Visual Concept Ontology for Image Annotations

Jan Botorek
Faculty of Informatics
Masaryk University
Brno, Czech Republic
xbotorek@fi.muni.cz

Petra Budikova
Faculty of Informatics
Masaryk University
Brno, Czech Republic
budikova@fi.muni.cz

Pavel Zezula
Faculty of Informatics
Masaryk University
Brno, Czech Republic
zezula@fi.muni.cz

ABSTRACT

In spite of the development of content-based data management, text-based searching remains the primary means of multimedia retrieval in many areas. Automatic creation of text metadata is thus a crucial tool for increasing the findability of multimedia objects. Search-based annotation tools try to provide content-descriptive keywords by exploiting web data, which are easily available but unstructured and noisy. Such data need to be analyzed with the help of semantic resources that provide knowledge about objects and relationships in a given domain. In this paper, we focus on the task of general-purpose image annotation and present the VCO, a new ontology of visual concepts developed as a part of image annotation framework. The ontology is linked with the WordNet lexical database, so the annotation tools can easily integrate information from both these resources.

Categories and Subject Descriptors

I.2.4 [**Artificial Intelligence**]: Knowledge Representation Formalisms and Methods

Keywords

visual concept ontology, WordNet, image annotation

1. INTRODUCTION

As more and more information becomes expressed by images or videos, effective management of multimedia data remains one of the priorities in information processing. In spite of the development of content-based retrieval, text-based searching continues to be the most frequent data access method. Consequently, multimedia objects need to be accompanied by content-describing text metadata to be findable. Unfortunately, the descriptions are often of uncertain quality or not available at all, since their manual creation is a tedious and time-consuming task. In this situation, automatic annotation of multimedia content and text metadata refinement methods are highly desirable.

In this paper, we focus on automatic annotation of images. We assume that the annotation task is defined by a query image I and a vocabulary V of candidate concepts, and the annotation function f_A assigns to each concept $c \in V$ its probability of being relevant for I. Traditional *model-based approaches* use manually-labeled training data to create classifiers for all concepts in V, which are then used to determine the relevance of any concept for image I. Such techniques are known to perform well in narrow-domain tasks (e.g. medical image classification), but are limited by the availability of high-quality training data, which may become a significant bottleneck when the vocabulary V is large. Therefore, an alternative *search-based annotation* has been proposed recently for general-purpose annotations. This method tries to exploit noisy but voluminous data that is easily available e.g. on the web. Specifically, content-based retrieval is applied to obtain images that are visually similar to I, and their metadata is used for reasoning about the probability of concepts from V [1, 10]. The whole process is schematically depicted in Figure 1.

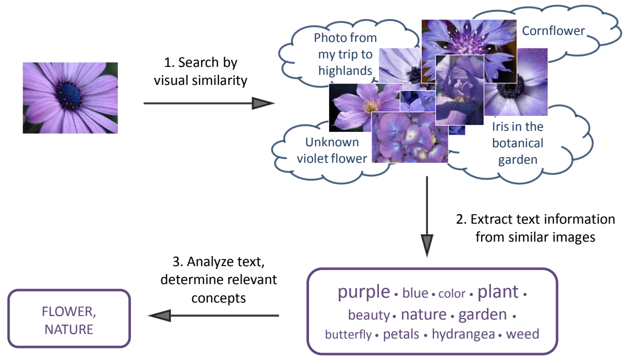

Figure 1: Search-based image annotation scheme.

The search-based approach allows us to take advantage of the digital data explosion, but also poses multiple challenges that need to be solved to achieve reasonable annotation quality. One of the crucial problems is determining the relevance of candidate concepts using the metadata of similar images (i.e. step 3 in Figure 1). It may be helpful to consider a parallel – let us imagine a person trying to guess what is depicted in a hidden image, using only visually similar images selected by a computer. The similar images may be noisy in terms of semantic content (due to the well-known *semantic gap* problem, an orange ball is likely to be evaluated as similar to both an orange fruit and the sun), the guessing person thus needs to look for connected topics and try to infer the most probable content of the hidden image. When we replace the person by some automated system, the task becomes even more difficult. The system does not understand an image per se and needs to work with its metadata, which again are likely to be noisy and further complicated by natural language properties such as ambiguity, use of synonyms, abstraction, etc. But most importantly, the computer lacks human experience with real world that allows us to decide which topics are related and infer further information.

