News image annotation on a large parallel text-image corpus

Pierre Tirilly, Vincent Claveau, Patrick Gros

Université de Rennes 1/IRISA, CNRS/IRISA, INRIA Rennes-Bretagne Atlantique
Campus de Beaulieu
35042 Rennes Cedex, France
ptirilly@irisa.fr, vclaveau@irisa.fr, pgros@inria.fr

Abstract

In this paper, we present a multimodal parallel text-image corpus, and propose an image annotation method that exploits the textual information associated with images. Our corpus contains news articles composed of a text, images and image captions, and is significantly larger than the other news corpora proposed in image annotation papers (27,041 articles and 42,568 captionned images). In our experiments, we use the text of the articles as a textual information source to annotate images, and image captions as a groundtruth to evaluate our annotation algorithm. Our annotation method identifies relevant named entities in the texts, and associates them with high-level visual concepts detected in the images (in this paper, faces and logos). The named entities most suited to image annotation are selected using an unsupervised score based on their statistics, inspired from the weights used in information retrieval. Our experiments show that, although it is very simple, our annotation method achieves an acceptable accuracy on our real-world news corpus.

1. Introduction

In the last decade, the amount of available image data has grown continuously, due to dissemination of image acquisition and exchange devices (numeric cameras, cellphones, internet...). Image professionals, such as journalists, as well as private individuals need therefore new solutions to efficiently and effectively store, index and retrieve the images they produce. One particularily challenging problem is to automatically determine the semantic content of pictures, so that images can be annotated with textual information, and query images databases intuitively using words, simarly to what is done to retrieve textual data.

Two reasons make annotating images with language resources difficult. The first reason is the *semantic gap*, *i.e.* the difference between the low-level or perceptual features (color, texture, shape) that we can extract from an image and what this image really means. The second reason is the need for large annotated image corpora, to train machine learning algorithms or evaluate annotation techniques on real-world and large-scale applications. Building such corpora manually is very costly and many authors content themselves with artificial data, such as Corel collections, or small-sized collections. In this paper, we present an image annotation scheme that associates pictures with textual information extracted from surrounding text (Section 3.), relying on a large parallel text-image corpus consisting of news articles, and its associated groundtruth (Section 2.).

2. Parallel Text-Image News Corpus

Our corpus contains 27,041 news articles from March to November 2006. Each article is made of one text in French and one or more images that illustrate the text. The whole corpus contains 42,568 images. Each image comes with a caption, that is often divided into two parts. The first caption sentence, in bold, describes the image precisely. The rest of the caption reminds the context of the article. Figure 1 shows a document from this corpus.

There are two ways of exploiting such a bimodal corpus for image indexing:

1. using the textual and visual modalities as complementary descriptors to perform image retrieval or article retrieval, as done by Tollari and Glotin (2007).

2. using the textual information to annotate images, as done by Deschacht *et al* (2007).

In our corpus, we can use article texts, image captions, or both, as textual information. These three approaches require a solid groundtruth to evaluate the proposed indexing algorithms. Moreover, in the case of image annotation, most of the existing algorithms rely on machine learning techniques that also require groundtruth data to optimize their parameters. However, such groundtruth has to be built manually, which is very time-consuming, especially for large-scale corpora. In this work, we by-pass this limitation by exploiting directly the available information: we use the article texts to perform image annotation, then image captions as a groundtruth to evaluate our algorithm. This method allows us to work with large-scale corpora of any size at no manual annotation cost.

3. Image Annotation using high-level visual and textual concepts

3.1. Annotation algorithm

Our annotation algorithm associates high-level visual concepts detected in images with corresponding textual concepts extracted from the article text that comes with the images. Given an image and its associated article text, we use the following algorithm:

1. Detect n visual concepts in the image.

2. Detect textual concepts in the text and assign to each concept a score that reflects its importance as a potential annotation.

