Automatic Annotation of Web Images combined with Learning of Contextual Knowledge

Thomas Scholz, Sadet Alčić, and Stefan Conrad

Institute for Computer Science
Databases and Information Systems
Heinrich Heine University
D-40225 Düsseldorf, Germany
thomas.scholz@uni-duesseldorf.de
{alcic,conrad}@cs.uni-duesseldorf.de

Abstract. This paper introduces an approach for automatic image annotation which is based on contextual knowledge extracted from semantic rich web documents containing images. The knowledge itself is organized in ontologies and extended by learning algorithms for new contextual information. For this purpose, the contextual background of a picture is used for the annotation process and later in the image retrieval process. The paper shows the design of our system and how the different parts work together to enable and improve the annotation procedure. We created learning algorithms for harvesting new contextual information and thus improve context analysis. Finally, we evaluate our methods with a set of sports web pages.

1 Introduction

Today, we are facing a very huge and rapidly increasing number of web images. Controlled access to this large repository is challenging. Content-based Image Retrieval (CBIR) uses low-level feature extraction for retrieval modes. Compared to the human way to handle images and pictured objects this presents a totally different approach, which lacks in high-level semantics. The problem is known as the *Semantic Gap* [4]. There are many low-level features that can be extracted from images, but in general it is difficult to find the corresponding interpretation. Any additional information to the context of the image can improve the retrieval quality. Unfortunately, such information is very difficult to be estimated only based on low-level features.

An approach on a higher level are manually applied annotations provided by human annotators. They are very useful, but expensive in time and human effort. Further, the problem with the Semantic Gap still is not solved but relinquished to human.

In a web environment text contents often provide semantic information on a higher level. According to [1] contextual information can increase the quality of annotations which are made by human. This applies more than ever when ontologies with hierarchical structures build the backbone of the contextual knowledge.

Having considered this advantages of contextual information for manual image annotation, our basic idea is using contextual information for automatic image annotation. In this way we want to get additional information and thus improve the retrieval quality for web images.

The preliminary considerations of our approach can be summarized as follows:

1. The algorithms and data structures for the whole annotation process should be simple.
2. Only a few start information should be needed.
3. The created system should be able to learn new information.

The remainder of this paper is as follows: In Sect. 2 we review some related works and highlight the differences to our work. After that we introduce our image annotating approach in Sect. 3 by giving a short overview the system's components before going into detail. In Sect. 4 we evaluate our methods by checking the resulting metadata and putting the automatic annotated images into a retrieval situation. And finally, the paper is concluded with the evaluation results and a short outline to future works.

2 Related Works

The early approach to searching in image databases was the Content-based Image Retrieval (CBIR) where a selection of low-level features formed all capabilities for query answering [10]. Although CBIR can be enhanced by relevance feedback techniques [11, 12], this way of browsing in image databases is limited by the *Semantic Gap*. While such representation is manageable for computers, handling of low-level features is very difficult for humans, e.g. in the retrieval situation, where a query has to be specified.

More promising approaches are image annotation techniques where image content is described by textual keywords which later can be used as basis for image retrieval. There the annotations are either associated with the whole image or with regions. In the latter case some kind of image segmentation is needed. There are different approaches to generate the resulting annotations which vary from manual annotations given by human annotators and semi automatic approaches to full automated approaches using relevance models between annotated training sets and their low-level descriptions [8].

One of the problems which occurs in image annotation approaches is the word disambiguation. A possible solution is often the use of an dictionary to extends keywords.

Another kind of extension is ontology-based image annotation and brings a new architecture of conceptual image annotations [6, 7]. The new semantic information are gained by the conceptual structures and relationships or leads to new models to describe images [9]. Approaches like [5] are learning from ontologies or discover knowledge in ontologies.

In our application we combine these two approaches: on the one hand an ontology model is used to improve automatic image annotation and on the other hand for storing new information which are the results of a learning procedure. Thus we are able to gain more contextual information and generate a growing and dynamical ontology.

3 System Design

3.1 Overview

The proposed system consists of two main modules: The DUNCAN component, which extracts the image context and annotations from the article, while the PAGANEL component learns new context information from the results of the DUNCAN module. An overview is shown in Fig. 1. In the following section all parts of the system will be discussed in detail.

