Search-based Image Annotation: Extracting Semantics from Similar Images

Petra Budikova, Michal Batko, Jan Botorek, and Pavel Zezula

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

Abstract. The importance of automatic image annotation as a tool for handling large amounts of image data has been recognized for several decades. However, working tools have long been limited to narrow-domain problems with a few target classes for which precise models could be trained. With the advance of similarity searching, it now becomes possible to employ a different approach: extracting information from large amounts of noisy web data. However, several issues need to be resolved, including the acquisition of a suitable knowledge base, choosing a suitable visual content descriptor, implementation of effective and efficient similarity search engine, and extraction of semantics from similar images. In this paper, we address these challenges and present a working annotation system based on the search-based paradigm, which achieved good results in the 2014 ImageCLEF Scalable Concept Image Annotation challenge.

1 Introduction

Acquiring and storing images is very easy nowadays – anyone with a decent mobile phone can take a picture and upload it to a web gallery in a few seconds. However, organizing and retrieving such data remains a challenging task. The most natural way of accessing data is a text search, but a lot of images are not associated with any text information. Therefore, automatic image annotation methods are being developed to improve the accessibility of visual information.

The image annotation task can be formalized as follows: given an *input image*, which may or may not be accompanied by *input metadata*, select suitable descriptive words from a given *vocabulary*. Depending on the target application, the annotation vocabulary may contain a few labels, or all words from a given language. In this paper, we focus on the problem of broad-domain annotation with no input metadata and large vocabularies, which applies to the above-mentioned task of annotating web images.

To address this problem, we have developed a search-based annotation system which exploits labeled web images to determine the annotation of an arbitrary input image. Such approach is not useful for narrow-domain classification tasks with few candidate classes, which are better served by traditional machine learning techniques. However, the search-based solution can be successfully used for broad-domain annotation tasks with sparse training data, as demonstrated by the success of our system in the ImageCLEF 2014 Image Annotation challenge.

The search-based annotation paradigm is based on techniques for content-based data retrieval. Visual similarity of image content is exploited to search for images similar to the picture being annotated, and textual metadata of the resulting images are used to form the annotation. While the idea of search-based annotation is rather straightforward, it is not easy to achieve satisfactory results. The challenges that need to be solved are several: acquisition of suitable image set for similarity searching, choosing a suitable visual content descriptor, implementation of effective and efficient search engine, and extraction of semantics from similar images. In the following sections, we address all these issues and propose a novel technique for analysis of image semantics.

The rest of the paper is organized as follows. In Section 2, we review recent work in the field of image annotation. Next, we introduce our annotation system and describe its components. The ImageCLEF 2014 annotation task is introduced in Section 4 and our results from the competition are analyzed in Section 5. Section 6 concludes the paper and outlines our future work.

2 Related Work

Recent work in the field of image annotation can be divided into two categories – model-based and search-based. Model-based techniques, which are surveyed in more detail e.g. in [21], require a training dataset consisting of reliably annotated images, which are used compute a statistical model for each concept. The state-of-the art model-based solution is represented by the neural network classifier developed by Alex Krizhevsky for the 2012 ImageNet challenge, which defeated other participants of the contest by a significant margin and achieved impressive results [12]. However, any model-based solution is limited in terms of vocabulary scalability: the classifiers can be created only for concepts for which reliable training data is available, and every new concept requires costly re-training.

On the other hand, search-based solutions sacrifice precision for broad applicability and attempt to utilize the voluminous but potentially erroneous information available in web image collections and social networks. The authors of [14] presented a simple solution based on this idea, which straightforwardly takes the tags from the most similar images and assigns them to the input image. The Arista system [20] exploits efficient duplicate search over a very large reference data set to select the most relevant images for annotation mining. In [1], a learning procedure is proposed which projects both visual and textual words into a latent meaning space, and the learned mapping is used to find nearest neighbors for annotation. Many works focus on advanced methods of extracting relevant keywords for visual-neighbor annotations, which include web search [22], analysis of co-occurring words [10], or concept ranking by random walks in similarity graphs [22]. Recently, several authors have also proposed to utilize semantic knowledge sources such as ontologies for improving annotation quality [11, 18]. In our approach, we combine the basic strategy of [14] with semantic knowledge bases and co-occurrence analysis similar to [10, 18]. The main improvement over existing work is a novel semantics-aware keyword selection process.

