Conceptual Image Retrieval over a Large Scale Database

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.
adrian.popescu@telecom-bretagne.eu,
{herve.le-borgne,pierre-alain.moellic}@cea.fr,

Abstract. Image retrieval in large-scale databases is currently based on a textual chains matching procedure. However, this approach requires an accurate annotation of images, which is not the case on the Web. To tackle this issue, we propose a reformulation method that reduces the influence of noisy image annotations. We extract a ranked list of related concepts for terms in the query from WordNet and Wikipedia, and use them to expand the initial query. Then some visual concepts are used to re-rank the results for queries containing, explicitly or implicitly, visual cues. First evaluations on a diversified corpus of 150000 images were convincing since the proposed system was ranked 4^{th} and 2^{nd} at the WikipediaMM task of the ImageCLEF 2008 campaign [1].

Key words: image retrieval, large-scale database, query reformulation.

1 Introduction

Existing Web-scale image search engines consider the text found *around* the images (caption, HTML tags...) as a relevant description to describe them, and thus match the query to those terms to propose results. The main advantages of this approach are its computational tractability and its applicability to large volume of data. Unfortunately, the descriptive text is often unrelated to image content and leads to an important imprecision of results. Query ambiguity is another important noise source. For instance, the word *bridge* can refer to the structure or to the card game, and the expected results are completely different for the two meanings. The use of semantic structures is a possible solution to cope with such problems, as long as these structures can cover the query space. We propose to expand the queries using conceptual relations from a prebuilt large-scale semantic structure, a process that enhances results and requires little computational overload.

Semantic structures, such as WordNet [2] were already used in image retrieval [3] but they do not ensure a sufficient coverage of the query space. For instance WordNet includes only few artefact instances for each concept (e.g. there is

* Adrian Popescu is currently with the Computer Science Dept., Télécom Bretagne

no WordNet entry for *Peugeot*) and these instances are popular Web queries. Wikipedia is a rich source of semi-structured information and has already been used to structure large quantities of knowledge [4, 5]. [5] proposed a method to clean the categorical tree of Wikipedia in order to obtain a sound taxonomy. Kazama et al. [6] successfully extracted *IsA* relations from the first sentence of articles using a syntaxic analysis. [7] explored the automatic enrichment of WordNet using Wikipedia content. They extract hyponymy, hyperonymy, holonymy and meronymy relations based on lexical patterns learned from a text corpus. The overall precision of the extraction process exceeds 50%, leaving a lot of incorrect relations in the extracted structure. DBPedia [4] is a translation of parts of Wikipedia articles to a database format, enabling structured queries over the content of the encyclopaedia. It parses structured parts of the articles (such as info boxes, tables, or categories), which contain a fairly detailed description of the concepts presented in the article.

Content based image retrieval (CBIR) is an alternative to text-based search, but it suffers from important drawbacks, such as the semantic gap [8] and its poor scalability. As a consequence, the use of image processing techniques in Web-scale image retrieval is currently limited to face detection (proposed by Google or Exalead). Previous works [9, 3] advocate that a combination of CBIR and text-based retrieval improves the quality of results. WordNet was exploited in CBIR applications [3], to create multimodal similarity vectors for the visual description of the images [10] or to limit the conceptual neighbourhood where visually similar images are searched [11]. Wang et al. [9] enriched an existing taxonomy of animals (620 terms) with visual information about animal's color and image properties (in/outdoor, photo/graph). The resulting structure outperformed Google Image and a purely textual version of the taxonomy when retrieving images from 20 animal species. However, this interesting approach was limited to a specific domain with quite stable visual properties (the colors of animals). Here we investigate a late fusion scheme of textual information and low level image descriptions, applied to diversified queries.

