

Approaches of Using a Word-Image Ontology and an Annotated Image Corpus as Intermedia for Cross-Language Image Retrieval

Yih-Chen Chang and Hsin-Hsi Chen*

Department of Computer Science and Information Engineering
National Taiwan University
Taipei, Taiwan

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

Abstract. Two kinds of intermedia are explored in ImageCLEFphoto2006. The approach of using a word-image ontology maps images to fundamental concepts in an ontology and measure the similarity between two images by using the kind-of relationship of the ontology. The approach of using an annotated image corpus maps images to texts describing concepts in the images, and the similarity of two images is measured by text counterparts using BM25. The official runs show that visual query and intermedia are useful. Comparing the runs using textual query only with the runs merging textual query and visual query, the latter improved 71%~119% of the performance of the former. Even in the situation which example images were removed from the image collection beforehand, the performance was still improved about 21%~43%.

1 Introduction

In recent years, many methods have been proposed to explore visual information to improve the performance of cross-language image retrieval (Clough, Sanderson, & Müller, 2005; Clough, et al., 2006). The challenging issue is the semantic differences among visual and textual information. For example, the visual information “red circle” may be ambiguous, i.e., images may containing “red flower”, “red balloon”, “red ball”, and so on, rather than the desired ones containing “sun”. That requires more contexts to retrieve the correct interpretation. The semantic difference between visual concept “red circle” and textual symbol “sun” is called a *semantic gap*.

Previous approaches have conducted text- and content-based image retrieval separately and then merged the results of two runs (Besançon, et al., 2005; Jones, et al., 2005; Lin, Chang & Chen, 2007). Content-based image retrieval may suffer from the semantic gap problem and report noise images. That may have negative effects on the final performance. Other approaches learn the relationships between visual and textual

* Corresponding author.

information and used the relationships for media transformation (Lin, Chang & Chen, 2007). The final retrieval depends on the performance of the relationship mining.

In this paper, we use two intermedia approaches to deal with the semantic gap. The main characteristic of these approaches is that human knowledge is imbedded in the intermedia and can be used to compute the semantic similarity of images. A word-image ontology and an annotated image corpus are explored and compared. Section 2 specifies how to build and use the word-image ontology. Section 3 deals with the uses of the annotated corpus. Sections 4 and 5 show and discuss the official runs in ImageCLEFphoto2006.

2 An Approach of Using a Word-Image Ontology

2.1 Building the Ontology

A word-image ontology is a word ontology aligned with the related images on each node. Building such an ontology manually is time consuming. In ImageCLEFphoto2006, the image collection has 20,000 colored images. There are 15,998 images containing English captions in <TITLE> and <DESCRIPTION> fields. The vocabularies include more than 8,000 different words, thus an ontology with only hundreds of words is not enough.

Instead of creating a new ontology from scratch, we extend WordNet, the well-known word ontology, to a word-image ontology. In WordNet, different senses and relations are defined for each word. For simplicity, we only consider the first two senses and kind-of relations in the ontology. Because our experiments in ImageCLEF2004 (Lin, Chang & Chen, 2006) showed that verbs and adjectives are less appropriate to be represented as visual features, we only used nouns here.

Before aligning images and words, we selected those nouns in both WordNet and image collection based on their TF-IDF scores. For each selected noun, we used Google image search to retrieval 60 images from the web. The returned images may encounter two problems: (1) they may have unrelated images, and (2) the related images may not be pure enough, i.e., the foci may be in the background or there may be some other things in the images. Zinger (2005) tried to deal with this problem by using visual features to cluster the retrieved images and filtering out those images not belonging to any clusters. Unlike Zinger (2005), we employed textual features. For each retrieved image, Google will return a short snippet. We filter out those images whose snippets do not exactly match the query terms. The basic idea is: “the more things a snippet mentions, the more complex the image is.” Finally, we get a word-image ontology with 11,723 images in 2,346 nodes.

