Available online at www.sciencedirect.com
ScienceDirect
Information Processing and Management 43 (2007) 488–502
INFORMATION PROCESSING & MANAGEMENT
www.elsevier.com/locate/infoproman

Integrating textual and visual information for cross-language image retrieval: A trans-media dictionary approach

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

Department of Computer Science and Information Engineering, National Taiwan University, No. 1, Sec. 4, Roosevelt Road, Taipei 106, Taiwan

Received 17 May 2006; accepted 25 July 2006
Available online 11 October 2006

Abstract

This paper explores the integration of textual and visual information for cross-language image retrieval. An approach which automatically transforms textual queries into visual representations is proposed. First, we mine the relationships between text and images and employ the mined relationships to construct visual queries from textual ones. Then, the retrieval results of textual and visual queries are combined. To evaluate the proposed approach, we conduct English monolingual and Chinese–English cross-language retrieval experiments. The selection of suitable textual query terms to construct visual queries is the major issue. Experimental results show that the proposed approach improves retrieval performance, and use of nouns is appropriate to generate visual queries.
© 2006 Elsevier Ltd. All rights reserved.

Keywords: Cross-language image retrieval; ImageCLEF; Language translation; Medium transformation; Trans-media dictionary

1. Introduction

Today multimedia data grows explosively. The Internet, for example, contains millions of images, videos, and music. Finding the requested information from large amounts of multimedia data is challenging. Two types of approaches, i.e., content-based and text-based, are usually adopted in image retrieval (Goodrum, 2000). Content-based image retrieval (CBIR) uses low-level visual features such as color, texture, and shape to represent images. Users can employ example images as queries, or directly specify the weight of low-level visual features to retrieve images. Images that are visually similar to an example image or contain the specified visual features are returned.

* Corresponding author. Tel.: +886 2 33664888x311; fax: +886 2 23628167.
E-mail addresses: denislin@nlg.csie.ntu.edu.tw (W.-C. Lin), ycchang@nlg.csie.ntu.edu.tw (Y.-C. Chang), hhchen@csie.ntu.edu.tw (H.-H. Chen).

0306-4573/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2006.07.015

In text-based approaches, text is used to describe images and formulate queries. Because images and image representations are in different types of media, media transformation is required. Images are transformed into text, and a text retrieval system is used to index and retrieve images. Textual features can also be derived from the text accompanying an image such as a caption or the surrounding text. Text-based approach encounters the following problems:

(1) Image captions are usually short. The short annotation cannot represent the image content completely.
(2) Image captions are not always available. Manually assigning captions to images is time consuming and costly.
(3) Some visual properties cannot be described directly in captions. For example, the styles of images, e.g., warm, cold, dark, sharp, or blurry, are usually not specified in captions.
(4) Users' queries may have different levels of semantics. Users may search for images at a higher semantic level or at a primitive level.

Since images are produced by people familiar with their own language, they can be annotated in different languages. In this way, text-based image retrieval has a multilingual nature. In addition, images are understandable by different language users. They can resolve the major argument in cross-language information retrieval, i.e., users that are not familiar with the target language cannot understand the retrieved documents. In such a situation, cross-language image retrieval has attracted researchers' attentions recently and is organized as one of evaluation tasks in the Cross-Language Evaluation Forum (CLEF) (Clough, Sanderson, & Müller, 2005). In addition to media transformation, language translation is also necessary to unify the language usages in queries and documents in cross-language image retrieval.

Textual and low-level visual features have different semantic levels. Textual feature is highly semantic, while low-level visual feature is less semantic and is more emotive. These two types of features are complementary and provide different aspects of information about images. In this paper, we explore the integration of textual and visual information in cross-language image retrieval. An approach that automatically transforms textual queries into visual representations is proposed. The generated visual representation is treated as a visual query to retrieve images. The retrieved results using textual and visual queries are combined to generate the final result.

The rest of this paper is organized as follows. Section 2 introduces the proposed model. The integration of textual and visual information is illustrated. Section 3 models the relationships between text and images. How to generate visual representation of a textual query is introduced. Section 4 specifies the experimental materials. Section 5 shows the experiment designs. Here, the selection of suitable textual query terms to construct visual queries is the major issue. In addition, three types of experiments are evaluated, including monolingual image retrieval, cross-language image retrieval and ideal visual queries. Finally, we conclude our work in Section 6.

2. Integrating textual and visual information

Several hybrid approaches that integrate visual and textual information have been proposed. A simple approach conducts text- and content-based retrieval separately and merges the retrieval results of the two runs (Besançon, Hède, Moellic, & Fluhr, 2005; Jones et al., 2005; Lin, Chang, & Chen, 2005). In contrast to this parallel approach, a pipeline approach employs textual or visual information to perform initial retrieval, and then uses the other feature to filter out the irrelevant images (Lowlands Team, 2001). In the above two approaches, users have to issue two types of queries, i.e., textual and visual. In these approaches, sometimes it is not intuitive to find an example image or to specify low-level visual features.

We take another approach as shown in Fig. 1 with our cross-language image retrieval system. This system automatically transforms textual queries into visual representations. First, the relationships between text and images are mined from a set of images annotated with text descriptions. A trans-media dictionary which is similar to a bilingual dictionary is set up from the training collections. When a user issues a textual query, the system automatically transforms the textual query into a visual one using the trans-media dictionary. The generated visual representation is treated as a visual query and is used to retrieve images. In this way, we have both textual and visual queries.