Image queries reformulation based on semantic resources has already been experimented. In [12], the authors compare a WordNet based query expansion to a ConceptNet based one and conclude that both semantic structures are complementary. The use of WordNet provides a better discrimination of the expanded queries whereas the use of ConceptNet supports better diversity. This was expectable since ConceptNet includes a larger number of inter-conceptual relations. In this paper, we advocate that only parts of the query should be reformulated. We consider that nouns are the most important part of image queries and focus the query expansion on them. For mono-conceptual queries, if knowledge exists about that particular concept, we should use it to expand the query. However, the reformulation is harder for more complex queries because the number of reformulations becomes rapidly unmanageable. This case is thus out of the scope of this work.

Section 2 presents our method based on conceptual structures reformulation. It is experimentally validated in section 3 and discussed in section 4.

2 System Description

Our approach integrates a textual query reformulation using automatically mined conceptual structures and a visual reformulation based on a list of visual concepts that can be automatically detected using image processing. In our approach, we distinguish a knowledge base building and a retrieval phase. The first, which aims at associating precise subtypes to nominal concepts, is performed off-line and its results are exploited during the retrieval, which has to be realized under real time constraints. In retrieval mode, a user request is analyzed and the system separates nominal and visual concepts which will be processed separately, leaving the rest of the query untouched. Each nominal concept in an initial query is reformulated using subtypes or synonyms in the knowledge base and the expanded query is probed against the textual descriptions of the image database. We consider the chance to mistake the annotation of an image is higher when the number of concept is low. Therefore, the images containing the largest number of terms are ranked better. The visual analysis consists in the detection of several visual concepts (from an existing list) and a classification of images with respect to these concepts. The multimedia reformulation of queries consists in re-ranking the text-based reformulation using the visual classification of images.

2.1 Automatic Building of Conceptual Structures

Building automatically conceptual neighbourhood of a good quality for nominal concepts is crucial for our approach. We first draw up a comprehensive list of terms that are to be probed against WordNet and Wikipedia in order to extract and rank their subtypes and synonyms. WordNet is used because it contains good quality structured knowledge, providing at low cost some lists of subtypes and synonyms as well as sense separation for ambiguous concepts. Unfortunately, WordNet has little information related to named entities (which often appear in Web queries) and is less complete than Wikipedia (for instance, there are just over 100 dog races in WordNet and around 600 in Wikipedia). The English version of the collaborative encyclopaedia currently includes over two million articles and, since its content is semi-structured, allows to extract good quality nominal hierarchies [5]. In order to increase the number of discovered subtypes, we first perform a WordNet-based concept expansion, then we reuse the subtypes to match the Wikipedia articles. For instance, when the system looks for subconcepts of *building*, it exploits the *isA* relation between *skyscraper* or *hotel* and *building* and therefore retains these subtypes as representative for *building*.

The concept matching procedure (table 1) relies on the analysis of the first sentence and of the "Categories" box of the articles. As illustrated in table 1, the information of the first sentence and that of the categories box is often complementary. We can extract *skyscraper* as parent concept from both parts of the article for *Empire State Building* and *Transamerica Pyramid*, but only from the Categories box for *50 California Street*. Nominal concepts often have a high number of subtypes and it is necessary to order them so as to favour

Table 1. Concept matching in Wikipedia. We present a ranked list of subtypes for *skyscraper*.

Concept	First sentence	Categories	Article length
Skyscaper	The *Empire State Building* is a 102-story Art Deco *skyscraper*...	*Skyscrapers in New York City*	165510
Skyscaper	The Transamerica Pyramid is the tallest and most recognizable *skyscraper*...	*Skyscrapers* in San Francisco	76403
Skyscraper	50 California Street is a massive *office tower*...	*Skyscrapers* in San Francisco	41049

Table 2. Type of elements that can be identified within a query

Element	Short denomination	possible instances
visual concepts	VIS	sky, night, day, portrait etc.
nominal concepts	NC	skyscraper, building, dog etc.
named entities	NE	Eiffel Tower, Ferrari, George W. Bush...
modifiers	MOD	white, red, gothic, historic
others	OTH	

those that are the most representative. Here we used the length of Wikipedia articles as a simple ranking measure, considering that subtypes described in more detail tend to be more representative. We illustrate the results of the ranking process in table 1, where the presented subtypes of *skyscraper* are ranked (first *Empire State Building* (165510), then *Transamerica Pyramid* (76403) and finally *50 California Street*). With the joint use of WordNet and Wikipedia, we obtain a large scale knowledge base, including good quality conceptual relations, which is usable during the retrieval phase.

