

Increasing cluster recall of cross-modal image retrieval*

Simon RÁCZ Bálint Daróczy Dávid Siklósi
Attila Pereszlényi Mátyás Brendel András Benczúr

Data Mining and Web search Research Group, Informatics Laboratory
Computer and Automation Research Institute of the Hungarian Academy of Sciences
{sracz, daroczyb, sdavid, peresz, mbrendel, benczur}@ilab.sztaki.hu

Abstract

We describe our approach to the ImageCLEF Photo and WikimediaMM 2008 tasks. The novelty of our method consists of combining image segment based image retrieval with our text based approach. We rank text hits by our own Okapi BM25 based information retrieval system and image similarities by using a feature vector describing the visual content of image segments. Images were segmented by a home developed segmenter. We use automatic query expansion by adding new terms from the top ranked documents. Queries were generated automatically from the title and the downweighted description words.

Categories and Subject Descriptors

H.3 [Information Storage and Retrieval]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries; H.2.3 [Database Management]: Languages—*Query Languages*

General Terms

Measurement, Performance, Experimentation

Keywords

Cross-modal retrieval, image annotation

1 Introduction

In this paper we describe our approach to the ImageCLEF Photo and WikimediaMM 2008 evaluation campaigns [5]. ImageCLEF Photo is over the IAPR TC-12 benchmark of 20,000 tourist photos [6] and WikimediaMM is over the INEX MM image database which contains approximately 150,000 images that cover diverse topics of interest. These images are associated with unstructured and noisy textual annotations in English. Both campaigns are ad-hoc image retrieval tasks: find as many relevant images as possible from the image collections.

The key feature of our solution in both cases is to combine text based and content based image retrieval. Our method is similar to the method we applied last year for ImageCLEF Photo [1].

*This work was supported by the EU FP7 project JUMAS – Judicial Management by Digital Libraries Semantics and by grants OTKA NK 72845 and NKFP-07-A2 *TEXTREND*.

Our CBIR method is based on segmentation of the image and on the comparison of features of the segments. We use the Hungarian Academy of Sciences search engine [2] as our information retrieval system that is based on Okapi BM25 and automatic query expansion.

2 The base text search engine

We use the Hungarian Academy of Sciences search engine [2] as our information retrieval system. Its ranking algorithm was developed originally for HTML pages. It uses the Okapi BM25 ranking [8] with the proximity of query terms taken into account [7, 3]. We deploy stopword removal and stemming by the Porter stemmer. We extended of stop word list with terms such as “photo” or “image” that are frequently used in annotations but does not have a distinctive meaning in this task.

We considered the text annotation of images including the title, the description and (in case of ImageCLEF Photo) the location separately. The ranking algorithm uses different weights depending on which part of the document contains the query term. Since many topics have location reference, we get the best results if the weight of hits inside the location is much higher than the weights of title and the description.

For queries we use the title and description of the topics with different weight. In addition to stop words we also removed sentences containing the phrase “not relevant”. We apply automatic query expansion based on the method described in [9]. For a given query Q , we ranked every w word in the top 10 documents according to the following formula.

$$Score(Q, w) = \sum_{t_i \in Q} (idf_i \log(\delta + \log(af(w, t_i))) idf_w / \log(n)) \quad (1)$$

1. $af(w, t_i) = \sum_{j=1}^{10} ft_{ij} fw_j$
2. $idf_i = \max\{1, \log((N - N_i + 0.5)/(N_i + 0.5))\}$
3. $idf_w = \max\{1, \log((N - N_w + 0.5)/(N_w + 0.5))\}$

Here ft_{ij} is the number of occurrences of t_i in the j th document and similarly fw_j is the number of occurrences of w in the j th document. At the 2th and 3th lines N is the total number of documents in the collection, N_i is the number of documents containing t_i and N_w is the number of documents containing w . We used a correcting term δ to avoid minus infinity scores.

After these computations we expanded our query with those new w words whose score was above zero and we gave them the weight $(Score(w) + 10)/500$ as query term weight. We also ensured that at most the first 15 words with highest rank got attached to the query.

3 The content based IR system

The task of the CBIR part was to help annotation based retrieval with a content based method. Our method for this year is similar to that of 2007 [1]. The basis of our method is to find segments on the images of the collection similar to the segments in the sample images. We used the Felzenszwalb and Huttenlocher [4] graph based segmentation algorithm with different number of segments for the two tasks that turned out most effective by manual investigation. The number of segments for different runs vary with an extreme case of even a single segment per image corresponding to global similarity measurement. Images were resized to the same size prior to segmentation.