2.2 Using the Ontology

2.2.1 Similarity Scoring

Each image contains several fundamental concepts specified in the word-image ontology. The similarity of two images is measured by the sum of the similarity of the fundamental concepts. In this paper we use kind-of relations to compute semantic

distance between fundamental concepts A and B in the word-image ontology. At first, we find the least common ancestor (LCA) of A and B . The distance between A and B is the length of the path from A through LCA to B . When computing the semantic distance of nodes A and B , the more the nodes should be traversed from A to B , the larger the distance is. In addition to the path length, we also consider the weighting of links in a path shown as follows.

(1) When computing the semantic distance of a node A and its child B , we consider the number of children of A . The more children A has, the larger the distance between A and B is. In an extreme case, if A has only one child B , then the distance between A and B is 0. Let $\#children(A)$ denote the number of children of A , and $base(A)$ denote the basic distance of A and its children. We define $base(A)$ to be $\log(\#children(A))$.

(2) When computing the semantic distance of a node A and its brother, we consider the level it locates. Assume B is a child of A . If A and B have the same number of brothers, then the distance between A and its brothers is larger than that between B and its brothers. Let $level(A)$ be the depth of node A . Assume the level of root is 0. The distance between node A and its child, denoted by $dist(A)$, is defined to be $C^{level(A)} \times base(A)$. Here C is a constant between 0 and 1. In this paper, C is set to 0.9.

If the shortest path between two different nodes N_0 and N_m is $N_0, N_1, \dots, N_{LCA}, \dots, N_{m-1}, N_m$, we define the distance between N_0 and N_m to be:

$$dist(N_0, N_m) = dist(N_{LCA}) + \sum_{i=1}^{m-1} dist(N_i) \quad (1)$$

The larger the distance of two nodes is, the less similar the nodes are.

2.2.2 Mapping into the Ontology

Before counting the semantic distance between two given images, we need to map the two images into nodes of the ontology. In other words, we have to find the fundamental concepts the two images consist of. A CBIR system is needed to do this. It regards an image as a visual query and retrieves the top k fundamental images (i.e., fundamental concepts) in the word-image ontology. In such a case, we have two sets of nodes $S_1 = \{n_{11}, n_{12}, n_{13}, \dots, n_{1k}\}$ and $S_2 = \{n_{21}, n_{22}, n_{23}, \dots, n_{2k}\}$, which correspond to the two images, respectively. We sum the CBIR scores of the k fundamental images with the given two images, respectively. The node with the highest score is selected as a basis, e.g., S_1 , to compute similarity. For each fundamental image n_{1i} ($1 \leq i \leq k$) for S_1 , we select the minimum distance between n_{1i} and n_{1j} ($1 \leq j \leq k$) of S_2 . The semantic distance between S_1 and S_2 is sum of the above k minimum distances.

Given a query with m example images, we regard each example image Q as an independent visual query, and compute the semantic distance between Q and images in the collection. Note that we determine what concepts are composed of an image in the collection before retrieval. After m image retrievals, each image in the collection has been assigned m scores based on the above formula. We choose the best score for each image and sort all the images to create a rank list. Finally, the top 1000 images in the rank list will be reported.

3 An Approach of Using an Annotated Image Corpus

An annotated image corpus is a collection of images along with their text annotations. The text annotation specifies the concepts and their relationships in the images. To measure the similarity of two images, we have to know how many common concepts there are in the two images. An annotated image corpus can be regarded as a reference corpus. We submit two images to be compared to the reference corpus using a CBIR system. The corresponding text annotations of the retrieved images are postulated to contain the concepts embedded in the two images. The similarity of text annotations measures the similarity of the two images indirectly.

The image collection in ImageCLEFphoto2006 can be considered as a reference annotated image corpus. Using image collection itself as intermedia has some advantages. It is not necessary to map images in the image collection to the intermedia. Besides, the domain can be more restricted to peregrine pictures. In the experiments, the <DESCRIPTION>, <NOTE>, and <LOCATION> fields in English form the annotated corpus.

To use the annotated image corpus as intermedia to compute similarity between example images and images in image collection, we need to map these images into intermedia. Since we use the image collection itself as intermedia, we only need to map example images in this work. An example image is considered as a visual query and submitted to retrieve images in intermedia by a CBIR system. The corresponding text counterparts of the top returned k images form a long text query and it is submitted to an Okapi system to retrieval images in the image collection. BM25 formula measures the similarity between example images and images in image collection.