2.2 Image Retrieval Phase

The query analysis is the key element of our image retrieval scheme. It separates the user requests in atomic parts, which can be one of the elements presented in table 2. This separation is necessary in order to process each query component adequately. For instance, we attempt a textual reformulation only for *nominal concepts* (NC) and a part of *named entities* (NE) but not for *visual concepts* (VIS), *modifiers* (MOD) and *others* (OTH). It is performed using existing lists of VISs, NCs, NEs and MODs and considers everything that is not in a list as being something else (OTH). At the end of the analysis, we remove stop words from the query. The list of visual concepts is arbitrary determined according to the hierarchy proposed by [13], corresponding to some concepts that can be processed by image processing algorithms. *Nominal concepts* and *named entities* are extracted from WordNet and Wikipedia, while *modifiers* are WordNet adjectives. In table 3, we present two examples of textual query reformulation using our technique. The textual reformulation works for concepts existing in the knowledge base only. If the query is composed of unknown concepts, it will not

Table 3. Examples of query reformulations

Initial query	Query analysis	Reformulated query
skyscraper	NC(skyscraper)	skyscraper + Empire State Building skyscraper + Transamerica Pyramid
bridges by night	NC(bridges) by VIS(night)	Golden Gate Bridge + bridge + night Pont Alexandre III + bridge + night

be reformulated and the results will be identical to a chain matching retrieval. During queries analysis, we chose to consider multiwords (such as *hunting dog* or *White House*) as single concepts because they refer to a single entity. For short queries, which are often ambiguous, we retain the default WordNet or Wikipedia sense of the concept. This choice is made because of the lack of information on the user's intent: we thus consider the most common sense of a term as the most adequate to answer the user need. If additional information is provided, we try to match the query to most appropriate word meaning.

The reformulated queries are compared to the textual descriptions of the images and the results are ranked to favour those images that are described by the highest number of concepts. The rank of a result is given by:

$$\begin{aligned} Rank = \alpha \times (\mathcal{N}_{NCinit} + \mathcal{N}_{NEinit} + \mathcal{N}_{VISinit}) + \\ \beta \times (\mathcal{N}_{NCrefo} + \mathcal{N}_{NErefo}) + \\ \gamma \times (\mathcal{N}_{MOD} + \mathcal{N}_{OTH}) \end{aligned} \quad (1)$$

where $\mathcal{N}_{Xy}$ is the number of concepts of a certain type (see table 2) appearing in the user query ($y = init$) or in its reformulated version ($y = refo$). Equation 1 gives a first ranking of results, favouring those results that are described by a high number of query related concepts. We studied different results configuration and decided that NCs, NEs and VISs in the initial queries should be given the highest weight, followed by NCs and NEs obtained after the query reformulation and by MODs and OTHs ($\alpha > \beta > \gamma$). Equation 1 differentiates between answers that are described by a different number of concepts or by different types of concepts. For instance, a picture annotated with *skyscraper* and *Empire State Building* is ranked higher than a second one annotated with *skyscraper* only, which is ranked higher than a third picture annotated with *Empire State Building* only. Equation 1 fails to separate queries having the same quantity and type of concepts (for instance, two pictures annotated with *Empire State Building*, respectively with *Transamerica Pyramid*). To discriminate these last types of answers, we use the subtypes based on Wikipedia articles length.

