Mining Visual Knowledge for Multi-Lingual Image Retrieval

Masashi Inoue
National Institute of Informatics
Tokyo, Japan
Email: m-inoue@nii.ac.jp

Abstract

Users commonly rely just on scarce textual annotation when their searches for images are semantic or conceptual based. Rich visual information is often thrown away in basic annotation-based image retrieval because its relationship to the semantic content is not always clear. To ensure that appropriate visual information is included, we propose using visual clustering within pre-processing and post-processing steps of text-based retrieval. A clustering algorithm finds pairs of images that are nearly identical and are, therefore, presumed semantically similar. The output from basic retrieval systems is a ranked list of images based only on lexical term matching. The obtained cluster knowledge is then used to modify the ranking result during the post-processing step. Low ranked images considered nearly identical to more highly ranked images are then pulled up. The modularity of this architecture allows us to integrate a data mining process without having to change core information retrieval systems. Evaluation on a cross-language image retrieval test collection showed that this method improved retrieval performance for certain queries in multilingual settings.

1 Introduction

1.1 Requirements for image retrieval

The retrieval of visual documents is a rapidly growing area of information access. We use image retrieval systems according to the various information needs. As a result, the desired image set differs for different tasks. For example, when creating a brochure for a beach resort to attract tourists, it is quite important to select some eye-catching and enticing pictures. Relevant images may include the photos of the resort, people enjoying whatever activities are available, or beautifully prepared dishes. However, a brochure containing such only topically relevant images may not be appealing enough to attract potential tourists. We may want to find the best photos because visually quite similar images sometimes create a completely different impression. That is, details do matter in visual communication. For this reason, we may want to have as many different candidate images as possible to select the best photo that matches the required context the best.

Current image retrieval systems experience some difficulties when performing comprehensive searches for appropriate images. When images are semantically retrieved based on their associated textual annotations, such as keywords or file names, some ideal images lacking usable textual annotations will not be considered relevant by the systems. One such important example is when the annotation is assigned in a language different from the query language. To avoid such 'misses', we used a nearly identical image clustering technique as a pre-processing step before the actual retrieval. Once we know which images are linked to which images, the retrieved images and un-retrieved images lacking appropriate annotations can be connected by using knowledge. Figure 1 is a schematic diagram of this retrieval process. The knowledge extracted from the visual information is later used to re-rank the initial outputs from the information retrieval (IR) systems. The modular nature of this architecture allows us to use existing IR systems without the need for big changes.

1.2 Finding nearly identical image pairs

Similarities between images are commonly exploited during visual content-based retrieval. However, visual similarity is unsuitable for enhancing ranked lists. Visually similar images are not always semantically similar and this differs from text, whose semantic content can adequately be represented by using words. For example, a photo of sunset can easily be confused with a photo of an apple because visually they are both perceived as red circles by machines. By only using nearly identical images, we can avoid including unwanted noise from unrelated images.

In Section 2, we explain our data mining method using pre- and post-processing steps and the IR system used.

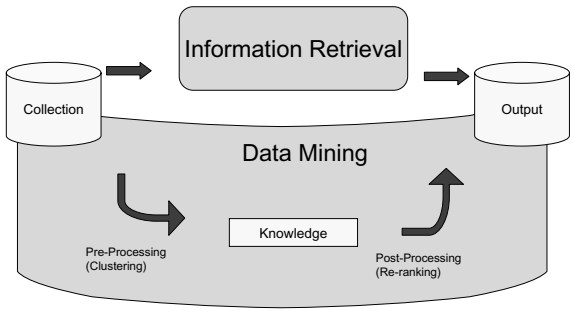

Figure 1. Combination of data mining (pre- and post-processing) and information retrieval.

In Section 3, we show the experimental results on the ImageCLEFphoto 2006 ad hoc test collection in multi-lingual setting. In Section 4, we discuss the characteristics of our approach and its relationship with other retrieval methods. Section 5 concludes the paper.

