

From Text to Image: Generating Visual Query for Image Retrieval

Wen-Cheng Lin, Yih-Chen Chang and Hsin-Hsi Chen

Department of Computer Science and Information Engineering
National Taiwan University
Taipei, TAIWAN
{denislin, ycchang}@nlg.csie.ntu.edu.tw; hh_chen@csie.ntu.edu.tw

Abstract

In this paper, we explore the help of visual features to cross-language image retrieval. We propose an approach that transforms textual queries into visual representations. The relationships between text and images are modeled. Visual queries are constructed from textual queries using the relationships. The retrieval results using textual and visual queries are combined to generate the final ranked list. We conducted English monolingual and Chinese-English cross-language retrieval experiments. The performances are quite good. The average precision of English monolingual textual run is 0.6304. The performance of cross-lingual retrieval is about 70% of monolingual retrieval. However, the help of generated visual query is limit. If appropriate query terms are selected to generate visual query, retrieval performance could be increased.

1. Introduction

Multimedia data has an explosive growth nowadays, and more and more people are searching and using it. People need effective and efficient tools to help them to find the information they need from a huge amount of data. Thus, how to retrieve multimedia data efficiently becomes an important research issue. Text retrieval, image retrieval, video retrieval, spoken data retrieval, music retrieval, etc., have been widely studied in recent years. Several evaluation tasks are organized to enhance researches in multimedia information retrieval technologies. TREC 2001 and 2002 contained a video track devoted to research in automatic segmentation, indexing, and content-based retrieval of digital video. Beginning in 2003, TRECVID (<http://www.itl.nist.gov/iaui/894.02/projects/trecvid/>) became an independent evaluation. In CLEF 2003, ImageCLEF (<http://ir.shef.ac.uk/imageclef2004/>) was organized to promote research in cross-language image retrieval.

Two types of approaches, i.e., content-based and text-based approaches, are usually adopted in multimedia retrieval. Content-based approaches use low-level features to represent multimedia objects. In image retrieval, low-level visual features such as color, texture and shape are often used. Text-based approaches use collateral text to describe the objects. Text can describe the content of multimedia objects in detail. Several hybrid approaches (The Lowlands Team, 2002; Westerveld, 2000, 2002) that integrate visual and textual information have been proposed. Experimental results showed that the optimal technique depends on the query. The combined approach could outperform text- and content-based approaches in some cases.

In ImageCLEF 2003, we adopted text-based approaches to deal with Chinese-English cross-language image retrieval problem (Lin, Yang and Chen, 2003). Image captions were used to represent images. Dictionary-based query translation approach was adopted to unify the languages in queries and image captions. Named entities that are not included in dictionary were translated using a similarity-based backward transliteration model. Experimental results showed that using similarity-based backward transliteration increased retrieval performances.

In this paper, we explore the help of visual features to cross-language image retrieval. We propose an approach that transforms textual queries into visual representations. We model the relationships between text and images. Visual queries are constructed from textual queries using the relationships. In addition to textual index, a visual index is built for retrieving images by visual query. The retrieval results using textual and visual queries are combined to generate the final ranked list. The rest of this paper is organized as follows. Section 2 discusses translingual transmedia information access. Section 3 models the relationships between text and images. The method of generating visual representation of textual query is introduced. Section 4 shows the query translation methods. Section 5 shows how to integrate textual and visual information. Section 6 discusses the experimental results. Finally, Section 7 provides concluding remarks.