The proposed retrieval scheme is flexible and is able to retrieve results that are described by the initial query and expanded concepts, by the initial query or the expanded concepts only or by parts of the initial query. Equation 1 and the use of the subtypes ranking order answers considering their closeness to the query.

2.3 Multimedia Query Reformulation and Matching

This section describes the visual analysis of queries that aims at (possibly) re-arranging the order of the answers returned by the textual reformulation with respect to the visual concepts in the query.

We used two systems to detect visual concepts within the images. The first one is the Viola-Jones face detector that is based on the boosting of Haar wavelets [14]. The second system [13] is a set of SVM-based classifiers learnt (RBF kernel) to determine the *type* of an image (clipart, map, painting or photo). In this last case (if the image is a photo), other sets of SVM determine whether the image is *indoor* or *outdoor*, *day* or *night*, as well as whether it is a *urban* or a *natural* scene. The multi-class classification scheme is solved using a one-versus-one approach. For each classifier, the images of the learning databases were chosen separately of the wikipedia corpus used in the experimental evaluation.

The queries were analysed to detect those containing (explicitly or implicitly) visual cues that can be detected using the visual analysis described above. Each visual concept was linked to a pre-defined list of textual concept that triggers its use. For instance, the presence of a person name (such as *Georges W Bush*) will trigger the use of the face detector. The presence of the word *map* in the query will claim for the use of the *image type detector* and favour the images tagged as *maps*; the word *cartoon* will trigger a search for images classified as *cliparts*. When a list of answers coming from the two first layers is reordered, the images detected as relevant according to the visual concept associated to the query are put at the head of the list without changing their relative order.

3 Experimental Validation

Our method has been evaluated in the context of the wikipedia MM task at ImageCLEF 2008 [1]. We submitted two runs, in order to compare our method to the state-of-the-art on the one hand, and to evaluate more specifically the influence of the multimedia query reformulation on the other hand.

Our system returns 170 documents to each query on average. Over the 75 queries to process, only 33 were reformulated with respect to the visual concepts. We quantified the change this brought about with the Levenshtein distance[15] between the index of the lists of results before and after this multimedia reformulation. The Levenshtein distance is a classic metric to measure the distance between two strings (so called "edit distance"), given as the minimum number of operations needed to transform one string into the other. In our case, we found an average Levenshtein distance of 98.1. The average "rank change", defined as the difference of rank within the lists before and after the multimedia reformulation, is 37.6.

Table 4 reports the main results of the two runs we submitted. The run *ceaTxt* is the output of the textual query reformulation and matching only, whereas the run *ceaConTxt* is the output of the full system including the multimedia query reformulation and matching. The textual reformulation is effective since

Table 4. Performances of our method at ImageCLEF wikipedia task. The results are given in terms of Mean Average Precision, and precision at ranks five and ten.

Run	MAP	P@5	P@10
ceaConTxt	0.2735	0.5467	0.4653
ceaTxt	0.2632	0.52	0.4427

our system is ranked 4th (MAP - 0.2632, P@10 - 0.4427) and the first purely textual approach (no reformulation and no feedback) is only ranked 10th (MAP - 0.2551, P@10 - 0.44). The difference between our two runs shows an interest for the multimedia reformulation and rearrangement that led to an improvement of one point in terms of MAP (from 0.263 to 0.273). It is worth noting that about half of the images were judged as relevant among the ten first answers returned by our system, demonstrating a practical interest for a real user.

4 Conclusions and Perspectives

We proposed a new image retrieval scheme that exploits both textual and visual information. The approach is based on a query reformulation using concepts that are semantically related to those in the initial query. We used Wikipedia and WordNet to extract a ranked list of related concepts for a large number of concepts and reformulate text queries. We also added an image processing which exploits visual cues in queries.

The results submitted at ImageCLEF 2008 were ranked 4^{th} and 2^{nd} with a mean average precision of 0.2632 and 0.2735. The small difference between the two submitted runs shows that the greater contribution to the final results was probably due to the use of conceptual structures, although a rigorous comparison would have required submitting a run with the third layer (visual concept detection) only. Nevertheless, the improvement of the results' precision accounts for the interest of introducing visual concept detection in the retrieval schema.

