Using Visual Concepts and Fast Visual Diversity to Improve Image Retrieval

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

Université Pierre et Marie Curie-Paris6
Laboratoire d'Informatique de Paris 6 - UMR CNRS 7606
104 avenue du président Kennedy, 75016 Paris, France
firstname.lastname@lip6.fr

Abstract. In this article, we focus our efforts (i) on the study of how to automatically extract and exploit visual concepts and (ii) on fast visual diversity. First, in the Visual Concept Detection Task (VCDT), we look at the mutual exclusion and implication relations between VCDT concepts in order to improve the automatic image annotation by Forest of Fuzzy Decision Trees (FFDTs). Second, in the ImageCLEFphoto task, we use the FFDTs learn in VCDT task and WordNet to improve image retrieval. Third, we apply a fast visual diversity method based on space clustering to improve the cluster recall score. This study shows that there is a clear improvement, in terms of precision or cluster recall at 20, when using the visual concepts explicitly appearing in the query and that space clustering can be efficiently used to improve cluster recall.

1 Introduction

Automatic image annotation is an important issue to improve image retrieval. In fact, users prefer to use words to express theirs need of information. The ImageCLEF track of the 2008 CLEF campaign permits us to study in the same context the image annotation and the image retrieval: the Visual Concept Detection Task (VCDT) [2] allows us to study how to extract visual concepts, and then in the Photo Retrieval task (ImageCLEFphoto) [1], we use the visual concept to match the text query with the visual concepts. In the other hand, the particularity of the 2008 ImageCLEFphoto edition was its focus on diversity. Most of the diversity methods propose to apply the diversification of the results after retrieving the images. This means that the diversification must be done on line and so must be very fast. So we proposed to use space visual clustering which is well known to be a fast clustering technique.

In Section 2, we present our Forests of Fuzzy Decision Trees methods and the cooccurrences analysis applied in the VCDT task. In Section 3, we describe the techniques we use in the ImageCLEFphoto task, especially how we use the VCDT concepts in this task and our diversification method. Finally, in the last section, we conclude.

2 The Visual Concept Detection Task (VCDT)

2.1 Forests of Fuzzy Decision Trees (FFDTs)

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 [4]. In [5] 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 $t \leq n$.

2.2 Cooccurrences analysis

DTs learn each concept independently, but concepts can be related. For instance, when a scene can not be simultaneously *indoor* and *outdoor*, or if we observe that is *overcast*, it implies that we have the concept *sky*. Here, we propose to use cooccurrence analysis to automatically find these relations. Once we have discovered the relation, we need a rule to resolve the conflicting annotations. In fact, each concept is annotated by a FFDT, with a certain confidence degree. For instance, for each image, we will have a degree of having the concept *outdoor* and a certain degree of having *indoor*. We know that both can not appear simultaneously, something has to be done. We propose to use simple rules. In this paper, we study two type of relations between concepts: exclusion and implication.

Exclusion discovery and rule To discover the *exclusions*, we need to look at what concept *never* appear together. For this, we calculate a cooccurrence matrix COOC. Since there may be some noise (e.g. annotation mistakes), we use a threshold α to decide which pair of concepts never appear together. Once we know which concepts are related, we apply a resolution rule to the scores provided by the FFDT. We choose the rule, that for mutually excluding concepts, eliminates (i.e. gives a confidence of zero) to the label having the lowest confidence. For instance, if we have *outdoor* with a degree of confidence of 42/50 and *indoor* with a degree of 20/50 then we will say that it is certainly not *indoor* and its degree should equal 0. For each test image I, let d(I,C) be the FFDT degree of I for concept C, we then apply the following algorithm:

```
for each couple of concepts (A,B) where $COOC(A,B) \leq \alpha$ (discovery)
    if d(I,A) > d(I,B) then d(I,A)=0 else d(I,B)=0 (resolution rule)
```

where COOC is the concept cooccurrence matrix.

Implication discovery and rule To discover *implications*, we need to look, by definition of the implication, at the cooccurrence of the absence of concepts and of the presence of concepts. The resulting cooccurrence matrix COOCNEG is non symmetric, which reflects the fact that one concept may imply another one, but the reciprocal may not be true. The resolution rule says that if a concept implies another one, the confidence degree of the latter should be at least equal to the former. Since there may be some noise, we use a threshold β to decide which concepts imply other ones. For each test image I, let d(I,C) be the FFDT degree of I for concept C, we then apply the following algorithm:

```
for each couple of concepts (A,B) where $COOCNEG(A,B) \leq \beta$ (discovery)
    d(I,B)=max(d(I,A),d(I,B)) (resolution rule)
```

where COOCNEG is the concept cooccurrence asymmetric matrix between a concept and the negation of an other concept.