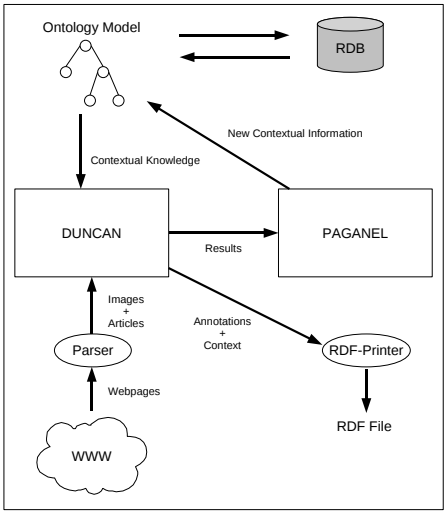

Fig. 1. System Overview

3.2 System Components

WWW and Parser At the beginning of the automatic annotation process, a parser considers pairs of images and corresponding articles of a given web page like proposed in [13].

Then the stopwords of the article are removed. Now three different types of textual information are available: *Metadata* of the image (e.g., alternative text),

image caption (if the webpage has such design, otherwise this text is empty) and the *full text* of the article.

Ontology Model Our ontology model for image annotation is quite simple. It has two types of entities: classes and attributes. Classes represent a context, while attributes are extra information which allow to determine a context and contain extra information for the annotation process. In the tree representation the attributes are leaves while the classes are inner nodes.

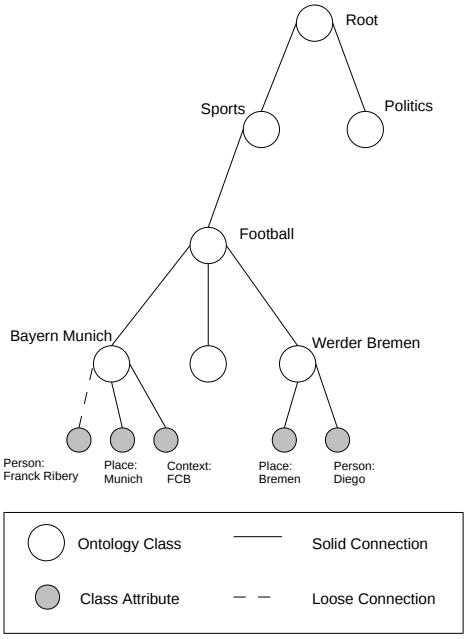

Fig. 2. A sample ontology

The connection between the classes of our ontology are solid, and thus reliable. Loose connections can occur between classes and new learned attributes. This distinction is very important for the learning process, because only the attributes with a higher occurrence will get a solid connection to their corresponding class and are used in the context detecting step.

DUNCAN Module The main tasks of the DUNCAN module are to find a context class, to extract the image annotations of the text (person, place, object and action) and to send the results to the RDF-Printer and the PAGANEL component.

DUNCAN uses the ontology model to determine a concrete context class for an image. Thus all three text types are needed in order to search for class attributes in the text. To every context class a score value is maintained. This value is increased if the text has attributes which belong to the context class. The score value is provided by the following function p

$$p(a, c) = \sum_{i=1}^{n} f(w_i, c) * (n - i) \tag{1}$$

where w_i represents the *i-th* word in the article a (if an expression is larger than one word, the first word defines the position), n is the number of words in the whole article and c the context class of the ontology.

The function f

$$f(w, c) = \begin{cases} 1 & : \quad w \subseteq \Phi(c) \\ 0 & : \quad w \not\subseteq \Phi(c) \end{cases} \tag{2}$$

is an indicator function which shows if the word w is an attribute of the context c or not. $\Phi(c)$ is the set of attributes which solidly belong to the context c. p is further designed in such way that words at the beginning of an article assign more importance for the context than words at the end. The class with the highest score provides the final context of the image. The people names and location information are extracted from articles using the OpenNLP library [14]. The objects and actions were extracted from the surrounding text of the image (alternative text, image caption and heading of the article) using a dictionary to unmask words as actions or objects.

Finally, the results (context class with annotations) are sent to the RDF-Printer to create a RDF-File and to the PAGANEL component for the learning process.

PAGANEL Module The main task of the PAGANEL component is to extend the ontology by learning new contextual information. The results of the DUNCAN module form the basis of this learning process. New attributes are obtained from the last annotations while the ontology class is given by the determined context. For example, if the last image is annotated with the person "Franck Ribery" and the context "Bayern Munich", PAGANEL extends the ontology class "Bayern Munich" with the attribute "Franck Ribery" of type *person*.

Additionally, PAGANEL retrieves the header of the article, which is used to extend the ontology class with attributes of the general type *context*. Tab. 1 shows an excerpt of class attributes of our example ontology.

