

# A Corpus-Based Relevance Feedback Approach to Cross-Language Image Retrieval

Yih-Chen Chang<sup>1</sup>, Wen-Cheng Lin<sup>2</sup>, and Hsin-Hsi Chen<sup>1</sup>

<sup>1</sup> Department of Computer Science and Information Engineering  
National Taiwan University  
Taipei, Taiwan

ycchang@nlg.csie.ntu.edu.tw, hhchen@csie.ntu.edu.tw

<sup>2</sup> Department of Medical Informatics  
Tzu Chi University  
Hualien, Taiwan  
denislin@mail.tcu.edu.tw

**Abstract.** This paper regards images with captions as a cross-media parallel corpus, and presents a corpus-based relevance feedback approach to combine the results of visual and textual runs. Experimental results show that this approach performs well. Comparing with the mean average precision (MAP) of the initial visual retrieval, the MAP is increased from 8.29% to 34.25% after relevance feedback from cross-media parallel corpus. The MAP of cross-lingual image retrieval is increased from 23.99% to 39.77% if combining the results of textual run and visual run with relevance feedback. Besides, the monolingual experiments also show the consistent effects of this approach. The MAP of monolingual retrieval is improved from 39.52% to 50.53% when merging the results of the text and image queries.

## 1 Introduction

In cross-language image retrieval, users employ textual queries in one language and example images to access image database with text descriptions in another language. It becomes practical because many images associating text like captions, metadata, Web page links, and so on, are available nowadays. Besides, the neutrality of images to different language users resolves the arguments that users not familiar with the target language still cannot afford to understand the retrieved documents in cross-language information retrieval.

Two types of approaches, i.e., content-based and text-based approaches, are usually adopted in image retrieval [1]. Content-based image retrieval (CBIR) uses low-level visual features to retrieve images. In such a way, it is unnecessary to annotate images and translate users' queries. However, due to the semantic gap between image visual features and high-level concepts [2], it is still challenging to use a CBIR system to retrieve images with correct semantic meaning. Integrating textual information may help a CBIR system to cross the semantic gap and improve retrieval performance.

Recently, many approaches have tried to combine text- and content-based methods for image retrieval. A simple approach is conducting text- and content-based retrieval

separately and merging the retrieval results of the two runs [3, 4]. In contrast to the parallel approach, a pipeline approach uses textual or visual information to perform initial retrieval, and then uses the other feature to filter out irrelevant images [5]. In these two approaches, textual and visual queries are formulated by users and do not directly influence each other. Another approach, i.e., transformation-based approach [12], mines the relations between images and text, and uses the mined relations to transform textual information into visual one, and vice versa.

To formulate the cross-media translation between visual and textual representations, several correlation-based approaches have been proposed. Mori, Takahashi and Oka [6] divided images into grids, and then the grids of all images were clustered. Co-occurrence information was used to estimate the probability of each word for each cluster. Duygulu, *et al.* [7] used blobs to represent images. First, images are segmented into regions using a segmentation algorithm like Normalized Cuts [8]. All regions are clustered and each cluster is assigned a unique label (blob token). The Expectation-Maximization (EM) algorithm [9] is used to construct a probability table that links blob tokens with word tokens. Jeon, Lavrenko, and Manmatha [10] proposed a cross-media relevance model (CMRM) to learn the joint distribution of blobs and words. They further proposed continuous-space relevance model (CRM) that learned the joint probability of words and regions, rather than blobs [11]. Lin, Chang and Chen [12] transformed a textual query into visual one using a transmedia dictionary.

The above approaches use the relation between text and visual representation as a bridge to translate image to text. However, it is hard to learn all relations between all visual and textual features. Besides, the degree of ambiguity of the relations is usually high. For example, visual feature “red circle” may have many meanings such as sun set, red flower, red ball, *etc.* Similarly, the word “flower” may have different looks of images, e.g., different color and shape. In contrast to the transmedia dictionary approach [12], this paper regards images with captions as a cross-media parallel corpus to transform visual features to textual ones. The text descriptions of the top- $n$  retrieved images of the initial image retrieval are used for feedback to conduct a second retrieval. The new textual information can help us determine the semantic meaning of a visual query, and thus improve retrieval performance.