4 Experiments

In the formal runs, we submitted 25 cross-lingual runs for eight different query languages. All the queries with different source languages were translated into target language (i.e., English) using Professional SYSTRAN system. We considered several issues, including (1) using different intermedia approaches (i.e., the text-image ontology and the annotated image corpus), and (2) with/without using visual query. In addition, we also submitted 4 monolingual runs which compared (1) the annotation in English and in German, and (2) using or not using visual query and intermedia. Finally, we submitted a run using visual query and intermedia only. The details of our runs are described as follows:

- (1) 8 cross-lingual and text query only runs:
 NTU-PT-EN-AUTO-NOFB-TXT, NTU-RU-EN-AUTO-NOFB-TXT,
 NTU-ES-EN-AUTO-NOFB-TXT, NTU-ZHT-EN-AUTO-NOFB-TXT,
 NTU-FR-EN-AUTO-NOFB-TXT, NTU-JA-EN-AUTO-NOFB-TXT,
 NTU-IT-EN-AUTO-NOFB-TXT, and NTU-ZHS-EN-AUTO-NOFB-TXT.

These runs are regarded as baselines and are compared with the runs using both textual and visual information.

- (2) 2 monolingual and text query only runs:
 NTU-EN-EN-AUTO-NOFB-TXT, and NTU-DE-DE-AUTO-NOFB-TXT.

These two runs serve as the baselines to compare with cross-lingual runs with text query only, and to compare with the runs using both textual and visual information.

- (3) 1 visual query only run with the approach of using an annotated image corpus:
NTU-AUTO-FB-TXTIMG-WEprf.

This run will be merged with the runs using textual query only, and is also a baseline to compare with the runs using both visual and textual queries.

- (4) 8 cross-lingual runs, using both textual and visual queries with the approach of an annotated corpus:

NTU-PT-EN-AUTO-FB-TXTIMG-T-WEprf,
NTU-RU-EN-AUTO-FB-TXTIMG-T-WEprf,
NTU-ES-EN-AUTO-FB-TXTIMG-T-WEprf,
NTU-FR-EN-AUTO-FB-TXTIMG-T-WEprf,
NTU-ZHS-EN-AUTO-FB-TXTIMG-T-WEprf,
NTU-JA-EN-AUTO-FB-TXTIMG-T-WEprf,
NTU-ZHT-EN-AUTO-FB-TXTIMG-T-WEprf, and
NTU-IT-EN-AUTO-FB-TXTIMG-T-WEprf.

These runs merge the textual query only runs in (1) and visual query only run in (3) with equal weight.

- (5) 8 cross-lingual runs, using both textual and visual queries with the approach of using word-image ontology:

NTU-PT-EN-AUTO-NOFB-TXTIMG-T-IOntology,
NTU-RU-EN-AUTO-NOFB-TXTIMG-T-IOntology,
NTU-ES-EN-AUTO-NOFB-TXTIMG-T-IOntology,
NTU-FR-EN-AUTO-NOFB-TXTIMG-T-IOntology,
NTU-ZHS-EN-AUTO-NOFB-TXTIMG-T-IOntology,
NTU-JA-EN-AUTO-NOFB-TXTIMG-T-IOntology,
NTU-ZHT-EN-AUTO-NOFB-TXTIMG-T-IOntology, and
NTU-IT-EN-AUTO-NOFB-TXTIMG-T-IOntology.

These runs merge textual query only runs in (1), and visual query runs with weights 0.9 and 0.1.

- (6) 2 monolingual runs, using both textual and visual queries with the approach of an annotated corpus:

NTU-EN-EN-AUTO-FB-TXTIMG, and NTU-DE-DE-AUTO-FB-TXTIMG

These two runs using both textual and visual queries. The monolingual run in (2) and the visual run in (3) are merged with equal weight.