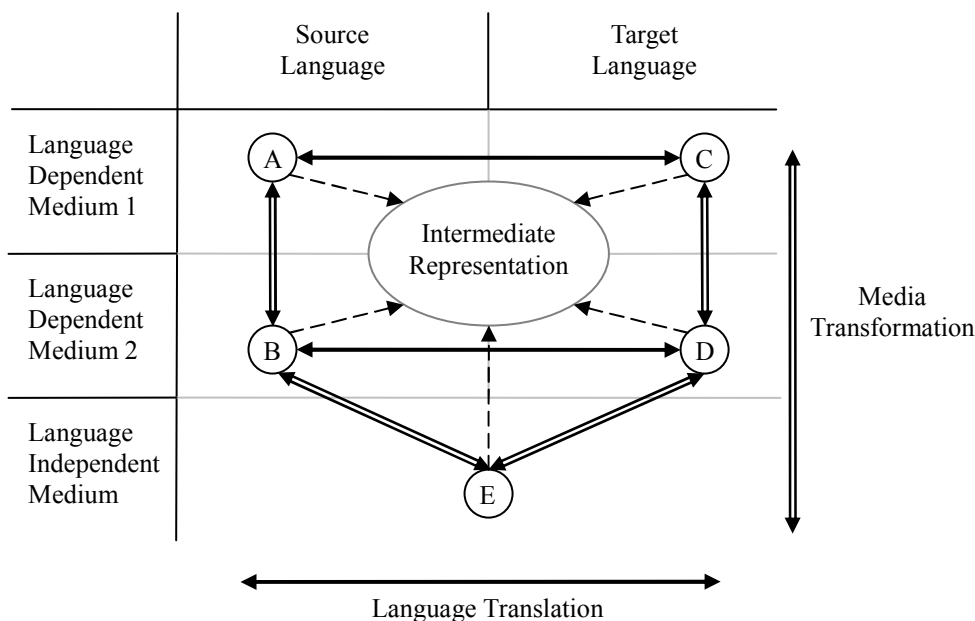


Figure 1. Media Transformation and Language Translation

2. Translingual Transmedia Information Retrieval

Multimedia data consist of different types of media. Some media are language dependent, e.g. text and speech, while the others are language independent, e.g. image and music. How to represent multimedia data is an important issue in multimedia retrieval. If queries and documents are in different types of media, media transformation is needed to unify the representations of queries and documents. If queries and documents are represented by text, but are in different languages, language translation is also needed. Media transformation combining with language translation problem is shown in Figure 1. Horizontal direction indicates language translation, and vertical direction means media transformation. There are several alternatives to unify the media forms and languages of queries and documents. We can transform queries into the same representation as documents, transform documents into the same representation as queries, or transform both of them into an intermediate representation. Take spoken cross-language access to image collection via captions (Lin, Lin and Chen, 2004) as an example. The data that users request are images while query is in terms of speech. Images are language independent, thus are in the E part in Figure 1. Spoken queries are in the A part. Images are represented by captions in a language different from that of query. In this way, the medium of target document is transformed into text, i.e., from E to D. We can transform spoken query in source language into text using a speech recognition system, then translate the textual query into target language and retrieve documents in target language. The transformation path is from A to B, then from B to D.

When using textual query to retrieve images, using text descriptions to represent images is a usually adopted approach to unify the representations of query and images. Image captions that describe the content of images are good material to represent images. In this way, traditional information retrieval systems can be used to retrieve captions. Although using text descriptions to represent images is effective, manually assigned captions are not always available. Automatic annotation has been studied to generate text descriptions of images automatically (Duygulu, *et al.*, 2002; Jeon, Lavrenko and Manmatha, 2003; Lavrenko, Manmatha and Jeon, 2003; Mori, Takahashi and Oka, 1999).

An alternative approach to unify representations is transforming textual queries into visual representations, i.e., from B to E in Figure 1. Content-based approaches can be used to retrieve images using the visual representations of queries. In cross-language information retrieval, translation ambiguity and target polysemy problems have to be tackled in translation process. If a word is not translated correctly, we can't capture the correct meaning of the word in the context. If the translation is polysemous, the undesired documents that contain the translation with other senses could be reported even if the translation is the correct one. Using visual queries could avoid these problems. Visual query draws images that user is looking for, and is used to retrieve images directly. In next section, we will introduce how to generate visual query from textual query.

