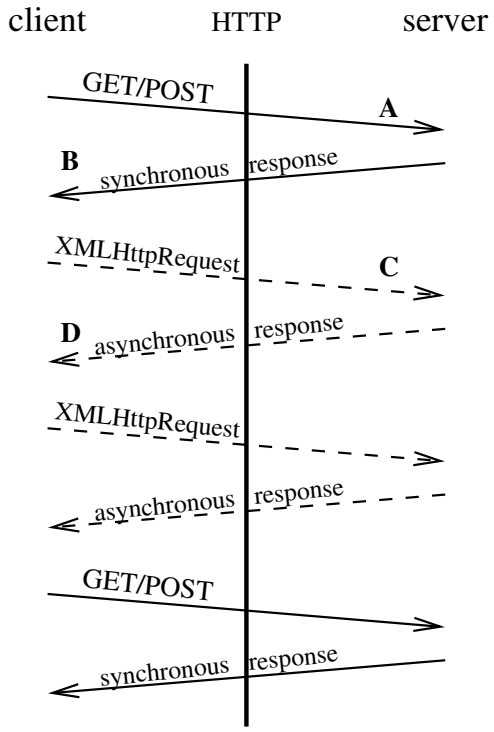

Fig. 3. The message exchange diagram of synchronous and asynchronous communications between the image search client and the content-based image retrieval server

Figure 3 shows the messaging diagram of synchronous and asynchronous communications between an image database server and a browser client. The image retrieval session is initiated by the client requesting the server to present some visual content for inspection. The server returns with a conventional synchronous HTTP action a set of images to the browser, which then presents them to the user. Together with the XML/HTML page containing the images, the server can also specify an URL from its own URL space where the client can send asynchronous user interaction data. To our knowledge, this is a novel idea not used in any existing content-based image retrieval systems.

[1] `http://www.w3.org/TR/XMLHttpRequest/`
[2] `http://www.adaptivepath.com/ideas/essays/archives/000385.php`

The message types transferred between the server and the client have been identified as **A**, **B**, **C** and **D** in Figure 3. Their exact roles in the communication of multimodal relevance feedback are omitted here due to the space limit.

3 Gaze-Enhanced Content-Based Image Retrieval

3.1 PicSOM CBIR System

We have integrated the multimodal information collector with an existing CBIR server named PicSOM[3] [10], which is a content-based image retrieval system developed since 1998, first at the Helsinki University of Technology and then at the Aalto University. PicSOM uses the principles of *query by example* and *relevance feedback* in implementing iterative and interactive image retrieval.

The unique approach used in PicSOM is to have several Self-Organizing Maps (SOMs) [11] in parallel to index and determine the similarity of images. These parallel SOMs have been trained with separate data sets obtained by using different feature extraction algorithms on the same objects. The extracted image features [12] include RGB histogram, DCT coefficients, edge statistics etc.

As the SOM maps visually similar images near to each other, this motivates to spread the relevance feedback given for the viewed images to their neighboring images on the map surface. Images marked as relevant are first given positive and those marked as non-relevant are given negative values on the map surface. These relevance values are then smoothed and spread around with low-pass filtering. Images with the largest resulting relevance scores are then shown to the user.

3.2 Using Gaze Patterns as Implicit Relevance Feedback

Based on the received gaze coordinates at the server, we calculate for each viewed image a 19-dimensional feature vector as specified in Table 1. These features have been used in image retrieval and ranking tasks before [6,7]. The relevance predictions for the viewed images are obtained with a simple logistic regression model created with separate training data.

In the PicSOM system, the gaze-based implicit relevance estimates are combined with the click-based explicit relevance feedback values. In this process the gaze-based regressor outputs are always in the range of $[0, 1]$ and the larger the value, the more probably the image is relevant. These values are then summed with the $+1$ and -1 values given for the clicked and non-clicked images, respectively. The combined relevance values are finally placed in the SOM units and spread to their neighbors with low-pass filtering similarly to PicSOM's normal operation.