We extracted 20 features for mean color, size, shape information, and histogram information. Our histograms had 5 bins in each channel. In addition we used contrast and 2D Fourier coefficients. Contrast means the maximal and minimal values of the L-channel in HSL color space. The

discrete Fourier transformation was sampled along a zig-zag order, i.e. the low frequency components were included. The DFT features were weighted 10 time higher compared to the other features. Similarity is measured in the above feature space as

$$\text{dist}(S_i, S_j) = d(F(S_i), F(S_j)) : S_i \in S(X), S_j \in S(I) \quad (2)$$

where $S(X)$ is the set of segments of an image X of the collection and $S(I)$ is the set of segments of the sample image I ; d is a distance function in the feature space and F assigns features to image segments.

Given the distance $\text{dist}(S, S')$ of two segments, the distance of image X to sample image I is computed from pairwise distances between pairs of segments $S(X)$ and $S(I)$ of images X and I , respectively. In the base method we averaged over $S_i \in S(I)$ such that for each S_i we took the closest segment from $S(X)$ as

$$\text{dist}(X, I) = \frac{1}{|S(I)|} \sum_j \min_i \{\text{dist}(S_i, S_j) : S_i \in S(X), S_j \in S(I)\}. \quad (3)$$

We introduce another method for computing image distances from segment distances that also takes the relative position of the segments into account. For a pair of images I and X first for all segments S in I the most similar segment $R(S)$ is searched in X . Then we look for the N closest neighbors S_1, \dots, S_N that have common border with S . We define $R(S_i)$ as the segment of X closest in relative spatial position to $R(S)$. The distance is computed as

$$\text{dist}(I, X) = a_1 \cdot \text{dist}(S, R(S)) + a_2 \cdot \frac{1}{N} \sum_{i=1}^N \text{dist}(S_i, R(S_i)) + a_3 \cdot \frac{1}{N} \sum_{i=1}^N \text{dist}^*(S_i, R(S_i)) \quad (4)$$

where dist^* is spatial distance within the image. The weights used are $a_1=1.0$, $a_2=0.8$ and $a_3=0.2$.

When combining TBIR and CBIR we considered TBIR much more reliable than CBIR. Since image distance decreases with relevance, we used CBIR scores by subtracting them from a sufficiently high constant that leaves the rank always positive.

4 The WikiMedia Task

In the WikiMedia Task part of the topics included a sample image. For these topics we combined the text score with a visual score described next. First we resized images to a maximum size of 800x800 by keeping the aspect ratio. We used a global one segment per image method in addition to a medium granularity segmentation with a minimum segment size of 1500 pixels. This latter method resulted in less than 100 segments per image.

The WikiMedia task, with over 100,000 images, already raises scalability issues for a CBIR. The total size of the feature vectors reached 10 Gigabytes, hence pairwise segment similarity scores were computed by a parallel matrix multiplication algorithm for which we utilized a computer cluster. For a single sample image with less than 100 segments, similarity computation with all 100,000 images took over 5 CPU hours by this method. This implies that for a larger collection similarity search data structures will have to be used.

There were several variants of our CBIR method for this task. Due to computational limitations we only used the more complex distance function based on the relative position of the segments. The comparison with the simple averaging method will be performed in the near future.

Variants submitted were named as follows. Since the distance of two images as defined in the previous section is asymmetric, we computed both $\text{dist}(I, X)$ and $\text{dist}(X, I)$. Label *avg* stands for the average of the two while in *avgw* we give 70% weight for the distance of the target image to the sample image $\text{dist}(I, X)$.

We also used global, single segment per image runs labeled *glob10*. Here the Fourier coefficients had weight 10.0 and the other features weight 1.0. The variant *avgw_glob10* combined *avgw* with *glob10* where the later had half the weight of the former.

	MAP	P20	CR20
text0	0.2988	0.3718	0.3592
text0.5	0.2469	0.3179	0.3703
text5.0	0.1739	0.2410	0.3421
text+img+qe0	0.3003	0.3769	0.3560
text+img+qe0.5	0.2495	0.3218	0.3703
text+img+qe5.0	0.1737	0.2423	0.3402

Table 1: Performance of the variants method evaluated by different measures

5 The Photo Task

In the Photo Retrieval Task we used medium sized segmentation as in the WikiMedia task. We only used plain segment distances with no relative position taken into account; a comparison will be performed in the near future. In distance computation we weighted the features as follows. For variant *w5* we had size 1; ratio 3; average color 1; shape 1; contrast 1. And for *w10* size 1; ratio 3; average color 20; histogram 50; shape 50 and contrast 10.

In the Photo Task 3 sample images were given for each topic. Consequently, for each image there are 3 distances. The 3 distances can be combined as average, minimum or maximum. We tried all of them, which is indicated in the name of the variant as *min*, *max* or *avg*.