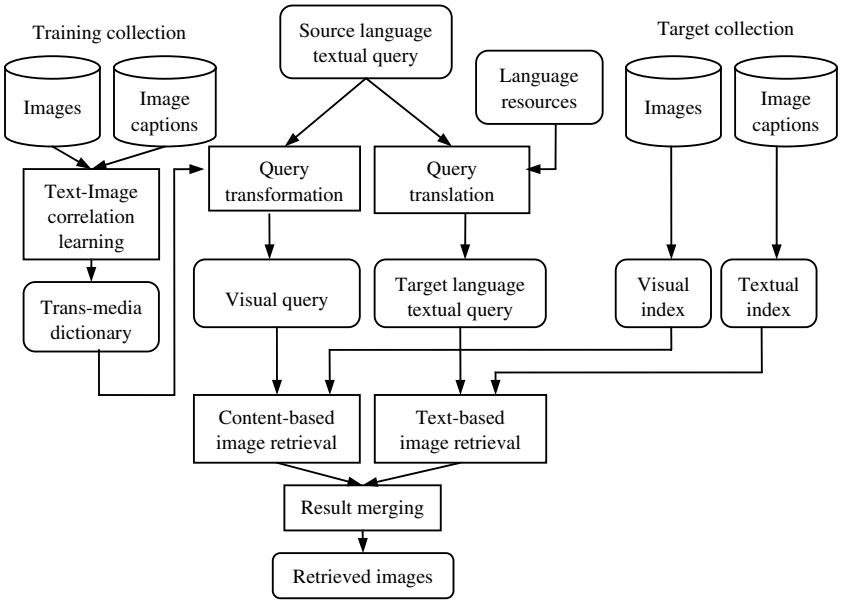

Fig. 1. Flow of cross-language image retrieval.

Given a collection of images and their captions, two kinds of indices are generated for image retrieval. One is textual index of image descriptions, and the other one is visual index of images. A textual query is used to retrieve image descriptions using the textual index. A visual query transformed from the textual query retrieves images using the visual index. The retrieval results of textual and generated visual queries are merged together.

The proposed approach can be applied to monolingual and cross-language image retrieval. In cross-language information retrieval, translation ambiguity and target polysemy problems (Chen, Bian, & Lin, 1999) have to be tackled in the translation process. If a word is not translated correctly, we cannot capture the correct meaning of the word. If the translation is polysemous, the undesired documents that contain the translation with other senses could be reported even if the translation is correct. Visual queries could be helpful to reduce these problems.

3. Visual representation of text

Given a set of images along with text descriptions, we can learn the relationships between images and the original (or the translated) text. For an image, a term in the description may relate to a portion of an image. If we divide an image into several smaller parts, e.g., blocks or regions, we could link the terms to the corresponding parts. This is analogous to word alignment in a sentence aligned parallel corpus. Here the word alignment is replaced with the textual-term/visual-term alignment. If we treat the visual representation of an image as a language, the textual description and visual parts of an image are an aligned sentence. The correlations between the vocabularies of two languages can be learned from the aligned sentences. Given a picture of sunset, for example, we can link textual feature ''sunset'' to visual feature ''red circle''.

In automatic annotation, several approaches have been proposed to model the correlations between 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, Barnard, Freitas, and Forsyth (2002) used blobs to represent images. First, images are segmented into regions using a segmentation algorithm. All regions are clustered and each cluster is assigned a unique label (blob token). EM algorithm constructs 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

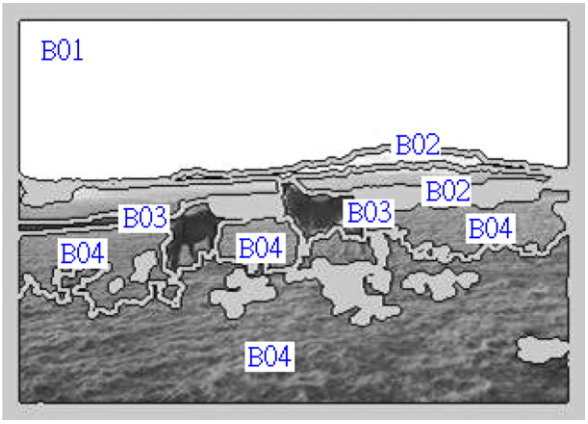

Fig. 2. Segmentation of a sample image.

proposed continuous-space relevance model (CRM) that learned the joint probability of words and regions, rather than blobs (Lavrenko, Manmatha, & Jeon, 2003).

This paper considers blobs as a visual representation of images, and adopts Blobworld (Carson, Belongie, Greenspan, & Malik, 2002) to segment an image into regions. Blobworld groups pixels in an image into regions which are coherent in low-level properties such as color and texture, and which roughly correspond to objects or part of objects. For each region, a set of features such as color, texture, shape, position, and size are extracted. The regions of all images are clustered by the K-means clustering algorithm. Each cluster is assigned a unique number, i.e., blob token, and each image is represented by the blob tokens.

Given the textual descriptions and blob tokens of images, we mine the correlation between textual and visual information. 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 follows:

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

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 co-occur in aligned image-text description pairs.

After the MIs between words and blobs are computed, we can generate related blobs for a given word w_i. The blobs whose MI values with w_i exceed a threshold are associated to w_i. The generated blobs can be regarded as the visual representation of w_i. In this way, a trans-media (word-blob) dictionary is established.

Fig. 2 shows an example image, which has nine regions. These regions cluster into four groups, each of which is assigned a blob visual term. Hence the nine regions can be represented by the following visual terms: B01, B02, B02, B03, B03, B04, B04, B04, and B04. The corresponding text description of the image is: "Mare and foal in field, slopes of Clatto Hill, Fife". After textual-term/visual-term alignment, the following possible pairs remain:

hill $\Longleftrightarrow$ B02, mare $\Longleftrightarrow$ B03, foal $\Longleftrightarrow$ B03, field $\Longleftrightarrow$ B04, slope $\Longleftrightarrow$ B04

4. Experimental materials

In the experiments, we adopt the 2004 and 2005 ImageCLEF test sets (Clough et al., 2005, in press). The image collection consists of 28,133 photographs from St. Andrews University Library's photographic

collection, which is one of the largest and most important collections of historic photography in Scotland. The majority of images (82%) are in black and white. All images are accompanied by a caption written in English by librarians. The information in a caption ranges from specific date, location, and photographer to a more general description of an image. Fig. 3 shows an example of an image and its caption in the St. Andrews image collection. The text descriptions are semi-structured and consist of several fields including document number, headline, record id, description text, category, and file names of images in a 368×234 large version and 120×76 thumbnail version.

