In press : Proceedings of the 30th European Conference On Information Retrieval
Glasgow, March-April 2008

Finding the Best Picture: Cross-Media Retrieval of Content

Koen Deschacht and Marie-Francine Moens

Katholieke Universiteit Leuven, Department of Computer Science,
Celestijnenlaan 200A, B-3001 Heverlee, Belgium
{Koen.Deschacht,Marie-Francine.Moens}@cs.kuleuven.be
http://www.cs.kuleuven.be/~liir/

Abstract. We query the pictures of Yahoo! News for persons and objects by using the accompanying news captions as an indexing annotation. Our aim is to find these pictures on top of the answer list in which the sought persons or objects are most prominently present. We demonstrate that an appearance or content model based on syntactic, semantic and discourse analysis of the short news text is only useful for finding the best picture of a person of object if the database contains photos each picturing many entities. In other circumstances a simpler bag-of-nouns representation has a good perfoprmance. The appearance models are tested in a probabilistic ranking function.

Key words: Cross-media Retrieval, Information Extraction, Ranking, Image search

1 Introduction

Repositories of multimedia content (e.g., provided by the World Wide Web) demand for effective means of retrieval without relying on manual annotations. In text-based image retrieval some form of textual description of the image contents is stored with the image, the image base is queried with a textual query, and correspondence is sought between the textual data when ranking the images. When people search for images, high precision on top of the answer list is very important. They might search for the best pictures of a person or object (in which the sought entities are most prominently present), or of a combination of them (e.g., picture of a meeting between Angela Merkel and George Bush), where a high recall of all best pictures is valuable, and a high recall of all images picturing the queried persons or objects is not.

Our goal is to find the best images of a person (persons) or object(s) in a database of photos (in our case found on the World Wide Web) that possibly picture many persons or objects and that have associated texts in the form of descriptive sentences. When a text describes an accompanying image, it is often the case that content described in the text is not present in the image and vice versa. In addition, retrieval of the images based on accompanying texts not always returns the best picture on top of the answer list. Our goal is to test

several approaches with varying complexity of analysis of the caption texts for their capability of being discriminative indexing descriptions of the images.

Generative probabilistic retrieval models are suited for cross-media information retrieval. They rely on a content model generated from a document. We call the content model a language model when it represents the content of a text and an appearance model when it represents the content of an image. When retrieving images that have accompanying texts, one can design several content models that probabilistically model the textual and/or the visual content. In the research reported here we build an appearance model solely on the basis of the text in an attempt to capture persons and objects that are present in the image and to compute their degree of prominence in the image. This appearance model is used in a probabilistic ranking function for retrieval. We illustrate our approach by querying the pictures of Yahoo! News.

This article is organized as follows. First we give an overview for our methodology with focus on the construction of the appearance model and its integration into a probabilistic retrieval model. Then, we describe and discuss our experiments and conclude with related research and prospects for future research.

2 Methods

2.1 The content models

The most simple content model of the text is made by tokenization of the text into words, which gives us a bag-of-word representatation (BOW-representation). A more advanced model considers only the nouns (including proper nouns) because in the search for persons and objects only nouns are important (bag-of-noun representation or BON). Part-of-speech tagging detects the syntactic word class and we use here the LTChunk tool[8]. In advanced content models we rely on more sophisticated natural language processing techniques. We perform pronoun resolution[1], word sense disambiguation [6], named entity recognition (NER)[2] and consider the visualness and salience of a noun phrase.

2.2 Computation of the visualness and salience

When we build an appearance model, entities that are not visual do not play a role because they cannot be part of an image. We compute the visualness (value between zero and one) of each noun and proper noun, where visualness is defined as the degree that a noun entity can be perceived visually by humans or a camera. Proper nouns that were classified as persons by the NER tool receive a visualness of 1. We compute the visualness of a common noun based on knowledge of the visualness of a few seed words and their semantic distance with the target nouns in WordNet.

[1] http://www.alias-i.com/lingpipe/
[2] Adaptation of Lingpipe NER tool.

San Francisco Giants' Barry Bonds, right, holds a bat while sitting in the dugout with Omar Vizquel, left, of Venezuela in the ninth inning against the Florida Marlins Tuesday, May 30, 2006 at Dolphin Stadium in Miami. Bonds did not play as the Marlins defeated the Giants 5-3.