3 Semantic Search-based Image Annotation

The annotation task may take many forms, as it appears in diverse applications that have different requirements on annotation vocabulary, efficiency, or flexibility. While most existing solutions focus on a single instance of the annotation problem, we believe that a more universal system can be designed that would be capable of adapting to diverse requirements. In our previous paper [2], we proposed a modular architecture for such system which allows to flexibly combine different image- and text-processing components.

In this paper, we present an instance of this architecture developed for broad-domain image annotation. Its fundamental modules and the flow of the data among them are schematically depicted in Figure 1 starting with a plain input image and finishing with the automatically generated annotation. There are four main phases of the annotation process. In the first phase, the annotation tool retrieves visually similar images from a suitable image collection. Then, the textual descriptions from the retrieved similar images are processed. Resulting sets of candidate keywords are analyzed using the WordNet lexical database and other sematic resources. Finally, the most probable concepts from the annotation vocabulary are selected as the final image description. In the following, we provide more details about the specific implementations of the respective parts and discuss different parameters of the annotation system that influence the overall performance.

3.1 Retrieval of Similar Images

The search-based approach to image annotation is based on the assumption that in a sufficiently large collection, images with similar content to any given query image are likely to appear. If these can be identified by a suitable content-based retrieval technique, their metadata such as accompanying texts, labels, etc. can be exploited to obtain text information about the query image. Important factors that influence the performance of search-based annotation are the reference collection size, reliability of reference image annotations, the quality of visual similarity measure, and the implementation of the similarity search engine.

Datasets The choice of image collection(s) over which the content-based retrieval is evaluated is a crucial factor of the whole annotation process. There should be as many images as possible in the chosen collection, the images should be relevant for the domain of the queries, and their descriptions should be rich and precise. Naturally, these requirements are in a conflict – while it is relatively easy to obtain large collections of image data (at least in the domain of general-purpose images appearing in personal photo-galleries), it is very difficult to automatically collect images with high-quality descriptions.

At the moment, our annotation system uses the Profiset image collection [5] as the baseline reference dataset. If additional training images are available for a specific task, they are added to this collection. The Profiset collection is freely available for research purposes and contains 20M high-quality images with rich