3.3 Automatic Speech Recognition

The speech recognition system used in the experiments has been developed by the speech group in the Department of Information and Computer Science at

[3] `http://www.cis.hut.fi/picsom`

Table 1. Eye movement features collected1 at the client side

Number	Name	Description
1	numMeasurements	log of total time of viewing the image
2	numOutsideFix	total time for measurements outside fixations
3	ratioInsideOutside	percentage of measurements inside/outside fixations
4	speed	average distance between two consecutive measurements
5	coverage	number of subimages covered by measurements[1]
6	normCoverage	coverage normalized by numMeasurements
7	pupil	maximal pupil diameter during viewing
8	nJumps1	number of breaks[2] longer than 60ms
9	nJumps2	number of breaks[2] longer than 600ms
10	numFix	total number of fixations
11	meanFixLen	mean length of fixations
12	totalFixLen	total length of fixations
13	fixPrct	percentage of time spent in fixations
14	nJumpsFix	number of re-visits (regressions) to the image
15	maxAngle	maximal angle between two consecutive saccades[3]
16	firstFixLen	length of the first fixation
17	firstFixNum	number of fixations during the first visit
18	distPrev	distance to the fixation before the first visit
19	durPrev	duration of the fixation before the first visit

[1]The image was divided into a regular grid of 4×4 subimages, and covering a subimage means that at least one measurement falls within it. [2]A sequence of measurements outside the image occurring between two consecutive measurements within the image. [3]A transition from one fixation to another.

Aalto University. The speech signal is sampled using 16 kHz sampling rate and 16 bits. The signal is then represented with 12 MFCC (mel-frequency cepstral coefficients) and the log-energy along with their first and second differentials. Above features are calculated in 16 ms windows with 8 ms overlap. Cepstral mean subtraction (CMS) and a maximum likelihood linear transformation, which is estimated in training, are applied to the features. For the acoustic model, we use state-clustered Hidden Markov triphone models that have 5062 states modeled with 32 Gaussians.

4 Experiments

In this section, we verify the functionality of the proposed multimodal information collector, and evaluate the gaze-enhanced CBIR system in an image tagging scenario. Automatic tagging is a useful but still not fully reliable means for associating keyword-type information to unannotated images. In the current state of the art, human effort is still needed for checking and correcting the tags [13]. The tag correction process can be seen as a special case of content-based image retrieval where the goal is to quickly correct the erroneously-tagged images.

4.1 CBIR-Assisted Image Tag Correction

Let us consider a CBIR setup where the viewed images are such that an automatic image annotation system has assigned all of them some particular tag or keyword based on their visual properties. The considered images are thus visually quite similar to each other, but due to imperfections in the assignment, there are bound to be semantic differences or tagging errors among them. The burden of a user who needs to check and correct the automatically assigned tags would be eased if the wrongly-tagged images could be found as early as possible.

This can be understood as a complementary setting for the conventional interactive CBIR setting. Now the relevant images are not those that resemble the target image, but those that are semantically different from the other, correctly-tagged ones. Nevertheless, CBIR techniques can be used to speed up retrieving of such images. This time, the search will be driven more by the negative relevance feedback, given to the correctly-tagged images. The system will then retrieve more and more images that are different from the typical correctly-tagged images and are thus more likely to be the wrongly-tagged ones.

4.2 Setups and Evaluations

Data. We used the *train* subset of the PASCAL Visual Object Classes Challenge 2007 (VOC2007) data set [14] with a total of 2501 annotated images that cover 20 overlapping categories. To ease the burden of users, we randomly selected 16 categories and divided them into two groups:

1. correctly-tagged: *car, dog, bicycle, person, motorbike, train*
2. wrongly-tagged: *sheep, horse, aeroplane, boat, bus, bottle, dining table, potted plant, sofa, tv-monitor*

Experiment Setup. We recruited 18 test subjects both males and females from several departments at the Aalto University. The mean age of the test subjects was 27.2 years old, ranging from 23 to 34 with good balances in between. Very few of the users had experiences in image tagging and only one user had experiences in gaze tracking.

