MUFIN at ImageCLEF 2011: Success or Failure?

Petra Budikova, Michal Batko, and Pavel Zezula

Masaryk University, Brno, Czech Republic
{budikova,batko,zezula}@fi.muni.cz

Abstract. In all fields of research it is important to discuss and compare various methods that are being proposed to solve given problems. In image retrieval, the ImageCLEF competitions provide such comparison platform. We have participated in the Photo Annotation Task of the ImageCLEF 2011 competition with a system based on the MUFIN Annotation Tool. Our approach is, in contrast to typical classifier solutions, based on a general annotation system for web images that provides general keywords for arbitrary image. However, the free-text annotation needs to be transformed into the 99 concepts given by the competition task. The transformation process is described in detail in the first part of this paper. In the second part, we discuss the results achieved by our solution. Even though the free-text annotation approach was not as successful as the classifier-based approaches, the results are competitive especially for the concepts involving real-world objects. On the other hand, our approach does not require training and is scalable to any number of concepts.

1 Introduction

In the course of the past few decades, multimedia data processing has become an integral part of many application fields, including medicine, art, security, etc. This poses a number of challenges to the computer science – we need techniques for efficient data representation, storing and retrieval. In many applications, it is also necessary to understand the semantic meaning of a multimedia object, i.e. to know what is represented in a picture or what a video is about. Such information is usually expressed in a textual form, e.g. as a text description that accompanies the multimedia object. The semantic description of an object can be obtained manually or (semi)-automatically. Since the manual annotation is an extremely labor-intensive and time-consuming task for larger data collections, automatic annotation or classification of multimedia objects is of high importance. Perhaps the most intensive is the research on automatic annotation of images, which is essential for semantic image retrieval [4].

In all fields of research it is important to discuss and compare various methods that are being proposed to solve given problems. In image retrieval, the ImageCLEF competitions provide such comparison platform. Each year, a set of specific tasks is defined that reflects the most challenging problems of current research. In 2011 as well as in the two previous years, one of the challenges was the Photo Annotation Task.

This paper presents the techniques used by the MUFIN group to handle the Annotation Task. We believe that our methods will be interesting for the community since our approach is different from the solutions presented in previous years [10]. Rather than creating a solution tailored for this specific task we employed a general purpose annotation system and used the returned keywords to identify the relevant concepts. The results show that while this approach logically lags behind the precision of the more finely tuned solutions, it is still capable of solving some instances pretty well.

The paper is organized as follows. First, we briefly review the related work on image annotation and distinguish two important classes of annotation methods. Next, the ImageCLEF Annotation Task is described in more detail and the quality of the training data is discussed. Finally, we present the methods used by the MUFIN group, analyze their performance and discuss the results.

2 Related Work

The reason why we are interested in annotation is to simplify access to the multimedia data. Depending on a situation, different types of metadata may be needed, both in content and form. In [6], three forms of annotation are discussed: free text, keywords chosen from a dictionary, and concepts from some ontology. While the free text annotation does not require any structure, the other two options pose some restrictions on the terminology used, and in particular make the selection of keywords smaller. On certain conditions, we then begin to call the task *classification* or *categorization* rather than *annotation*.

Even though the conditions are not strictly defined, the common understanding is that classification task works with a relatively small number of concepts and typically uses machine learning to create specialized classifiers for the given concepts. To train the classifiers, a sufficiently large training dataset with labeled data needs to be available. The techniques that can be engaged in the learning are numerous, including SVMs, kNN classifiers, or probabilistic approaches [7]. Study [10] describes a number of classification setups using different types of concept learning.

On the contrary, annotation usually denotes a task where a very large or unlimited number of concepts is available and typically no training data is given. The solution to such task needs to exploit some type of data mining, in case of image annotation it is often based on content-based image retrieval over collections of images with rich metadata. Such system is described for example in [8]. Typically, the retrieval-based annotation systems exploit tagged images from photo-sharing sites.