The 2004 test set contains 25 topics and the 2005 test set 28 topics. Each topic consists of (1) a title (a short sentence or phrase describing the search request in a few words), and (2) a narrative (a description of what constitutes a relevant or non-relevant image for each request). In addition to the text description for each topic, one and two example images are provided for 2004 and 2005 topic sets, respectively.

In this paper, each topic is a query to retrieve images from the St. Andrews photographic collections. In our experiments, queries are in Chinese. Fig. 4 illustrates a topic in English and in Chinese.

The Appendix lists all the topics for references. Clough et al. (2005) partitioned the topics in the 2004 topic set into five categories, including queries modified by photographer or date (1–4), queries modified by location (5–11), queries related to specific events (12–15), queries related with known-items (16–18), and queries related to general topics (19–25). Queries in the 2005 topic set are more general and more visual than those in the 2004 topic set (Clough et al., in press). More than half of them not only consider the objects in the images, but also their spatial or event relationships. The followings show some examples:

(1) Query "aircraft on the ground" specifies airplanes positioned on the ground. Pictures of aircraft flying are not relevant.
(2) Query "people gathered at bandstand" specifies a group of people at a bandstand. Pictures of people just walking past a bandstand are not relevant.
(3) Query "dog in sitting position" specifies dogs in a sitting position. Pictures of dogs that do not sit are not relevant.
(4) Query "steam ship docked" specifies at least one steam ship docked or moored. Pictures of steam ships at places other than the docks are not relevant.

<DOC>
<DOCNO> stand03_1041/stand03_9914.txt </DOCNO>
<HEADLINE> Azay le Rideau. Bridge. </HEADLINE>
<TEXT>
<RECORD_ID> JEAS-.000032.-.000045 </RECORD_ID>
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
<CATEGORIES>
[towers - round], [towers - conical roofed], [France urban views], [France all views]
</CATEGORIES>
<SMALL_IMG>
stand03_1041/stand03_9914.jpg
</SMALL_IMG>
<LARGE_IMG>
stand03_1041/stand03_9914_big.jpg
</LARGE_IMG>
</TEXT>
/DOC

Fig. 3. An image and its description in St. Andrews image collection.

<top>
<num> Number: 2 </num>
<title> Photos of Rome taken in April 1908 </title>
<narr>
Any view of Rome including buildings and specific locations (e.g. the coliseum) taken in April 1908 is relevant. Pictures by any photographer are relevant, but taken at any other time are not relevant.
</narr>
</top>

<top>
<num> Number: 2 </num>
<title>
1908年 (nián) 四 (sì) 月 (yuè) 拍 (pāi) 攝 (shè) 的 (de) 羅 (luó) 馬 (mǎ) 照 (zhào) 片 (piān)
<title>

Fig. 4. A topic in English and in Chinese.

(5) Query "fishermen in boat" specifies at least two people fishing in a boat or ship. Pictures of just one fisherman or fishermen not in boats (e.g., fishing from shore) are not relevant.

5. Experiments

5.1. Monolingual image retrieval

In the experiments, we adopt title field in topics to retrieve images. Okapi IR system (Robertson, Walker, & Beaulieu, 1998) is used to build both the textual and visual indices. For the textual index, the caption texts, i.e., the English captions, are used for indexing. All words are stemmed and stopwords are removed. For visual index, the blob tokens of each image are indexed. The weighting function used is BM25.

We evaluate our approach in monolingual image retrieval on the 2004 topic set at first. The correlations between text and images are learned from St. Andrews image collection. The title field of a topic is used as a query to retrieve images. For each textual query, a visual query is generated from the query terms according to the mined relationships. The first issue is which query terms are adopted to generate the visual query. Intuitively, we can generate visual representation for each query term. However, not all query terms are related to the visual content of images. Here, we employ part-of-speech (POS) to select suitable query terms to generate visual representations. Brill tagger (Brill, 1995) is used to tag English topics. Different types of POSes are explored to tell which types of query terms are useful. Nouns only (without named entities), nouns with named entities, verbs only, adjectives only, or nouns, verbs, along with adjectives are experimented.

For each selected query term, the top n blobs of MI values exceed a threshold t are regarded as its visual representation. The values of parameter n from 10 to 40 and t from 0.1 to 0.4 are experimented. The blobs corresponding to the selected query terms form a visual query. It is used to retrieve images using visual index. The results of textual and generated visual queries are merged into the final result. For each image, the similarity scores of textual and visual retrieval are normalized and linearly combined using weights 0.9 and 0.1 for the textual and visual runs, respectively. The top 1000 images of the highest combined scores are reported.

The performance of the proposed approach is shown in Fig. 5. Fig. 5(a)–(d) demonstrate 10, 20, 30, and 40 blobs are selected for each query term, respectively. Mean average precision (MAP) measures the retrieval performances. The approach of using nouns only, higher threshold and more blobs has better performance than that of using verbs and adjectives. The performances of using verbs or adjectives only in different setting of n and t are similar. This is because there are only a few verbs and adjectives in the topic set, e.g., only four adjectives in four topics and nine verbs in eight topics, and the MI values of blobs with verbs and adjectives tend to