Fig. 3 shows how the knowledge stored in the ontology is extended during the learning process. Elements, that are new in the ontology, get a loose connection to a corresponding class. This loose connection becomes solid if this relationship is learned more often (appearance count is over a specific threshold). The threshold depends on the length of the learning period.

Generally, the learning process consists of two periods: A class level period for each ontology class and a general level period for the whole ontology. PAGANEL

Table 1. Class Attribute Table

Class Attribute	Ontology Class	Type	Appearance
Franck Ribery	Bayern Munich	Person	2
Corner Kick	Football	Action	3
Jentson Button	Brawn GP	Person	1

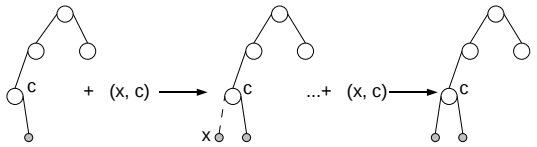

Fig. 3. Learning Process: Attributes are learned in 2 steps.

maintains the number of annotated images with a certain context. Hence, the system can establish learning periods for every context class. After a period PAGANEL deletes all attributes with loose connections. Moreover an attribute can loose the status of a solid connection and become a loose connection when its occurrence in the database is very low. For example all attributes with an occurrence score lower than 3 will get a score decreased by one after a class level learning period.

It must also be noted that this threshold is relative to the total number of annotated images because every context class has its own counter for the class level period. The counter for the whole ontology counts all annotated images. When this counter signals that a learning period is over, the knowledge of the ontology is getting reorganized.

At the beginning, the general distributed contextual information are collected at the upper class. This means: When a learned information appears with a certain percentage in all lower classes, the information (the class attribute) changes its connection to the higher class (see Fig. 4). The percentage can be e.g. 50%. In our football example the "header" can be learned in the context of the different football classes, but it is only an indication of football in general.

After that, contextual information, which appears in more than one ontology class, is moving to the class with the most appearance value. Some pieces of knowledge are learned in the wrong context, but the basic idea is that the growing number of results effects a more and more increasing accuracy in the learning process.

Finally, we add the possibility for attributes to change their context class (see Fig. 5). In our sports example a football player could play for a new football club. In this case our system increases appearance to the new context class and decreases the appearance in the old context class. So the old connection can get weaker and disappear at the end. Thus, we take into account that knowledge is time dependent and dynamic.

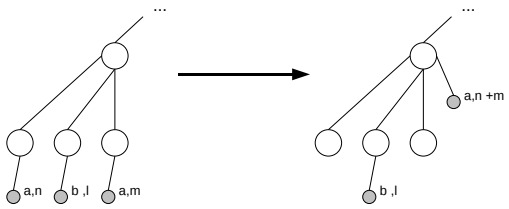

Fig. 4. Learning Process: The same distributed information moves to the upper class.

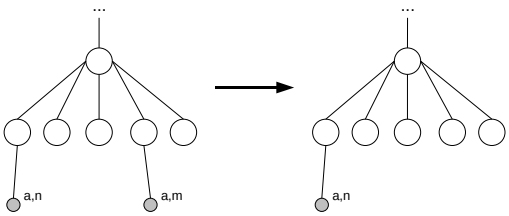

Fig. 5. Learning Process: The class attribute with the greater appearance takes the attributes in the same level

The change of a context also reduces the possibility of keeping wrong learned information for a long time in the ontology. This, and the implementation of loose and solid connections decelerate the learning progress, but increase the quality of new extracted knowledge. Eventually repetitions are even necessary for human, if they want to learn new things.

"Hoffenheim" is "Sinsheim" (this club has a small hometown so that they play in a bigger neighbor city).

4 Evaluation

4.1 Experiment Design

To evaluate our approach we focus on two aspects: First, we inspect the results of the annotation process. Secondly, the annotated images are tested in an image retrieval scenario.

At first we start with a training set of 130 images which are used to extend the initial context descriptions of the ontology classes. Then we crawled 500 images about German football from 10 different web domains (sports portals, news pages, pages from broadcast stations etc.) for our experiments. For the

contextual background an initial ontology to German football was designed with only a few class attributes per ontology class. This means that every context class has equal or less than 5 start information. One of these information contains the location. For example the ontology class "Bayern Munich" has got the class attribute "Munich" with type "place". Other start information are e.g. aliases of the football club. Persons are not inserted in the initial ontology. The PAGANEL component should learn persons by itself.

