Exploiting Visual Concepts to Improve Text-Based Image Retrieval

Sabrina Tollari, Marcin Detyniecki, Christophe Marsala,
Ali Fakeri-Tabrizi, Massih-Reza Amini, Patrick Gallinari

Abstract. In this paper, we study how to automatically exploit visual concepts in a text-based image retrieval task. First, we use Forest of Fuzzy Decision Trees (FFDTs) to automatically annotate images with visual concepts. Second, using optionally WordNet, we match visual concepts and textual query. Finally, we filter the text-based image retrieval result list using the FFDTs. This study is performed in the context of two tasks of the CLEF2008 international campaign: the Visual Concept Detection Task (VCDT) (17 visual concepts) and the photographic retrieval task (ImageCLEFphoto) (39 queries and 20k images). Our best VCDT run is the 4th best of the 53 submitted runs. The ImageCLEFphoto results show that there is a clear improvement, in terms of precision at 20, when using the visual concepts explicitly appearing in the query.

1 Introduction

Content-based (using only visual features) and text-based (using only textual features) image retrieval are two different approaches to retrieve images. A middle approach consists to combine text and visual information in the same framework. Previous works [2, 3, 8] show that combining text and visual information improves image retrieval, but most of this work use an early or late fusion of visual and textual modality. Another way to use both modalities is to use visual concepts to filter text-based results. In [9], such a method is proposed, but the user has to manually choose the visual concept to apply. In this paper, we particularly study how to automatically match visual concepts and textual query.

The international campaign ImageCLEF 2008[1] proposes (among other tasks) a Visual Concept Detection Task (VCDT) [4] and a general photographic retrieval task (ImageCLEFphoto) [1]. The 17 visual concepts of VCDT are: *indoor, outdoor, person, day, night, water, road or pathway, vegetation, tree, mountains, beach, buildings, sky, sunny, partly cloudy, overcast, animal*. These concepts are rather general and well adapted to images. The ImageCLEFphoto task proposes 39 queries (such as "church with more than two towers" or "people observing football match"). Each of the 20k images is part of the IAPR TC-12 photographic collection. It includes photographs of people, animals, cities, landscapes, pictures of different sports, and many other aspects of contemporary life.

[1] http://imageclef.org/ImageCLEF2008

In Section 2, we introduce our method of visual concept detection using Forests of Fuzzy Decision Trees. In Section 3, we describe how we match the visual concepts and the textual query, and how we filter the text-based results. In Section 4, we present experiments and results obtained during the ImageCLEF2008 campaign. Finally, in the last section, we conclude.

2 Visual Concept Detection Using Fuzzy Decision Trees

Automatic image annotation is a typical inductive machine learning approach. One of the most common methods in this research topic is the decision tree approach (DT). One limitation when considering classical DTs is their robustness and threshold problems when dealing with numerical or imprecisely defined data. The introduction of fuzzy set theory smoothes out these negative effects. In general, inductive learning consists on raising from the *particular* to the *general*. A tree is built, from the root to the leaves, by successively partitioning the training set into subsets. Each partition is done by means of a test on an attribute and leads to the definition of a node of the tree [6]. In [7] was shown that, when addressing unbalanced and large (in terms of dimension and size) data sets, it is interesting to combine several DTs, obtaining a Forest of Fuzzy Decision Trees (FFDTs). Moreover, when combining the results provided by several DTs the overall score becomes a degree of confidence in the classification.

During the learning step, a FFDT of n trees is constructed for each concept C. Each tree F_j of the forest is constructed based on a training set T_j, each being a balanced random sample of the whole training set.

During the classification step, each image I is classified by means of each tree F_j. We obtain a degree $d_j \in [0, 1]$, for the image I, to be a representation of the concept C. Thus, for each I, n degrees $d_j,\ j = 1 \ldots n$ are obtained from the forest. Then all these degrees are aggregated by a weighted vote, which mathematically corresponds to the sum of all the degrees: $d = \sum_{j=1}^{n} d_j$. Finally, to decide if an image presents a concept or not, we use a threshold value $t \leq n$.