3. Visual Representation of Text

Given a set of images with text descriptions, we can learn the relationships between images and text. For an image, each word in its description may relative to a portion of this image. If we divide an image into several small parts, e.g. blocks or regions, we could link the words in its description to the parts. This is analogous to word alignment in sentence aligned parallel corpus. If we treat the visual representation of image as a language, the textual description and image parts of an image is an aligned sentence. The correlations between the vocabularies of two languages can be learned from the aligned sentences. In automatic annotation task, several approaches are proposed to model the correlations of text and visual representation, and generate text descriptions from images. Mori, Takahashi and Oka (1999) divided images into grids, and then the grids of all images are clustered. Co-occurrence information is used to estimate the probability of each word for each cluster. Duygulu, *et al.* (2002) used blobs to represent images. First, images are segmented into regions using a segmentation algorithm like Normalized Cuts (Shi and Malik, 2000). All regions are clustered and each cluster is assigned a unique label (blob token). EM algorithm is used to construct a probability table that links blob tokens with word tokens. Jeon, Lavrenko, and Manmatha (2003) 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 learning the joint probability of words and regions, rather than blobs (Lavrenko, Manmatha, and Jeon, 2003).

As in Duygulu, *et al.* (2002), we use blobs as the visual representations of images. Blobworld (Carson, *et al.*, 2002) is used to segment an image into regions. The regions of all images are clustered into 2,000 clusters by K-means clustering algorithm. The correlations between annotated words and blobs are learned from a set of images with text descriptions. Mutual Information (MI) is adopted to measure the strength of correlation between an image blob and a word. Let x be a word and y be an image blob. The Mutual Information of x and y is defined as follow.

$$MI(x, y) = p(x, y) \times \log \frac{p(x, y)}{p(x)p(y)} \quad (1)$$

where $p(x)$ is the occurrence probability of word x in text descriptions,
 $p(y)$ is the occurrence probability of blob y in image blobs, and
 $p(x, y)$ is the probability that x and y occur in the same image.

Given a word w_i , we can generate relative blobs according the learned MI. We can set a threshold to select blobs. The blobs whose MI values with w_i exceed the threshold are associated to w_i . The generated blobs can be seen as the visual representation of w_i .

4. Query Translation

In the experiments, Chinese query set is used as source query. The Chinese queries are translated into English to retrieve English captions of images. We adopt the translation approach we used in ImageCLEF 2003 to translate Chinese queries (Lin, Yang and Chen, 2003). First, the Chinese queries are segmented by a word recognition system, and tagged by a POS tagger. Name entities are then identified (Chen, *et al.*, 1998). For each Chinese query term, we find its translations by looking up a Chinese-English bilingual dictionary. The bilingual dictionary is integrated from four resources, including the LDC Chinese-English dictionary, Denisowski's CEDICT¹, BDC Chinese-English dictionary v2.2² and a dictionary used in query translation in MTIR project (Bian and Chen, 2000). The dictionary gathers 200,037 words, where a word may have more than one translation. If a query term has more than one translation, we use the first-two-highest-frequency method to select translations. The first two translations with the highest frequency of occurrence in the English image captions are considered as the target language query terms.

For the named entities that are not included in the dictionary, we use similarity-based backward transliteration scheme to translate them. First, we adopt the transformation rules (Chen, H.H., Yang, C. and Lin, 2003) to identify the name part and keyword part of a name. The keyword parts are general nouns, e.g., “湖” (lake), “河” (river) and “橋” (bridge), and translated by dictionary looking up as described above. The name parts are transliterations of foreign names, and are transliterated into English in the following way.

- (1) The personal names and the location names in the English image captions are extracted. We collect a list of English names that contains 50,979 personal names and 19,340 location names. If a term in the

¹ The dictionary is available at <http://www.mandarintools.com/cedict.html>

² The BDC dictionary is developed by the Behavior Design Corporation (<http://www.bdc.com.tw>)

captions can be found in the name list, it is extracted. Total 3,599 names are extracted from the image captions.