2 System description

2.1 Data mining procedure

The first step of our framework is to extract visual knowledge by clustering. Clustering algorithms can be categorized into two types: macro-clustering, or global partitioning, and micro-clustering, or local pairing. The entire feature space is divided into sub-regions in macro-clustering. Whereas, in micro-clustering, the data points that are nearby are linked so that they form a small group in a restricted region of the feature space. Since we only considered nearly identical image pairs, the resulting clusters became micro-clusters that only contained considerably close neighbors. The micro-clustering technique has been tested for text processing in which terms were used in calculating similarities [1]. We used micro-clustering on visual features and a similarity measure suitable for them.

The process of clustering is explained in the top half of Table 1. First, nearly identical images are sought for each image. They are then placed into a cluster that is associated with a seed image. The resulting cluster information is stored as metadata of each image for later use. Then, usual IR is conducted. After retrieval, the ranked image lists of images produced by the retrieval engine are modified using the cluster information. The post-processing is explained in the bottom half of Table 1. A ranked list is searched from the top, and when an image belonging to a cluster is found, all other members in the cluster will be given the same position as its highly ranked one. This process is continued until the number of images in the new list exceeds a prespecified number. Note that the retrieval results are evaluated on a certain number of top-ranked documents and not on the entire ranked list.

2.2 Features and similarity metric

Visual feature values were extracted from all images in the target collection. Simple color histograms were used as the feature. Images we used were provided in true color JPEG format, and this enabled histograms to be created for the red (R), green (G), and blue (B) elements in the images. This results in each image having three vectors: $\mathbf{x}_r$, $\mathbf{x}_g$, and $\mathbf{x}_b$. The length of each vector, or the size of the histogram, $i = 256$. These parameters are combined and used to define a single feature matrix for each image: $X = [\mathbf{x}_r, \mathbf{x}_g, \mathbf{x}_b]$. Thus, the size of feature matrix is i by j where $j = 3$.

Similarities between images were calculated using the above feature values. The similarity measure used was a two-dimensional correlation coefficient, r, between the matrices. If matrices A and B are assumed, the correlation coefficient is given as

$$r = \frac{\sum_i \sum_j (A_{ij} - \bar{A})(B_{ij} - \bar{B})}{\sqrt{(\sum_i \sum_j (A_{ij} - \bar{A})^2)(\sum_i \sum_j (B_{ij} - \bar{B})^2)}}$$

where $\bar{A}$ and $\bar{B}$ are the mean values of A and B respectively.

The image pairs whose r is larger than the threshold T are considered identical and grouped into the same cluster. That is, the actual threshold $d < T$ in Table 1 was performed as $r > T$ for the similarity measure. The threshold is determined manually by inspecting the distribution of similarity scores so that only a relatively small group of images constitute each separate cluster. The important point of this processing is to avoid including unrelated images in the same cluster.

2.3 Retrieval engine

Our method is unique solely in the combination of data mining with image retrieval. The basic ranking was conducted using an existing text search engine. We used an information retrieval toolkit, the Lemur Toolkit, for that purpose[1]. As for the configuration, we used a unigram language-modeling algorithm for building the document models, and Kullback-Leibler divergence for ranking. The document models were smoothed using Dirichlet prior [8].

[1] http://www.lemurproject.org/

Table 1. Procedures for pre-processing (knowledge extraction by clustering) and post-processing (knowledge injection by re-ranking)

	Pre-Processing
1	**for** image $n \in C$ (C is the document collection)
2	**for** image $m \in C$ except n
3	calculate distance d between n and m
4	**if** $d <$ T (T is pre-specified threshold)
5	put m in the cluster c_n
6	**if** the size of cluster $\|c_n\| >$ S (S is the pre-specified threshold)
7	**break** and continue from step 1 with next n
8	**otherwise**, continue from step 2 with next m
9	**end**
	Post-Processing
1	**for** image index $l \in L$ (L is the initial ranked list)
2	find $n = r$
3	**for** index $c \in c_n$
4	place c just after l
5	**if** $\|\hat{L}\| =$ E (E is pre-specified length of list to be evaluated)
6	**break** and continue from step 1
7	**otherwise**, continue from step 3
8	**end**

3 Experimental setup and results

3.1 Task and test collection

Our experimental task was ad-hoc image retrieval based on text: the document collection to be searched was known, but query topics were unknown to the system in advance. In addition, the collection was linguistically heterogeneous or multi-lingual. The goal was to find as many relevant images as possible in a given document collection.