3 Using Concepts to Improve Text-Based Retrieval

Once we are able to determine visual concepts present in an image, the difficulty is to determine how to use them in an image retrieval task. We propose two kinds of matching between visual concepts and textual queries. The first matching (called *direct matching*) is applied when the name of a visual concept appears in the text of the query. The second matching (called *WN matching*) is applied when the name of a visual concept appears (i) in the text of the query or (ii) in the list of words semantically in relation (according to WordNet [5]) with the words of the query. For example, the text of query 5 of ImageCLEFphoto2008 is "animal swimming". Using direct matching, the system automatically determines that it must use the FFDT for the concept *animal*. In addition, if we use WordNet (WN matching) and the relation of synonymy, the system automatically determines

	EER	Gain	AUC	Gain
53 runs (average)	33.92	-	63.64	-
Random	50.17	-48%	49.68	-22%
FFDT	24.55	+28%	82.74	+30%

Table 1. Equal Error Rate (EER) and Area under ROC curve (AUC) obtained in the ImageCLEF2008's Visual Concept Detection Task (VCDT)

that it must use the FFDT for *animal* and also for *water*, because according to WordNet, synonyms of "swimming" are: "water sport, aquatics".

When a visual concept C matches (by direct or by WN matching) the text of the query q, then we propose to filter the image list result of a text-based retrieval, according to the degree d, given by the FFDT, that C appears in the image. We put forward the following algorithm. Let R be the numbers of images which could be filter. The system browses the retrieved images from rank 1 to rank R. If the degree of an image is lower than the threshold t, then the image is re-ranked at the end of the current R images list. In this way, we keep the relevant images in the top R.

4 Experiments and Results

Visual Concepts Detection Task The VCDT corpus contains 1827 training images and 1000 test images. There are 17 concepts. This task corresponds to a multi-class multi-label image classification. Images of the training set are labeled in average by 5.4 concepts. All the FFDTs are composed of 50 trees. The degrees of confidence are the direct result of the corresponding FFDT, for each concept.

In order to obtain spatial-related information, the images are segmented into 9 overlapping regions. A large central region represents the purpose of the picture; top and bottom regions correspond to a spatial focus of these areas; left and right top, left and right middle, left and right bottom regions are described in terms of color difference between the right and the left, in order to explicit any *recurrent* symmetries. In fact, objects can appear on either side and decision trees are not able to automatically discover this type of relations. For each region, an HSV histogram is computed.

Table 1 compares results of the ImageCLEF2008's VCDT task. Our run, based on FFDT, ranked 4th run over the 53 submitted runs (third team of 11 international teams). Our method provides a gain of 28%, in terms of Equal Error Rate (EER), compared to the average of the 53 submitted runs.

Image Retrieval Task The ImageCLEFphoto2008 corpus contains 20000 images and 39 queries. Each image is associated with an alphanumeric caption stored in a semi-structured format. These captions include title of the image, creation date, location, name of the photographer, a semantic description of the contents of the

	All 39 queries		Queries modified by filtering				
	P20	gain	Number of queries	P20	gain	Number of filters	Number of images re-ranked
Text only	0.250	-	12	0.146	-		
Direct matching	0.276	+10%	12	0.233	+60%	12	250
Text only	0.250	-	25	0.210	-		
WN matching	0.255	+2%	25	0.228	+9%	33	749

Table 2. Comparison of direct and WN matching for visual concept filtering applied on the first 50 images of a text only result ($R = 50$). For direct matching, only 12 queries were concerned, while for WN matching there were 25. A random permutation, for each query, of the first 50 text results gives a precision at 20 (P20 score) of 0.215

image and additional notes. For text-based retrieval, we use all this elements, but to match concepts and queries, we only use the title field.

To determine if an image shows a visual concept, we choose to set the threshold t to the mean[2] of all the degrees values for a given concept. Since our method depends on the presence of a concept in the text query, it does not apply to every query. For the other queries, result images from text retrieval are not modified.