2.3 VCDT Experiments

Visual descriptors The visual descriptors used in this paper are exclusively color based. In order to obtain spatial-related information, the images were segmented into 9 overlapping regions. For each region, we compute a color histogram in the HSV space. The number of bins of the histogram (i.e. numbers of colors) reflects the importance of the region by being valued. The large central region (the image without borders) represents the purpose of the picture. Two other regions, top and bottom, correspond to a spatial focus of these areas. We believe that they are particularly interesting for general concepts (i.e. not objects), as for instance: sky, sunny, vegetation, etc. The remaining regions (left and right top, left and right middle, left and right bottom) are described in terms of color difference between the right and the left. The idea is to explicit any *systematic* symmetries. In fact, objects can appear on either side. Moreover, decision trees are not able to automatically discover this type of relations.

Corpus The VCDT corpus contains 1827 train images and 1000 test images. There are 17 concepts. A train image is labeled in average by 5.4 concepts (standard deviation=2.0, between 0 (2 images) to 11 concepts by image). A concept label in average 584 train images (standard deviation=490, between 68 to 1607 train images by concept). All the forests are composed of 50 trees. This task corresponds to a multi-class multi-label image classification.

			Without class decision		With class decision (t=25)	
	Excl.	Impl.	EER(AUC)	EER(AUC)	EER(AUC)	EER(AUC)
	rule	rule		gains %		gains %
FFDT			24.55 (82.74)	-	26.20 (57.09)	-
FFDT	X		27.37 (71.58)	-11 (-13)	28.83 (54.19)	-10 (-5)
FFDT		X	25.66 (82.48)	-5 (0)	27.51 (54.89)	-5 (-4)
FFDT	X	X	27.32 (71.98)	-11 (-13)	28.93 (53.78)	-10 (-6)
Random			50.17(49.68)	-104(-40)	50.26 (24.89)	-48(-56)

Table 1. Results of VCDT task (EER: Equal Error Rate - AUC: Area under ROC)

Exclusive and implication relations A preliminary step before extracting visual concepts is to study cooccurrence values to discover exclusions and implications. For the 17 concepts, there are 136 cooccurrences values. Those values vary from 0 to 1443 (there are 1827 train images). We set $\alpha = 5$ (two concepts are considered exclusive if at the maximum 5 of the 1827 training images were annotated as presenting the two concepts in the training sets). For the same reason, we set $\beta = 5$ (a concept implies an other concept if at the maximum 5 training images are not annoted by the first concept, but annoted by the second one). Our system automatically discovered 25 exclusive relations and 12 implication relations. We found not only most of the relations suggested in the schema describing the training data, but also several other ones. For the latter, some are logic and some are the result of the fact that some labels are not very frequent. We notice, for instance, that *sunny* and *night* never appear together, but also that there is never a *beach* and a *road* together.

In order to appreciate the effect of the implication and exclusion rules, we look at the results in Table 1. Based on these scores, the exclusion and implication rules seem to worsen the results provided by the FFDTs. We believe that this is due to the fact that these scores are not adapted to boolean classification (and our rules provide boolean decisions). The area under the curve and the equal error rate are interesting when the classification is accompanied by a degree of confidence. Moreover, this measure penalize boolean decision over degrees.

3 The Photo Retrieval Task 2008

3.1 Using VCDT concepts in ImageCLEFphoto

Previous works 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. Following the idea of VCDT and ImageCLEFphoto tasks, we propose to use VCDT visual concepts to filter ImageCLEFphoto text runs in order to answer if visual concept filtering can improve text only retrieval.

The difficulty is to determine how to use the visual concepts of VCDT in ImageCLEFphoto 2008. In the VCDT task, we have obtained a FFDT by concept (see Section 2). Each of these FFDTs can give a degree that the corresponding visual concept appears in a new image. In order to make a decision, we put a threshold t to determine if an image contains the given concept according to the corresponding FFDT. First, if the name of a concept appears in the <title> element (VCDT filtering), we propose to filter the rank images list according to the FFDT of this concept. Second, if the name of a concept appears in the <title> element or in the list of synonyms (according to WordNet [3]) of the words in the <title> element (VCDTWN filtering), we also propose to filter the rank images list according to the FFDT of this concept. For example, the <title> of topic 5 is "animal swimming". Using only VCDT filtering, the system automatically determine that it must use the FFDT of the concept *animal.* If, in addition, we use WordNet (VCDTWN filtering), the system automatically determine that it must use the FFDT of the concept *animal* and of the concept *water* (because according to WordNet, the synonym of "swimming" is: "water sport, aquatics").

For each query, we obtain a list of images ranked by their text relevance according to a language model (LM) or TF-IDF text models. Then, using the decision of the FFDTs, we rerank the first 50 ranked images: the system browses the retrieves images from rank 1 to rank 50. If the degree of an image is lower than the threshold t, then this image is reranked at the end of the current 50 images list.