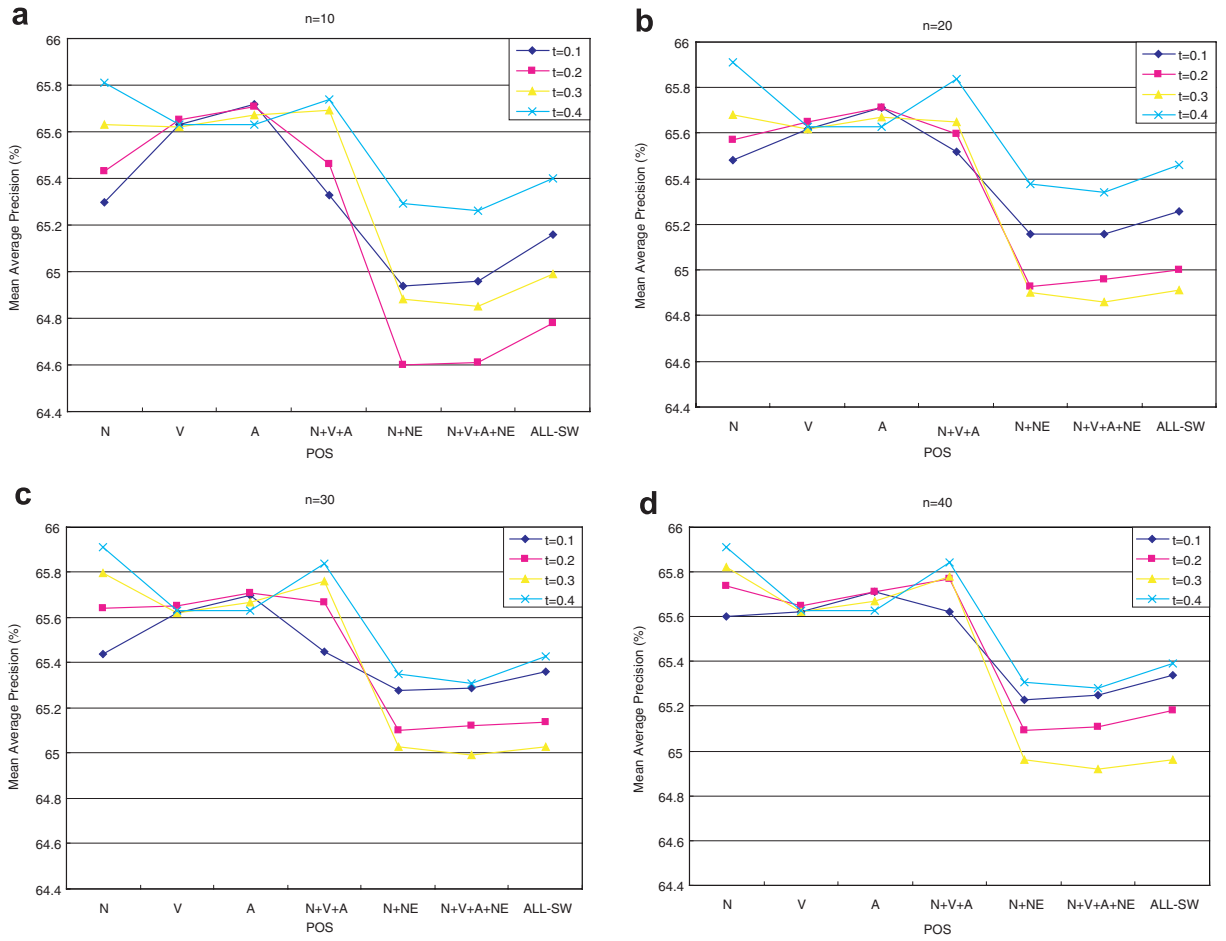

Fig. 5. Performance of monolingual image retrieval (a) blob number: 10, (b) blob number: 20, (c) blob number: 30, and (d) blob number: 40.

be low. When using nouns, verbs and adjectives, the performance is slightly worse than using nouns only. The performance is dropped when named entities are added. It is even worse than using all words with stopword removal (ALL-SW).

The best performance is 0.6591 when using nouns only, $n = 20$, and $t = 0.4$. Comparing with the approach of using textual query only, the MAP is increased. The performances of textual query and generated visual query are shown in Table 1. The results show that the proposed approach increases retrieval performance. Although the generated visual queries are not so good enough, the integration of them is useful to improve retrieval performance. Several factors may affect the visual query construction. First, the image segmentation has a large effect. Because the majority of images in the St. Andrews image collection are in black and white, that characteristic makes image segmentation more difficult. Second, clustering affects the performance of the blobs-based

Table 1
Integrating textual and generated visual queries in monolingual cases

Query type	MAP (2004/2005 topic sets)
Textual query	0.6304/0.3952
Generated visual query (14 topics/12 topics)	0.0036/0.0215
Textual query + generated visual query (N, $n = 20$, $t = 0.4$)	0.6591/0.3945

approach. If image regions that are not similar enough are clustered together, the cluster (blob) may have several different meanings. That is analogous to the polysemy problem on word level (Chen et al., 1999).

Table 1 also shows the performance of the same approach on the 2005 topic set. The MAP of using textual queries is decreased to 0.3952. It confirms that the 2005 topic set containing more general and visual queries is more challenging than the 2004 topic set. The MAP of the generated visual queries is 0.0215, which is larger than that of the 2004 topic set. However, the overall performance of integrating textual and generated visual queries is a little worse than that of textual queries only. This might be due to that the spatial and action relationships among objects in the images are crucial in the 2005 topic set and our approach only lists the possible objects, but does not consider their relationships.

5.2. Cross-language image retrieval

In the experiments of cross-language image retrieval, Chinese queries are used as source queries and translated into English to retrieve English captions of images. First, we deal with the 2004 topic set. The Chinese queries are segmented by a word recognition system and tagged by a POS tagger. Named entities are identified by a Chinese NER tool (Chen, Ding, Tsai, & Bian, 1998). For each Chinese query term, we find its translation equivalents using a Chinese–English bilingual dictionary. If a query term has more than one translation, the first two translations with the highest frequency of occurrences in the English image captions are considered as the target language query terms.

For those named entities that are not included in the dictionary, a similarity-based backward transliteration scheme (Lin & Chen, 2002) is adopted. First, transformation rules (Chen, Lin, Yang, & Lin, 2006) tell out the name and the keyword parts of a named entity. The keyword parts are general nouns, and are translated by dictionary lookup as described above. The name parts, which are transliterations of foreign names, are transliterated into English using similarity-based backward transliteration. Total 3599 English names from the image captions are extracted. Given a transliterated name, 300 candidate names are selected from the 3599 names using an IR-based candidate filter (Lin et al., 2005). We transform the transliterated name and candidate names to International Phonetic Alphabet (IPA), and compute the similarities between IPA representations of the transliterated name and candidate names. The top 6 candidate names with the highest similarity are chosen as the original names.

