

Gaze- and Speech-Enhanced Content-Based Image Retrieval in Image Tagging

He Zhang¹, Teemu Ruokolainen¹, Jorma Laaksonen¹,
Christina Hochleitner², and Rudolf Traunmüller²

¹ Department of Information and Computer Science
Aalto University School of Science, Espoo, Finland
{he.zhang, teemu.ruokolainen, jorma.laaksonen}@aalto.fi
² celum gmbh., Linz, Austria
{christina.hochleitner, rudolf.traunmueller}@celum.com

Abstract. We describe a setup and experiments where users are checking and correcting image tags given by an automatic tagging system. We study how much the application of a content-based image retrieval (CBIR) method speeds up the process of finding and correcting the erroneously-tagged images. We also analyze the use of implicit relevance feedback from the user’s gaze tracking patterns as a method for boosting up the CBIR performance. Finally, we use automatic speech recognition for giving the correct tags for those images that were wrongly tagged. The experiments show a large variance in the tagging task performance, which we believe is primarily caused by the users’ subjectivity in image contents as well as their varying familiarity with the gaze tracking and speech recognition setups. The results suggest potentials for gaze and/or speech enhanced CBIR method in image tagging, at least for some users.

Keywords: content-based image retrieval, automatic image tagging, gaze tracking, speech recognition

1 Introduction

Automatic tagging is a useful but still not fully reliable means for associating keyword-type information to unannotated images. In the current state of art, human effort is still needed for checking and correcting the tags [8, 1]. The tag correction process can be seen as a special case of content-based image retrieval (CBIR) where the goal is to quickly correct the erroneously-tagged images.

We present the results of a study that has addressed the image tag correction and assignment task from a multitude of viewpoints. 1) we have studied whether neural-network-based CBIR techniques can speed up finding of erroneously-tagged images. 2) we have used implicit relevance feedback from the user’s gaze tracking patterns to boost the performance of CBIR compared to using only explicit feedback from pointer clicks. 3) we have applied automatic speech recognition for providing corrected image tags. Section 2 describes the CBIR system and Section 3 the automatic speech recognizer used in the experiment. The results are presented in Section 4 and conclusions drawn in Section 5.

Table 1. PicSOM’s visual features extracted from images [11].

Feature	Tiling	Dim.
DCT coefficients of average color in rectangular grid	global	12
CIE L*a*b* color of two dominant color clusters	global	6
Histogram of local edge statistics	4×4	80
Haar transform of quantized HSV color histogram	global	256
Average CIE L*a*b* color	5	15
Three central moments of CIE L*a*b* color distribution	5	45
Histogram of four Sobel edge directions	5	20
Co-occurrence matrix of four Sobel edge directions	5	80
Magnitude of the 16×16 FFT of Sobel edge image	global	128
Histogram of relative brightness of neighboring pixels	5	40
Histogram of interest point SIFT features	global	256

2 Gaze-Enhanced Content-Based Image Retrieval

The typical setting for interactive content-based image retrieval is that the user identifies images that are the most relevant for his current search task (or visually the most similar ones to the target image he has in mind) and passes this information to the system as relevance feedback on the images the system has retrieved. One can differentiate between positive and negative relevance feedback. The former is given for the relevant images and the user is expecting to see more similar ones, whereas the latter is given for the non-relevant images the user does not want to view. Based on the relevance feedback the system is then able to find more similar images and present them to the user (see for example [3] for an extensive survey).

2.1 PicSOM CBIR System

PicSOM³ [7] 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) [6] 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 and their dimensionalities are listed in Table 1.

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 <http://www.cis.hut.fi/picsom>

Table 2. Eye movement features collected in PinView [5].

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

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

2.2 Using Gaze Patterns as Implicit Relevance Feedback

The PinView project⁴ has studied the use of gaze patterns as a form of on-line implicit relevance feedback in interactive CBIR [2]. Based on the analyzed gaze patterns we calculate for each viewed image a 19-dimensional feature vector as specified in Table 2. Based on these features, relevance predictions for the 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.

⁴ <http://www.pinview.eu/>

2.3 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, now 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.

3 Automatic Speech Recognition

Automatic speech recognition (ASR) aims at providing a textual transcript for a given uttered speech signal. The acoustic and language models and the lexicon have been trained earlier with available training corpora. During recognition, spectral feature vectors are first extracted from the speech signal at regular time intervals. Then, given the acoustic information and the probabilities provided by the models, the decoder finds the best transcript hypothesis. Finally, the best hypothesis is output as the recognition result.

The speech recognition system used in the experiments has been developed in the Department of Information and Computer Science at 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. State durations are modeled with gamma probability functions described in [10]. The Hidden Markov model was trained with Wall Street Journal (WSJ) speech corpus; the WSJ corpus consists of dictated newswire articles spoken by American English speakers.

The speech recognition system supports the n -gram models [9] as the default language model type. Therefore, the language model based on the keyword list of the experiment is created in the following straightforward manner: each keyword is considered a word in the vocabulary and each word is assigned an equal probability. In other words, we model the keyword list as a unigram model with equal probability assigned to each word.

4 Experiments

The experiments were conducted to validate the relative performance of different variants of the system and to verify how well the system works in practice.