Barry Bonds 0.75	bat 0.259
dugout 0.254	Omar Vizquel 0.214
Dolphin 0.172	Stadium 0.084

Fig. 1. Image-text pair (source: AP Photo/Yahoo! News) with the probabilities that the text entities appear in the image.

We also compute the salience (value between zero and one) of each noun and proper name, assuming that salient entities in texts that accompany images have a better chance of being present in the images. Computation of visualness and salience is described and evaluated in detail in [5].

2.3 Computation of the appearance model

We assume that entities found in a text T_j might be present in the accompanying image I_j, and that the probability of the occurrence of an entity e_i in the image, given a text T_j, $P(e_{i-im}|T_j)$, is proportional with the degree of visualness and salience of e_i in T_j. In our framework, $P(e_{i-im}|T_j)$ is computed as the product of the salience of the entity e_i and its visualness score, as we assume both scores to be independent, normalized by the sum of appearance scores of all entities in T_j. We have here used a very simple smooting method in order to counter errors in the named entity recognition, where we give all words which receive a zero score in the appearance model a fixed score of 0.01. $P(e_{i-im}|T_j)$ defines a ranking of the text's entities. Figure 1 gives an example of such a ranking generated from the text by our system.

In [5] the impact and a detailed error analysis of each step in the construction of the appearance model is given. It was shown that both the salience and visualness substantially contribute to an improved appearance model for describing and ranking the entities according to prominence in an accompanying image. We now want to find out how good this model is for discriminatively indexing the images in a cross-media retrieval task compared to more simpler models.

2.4 Integration in a probabilistic retrieval model

Statistical language modeling has become a successful retrieval modeling approach [4]. A textual document is viewed as a model and a textual query as a

string of text randomly sampled from that model. In case of our text-based image retrieval, the content model of image I_j is solely generated from the accompanying text T_j. Let the query be composed of one or more entities where the queried entity e_i is in the form of a person proper name or common noun representing a person or object. Our baseline appearance model considers a bag-of-words (BOW) representation of the text as content model of the image resulting in following retrieval model or ranking function:

$$P(e_1, ..., e_m|I_j) = \prod_{i=1}^{m}((1-\lambda)P(e_i|T_j) + \lambda P(e_i|C)) \quad (1)$$

where q_i is the ith query term in a query composed of m terms, and $P(e_i|T_j)$ is specified by the appearance model built from the text, and C represents the collection of documents. We estimate $P(e_i|T_j)$ by maximum likelihood estimation of the occurrence of the query term in the text. An intermediate model (BON) uses a bag-of-noun representation of the text and computes the probability that the text generates the entity $P(e_i|T_j)$ by maximum likelihood estimation of the occurrence of the query term in the text filtered by all words except nouns (including proper nouns). Both models do not take into account that an entity mentioned in the text can actually be shown in an image. In a limited way they consider salience as in longer texts the maximum likelihood of a term will be lower - especially when terms mostly occur only once - and thus the entities mentioned are likely to be less important.

We also integrate the appearance model (AP) described above:

$$P(e_i, ..., e_m|I_j) = \prod_{i=1}^{m}((1-\lambda)P(e_{i-im}|T_j) + \lambda P(e_i|C)) \quad (2)$$

Variations of this model only consider the factor salience (APS) or visualness (APV) when generating $P(e_{i-im}|T_j)$. We used the Lemur toolkit[3] and adapted it to suit our appearance models (AM). We used Jelinek-Mercer smoothing with the linear interpolation weight λ set to 0.1.

3 Experiments, results and discussion

3.1 The data collection, queries and ground-truth answer lists

Because of the lack of a standard dataset that fits our tasks and hypotheses, we annotated our own ground truth corpus. Our dataset from the Yahoo! News website[4] is composed of 700 image-text pairs. Every image has an accompanying news text which describes the content of the image. This text will in general discuss one or more persons in the image, possibly one or more other objects, the location and the event for which the picture was taken. Not all persons

[3] http://www.lemurproject.org/
[4] http://news.yahoo.com/

or objects who are pictured in the photograph are necessarily described in the news text. The inverse is also true. The texts are short and contain maximum 3 sentences. On average the texts have a length of 40.98 words, and contain 21.10 words that refer to noun phrase entities of which 2.77 refer to distinct persons and objects present in the image (see table 1 for the distribution of visible entities in the documents).