Visual queries are generated from Chinese queries. In order to learn the correlations between Chinese words and blob tokens, English image captions are translated into Chinese by SYSTRAN machine translation system. Similarly, POS selects query terms for visual query construction. Fig. 6(a)–(d) shows the values of parameter n from 10 to 40 and t from 0.01 to 0.04 are experimented. The performances of term selection strategies are similar to that of monolingual image retrieval. Using nouns only to generate visual query has better performance than using verbs and adjectives only. When $n \geqslant 30$, using nouns, verbs and adjectives together performs better than using nouns only. The best performance is 0.4441 when using nouns, verbs and adjectives, $n = 30$, and $t = 0.02$. The performances of textual query and generated visual query are shown in Table 2. In cross-language experiment, the improvement of retrieval performance is not as well as monolingual experiment. One of the reasons is that the quality of training data is not good. We use a famous machine translation system to translate image captions. However, there are still many translation errors that affect the correctness of learned correlations. Table 2 also lists the performance on the 2005 topic set. The MAP of the 2005 test set is worse than that of the 2004 test set. That is consistent to the monolingual case due to that the former topic set covers more general and visual information needs than the latter one (Clough et al., in press).

The improvement is verified by Wilcoxon signed-rank test, which considers the sign and the magnitude of the rank of the difference between pairs of measurements. In the 2004 test set, we can generate visual queries for 18 topics. Table 3 shows that 14 topics have nonzero performance differences after integrating visual queries. The observed value of z is computed as follows, where W is the sum of signed ranks, and N is sample size.

$$z = \frac{W - 0.5}{\sigma_W} \tag{2}$$

$$\sigma_{\mathrm{w}} = \sqrt{\frac{N(N+1)(2N+1)}{6}} \tag{3}$$

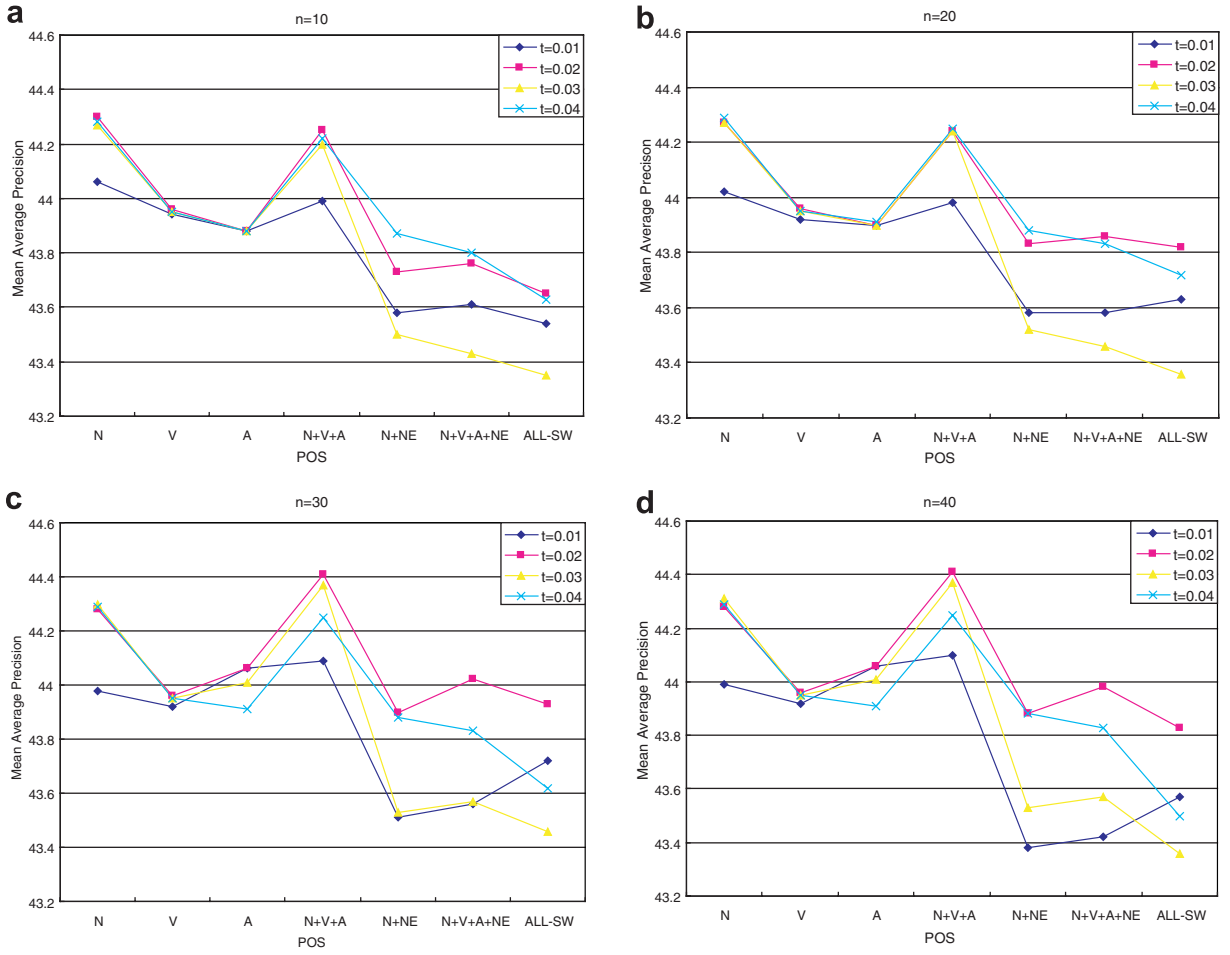

Fig. 6. Performance of cross-language image retrieval (a) blob number: 10, (b) blob number: 20, (c) blob number: 30, and (d) blob number: 40.