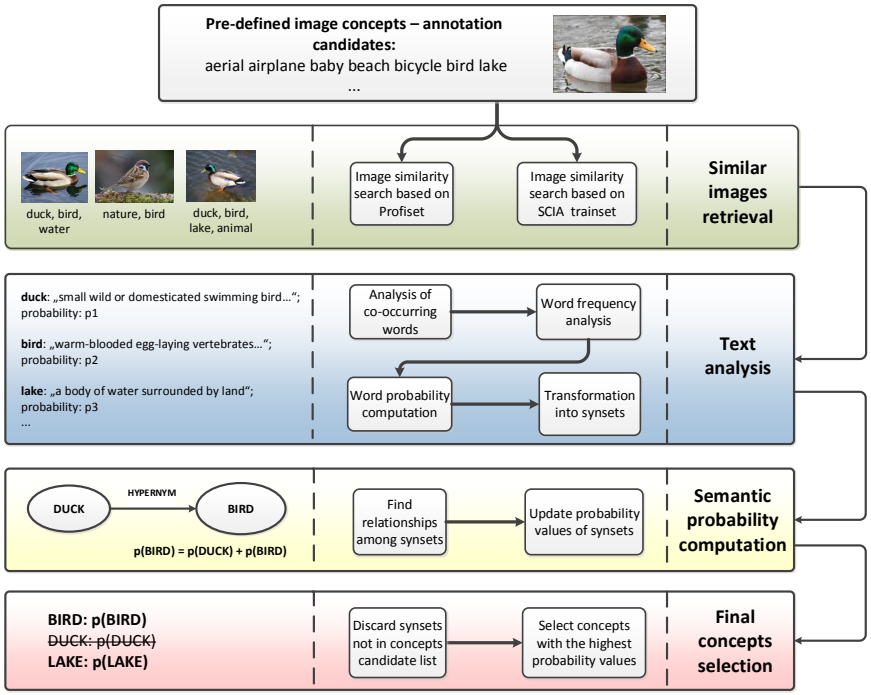

Fig. 1. Annotation tool architecture

annotations (about 20 keywords per image in average) obtained from a photo-stock website. The Profiset annotations have no fixed vocabulary and their quality is not centrally supervised, however the authors of annotations were interested in selling their photos and thus motivated to provide relevant keywords.

Visual descriptors Visual content descriptors and associated similarity function are used to evaluate the visual similarity of images. The content-based retrieval engine we employ can work with any descriptors that satisfy the metric space postulates, i.e. the similarity function is reflexive, symmetric, and satisfies the triangle inequality. Historically, the MPEG7 [15] multimedia standard defined several global visual features which were known to provide reasonably effective results with high efficiency. The annotation tool thus, as one option, uses a combination of five MPEG7 visual descriptors according to the best configuration provided in [13].

Recently, new visual descriptors called DeCAF features were proposed in [8]. Based on the successful image classifier developed by Krizhevsky [12], these features have been shown to perform promisingly in various image processing tasks. Therefore, we decided to use them as another option for our similarity search module. Specifically, we utilize the DeCAF_7 feature, which is produced by the last hidden layer of the neural network classifier. The DeCAF_7 representation

of a single image consists of a 4096-dimensional vector of real numbers and its extraction is a rather heavy computational task [8]. However, once the descriptors are extracted from a dataset, they can be efficiently indexed and searched. To compute the distance of two DeCAF_7 features, we utilize the Euclidean distance.

Indexing and searching In our solution, we utilize the MUFIN similarity search system [3] to index and search images. The MUFIN system exploits state-of-the-art metric indexing structures and enables fast retrieval of similar images from very large collections. For the combination of the five MPEG7 descriptors, we employ the M-Index technique [16]. For the bigger DeCAF descriptors, which need to read more data from the disk, we use the PPP-Codes technique [17]. Both indexing structures allows us to search a collection of 20M images in 1-2 seconds.

For each image to be annotated, a fixed number k of most similar images is selected and used for further processing. The number k needed to be chosen carefully, as it influences the quality of results. If we could suppose that all found objects are relevant for the query, a high k would be advantageous. However, this is often not the case in similarity-based image retrieval, where semantically irrelevant images are likely to be evaluated as visually similar to the query. It is therefore necessary to determine such k that the selected images provide sufficient amount of information but do not introduce too much noise.

3.2 Text Processing

In the second phase of the annotation process, the descriptions of images returned by content-based retrieval need to be analyzed in order to select the most probable concepts from target vocabulary. During this phase, we utilize various semantic resources to reveal the common topics depicted in the images. In the current implementation, our solution relies mainly on the WordNet semantic structure. The following sections explain how we link keywords from similar images' annotations to WordNet synsets, how the probability of individual synsets is computed, and how the synsets are transformed into the final annotation.

Selection of initial keywords Having retrieved the set of similar images, we first divide their text metadata into separate words and compute the frequency of each word. This way, we obtain a set of *initial keywords*. For each keyword, we compute its *initial probability*, which depends on the frequency of the keyword in descriptions of similar images. Only the n most probable keywords are kept for further processing.

Matching keywords to WordNet The set of keywords with their associated probabilities contains rich information about query image content, but it is difficult to work with this representation since we have no information about semantic connections between individual words. Therefore, we need to transform the keywords into semantically connected objects. We have decided to base our furter analysis on the WordNet lexical database [9], which is a comprehensive semantic tool interlinking dictionary, thesaurus and language grammar book.