3. Keep the m concepts whose scores are above a given threshold T.

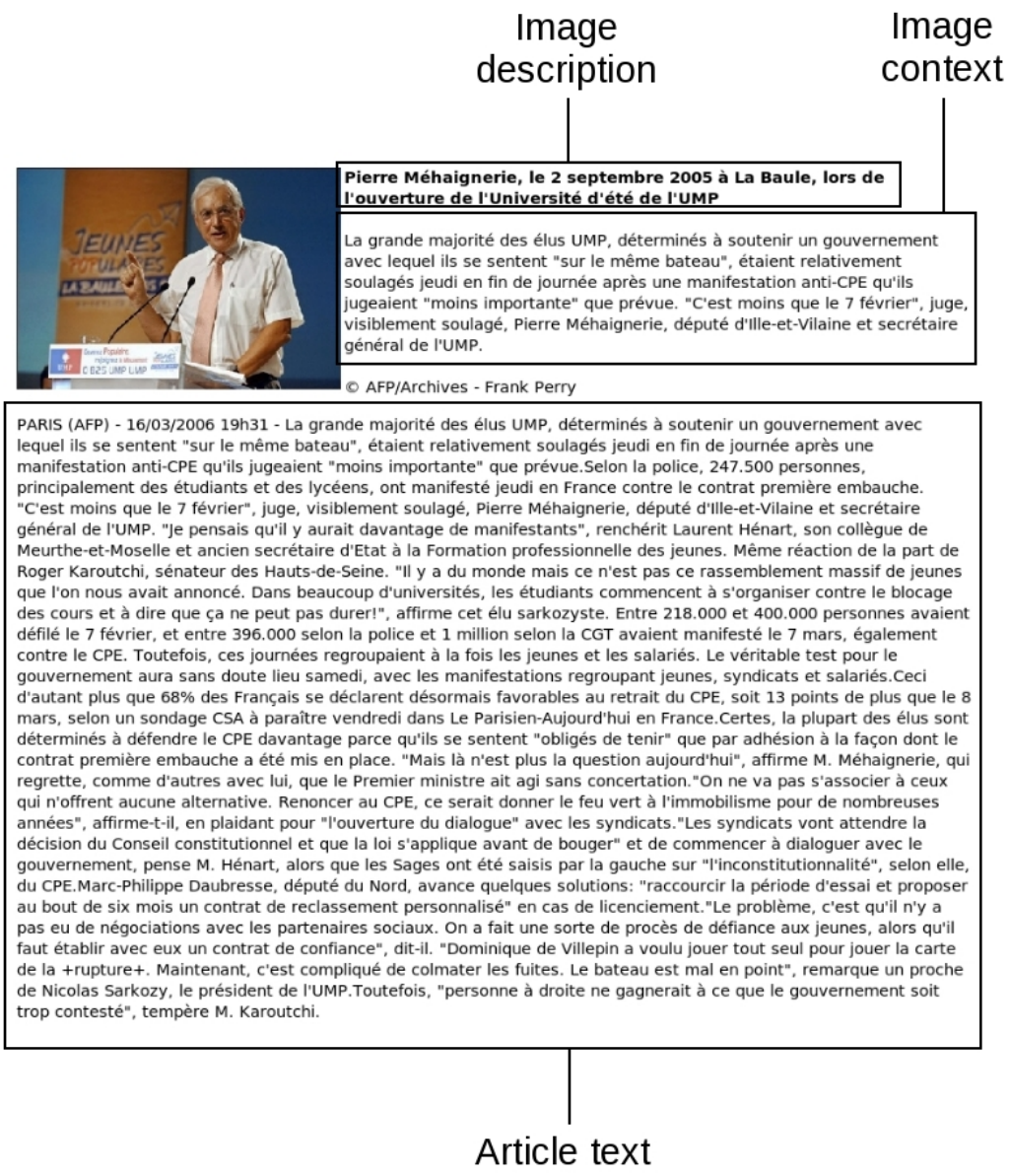


Figure 1: A document from our news corpus

4. Annotate the image with the $\min(n, m)$ textual concepts that have the highest scores.

If we have to choose between several concepts with same scores (ambiguity), none of them is used to perform the annotation. The threshold T has a impact on the annotation results: the number of annotated images decreases when T increases, because all the concepts detected in a text document may have a score lower than T, but we also expect that more annotation will be correct, since we keep more significative concepts. In this paper, the textual concepts that we use are named entities. The visual concepts that we consider in this work correspond therefore to one or more categories of named entities each (see Section 3.3.).

3.2. Visual concepts

We consider two kinds of visual concepts that can be associated with named entities: faces and logos. This two concepts can be efficiently and effectively detected in images.