To be able to make sense of the similar images' metadata, the annotation system thus needs two types of knowledge sources: 1) computer-processable information about linguistic relations between words, and 2) ontologies that describe the semantics of real-world concepts and links between them. The first resource can be considered domain-independent, but the latter should be tailored for a particular domain to reflect the importance of individual concepts and their relations in the given context. For general-purpose image annotation, the domain of interest consists of so-called *visual concepts*, i.e. objects or abstract notions that are typically depicted in photos. Although some existing ontologies cover this domain, none of them entirely suits the needs of the annotation task [9]. Specifically, these ontologies are either too complex or unsuitably structured, so that some important semantic connections between visual concepts are missing.

Therefore, we propose a new *Visual Concept Ontology (VCO)*, which organizes common visual concepts that appear in image annotations. The VCO is based on the WordNet lexical database, a linguistic tool frequently used in information mining. By encapsulating relevant parts of the WordNet hierarchy into a more concise structure, the VCO becomes a strong tool for analysis of image descriptions.

2. EXISTING SEMANTIC RESOURCES

Before introducing the VCO, let us briefly survey existing semantic resources and discuss their limitations with respect to the image annotation task. We focus on ontologies and tools that can be used for processing of general-purpose images (such as those found in personal photo collections), leaving out domain-specific resources for medicine, arts, etc.

WordNet. The WordNet [3] 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, e.g. hypernymy/hyponymy (super-type and sub-type relation) or meronymy (part-whole relation). The WordNet structure is divided into four independent categories according to the parts of speech encoded – nouns, verbs, adjectives, and adverbs. The noun hierarchy, which is best developed, forms a tree structure corresponding to the hypernymy and hyponymy relations among synsets. In the current version of WordNet, there are more than 82,000 synsets for 117,798 unique nouns stored in the noun hypernymy tree structure.

In context of search-based image annotation, the WordNet is very useful for identifying synonymous words and discovering relationships between concepts [1, 10]. However, the WordNet relationships only cover linguistic dependencies, which are not satisfactory for image metadata analysis – for instance, there is no relationship linking "roof" and "house" although these words are clearly semantically related. Another problem of the WordNet structure is uneven depth of individual branches of the noun hierarchy, which makes it difficult to quantify the relatedness of any two synsets.

ImageNet. A well-known resource for image data processing is the ImageNet [2], a database of images organized according to the WordNet hierarchy. The ImageNet project aims at illustrating each WordNet synset by several hundred images, using a crowdsourcing platform to supervise the selection of images. Currently, the database contains about 14 million images for nearly 22 000 noun synsets.

The ImageNet has been successfully engaged in many image processing tasks, but its applicability for general-purpose annotations is limited. The database can be used as the reference dataset, from which the similar images are retrieved; when sufficiently similar images are found for the input image I, the annotation will be very precise. However, more complex scenes with multiple topics are not likely to be found among ImageNet objects and therefore will not be annotated correctly. Less organized but significantly larger collections of web images are therefore more promising for the general-purpose annotations.

Ontologies. Ontologies are a standard tool for describing knowledge about concepts from a given domain. In contrast to narrow-domain ontologies, which are created by a few domain experts, broad-domain knowledge bases are often developed by cooperation of large user groups (DBpedia[1], Freebase[2], etc.). The crowd-sourcing paradigm can facilitate wide coverage, but it is problematic to keep the overall philosophy of the resource uniform, which complicates any utilization of such ontology in automatic processing. Therefore, some recent ontologies have been developed on top of existing, well-structured resources, especially the WordNet. The YAGO ontology [8] has been automatically created by aligning the WordNet with Wikipedia facts, whereas the existing Kyoto ontology [4] has been semi-automatically mapped to the WordNet. However, all these resources were designed primarily for text processing, which has different requirements on the selection and categorization of concepts than analysis of image annotations.