The basic building block of WordNet hierarchy is a *synset*, an object which unifies synonymous words into a single item. On top of synsets, different semantic relations are encoded in the WordNet structure.

Each initial keyword is therefore mapped to a corresponding WordNet synset. Since there are often more possible meanings of a given word and thus more candidate synsets, we use a probability measure based on the cntlist[1] frequency values to select the most probable synset for each keyword. The cntlist measure is based on the frequency of words in a particular sense in semantically tagged corpora and expresses a relative frequency of a given synset in general text. To avoid false dismissals, several highly probable synsets may be selected for each keyword. Each selected synset is assigned a probability value computed as a product of the WordNet normalized frequency and the respective keyword's initial probability.

Exploitation of WordNet relationships By transforming keywords into synsets, we are able to group words with the same meaning and thus increase the probability of recognizing a significant topic. Naturally, this can be further improved by analyzing semantic relationships between the candidate synsets. In our solution, we exploit four WordNet relationships to create a *candidate synset graph*: *hypernymy* – the generalization, is-a relationship; *hyponymy* – the specialization relationship, the opposite of hypernymy; *holonymy* – the has-parts relationship, upward direction in the part/whole hierarchy; and *meronymy* – the is-a-part-of relationship, the opposite of holonymy.

To build the candidate synset graph, we first apply the upward-direction relationships (i.e. hypernymy and holonymy) in a so-called *expansion mode*, when all synsets that are linked to any candidate synset by these relationships are added to the graph; this way, the candidate graph is enriched by upper level synsets in the potentially relevant WordNet subtrees. However, we are not interested in some of the upper-most levels that contain very general concepts such as *entity*, *physical entity*, etc. Therefore, we also utilize the Visual Concept Ontology (VCO) [4] in this step, which was designed as a complementary tool to WordNet and provides a more compact hierarchy of concepts related to image content. Synsets not covered by the VCO are considered to be too general and therefore are not included in the candidate graph.

After the expansion, the other two relationships are utilized in an *enhancement mode* that adds new links to the graph using relationships between synsets that already are in the graph. Finally, the candidate graph is submitted to an iterative algorithm that updates the probabilities of individual synsets so that synsets with high number of links receive higher probabilities and vice versa.

Final concept selection At the end of the candidate graph processing, the system produces a set of candidate synsets with updated probabilities. If the annotation vocabulary is unlimited, the m most probable synsets are displayed as the final annotation. Otherwise, the synsets are confronted with the annotation vocabulary and the m most probable concepts from the intersection are

[1] https://wordnet.princeton.edu/wordnet/man/cntlist.5WN.html

displayed. The parameter m can be provided by the user, otherwise an experimentally determined value is used that provides the optimal trade-off between annotation precision and recall.

4 ImageCLEF 2014 Annotation Challenge

In 2014, we entered the ImageCLEF Scalable Concept Image Annotation (SCIA) challenge [19] to compare our annotation system to other state-of-the-art solutions. This section briefly introduces the task and describes the necessary adjustments of our annotation system.

4.1 Scalable Concept Image Annotation Task

The SCIA challenge is a standard annotation task, where relevant concepts from a fixed set of candidate concepts need to be assigned to an input image. The *input images* are not accompanied by any descriptive metadata, so only the visual image content serves as annotation input. For each test image, there is a *list of SCIA concepts* from which the relevant ones need to be selected. Each concept is defined by one keyword and a link to relevant WordNet nodes.

As the 2014 SCIA challenge focused especially on the concept-wise scalability of annotation techniques, the participants were not provided with hand-labeled training data and were not allowed to use resources that require significant manual preprocessing. Instead, they were encouraged to exploit data that can be crawled from the web or otherwise easily obtained, so that the proposed solutions should be able to adapt easily when the list of concepts is changed. Accordingly, the training dataset provided by organizers consisted of 500K images downloaded from the web, and the accompanying web pages. The raw images and web pages were further preprocessed by competition organizers to ease the participation in the task, resulting in several visual and text descriptors as detailed in [19].