The rest of the paper is organized as follows. Section 2 presents the proposed approach and Section 3 shows the experimental results in bilingual ad hoc retrieval task at ImageCLEF2005. Section 4 provides some discussion and Section 5 ends the paper with concluding remarks.

## 2 A Corpus-Based Relevance Feedback Approach

In this paper, we translate visual and textual features without learning correlations. We treat the images along with their text descriptions as an aligned cross-media parallel corpus, and a corpus-based method transforms a visual query to a textual one. Figure 1 shows the concept of this approach.

In cross-language image retrieval, given a set of images  $I = \{i_1, i_2, \dots, i_m\}$  with text descriptions  $T_{L1} = \{t_1, t_2, \dots, t_m\}$  in language  $L1$ , users issue a textual query  $Q_{L2}$  in

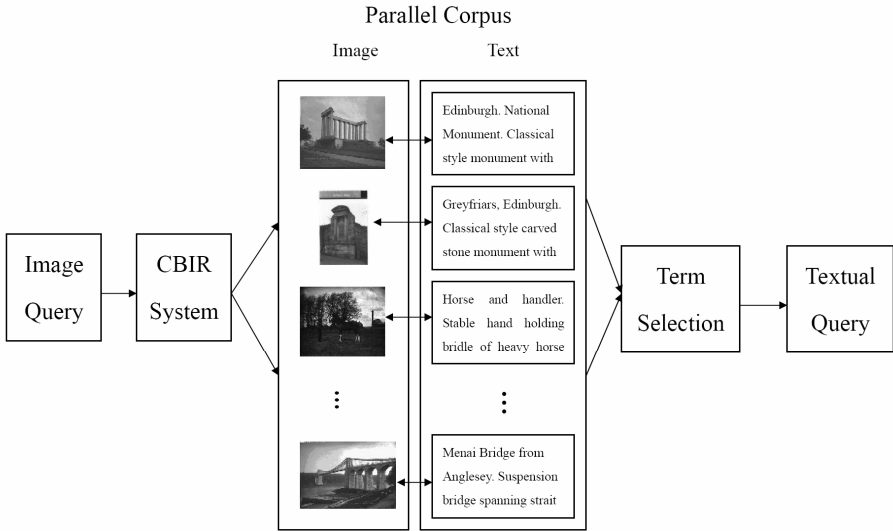


Fig. 1. Key concept of a corpus-based approach

language  $L2$  ( $L2 \neq L1$ ) and example images  $E = \{e_1, e_2, \dots, e_p\}$  to retrieve relevant images from  $I$ . At first, we submit example images  $E$  as initial query to a CBIR system, e.g., VIPER [13], to retrieve images from  $I$ . The retrieved images are  $R = \{r_{i1}, r_{i2}, \dots, r_{in}\}$  and their text descriptions are  $T_{R,L1} = \{t_{ri1}, t_{ri2}, \dots, t_{rin}\}$  in language  $L1$ . Then, we select terms from the text descriptions of the top  $k$  retrieved images to construct a new textual query. The new textual query can be seen as a translation of initial visual query by using a corpus-based approach. We submit the new textual query to a text-based retrieval system, e.g., Okapi [14], to retrieve images from  $I$ . That is latter called a *feedback run*.

Figure 2 shows how to integrate the feedback process into a cross-language image retrieval system. In addition to the visual feedback run, we also conduct a text-based run using the textual query in the test set. We use the method proposed in ImageCLEF 2004 [15] to translate textual query  $Q_{L2}$  into query  $Q_{L1}$  in language  $L1$ , and submit the translated query  $Q_{L1}$  to the Okapi system to retrieve images. The results of textual run and visual feedback run can be combined. The similarity scores of images in the two runs are normalized and linearly combined using equal weight.

### 3 Experimental Results

In the experiments, we used historic photographs from the St. Andrews University Library<sup>1</sup> [16]. There are 28,133 photographs, which are accompanied by a textual description written in British English. The ImageCLEF test collection contains 28

<sup>1</sup> <http://www-library.st-andrews.ac.uk/>

topics, and each topic has text description in different languages and two example images. In our experiments, queries are in traditional Chinese. Figure 3 shows an image and its description. Figure 4 illustrates a topic in English and in Chinese.