3 ImageCLEF Photo Annotation Task

In the ImageCLEF Photo Annotation Task, the participants were asked to assign relevant keywords to a number of test images. The full setup of the contest is described in the Task overview [11]. From our perspective, two facts are important: (1) the keywords to be assigned were chosen from a fixed set of 99 concepts, and (2) a set of labeled training images was available. Following our definition of terms from the previous section, the task thus qualifies as a classification problem. As such, it is most straightforwardly solved by machine learning approaches, using the training data to tune the parameters of the model. The quality of the training data is then crucial for the correctness of the classification.

As explained in [10], it is difficult to obtain a large number of labeled images both for training and contest evaluation. It is virtually impossible to gather such data only with the help of a few domain experts. Therefore, only a part of the data was labeled by domain experts. The rest was annotated in a crowdsourcing way, using workers from the Amazon Mechanical Turk portal. Even though the organizers of the contest did their best to ensure that only sane results would be accepted, the gathered data still contain some errors. In the following, we would like to comment on some of them that we have noticed, so that they could be corrected in the future. Also, we will discuss later how these errors may have influenced the performance of the annotation methods.

3.1 Training data deficiencies

During the preparation of our solution for the Photo Annotation Task, we have identified to following types of errors in the labeled training data:

- *Logical nonsense*: Some annotations in the training dataset contradict the laws of the real world. The most significant nonsense we found was a number of images with the following triplet of concepts: `single_person`, `man`, `woman`. Such combination appeared for 784 images. Similarly, `still_life` and `active` are concepts that do not match together. Though the emotional annotations are more subjective and cannot be so easily discarded as nonsense, we also believe that annotating an image by both `cute` and `scary` concepts is an oxymoron.
- *Annotation inconsistence*: Since the contest participants were not provided by any explanation of the concepts, it was not always clear to us what counts as relevant for a given concept and what does not. One such unclear concept was `single_person`. Should a part of a body be counted as a person or not? It seems that this question was also unclear to the trainset annotators as the concept `bodypart` sometimes co-occurred with `single_person` and sometimes with `no_person`.
- *Concept overuse*: The emotional and abstract concepts are definitely difficult to assign. Even within our research group, we could not decide what determines the "cuteness" of an image or what is the definition of a "natural" image. Again, the labeled training data were not of much help as we could not discover any inner logic in them. We suspect that the Amazon Turk workers solved this problem by using the emotional and abstract terms as often as possible. For illustration, from among the 8000 training images 3910 were labeled `cute`, 3346 were considered to be `visual_arts` and 4594 images were perceived as `natural`.

Each error type is illustrated by a few examples in Figure 1.

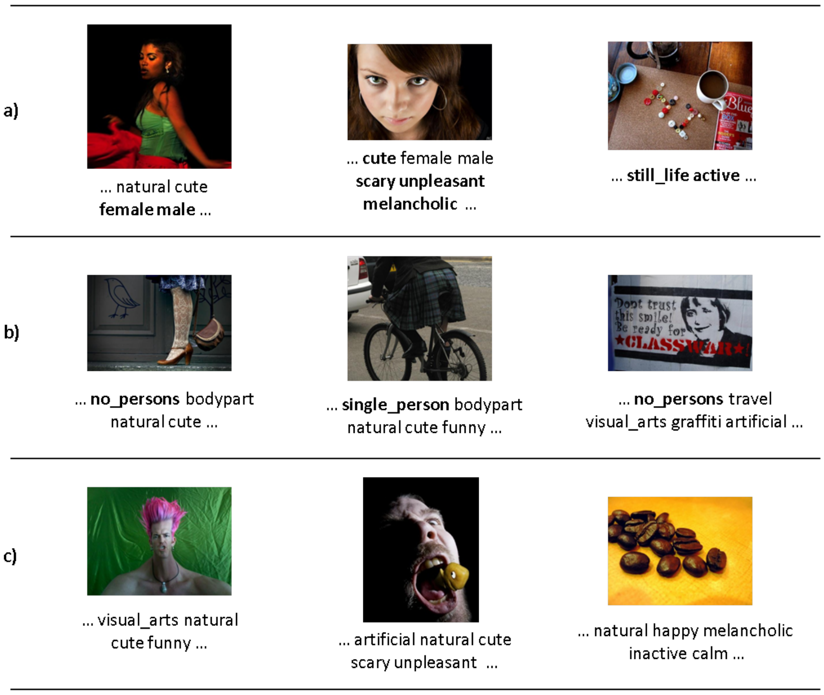
a)
... natural cute female male ...
... cute female male scary unpleasant melancholic ...
... still_life active ...
b)
... no_persons bodypart natural cute ...
... single_person bodypart natural cute funny ...
... no_persons travel visual_arts graffiti artificial ...
c)
... visual_arts natural cute funny ...
... artificial natural cute scary unpleasant ...
... natural happy melancholic inactive calm ...