- (2) For each Chinese name, 300 candidates are selected from the 3,599 English names by using an IR-based candidate filter. The document set is the International Phonetic Alphabet (IPA) representations of the 3,599 English names. Each name is treated as one document. The query is the IPA representation of the Chinese name. The phonemes of the Chinese name are expanded with their most co-transliterated English phonemes. The co-transliterated Chinese-English phoneme pairs are trained from a Chinese-English personal name corpus, which has 51,114 pairs of Chinese transliterated names and the corresponding English original names. Mutual Information is adopted to measure the strength of co-transliteration of two phonemes. A Chinese phoneme x is expanded with the English phonemes that have positive MI values with x . The augmented phonemes are weighted by $MI(x, y)/\text{the number of augmented terms}$. After retrieving, top 300 English names are reported as candidates.
- (3) The similarities of the Chinese name and the 300 candidates are computed at phoneme level. First, the Chinese name and candidate names are transformed into IPA. For each candidate word, the score of optimal alignment, i.e., the alignment with the highest score, between its IPA string and the IPA string of the Chinese name is computed as their similarity score. Given two strings S_1 and S_2 , let Σ be the alphabet of S_1 and S_2 , $\Sigma' = \{\Sigma, ' _ '\}$, where $' _ '$ stands for space. Space could be inserted into S_1 and S_2 such that they are of equal length and denoted as S_1' and S_2' . S_1' and S_2' are aligned when every character in either string has a one-to-one mapping to a character or space in the other string. The similarity score of an alignment is measured by the following formula.

$$\text{Score} = \sum_{i=1}^l s(S_1'(i), S_2'(i)) \quad (2)$$

where $s(a, b)$ is the similarity score between the character a and b in Σ' ,
 $S'(i)$ is the i^{th} character in the string S' , and
 l is the length of S_1' and S_2' .

The similarity score $s(a, b)$ is automatically learned from a bilingual name corpus (Lin and Chen, 2002). After the similarities are computed, the top 6 candidates with the highest similarities are considered as the translations of the Chinese name.

5. Combining Textual and Visual Information

As described in Section 1, there are two types of approaches, i.e., content-based and text-based approaches, to retrieve images. Content-based approaches use low-level features to represent images, while text-based approaches use collateral texts to describe images. Given an image collection, we can build two kinds of index for image retrieval. One is textual index of captions; the other one is visual index of images. In our experiments, we use blobs as the visual representation of images. Images are segmented into regions at first, then the regions of all images are clustered. Each cluster is assigned a unique label. Each image is represented by the blobs that its regions belong to. We can treat blobs as a language in which each blob token is a word. In this way, we can use text retrieval system to index and retrieve images using blobs language. In the experiments, both textual index and visual index are built by Okapi IR system. The weighting function is BM25 (Robertson, Walker, and Beaulieu, 1998).

Given a textual query, we can retrieve images using textual index. In addition to textual information, we can generate visual representation of textual query to retrieve images using visual index. The retrieval results of text-based and content-based approaches are merged to generate final result. For each image, the similarity scores of textual and visual retrieval are normalized and combined using linear combination. In ImageCLEF topic set, each topic also contains an example image. The example image can be used as visual query to retrieve images. The example image is represented as blobs, then is submitted to IR system to retrieve images using visual index. The retrieval result can be combined with the results of textual query and generated visual query using liner combination.

6. Experiment Results

In the experiments, both textual index and visual index were built by Okapi IR system. For textual index, the caption text, <HEADLINE> and <CATEGORIES> sections of English captions were used for indexing. For visual index, the blob tokens of each image were indexed. The weighting function is BM25. Chinese queries were used as source queries. Query translation was adopted to unify the languages used in queries and captions. Generated visual queries were generated from Chinese queries. In order to learn the correlations between Chinese words and blob tokens, image captions were translated into Chinese by SYSTRAN system.

We submitted four Chinese-English cross-lingual runs and one English monolingual run in CLEF 2004 image track. In English monolingual run, only textual queries were used. In the four cross-lingual runs, using example image or not using example image, and using generated visual query or not using generated visual query will be compared. The details of the cross-lingual runs are described as follows.

1. NTU-adhoc-CE-T-W

This run used textual queries only to retrieve images. Translation method described in Section 4 was used to translate Chinese queries into English to retrieve images using textual index.

2. NTU-adhoc-CE-T-WI

This run used both textual query and generated visual query. We used nouns, verbs, and adjectives in textual query to generate blobs as generated visual query. Named entities were not used to generate blobs. The top 30 blobs with MI value exceed a threshold, i.e., 0.01, were selected. Textual query used textual index, and generated visual query used visual index to retrieve images. For each image, the similarity scores of textual and visual retrieval were normalized and linear combined using weights 0.9 and 0.1 for textual and visual runs respectively. The top 1000 images with highest combined scores were taken as final results.