The text-based retrieval system is Okapi IR system, and the content-based retrieval system is VIPER system. The <HEADLINE> and <CATEGORIES> sections, and the record body of English captions are used for indexing. The weighting function is BM25. Chinese queries and example images are used as the source queries.

In the formal runs, we submitted four Chinese-English cross-lingual runs, two English monolingual runs and one visual run in CLEF 2005 image track. In English monolingual runs, using narrative or not using narrative will be compared. In the four cross-lingual runs, combining with visual run or not combining with visual run, and using narrative or not using narrative will be compared. The details of the cross-lingual runs and visual run are described as follows.

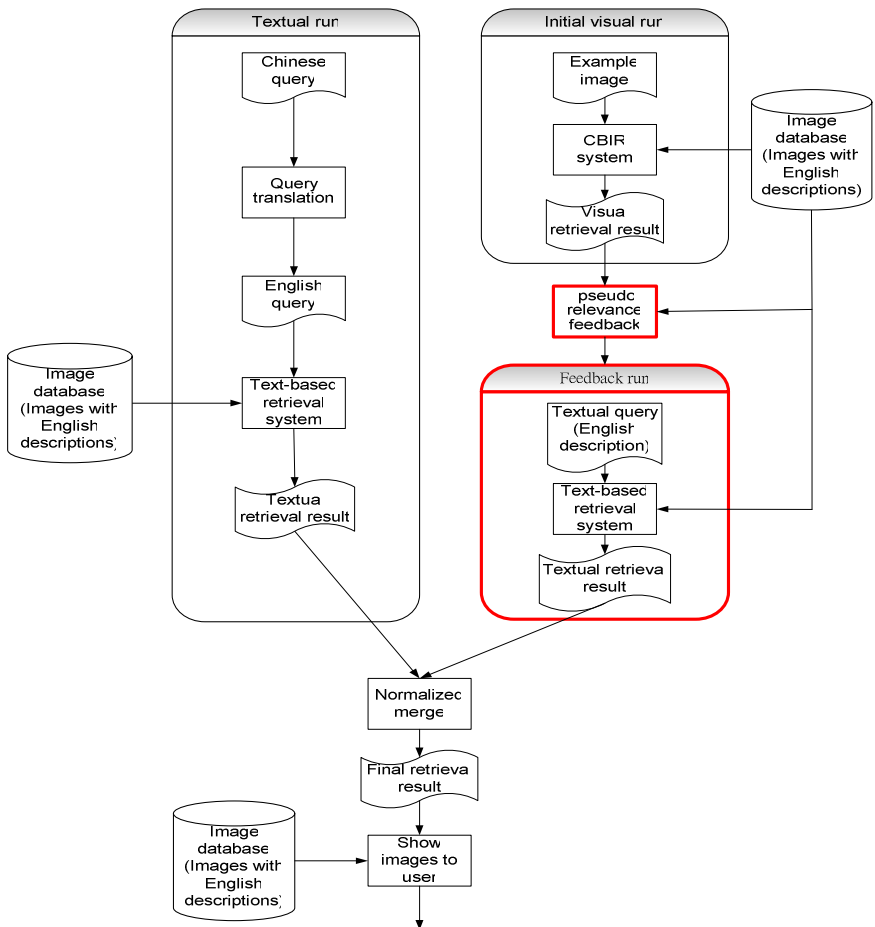


Fig. 2. A cross-language image retrieval system


|   |  |
|---|--|
|  | <pre> &lt;DOC&gt; &lt;DOCNO&gt; stand03_1041/stand03_9914.txt &lt;/DOCNO&gt; &lt;HEADLINE&gt; Azay le Rideau. Bridge. &lt;/HEADLINE&gt; &lt;TEXT&gt; &lt;RECORD_ID&gt; JEAS-.000032.-.000045 &lt;/RECORD_ID&gt;   Azay le Rideau.   Round tower with conical roof attached to large three-storey   building; low bridge spanning still water to right.   1907   John Edward Aloysius Steggall   Indre et Loire, France   JEAS-32-45 pc/jf &lt;CATEGORIES&gt;   [towers - round], [towers - conical roofed], [France urban   views], [France all views] &lt;/CATEGORIES&gt; &lt;SMALL_IMG&gt;   stand03_1041/stand03_9914.jpg &lt;/SMALL_IMG&gt; &lt;LARGE_IMG&gt;   stand03_1041/stand03_9914_big.jpg &lt;/LARGE_IMG&gt; &lt;/TEXT&gt; &lt;/DOC&gt; </pre> |
|---|--|