Table 2 shows that, for all queries, direct matching improves the precision at 20 (P20 score) by 10% compare to a text-based retrieval based on TF-IDF, while WN matching improves P20 by 2%. When using direct matching, only 12 queries are modified, 12 filters are applied and the total number of images that were filtered out (i.e. put at the end of the list) is 250. Using WN matching, 25 queries are concerned. Several queries are modified several times. The total number of times that a filter is applied is 33, for a total of 749 filtering actions. Thus, we separate the study into three groups: all the queries, the 12 queries of direct matching and the 25 queries of WN matching. On Table 2, we observe an improvement of +60%, with respect to TF-IDF scores, for P20 on the 12 queries modified (the P20 scores of all the 12 queries are improved). When using WordNet there is still an improvement with respect to TD-IDF but weaker (+9% for P20). The presented scores correspond to the use of the synonymy relation of WordNet. We also tested hypernymy and hyponymy and the corresponding results were below the synonymy ones. We also try to use all the text of each query (not only the title), but the results are similar or below the scores using only the words of the title. We believe that if WN matching does not work, is because WordNet is not well adapted for images. The WN matching matches concepts, which are not in relation, in the domain of images, with the queries. It could be interesting to have an ontology adapted to images.

If we compare, on Table 2, the P20 score for all the 39 topics (0.250) with the P20 score just on concerned topics (12 topics for direct matching (0.146) and 25 for WN matching (0.210)), we notice that the first is higher than the others. An explanation should be that the modified queries - which contain a visual concept

[2] In this paper, we use the mean operator instead of the median, as submitted to ImageCLEFphoto 2008. Results are slightly different, but conclusions are the same.

in their text - have a strong visualness [8], i.e. particularly for those queries, a usefull information is contained in the visual content of images, sometimes this information is even more usefull than the text information.

5 Conclusion

In this article, we focus on how to automatically exploit visual concepts in an image retrieval task. We show that automatic learning of visual concepts and then its exploitation, by filtering of text-based image retrieval is effective. This study provides evidence for a recurrent and clear improvement, in terms of precision at 20, when using the visual concepts explicitly appearing in the query. Since explicit indication of the concept is not always available, we tested a matching expansion based on WordNet relations. The number of modified queries increased but the performance declined, staying above the text only baseline. We deduce that visual concept filtering is a promising approach, but the challenge lies in how to automatically detect, from the query, the visual concept to be used. We believe that errors coming from the matching expansion are due to the lack of visual awareness in the used semantic lexicon. On future work, we will focus on how to use the relation between concepts to improve the concepts detection and the image retrieval. We will also study how the concepts detected in the query images of each query can be used to improve image retrieval using concepts.

Acknowledgment This work was partially supported by the French National Agency of Research (ANR-06-MDCA-002 AVEIR project).

References

1. T. Arni, P. Clough, M. Sanderson, and M. Grubinger. Overview of the ImageCLEFphoto 2008 photographic retrieval task. In *Evaluating Systems for Multilingual and Multimodal Information Access – 9th Workshop of the Cross-Language Evaluation Forum*, LNCS, 2008.
2. K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei, and M. I. Jordan. Matching words and pictures. *Machine Learning Research*, 3:1107–1135, 2003.
3. R. Datta, D. Joshi, J. Li, and J. Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 40(2), 2008.
4. T. Deselaers and T. M. Deserno. The visual concept detection task in ImageCLEF 2008. In *Evaluating Systems for Multilingual and Multimodal Information Access – 9th Workshop of the Cross-Language Evaluation Forum*, LNCS, 2008.
5. C. Fellbaum. *WordNet - An Electronic Lexical Database*. Bradford books, 1998.
6. C. Marsala and B. Bouchon-Meunier. Forest of fuzzy decision trees. In *International Fuzzy Systems Association World Congress*, volume 1, pages 369–374, 1997.
7. C. Marsala and M. Detyniecki. Trecvid 2006: Forests of fuzzy decision trees for high-level feature extraction. In *TREC Video Retrieval Evaluation Online Proceedings*, 2006.
8. S. Tollari and H. Glotin. Web image retrieval on ImagEVAL: Evidences on visualness and textualness concept dependency in fusion model. In *ACM CIVR*, 2007.
9. A. Yavlinsky, D. Heesch, and S. M. Rüger. A large scale system for searching and browsing images from the world wide web. In *CIVR 2006*, pages 537–540, 2006.