Face detection: We use the face detector provided in the openCV library (Lienhart et al., 2003). This detector yields good results (about 80% precision with 90% recall) on its authors' dataset, which contains only frontal faces. Although we did not perform extensive experiments, we believe its performance is worse on our dataset, since it contains more changes in face orientations, sizes and occlusions. This can affect the performance of our annotation system: missed faces will reduce our system's recall, as it will make it annotate less images, whereas falsely detected faces may reduce precision (depending on the presence or not of the corresponding textual concept in the text). At last, since the detector often detects the same face several times, we only count the overlapping detection windows once (see Figure 5, where the initial detected windows in red are reduced to a single green window).

Logo detection: We developed ourselves a logo detector based on the visual word framework proposed by Sivic and Zisserman (Sivic and Zisserman, 2003). Tested on a subset of 413 images from our dataset, it achieves 95% detection precision with 60% recall (Tirilly et al., 2010). We consider at most one logo per image, *i.e* if several logos are detected, we consider that only one has been detected because a logo can appear several times on an image (unlike faces, which are necessarily all distinct on a given image).

3.3. Named entity detection and scoring

We use NEMESIS to detect and categorize named entities (Fourour, 2002). According to its author, it achieves 95% precision with 90% recall for the detection and categorization of anthroponyms and toponyms on a news corpus from the French newspaper *Le Monde*. However, we observed in our experiments that it makes more errors on other categories: this might reduce both recall and precision of our system since it can make it miss textual concepts or detect wrong textual concepts. For each image feature, we consider only the named entities from the suited categories:

- we associate faces with named entities from the *anthroponym* category;
- we associate logos with named entities from the following categories: *brands and products*, *organizations*, *firms*, *events*, *artistic groups*, *institutions*.

We assign to each detected named entity a score that reflects the fact that this entity might be a good image annotation or not. These scores are calculated using named entity statistics in the current document and in the whole corpus, similarly to the term weights used in information retrieval (Salton and McGill, 1983). We define the following values that we use to define entity scores:

- **entity frequency** f_{ij}: the occurrence number of entity i in document j;
- **document frequency df**$_i$: the number of documents in which entity i occurs;
- **annotation frequency af**$_i$: the number of images automatically annotated with entity i after an annotation step. It is based on correct as well as false annotations (unsupervised approach);
- **learned annotation frequency laf**$_i$: the number of images annotated with entity i computed on a training set. It contains only correct annotations but require a groundtruth to be calculated (supervised approach).

We then define named entity scores (Table 1). For each $x \in \{\mathrm{df}, \mathrm{af}, \mathrm{laf}\}$, we define two possible scores:

- a direct score f-x that favors entities with a high x;
- an inverse score f-ix, inspired from the classic IDF weights of information retrieval (Salton and McGill, 1983), which emphasizes entities with a small x.

3.4. Experiments

We test our approach on our multimodal corpus. We detect named entities in the article texts and use the first sentence of each image caption (referred as "image description" on Figure 1) as groundtruth: if the entity used to annotate the image is present in the description, then the annotation is considered as right. We measure the performance of our algorithm using annotation precision, defined as follows:

$$\text{precision} = \frac{1}{N}\sum_{i=1}^{N}\frac{\text{number of correct annotations in image } i}{\text{number of annotations in image } i}$$

where N is the total number of annotated images. We make the named entity selection threshold vary, so we can get several *(number of annotated images, precision)* points and present the results with curves similar to recall-precision curves commonly used in information retrieval.
Figures 2, 3 and 4 report annotation results for the two visual concepts we considered and the different scoring techniques we propose. Figure 5 shows some examples of successfull annotations. We can make a few interesting observations:

- af values are computed on images that we annotated once with a simple f score. We considered two thresholds for this first annotation step, $f_{ij} = 1$ and $f_{ij} = 10$, corresponding respectively to af-1 and af-10. Using a low threshold allows us to initially annotate many images, and thus provides more information to compute af, but this information contains many mistakes. On the contrary, using a high threshold reduces the number of mistakes, but also provides less information. This is why af-10-based scores yield results that are much more similar to initial annotation results (f curve) than af-1-based scores;
- the simple f score yields many ambiguity cases, because the possible scores are reduced to integer values. The other scores are real-valued and vary much more. It allows us to avoid many ambiguity cases and therefore annotate much more images;
- the difference between direct and inverse scores shows a trend in our corpus that must be valid in other news corpora: a named entity that appears frequently in the corpus tends to appear frequently in images, *i.e.* images do not tend to contain rare information but rather common information;
- the results of the supervised approach laf, which is not subject to initial annotation errors, tend to show that annotation frequency is much less significative than raw intra-document frequency to annotate images, although it may be a good method to solve ambiguity cases, especially for logo annotations.