Because of the may images with only one person pictured, we refer to this dataset as the EASYSET. From this set we select a subset of pictures where three or more persons or objects are shown, which varying degree of prominence in the image. We call this dataset that comprises 380 image-text pairs the DIFFICULTSET (see example in figure 2). Tests on the latter set allows us to better understand the behavior of our different indexing methods when many persons or objects with varying degree of prominence are shown in the photographs.

Table 1. Number of image-text pairs for a given number of entities in the image.

Entities	0	1	2	3	4	5	6	7	≥ 8
Documents	2	168	353	151	133	47	17	8	24

We annotated the images with the names of the persons and objects shown, and ranked these entities according to prominence in the image. Queries were randomly generated from the manual annotations of the images and were filtered in order to have images in which the queried persons or objects were present at several levels of prominence. In this way we obtained 53 queries that contain one name of a person or object, and 26 queries with two entities (23 queries with two person names and 3 queries with a person and object name). Larger queries do not seem to make sense, as people often search for a picture of one person, perhaps a person with an object (e.g., car, flag), or 2 persons meeting each other.

For each query we generated a list of images sorted according to relevancy for the query (ground truth answer list) where the prominence of the entities in the image is taken into account and where we give priority to images with fewer persons or objects, and take into account the centrality and size of the person(s) or object(s) of interest.

Note that all queries have at least one relevant image in the data sets, which makes a comparison among the methods for finding the best picture more transparent, and, most importantly, that multiple images can occupy the same relevance rank in the ground truth answer list.

3.2 Evaluation

Our aim is to retrieve the best pictures, i.e., the images on rank 1 in the ground truth answer list for a certain query, on top of the machine generated list. We use the mean R-precision or R-recall where R is defined as the number of relevant pictures on rank 1 in the ground truth answer list. This corresponds with

U.S. President George W. Bush (2nd R) speaks to the press following a meeting with the Interagency Team on Iraq at Camp David in Maryland, June 12, 2006. Pictured with Bush are (L-R) Vice President Dick Cheney, Defense Secretary Donald Rumsfeld and Secretary of State Condoleezza Rice.

Fig. 2. Image-text pair (source: Reuters/Yahoo! News).

Table 2. Results in terms of Mean R-precision (MRP) and Mean Average Precision (MAP) for the ranking models based on the different text representations for the EASYSET and DIFFICULTSET where the query is composed of one entity.

Content model	MRP EASYSET	MAP EASYSET	MRP DIFFICULTSET	MAP DIFFICULTSET
BOW	53.84 %	56.90 %	50.00 %	70.48 %
BON	69.23 %	61.25 %	58.00 %	74.14 %
AP	57.69 %	59.28 %	60.00 %	75.57 %
APS	57.69 %	57.14 %	56.00 %	74.07 %
APV	61.54 %	57.00 %	60.00 %	75.27 %

precision@1 taking into account that the first position or rank might contain several best images. We also compute the classical average precision (AP) for the R relevant pictures. The above precision values are averaged over the queries and named in the tables below respectively as MRP and MAP.

3.3 Results

The results in terms of MRP and MAP are shown in tables 2 and 3. First, we see that the visualness measure in generally improves the retrieval model when the query is composed of one entity. This measure enables to determine how many entities in a given text are likely to appear in the image, and thus to create a more fine-grained ranking (since images with a small number of entities are prefered above images with a large number of entities). We see furthermore from tables 2 and 3 that this effect is most important when testing on the difficult set. This seems intuitive, since the difficult set contains only documents with large numbers of entities, for which it is important to determine what entities appear exactly in the image. The results also show that prominence is sufficiently captured by the maximum likelihood estimation of the term occurrence in the text. The longer the captions, the more content probably is shown in the image and the less important the individual entities in the image are. This simple heuristic yields good results when using captions for indexing images, while more advanced salience detection techniques are superfluous. When the queries

Table 3. Results in terms of Mean R-precision (MRP) and Mean Average Precision (MAP) for the ranking models based on the different text representations for the EASYSET and DIFFICULTSET where the query is composed of two entities.

Content model	MRP EASYSET	MAP EASYSET	MRP DIFFICULTSET	MAP DIFFICULTSET
BOW	53.85 %	60.60 %	69.23%	68.08 %
BON	69.23 %	64.93 %	73.07 %	72.08 %
AP	57.69 %	59.28 %	57.70 %	63.83 %
APS	57.69 %	54.44 %	61.54 %	60.30 %
APV	53.85 %	52.81 %	61.54 %	62.78 %

contain more terms, the simpler bag-of-words or bag-of nouns models have better retrieval performance, possibly explained by the fact that short caption texts that contain the query entities retrieve the best pictures.

