Combining Text and Image Queries at ImageCLEF2005

Yih-Cheng Chang[1], Wen-Cheng Lin[1,2] and Hsin-Hsi Chen[1]

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

[2]Department of Medical Informatics
Tzu Chi University
Hualien, Taiwan
ycchang@nlg.csie.ntu.edu.tw; denislin@mail.tcu.edu.tw; hhchen@csie.ntu.edu.tw

Abstract

This paper presents our methods for the tasks of bilingual ad hoc retrieval and automatic annotation in ImageCLEF 2005. In ad hoc task, we propose a feedback method for cross-media translation in a visual run, and combine the results of visual and textual runs to generate the final result. Experimental results show that our feedback method performs well. Comparing to initial visual retrieval, average precision is increased from 8% to 34% after feedback. The performance is increased to 39% if we combine the results of textual run and visual run with pseudo relevance feedback. In automatic annotation task, we propose several methods to measure the similarity between a test image and a category, and a test image is classified to the most similar category. Experimental results show that the proposed approaches have good performance, but the simplest 1-NN method has the best performance. We will analyze these results in the paper.

ACM Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval---Retrieval models, Relevance feedback

Free Keywords: Cross language image retrieval, cross-media translation, automatic image annotation, classification

1 Introduction

While digital images have an explosive growth, cross-language image retrieval and automatic annotation become very important nowadays. An automatic annotation system can help us to annotate large amount of images, and a cross-language image retrieval system retrieves images that are annotated in different languages.

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

Recently many approaches tried to combine text- and content-based methods for image retrieval. A simple approach is conducting text- and content-based retrieval separately and merging the retrieval results of the two runs [3,4]. In contrast to the parallel approach, a pipeline approach uses textual or visual information to perform initial retrieval, and then uses the other features to filter out irrelevant images [5]. In these two approaches, textual and visual queries are formulated by users and do not directly influence each other. Another approach, i.e., transformation-based approach, tries to mine the relations between images and text, and uses the mined relations to transform textual information into visual one, and vice versa [6]. In this paper we try another method to transform visual features to textual ones. We use a feedback method to transform a visual query into textual one. The text descriptions of the top retrieved images of the initial retrieval are used for feedback to conduct a second retrieval. The new textual information can help us cache the semantic meaning of a visual query, and thus improve retrieval performance.

The correlation between images and text can be used to annotate images. However, the training data of automatic annotation task has no textual information, thus we use only visual features to classify images. In automatic annotation task, we try several classification methods. A nearest neighbor (1-NN) method is

considered as our baseline. We propose several methods to measure the similarity between a test image and a class, and a test image is classified to the most similar class. We propose a method that measures the similarity between an image and a class by averaging the similarity scores of the top *n* most similar images in the class. Besides, we also propose an approach that divides a class into several smaller classes and classifies a test image according to the similarities between the test image and the centroids of the smaller classes.

The rest of the paper is organized as fellows. Section 2 and 3 introduce the proposed approaches and experimental results of bilingual ad hoc retrieval task and automatic annotation task, respectively. Section 4 concludes the remark.

2 Bilingual Ad Hoc Retrieval Task

2.1 Feedback Method for Cross-Media Translation

To do cross-media translation between visual and textual representations, several correlation-based approaches have been proposed in automatic annotation task. Those approaches model the correlation between text and visual representation, and use the mined relation to translate images to text descriptions. Mori, Takahashi and Oka [7] divided images into grids, and then the grids of all images were clustered. Co-occurrence information was used to estimate the probability of each word for each cluster. Duygulu, *et al.* [8] used blobs to represent images. First, images are segmented into regions using a segmentation algorithm like Normalized Cuts [9]. 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 [10] proposed a cross-media relevance model (CMRM) to learn the joint distribution of blobs and words. They further proposed continuous-space relevance model (CRM) that learned the joint probability of words and regions, rather than blobs [11].

The above approaches use the relation between text and visual representation as a bridge to translate image to text. However, it is hard to learn all relations between all visual and textual features. In the experiments mentioned above, relations are learned from only hundreds of keywords in textual annotation. Another problem is that the degree of ambiguity of the relations is usually high. For example, visual feature "red circle" may have many meanings such as sun set, red flower, and red ball. Similarly, the word "flower" may look very different, e.g. different color and shape, in images.

In this paper we translate visual and textual features without learning correlations. We treat the retrieved images and their text descriptions as aligned documents, and a corpus-based method that uses pseudo relevance feedback is adopted to translate visual or textual features and generate a new query.

