Multimodal Retrieval by Text–Segment Biclustering*

András Benczúr, István Bíró, Mátyás Brendel, Károly Csalogány, Bálint Daróczy, and Dávid Siklósi

Data Mining and Web search Research Group, Informatics Laboratory
Computer and Automation Research Institute of the Hungarian Academy of Sciences
{benczur,ibiro,mbrendel,cskaresz,daroczyb,sdavid}@ilab.sztaki.hu
http://datamining.sztaki.hu/

Abstract. We describe our approach to the ImageCLEFphoto 2007 task. The novelty of our method consists of biclustering image segments and annotation words. Given the query words, it is possible to select the image segment clusters that have strongest cooccurrence with the corresponding word clusters. These image segment clusters act as the selected segments relevant to a query. We rank text hits by our own tf.idf-based information retrieval system and image similarities by using a 20-dimensional vector describing the visual content of an image segment. Relevant image segments were selected by the biclustering procedure. Images were segmented by graph-based segmentation. We used neither query expansion nor relevance feedback; queries were generated automatically from the title and the description words. The later were weighted by 0.1.

1 Introduction

In this paper we describe our approach to the ImageCLEF Photo 2007 evaluation campaign [1].The key feature of our solution is to combine text based image retrieval and content based image retrieval introducing biclustering algorithm of image segments and annotation words to form interrelated clusters. Our CBIR method is based on the segmentation of the image and on the comparison of features of the segments. The biclustering algorithm is used to filter out irrelevant segments. The text retrieval system is described in [2] with two differences in using:

- A wider set of stop words including “photo”, “image” etc.;
- Heavy weight to hits in the location field but using e.g. South as location stop word.

Query terms from the topic title have ten times higher weight than the narrative terms, however, sentences containing phrase “not relevant” were automatically removed.

As our main result, we demonstrate that biclustering of image segments and annotation words additively improves retrieval performance by over 2%. In future work query expansion and feedback will be used to test whether the method can improve performance over the state of the art.

* This work was supported by a Yahoo! Faculty Research Grant and by grants *MOLINGV* NKFP-2/0024/2005, NKFP-2004 project Language Miner.

2 The Content-Based Information Retrieval System

Our CBIR system relies on so called blobs, regions or segments similar to for example those of [3–6]. For each topic three sample images were given; we used the minimum of their distances from the target image for ranking. Distances were computed based on the segments of the target and sample image; for segmentation we used the code of the Felzenszwalb and Huttenlocher [7] graph-based method.

First we describe how we measure the distance between segments. For each segment we use a minimalistic 20-dimensional real valued feature vector and Euclidean distance after normalization. Out of the 20 values, 15 consist of histograms with 5 bins in each of the RGB channel and an additional 3 values contain the average intensity. The single shape information consists of the ratio of the logarithm of segment width and height. Finally the last value is the logarithm of the size in pixels.

Given the distance $\mathrm{dist}(S, S')$ 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. Since segments of I are considered as the description of the search goal, we averaged over $S_i \in S(I)$ such that for each S_i we took the closest segment from $S(X)$ as

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

3 Image Segment – Annotation Word Biclustering

Our method is special in using the annotation text to guide the CBIR via biclustering, a technique used in a wide variety of applications [8]. Our assumption is that biclusters indicate connection between the features and the text such as blue color and "pool", white color and "snow", black and white histogram and "black and white". This can be used to select relevant segments of the sample image. Hence we compute an interrelated segment and word clustering together with a weight for each pair of a segment and a word cluster.

The output of segment–word biclustering is used to refine the CBIR method. In equation (1) we use only those segments of the three sample images where there is a topic title word in a text cluster with large weight for the given segment cluster. In the rare case when none of the segments is selected, we keep all as a fallback mechanism.

We used the biclustering algorithm of [8]. Let X and Y be discrete random variables that take values in the sets $\{\text{segments}\}$ and $\{\text{annotation words}\}$ respectively. Let $p(X, Y)$ denote the joint probability distribution of X and Y. Let the k clusters of X be $\{\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_k\}$, and let the ℓ clusters of Y be $\{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_\ell\}$. We are interested in finding maps C_X and C_Y,

$$C_X : \{x_1, x_2, \ldots, x_k\} \mapsto \{\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_k\},\ C_Y : \{y_1, y_2, \ldots, y_\ell\} \mapsto \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_\ell\}. \tag{2}$$

For brevity we write $\hat{X} = C_X(X)$ and $\hat{Y} = C_Y(Y)$ where $\hat{X}$ and $\hat{Y}$ are random variables that are a deterministic function of X and Y, respectively. Finally let $D(p \parallel q)$ denote the *Kullback–Leibler* divergence of probability distributions p and q.

Table 1. Comparison of performance of various methods evaluated by different measures.