We used the ImageCLEFphoto 2006 test collection. It consisted of the document set, the query set, and the relevant assessment to the queries. The document set is called the IAPR TC-12 Benchmark that contains $20,000$ photos taken at various locations around the world. Subjects of the pictures in this collection include accommodation, facilities, and ongoing social projects. The illumination, viewing angle and background in each image varies, even when taken of the same content. Details of the data creation process are given elsewhere [3]. Each annotation has seven fields, but only the title and description fields corresponding to the image contents were used in the current experiment.

To simulate users' search needs, a total of 60 topics were provided. Half of the topics were semantic, 20 were neutral, and 10 were visual. A title field that described the search topic in a few words, either in English or German, was used as a query. The complete list of topics can be found in the first table of the Appendix. A list of relevant images was provided for each search topic, and its size differed from topic to topic.

3.2 Cross-language experiment

Baseline runs conducted involved cross-language retrieval against linguistically homogeneous collections. The query and collection languages were English and German. When the query language and the collection language were different, we employed a query translation procedure to enable lexical matching. The Systran machine translation (MT) system[2] was applied to queries.

The results of these runs are summarized in Table 2. The mean average precision (MAP) scores were used to evaluate performance. The average precision is the mean ratio of relevant documents at each occurrence of relevant document to the total number of documents from the top list. The MAP score summarizes the average precision over all 60 topics. The collection language in the baselines runs was used to determine retrieval performance, as shown in the table. The results show that searching in the English collection was better for both query languages. The translated queries from German to English in the English collection performed better than mono-lingual German queries in the German collection.

[2] http://babelfish.altavista.com/

Table 2. Summary of runs using linguistically homogeneous collection (MAP scores).

Query Language	English Collection	German Collection
English	0.1193	0.0634
German	0.1069	0.0892

3.3 Multilingual experiment

The advantage of using visual similarity based pre-clustering will become clearer when considering the application in linguistically heterogeneous image collections. Without translations, images annotated using different language from the query language can only be accessed using the visual linkages provided by the clustering. To simulate the retrieval of a linguistically heterogeneous collection, instead of viewing the collection as a single bilingual collection of $20,000$ English and German documents, a mixed collection was constructed by taking $10,000$ randomly chosen images from each of the English and German annotations. There was no overlapping of images that came with both English and German annotations. English and German queries without translation were tested on this single collection with or without the micro-clustering pre-processing and the re-ranking post-processing. The threshold $\mathtt{T}$ used to determine nearly identical images was set to 0.9. The cluster size $\mathtt{S}$ was limited to 10. The size of the ranked list $\mathtt{E}$ used for evaluations was $1,000$ in our study. The same configuration was used in all runs.

The generated clusters were small and often contained two images; a cluster being formed by a pair of images. We obtained many quite small yet highly restricted micro-clusters. The unlimited sized cluster had a mean of 12.72, a standard deviation of 43.81, a minimum of 0, a median of 368, and a maximum of 0. Some clusters originally contained more than 100 members but were truncated to $\mathtt{S}=10$. Such non-micro clusters were not considered ideal because when one of their members appeared at the top of the list, the cluster dominated the entire list after re-ranking.

Table 3 lists the experimental results. The results show that there was no improvement with visual knowledge in terms of MAP scores. However, interesting trends can be seen when we look closely at the changes given by the pre- and post-processing steps for the different groups of search topics. We considered three categories of queries. The first one was "Sports", which contains photos of people doing sports or watching sports. The second one was "Building", which contains photos of various types of architecture and the interiors or exteriors of buildings. The third one was "Scene", which contains photos of various natural scenes and natural landmarks, such as deserts and mountains. The actual queries in these categories are listed in the Appendix. Figure 2 shows a plot of the difference in MAP scores between baseline retrieval and retrieval with the data mining processes. The dots represent the value for individual queries, and the crosses represent the mean values in the categories. A 0 value on the vertical axis means that there was no change in performance. If a dot is placed in the positive region, the performance improved for that query, and if a dot is found in the negative region, there was a corruption in the retrieval. As the results show in the figure, the changes of performance vary among different query subjects. Extracted visual knowledge by micro-clustering was helpful for the "Sports" category, did not effect either way the "Building" category, but severely deteriorated the effectiveness of the "Scene" category. The mean MAP changes for these categories were 0.0054, -0.0098, and -0.0510, respectively.