Table 2
Integrating textual and generated visual queries in cross-language cases

Query type	MAP (2004/2005 sets)
Textual query	0.4395/0.2399
Generated visual query (18 topics/27 topics)	0.0110/0.0133
Textual query + generated visual query ($N + V + A$, $n = 30$, $t = 0.02$)	0.4441/0.2401

For the example, with $N = 14$, $W = 61$, and $\sigma_W = 31.86$, the result 1.90, which passes the critical value 1.645, is significant beyond level 0.05.

The performance of generated visual query is not as good as our expectation. One of the reasons is that we use only a part of query terms to generate visual query, thus some information is lost. 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 query 13 of the 2004 topic set, 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. Terms "聖" (St), "高爾夫球" (golf) and "公開賽" (Open Championship) are tagged as nouns, thus they are selected to generate visual query. Of the top 10 returned images shown in Fig. 7, nine images are about the Open Championship golf tournament,

Table 3
Wilcoxon signed-rank test

Topic id	MAP of textual query	MAP of textual and visual query	Signed rank
1	0.4532	0.5062	+14
3	0.0508	0.0491	−6
5	0.2295	0.2309	+5
6	0.9582	0.9582	–
8	0.8689	0.8683	−3
11	0.2055	0.2323	+12
12	0.5477	0.5477	–
13	0.8776	0.9134	+13
14	0.9263	0.9263	–
16	0	0	–
17	0.0003	0.0004	+1
18	0.0961	0.1038	+9
19	0.0024	0.0031	+4
20	0.4946	0.5025	+10
21	0.0330	0.0350	+7
23	0.7249	0.7024	−11
24	0.0082	0.0080	−2
25	0.6217	0.6258	+8

Fig. 7. Top 10 images returned by generated visual Query 13.

but they are not the one held in 1939. The date/time expression is hard to be captured by visual features. It shows that using visual information only is not enough, integrating textual information is needed.

Since the performance of generated visual query depends on image segmentations, blob clustering, and so on, we create an ideal query from relevant images to test if a visual query can help increase the performance of image retrieval. A useful visual query will exist if the relevant images for a query share some image features. The common image features can help us retrieve the relevant images well. We use χ^2 score to select blobs from relevant images of the 2004 topic set. For each query we generate 10 blobs whose χ^2 scores are larger than 7.88 ($v = 1$, $p = 0.005$). The selected blobs form a visual query to retrieve images. The retrieval result is combined with that of a textual query. The performances are shown in Table 4. The results show that a good visual query can improve performance of image retrieval.

5.3. More experiments and discussions

Fig. 8 shows the different combinations of media transformation (denoted by thick lines) and language translation (denoted by thin lines) using a trans-media dictionary approach. The first column and the last

Table 4
Performances of ideal visual queries

Query type	MAP
Ideal visual query	0.1478
English query + ideal visual query	0.7082
Chinese query + ideal visual query	0.4780

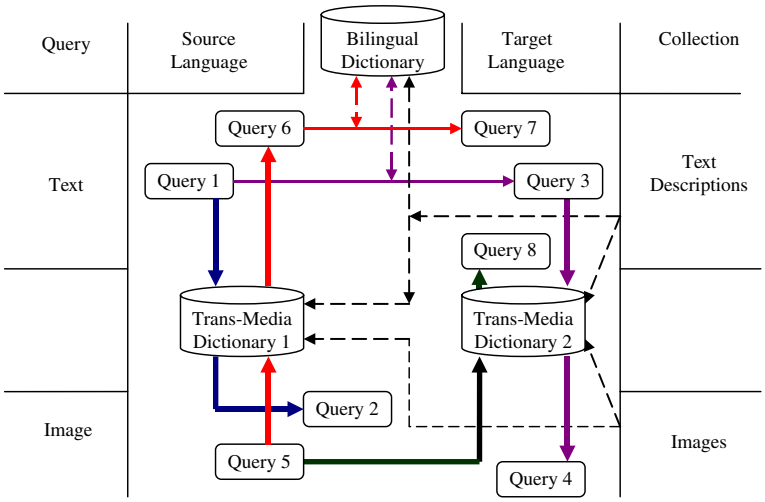

Fig. 8. A generalized transmedia dictionary approach.

column denote the query and the data collection, respectively. In addition, the first row and the last row denote text and image, respectively. Query 1 is a textual query in source language, and Query 5 is a visual query. The text descriptions of images are in target language. There are two alternatives (denoted by dotted lines) to construct a trans-media dictionary. The first one is translating the text description through a bilingual dictionary, which is similar to document translation in traditional CLIR, and then mining the relationship between textual terms and visual terms. The second one is: mining the relationship between textual terms and visual terms directly without translation. Trans-Media Dictionaries 1 and 2 in Fig. 8 belong to the first and the second alternatives, respectively.

By looking up Trans-Media Dictionary 1, Query 1, a textual query, is transformed into Query 2, a visual query. In the counter part, Query 5, a visual query, is transformed into Query 6, a textual query in source language, and then Query 6 is translated into Query 7, a textual query in target language. Before looking up Trans-Media Dictionary 2, Query 1 has to be translated into Query 3, a textual query in target language. Then Query 3 is transformed into Query 4, a visual query. Consulting Trans-Media Dictionary 2, Query 5 can also be transformed into Query 8, a textual query in target language. The combinations are summarized as follows.

