Multimodal Image Retrieval over a Large Database

Débora Myoupo*, Adrian Popescu**, Hervé Le Borgne*, and Pierre-Alain Moëllic*

*CEA, LIST, Laboratoire d'ingénierie de la connaissance multimédia et multilingue
F-92265 Fontenay-aux-Roses, France.
**Computer Science Dept., Télécom Bretagne
debora.myoupo@gmail.com,adrian.popescu@telecom-bretagne.eu,
{herve.le-borgne,pierre-alain.moellic}@cea.fr,

Abstract. We introduce a new multimodal retrieval technique which combines query reformulation and visual image reranking in order to deal with results sparsity and imprecision, respectively. Textual queries are reformulated using Wikipedia knowledge and results are then reordered using a k-NN based reranking method. We compare textual and multimodal retrieval and show that introducing visual reranking results in a significant improvement of performance.

1 Introduction

The combination of textual and visual retrieval techniques has been studied in a number of existing works [11, 5, 8] but it remains a stimulating research area, particularly when applied to large scale datasets. Some open questions include: how to deal with difficult queries? What should be the text and image features in multimodal approaches? How to design retrieval frameworks which are fast enough for real-time search? In this paper, we present techniques which propose possible answers to these questions. To answer the first question, we employ a query expansion technique which draws on our work in [8]. The categorical structure of Wikipedia is exploited in order to find and rank concepts from the encyclopedia which are semantically similar to the initial query. While we agree that textual queries are the main way to access Web images, we advocate for a more central role of image processing. Image reranking, built around a visual model of a query and an external class, is applied to query results before displaying them. We verify our assumptions on the Wikipedia collection, a large image dataset [10] and present results which show that multimodal search outperforms textual search while introducing a small computational overload.

With over 3,000,000 articles in its English version, Wikipedia is a rich resource and is used in a variety of research tasks, such as: sense disambiguation, ontology extraction or semantic relatedness. The last problem can be formulated as follows: given an input (a concept or a larger text), find the concepts which are most closely related to the input. Wikipedia based techniques to find

semantic relatedness include WikiRelate! [9], Explicit Semantic Analysis (ESA) [4] and Wikipedia Link-based Measure (WLM) [6]. WikiRelate! modifies techniques previously applied to WordNet in order to suit Wikipedia's structure. The authors of [4] map queries to Wikipedia concepts representation in order to find related concepts. ESA is interesting because it finds related concepts for any given query and not only for mono-conceptual queries and is thus suited for use in Web information retrieval. WLM exploits only Wikipedia links to find related concepts. When compared [6], ESA achieves the best performances, followed by WLM and WikiRelate!. However, WLM is considered as computationally simpler than ESA.

Image reranking can be performed using textual information associated to images, visual description or a combination of both. Here, we consider the case of visual reranking applied to textual query results. In [5], the authors adapt the PageRank algorithm to image retrieval in order to find authority nodes in a visual similarity graph. Both homogeneous and heterogeneous visual concepts are discussed but the approach is only tested on product images and it largely outperforms the Google standard search. Van Leuken et al. [11] propose techniques for diversifying image search results based on visual clustering. Clustering is applied to both ambiguous and non-ambiguous queries and it is evaluated against manually clustered search results. Tests show that the approach tends to reproduce manual clustering in a majority of cases. Deselaers et al. [3] discuss the joint optimization of search precision and diversity, with a focus on diversity. They implement a dynamic programming algorithm applied on top of a greedy selection and test their approach on a heterogeneous test database (ImageCLEF 2008 photo retrieval task [1]). An improvement of diversity, accompanied by a small precision loss is reported when comparing results to ImageCLEF runs. The authors of [7] implement a shared nearest neighbors algorithm (s-NN) which clusters both tags and visual content for any given query. Unfortunately, the technique in is not fast enough to be performed at query time. Though effective, techniques like s-NN [7] or dynamic programming [3] are computationally expensive and are hard to apply under real time constraints. Visual clustering and visual reranking serve the same purpose: surfacing relevant (and diversified) images based on their visual properties. When visual clustering is applied, diversity is obtained by selecting images from different clusters whereas in our approach diversity is handled using query expansion.

2 Retrieval method

Our retrieval method has two steps: the first one is a text-based retrieval, with automatic query expansion whereas the second exploits the visual properties of a query to improve results of the text search.

2.1 Conceptual neighbourhood building

The main resource we exploited was Wikipedia, which provides its dumps for free use. We download the April 2009 English dump, which contains over 2.6 million

articles and is provided as a single XML file. Next, we split the dump into individual articles in order to process the information faster. The information in Wikipedia covers large number of conceptual domains, with a high number of articles describing known people, places, entertainment and organisations. Hereafter we use concept to denote Wikipedia article titles.