The actual competition task consisted of annotating 7291 images with different concept lists. Altogether, there were 207 concepts, with the size of individual concept lists ranging from 40 to 207 concepts. Prior to releasing the test image set, participants were provided with a development set of query images and concept lists, for which a ground truth of relevant concepts was also published. The development set contained 1940 images and only 107 concepts.

4.2 DISA Participation

Our annotation system entered the competition under the name DISA, referring to the name of our lab. The DISA solution consisted of the system described in Section 3 with one minor extension – the 500K set of training images provided by SCIA organizers (the SCIA trainset) was used as a second collection of images for similarity searching. In comparison to Profiset, the SCIA trainset is smaller and the quality of text data is much lower; on the other hand, it has been designed to contain images for all keywords from the SCIA task concept lists, which makes it a very good fallback for topics not sufficiently covered in Profiset.

5 Evaluation

Participation in the SCIA challenge allowed us to compare our system to other solutions and also to evaluate the performance of various settings of our system. As explained in Section 3, the annotation tool has multiple components that have various parameters. The following sections describe the most interesting findings, more details can be found in the reports on DISA participation in SCIA 2014 [7, 6]. Let us also mention that the implementation with DeCAF descriptors was not ready before the SCIA competition deadline, therefore it did not enter the competition. However, the organizers kindly agreed to evaluate the DeCAF implementation for us afterward (out of the contest).

The quality of annotations was measured in terms of precision (P), recall (R), F-measure (F), and mean average precision (MAP). All these measures can be computed from two different perspectives: concept-based and sample-based. A concept-based precision (or any other measure) is computed for each concept, whereas sample-based precision is computed for each image to annotate. In both cases, the arithmetic mean was used as a global measure of performance. More details about the measures can be found in [19].

Visual descriptors As expected, the choice of visual descriptors used in the similarity searching phase is crucial for the overall performance of the annotation system. Using the cutting-edge DeCAF_7 features, the quality of results was 10-20 % higher than with older MPEG7 features. The values of individual measures are provided in Table 1.

Knowledge base size and quality To analyze the influence of dataset size and quality on the annotation system performance, we utilized several test image collections that were employed in the similarity search phase. Apart from the SCIA 500K dataset and Profiset 20M, we created random subsets of Profiset with 500K, 2M and 5M images. The performance of the annotation system on individual datasets is depicted in Figure 2. For each set of experiments, optimal settings of the semantic analysis phase were chosen so that the influence of similarity search parameters is clearly visible.

The first two groups of results compare the performance of DeCAF on SCIA 500K and Profiset 500K. We can clearly see that the higher-quality Profiset database provides better results in all three metrics. For both collections, the result quality grows with number k of similar images taken into consideration.

The following result groups provide comparison of DeCAF performance on high-quality datasets of different sizes. We can observe that increasing dataset size continually improves the result quality, so we can assume that even better results could be achieved if we had a larger reference dataset with high-quality data. Again, better results are generally achieved for larger k.

Finally, the last group of results depicts the results achieved by combination of Profiset 20M and SCIA 500K data. The slight improvement over Profiset 20M is caused by the fact that the SCIA 500K dataset covers all topics considered