3. NTU-adhoc-CE-T-WE

This run used textual query and example image. Example image was represented as blobs. Textual query used textual index and example image used visual index to retrieve images. For each image, the similarity scores of textual and visual retrieval were normalized and linear combined using weights 0.7 and 0.3 for textual and example image runs respectively.

4. NTU-adhoc-CE-T-WEI

This run used textual query, generated visual query, and example image. Each topic had three retrieval runs. For each image, the similarity scores of three runs were normalized and linear combined using weights 0.7, 0.2, and 0.1 for textual query, example image, and generated visual query runs, respectively.

The performances of official runs are shown in Table 1. We found that we made a mistake when building textual index. Long captions were truncated, thus some words were not indexed. After fixing the error, we redid the experiments. The performances of unofficial runs are shown in Table 2. From Table 2, the performance of monolingual retrieval is good. The average precision of monolingual run is 0.6304. The cross-lingual runs also have good performances which are the top 4 Chinese-English runs. When using textual query only, the average precision of run NTU-CE-T-W-new is 0.4395, which is 69.72% of monolingual retrieval. Comparing to the results using ImageCLEF 2003 test set, the performance of this year is better. In ImageCLEF 2003 test set, the performance of Chinese-English cross-lingual textual run is 55.56% of English monolingual run when using intersection strict relevance set. One of the reasons is that several named entities are not translated into Chinese in Chinese query set of ImageCLEF 2004. These English names don't need to be translated when translating queries, thus there is no translation error. There are six topics, i.e., Topic 1, 3, 4, 11, 12, and 14, containing original English named entities. Total three of these topics, i.e., Topic 1, 12 and 14, have an average precision higher than 40%. The average precisions of each query are shown in Figure 2.

Table 1. Results of official runs

Run	Merging Weight			Average Precision
	Textual Query	Example Image	Generated Visual Query	
NTU-adhoc-CE-T-W	1.0	-	-	0.3977
NTU-adhoc-CE-T-WI	0.9	-	0.1	0.3969
NTU-adhoc-CE-T-WE	0.7	0.3	-	0.4171
NTU-adhoc-CE-T-WEI	0.7	0.2	0.1	0.4124
NTU-adhoc-EE-T-W				0.5463

Table 2. Performances of unofficial runs

Run	Merging Weight			Average Precision
	Textual Query	Example Image	Generated Visual Query	
NTU-CE-T-W-new	1.0	-	-	0.4395
NTU-CE-T-WI-new	0.9	-	0.1	0.4409
NTU-CE-T-WE-new	0.7	0.3	-	0.4589
NTU-CE-T-WEI-new	0.7	0.2	0.1	0.4545
NTU-EE-T-W-new				0.6304