Fig. 3. An image and its description


|  |  |
|--|--|
|  | <pre> &lt;top&gt; &lt;num&gt; Number: 17 &lt;/num&gt; &lt;title&gt; man or woman reading &lt;/title&gt; &lt;narr&gt;   Relevant images will show men or women reading books   or a paper. People performing any other activity are not   relevant. &lt;/narr&gt; &lt;/top&gt;  &lt;top&gt; &lt;num&gt; Number: 17 &lt;/num&gt; &lt;title&gt;   正在閱讀的男人或女人 &lt;/title&gt; &lt;/top&gt; </pre> |
|--|--|

Fig. 4. A Topic in English and in Chinese

## (1) NTU-adhoc05-CE-T-W

This run employs textual queries (title field only) to retrieve images. We use the query translation method as proposed for CLEF 2004 [15] to translate Chinese queries into English ones, and the Okapi IR system retrieves images based on a textual index.

## (2) NTU-adhoc05-CE-TN-W-Ponly

This run uses textual queries (title plus narrative fields). Only the positive information in narrative field is considered. The sentences that contain phrase “are not relevant” are removed to avoid noise [17].

## (3) NTU-adhoc05-EX-prf

It is a visual run with pseudo relevance feedback. VIPER system provided by ImageCLEF retrieves the initial results, and the text descriptions of the top 2 images are used to construct a textual query. The textual query is submitted to Okapi IR system to retrieve images.

## (4) NTU-adhoc05-CE-T-WEprf

This run merges the results of NTU-adhoc05-CE-T-W and NTU-adhoc05-EX-prf. The similarity scores of images in the two runs are normalized and linearly combined with equal weight 0.5.

## (5) NTU-adhoc05-CE-TN-WEprf-Ponly

This run merges the results of NTU-adhoc05-CE-TN-W-Ponly and NTU-adhoc05-EX-prf.

## (6) NTU-adhoc05-EE-T-W

This run is a monolingual run by using title field only.

## (7) NTU-adhoc05-EE-TN-W-Ponly

This run is a monolingual run by using title and narrative fields.

Two unofficial runs shown as follows are also conducted for comparison.

## (8) NTU-adhoc05-EE-T-WEprf

This run merges the results of NTU-adhoc05-EE-T-W and NTU-adhoc05-EX-prf.

## (9) VIPER

This run is the initial visual run.

Tables 1 and 2 show the experimental results of official runs and unofficial runs, respectively. The Mean Average Precision (MAP) of the textual query using title and narrative is better than that of the textual query using title only, but the difference is not significant. That is,

NTU-adhoc05-CE-TN-W-Ponly > NTU-adhoc05-CE-T-W,

NTU-adhoc05-CE-TN-WEprf-Ponly > NTU-adhoc05-CE-T-WEprf, and

NTU-adhoc05-EE-TN-W-Ponly > NTU-adhoc05-EE-T-W.

Besides, the MAP of integrating textual and visual queries by using corpus-based relevance feedback approach is much better than that of textual query only. That is,

NTU-adhoc05-CE-T-WEprf > NTU-adhoc05-CE-T-W,

NTU-adhoc05-CE-TN-WEprf-Ponly > NTU-adhoc05-CE-TN-W-Ponly, and

NTU-adhoc05-EE-T-WEprf > NTU-adhoc05-EE-T-W.