A few attempts have also been made to establish an ontology for visual information. The LSCOM ontology [5] specializes on concepts appearing in video news, and a simple "Photo Tagging Ontology" covering 100 concepts was issued with the ImageCLEF annotation task [7]. However, the first one specializes only on the most frequent video news topics, whereas the second is too shallow to be applicable outside the ImageCLEF competitions.

3. VISUAL CONCEPT ONTOLOGY

As we discussed in the introduction, knowledge bases that provide information about visual concepts and their relationships are much needed for search-based annotation. During the development of a general-purpose annotation system [1], we discovered that none of the existing resources fulfills the specific needs of such application. Therefore, a new Visual Concept Ontology was created for the specific needs of image content description.

The Visual Concept Ontology organizes frequent visual concepts into a hierarchical structure and maintains semantic links between them. Individual concepts are not defined formally in the VCO, but are specified by links to related WordNet synsets. The VCO taxonomy simplifies and flattens the WordNet hierarchy, removing concepts not relevant to the visual domain and adding semantical connections between interrelated WordNet subtrees. At the moment, VCO only covers nouns from the WordNet structure, as these are

[1] http://dbpedia.org
[2] http://www.freebase.com/

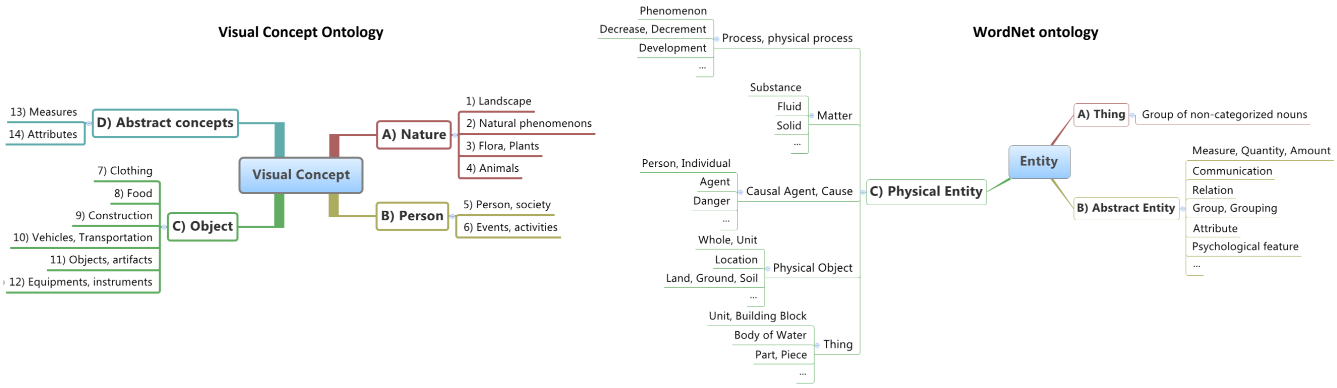

Figure 2: Comparison of VCO and WordNet top-level concepts.

most frequently used in image descriptions and have the strongest support in the WordNet. The ontology is encoded by the Web Ontology Language OWL, a state-of-the-art formalism in knowledge representation. Individual VCO concepts are represented by OWL classes, whereas the related WordNet synsets are modeled as OWL individuals.

3.1 Selection of Concepts

The objective of the VCO is to identify basic semantic categories of objects occurring in a picture and organize them in a hierarchical structure. We are not interested in very specific categories such as "poodle dog", which are well organized in the WordNet, but rather in determining reasonable top-level categories of visual objects. To choose suitable concepts for populating the VCO taxonomy, we utilized a combination of top-down and bottom-up analysis of the WordNet hierarchy. The ontology construction process included the following four phases:

1. Significant synset extraction In the first phase, we extracted significant synsets from the WordNet noun hypernymy tree. For each synset, the number of all hyponyms was computed. If the number was higher than a specified constant (experimentally set to 300), the synset was declared to be significant and was included into a group of VCO concept candidates. In this phase, more than 400 candidate synsets were identified.