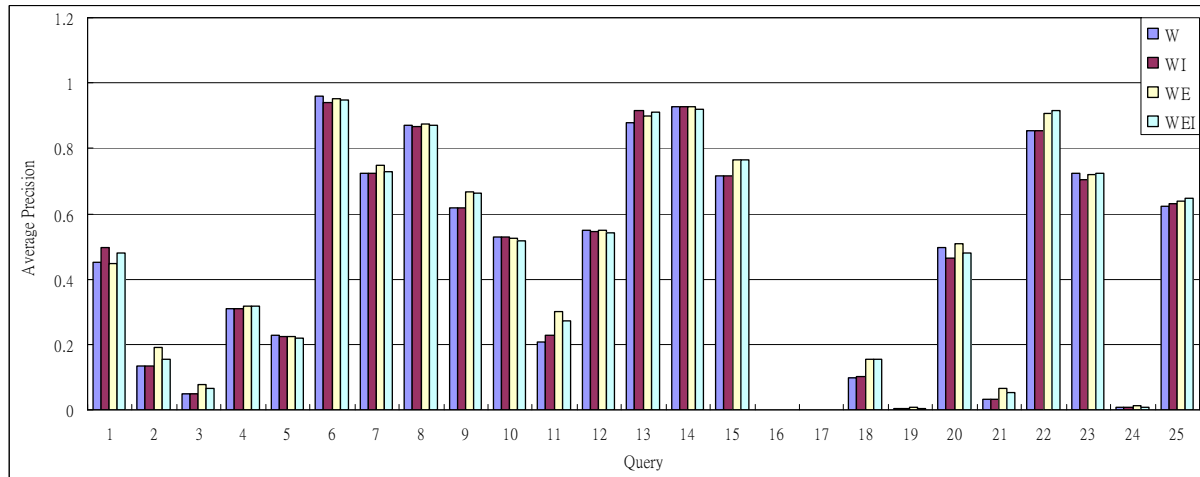


Figure 2. Average precisions of each topic in unofficial cross-lingual runs

Combining textual query and example image, average precision is increased to 0.4589. When using example image only to retrieve images, the average precision is 0.0523 which is contributed mostly by example image itself. In the result list of example image run, the top one entry is the example image itself and is relevant to the topic except Topic 17 (the example image of Topic 17 is not in the pisec-total relevant set of Topic 17). This makes example image having high score after combining the results of textual query and example image runs. In runs NTU-CE-T-WI-new and NTU-CE-T-WEI-new, the help of generated visual query is not clear. When using generated visual query only to retrieve images, the average precision is only 0.0103. The performance is average across 24 topics, since Topic 10 doesn't generate any blob. The poor performance of generated visual query run is helpless to increase final retrieval performance. Although the over all performance is not good, there are 8 topics gaining better performances after combining textual query and generated visual query runs. These topics have better performances than the others in generated visual query run.

The performance of generated visual query run doesn't meet our expectation. There are several factors that affect the performance of visual query. First, the performance of image segmentation is not good enough. The objects in an image can't be segmented perfectly. Furthermore, the majority of images in the St. Andrews image collection are in black and white, this makes image segmentation more difficult. Second, the performance of clustering affects the performance of blobs-based approach. If image regions that are not similar enough are clustered into the same cluster, this cluster (blob) may have several different meanings. This is analogous to polysemy problem. The third factor is the quality of training data. The St. Andrews image collection has only English captions. In order to learn the correlations between Chinese words and blob tokens, image captions were translated into Chinese by SYSTRAN system. However, there are many translation errors that affect the correctness of learned correlations. We conducted a monolingual experiment that using English captions to learn the correlations between text and images. Visual queries were generated from English queries. The results of English textual run and English generated visual query run were combined. The average precision is increased from 0.6304 to 0.6561. Another problem is that we used all words in captions to learn correlations. Many words, e.g. stopwords, date expressions, and names of photographers, are not relative to the content of images. These words should be excluded in training stage. The fourth problem is which word in a query should be used to generate visual query. In the experiments, nouns, verbs, and adjectives in textual query were used to generate visual query. While not all of these words are relative to the content of images or discriminative. For example, “照片” (picture) is not relative to the content of images, neither discriminative. Thus, “照片” (picture) should not be used to generate visual query. We conducted an experiment that manually selected query terms to generate visual query. There are 7 topics don't generate any blob. The average precision across 18 topics is 0.0146. In some topics, the retrieved images are not relevant to the topics, while they are relevant to the query terms that are used to generate visual query. Take Topic 13, i.e., 1939 年聖安德魯斯高爾夫球公開賽 (The Open Championship golf tournament, St. Andrews 1939), as an example, “高爾夫球” (golf) and “公開賽” (Open Championship) are chosen to generate visual query. The top 10 images



Figure 3. The top 10 images of Topic 13 in generated visual query run using manually selected terms

shown in Figure 3 are all about the Open Championship golf tournament, but are not the one held in 1939. It shows that using visual information only is not enough, integrating textual information is needed. After the result of manually selecting run merging with textual query run, the performance is slightly increased to 0.4427.

7. Conclusion

In this paper, we explore the help of visual features to cross-language image retrieval. We propose an approach that transforms textual queries into visual representations. The relationships between text and images are modeled. We use blobs as the visual representation of image. The textual description and visual representation of an image is treated as an aligned sentence. The correlations between words and blobs are learned from the aligned sentences. Visual queries are generated from textual queries using the relationships. In addition to textual index, a visual index is built for retrieving images by visual query. The retrieval results using textual and visual queries are combined to generate the final ranked list.