Fig. 1. Trainset labeling errors: a) logical nonsense, b) annotation inconsistence, c) concept overuse.

4 Our solution

The MUFIN Annotation Tool came into existence in the beginning of this year as an extension of the MUFIN Image Search[1], a content-based search engine that we have been developing for several years. Our aim was to provide an online annotation system for arbitrary web images. The first prototype version of the system is available online[2] and was presented in [3]. As the first experiments with the tool provided promising results, we decided to try our tool in the ImageCLEF Annotation Task.

The fundamental difference in the basic orientation of MUFIN Annotation Tool and the Annotation Task is that our system provides *annotations* while the task asks for *classification*. Our system provides free-text annotation of images,

[1] http://mufin.fi.muni.cz/imgsearch/

[2] http://mufin.fi.muni.cz/annotation/

using any keywords that seem relevant using the content-based searching. To be able to use our tool for the task, we needed to transform the provided keywords into the restricted set of concepts given by the task. Moreover, even though the MUFIN tool is quite good at describing the image content it does not give much information about emotions and technical-related concepts (we will discuss the reasons later). Therefore, we also had to extend our system and add new components for specialized processing of some concepts. The overall architecture of the processing engaged in the Annotation Task is depicted in Figure 2. In the following sections, we will describe the individual components in more detail.

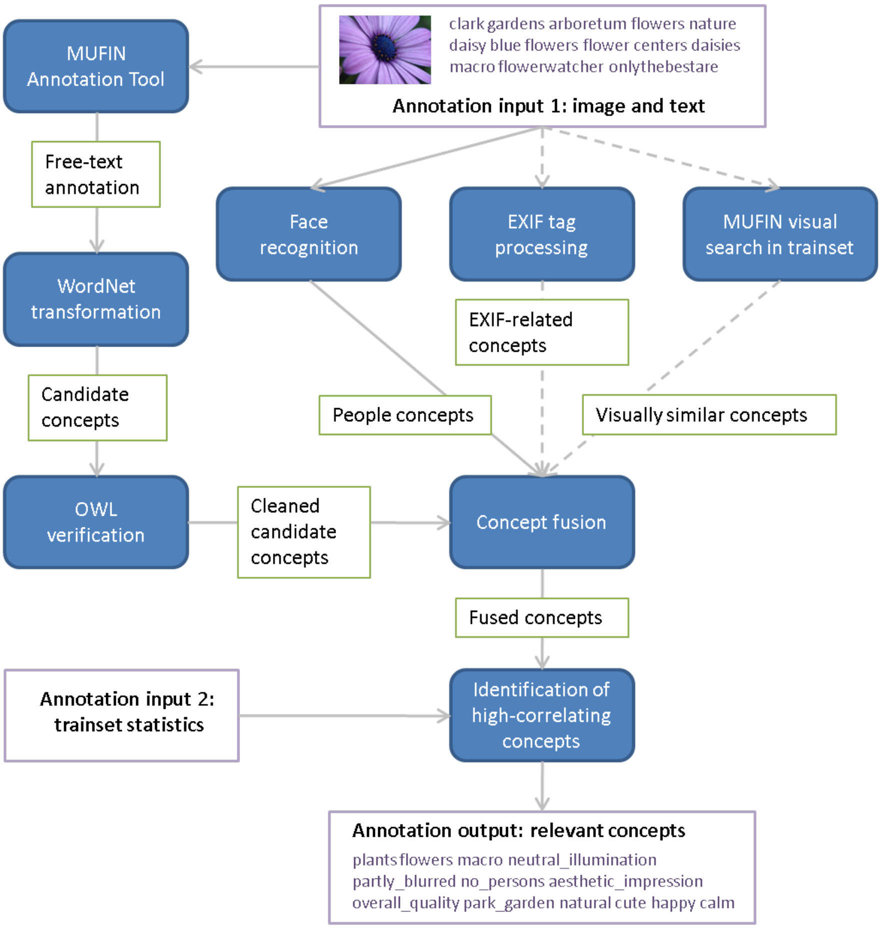
MUFIN Annotation Tool
clark gardens arboretum flowers nature daisy blue flowers flower centers daisies macro flowerwatcher onlythebestare
Annotation input 1: image and text
Free-text annotation
Face recognition
EXIF tag processing
MUFIN visual search in trainset
WordNet transformation
EXIF-related concepts
Candidate concepts
People concepts
Visually similar concepts
OWL verification
Cleaned candidate concepts
Concept fusion
Fused concepts
Annotation input 2: trainset statistics
Identification of high-correlating concepts
Annotation output: relevant concepts
plants flowers macro neutral_illumination partly_blurred no_persons aesthetic_impression overall_quality park_garden natural cute happy calm