(p1) Query 1 ⇒ Translation ⇒ Query 3
(p2) Query 1 ⇒ Transformation ⇒ Query 2
(p3) Query 1 ⇒ Translation ⇒ Query 3 ⇒ Transformation ⇒ Query 4
(p4) Query 5 ⇒ Transformation ⇒ Query 6 ⇒ Translation ⇒ Query 7
(p5) Query 5 ⇒ Transformation ⇒ Query 8
(p6) Query 5

The following explores these combinations to study the effects of textual and visual features in information retrieval. Because the experiments in Section 5.2 depict that selecting nouns, verbs and adjectives for

Table 5
Performance of different combinations

Paths	Thresholds							
	The 2004 topic set				The 2005 topic set			
	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4
p1 (Query 3)	0.4395				0.2399			
p2 (Query 2)	0.0126	0.0094	0.0074	0.0095	0.0095	0.0139	0.0158	0.0166
p3 (Query 4)	0.0149	0.0072	0.0068	0.0060	0.0027	0.0025	0.0030	0.0035
p4 (Query 7)	0.0090	0.0093	0.0088	0.0078	0.0295	0.0295	0.0259	0.0261
p5 (Query 8)	0.0463	0.0356	0.0303	0.0311	0.0406	0.0280	0.0278	0.0293
p6 (Query 5)	0.0523				0.0633			

transformation and adopting more blobs (e.g., 40) has the better performance, the following experiments only consider those cases. Table 5 shows the experimental results. The MAPs of Query 3 and Query 5 on the (2004, 2005) topic sets are (0.4395, 0.2399) and (0.0523, 0.0633), respectively. Query 2 and Query 4 are in the same medium, i.e., image. However, Query 4 is generated through both query translation and medium transformation, thus more noise is introduced and the MAP of Query 4 is worse than that of Query 2. Similarly, Query 8 and Query 7 are in the same medium, i.e., text. Query 7 is generated from Query 5 by two operations, so that the MAP of Query 7 is worse than that of Query 8. Compared Query 4 and Query 7, both resulting from query translation and medium transformation, the former is worse than the latter. In other words, medium transformation first and then language translation is better than language translation first and then medium transformation.

6. Conclusion

This paper explores the uses of both textual information and visual features for cross-language image retrieval. We conduct English monolingual and Chinese–English cross-language retrieval experiments to evaluate our approach. Experimental results show that combining retrieval results of textual and generated visual queries improves retrieval performance. The generated visual query has little impact in the cross-lingual experiments. One of the reasons is that using a machine translation system to translate English captions into Chinese introduces many translation errors that affect the correctness of learned correlations. We also construct an ideal visual query from relevant images. Using the ideal visual query increases retrieval performance about 12.3% in monolingual and 8.8% in cross-language image retrieval. The results show that a good visual query can improve performance of image retrieval.

We use POS to select query terms for constructing a visual query. Experiments show that nouns are appropriate to generate visual queries, while using named entities is useless. Nouns usually indicate the objects in images, which is the kernel of an image, thus it is reasonable to link nouns to the image regions which correspond to objects. Named entities, such as personal name, location name, and date, do not have strong relations with image regions, and cannot be represented well by visual representations. In this way, the visual representations of named entities introduce noise and decrease the retrieval performance. Similarly, verbs that indicate actions are rarely represented by visual features. Thus, verbs are not feasible for visual query generation. Some adjectives that are relative to visual features could be used to generate visual queries. For example, red is relative to color, a low-level visual feature. In the experiments, we use syntactic information to select query terms. Semantic information which may provide more clues for term selection is not used. We will investigate query term selection on the semantic level in the future.

Acknowledgement

Research of this paper was partially supported by National Science Council, Taiwan, under the contracts NSC94-2213-E-002-076 and NSC95-2752-E-001-001-PAE.

Appendix. Topics in the experiments

The following lists the 25 topics and 28 topics in 2004 and 2005 topic sets, respectively. The Chinese topics enclosed in parentheses are translated from English ones by human.

(a) **The 2004 topic set**

(1) Portrait pictures of church ministers by Thomas Rodger (Thomas Rodger 拍攝的牧師肖像).
(2) Photos of Rome taken in April 1908 (1908 年四月拍攝的羅馬照片).
(3) Views of St. Andrews cathedral by John Fairweather (John Fairweather 拍攝的聖安德魯斯教堂).
(4) Men in military uniform, George Middlemass Cowie (George Middlemass Cowie 拍攝的穿軍服的男人).
(5) Fishing vessels in Northern Ireland (北愛爾蘭的漁船).
(6) Views of scenery in British Columbia, Canada (加拿大英屬哥倫比亞區的風景).
(7) Exterior views of temples in Egypt (埃及廟宇的外部景觀).
(8) College or university buildings, Cambridge (劍橋學院或大學的建築).
(9) Pictures of English lighthouses (英格蘭燈塔照片).
(10) Busy street scenes in London (倫敦的繁忙街景).
(11) Composite postcard views of Bute, Scotland (蘇格蘭Bute地區風景的綜合明信片).
(12) Tay Bridge rail disaster, 1879 (1879 年Tay Bridge的火車災難).
(13) The Open Championship golf tournament, St. Andrews, 1939 (1939 年聖安德魯斯高爾夫球公開賽).
(14) Elizabeth the Queen Mother visiting Crail Camp, 1954 (1954 年伊莉莎白女王的母親訪問Crail Camp).
(15) Bomb damage due to World War II (二戰轟炸造成的損害).
(16) Pictures of York Minster (約克大教堂的照片).
(17) All views of North Street, St. Andrews (聖安德魯斯北街的所有景觀).
(18) Pictures of Edinburgh Castle taken before 1900 (1900 年之前拍攝的愛丁堡城堡的照片).
(19) People marching or parading (列隊行進或遊行中的人們).
(20) River with a viaduct in background (背景有高架橋的河流).
(21) War memorials in the shape of a cross (十字型戰爭紀念碑).
(22) Pictures showing traditional Scottish dancers (跳傳統蘇格蘭舞的照片).
(23) Photos of swans on a lake (天鵝在湖上的照片).
(24) Golfers swinging their clubs (正在揮動球杆的高爾夫球員).
(25) Boats on a canal (運河上的船隻).

(b) **The 2005 topic set**

(1) Aircraft on the ground (地面上的飛機).
(2) People gathered at bandstand (演奏臺旁聚集的群眾).
(3) Dog in sitting position (狗的坐姿).
(4) Steam ship docked (靠碼頭的蒸汽船).
(5) Animal statue (動物雕像).
(6) Small sailing boat (小帆船).
(7) Fishermen in boat (在船上的漁夫們).
(8) Building covered in snow (被雪覆蓋的建築物).
(9) Horse pulling cart or carriage (馬拉動運貨車或四輪車的圖片).
(10) Sun pictures, Scotland (蘇格蘭的太陽).
(11) Swiss mountain scenery (瑞士山景).
(12) Postcards from Iona, Scotland (蘇格蘭愛奧那島的明信片).
(13) Stone viaduct with several arches (多拱形石頭拱橋).
(14) People at the marketplace (市場的群眾).
(15) Golfer putting on green (高爾夫果嶺上撥球).
(16) Waves breaking on beach (海浪沖擊海灘).