2. General concept removal In the next step, we manually examined the candidates and removed very general synsets such as "entity" or "thing". Similar to stopwords in natural language, these general concepts would be of little use in image understanding tasks. Therefore, we checked individual branches of the border candidates tree and selected synsets that were specific enough to be included in the ontology.

This cleaning needed to be done manually, since individual branches of the WordNet noun tree significantly differ in depth. To reach the "white rhinoceros" synset from the "entity" root synset, it is necessary to traverse 15 more general synsets, but only 7 synsets separate the "river Thames" from the root. Therefore, it was not possible to simply remove the first k levels of the WordNet hierarchy.

In this phase, we also removed most descendants of the "abstraction" synset that are highly unlikely to be visually represented in images (e.g. "communication" synset with descendants like "Soprano" or "poetry", or "grouping" synset covering concepts of "overpopulation" or "baby-boom"). However, several abstract concepts that can be meaningfully applied to images were retained, especially those regarding time-related concepts ("Christmas","evening") and human emotions ("joy", "sadness").

3. Organization of classes Having selected significant synsets with visually-relevant content, we were able to establish the top-level VCO classes and individual sub-trees. The remaining candidate synsets were explored and semantically similar ones were merged a joint class. In this step, we thus established links between semantically interrelated synsets placed in different parts of the WordNet hierarchy – e.g. the "roof" and "house" synsets that are far apart in the WordNet.

4. Confrontation with existing ontologies In the last phase, we examined other broad-domain ontologies to check whether some important concepts were not missed. We particularly focused on inspection of basic classes and overall architecture of the taxonomies. Specifically, YAGO [8], Sumo [6], LSCOM [5] and the ImageCLEF Photo Tagging Ontology [7] were utilized to check and fine-tune the VCO taxonomy.

As a result of these four steps, we obtained 14 top-level classes that were further divided into 90 more specific subclasses. On top of these, a final high-level generalization was performed, producing 4 super-classes: *nature, person, object* and *abstract concepts*. The resulting hierarchy is outlined in Figure 2; in contrast to the WordNet, the VCO taxonomy is semantic rather than lexicographic and the top-level categories are suitable for human-readable labeling of image content. Each VCO class is linked to one or several WordNet synsets, which define its semantics and allow to access a detailed hierarchy of synsets relevant for the given topic.

3.2 Relationships

Relationships between ontology classes are a vital tool for describing semantics. In the VCO, we define two basic types of relationships – *class-to-class* and *class-to-individual*. The class-to-class relationships describe semantic links between individual VCO concepts, especially the standard sub/super-class relation between classes. Class-to-individual relationships connect VCO classes with *individuals* – OWL objects that represent related WordNet synsets. Inspired by the Kyoto ontology which also models synset-to-concept links, we defined the following two types of class-to-individual relationship: *equivalenceOf* link connects an individual to a class that is semantically equal to the synset referenced by the individual, *superClassOf* links classes with individuals that are not semantically equal to it but semantically belong to the given class.

4. POSSIBLE APPLICATIONS

4.1 Hierarchical Image Annotation

As we have discussed in the introduction, search-based annotation techniques are primarily intended for tasks with wide annotation vocabularies. Specifically, we aim at developing a general annotation tool that would work with the full English vocabulary and provide keyword annotation with user-selected level of detail. The overall architecture of this tool is described in [1]. The VCO ontology will be used for two purposes within the tool. First, it will be exploited during analysis of the descriptions of similar images, as described earlier. Complementing the WordNet relationships with VCO categorization will provide a richer set of links between concepts, and the VCO structure will also be used to prevent too extensive searching for connections. Second, after choosing the most probable VCO and WordNet subtrees, the final annotation will be presented to the user using the VCO taxonomy, so that the user will be able to easily access descriptive keywords on different levels of detail.