5 Results and Discussions

Table 1 shows experimental results of official runs in ImageCLEFphoto2006. We compare performance of the runs using textual query only, and the runs using both textual and visual queries (i.e., Text (T) Only vs. Text + Annotation (A) and Text + Ontology (O)). In addition, we also compare the runs using word-image ontology and the runs using annotated image corpus (i.e., Text + Ontology vs. Text + Annotation). The runs whose performance is better than that of baseline (i.e., Text Only) will be marked in bold. The results show all runs using annotated image corpus are better than the baseline. In contrast, only two runs using word-image ontology are better.

Table 1. Performance of Official Runs

Query Language	MAP	Description	Runs
Portuguese	0.1630	T	NTU-PT-EN-AUTO-NOFB-TXT
	0.2854	T+A	NTU-PT-EN-AUTO-FB-TXTIMG-T-Weprf
	0.1580	T+O	NTU-PT-EN-AUTO-NOFB-TXTIMG-T-IOntology
Russian	0.1630	T	NTU-RU-EN-AUTO-NOFB-TXT
	0.2789	T+A	NTU-RU-EN-AUTO-FB-TXTIMG-T-Weprf
	0.1591	T+O	NTU-RU-EN-AUTO-NOFB-TXTIMG-T-IOntology
Spanish	0.1595	T	NTU-ES-EN-AUTO-NOFB-TXT
	0.2775	T+A	NTU-ES-EN-AUTO-FB-TXTIMG-T-Weprf
	0.1554	T+O	NTU-ES-EN-AUTO-NOFB-TXTIMG-T-IOntology
French	0.1548	T	NTU-FR-EN-AUTO-NOFB-TXT
	0.2758	T+A	NTU-FR-EN-AUTO-FB-TXTIMG-T-Weprf
	0.1525	T+O	NTU-FR-EN-AUTO-NOFB-TXTIMG-T-IOntology
Simplified Chinese	0.1248	T	NTU-ZHS-EN-AUTO-NOFB-TXT
	0.2715	T+A	NTU-ZHS-EN-AUTO-FB-TXTIMG-T-Weprf
	0.1262	T+O	NTU-ZHS-EN-AUTO-NOFB-TXTIMG-T-IOntology
Japanese	0.1431	T	NTU-JA-EN-AUTO-NOFB-TXT
	0.2705	T+A	NTU-JA-EN-AUTO-FB-TXTIMG-T-Weprf
	0.1396	T+O	NTU-JA-EN-AUTO-NOFB-TXTIMG-T-IOntology
Traditional Chinese	0.1228	T	NTU-ZHT-EN-AUTO-NOFB-TXT
	0.2700	T+A	NTU-ZHT-EN-AUTO-FB-TXTIMG-T-Weprf
	0.1239	T+O	NTU-ZHT-EN-AUTO-NOFB-TXTIMG-T-IOntology
Italian	0.1340	T	NTU-IT-EN-AUTO-NOFB-TXT
	0.2616	T+A	NTU-IT-EN-AUTO-FB-TXTIMG-T-Weprf
	0.1287	T+O	NTU-IT-EN-AUTO-NOFB-TXTIMG-T-IOntology

The reason why the word-image ontology does not perform as well as we expected may be that the images in the word-image ontology come from the web and the images in the web still contain much noise even after filtering. To deal with this problem, a better method of image filtering is necessary.

Since the example images in this task are in the image collection, the CBIR system always correctly maps the example images into themselves at mapping step. We made some extra experiments to examine the performance of our intermedia approach. In the experiments, we took out the example images from the image collection when mapping example images into intermedia. Table 2 shows the experiment results of the unofficial runs. Comparing Table 1 and Table 2 we find the performance of Table 2 is lower than that of Table 1. It shows the performance of CBIR in mapping stage will influence the final result and that is very critical. From Table 2, we also find that the approaches of annotated image corpus are better than the runs using textual query only. It shows even though there are some errors in mapping stage, the annotated image corpus can still work well.