4.2 Annotation Quality

In the first part of the evaluation we review the concrete annotations and the allocation of the images to a context class. There we check if the image is annotated with the correct person, has the right location, context and so on. In addition we measure the quality of the learning process by proving the classification of a person into the right context class.

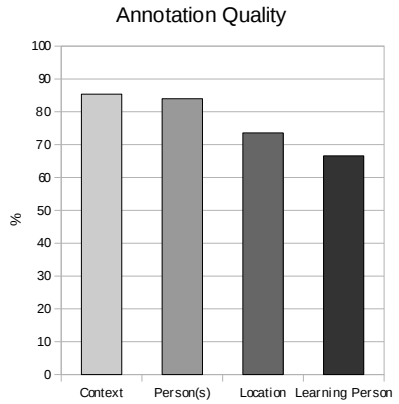

Fig. 6. Annotation Quality

The results of the annotation quality (Fig. 6) are remarkable: The correct context is found with a precision of 85,4%, while 84,0% of the images are annotated with the correct person names. In 73,6% of the cases our system determines the right location of the image.

The quality of the context analysis is very interesting because there are some articles which are quite difficult to analyze in view of the image's contextual environment. Sometimes articles concern a game of two football clubs or speculations which tell about a player who is going to change his football club and the image shows him in the actual club jersey or the player plays for the national team or something like that.

We think that the evaluation process validates the design of formula 1 which controls our context determination. Sometimes articles are quite long and only the beginning of the article is about or belongs to directly to the article's image and later other topics are mentioned in the article.

The efficiency for finding locations can be explained by the following example: Sometimes, e.g. when two teams play against each other, the probability of choosing the right location is 50% (when the right context was determined of course). But the articles reports not only about the games, they tell about press conferences and trainings or are interviews and portraits. For some reason the authors take more home pictures and this explains the performance.

Also the object and action annotations performed very well (they were right in the most cases), but unfortunately only 3,8% and less than 8% of the images get object respectively action annotation from the article. Here is the reason that these annotation are not in the article text which surrounds the image (image caption, head of the article). Certainly one reason is that the author has not to describe that for example the player is running and that the ball is on the pitch. The extraordinary things and actions are mentioned. Another reason is that the images do not have an object or an action. So these results are not so convincing.

At a first look the results of the person learning (66,6%) seem to be worse, but there are regarded two things. On the one hand, the context and the person annotation have to be both correct for a useful class attribute. And on the other hand, a failure in person learning does not immediately lead to bad knowledge in the ontology model. The same person must appear a second time in the same context to get a solid connection. Further, after a learning period solid connections may get loose and loose connections are deleted from the ontology.

Table 2. Class Attribute Table of "Werder Bremen"

ID	Class Attribute	Ontology Class	Type	App.
380	diego	werder bremen	person	5
413	bremen	werder bremen	context	6
525	claudio pizarro	werder bremen	person	2
1538	thomas schaaf	werder bremen	person	7
1595	point	werder bremen	context	3
1899	werder bremen	werder bremen	context	9

To get an impression of the whole learning success, Tab. 2 shows all new learned class attributes which have a solid connection to the ontology class "werder bremen". PAGANEL learned persons like the football players "Diego" and "Claudio Pizarro" and the coach of the team "Thomas Schaaf". It obtained new contextual information from the heads of the articles like "werder bremen", "bremen" and "point", too (the attribute "point" is of course to general for this context).

The evaluation procedure makes clear that the person learning and the insertion of new contextual information helps the context determination.

4.3 Image Retrieval Quality

After looking at the concrete results of the annotation process, we took the collected images with the created RDF files for our retrieval experiments because the results shall fuse that our annotation files serve the purpose in practice. Besides, we want to illustrate that the combination of contextual classification and automatic annotations improve image retrieval.

Fig. 7. The GLENARVAN Retrieval System [1]

We make use of the GLENARVAN retrieval system (Fig. **??**) to combine annotations and contextual information. GLENARVAN works with contextual queries q as a tuple which has the form:

$$q = (s, l) \tag{3}$$

where s denotes a query string consisting of a set of keywords, while l defines the context. Two similarity values are computed, the first one based on a lexical similarity and the second one on a contextual similarity. A dictionary and string comparison (e.g. the edit distance) are used for lexical similarity, while contextual similarity calculates the distance of two ontology classes (see [1]). The multiplication of the two values results in the total similarity. GLENARVAN has a result threshold (not shown in Fig. **??**). The result images must have at least 50% of the best picture's similarity value.

For the evaluation 50 queries are formulated which involve different people in their contexts. We keep the annotation type "persons" tight because we want to relate this results to the results of part one.