(17) Man or woman reading (正在閱讀的男人或女人).
(18) Woman in white dress (身穿白色晚禮服的女人).
(19) Composite postcards of Northern Ireland (北愛爾蘭複合風景明信片).
(20) Royal visit to Scotland (not Fife) (皇室到蘇格蘭拜訪(不包括伐夫郡)).
(21) Monument to poet Robert Burns (詩人羅伯特彭斯的紀念碑).
(22) Building with waving flag (建築物上飄揚的旗子).
(23) Tomb inside church or cathedral (教堂和大教堂裡的墓).
(24) Close-up picture of bird (近距離的小鳥圖片).
(25) Arched gateway (拱形的通道).
(26) Portrait pictures of mixed sex group (一群男女在一起的相片).
(27) Woman or girl carrying basket (提籃子的女人或女孩).
(28) Colour pictures of woodland scenes around St. Andrews (聖安德魯斯周圍林地景色的彩色圖片).

References

Besançon, R., Hède, P., Moellic, P. A., & Fluhr, C. (2005). Cross-media feedback strategies: Merging text and image information to improve image retrieval. In C. Peters, P. Clough, J. Gonzalo, G. J. F. Jones, M. Kluck, & B. Magnini (Eds.), *Proceedings of 5th workshop of the cross-language evaluation forum. LNCS 3491* (pp. 709–717). Berlin: Springer.

Brill, E. (1995). Transformation-based error-driven learning and natural language processing: A case study in part of speech tagging. *Computational Linguistics, 21*(4), 543–565.

Carson, C., Belongie, S., Greenspan, H., & 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.

Chen, H. H., Bian, G. W., & Lin, W. C. (1999). Resolving translation ambiguity and target polysemy in cross-language information retrieval. In *Proceedings of 37th annual meeting of the association for computational linguistics* (pp. 215–222). East Stroudsburg, PA: Association for Computational Linguistics.

Chen, H. H., Ding, Y. W., Tsai, S. C., & Bian, G. W., (1998). Description of the NTU system used for MET2. In *Proceedings of seventh message understanding conference*. Available from: <http://www.itl.nist.gov/iaui/894.02/related_projects/muc/index.html>.

Chen, H. H., Lin, W. C., Yang, C., & Lin, W. H. (2006). Translating–transliterating named entities for multilingual information access. *Journal of American Society for Information Science and Technology, Special Issue on Multilingual Information Systems, 57*(5), 645–659.

Clough, P., Müller, H., Deselaers, T., Grubinger, M., Lehmann, T.M., Jensen, J., et al. (in press). The CLEF 2005 cross-language image retrieval track. In *Proceedings of 6th workshop of the cross-language evaluation forum 2005*. Lecture Notes in Computer Science.

Clough, P., Sanderson, M., & Müller, H. (2005). The CLEF 2004 cross-language image retrieval track. In C. Peters, P. Clough, J. Gonzalo, G. J. F. Jones, M. Kluck, & B. Magnini (Eds.), *Proceedings of 5th workshop of the cross-language evaluation forum. LNCS 3491* (pp. 597–613). Berlin: Springer.

Duygulu, P., Barnard, K., Freitas, N., & 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).

Goodrum, A. A. (2000). Image information retrieval: an overview of current research. *Information Science, 3*(2), 63–66.

Jeon, J., Lavrenko, V., & 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* (pp. 119–126). New York: ACM Press.

Jones, G. J. F., Groves, D., Khasin, A., Lam-Adesina, A., Mellebeek, B., & Way, A. (2005). Dublin City University at CLEF 2004: experiments with the ImageCLEF St Andrew's collection. In C. Peters, P. Clough, J. Gonzalo, G. J. F. Jones, M. Kluck, & B. Magnini (Eds.), *Proceedings of 5th workshop of the cross-language evaluation forum. LNCS 3491* (pp. 653–663). Berlin: Springer.

Lavrenko, V., Manmatha, R., & Jeon, J., (2003). A model for learning the semantics of pictures. In *Proceedings of the seventeenth annual conference on neural information processing systems*, December 9–11, 2003, Vancouver, British Columbia, Canada.

Lin, W. C., Chang, Y. C., & Chen, H. H. (2005). From text to image: generating visual query for image retrieval. In C. Peters, P. Clough, J. Gonzalo, G. J. F. Jones, M. Kluck, & B. Magnini (Eds.), *Proceedings of 5th workshop of the cross-language evaluation forum. LNCS 3491* (pp. 664–675). Berlin: Springer.

Lin, W. H., & Chen, H. H. (2002). Backward machine transliteration by learning phonetic similarity. In *Proceedings of sixth conference on natural language learning* (pp. 139–145). East Stroudsburg, PA: Association for Computational Linguistics.

Lowlands Team (2001). Lazy users and automatic video retrieval tools in (the) Lowlands. In Voorhees, E. M., Harman, D. K. (Eds.). *Proceedings of the tenth text retrieval conference* (pp. 159–168). Gaithersburg, MD: National Institute of Standards and Technology.

Mori, Y., Takahashi, H., & 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, October 30, 1999, Orlando, FL, in conjunction with ACM multimedia conference*. New York: ACM Press.

Robertson, S. E., Walker, S., & Beaulieu, M. (1998). Okapi at TREC-7: automatic ad hoc, filtering, VLC and interactive. In E. M. Voorhees & D. K. Harman (Eds.), *Proceedings of the seventh text retrieval conference* (pp. 253–264). Gaithersburg, MD: National Institute of Standards and Technology.