Moreover, the performance of our annotation scheme strongly depends on the initial performance of the visual and textual detectors. A qualitative examination of the results shows that many mistakes are due to initial detection errors.

f	frequency	f_{ij}
f-idf	frequency and inverse document frequency	$f_{ij}.\log(\frac{N}{\mathrm{df}_i})$
f-df	frequency and document frequency	$f_{ij}.(1+\frac{\mathrm{df}_i}{N})$
f-iaf	frequency and inverse annotation frequency	$f_{ij}.\log(\frac{N}{\mathrm{af}_i})$
f-af	frequency and annotation frequency	$f_{ij}.(1+\frac{\mathrm{af}_i}{N})$
f-ilaf	frequency and inverse learned annotation frequency	$f_{ij}.\log(\frac{N}{\mathrm{laf}_i})$
f-laf	frequency and learned annotation frequency	$f_{ij}.(1+\frac{\mathrm{laf}_i}{N})$

Table 1: Named entity scores

4. Related work

4.1. Multimodal text-image corpora

Most of the bimodal corpora available to perform image annotation only provide images and keywords, but not full texts. Moreover, these corpora are often artificial and contain mostly categorized pictures with few intra-category changes, which make the annotation problem much easier to solve than in the case of real-world corpora. For instance, the most popular corpus of this kind in the image annotation literature, the COREL image collection, is known to contain many biases (Müller et al., 2002). This consideration motivates the need for new bimodal corpora, containing real texts and corresponding to real application cases. A few corpora of this kind exist in the literature, mostly news corpora. However, they often do not provide complete texts but only captions, which describe the image only (Berg et al., 2005; Deschacht and Moens, 2007; Feng et al., 2004; Westerveld, 2000). Moreover, their size is quite limited (a few hundreds documents). The only news corpora providing full article texts are, to our knowledge, the one used by Jiang and Tan, which is limited to one topic (terrorism) and contains 300 documents (Jiang and Tan, 2009), and the corpus proposed by Feng and Lapata, which contains 3,361 BBC news articles (Feng and Lapata, 2008). We can also mention some real-world benchmarking corpora from conference challenges:

- the ImagEval corpus, which contained full texts, images and groundtruth, but is no longer available;
- the ImageClef Wikipedia tasks, that provides a large-scale corpus of Wikipedia images with user tags;
- the MIR FlickR corpus, that provides user-tagged images from FlickR.

However, none of these corpora has exactly the same properties as ours, which is large-scaled, contains full illustrated texts in French and corresponds to a real-world application.

4.2. Image annotation

Most of the image annotation papers rely on supervised machine learning, such as the seminal work of (Barnard and Forsyth, 2001). However, these approaches require massive training data, which is rarely available in real-world applications and makes the system's capabilities dependent to this data, *i.e.* the annotation system needs many images that are very similar to the image to annotate to perform a correct annotation. The aim of our approach is to avoid relying on such training data. The work that is the closest to ours is certainly the one of Berg *et al.* who annotate faces using the names found in the captions describing pictures (Berg et al., 2005). The main difference with our work is that we try to detect textual concepts from full texts, whereas Berg *et al.* rely on the specific form of the captions they use to extract the people names. Moreover, we extend their principle to use associated high-level concepts such as names and faces to the case of logos and their associated named entities.

5. Conclusion

In this paper, we proposed to make the most of a large text-image corpora with an associated groundtruth in order to use language resources to annotate images in a more semantic way that what can be done by relying on pure image processing. More specifically, we proposed an image annotation method relying on high-level visual features extracted from images and concepts extracted from text. This method is very simple but still quite accurate. It is computationally cheap and can be applied to large-scale indexing of images.