Table 3 shows the experiment results of monolingual runs. Using both textual and visual queries are still better than runs using textual query only. The performance of the runs by taking out the example images from collection beforehand is still better than the runs use textual query only. From this table, we also find the runs using textual query only

Table 2. MAP of Unofficial Runs by Removing Example Images from the Collection

Query Language	Text Only	Text + Annotated image corpus
Portuguese	0.1630	0.1992
Russian	0.1630	0.1880
Spanish	0.1595	0.1928
French	0.1548	0.1848
Simplified Chinese	0.1248	0.1779
Japanese	0.1431	0.1702
Traditional Chinese	0.1228	0.1757
Italian	0.1340	0.1694

does not perform well even in monolingual runs. This may be because the image captions of this year are shorter and we do not have enough information when we use textual information only. In addition, when image captions are short too, the little differences in vocabularies between query and document may influence the results a lot. Therefore, German monolingual run and English monolingual runs perform very differently.

Table 4 shows the experiment of runs that use the visual query and annotated image corpus only, i.e., the textual query is not used. When example images were kept in the image collection, we can always map the example images into the right images. Therefore, the translation from visual information into textual information will be more correctly. The experiment shows the performance of visual query runs is better than that of textual query runs when the transformation is correct.

Table 3. Performance of Monolingual Image Retrieval

Query Language	MAP	Description	Runs
English	0.1787	T	NTU-EN-EN-AUTO-NOFB-TXT
(+example images)	0.2950	T+A	NTU-EN-EN-AUTO-FB-TXTIMG
(-example images)	0.2027	T+A	NTU-EN-EN-AUTO-FB-TXTIMG-NoE
German	0.1294	T	NTU-DE-DE-AUTO-NOFB-TXT
(+example images)	0.3109	T+A	NTU-DE-DE-AUTO-FB-TXTIMG
(-example images)	0.1608	T+A	NTU-DE-DE-AUTO-FB-TXTIMG-NoE

Table 4. Performance of Visual Query

MAP	Description	Runs
0.1787	T (monolingual)	NTU-EN-EN-AUTO-NOFB-TXT
0.2757	V+A (+example images)	NTU-AUTO-FB-TXTIMG-Weprf
0.1174	V+A (-example images)	NTU-AUTO-FB-TXTIMG-Weprf-NoE

6 Conclusion

The experiments show visual query and intermedia approaches are useful. Comparing the runs using textual query only with the runs merging textual query and visual query,

the latter improved 71%~119% of performance of the former. Even in the situation which example images are removed from the image collection, the performance can still be improved about 21%~43%. We find that the visual query in image retrieval is important. The performance of the runs using visual query only can be even better than the runs using textual only if we translate visual information into textual one correctly. The word-image ontology built automatically still contains much noise. We plan to investigate how to filter out the noise and explore different methods for cross-language image retrieval.

Acknowledgements

Research of this paper was partially supported by National Science Council, Taiwan, under the contracts NSC95-2221-E-002-334 and NSC95-2752-E-001-001-PAE.

References

1. Besançon, R., Héde, P., Moellic, P.A., Fluhr, C.: Cross-media feedback strategies: Merging text and image information to improve image retrieval. In: Peters, C., Clough, P.D., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) CLEF 2004. LNCS, vol. 3491, pp. 709–717. Springer, Heidelberg (2005)
2. Clough, P., Sanderson, M., Müller, H.: The CLEF 2004 cross language image retrieval track. In: Peters, C., Clough, P.D., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) CLEF 2004. LNCS, vol. 3491, pp. 597–613. Springer, Heidelberg (2005)
3. Clough, P., Müller, H., Deselaers, T., Grubinger, M., Lehmann, T.M., Jensen, J., Hersh, W.: The CLEF 2005 cross-language image retrieval track. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 535–557. Springer, Heidelberg (2006)
4. Jones, G.J.F., Groves, D., Khasin, A., Lam-Adesina, A., Mellebeek, B., Way, A.: Dublin City University at CLEF 2004: Experiments with the ImageCLEF St Andrew's Collection. In: Peters, C., Clough, P.D., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) CLEF 2004. LNCS, vol. 3491, pp. 653–663. Springer, Heidelberg (2005)
5. Lin, W.C., Chang, Y.C., Chen, H.H.: Integrating textual and visual information for cross-language image retrieval: A trans-media dictionary approach. *Information Processing and Management, Special Issue on Asia Information Retrieval Research* 43(2), 488–502 (2007)
6. Zinger, S.: Extracting an ontology of portable objects from WordNet. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, Springer, Heidelberg (2006)