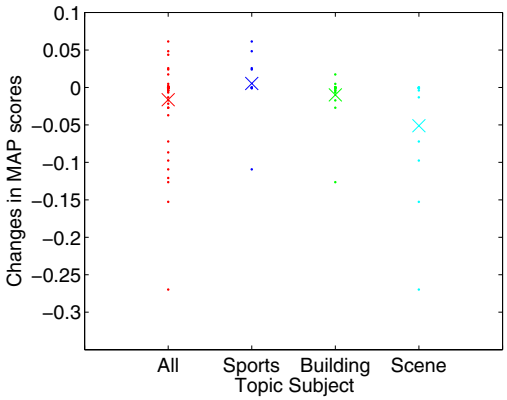

Figure 2. Changes in retrieval performance when clustering pre-processing was applied. Dots represent variation in mean average precision. If dots are placed above zero, results have improved. Crosses indicate the mean of variations for all queries on subjects.

4 Discussion

4.1 Limitations and promises

At this point, improved average performance over all queries was not possible by incorporating visual pre-

Table 3. Summary of runs using linguistically heterogeneous collection (MAP scores).

Query Language	Half English and Half German Collection	
	Without visual knowledge	With visual knowledge
English	0.0838	0.0586
German	0.0509	0.0374

processing. This failure might be because the clusters of topically irrelevant images were used. The baseline IR performance may influence the result. If the initial ranked list contains many irrelevant images in its top-ranked images, post-processing may introduce additional irrelevant images. Examining the influence of the initial quality of ranked lists is the basis of our future work. Also, our visual features and similarity metrics were simple; therefore we might be able to change them so that it only captures the image pairs in the categories, such as the "Sports" category in which pre-processing is proven as useful.

A limitation of using visual information in its current form is that the linkage information obtained is only for part of the collections. Our method affects only top-ranked images having multiple nearly identical images. However, there were many clusters that only contained one image. Thus, we cannot expect significant improvement in performance unless the target collection contains many nearly identical but textually different images. A trade-off exists between the quality of clustering and the degree of expanding the search target, and the threshold we used may have been conservative to avoid including any unwanted noise images. Additional investigation is needed to clarify the effect of threshold values.

A promising fact is that our method improved retrieval performance for some queries despite the above-mentioned limitations. If we can use query classification (e.g., [7]) successfully, it may be possible to use the extracted visual knowledge only on queries that have properties suitable for re-ranking. By not applying re-ranking method for the query categories such as the "Scene" in the current research, the average overall performance will be improved by re-ranking.

4.2 Related methods

Research has been performed in image retrieval that utilizes clustering. For example, Chen et al. applied an image clustering method to present visually similar images as groups rather than a list [2]. Their method is different from our method because clustering was used after querying and not as a pre-processing step for annotation-based retrieval.

Various image retrieval methods have been examined that combine complementary properties of visual information and textual information, mostly for interactive image retrieval (e.g., [4]). As for annotation-based ad hoc retrieval, the similarity between images have been utilized where the visual knowledge is transferred into the word association knowledge [5]. This approach integrates the obtained knowledge into the retrieval model. In contrast, our method separates the knowledge extraction and knowledge injection stages from the core IR processes?. Jing et al. considered the use of visual information in both interaction and pre-processing [6]. However, their pre-processing on visual clustering was used for automatic annotation of images and not for retrieval.

For a multi-lingual document collection, it is possible to use a query translation method. The advantage of using a translation method as compared to our method is that once the query has been translated, users can access all images annotated in the translated language. Our method only allows access to clustered images when one of them is annotated in the query language. A disadvantage of the query translation method is the difficulty in combining the outputs from the retrieval systems for different languages. Integrating multiple rankings into a single rank may require careful weighting. Another disadvantage is errors in machine translations. Inclusion of additional images by machine translation may have adverse effect on retrieval performance for some languages.