Each subject was asked to perform six tagging tasks. For each task, the user had to check and correct the tags of one particular category from group 1. Before each task, the system randomly selected 40 images of that category and another 40 images of the ten categories from the wrongly-tagged group. Thus half of the images were always tagged correctly. During each task, the system showed a total of 40 images, contained in five image pages each having eight images. After each task, the user was asked of his or her subjective opinions whether the corresponding variant facilitated the tagging task, and whether it was reliable and fast enough.

Feedback Modalities. The following relevance feedback modality types or variants of the system were compared:

1. *Baseline*: The user corrects the image tag by selecting the corresponding category name from the drop-down menu under the image. No CBIR or speech recognition techniques are used.

2. *Explicit*: The user clicks the pointer over the wrongly-tagged image and speaks the desired category name into the microphone. Only explicit relevance feedback from pointer clicks are used.
3. *Implicit*: The tag correction is similar as in *explicit*. However, the user's eye movements are unobtrusively recorded by a Tobii eye tracker[4]. Both explicit pointer relevance and implicit gaze relevance feedback are used.

For the baseline variant all the 40 images presented to the user were randomly chosen, whereas for the other two variants only the eight images in the first page were random while the images in the remaining four pages were selected by the relevance feedback information.

The Evaluation and Results of Image Retrieval. The measure of performance is the number of images that the user corrects in one tagging task, which gives reflection on how well the system retrieves wrongly-tagged images. Table 2a gives the quantitative performance of the three variants for each user. Although the relative performance of the variants varies between users, it is clear that explicit and implicit feedback are better than the baseline. This can be seen from

Table 2. (a) The rounded average numbers of images that each user corrected when using the three variants of the system. The best performance(s) are marked in bold for each user. (b) The rounded average numbers of images corrected for each category averaged over 18 users. (c) Means and variances over the 18 users for the three system variants.

(a)

User	Baseline	Explicit	Implicit
1	16	23	**24**
2	**22**	21	**22**
3	**26**	25	**26**
4	19	**27**	**27**
5	24	**27**	23
6	23	**27**	26
7	25	25	**26**
8	23	**29**	19
9	23	**26**	24
10	18	24	**28**
11	15	**26**	22
12	25	22	**26**
13	22	**28**	27
14	22	**28**	24
15	**27**	25	22
16	28	**29**	27
17	27	**28**	25
18	26	24	**27**

(b)

Category	Baseline	Explicit	Implicit
car	20	**23**	22
dog	22	**28**	26
bicycle	23	**27**	25
person	23	26	**31**
motorbike	**26**	24	23
train	22	**26**	23

(c)

	Baseline	Explicit	Implicit
mean	22.83	25.78	24.72
var	14.15	5.48	5.74

[4] `http://www.tobii.com/`

the averages, and from the fact that the baseline has the best performance in only three cases. For users 10 and 12, the implicit variant of the system retrieved about 17% more of the wrongly-tagged images than the explicit variant.

Table 2b gives the quantitative performance for each tagging category averaged over all the users. Similarly, the performances of the explicit and implicit variants are better than that of the baseline type, except for the *motorbike* category. The reason is probably because of the overlapping categories of the images in the VOC2007 database. For example, an image tagged as *motorbike* usually contains a person riding on it, which might cause users to tag it as *person*. However, for the *person* category, the implicit variant of the system retrieved about 20% more of the wrongly-tagged images than the explicit variant did.

The Evaluation of User Experience. A close examination of the qualitative feedback from the users (questionaires) indicates that most of the test subjects (between 66% and 75%) believed that all the variants help to facilitate the tagging tasks, though they had to spend extra efforts in adapting to the eye tracker and microphone. As for reliability, about 82% of the test subjects considered the explicit variant with speech input to be the most reliable one, whereas respectively 56% and 50% of the subjects marked the implicit variant and baseline variant to be reliable. As for speed, the implicit variant with gaze tracking received the highest vote of 64%, followed by the explicit variant of 56% and the baseline variant of 43%.