In cross-language image retrieval, giving a set of images $I=\{i_1,i_2,\ldots,i_m\}$ with text descriptions $T_{I,L1}=\{t_1,t_2,\ldots,t_m\}$ in language L1, users use textual query Q_{L2} in language L2 (L2≠L1) and example images $E=\{e_1,e_2,\ldots,e_p\}$ to retrieve relevant images from I.

We use a feedback method in a visual run to translate the visual query into textual one as follows. We first use an example image e_i as initial query, and use a CBIR system, i.e. VIPER [12], to retrieve images from I. The retrieved images are $R=\{r_{i1},r_{i2},\ldots,r_{in}\}$ and their text descriptions are $T_{R,L1}=\{t_{ri1},t_{ri2},\ldots,t_{ril}\}$ in language L1. Then we use the text descriptions of the top *k* retrieved images to construct a new textual query. The new textual query can be seen as a translation of initial visual query. In the feedback run, we submit the new textual query to a text-based retrieval system, i.e. Okapi [13], to retrieve images from I.

In addition to the visual feedback run, we also conduct a text-based run using the textual query in the test set. We use the method we proposed last year [14] to translate textual query Q_{L2} into query Q_{L1} in language L1, and submit the translated query Q_{L1} to Okapi system to retrieve images. The results of textual run and visual feedback run can be combined. The similarity scores of images in the two runs are normalized and linearly combined using equal weight.

2.2 Experimental Results

In the experiments, the text-based retrieval system used is Okapi IR system, and the content-based retrieval system used is VIPER system. For textual index in Okapi, the caption text, <HEADLINE> and <CATEGORIES> sections of English captions are used for indexing. The weighting function used is BM25. Chinese queries and example images are used as our source queries.

We submitted four Chinese-English cross-lingual runs, two English monolingual runs and one visual run in CLEF 2005 image track. In English monolingual runs, using narrative or not using narrative will be compared.

In the four cross-lingual runs, combining with visual run or not combing with visual run, and using narrative or not using narrative will be compared. The details of the cross-lingual runs and visual run are described as follows.

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

This run use textual queries without narrative to retrieve images. We use query translation method we used last year to translate Chinese queries into English to retrieve images using textual index.

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

This run use textual queries with narrative. We only use the positive information in narrative. The sentences that contain phrase "不算相關 (are not relevant)" are removed.

(3) NTU-adhoc05-EX-prf

This run is a visual run with pseudo relevance feedback (the query becomes textual one after feedback). We use the retrieval results of VIPER system provided by ImageCLEF as our initial retrieval results, and use the text descriptions of the top 2 images to construct a new textual query in feedback run. The caption text in descriptions is used to construct a query. The textual query is submitted to Okapi IR system to retrieve images.

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

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

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

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

From Table 1, the average precision of monolingual retrieval using title field only is 0.3952. Comparing to the performance of last year (0.6304), this year's query set is much harder. After adding narrative information, average precision is increased slightly. The performance of Chinese-English cross-lingual textual run is about 60.7% of English monolingual run. It shows that there are still many errors in language translation. From Table 2, the performance of initial visual run, i.e. VIPER, is not good enough. Text-based runs, even cross-lingual runs, perform much better than initial visual run. It shows that semantic information is very important for the queries of this year. After feedback, the performance is increased dramatically from 0.0829 to 0.3452. The result shows that the feedback method transforms visual information into textual one well. Combining textual and visual feedback runs further improves retrieval performance. The combined runs perform better than the individual runs. The results show that it needs more information to define users' information need. The feedback textual query has additional information and help user's textual query perform better.

Table 1. Results of official runs

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

Table 2. Performances of unofficial runs (NTU-adhoc05-EE-T-WEprf merges the results of NTU-adhoc05-EE-T-W and NTU-adhoc05-EX-prf)