Future work include improvements to our annotation method. We plan to check and propagate annotations by grouping the detected images parts using visual features, similarly to what is done in (Berg et al., 2005). We will also use more complex text analysis to estimate the real presence of an entity in the article texts, for instance by using anaphora resolution to follow each entity through the whole article.

6. References

Kobus Barnard and David Forsyth. 2001. Learning the semantics of words and pictures. In *Proceedings of the International Conference of Computer Vision (ICCV)*, Vancouver, Canada.

Tamara L. Berg, Alexander C. Berg, Jaety Edwards, Michael Maire, Ryan White, Yee-Whye Teh, Erik Learned-Miller, and David Forsyth. 2005. Names and faces in the news. In *Proceedings of the international conference on Computer Vision And Pattern Recognition (CVPR)*, San Diego, USA.

Koen Deschacht and Marie-Francine Moens. 2007. Text analysis for automatic image annotation. In *Proceedings of the annual meeting of the Association of Computational Linguistics (ACL)*, Prague, Czech Republic.

Yansong Feng and Mirella Lapata. 2008. Automatic image annotation using auxiliary text information. In *Proceed-*

ings of the annual meeting of the Association of Computational Linguistics (ACL), Colombus, USA.

Huamin Feng, Rui Shi, and Tat-Seng Chuan. 2004. A bootstrapping framework for annotating and retrieving WWW images. In *Proceedings of ACM Multimedia*, New York, USA.

Nordine Fourour. 2002. Nemesis, un système de reconnaissance incrémentielle des entités nommées pour le français. In *Actes de la conférence Traitement Automatique des Langues Naturelles (TALN)*, Nancy, France.

Tao Jiang and Ah-Hwee Tan. 2009. Learning image-text associations. *IEEE Transactions on Knowledge and Data Engineering*, 21(2).

Rainer Lienhart, Alexander Kuranov, and Vadim Pisarevsky. 2003. Empirical analysis of detection cascades of boosted classifiers for rapid object detection. In *Proceedings of the annual Pattern Recognition Symposium of the German Association for Pattern Recognition (DAGM)*, Magdeburg, Germany.

Henning Müller, Stéphane Marchand-Maillet, and Thierry Pun. 2002. The truth about corel-evaluation in image retrieval. In *Proceedings of the Conference on Image and Video Retrieval (CIVR)*, London, United Kingdom.

G. Salton and M. J. McGill. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill.

Josef Sivic and Andrew Zisserman. 2003. Video Google: A text retrieval approach to object matching in videos. In *Proceedings of ICCV*, Nice, France.

Pierre Tirilly, Vincent Claveau, and Patrick Gros. 2010. Dtection de logos pour l'annotation d'images de presse. In *Proceedings of the conference Reconnaissance de Formes et Intelligence Artificielle (RFIA)*, Caen, France.

Sabrina Tollari and Hervé Glotin. 2007. Web image retrieval on imageval: Evidences on visualness and textualness concept dependency in fusion model. In *Proceedings of the Conference on Image and Video Retrieval (CIVR)*, Amsterdam, The Netherlands.

Thijs Westerveld. 2000. Image retrieval: Content versus context. In *Proceedings of the Conference on Computer Assisted Information Retrieval (RIAO)*, Paris, France.

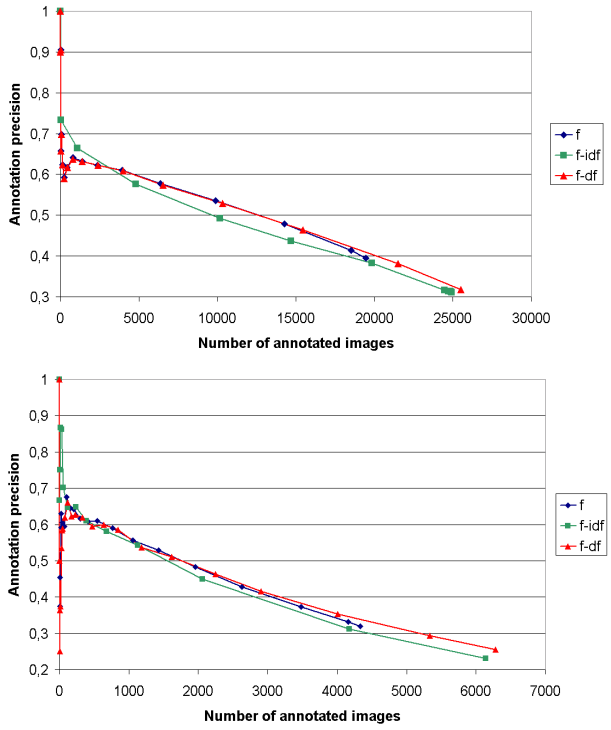


Figure 2: Annotation performance of df-based scores for face (left) and logo images (right).