4 Related and future work

Since the early days of image retrieval, text-based approaches are common because users often express an information need in terms of a natural language utterance. Especially in a Web context text-based image retrieval is important given that users are acquainted with keyword searches. Recognizing content in the image that relies on descriptions of surrounding texts is researched, for instance, by [9, 1]. [3] demonstrated the importance of content that surround the images on Web pages for their effective retrieval and have investigated how multiple evidence from selected content fields of HTML Web pages (e.g. meta tags, description tags, passages) contribute to a better indexing. Also [10] combine textual and visual evidence in Web image retrieval. The textual analysis in the above research does not go further than a bag-of-words representations scheme. The most interesting work here to mention is the work of Berg et al. [2] who also process the image-text pairs found in the Yahoo! news corpus. They consider pairs of person names recognized with named entity recognition (text) and faces (image) and use clustering with the Expectation Maximization algorithm to find all faces belonging to a certain person. Bayesian networks have been successfully used in image retrieval by [3] who integrate evidence of multiple fields of HTML Web pages. These authors found that a combination of description tags with a 40-term textual passage that most closely accompanies the image, provides best retrieval performance. However, still a bag-of-words approach is used. Our work can perfectly complement the above research as we provide a more accurate appearance model. The future lies in combining evidence from the different media relying on advanced technology for text and image analysis (cf. [7]). Our current and future work which combines visual and textual features goes in this direction. When we obtain evidence from different sources other probabilistic ranking models, such as inference or Bayesian networks are valuable [3].

5 Conclusions

We have built and tested several probabilistic appearance or content models from texts that accompany images. A simple bag-of-words approach is compared with a bag-of-nouns approach and with a more fine-grained identification of what content of the text can be visualized and of how prominent the content is in the image. The appearance models were integrated in a probabilistic retrieval function. The models based on more advanced text analysis taking into account syntactic, semantic and discourse analysis - although successful in automatically annotating the images - are not necessarily more discriminative for indexing purposes, except when querying a difficult data set for one person or object, where the images contain three or more persons or objects. A bag-of-nouns representation yielded overall the best results, especially when the query becomes more elaborated (with more entities), the overlap with the caption by using simpler representations is sufficient.

Acknowledgments

The work reported was supported by the CLASS project (EU-IST 027978). We thank Yves Gufflet (INRIA, France) for collecting the Yahoo! News dataset.

References

1. K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei, and M. I. Jordan. Matching Words and Pictures. *Journal of Machine Learning Research*, 3(6):1107–1135, 2003.
2. T. L. Berg, A. C. Berg, J. Edwards, and D. Forsyth. Who's in the Picture? In *Neural Information Processing Systems*, pages 137–144, 2004.
3. T. Coelho, P. Calado, L. Souza, and B. Ribeiro-Neto. Image Retrieval Using Multiple Evidence Ranking. *Image*, 16(4):408–417, 2004.
4. W. B. Croft and J. Lafferty. *Language Modeling for Information Retrieval*. Kluwer Academic Publishers, Boston, MA, 2003.
5. K. Deschacht and M. Moens. Text Analysis for Automatic Image Annotation. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 1000–1007, 2007.
6. K. Deschacht and M.-F. Moens. Efficient Hierarchical Entity Classification Using Conditional Random Fields. In *Proceedings of the 2nd Workshop on Ontology Learning and Population*, pages 33–40, Sydney, July 2006.
7. W. H. Hsu, L. Kennedy, and S.-F. Chang. Reranking methods for visual search. *IEEE Multimedia Magazine*, 13(3), 2007.
8. A. Mikheev. Automatic Rule Induction for Unknown-Word Guessing. *Computational Linguistics*, 23(3):405–423, 1997.
9. Y. Mori, H. Takahashi, and R. Oka. Automatic Word Assignment to Images Based on Image Division and Vector Quantization. In *RIAO-2000 Content-Based Multimedia Information Access*, Paris, April 12-14 2000.
10. S. Tollari and H. Glotin. Web image retrieval on ImagEVAL: Evidences on visualness and textualness concept dependency in fusion model. In *ACM International Conference on Image and Video Retrieval (CIVR)*, July 2007.