Fig. 2. Concept retrieval schema.

4.1 MUFIN Annotation Tool

The MUFIN Annotation Tool is based on the MUFIN Image Search engine, which retrieves the nearest neighbors of a given image based on visual and text similarity. The image search system, described in full detail in [1], enables fast retrieval of similar images from very large collections. The visual similarity is defined by five MPEG-7 global descriptors – Scalable Color, Color Structure, Color Layout, Edge Histogram, and Region Shape – and their respective distance functions [9]. If some text descriptions of images are available, their tf-idf similarity score is also taken into consideration. In the ImageCLEF contest, free-text image descriptions and EXIF tags were available for some images. Together with the image, these were used as the input of the retrieval engine.

To obtain an annotation of some input image, we first evaluate the nearest neighbor query over a large collection of high-quality images with rich and trustworthy text metadata. In particular, we are currently using the Profimedia dataset, which contains 20M images from a microstock site [2]. When the query is processed by the MUFIN search, we obtain a set of images with their respective keywords. In the Profimedia dataset, each image is accompanied by a set of title words (typically 3 to 10 words) and keywords (about 20 keywords per image in average). Both the title words and the keywords of all images in the result set are merged together (the title words receiving a higher weight) and the frequencies of individual lemmas are identified. A list of stopwords and the WordNet lexical database [5] are used to remove irrelevant word types, names, etc. The Annotation Tool then returns the list of the most frequent keywords with the respective frequencies, which express the confidence of the annotation.

4.2 Annotation to concept transformation

To transform the free-text annotation into the ImageCLEF concepts it was necessary to find the semantic relations between the individual keywords and concepts. For this purpose, we used the WordNet lexical database, which provides structured semantic information for English nouns, verbs, adjectives, and adverbs. The individual words are grouped into sets of cognitive synonyms (called synsets), each expressing a distinct concept. These synsets are interlinked by different semantic relations, such as hypernym/hyponym, synonym, meronym, etc. It is thus possible to find out whether any two synsets are related and how. In our case, we were interested in the relationships between the synsets of our annotation keywords and the synsets of the ImageCLEF concepts. To obtain them, we first needed to find the relevant synsets for our keywords and the task concepts.

The WordNet synset is defined as a set of words with the same meaning, accompanied by a short description of the respective semantic concept. Very often, a word has several meanings and therefore is contained in several synsets. For instance, the word "cat" appears both in a synset describing the domestic animal and a synset describing an attractive woman. If we have only a keyword and no other information about the word sense or context, we need to consider all synsets that contain this keyword. This is the case of keywords returned by the MUFIN Annotation Tool. We do not try here to determine whether the synsets are really relevant but rely on the majority voting of a large number of keywords that are processed.