5 Conclusion

The use of visual information in annotation-based image retrieval is challenging. We developed a method that uses nearly identical visual knowledge obtained from data mining pre-processing in the re-ranking of retrieval results.

Visual information is independent of languages and can be used to link the images annotated by different languages. Results using the multi-lingual photo collection showed that our method improved the retrieval effectiveness for some queries when they were categorized. More detailed experiments may be needed to verify this finding.

Refinement of our approach may be possible in the following directions: the use of more sophisticated visual features, the use of collection dependent metrics for comparing images, developing more advanced clustering techniques, and making the threshold values in the data mining process adaptive.

Acknowledgment

This research was partly supported by a MEXT Grant-in-Aid for Scientific Research on Priority Areas (Cyber Infrastructure for the Information-explosion Era) and a Grant-in-Aid for Scientific Research from the Japan Society for the Promotion of Science (17700166).

References

[1] Akiko Aizawa. An approach to microscopic clustering of terms and documents. In *Proceedings of PRICAI 2002: Trends in Artificial Intelligence : 7th Pacific Rim International Conference on Artificial Intelligence*, volume 2417 of *Lecture Notes in Computer Science*, pages 404–413, 2002.

[2] Yixin Chen, James Z. Wang, and Robert Krovetz. CLUE: Cluster-based retrieval of images by unsupervised learning. *IEEE Transactions on Image Processing*, 14(8):1187–1201, Aug. 2005.

[3] Paul Clough, Michael Grubinger, Thomas Deselaers, Allan Hanbury, and Henning Müller. Overview of the ImageCLEF 2006 photo retrieval and object annotation tasks. In *CLEF working notes*, Alicante, Spain, September 2006.

[4] Marin Ferecatu, Nozha Boujemaa, and Michel Crucianu. Hybrid visual and conceptual image representation within active relevance feedback context. In *MIR '05: Proceedings of the 7th ACM SIGMM international workshop on Multimedia information retrieval*, pages 209–216, 2005.

[5] Masashi Inoue and Naonori Ueda. Retrieving lightly annotated images using image similarities. In *SAC '05: Proceedings of the 2005 ACM symposium on Applied computing*, pages 1031–1037, March 2005.

[6] F. Jing, M. Li, H.-J. Zhang, and B. Zhang. A unified framework for image retrieval using keyword and visual features. *IEEE Transactions on Image Processing*, 14(7):979–989, July 2005.

[7] Dou Shen, Rong Pan, Jian-Tao Sun, Jeffrey Junfeng Pan, Kangheng Wu, Jie Yin, and Qiang Yang. Q2c@ust: our winning solution to query classification in kddcup 2005. *SIGKDD Explor. Newsl.*, 7(2):100–110, 2005.

[8] Chengxiang Zhai and John Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22(2):179–214, 2004.

Appendix

List of Sports queries

ID	Topic Title
12	people observing football match
14	scenes of footballers in action
18	sport stadium outside Australia
19	exterior view of sport stadia
22	tennis player during rally
23	sport photos from California
27	motorcyclists racing at the Australian Motorcycle Grand Prix
33	people on surfboards
52	sports people with prizes

List of Building queries

ID	Topic Title
1	accommodation with swimming pool
2	church with more than two towers
9	tourist accommodation near Lake Titicaca
13	exterior view of school building
15	night shots of cathedrals
17	lighthouses at the sea
18	sport stadium outside Australia
19	exterior view of sport stadia
21	accommodation provided by host families
24	snowcapped buildings in Europe
28	cathedrals in Ecuador
29	views of Sydney's world-famous landmarks
30	room with more than two beds
50	indoor photos of churches or cathedrals
53	views of walls with unsymmetric stones
54	famous television (and telecommunication) towers
57	photos of radio telescopes

List of Scene queries

ID	Topic Title
6	straight road in the USA
10	destinations in Venezuela
31	volcanos around Quito
36	photos with Machu Picchu in the background
37	sights along the Inka-Trail
38	Machu Picchu and Huayna Picchu in bad weather
40	tourist destinations in bad weather
41	winter landscape in South America
42	pictures taken on Ayers Rock
43	sunset over water
44	mountains on mainland Australia
60	salt heaps in salt pan