Number of features of our system are currently still under investigation. The detection of associated concepts is currently limited to the use of Wikipedia and WordNet. We plan to extend our approach so as to exploit search engine snippets, in order to improve the coverage of the resources. As well, while simple and generally effective, the current ranking procedure can certainly be improved if, for instance, we favour unambiguous hyponyms over ambiguous ones. Finally, we are currently exploring a finer grained filtering of visual concepts.

Acknowledgments. We thank the Direction Générale des Entreprises for funding us through the regional business cluster Systematic (project POPS) and Cap Digital (project Mediatic).

References

1. Tsikrika, T., Kludas, J.: Overview of the WikipediaMM task at ImageCLEF 2008. In Peters, C., Giampiccol, D., Ferro, N., Petras, V., Gonzalo, J., Peñas, A., Deselaers, T., Mandl, T., Jones, G., Kurimo, M., eds.: Evaluating Systems for Multilingual and Multimodal Information Access – 9th CLEF Workshop. Lecture Notes in Computer Science, Aarhus, Denmark (2008)
2. Fellbaum, C.: WordNet : an electronic lexical database. MIT press, Cambridge, MA, USA (1998)
3. Yang, J., Wenyin, L., Zhang, H., Zhuang, Y.: Thesaurus-aided approach for image browsing and retrieval. Multimedia and Expo, 2001. ICME 2001. IEEE Intl. Conference on (2001) 1135–1138
4. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: Dbpedia: A nucleus for a web of open data. In: Proc. of the 6th Intl. Semantic Web Conference. (2008) 722–735
5. Ponzetto, S., Strube, M.: Deriving a large scale taxonomy from wikipedia. In: Proc. of the 22nd National Conference on Artificial Intelligence (AAAI-07), Vancouver, B.C. (2007) 1440–1447
6. Kazama, J., Torisawa, K.: Exploiting wikipedia as external knowledge for named entity recognition. In: Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. (2007) 698–707
7. Ruiz-Casado, M., Alfonseca, E., Castells, P.: Automatising the learning of lexical patterns: An application to the enrichment of wordnet by extracting semantic relationships from wikipedia. Data Knowl. Eng. **61** (2007) 484–499
8. Datta, R., Joshi, D., Li, J., Wang, J.: Image retrieval : Ideas, influences and trends of the new age. ACM Transactions on Computing Surveys (2008)
9. Wang, H., Liu, S., Chia, L.T.: Does ontology help in image retrieval?: a comparison between keyword, text ontology and multi-modality ontology approaches. In: Proc. of the 14th ACM Intl. Conference on Multimedia, New York, NY, USA, ACM (2006) 109–112
10. Ferecatu, M., Boujemaa, N., Crucianu, M.: Semantic interactive image retrieval combining visual and conceptual content description. Multimedia systems **13** (2007) 309–322
11. Popescu, A., Millet, C., Moëllic, P.A.: Ontology driven content based image retrieval. In: Proc. of the 6th ACM Intl. Conference on Image and Video Retrieval, New York, NY, USA, ACM (2007) 387–394
12. Hsu, M.H., Tsai, M.F., Chen, H.H.: Query expansion with conceptnet and wordnet: An intrinsic comparison. In: Proc. of the 3rd Asia Information Retrieval Symposium – Information Retrieval Technology. Lecture Notes in Computer Science (2006) 1–13
13. Millet, C.: Automatic image annotation: consistent annotation, and creating automatically a learning database. ENST, Paris (2008) PhD thesis.
14. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. Computer Vision and Pattern Recognition, 2001. Proc. of the 2001 IEEE Computer Society Conference on **1** (2001) I–511–I–518 vol.1
15. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady **10** (1966) 707710