The situation is however different in case of the ImageCLEF concepts where it is much more important to know the correct synsets. Fortunately, we have actually two possible ways of determining the relevant synsets. First, we can sort them out manually since the number of concepts is relatively small. The other, more systematic solution, will run the whole annotation process with all the candidate synsets of the concepts, log the contributions of the respective individual synsets, evaluate their performance, and rule out those with a low success rate.

Once the synsets are determined, we can look for the relationships. Again, there are different types of relations and some of them are relevant for one concept but irrelevant for another. Again, we have the same two possibilities of choosing the relevant relationships as in the case of synsets. In our implementation, we have used the manual cleaning approach for synsets and automatic selection approach for relationships.

With the relevant synsets and relationships, we count a relevance score of each ImageCLEF concept during the processing of each image. The score is increased each time a keyword-synset is related to concept-synset. The increase is proportional to the confidence score of the keyword as produced by the MUFIN Annotation Tool.

Finally, the concepts are checked against the OWL ontology provided within the Annotation Task. The concepts are visited in a decreasing order of their scores and whenever a conflict between two concepts is detected, the concept with a lower score is discarded.

4.3 Additional image processing

The mining in keywords of similar images allows us to obtain such information as is usually contained in the image descriptions. This is most often related to image content, so the concepts related to nature, buildings, vehicles, etc. can be identified quite well. However, the Annotation Task considers also concepts that are less often described in the text. To get some more information about these, we employed the following three additional information sources:

- Face recognition: The face recognition algorithms are well-known in the image processing. We employ face recognition to determine the number of persons in an image.
- EXIF tag processing: Some of the input photos are accompanied by EXIF tags that provide information about various image properties. When available, we use these tags to decide the relevance of concepts related to illumination, focus, and time of the day.
- MUFIN visual search: Apart from the large Profimedia collection, we also have the training dataset that can be searched with respect to the visual

similarity of images. However, since the trainset is rather small and some concepts are represented by only a few images, there is quite a high probability that the nearest neighbors will not be relevant. Therefore, we only consider neighbors within a small range of distances (determined by experiments).

4.4 Trainset statistics input

Definitely the most difficult concepts to assign are the ones related to user's emotions and also the abstract concepts such as `technical`, `overall_quality`, etc. As discussed in Section 3.1, it is not quite clear either to us or to the people who annotated the trainset what these concepts precisely mean. Therefore, it is very difficult to determine their relevance using the image visual content. The text provided with the images is also not helpful in most cases.

We finally decided to rely on the correlations between image content and the emotions it most probably evokes. For example, images of babies or nature are usually deemed cute. A set of such correlation rules was derived from the trainset and used to choose the emotional and abstract concepts.

4.5 Our submissions at ImageCLEF

In our primary run, all the above-described components were integrated as depicted in Figure 2. In addition, we submitted three more runs where we tried various other settings to find out whether the proposed extensions really provided some added quality to the search results. The components that were left out in some experiments are depicted by dashed lines in Figure 2. We also experimented with the transformation of our concept scores into the confidence values expressed as percentage, which was done either as concept-specific or concept-independent. The individual run settings were as follows:

- MufinSubmission100: In this run, all the available information was exploited including photo tags, EXIF tags, visual image descriptors, and trainset statistics. Concept-specific mapping of annotation scores to confidence values was applied.
- MufinSubmission101: The same settings were used for this run as in the previous case but concept-independent mapping of annotation scores was applied.
- MufinSubmission110: In this run, we did not use the EXIF tags for the processing of concepts related to daytime and illumination as described in Section 4.3. The MUFIN visual search in the trainset was omitted as well. However, the textual EXIF tags were used as a part of the input for the MUFIN Annotation Tool. Concept-independent mapping of annotation scores was applied.
- MufinSubmission120: In this run, the EXIF tags were not applied at all, neither as a part of the text-and-visual query in the basic annotation step nor in the additional processing. Again, the MUFIN visual search was skipped. Concept-independent mapping of annotation scores was applied.