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

Figure 1. Average precision of each query

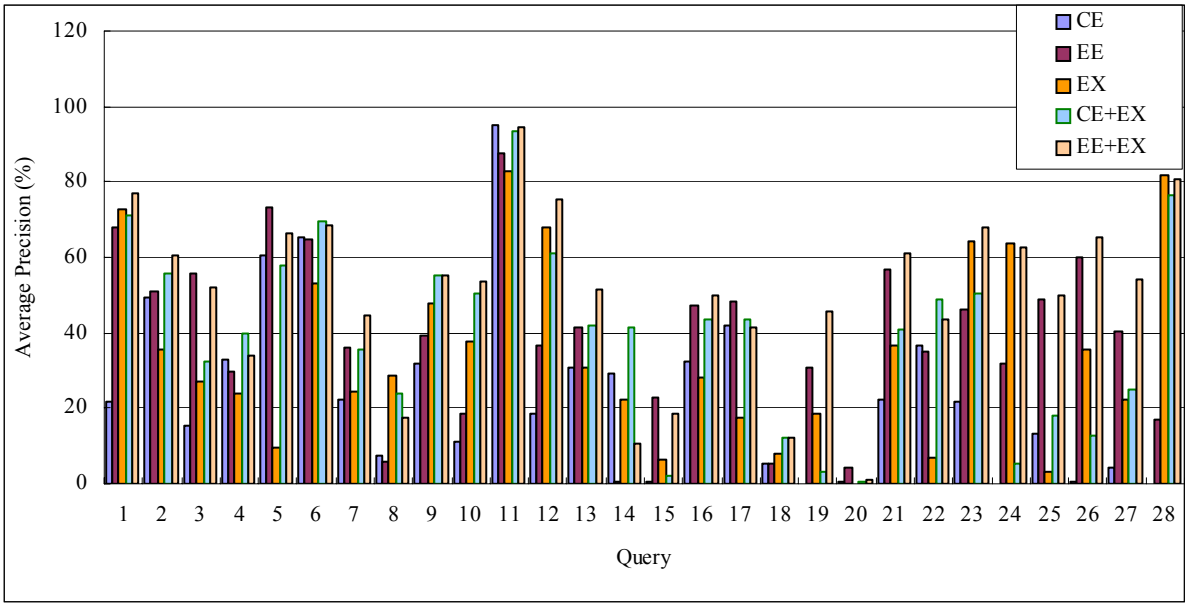

Figure 1 shows the performances of each query in run NTU-adhoc05-CE-T-W, NTU-adhoc05-EE-T-W, NTU-adhoc05-EX-prf, NTU-adhoc05-CE-T-Weprf, and NTU-adhoc05-EE-T-WEprf. For most queries, monolingual run has better performance than visual feedback run. We can say that there are translation errors in cross-media translation. There are ten topics in which the performance of visual feed back run is better than that of monolingual run. This is probably because that the user's information need is not detailed described in a textual query, i.e. some information is lost. Also, the words used in textual query and image descriptions may be inconsistency. Thus, it is hard to retrieve all relevant images by the textual queries formulated by users. We can use additional information that is not provided directly by users to retrieve more relevant images. The constructed query in feedback run has additional information that comes from example images. For example, when a user wants to find images that have aircraft on land, using query "aircraft in military air base" may be better than using "aircraft on the ground". This is because that the descriptions of images don't mention that aircraft is on the ground directly, but aircrafts in military air base are very likely to be parked and thus are on the ground. The additional information "military air base" is obtained because that it is mentioned in the descriptions of images retrieved by example images using a CBIR system. Comparing the performances of runs that combining visual feedback run or not, we can find that most topics perform better after combining. This is probably because that the additional information in feedback run helps our system retrieves images more precisely, and that queries constructed in feedback run could recover translation errors in a cross-lingual run.

3 Automatic Annotation Task

3.1 Classification Approaches

The automatic annotate task in ImageCLEF 2005 can be seen as a classification task, since each image can only be annotated with one word (category). In classification task, *k*-nearest neighbor (*k*-NN) method is a usually adopted approach [15]. Performance for different categories in *k*-NN method usually depends on the number of training data in each category. Test images tend to be classified to the categories that have many training data (We will show this later). To solve this problem, computing several representative data is used to normalize the number of training data in each category. We can reduce the number of training data to 1 using a centroid to represent a category. But sometimes using only one centroid to represent a whole category is not sufficient if the images in the category are very different. For example, the images of the flank and the front of skull look very different. Using two centroids of two smaller classes to represent category "skull" is better than using only one centroid that is between the flank and front of skull. In this task, we use clustering to help us to find the representative data of each category. We assume that the images that belong to the same cluster and the same category are very similar, and can be represented by a centroid. The detail of our method is described as

follows.

(1) First we use *k*-means algorithm to cluster all training data. The images in a cluster may belong to different categories.
(2) After clustering, we compute the centroids of each category in each cluster.
(3) Given a test image, we compute the distances between it and each centroids, and the test image is classified to the category with the shortest distance.