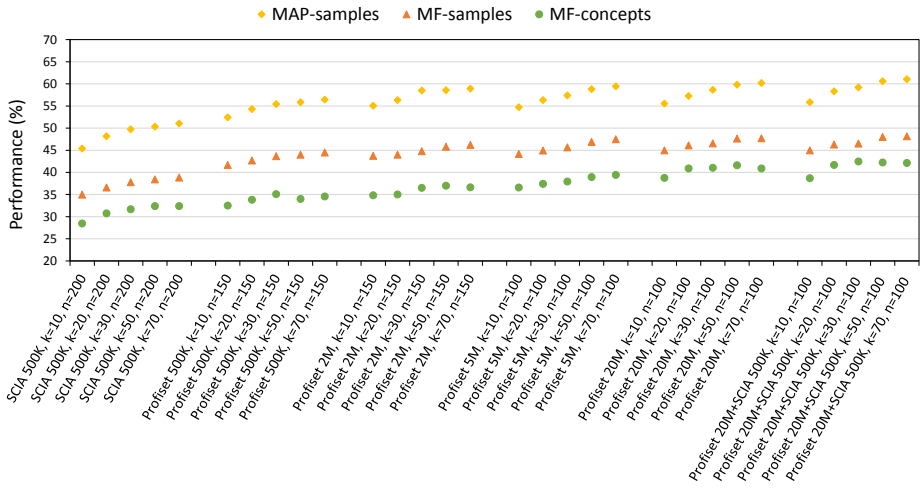

Fig. 2. Influence of the dataset quality and size on the annotation performance.

in the annotation task. This increases the chance of correctly identifying less common concepts that do not appear in the Profiset collection.

Semantic analysis Next, we focus on the semantic analysis part of our annotation process that utilizes WordNet relationships. Table 1 compares MPEG7-based and DeCAF-based similarity search combined with different levels of semantic analysis.

The base semantic analysis uses only the frequency of the words occurring in the retrieved similar images. In the next step, we have used WordNet synsets instead of the original words. Therefore, synonyms present the similar images keywords are grouped together (see Section 3.2) thus increasing their probability to enter the final annotation. The two final steps then utilize the relationships between to synsets to find the most probable words for the annotation (see Section 3.2). We can observe that for both the MPEG and DeCAF data, adding semantic analysis steps consistently increases the final result quality.

Efficiency The annotation of a single image requires on average about 4-5 seconds. The overall processing time is determined by the costs of four computationally intensive phases: 1) extraction of visual features from the query image, 2) the similarity search, 3) retrieval of words for similar images (these are not stored in the similarity index to minimize its size), and 4) the computation of synset probabilities over the candidate synset graph.

The annotation tool with the parameter setup as described above needed about 1 second for extraction of DeCAF descriptor from a common size image. The similarity search in 20M images took about 1-2 seconds, the retrieval of the words from the 70 similar images needed about half a second and the semantic analysis restricted to 100 most probable synsets required another 0.5-1 second.

Table 1. Experiments on SCIA development dataset: MPEG and DeCAF similarity search over 20M Profiset combined with different levels of semantic analysis.

Semantic analysis	MP-c	MR-c	MF-c	MP-s	MR-s	MF-s	MAP-s
MPEG, basic word frequency analysis	18.2	32.9	19.0	23.8	40.8	27.6	34.7
MPEG, mapping words to synsets, synset frequency analysis	29.1	29.2	22.4	28.3	39.5	30.3	38.4
MPEG, semantic probability computation using hypernymy, hyponymy	29.2	26.7	21.2	30.1	44.2	33.1	42.1
MPEG, full semantic probability comp. (hypernymy, hyponymy, meronymy, holonymy)	29.5	27.5	21.8	30.4	45.2	33.5	42.7
DeCAF, basic word frequency analysis	32.5	46.8	33.6	37.4	49.9	39.6	49.5
DeCAF, mapping words to synsets, synset frequency analysis	48.9	48.8	40.6	42.7	55.6	44.9	55.6
DeCAF, semantic probability computation using hypernymy, hyponymy	48.0	48.5	41.5	44.6	61.0	48.1	60.8
DeCAF, full semantic probability comp. (hypernymy, hyponymy, meronymy, holonymy)	47.7	49.0	41.7	44.7	61.5	48.3	61.1

Table 2. The SCIA competition results table from [19] with a new line for DISA DeCAF results. Only the best result for each group is given. The systems are ranked by overall performance as defined in [19].