5 Discussion of results

As detailed in [11], the following three quality metrics were evaluated to compare the submitted results: Mean interpolated Average Precision (MAP), F-measure (F-ex), and Semantic R-Precision (SR-Precision). As we expected, our best submission was MufinSubmission100 which achieved 0.299 MAP, 0.462 F-ex, and 0.628 SR-precision. The other submissions that we have tried received slightly worse scores. After the task evaluation and the release of the algorithm for computing the MAP metric, we also re-evaluated our system with better settings of the MUFIN Annotation Tool that we have improved since the ImageCLEF task submission. Using these, we were able to gain one or two percent increase of the MAP score. With respect to the MAP measure, our solution ranked at position 13 among the 18 participating groups.

Apart from the overall results, it is also interesting to take a closer look at the performance of the various solutions for individual concepts. The complete list of concept results for each group is available on the ImageCLEF web pages[3]. Here we focus on the categories and particular examples of concepts where MUFIN annotation performed either well or poorly and discuss the possible reasons.

First of all, we need to define what we consider a good result. Basically, there are two ways: either we only look at the performance, e.g. following the MAP measure, or we consider the performance in relation to the difficulty of assigning the given concept. The assigning difficulty can be naturally derived from the competition results – when no group was able to achieve high precision with some concept, then the concept is problematic. Since we believe that the second way is more suitable, we express our results as a percentage of the best MAP achieved for the given concept. Table 1 and Figure 3 summarize the results of MufinSubmission100 expressed in this way.

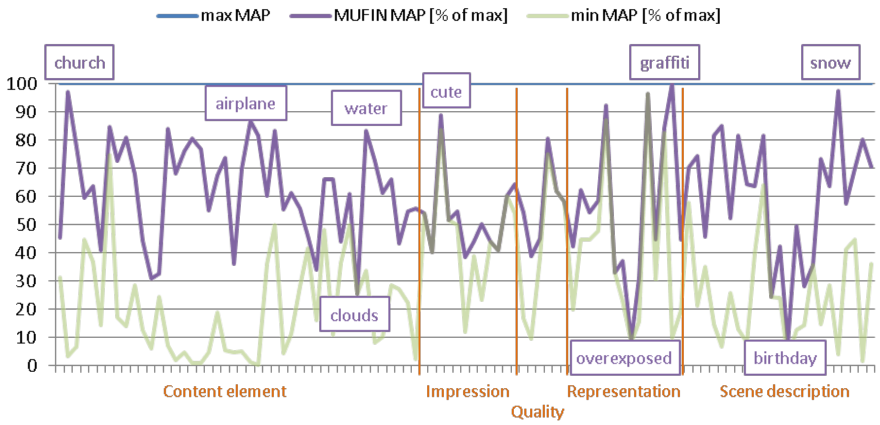

Fig. 3. MUFIN relative MAP performance per concept.

[3] http://imageclef.org/2011/Photo

Content element	Landscape elements	58.2 %	62.4 %
	Pictured objects	63.0 %	
	Urban elements	73.3 %	
Impression	Expressed impression	49.5 %	52.6 %
	Felt impression	58.6 %	
Quality	Aesthetics	60.0 %	56.3 %
	Blurring	54.6 %	
Representation	Art	58.6 %	56.4 %
	Impression	55.1 %	
	Macro	54.4 %	
	Portrait	62.1 %	
	Still life	42.0 %	
Scene description	Abstract categories	64.0 %	60.8 %
	Activity	49.4 %	
	Events	24.7 %	
	Place	72.5 %	
	Seasons	69.4 %	
	Time of day	67.9 %	

Table 1. MUFIN average relative MAP performance per category – averages over the concepts in the respective category are shown.

Table 1 shows the results averages in groups of semantically close categories as specified by the ontology provided for ImageCLEF. We can observe that the MUFIN approach is most successful in categories that are (1) related to visual image content rather than higher semantics, and (2) probable to be reflected in image tags. These are, in particular, the categories describing the depicted elements, landscape, seasons, etc. Categories related to impressions, events, etc. represent the other end of the spectrum; they are difficult to decide using only the visual information and (especially the impressions) are rarely described via tags.