Wikipedia images come with brief textual descriptions. Query expansion is an appealing way to improve recall and, if performed in a judicious way, to also improve results precision. Topics are preprocessed in order to eliminate stop words and visual terms from a closed list (including image(s), photograph(s) etc.). Then, we lemmatize remaining terms and arrange them using their term frequency in Wikipedia in order to favor rare terms (which are more likely to be discriminant than frequent words). Also, when a non-ambiguous Wikipedia concept is found in a query, we consider that all Wikipedia redirects toward the concept's article are synonyms. Finally, we compare terms in the query to Wikipedia categories and retain articles that match at least one term. A limit of 5000 articles is imposed to speed up processing.

The relatedness of the 5000 discovered concepts to the initial query is variable and we need to rank them by pertinence. We rank concepts by counting the number of terms from the initial query which appear in each article's categories and by favoring related concepts which match rare terms in the query. At this point, there are usually several concepts with the same score and in order to differentiate them we refine the ranking by answering the following questions:

- is the concept ambiguous?
- do all terms in the initial query appear in the first paragraph of the Wikipedia article?
- do all terms in the initial query appear in the article's text?

Concepts are considered ambiguous if their Wikipedia article contains links to disambiguation pages. The refined list of related concepts will favor unambiguous elements and concepts which contain all terms in the query either in the first paragraph of article (which is often a definition) or in the remaining text. When ties appear, they are broken by counting the total number of query terms in the article. Similarly to [4] or [6] our method finds semantically similar concepts from Wikipedia for an input text. However, whereas the relations in [4] or [6] are untyped, our method will primarily find concepts which are related to (a part of) the query via hyponymy. In image retrieval, such relations are preferable because specific concepts illustrate well more generic ones.

3 Textual retrieval

Given a query, image results from the Wikipedia collection are retrieved by searching for images which are described either by terms in the initial query or by related concepts from Wikipedia. We assume that the relatedness of an image to a query is proportional to the number of terms/concepts which describe it and are also in the query (in its extended form). Therefore, relevant images are

found by launching queries with the initial terms and the expanded queries in the following order:

- all terms in the initial query and a related concept
- the initial query
- parts of the initial query (starting with largest subparts - and favoring rare terms) and a related concept
- related concept or parts of the initial query

The above types of queries are called blocks. In order to differentiate between images in the same block, weights are applied to the terms in the initial query (rare terms are favored) and to related concepts. Individual image scores are then calculated by adding these scores. For instance, if a query contains three terms, and two images are annotated with two different terms out of three, the one which is annotated with the rarest terms is favored. Since blocks are ordered by importance, an image is retained only once, that is why they first appear. The result of textual retrieval is noted $\mathcal{R}_t$.

4 Visual reranking

We introduce a reranking method which is computationally inexpensive if the images are preindexed and which is based on a contrastive model. We assume that an image which is visually close to a visual model of a query is more likely to be a good answer than another image which is less similar to the visual model. To evaluate the similarities between the $\mathcal{R}_t$ images and a topic, we need to obtain a low-level description of that particular topic. We create a *visual model* (a *positive set* of images $\mathcal{R}_{pos}$ which depicts the topic) using Web images. A *negative set* $\mathcal{R}_{neg}$ containing diversified images is constructed and used as an outlier for all topics in order to discard images which are not visually close to the topic's visual model. Then we use the *visual coherence* to evaluate the relevance of both sets ($\mathcal{R}_{all} = \mathcal{R}_{pos} \cup \mathcal{R}_{neg}$) and rerank them. Finally, we rerank the images in each block generated by the textual retrieval using their visual similarity to the query.

4.1 Visual coherence

The *visual coherence* (VC) of an image is a metric measuring its relatedness to a visual model of the query. This metric is computed using a *positive set*, containing N_{pos} topic images and a *negative set* $\mathcal{R}_{neg}$ of N_{neg} non relevant images. We compute the *visual coherence score* based on the following scores :

False neighbours: For an image, we search for its N_{neigh} closest neighbours in $\mathcal{R}_{pos}$ as well as its N_{neigh} closest neighbours in $\mathcal{R}_{neg}$. Elements of the two lists of N_{neigh} neighbours are ordered using the euclidean distance in the feature space. The first part of the VC score is defined as the number of neighbours which belong to the negative set among the first N_{neigh} images of this $2 \times N_{neigh}$ size list. Ideally, very good depictions of the concept will

not have any pictures from $\mathcal{R}_{neg}$ among their N_{neigh} nearest neighbours. Inversely, the noisy images should have a lot of neighbors from $\mathcal{R}_{neg}$ among their neighbours.