5 Conclusions and Future Work

We have developed a novel client-side multimodal information collector that has been implemented as a Firefox browser extension for asynchronously transmitting versatile user interaction modalities, such as eye movements on the displayed images, to the remote CBIR server. The collector has been integrated with an existing neural-network-based CBIR system that is made capable of handling XMLHttpRequest messages from the client.

We have verified the functionality of the collector by evaluating the performance of the gaze-enhanced CBIR system with real user image tagging tasks. The quantitative results showed that both the explicit variant using pointer clicks and the implicit variant using gaze tracking patterns can to some extent speed up the search and correction of wrongly-tagged images, compared to the baseline variant with drop-down menus. The qualitative results revealed that the implicit variant enhanced by gaze and speech was believed to have the highest speed among the three.

Our next step is to improve the client-server system by fusing the user's eye movements and mouse tracks with more users involved. This time the image database will be expanded to contain millions of images sampled from Flickr and ImageNet.

Acknowledgements. This work is supported by the ICS Department of Aalto University and has received funding from the European Community's Seventh

Framework Programme (FP7/2007–2013) under *grant agreement* n° 21652, Personal Information Navigator Adapting Through Viewing, PinView. We gratefully acknowledge Ms. Na Li for helping the image tagging experiments.

References

1. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Image retrieval: Ideas, influences, and trends of the new age. ACM Computing Surveys 40(2), 1–60 (2008)
2. Kelly, D., Teevan, J.: Implicit feedback for inferring user preference: a bibliography. SIGIR Forum 37(2), 18–28 (2003)
3. Zhang, H., Koskela, M., Laaksonen, J.: Report on forms of enriched relevance feedback. Technical Report TKK-ICS-R10, Helsinki University of Technology (2008)
4. Hardoon, D.R., Shawe-Taylor, J., Ajanki, A., Puolamäki, K., Kaski, S.: Information retrieval by inferring implicit queries from eye movements. In: Eleventh International Conference on Artificial Intelligence and Statistics (2007)
5. Rayner, K.: Eye movements in reading and information processing: 20 years of research. Psychological Bulletin 124(3), 372–422 (1998)
6. Klami, A., Saunders, C., de Campos, T., Kaski, S.: Can relevance of images be inferred from eye movements? In: Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval, pp. 134–140. ACM (2008)
7. Hardoon, D., Pasupa, K.: Image ranking with implicit feedback from eye movements. In: Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications, pp. 291–298. ACM (2010)
8. Maglio, P.P., Campbell, C.S.: Attentive agents. Commun. ACM 46(3), 47–51 (2003)
9. Gruenstein, A., McGraw, I., Badr, I.: The WAMI Toolkit for developing, deploying, and evaluating web-accessible multimodal interfaces. In: Proceedings of Tenth International Conference on Multimodal Interfaces (ICMI 2008), Chania, Greece (October 2008)
10. Laaksonen, J., Koskela, M., Oja, E.: PicSOM—Self-organizing image retrieval with MPEG-7 content descriptions. IEEE Transactions on Neural Networks, Special Issue on Intelligent Multimedia Processing 13(4), 841–853 (2002)
11. Kohonen, T.: Self-Organizing Maps, 3rd edn. Springer Series in Information Sciences, vol. 30. Springer, Berlin (2001)
12. Viitaniemi, V., Laaksonen, J.: Evaluating the performance in automatic image annotation: example case by adaptive fusion of global image features. Signal Processing: Image Communications 22(6), 557–568 (2007)
13. Ames, M., Naaman, M.: Why we tag: motivations for annotation in mobile and online media. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 971–980. ACM, New York (2007)
14. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2007, VOC 2007 (2007)