We conducted English monolingual and Chinese-English cross-language retrieval experiments. The performances are quite good. The average precision of English monolingual textual run is 0.6304. The performance of cross-lingual retrieval is about 70% of monolingual retrieval. Combining textual query run with generated visual query run, the performance is increased in English monolingual experiment. However, generated visual query has little impact in cross-lingual experiments. One of the reasons is that using MT system to translate English captions into Chinese has many translation errors that affect the correctness of learned correlations. Although the help of generated visual query is limit, using generated visual query could retrieve images relevant to the query terms that the visual query is generated from. Without the help of other query terms, the retrieved images are not relevant to the topics. How to select appropriate terms to generate visual query and how to integrate textual and visual information effectively will be further investigated.

Reference

1. Bian, G.W. and Chen, H.H. (2000). Cross Language Information Access to Multilingual Collections on the Internet. *Journal of American Society for Information Science*, 51(3), 281-296.
2. Carson, C., Belongie, S., Greenspan, H. and Malik, J. (2002). Blobworld: Image Segmentation Using Expectation-Maximization and Its Application to Image Querying. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(8), 1026-1038.
3. Chen, H.H., Ding, Y.W, Tsai, S.C. and Bian, G.W. (1998). Description of the NTU System Used for MET2. In *Proceedings of Seventh Message Understanding Conference*.
4. Chen, H.H., Yang, C. and Lin, Y. (2003). Learning Formulation and Transformation Rules for Multilingual Named Entities. In *Proceedings of ACL 2003 Workshop on Multilingual and Mixed-language Named Entity Recognition: Combining Statistical and Symbolic Models* (pp. 1-8).
5. Duygulu, P., Barnard, K., Freitas, N. and Forsyth, D. (2002). Object Recognition as Machine Translation: Learning a lexicon for a fixed image vocabulary. In *proceedings of Seventh European Conference on Computer Vision*, Vol. 4 (pp. 97-112).
6. Jeon, J., Lavrenko, V. and Manmatha, R. (2003). 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 (SIGIR 2003)* (pp. 119-126).
7. Lavrenko, V., Manmatha, R. and Jeon, J. (2003). A Model for Learning the Semantics of Pictures. In *Proceedings of the Seventeenth Annual Conference on Neural Information Processing Systems*.
8. Lin, W.H. and Chen, H.H. (2002). Backward Machine Transliteration by Learning Phonetic Similarity. In *Proceedings of Sixth Conference on Natural Language Learning* (pp. 139-145).

9. Lin, W.C., Lin, M.S. and Chen, H.H. (2004). Cross-language Image Retrieval via Spoken Query. In *Proceedings of RIAO 2004: Coupling Approaches, Coupling Media and Coupling Languages for Information Retrieval* (pp. 524-536).
10. Lin, W.C., Yang, C., and Chen, H.H. (2003). Foreign Name Backward Transliteration in Chinese-English Cross-Language Image Retrieval. In *Working Notes of CLEF 2003*.
11. Mori, Y., Takahashi, H. and Oka, R. (1999). 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*.
12. Robertson, S.E., Walker, S. and Beaulieu, M. (1998). Okapi at TREC-7: automatic ad hoc, filtering, VLC and interactive. In *Proceedings of the Seventh Text REtrieval Conference (TREC-7)* (pp. 253-264).
13. Shi, J. and Malik, J. (2000). Normalized Cuts and Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8), 888-905.
14. The Lowlands Team (2002). Lazy Users and Automatic Video Retrieval Tools in (the) Lowlands. In *Proceedings of the Tenth Text REtrieval Conference (TREC 2001)* (pp. 159-168).
15. Westerveld, T. (2000). Image Retrieval: Content versus Context. In *Proceedings of RIAO 2000*, Vol. 1 (pp. 276-284).
16. Westerveld, T. (2002). Probabilistic Multimedia Retrieval. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2002)* (pp. 437-438).