3.2 Promote Diversity by fast clustering visual space

For a given query *similar* documents are naturally closely ranked. When a user makes a query, he should want that the first relevant documents are as diverse as possible. So the ImageCLEFphoto 2008 task is very interesting to improve image retrieval, but the definition of diversity in the ImageCLEFphoto 2008 task is not very clear, in particular in term of granularity. In most cases, it is strongly related to the text. For us, there are two kinds of diversification in the ImageCLEFphoto 2008. The first one is knowledge based: *city, state, country, venue, landmark....* The second one is based on visual information: *weather condition, group composition, statue....* For this clusters, visual diversification should improve results. As in real applications, it is not obvious to determine automatically which kind of diversification applying for a given query [6], we choose to apply, for all query (even if it is suboptimal), the same kind of diversification (the visual one) by clustering the visual space.

Visual clustering has been studied for a long time now. Two approaches are generally proposed: data clustering and space clustering. The first approach requires lots of calculation time and should be adapted to distribution of the first images ranked by a given query. The second approach, since it is done independently of the data, is often less efficient, but can be applied extremely fast. We choose to cluster the visual space based on the hue dimension of the HSV space. For each image, we binarize its associated 8 bin hue histogram. Each binary vector correspond to a cluster. The number of clusters is 256 (not all are instantiated), a reasonable number for a re-ranking at P20.

	Visual	All 39 topics		Topics modified by filtering		
	concept	P20	CR20	Nb	P20	CR20
Text	filtering	(gain %)	(gain %)	topics	(gain %)	(gain %)
LM	-	0.185 (-)	0.247 (-)	11	0.041 (-)	0.090 (-)
				25	0.148 (-)	0.254 (-)
	VCDT	0.195 (+6)	0.257 (+4)	11	0.077 (+88)	0.126 (+40)
	VCDTWN	0.176 (-5)	0.248 (+1)	25	0.134 (-9)	0.257 (+1)
TF-IDF	-	0.250 (-)	0.300 (-)	11	0.155 (-)	0.161 (-)
				25	0.210 (-)	0.305 (-)
	VCDT	0.269 (+8)	0.313 (+5)	11	0.223 (+44)	0.209 (+30)
	VCDTWN	0.260 (+4)	0.293 (-2)	25	0.226 (+8)	0.294 (-4)

Table 2. Comparison of VCDT and VCDTWN filtering. For VCDT filtering, only 11 topics are modified. For VCDTWN, only 25 topics are modified

We use the visual space clusters to rerank the 50 retrieve images. For each query, the system browses the retrieves images from rank 1 to rank 50. If an image has the same visual space cluster as an image of highest rank, then this image is reranked at the end of the current 50 images list. In this way, if in the 50 first images, there are n different visual space clusters, then at the end of the rerank process, the first n images correspond to strictly different visual space clusters. We call this diversification method: DIVVISU.

In order to have a point of comparison, we also propose to randomly permute the first 40 retrieve images. We call this naive method of diversification: DIVALEA.

3.3 ImageCLEFphoto Experiments and Results

The ImageCLEFphoto2008 corpus contains 20k images and 39 topics. Each image is associated with an alphanumeric caption stored in a semi-structured format. These captions include the title of the image, its creation date, the location at which the photograph was taken, the name of the photographer, a semantic description of the contents of the image (as determined by the photographer) and additional notes. In the text retrieval, we use all this elements. We build 18 runs: on the beginning, we build two runs based on classical text models (language model and TF-IDF), then we apply, on each of these runs, VCDT filtering or VCDTWN filtering, and finally we apply DIVVISU and DIVALEA diversity methods.

VCDT and VCDTWN filtering To determine if an image should or not contains a visual concept, we choose to set the threshold t to the median of all the degrees values for a given concept (this value varies from 7.3 (*overcast*) to 28.8 (*outdoor*)). We do not use cooccurrence analysis (neither exclusion nor implication rules) in the ImageCLEFphoto task because it was not conclusive in the VCDT task. Table 2 shows that, for all topics, VCDT filtering improves P20 by 8% and VCDTWN filtering improves P20 by 4% in comparison to TF-IDF P20. Since our method depends on the presence of a concept in the text query, it