System	MAP-samples				MF-samples				MF-concepts				
	all	ani.	food	207	all	ani.	food	207	all	ani.	food	207	unseen
KDEVIR 9	36.8	33.1	67.1	28.9	37.7	29.9	64.9	32.0	54.7	67.1	65.1	31.6	66.1
DISA DeCAF	48.7	51.0	67.1	32.3	39.9	44.4	48.5	26.7	41.1	45.3	42.1	22.4	44.9
MIL 3	36.9	30.9	68.6	23.3	27.5	20.6	53.1	18.0	34.7	34.7	50.4	16.9	36.7
MindLab 1	37.0	43.1	63.0	22.1	25.8	17.0	45.2	18.3	30.7	35.1	35.3	16.7	34.7
MLIA 9	27.8	18.8	53.6	16.7	24.8	12.1	46.0	16.4	33.2	32.7	37.3	16.9	34.8
DISA 4	34.3	46.6	39.6	19.0	29.7	40.6	31.2	16.9	19.1	23.0	22.3	7.3	19.0
RUC 7	27.5	25.2	44.2	15.1	29.3	28.0	28.2	20.7	25.3	20.1	23.1	10.0	18.7
IPL 9	23.4	30.0	48.5	18.9	18.4	20.2	29.8	17.5	15.8	15.8	33.3	12.5	22.0
IMC 1	25.1	35.7	35.6	12.9	16.3	14.3	21.0	10.9	12.5	10.2	15.1	6.1	11.2
INAOE 5	9.6	6.9	15.0	8.5	5.3	0.4	0.5	6.4	10.3	1.0	0.8	17.9	19.0
NII 1	14.7	23.2	22.0	4.6	13.0	18.9	18.7	4.9	2.3	3.0	2.1	0.9	1.8
FINKI 1	6.9	N/A	N/A	N/A	7.2	8.1	12.3	4.1	4.7	6.3	9.0	2.9	4.7

SCIA task results After fine-tuning the various annotation parameters on SCIA development data, the DISA team submitted several competition runs. The results of the ImageCLEF 2014 SCIA Task are summarized in Table 2, more details can be found in [7, 19]. Altogether, the DISA team ranked fifth out of eleven participating teams.

In comparison with other competing groups, our best solution ranked rather high in both sample-based mean F-measure and sample-based MAP. Especially the sample-based MAP achieved by the run DISA_04 was very close to the overall best result (DISA_04 – MAP 34.3, best result kdevir_09 – MAP 36.8). The results for concept-based mean F-measure were less competitive, which did not come as a surprise. In general, the search-based approach works well for frequent terms, whereas concepts for which there are few examples are difficult to recognize.

Furthermore, the MPEG7 similarity is more suitable for scenes and dominant objects rather than details which were sometimes needed by SCIA.

Table 2 also shows that with the DISA DeCAF run, the DISA team would rank as second while outperforming the winner in most sample-based quality measures. However, it is clear that the KDEVIR solution still significantly outperforms ours in terms of concept-based MF. The evaluation results also show that DISA DeCAF achieved better results than some other groups who also employed the neural network approach. This confirms the importance of the semantic analysis step developed by our group.

6 Conclusions and Future Work

In this paper, we have described our approach to general image annotation task. The presented tool applies similarity-based retrieval on annotated image collections to retrieve images similar to a given query, and then utilizes semantic resources to detect dominant topics in the descriptions of similar images. We have presented experimental results with various settings of our tool as well as the tool performance in 2014 Scalable Concept Image Annotation challenge. The results show that the search-based approach to annotation can be successfully used to identify dominant concepts in images. As opposed to training-based annotators that can provide better results for a limited set of pre-trained concepts, the strength of the similarity-search approach lies in the fact that it requires minimum training and easily scales to new concepts.

The experiments and the competition revealed several directions in which the system can be further improved. First, we plan to extend the set of semantic relationships exploited in the annotation process, using e.g. specialized ontologies or Wikipedia. We also intend to develop a more sophisticated method of the final selection of concepts. Furthermore, we would like to improve the response times of our implementation. In particular, the feature extraction can be made faster by introducing GPU processing, while SSD disks can be used for keyword data storage. We will also focus on a more efficient implementation of the semantic analysis phase.