Best performing pure CBIR MAP score 0.0243 was obtained by *w5_min* with a runner up 0.0212 of *w10_min*. Official runs submitted were unfortunately combined with the *avg* variants that performed below a MAP of 0.003 only. In combination with the TBIR score we were able to improve the MAP from 0.2978 to 0.3014 by *w10_min*. Official runs performed slightly below this value.

5.1 Increasing cluster recall

We modified our method to achieve greater diversity within the top 20. For each topic in the ImageCLEF Photo set, relevant images were manually clustered into sub-topics. Evaluation was based on two measures: precision at 20 and instance recall at rank 20 (also called S-recall) which calculates the percentage of different clusters represented in the top 20.

Our method works as follows. Let $\text{Orig}(i)$ be the i th document ($0 \leq i < 999$) and $\text{OrigSc}(i)$ be the score of this element on the original list for a given query Q_j . We modified these scores by giving penalties to the scores of the documents based on their Kullback-Leibler distance. We used the following algorithm.

1. $\text{New}(0) = \text{Orig}(0)$ and $\text{NewSc}(0) = \text{OrigSc}(0)$
2. For $i = 1$ to 20
 - (a) $\text{New}(i) = \text{argmax}_k \{ \text{CL}_i(k) \mid i \leq k < 999 \}$
 - (b) $\text{NewSc}(i) = \max \{ \text{CL}_i(k) \mid i \leq k < 999 \}$
 - (c) For $\ell = 0$ to $(i - 1)$

$$\text{NewSc}(\ell) = \text{NewSc}(\ell) + c(i)$$

Here $\text{CL}_i(k) = \text{OrigSc}(k) + \alpha \sum_{l=0}^{i-1} \text{KL}(i, k)$. where α is a tunable parameter and $\text{KL}(i, k)$ is the Kullback-Leibler distance of the i th and k th documents. We used a correction term $c(i)$ at Step 2c to ensure that the new scores will be also in descending order.

5.2 results

Table 1 shows the results of the text based and the mixed method. The results were evaluated by the ImageCLEF organizers using the following measures: Mean Average Precision (MAP), Precision at 20 (P20) and Cluster Recall at 20 (CR20).

The number in the names indicate the value of α from the previous section. As it can be seen the algorithm used to increase CR20 was successful with $\alpha = 0.5$. The cost of increasing CR20 was worse MAP and P20. The value $\alpha = 5.0$ turned out to be too much. Using CBIR and query expansion improved the algorithm a little, and the same trend is true for α .

Unfortunately at the moment we do not have evaluations for text+qe and text+image without qe to check which part is responsible for the improvement.

Conclusion and future work

We have demonstrated that image based retrieval based on segmentation can improve the performance of an IR system. In future work we will conduct a more thorough comparison of the different CBIR techniques introduced. We also plan to improve on the performance of query expansion.

References

- [1] András Benczúr, István Bíró, Mátyás Brendel, Csalogány Károly, Bálint Daróczy, and Dávid Siklósi. Cross-modal retrieval by text and image feature biclustering. In *Working Notes for the CLEF 2007 Workshop*, Budapest, Hungary, 2007.
- [2] András A. Benczúr, Károly Csalogány, Eszter Friedman, Dániel Fogaras, Tamás Sarlós, Máté Uher, and Eszter Windhager. Searching a small national domain—preliminary report. In *Proceedings of the 12th World Wide Web Conference (WWW)*, 2003.
- [3] Stefan Büttcher, Charles L. A. Clarke, and Brad Lushman. Term proximity scoring for ad-hoc retrieval on very large text collections. In *SIGIR '06*, pages 621–622, New York, NY, USA, 2006. ACM Press.
- [4] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59, 2004.
- [5] Michael Grubinger, Paul Clough, Allan Hanbury, and Henning Müller. Overview of the Image-CLEFphoto 2007 photographic retrieval task. In *Working Notes of the 2007 CLEF Workshop*, Budapest, Hungary, September 2007.
- [6] Michael Grubinger, Paul Clough, Henning Müller, and Thomas Deselears. The IAPR TC-12 benchmark - a new evaluation resource for visual information systems. In *OntoImage*, pages 13–23, 2006.
- [7] Yves Rasolofo and Jacques Savoy. Term proximity scoring for keyword-based retrieval systems. In *ECIR*, pages 207–218, 2003.
- [8] Stephen E. Robertson and Karen Sparck Jones. Relevance weighting of search terms. In *Document retrieval systems*, pages 143–160. Taylor Graham Publishing, London, UK, UK, 1988.
- [9] J. Xu and W.B. Croft. Query expansion using local and global document analysis. *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 4–11, 1996.