The second method we used is to compute the similarities between a test image and each category, and then classify the test image to the most similar category. The similarity between a test image and a class is measured by averaging the similarity values between the test image and the top 2 most similar images in the class. A test image is classified to the class that has the highest similarity.

3.2 Experimental Results

In this task we submit three runs. The three runs use the same image features. The difference between them is the classification method used. The image features are extracted in the following way. First we resize images to 256 x 256 pixels and segment each image into 32 x 32 blocks (each block is 8 x 8 pixels). Then we compute the average gray value of each block to construct a vector with 1024 elements. We use this vector to represent an image, and the similarity between two images is measured by cosine formula. The details of each run are described as follows.

(1) NTU-annotate05-1NN
This run is our based line. It uses 1-NN method to classify each image.
(2) NTU-annotate05-Top2
This run uses the second method described in Section 3.1. We compute the similarity between a test image and category using the top 2 nearest images in each category, and classify the test image to the most similar category.
(3) NTU-annotate05-SC
This run uses the first method described in Section 3.1. Training data is clustered using *k*-means algorithm (*k*=1000). We compute the centroids of each category in each cluster, and classify a test image to the category of the nearest centroid.

The results of official runs are shown in Table 3. The results show that 1-NN method is very useful. 1-NN has the same performance as run NTU-annotate05-Top2, but it doesn't need to compute average similarity, thus it is faster than top2 method. The performance of run NTU-annotate05-SC is worse than run NTU-annotate05-1NN. Normalizing the number of training data in each category may have a trade-off. Normalization may increase the performance of categories that have less training data, but decrease the performance of categories that have more training data. Table 4 shows the error rate of individual categories. The categories that have a lot of training data are listed in the upper part of Table 4, and the categories that have a few training data are in the lower part. From Table 4, the performances of categories with a lot of training data are better than that of categories with a few training data. For the categories with a lot of training data, 1-NN method performs better than normalization method (run SC). In contrast, normalization method performs much better than 1-NN method for the categories with a few training data. It shows that normalization method could reduce the problem that prefers classifying images to large categories. The reason that the overall performance of normalization method is worse than that of 1-NN method is that large categories have more test images and thus have more influence on the final result.

Table 3. Results of official runs

Run	NTU-annotate05-1NN	NTU-annotate05-Top2	NTU-annotate05-SC
Error Rate	21.7 %	21.7 %	22.5 %

Table 4. Error rate of individual categories. The upper part shows the top 10 categories that have a lot of training data, and the lower part shows the categories that have a few training data

Cat.	#Training image	#Test image	Error rate (1NN)	Error rate (Top2)	Error rate (SC)
12	2563	297	0.003367	0.003367	0.016835
34	880	79	0.012658	0.012658	0.000000
6	576	67	0.194030	0.223881	0.253731
1	336	38	0.000000	0.000000	0.078947
25	284	36	0.138889	0.166667	0.194444
28	228	16	0.312500	0.312500	0.250000
5	225	25	0.080000	0.080000	0.080000
17	217	24	0.125000	0.125000	0.208333
3	215	24	0.291667	0.291667	0.250000
18	205	12	0.416667	0.500000	0.416667
Avg.	572.9	61.8	0.157478	0.171574	0.174896
51	9	1	0.000000	0.000000	0.000000
52	9	1	0.000000	0.000000	0.000000
55	10	2	1.000000	1.000000	1.000000
53	15	3	0.333333	0.000000	0.333333
15	15	3	0.666667	0.666667	0.666667
24	17	4	1.000000	0.750000	0.750000
35	18	4	0.750000	1.000000	0.500000
37	22	2	1.000000	1.000000	1.000000
16	23	1	1.000000	1.000000	0.000000
46	30	1	1.000000	1.000000	0.000000
Avg.	16.8	2.2	0.675000	0.641667	0.425000

4 Conclusions

In bilingual ad hoc retrieval task, we propose a simple and useful feedback method for cross-language image retrieval. We transform visual features into textual ones without learning correlations. Experimental results show that the proposed feedback approach performs well. Comparing to initial visual retrieval, average precision is increased from 8% to 34% after feedback. The feedback textual query has additional information that comes from example images, and help user's textual query perform better. After combining textual and visual feedback runs, average precision is increased from 0.2399 to 0.3977 and from 0.3952 to 0.5053 in cross-lingual and monolingual experiments, respectively. We will test our method in other image collections in the future.