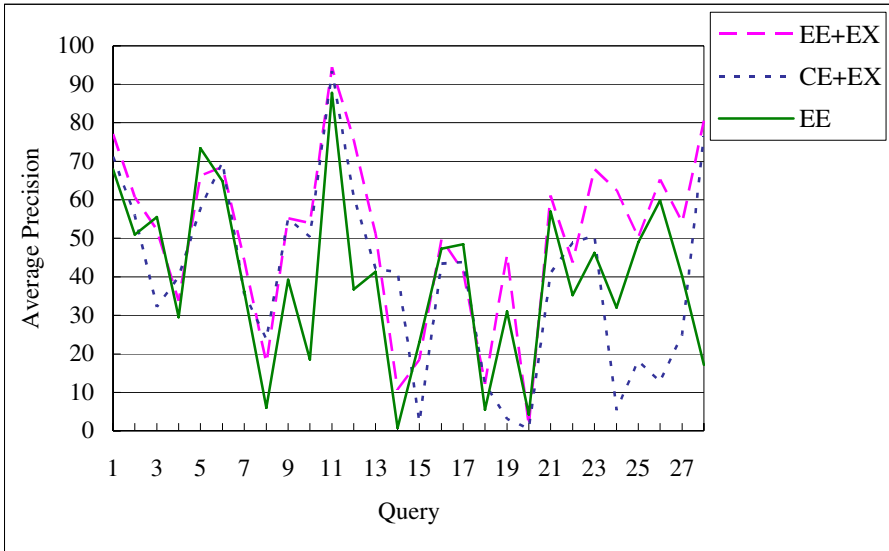
Although the MAP of initial visual run is only 8.29%, the effects from relevance feedback improve the performance significantly. Figure 5 illustrates the average precision of each query using NTU-adhoc05-EE-T-WEprf (EE+EX), NTU-adhoc05-CE-T-WEprf (CE+EX), NTU-adhoc05-EE-T-W (EE), NTU-adhoc05-EX-prf

**Table 1.** Results of official runs

| Run                           | Features in Query                   |                               | MAP    |
|-------------------------------|-------------------------------------|-------------------------------|--------|
|                               | Text                                | Visual                        |        |
| NTU-adhoc05-CE-T-W            | Chinese (Title)                     | None                          | 0.2399 |
| NTU-adhoc05-CE-TN-W-Ponly     | Chinese (Title+ Positive Narrative) | None                          | 0.2453 |
| NTU-adhoc05-CE-T-WEprf        | Chinese (Title)                     | Example image                 | 0.3977 |
| NTU-adhoc05-CE-TN-WEprf-Ponly | Chinese (Title+ Positive Narrative) | Example image                 | 0.3993 |
| NTU-adhoc05-EX-prf            | English (feedback query)            | Example image (initial query) | 0.3425 |
| NTU-adhoc05-EE-T-W            | English                             | None                          | 0.3952 |
| NTU-adhoc05-EE-TN-W-Ponly     | English (Title+ Positive Narrative) | None                          | 0.4039 |

**Table 2.** Performances of unofficial runs

| Run                        | Features in Query |               | MAP    |
|----------------------------|-------------------|---------------|--------|
|                            | Text              | Visual        |        |
| NTU-adhoc05-EE-T-WEprf     | English (Title)   | Example image | 0.5053 |
| Initial Visual Run (VIPER) | None              | Example image | 0.0829 |