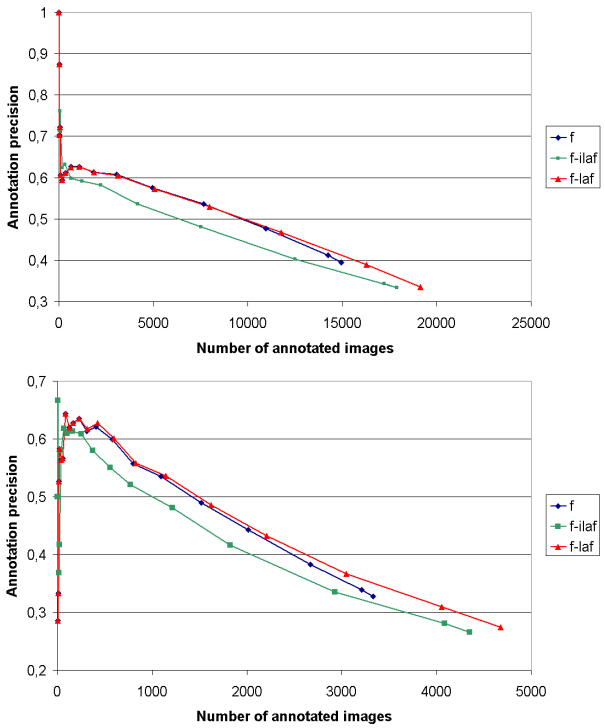


Figure 4: Annotation performance of laf-based scores for face (left) and logo images (right) using 10,000 captions as training set.

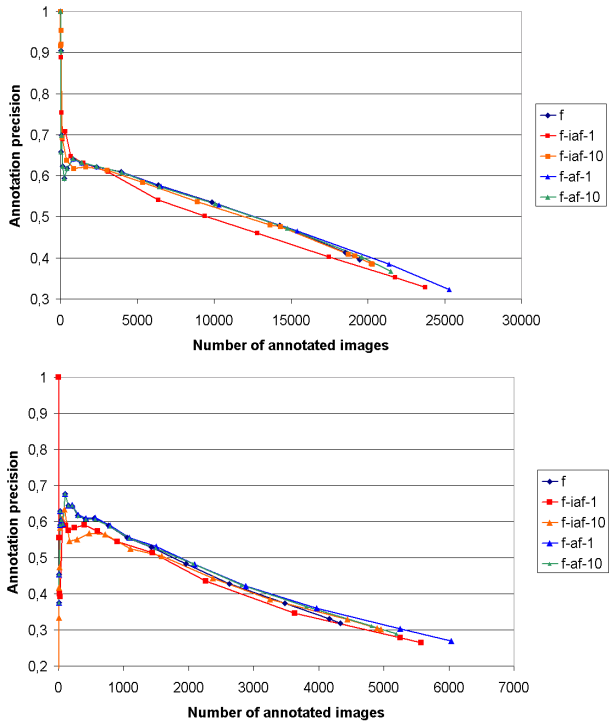
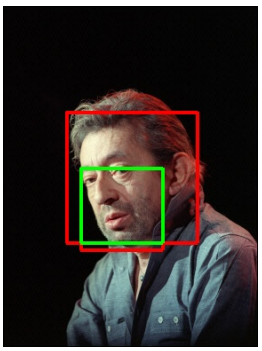


Figure 3: Annotation performance of af-based scores for face (left) and logo images (right).

Libération 8
CE 2
SCPL 2
Rotschild 2
Société Civile des
Personnels de Libération 1
Le Monde 1
Comité d'Entreprise 1



Gainsbourg 17
Nelson 2
Melody 2
Birkin 2
Hardy 2

Figure 5: Some examples of annotations obtained with our method. Candidates annotations are given with their frequency. Final annotations are the annotations in bold.