In automatic annotation task, we propose a method that normalizes the number of training data of each category. The normalization approach may have a trade-off. It may increase the performance of categories that have less training data, but decrease the performance of categories that have more training data. We will try our method in different collections and study what is the suitable time to use normalization since using normalization may have a trade-off.

Acknowledgement

Research of this paper was partially supported by National Science Council, Taiwan, under the contracts NSC93-2752-E-001-001-PAE and NSC94-2752-E-001-001-PAE.

References

1. Goodrum, A.A.: Image Information Retrieval: An Overview of Current Research. Information Science, 3(2). (2000) 63-66.
2. Eidenberger, H. and Breiteneder, C.: Semantic Feature Layers in Content-based Image Retrieval: Implementation of Human World Features. In: Proceedings of International Conference on Control, Automation, Robotic and Vision. (2002).

3. Besançon, R., Hède, P., Moellic, P.A., and Fluhr, C.: Cross-Media Feedback Strategies: Merging Text and Image Information to Improve Image Retrieval. In: Multilingual Information Access for Text, Speech and Images: 5th Workshop of the Cross-Language Evaluation Forum, CLEF 2004, LNCS 3491. Springer-Verlag GmbH (2005) 709-717.
4. Jones, G.J.F., Groves, D., Khasin, A., Lam-Adesina, A., Mellebeek, B., and Way, A.: Dublin City University at CLEF 2004: Experiments with the ImageCLEF St. Andrew's Collection. In: Multilingual Information Access for Text, Speech and Images: 5th Workshop of the Cross-Language Evaluation Forum, CLEF 2004, LNCS 3491. Springer-Verlag GmbH (2005) 653-663.
5. Baan, J., van Ballegooij, A., Geusenbroek, J.M., den Hartog, J., Hiemstra, D., List, J., Patras, I., Raaijmakers, S., Snoek, C., Todoran, L., Vendrig, J., de Vries, A., Westerveld, T., and Worring, M.: Lazy Users and Automatic Video Retrieval Tools in (the) Lowlands. In: Proceedings of the Tenth Text REtrieval Conference (TREC 2001). National Institute of Standards and Technology (2002) 159-168.
6. Lin, W.C., Chang, Y.C. and Chen, H.H.: Integrating Textual and Visual Information for Cross-Language Image Retrieval. In: Proceedings of the Second Asia Information Retrieval Symposium (AIRS 2005). (2005).
7. Mori, Y., Takahashi, H. and Oka, R.: Image-to-Word Transformation Based on Dividing and Vector Quantizing Images with Words. In: Proceedings of the First International Workshop on Multimedia Intelligent Storage and Retrieval Management. (1999).
8. Duygulu, P., Barnard, K., Freitas, N. and Forsyth, D.: Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary. In: Proceedings of Seventh European Conference on Computer Vision, Vol. 4. (2002) 97-112.
9. Shi, J. and Malik, J.: Normalized Cuts and Image Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(8). (2000) 888-905.
10. Jeon, J., Lavrenko, V. and Manmatha, R.: Automatic Image Annotation and Retrieval using Cross-Media Relevance Models. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2003). ACM Press (2003) 119-126.
11. Lavrenko, V., Manmatha, R. and Jeon, J.: A Model for Learning the Semantics of Pictures. In: Proceedings of the Seventeenth Annual Conference on Neural Information Processing Systems. (2003).
12. Squire, D.M., Müller, W., Müller, H., and Raki, J.: Content-based query of image databases, inspirations from text retrieval: Inverted files, frequency-based weights and relevance feedback. In: Scandinavian Conference on Image Analysis. (1999) 143–149.
13. Robertson, S.E., Walker, S. and Beaulieu, M.: Okapi at TREC-7: Automatic Ad Hoc, Filtering, VLC and Interactive. In: Proceedings of the Seventh Text REtrieval Conference (TREC-7). National Institute of Standards and Technology (1998) 253-264.
14. Lin, W.C., Chang, Y.C. and Chen, H.H.: From Text to Image: Generating Visual Query for Image Retrieval. In: Multilingual Information Access for Text, Speech and Images: 5th Workshop of the Cross-Language Evaluation Forum, CLEF 2004, LNCS 3491. Springer-Verlag GmbH (2005) 664-675.
15. Lehmann, T.M., Güld, M.O., Deselaers, T., Keysers, D., Schubert, H., Spitzer, K., Ney, H., and Wein, B.B.: Automatic categorization of medical images for content-based retrieval and data mining. Computerized Medical Imaging and Graphics, 29(2-3). Elsevier, (2005) 143-155.