Distances: The second score is the sum of the distances of their N_{sum} closest neighbours in $\mathcal{R}_{pos}$. A small value of this sum implies that the image is visually similar to the concept in the descriptor's space. The distances computation is based on the same descriptor as in the previous step.

To rank a set of images according to their visual relatedness to a query, we sort them using only the first score (in $0 \dots N_{neigh}$). If two images have the same number of neighbors, we use the second score of the visual coherence to refine the ranking. The algorithm depends primarily on the low-level descriptor.

4.2 Visual model creation

We first download a set of N_{neg} which constitute the *negative set* $\mathcal{R}_{neg}$. Since the *negative set* will be used to rerank images for diversified topics, it needs to be itself diversified. To insure diversity, a large number of concepts from different conceptual domains, such as *mountain*, *dog*, *car*, *football* or *protest* are manually selected and represented in the *negative set* in order to depict the noise that is to say everything except the WikipediaMM topics. Then, we query the Web for with each WikipediaMM topic, download N_q raw images and index them with low-level descriptors. For the visual model to be effective, the images in the *positive set* need to be as accurate as possible and it is necessary to filter out noisy results returned by the Web search engine. To find relevant images, we compute the *VC* on the raw set of topic images(for each image, the N_q-1 other images temporary constitutes $\mathcal{R}_{pos}$) and keep the top N_{pos} results only. Since the visual coherence computation depends on the features, a visual model is defined for each low-level descriptor. The visual model of a query is composed of an ordered list of N_{pos} *positive images* and a set of N_{neg} *negative images* and can be used for image reranking.

4.3 Textual results reranking

Let $\mathcal{R}_t$ be the list of images returned by the textual matching procedure. It can now be reordered using the visual model using the same k-NN method used to build the model itself. We compute the signature of each image from $\mathcal{R}_t$ and its distance to all the images of the visual model. Our assumption is that a relevant result from the textual run will be related to some of the images composing the visual model. We run experiments using:

Single Descriptor Reranking We use the visual model created with one descriptor to compute the visual coherence scores of the $\mathcal{R}_t$ list. Then, the list is simply ranked according to the VC score.

Late-fusion Reranking The picture is reranked according to the sum of its ranks into the lists coming from several Descriptor Rerankings (i.e. using several descriptors).

Results of the textual retrieval have weights which express their relatedness to the query. These relatedness is used to define three blocks of answers by decreasing relevance to the topic.

5 Experimental validation

We evaluate our multimodal retrieval method on the WikipediaMM collection [10]. For the visual models creation, we used both Google and Yahoo! image search engines to improve the model variety. We kept the top 50 pictures of each search engine results, leading to a raw set of $N_q = 100$ images, then a new positive set containing $N_{pos} = 50$ images. Ideally, the *negative set* should be redefined for each query in order to avoid overlaps. However, since the negative set contains few images per concept, the overlap with query images is always small or null and the same *negative set* can be used for all queries. We fixed the number of negative examples (N_{neg}) to 300.

To compute the visual coherence score, N_{neigh} and N_{sum} are both fixed to 10 : the diversity of a concept can make two pictures semantically relevant but different from a visual point of view and it can be interesting to experiment with different low-level features. We experimented with a global descriptor (the color and texture based Local Edge Pattern (LEP), derived from [2]) and a local descriptor (bag of SIFTs similar to [12]). Hence, for each query, we obtain three visual models : LEP visual model, Bag of feature visual model and the Early-fusion visual model which results from the mix of the previous two. Finally, we get four content-based rerankings : the texture LEP reranking, the Bag of features reranking, the Late-fusion reranking .The experimental results are reported in table 1 [1].

Table 1. Evaluation results for the different methods, on the WikipediaMM collection.

Run	Reranking procedure	MAP	P@10	P@20
cealateblock	$\mathcal{R}_t$ textual ranking + Late-fusion reranking	0.2430	0.4000	0.3022
ceabofblock	$\mathcal{R}_t$ textual ranking + Bag of features reranking	0.2192	0.3867	0.2989
ceatlepblock	$\mathcal{R}_t$ textual ranking + Texture LEP reranking	0.2118	0.3511	0.2811
ceatxt	- none -	0.1870	0.2689	0.2267

The results in table 1 show that our multimodal retrieval technique is more effective than a textual search strategy. When looking at MAP scores, the improvement is 0.025 for a visual reranking with global features, 0.032 for a visual reranking with local descriptors and 0.056 for a late fusion of decriptors. The advantages of multimodal search are even more salient if we look at P@10 or P@20. For P@20, the textual retrieval score is 0.2267 whereas the late fusion score is

[1] Although the method was similar, official ImageCLEF 2009 results were hampered by bugs which were later corrected.

0.3022. Precision is particularly relevant in Web image search, where users often look only at the first page of results and neglect the other results pages. Our visual reranking method is appropriate for Web image retrieval because it maximizes the quality of top results.