**Fig. 5.** Average precision of each query

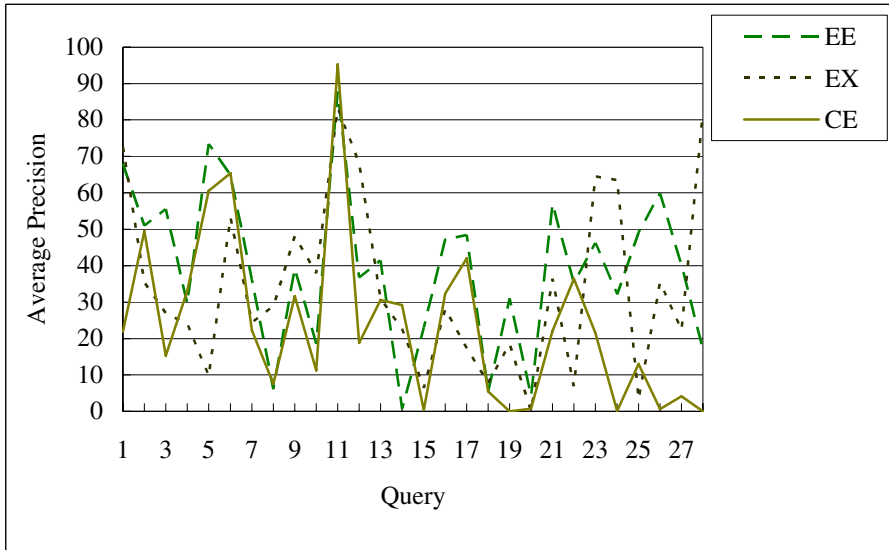


Fig. 5. Average precision of each query (*Continued*)

(EX), and NTU-adhoc05-CE-T-W (CE). In summary,  $EE + EX > CE + EX \cong EE > EX > CE > \text{visual run}$ .

## 4 Discussion

The MAP of monolingual retrieval using the title field only is 39.52%. Comparing with our performance at ImageCLEF 2004 [15], i.e., 63.04%, topics of this year is more general and more visual than those of last year, e.g., waves breaking on beach, dog in sitting position, *etc.* The MAP of Chinese-English cross-lingual run (23.99%) is 60.70% of that of English monolingual run (39.52%). It shows that there are still many errors in language translation.

The MAP of initial visual run, i.e., VIPER, is not good enough. Text-based runs, even cross-lingual runs, perform much better than initial visual run. It shows that semantic information is very important for the queries of this year. After relevance feedback, the performance is increased dramatically from 8.29% to 34.25%. The result shows that the feedback method transforms visual information into textual one. Combining textual and visual feedback runs further improves retrieval performance.

Figure 6 shows the first three returned images of query “aircraft on the ground”. For monolingual case, the images containing aircrafts not on the ground are reported wrongly. For cross-lingual case, “地面上的飛機” is translated to “aircraft above the floor”, which captures wrong images. For visual case, the feedback query “aircraft in military air base” captures more relevant images. This is because aircrafts in military air base are very likely to be parked and thus are on the ground.



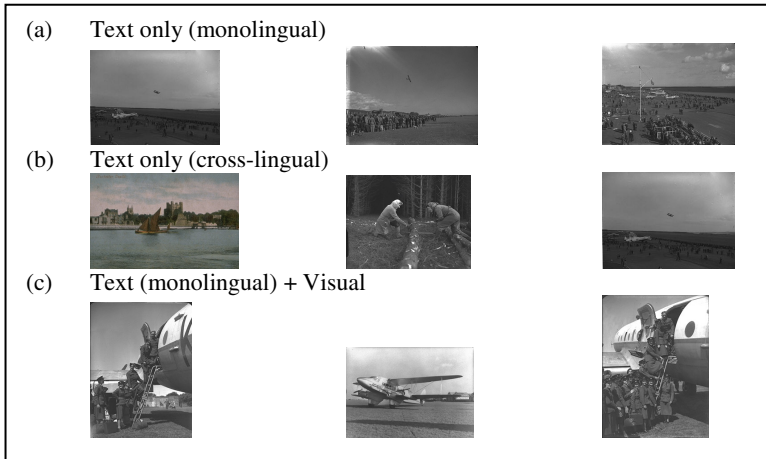


Fig. 6. Retrieval results of query “Aircraft on the Ground”

## 5 Conclusion

An approach of combining textual and image features is proposed for Chinese-English image retrieval. A corpus-based feedback cycle is performed after CBIR. Comparing with the MAP of monolingual IR (i.e., 39.52%), integrating visual and textual queries achieves better MAP in cross-language image retrieval (39.77%). It indicates part of translation errors is resolved. The integration of visual and textual queries also improves the MAP of the monolingual IR from 39.52% to 50.53%. It reveals the integration provides more information. The MAP of Chinese-English image retrieval is 78.2% of the best monolingual text retrieval in ImageCLEF 2005. The improvement is the best among all the groups.

## Acknowledgements

Research of this paper was partially supported by National Science Council, Taiwan, under the contracts NSC 94-2752-E-001-001-PAE and NSC 94-2213-E-002-076.