Acknowledgments

This work was supported by the Czech national research project GBP103/12/G084. The hardware infrastructure was provided by the METACentrum under the programme LM 2010005.

References

1. Ballan, L., Uricchio, T., Seidenari, L., Bimbo, A.D.: A cross-media model for automatic image annotation. In: International Conference on Multimedia Retrieval (ICMR). pp. 73–80 (2014)

2. Batko, M., Botorek, J., Budikova, P., Zezula, P.: Content-based annotation and classification framework: a general multi-purpose approach. In: 17th International Database Engineering & Applications Symposium (IDEAS 2013). pp. 58–67 (2013)
3. Batko, M., Falchi, F., Lucchese, C., Novak, D., Perego, R., Rabitti, F., Sedmidubský, J., Zezula, P.: Building a web-scale image similarity search system. Multimedia Tools and Applications 47(3), 599–629 (2010)
4. Botorek, J., Budíková, P., Zezula, P.: Visual concept ontology for image annotations. CoRR abs/1412.6082 (2014), http://arxiv.org/abs/1412.6082
5. Budikova, P., Batko, M., Zezula, P.: Evaluation platform for content-based image retrieval systems. In: International Conference on Theory and Practice of Digital Libraries (TPDL 2011). pp. 130–142 (2011)
6. Budíková, P., Botorek, J., Batko, M., Zezula, P.: DISA at imageclef 2014 revised: Search-based image annotation with decaf features. CoRR abs/1409.4627 (2014), http://arxiv.org/abs/1409.4627
7. Budikova, P., Botorek, J., Batko, M., Zezula, P.: DISA at ImageCLEF 2014: The search-based solution for scalable image annotation. In: CLEF 2014 Evaluation Labs and Workshop, Online Working Notes (2014)
8. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. International Conference on Machine Learning pp. 647–655 (2014)
9. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database. The MIT Press (1998)
10. Hu, J., Lam, K.M.: An efficient two-stage framework for image annotation. Pattern Recognition 46(3), 936–947 (2013)
11. Ke, X., Li, S., Chen, G.: Real web community based automatic image annotation. Computers & Electrical Engineering 39(3), 945–956 (2013)
12. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems (NIPS 2012). pp. 1106–1114 (2012)
13. Lokoc, J., Novák, D., Batko, M., Skopal, T.: Visual image search: Feature signatures or/and global descriptors. In: 5th International Conference on Similarity Search and Applications (SISAP 2012). pp. 177–191 (2012)
14. Makadia, A., Pavlovic, V., Kumar, S.: A new baseline for image annotation. In: ECCV 2008. pp. 316–329 (2008)
15. MPEG-7: Multimedia content description interfaces. Part 3: Visual. ISO/IEC 15938-3:2002 (2002)
16. Novak, D., Batko, M., Zezula, P.: Large-scale similarity data management with distributed metric index. Inf. Processing & Management 48(5), 855–872 (2012)
17. Novak, D., Zezula, P.: Rank aggregation of candidate sets for efficient similarity search. In: Database and Expert Systems Appl. (DEXA 2014). pp. 42–58 (2014)
18. Tousch, A.M., Herbin, S., Audibert, J.Y.: Semantic hierarchies for image annotation: A survey. Pattern Recognition 45(1), 333–345 (2012)
19. Villegas, M., Paredes, R.: Overview of the ImageCLEF 2014 Scalable Concept Image Annotation Task. In: CLEF 2014 Evaluation Labs and Workshop, Online Working Notes (2014)
20. Wang, X.J., Zhang, L., Ma, W.Y.: Duplicate-search-based image annotation using web-scale data. Proceedings of the IEEE 100(9), 2705–2721 (2012)
21. Zhang, D., Islam, M.M., Lu, G.: A review on automatic image annotation techniques. Pattern Recognition 45(1), 346–362 (Jan 2012)
22. Zhang, X., Li, Z., Chao, W.H.: Improving image tags by exploiting web search results. Multimedia Tools and Applications 62(3), 601–631 (2013)