Visual reranking results show that local descriptors are slightly better than global ones (MAP 0.2192 vs. 0.2118) on the query set. More importantly, late fusion results (MAP 0.2430) are significantly better than local or global reranking results and lead us to conclude that descriptors fusion is the best reranking strategy among those tested here. An explanation of the obtained results is that the 45 queries in Wikipedia MM 2009 are diversified and no single descriptor is fitted for all queries.

Our best MAP result (0.2430) is very close to the best of the campaign WikipediaMM 2009 [10]. A qualitative comparison to the best systems highligth the weakness of the pure textual retrieval of our method. However, there is no insurance that the visual reranking would be as efficient as in our case when the initial textual results are better. Document expansion and textual reranking are not included in our current approach but they can be easily integrated and this integration would probably bring further improvements.

The use of the content-based reranking (i.e the second step) improves the results compared to the text-based baseline. However, we noticed during our preliminary experiments that the results can decrease when the reranking is global (i.e without keeping a relative order according to each textual block). We expected such results because we retain up to 1000 results for each topic and the test collection contains around 150000 images. Consequently, a large part of the 1000 answers are not relevant and it is useful to exploit the textual ranking.

Late-fusion reranking significantly improves MAP results (around 2 points) in comparison to simple descriptor reranking. It confirms that when dealing with diversified topics, it is interesting to merge local and global descriptors in order to obtain better results. During preliminary works, some "early-fusion reranking" schemes were tested, leading to quite similar results.

6 Conclusion and future work

We proposed a new multimodal retrieval technique which combines query reformulation and visual image reranking. It is both effective and can easily be scaled-up to larger image repositories. The method has two main steps consisting in retrieving a list of images using the textual information only, then reranking it using the visual information. Textual queries are reformulated using Wikipedia knowledge and results are then reordered using a k-NN based reranking method with a visual query model. We compared textual and multimodal retrieval and showed that the second step lead to a significant improvement of performances. The late fusion method we proposed had a mean average precision (0.2430) comparable to the best score obtained during the official campaign on the large corpus (150,000 images) of the Wikipedia task [10] at ImageCLEF 2009.

One interesting direction for future work is to replace the refinement part of the concept ranking with techniques such as explicit semantic analysis [4] and to assess the impact of this new concept ranking on the performances of the system. We will also test the effects of the query expansion on larger datasets (such as the Web corpus) in order to compare it to standard search engine results. We are confident that results are likely to improve in terms of precision and diversity for queries that map well on Wikipedia categories because Wikipedia related concepts cover various aspects of a topic.

7 Aknowledgement

We thank the DGCIS for funding us through the regional business cluster Systematic (project POPS) and Cap Digital (project Mediatic and Romeo), and the french National Agency for Research (ANR) project Georama.

References

1. T. Arni, P. Clough, M. Sanderson, M. Grubinger. Overview of the ImageCLEFphoto 2008 photographic retrieval task. In CLEF 2008 Workshop Working Notes, 2008
2. Y.-C. Cheng, S.-Y. Chen. Image classification using color, texture and regions. In Image Vision Computing, 2003
3. T. Deselaers, T. Gass, P. Dreuw, H. Ney. Jointly Optimising Relevance and Diversity in Image Retrieval. In Proc. of CIVR, 2009
4. E. Gabrilovich, S. Markovich. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In Proc. of IJCAI, 2007
5. Y. Jing, S. Baluja. VisualRank: Applying PageRank to Large-Scale Image Search. PAMI, 2008
6. D. Milne, I. H. Witten. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In Proc. of WIKIAI, 2008
7. P.-A. Moëllic, J.-E. Haugeard, G. Pitel. Image clustering based on a shared nearest neighbors approach for tagged collections. In Proc. of CIVR, 2008
8. A. Popescu, H. Le Borgne, P.-A. Moëllic, Conceptual Image retrieval over a Large Scale Database. In Evaluating Systems for Multilingual and Multimodal Information Access, In Proc. of the 9th Workshop of the Cross-Language Evaluation Forum, Lecture Notes in Computer Science, 2009
9. M. Strube, S. P. Ponzetto. WikiRelate! Computing Semantic Relatedness Using Wikipedia. In Proc. of AAAI, 2006
10. T. Tsikrika, J. Kludas. Overview of the wikipediaMM task at ImageCLEF 2009. CLEF working notes, 2009
11. R. H. Van Leuken, L. Garcia, X. Olivares, R. van Zwol. Visual Diversification of Image Search Results. In Proc. of WWW, 2009
12. J. Zhang, M. Marszalek, S. Lazebnik, C. Schmid. Local features and kernels for classifcation of texture and object categories: An in-depth study. Technical Report RR-5737, INRIA, 2005