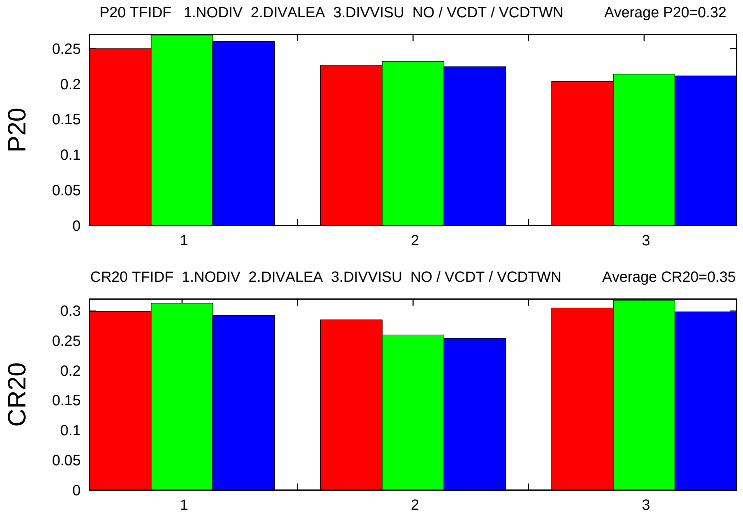

Fig. 1. Comparison of diversification methods 1. no diversification, 2. random diversification (DIVALEA) 3. diversification by visual space clustering (DIVVISU). For each diversification method, scores for TF-IDF only (1st bar), TF-IDF+VCDT (2nd bar) and TF-IDF+VCDTWN filtering (3rd bar) are given

does not apply to every topic. Using VCDT filtering, only 11 topics where filtered. Using VCDTWN filtering, 25 topics where modified. For the other topics, result images from text retrieval keep the same ranked. Thus, we separate the study into three groups: all the topics, the 11 topics modified by VCDT filtering and the 25 topics for which we applied VCDTWN filtering. On Table 2, we observe an improvement on TF-IDF scores of +44% for P20 and +30% for the 11 topics modified by VCDT filtering, but not by VCDTWN filtering (+8% for P20 and -4% for CR20). Using VCDT filtering, all the modified topics are improved, but using VCDTWN filtering, some topics are improved and others are worsened. Then, we conclude that the way we use WordNet is not adapted for this task. Further study is needed.

Diversification Figure 1 compares diversification method scores. DIVALEA and DIVVISU give lower P20 than no diversification, but DIVVISU slightly improves CR20 (in average +2%). So our DIVVISU diversification method works slightly well for diversification, but lowers precision as many others diversity methods (see [7]).

4 Conclusion

In this article, we focus our efforts (i) on the study of how to automatically extract and exploit visual concepts and (ii) on fast visual diversity. First, in VCDT task, we look at the mutual exclusion and implication relations between the concepts, in order to improve the automatic labelling. Our best VCDT run is the 4th ones under 53 submitted runs (3rd team under 11 teams). In our experiments, the use of the relations do not improve nor worsen the quality of the labeling. Second, in ImageCLEFphoto task, we analyse the influence of extracted visual concepts models to the diversity and precision, in a text retrieval context. This study shows that there is a clear improvement, in terms of precision or cluster recall at 20, when using the visual concepts explicitly appearing in the query. Third, we show that our fast visual diversity method based on fast clustering improved the cluster recall at 20. In our future researches, we will focus on how using image query to improve image retrieval using concept.

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 C. Peters, D. Giampiccol, N. Ferro, V. Petras, J. Gonzalo, A. Peñas, T. Deselaers, T. Mandl, G. Jones, and M. Kurimo, editors, *Evaluating Systems for Multilingual and Multimodal Information Access – 9th Workshop of the Cross-Language Evaluation Forum*, Lecture Notes in Computer Science, Aarhus, Denmark, 2008 (printed in 2009).
2. T. Deselaers and T. M. Deserno. The visual concept detection task in ImageCLEF 2008. In C. Peters, D. Giampiccol, N. Ferro, V. Petras, J. Gonzalo, A. Peñas, T. Deselaers, T. Mandl, G. Jones, and M. Kurimo, editors, *Evaluating Systems for Multilingual and Multimodal Information Access – 9th Workshop of the Cross-Language Evaluation Forum*, Lecture Notes in Computer Science, Aarhus, Denmark, 2008 (printed in 2009).
3. C. Fellbaum, editor. *WordNet - An Electronic Lexical Database*. Bradford books, 1998.
4. C. Marsala and B. Bouchon-Meunier. Forest of fuzzy decision trees. In *Proceedings of the Seventh International Fuzzy Systems Association World Congress*, volume 1, pages 369–374, 1997.
5. 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.
6. S. Tollari and H. Glotin. Web image retrieval on ImagEVAL: Evidences on visualness and textualness concept dependency in fusion model. In *ACM Conference on Image and Video Retrieval (CIVR)*, pages 65–72, 2007.
7. Sabrina Tollari, Philippe Mulhem, Marin Ferecatu, Hervé Glotin, Marcin Detyniecki, Patrick Gallinari, Hichem Sahbi, and Zhong-Qiu Zhao. A comparative study of diversity methods for different text and image retrieval approaches. In *Evaluating Systems for Multilingual and Multimodal Information Access – 9th Workshop of the Cross-Language Evaluation Forum*, Lecture Notes in Computer Science, Aarhus, Denmark, 2008 (printed in 2009).