## References

1. Goodrum, A.A.: Image Information Retrieval: An Overview of Current Research. *Information Science*, 3(2). (2000) 63-66.
2. Eidenberger, H. and Breiteneder, C.: Semantic Feature Layers in Content-based Image Retrieval: Implementation of Human World Features. In: *Proceedings of International Conference on Control, Automation, Robotic and Vision*. (2002).
3. Besançon, R., Hède, P., Moellic, P.A., and Fluhr, C.: Cross-Media Feedback Strategies: Merging Text and Image Information to Improve Image Retrieval. In: *5th Workshop of the Cross-Language Evaluation Forum, LNCS 3491*. (2005) 709-717.

4. Jones, G.J.F., Groves, D., Khasin, A., Lam-Adesina, A., Mellebeek, B., and Way, A.: Dublin City University at CLEF 2004: Experiments with the ImageCLEF St. Andrew's Collection. In: 5th Workshop of the Cross-Language Evaluation Forum, LNCS 3491. (2005) 653-663.
5. Baan, J., van Ballegooij, A., Geusenbroek, J.M., den Hartog, J., Hiemstra, D., List, J., Patras, I., Raaijmakers, S., Snoek, C., Todoran, L., Vendrig, J., de Vries, A., Westerveld, T., and Worring, M.: Lazy Users and Automatic Video Retrieval Tools in the Lowlands. In: Proceedings of the Tenth Text REtrieval Conference. National Institute of Standards and Technology (2002) 159-168.
6. Mori, Y., Takahashi, H. and Oka, R.: Image-to-Word Transformation Based on Dividing and Vector Quantizing Images with Words. In: Proceedings of the First International Workshop on Multimedia Intelligent Storage and Retrieval Management. (1999).
7. Duygulu, P., Barnard, K., Freitas, N. and Forsyth, D.: Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary. In: Proceedings of Seventh European Conference on Computer Vision, Vol. 4. (2002) 97-112.
8. Shi, J. and Malik, J.: Normalized Cuts and Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8). (2000) 888-905.
9. Dempster, A.P., Laird, N.M., and Rubin, D.B.: Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1). (1977) 1-38.
10. Jeon, J., Lavrenko, V. and Manmatha, R.: Automatic Image Annotation and Retrieval using Cross-Media Relevance Models. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. (2003) 119-126.
11. Lavrenko, V., Manmatha, R. and Jeon, J.: A Model for Learning the Semantics of Pictures. In: Proceedings of the Seventeenth Annual Conference on Neural Information Processing Systems. (2003).
12. Lin, W.C., Chang, Y.C. and Chen, H.H.: Integrating Textual and Visual Information for Cross-Language Image Retrieval. In: Proceedings of the Second Asia Information Retrieval Symposium, LNCS 3689. (2005) 454-466.
13. Squire, D.M., Müller, W., Müller, H., and Raki, J.: Content-Based Query of Image Databases, Inspirations from Text Retrieval: Inverted Files, Frequency-Based Weights and Relevance Feedback. In: Scandinavian Conference on Image Analysis. (1999) 143-149.
14. Robertson, S.E., Walker, S. and Beaulieu, M.: Okapi at TREC-7: Automatic Ad Hoc, Filtering, VLC and Interactive. In: Proceedings of the Seventh Text REtrieval Conference. National Institute of Standards and Technology (1998) 253-264.
15. Lin, W.C., Chang, Y.C. and Chen, H.H.: From Text to Image: Generating Visual Query for Image Retrieval. In: 5th Workshop of the Cross-Language Evaluation Forum, LNCS 3491. (2005) 664-675.
16. Clough, P., Müller, H., Deselaers, T., Grubinger, M., Lehmann, T.M., Jensen, J., and Hersh, W.: The CLEF 2005 Cross-Language Image Retrieval Track, Proceedings of the Cross Language Evaluation Forum 2005, Springer Lecture Notes in Computer science, 2006 - to appear.
17. Feng, K.M. and Chen, H.H.: Effects of Positive and Negative Information Needs on Information Retrieval. *Bulletin of the College of Engineering, National Taiwan University*, 90. (2004) 35-42.