	MAP	P10	P20	P30	BPREF	GMAP	manual MAP
text + visual + bicluster	0.2238	0.3283	0.2875	0.2556	0.2003	0.0449	0.2545
text + visual	0.2076	0.3183	0.2683	0.2372	0.1924	0.0419	0.2441
text only	0.2020	0.3033	0.2492	0.2200	0.1747	0.0463	0.2295
visual + bicluster only	0.0138	0.0467	0.0433	0.0367	0.0240	0.0019	
visual only	0.0129	0.0683	0.0400	0.0317	0.0427	0.0021	

Table 2. Performance of best method as function of the CBIR weight in ranking.

weight of image	1	10	100	1000	2000	5000	10000
MAP	0.2146	0.2152	0.2151	0.2238	0.2120	0.2027	0.1951

The algorithm of [8] iterates between computing segment (row) and word (column) clusters. As in iteration t of [8], the new cluster index of word y becomes

$$C_Y^{(t+1)}(y) = \text{argmin}_{\hat{y}} \; D\left(\frac{p(X,y)}{p(y)}\middle\|p(X)\cdot\frac{p(\hat{x},\hat{y})}{p(\hat{x})\cdot p(\hat{y})}\right), \tag{3}$$

resolving ties arbitrarily. We slightly modify this procedure for computing the new cluster index of segment x by using the 20-dimensional segment feature vector $f_1(x)$, ..., $f_s(x)$. We combine the Kullback-Leibler distance over the word incidence matrix with Euclidean distance in 20 dimensions as

$$C_X^{(t+1)}(x) = \underset{\hat{x}}{\text{argmin}}\left\{D\left(\frac{p(x,Y)}{p(x)}\middle\|p(Y)\cdot\frac{p(\hat{x},\hat{y})}{p(\hat{x})\cdot p(\hat{y})}\right) + \sqrt{\sum_{i=1}^{s}\left(f_i(x)-f_i(\hat{x})\right)^2}\right\} \tag{4}$$

where $f_i(\hat{x})$ is the cluster average. We resolve ties arbitrarily.

4 Results

Table 1 shows the results of the text based, content based and the mixed method in our original submission. Visual only results are very poor; however when combined with text, our CBIR yields significant improvements in all measures. Surprisingly, our CBIR improves more over text only retrieval than its performance when used alone. This fact is further justified by using manually constructed text queries for the worst performing 6 topics (last column of Table 1). Table 2 shows the performance in function of the weight of the image based method when combining it with the text based query.

Table 3 shows a detailed analysis of the text only method for the topics. Best performances (57: "radio telescope", 21: "accommodation, host family", 10: "destination,

Table 3. Performance of text only method, some selected topics.

topic	57	21	10...	...15...	...11...	...41	30	18
MAP	0.9650	0.9306	0.7852	0.4563	0.3119	0.00	0.00	0.00

Table 4. Some of the topics with the best improvement and worst deterioration when adding image similarities (top), when, in addition, using segment selection by biclustering (middle) and the overall visual improvement (bottom).

topic	11	27	15	6	17	53	32	43	48	10	8
MAP improv. by image	0.26	0.13	0.05	0.06	-0.08	-0.12	-0.16	0.08	0.12	-0.14	0.08
additional improv. by bic	0.31	0.14	0.14	0.06	0.17	0.18	0.16	-0.07	-0.16	0.06	-0.19
total MAP improvement	0.57	0.27	0.19	0.12	0.11	0.06	0.0	0.01	-0.04	-0.09	-0.11

Venezuela") are achieved when title contains specific words that match the annotation style. Worst performance corresponds to the need for either expanding terms (41: "South America", 18: "outside Australia" – the latter easily solved manually by negation) or understanding the visual semantics of the topic (30: "more than two beds").

Image content with biclustering increases performance by more than 2% in average, including some topics with large improvement and only a slight deterioration for others as seen in table 4. Biclustering only sporadically deteriorates the CBIR performance. We see the largest improvement from topics where the content-based feature is related to color histogram such as "black and white" in topic 11 or "night shots" in 15. We see improvements with different explanation as well (27: "motorcyclist racing", 6: "straight road") that must have also utilized certain semantical image content amplified in addition by biclustering. As an interesting example, Topic 53 "asymmetric stones" has a deterioration of 0.12 by visual similarity but an improvement of 0.18 by biclustering that sums up to +0.06 with a possible reason that biclustering removes segments belonging to the people in one of the sample images.

References

1. Grubinger, M., Clough, P., Hanbury, A., Müller, H.: Overview of the ImageCLEF 2007 photographic retrieval task. In: Working Notes of the 2007 CLEF Workshop. (2007)
2. Schönhofen, P., Benczúr, A., Bíró, I., Csalogány, K.: Cross-language retrieval with wikipedia. In this volume (2007)
3. Chen, Y., Wang, J.Z.: Image categorization by learning and reasoning with regions. J. Mach. Learn. Res. **5** (2004) 913–939
4. Prasad, B.G., Biswas, K.K., Gupta, S.K.: Region-based image retrieval using integrated color, shape, and location index. Comput. Vis. Image Underst. **94**(1-3) (2004) 193–233
5. Carson, C., Belongie, S., Greenspan, H., Malik, J.: Blobworld: Image segmentation using expectation-maximization and its application to image querying. IEEE Trans. Pattern Anal. Mach. Intell. **24**(8) (2002) 1026–1038
6. Lv, Q., Charikar, M., Li, K.: Image similarity search with compact data structures. In: CIKM '04: Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management, New York, NY, USA, ACM Press (2004) 208–217
7. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient graph-based image segmentation. International Journal of Computer Vision **59** (2004)
8. Dhillon, I.S., Mallela, S., Modha, D.S.: Information-theoretic co-clustering. In: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. (2003) 89–98