Data. We used the *train* subset of the PASCAL Visual Object Classes Challenge 2007 (VOC2007) data set [4] 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. Almost none 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 every task, the user had to check and correct the tags of one particular category. Before each task, the system randomly selected 40 images of that category and another 40 images of the ten categories from the wrongly-tagged class. 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 tracker⁵. 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 3a gives the quantitative performance of the three variants for each user. Although the relative performance of the variants varies between users, it is clear

⁵ <http://www.tobii.com/>

Table 3. (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) Pairwise t -test p -values for variants in user-wise experiment. (d) Pairwise t -test p -values for variants in category-wise experiment.

(a)				(b)			
User	Baseline	Explicit	Implicit	Category	Baseline	Explicit	Implicit
1	16	23	24	car	20	23	22
2	22	21	22	dog	22	28	26
3	26	25	26	bicycle	23	27	25
4	19	27	27	person	23	26	31
5	24	27	23	motorbike	26	24	23
6	23	27	26	train	22	26	23
7	25	25	26	Average	23	26	25
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				
Average	23	26	25				

(c)		
	Explicit	Implicit
Baseline	0.0080	0.0812
Explicit	—	0.1901

(d)		
	Explicit	Implicit
Baseline	0.0218	0.1665
Explicit	—	0.6747

that the explicit and implicit feedback variants are better than the baseline (see t -test values in Table 3c). The difference between the explicit variant using only clicks and the implicit variant using also gaze is not significant (with a pairwise t -test p -value of 0.1901), though the former on average ranks higher than the latter. The worst implicit result was that of user 8 with whom the gaze tracking obviously failed. For users 10 and 12, the implicit variant of the system retrieved about 17% more of the wrongly-tagged images than the explicit variant did.

Table 3b 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 (see t -test values in Table 3d), 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*. Again the difference between the explicit variant and the implicit variant is not significant (with a pairwise t -test p -value of 0.6747), though the former slightly outperforms the latter. 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 and results of speech recognizer. Speech recognition accuracy is commonly reported with the word error rate (WER) defined as the number of erroneously recognized words divided by the total number of spoken words. In our experiments, obtaining the WER would have required a manual annotation of the speech recorded during the experiments, i.e. which word was uttered at which time instance, followed by a (partly) manual search for the corresponding recognized words from the speech recognizer’s output.

Due to tediousness of such task, we instead investigated other means of evaluation; particularly, we consider the increase in the number of correctly labeled images after the annotations divided by the total number of annotations conducted by the users. This gives us an efficiency score with maximum value 1.0 in case each annotation results in a correctly labeled image; value zero in case there is no change in the number of correctly labeled images; and negative values in case the annotation results in a decrease in correctly labeled images.

Averaged over all the 18 users with four experiments using speech each, we obtain a mean efficiency of 0.36. This value means that, on average, a successful labeling required roughly three annotation trials. The highest efficiency achieved in an individual experiment was 0.60 and the lowest 0.05; the corresponding numbers of trials per successful labeling are roughly 1.5 and 20, respectively, indicating a large variance in user-wise performance. However, according to our observations, the reason for the varying performance need not be technical in nature. Instead, we suggest that the most important factor leading to performance variance is the lack of user experience with speech recognizers.

The evaluation of user experience. A close examination of the qualitative feedback from the users 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, while 56% marked the implicit variant, and only 50% marked the 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

We investigated the use of neural-network-based CBIR system enhanced by gaze feedback modality and speech input for an image tagging task. Three relevance feedback variants were evaluated combined with automatic speech recognition being applied for providing the corrected image tags. Our 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.

Based on the quantitative evaluation, the explicit variant generally outperformed the other two, whereas from the qualitative evaluation, the implicit variant was believed to have the highest speed among the three and was considered

more reliable than the baseline, even though most of the test subjects reported uncomfot in sitting still in front of the eye tracker and difficulties in adapting to the microphone. These results imply, in addition to the quantitative evaluation of the tagging system, that people’s subjective characteristics and preferences vary with respect to willingness and ability to interact with novel multimodal interfaces. This will need to be taken into account in future studies.

Acknowledgments The research leading to these results has received funding from the European Community’s Seventh Framework Programme (FP7/2007–2013) under *grant agreement* n° 216529, Personal Information Navigator Adapting Through Viewing, PinView.

References

1. 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, NY, USA (2007)
2. Auer, P., Hussain, Z., Kaski, S., Klami, A., Kujala, J., Laaksonen, J., Leung, A.P., Pasupa, K., Shawe-Taylor, J.: Pinview: Implicit feedback in content-based image retrieval. In: Diethe, T., Cristianini, N., Shawe-Taylor, J. (eds.) Proceedings of Workshop on Applications of Pattern Analysis. JMLR Workshop and Conference Proceedings, vol. 11, pp. 51–57 (2010)
3. 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 (April 2008)
4. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. <http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html> (2007)
5. Klami, A., Kaski, S., Pasupa, K., Saunders, C., de Campos, T.: Prediction of relevance of an image from a scan pattern. PinView FP7-216529 Project Deliverable Report D2.1 (December 2008), available online at <http://www.pinview.eu/deliverables.php>
6. Kohonen, T.: Self-Organizing Maps, Springer Series in Information Sciences, vol. 30. Springer-Verlag, Berlin, third edn. (2001)
7. 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 (July 2002)
8. Lerman, K., Jones, L.: Social browsing on flickr. CoRR abs/cs/0612047 (2006)
9. Manning, C., Schütze, H., MITCogNet: Foundations of statistical natural language processing, vol. 59. MIT Press (1999)
10. Pylkkönen, J., Kurimo, M.: Duration modeling techniques for continuous speech recognition. In: Eighth International Conference on Spoken Language Processing. ISCA (2004)
11. 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 (July 2007)