However, the average MAP values do not differ that much between categories. The reason for this is revealed if we take a closer look at the results for individual concepts, as depicted in Figure 3. Here we can notice low peaks in otherwise well performing categories and vice versa. For instance, the `clouds` concept in the *landscapes* category performs rather poorly. This is caused by the fact that clouds appear in many images but only as a part of a background, which is not important enough to appear in the annotation. On the contrary, airplanes are more interesting and thus regularly appear in the annotations. In fact, we again encounter the difference between the annotation and classification tasks – in annotation we are usually interested in the most important/interesting tags while in classification all relevant tags are wanted.

Several more extremes are pointed out in Figure 3. For instance, the concept `cute` performs well because of its high frequency in the dataset. On the other hand, for the concept `overexposed` a specialized classifier is much more suitable than the annotation mining. The detailed discussion of the best fitting methods for individual categories is beyond the scope of this paper. However, we believe that it is worth further studies to sort out the different types of concepts as well as annotation approaches and try to establish some relationships between them.

6 Conclusions

In this study, we have described the MUFIN solution of the ImageCLEF Photo Annotation Task, which is based on free-text annotation mining, and compared it to more specialized, classifier-based approaches. The method we presented has its pros and cons. Mining information from annotated web collections is complicated by a number of features related to the way the annotations are created. As discussed in [12], we need to expect errors, typing mistakes, synonyms, etc. However, there are also ways of overcoming these difficulties. In our approach, we have exploited a well-annotated collection, the semantical information provided by WordNet, and a specialized ontology. Using these techniques, we have been able to create an annotation system that shows precision comparable to average classifiers, which are usually trained for specific purposes only. The main advantage of our solution lies in the fact that it requires minimum training (and is therefore less dependent on the availability of high-quality training data) and is scalable to any number of concepts.

Acknowledgments

This work has been partially supported by Brno PhD Talent Financial Aid and by the national research projects GAP 103/10/0886 and VF 20102014004. The hardware infrastructure was provided by the METACentrum under the programme LM 2010005.

References

1. Batko, M., Falchi, F., Lucchese, C., Novak, D., Perego, R., Rabitti, F., Sedmidubsky, J., Zezula, P.: Building a web-scale image similarity search system. Multimedia Tools and Applications 47, 599–629 (2010), http://dx.doi.org/10.1007/s11042-009-0339-z, 10.1007/s11042-009-0339-z
2. Budikova, P., Batko, M., Zezula, P.: Evaluation platform for content-based image retrieval systems. In: To appear in Theory and Practice of Digital Libraries (TPDL 2011) (26-28 September 2011)
3. Budikova, P., Batko, M., Zezula, P.: Online image annotation. In: 4th International Conference on Similarity Search and Applications (SISAP 2011). pp. 109–110 (2011)
4. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Image retrieval: Ideas, influences, and trends of the new age. ACM Comput. Surv. 40(2) (2008)
5. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database. The MIT Press (1998)
6. Hanbury, A.: A survey of methods for image annotation. J. Vis. Lang. Comput. 19(5), 617–627 (2008)
7. Kwasnicka, H., Paradowski, M.: Machine learning methods in automatic image annotation. In: Advances in Machine Learning II, pp. 387–411 (2010)

8. Li, X., Chen, L., 0001, L.Z., Lin, F., Ma, W.Y.: Image annotation by large-scale content-based image retrieval. In: Nahrstedt, K., Turk, M., Rui, Y., Klas, W., Mayer-Patel, K. (eds.) ACM Multimedia. pp. 607–610. ACM (2006)
9. MPEG-7: Multimedia content description interfaces. Part 3: Visual. ISO/IEC 15938-3:2002 (2002)
10. Nowak, S., Huiskes, M.J.: New strategies for image annotation: Overview of the photo annotation task at imageclef 2010. In: Braschler, M., Harman, D., Pianta, E. (eds.) CLEF (Notebook Papers/LABs/Workshops) (2010)
11. Nowak, S., Nagel, K., Liebetrau, J.: The CLEF 2011 Photo Annotation and Concept-based Retrieval Tasks. In: CLEF 2011 working notes (2011)
12. Trant, J.: Studying social tagging and folksonomy: A review and framework. J. Digit. Inf. 10(1) (2009)