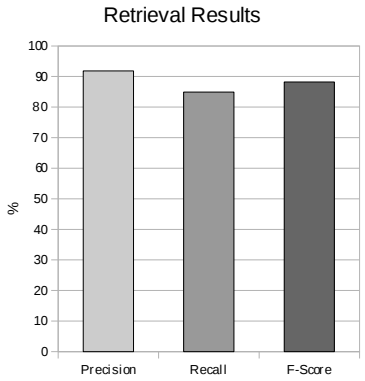

Fig. 8. Retrieval Quality

The results of the retrieval evaluation are summarized in Fig. 8. With 91,84% the precision performs very well and the recall value of 84,91% is also considerable. The F-Score amounts 88,24%.

The evaluation makes clear that the created RDF-files are very useful in a retrieval process especially when the additional contextual information are considered.

5 Conclusion and Future Works

In this paper we have summarized the problems of automatic image annotation and the handling contextual background knowledge. We have presented a new approach which combines automatic annotation and annotations based on ontologies. Besides, we added learning algorithms to obtain new contextual information. Finally, the evaluation illustrates that our system produces advantageous RDF-files which retrieve reliably image data.

The paper shows that the combination of ontology based annotation and learning of new contextual information is a favorable solution for automatic image annotation which can stand in a real retrieval situation.

Prospectively we will work on a way to combine this approach with a large set of external knowledge. We plan to achieve two things:

1. We want to double check the context.
2. We want to develop more possibilities of person, objects and action recognition.

The advantage of a large database would be the co-occurrence of specific expressions which appear again in the same context. By a comparison the determination of the context could be verified. Here links between the separate pieces of knowledge may be a second method. In this way the automatic and conceptual image annotation and contextual learning can be more improved.

References

1. J. Vompras, T. Scholz, and S. Conrad. Extracting contextual information from multiuser systems for improving annotation-based retrieval of image data. In *MIR '08: Proceeding of the 1st ACM international conference on Multimedia information retrieval*, pages 149–155, New York, NY, USA, 2008. ACM.
2. D. Brickley and R. V. Guha. Resource Description Framework (RDF) Schema Specification. World Wide Web Consortium. 2000.
3. T. Berners-Lee, J. Hendler, and O. Lassila. The Semantic Web. In *Scientific American*, page 279, 2001.
4. A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(12):1349–1380, 2000.
5. A. Mallik, P. Pasumarthi, and S. Chaudhury. Multimedia ontology learning for automatic annotation and video browsing. In *MIR '08: Proceeding of the 1st ACM international conference on Multimedia information retrieval*, pages 387–394, New York, NY, USA, 2008. ACM.
6. A. T. G. Schreiber, B. Dubbeldam, J. Wielemaker, and B. Wielinga. Ontology-based photo annotation. *IEEE Intelligent Systems*, 16(3):66–74, 2001.
7. E. Hyvönen and K. Viljanen. Ontogator: combining view- and ontology-based search with semantic browsing. In *In Proceedings of XML*, 2003.
8. J. Liu, M. Li, W.-Y. Ma, Q. Liu, and H. Lu. An adaptive graph model for automatic image annotation. In *MIR '06: Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, pages 61–70, New York, NY, USA, 2006. ACM.
9. T. Osman, D. Thakker, G. Schaefer, and P. Lakin. An integrative semantic framework for image annotation and retrieval. In *WI '07: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, pages 366–373, Washington, DC, USA, 2007. IEEE Computer Society.
10. Y. Ishikawa, R. Subramanya, and C. Faloutsos. MindReader: Querying databases through multiple examples. In *Proceedings of 24th International Conference on Very Large Data Bases, VLDB'98*, pages 218–227, 1998.
11. T. S. Huang, Y. Rui, M. Ortega, and S. Mehrotra. Relevance Feedback: A Power Tool for Interactive Content-Based Image Retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 25–36, 1998.
12. Y. Rui, T. S. Huang, and S. Mehrotra. Content-Based Image Retrieval with Relevance Feedback in MARS. In *Proceedings of the 1997 International Conference on Image Processing (ICIP '97)*, pages 815–818, 1997.
13. S. Alcic and S. Conrad. 2-DOM: A 2-dimensional Object Model towards Web Image Annotation In *3rd International Workshop on Semantic Media Adaptation and Personalization (SMAP 08)* December 15-16, 2008. Prague, Czech Republic.
14. OpenNLP Project Website. 2009 `http://opennlp.sourceforge.net`