4.2 Ground Truth Construction

As a second possible application, we would like to discuss the problem of ground truth construction for the evaluation of annotation techniques. Establishing evaluation data is known to be a difficult task, which is typically solved by manual effort of a group of experts or using the crowdsourcing platforms. However, the former approach can only generate small test collections, whereas the latter is likely to provide uncomplete or erroneous ground truth. Using the VCO, we believe that the ground truth construction could be facilitated in a semi-automatic way, removing much of the manual effort and guaranteeing a uniform vocabulary. Specifically, we suggest to take some dataset of richly annotated images, e.g. the Profiset collection[3] of web-stock images, analyze the annotations using WordNet relationships, and replace them by the appropriate top-level categories and subcategories from VCO. If the original annotations are rich and precise enough, which is typically true for stock data, there is a high chance that the automatically selected categories will provide precise annotation of the image content. Any imprecisions can be removed by a quick manual checking of the classification results. In contrast to the original annotations, which are not applicable as a ground truth due to the unsupervised vocabulary, the new metadata will be selected from a fixed vocabulary with a given level of detail. Figure 3 shows the transformed metadata of three random pictures from the Profiset collection.

5. CONCLUSIONS

This paper introduces the VCO ontology, which was created as a part of search-based image annotation framework. The VCO provides a mapping between fundamental visual concepts and the WordNet hierarchy, thus creating a useful connection between the image domain semantics and natural

[3]http://disa.fi.muni.cz/results/software/profiset-testbed/

Original keywords: colors, colours, petals, tulips, blooms, bulbs, flowering, flowers, plants, spring, white
VCO concepts: Flora (Plants [Herbs], Plant_parts)

Original keywords: animal, brochure, butterfly, wildlife, central_american, colors, colours, costa_rican, insect, invertebrate, lepidoptera, american, iole, marpesia, …
VCO concepts: Animals (Invertebrates [Insect])

Original keywords: adult, area, back, competition, day, defense, effort, field, football, goal, grass, group, horizontal, image, kicking, length, low, match, men, outdoors, …
VCO concepts: Events&Activities (Activity [Human_activity], Social_event [Sport_contest]) ▪ Clothing ▪ Person&Society (Group_of_people, Single_person [Body_parts])

Figure 3: Author keywords vs. covering concepts.

language. The ontology can be used to analyze image descriptions and assist in categorization of image content, but also to automatically extract category-based ground truth from web images with free-text annotations. The complete ontology is available at `http://disa.fi.muni.cz/vco`.

Acknowledgments

This work was supported by the Czech national research project GBP103/12/G084.

6. REFERENCES

[1] M. Batko, J. Botorek, P. Budíková, and P. Zezula. Content-Based Annotation and Classification Framework: A General Multi-Purpose Approach. In *IDEAS 2013*, pages 58–67. ACM, 2013.

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and F.-F. Li. ImageNet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. IEEE, 2009.

[3] C. Fellbaum, editor. *WordNet: An Electronic Lexical Database*. The MIT Press, 1998.

[4] E. Laparra, G. Rigau, and P. Vossen. Mapping WordNet to the Kyoto ontology. In *LREC-2012*, pages 2584–2589, 2012.

[5] M. R. Naphade, J. R. Smith, J. Tesic, S.-F. Chang, W. H. Hsu, L. S. Kennedy, A. G. Hauptmann, and J. Curtis. Large-scale concept ontology for multimedia. *IEEE MultiMedia*, 13(3):86–91, 2006.

[6] I. Niles and A. Pease. Towards a standard upper ontology. In *FOIS*, pages 2–9, 2001.

[7] S. Nowak, K. Nagel, and J. Liebetrau. The CLEF 2011 Photo Annotation and Concept-based Retrieval Tasks. In *CLEF 2011 working notes*, 2011.

[8] F. M. Suchanek, G. Kasneci, and G. Weikum. YAGO: A Large Ontology from Wikipedia and WordNet. *J. Web Sem.*, 6(3):203–217, 2008.

[9] A.-M. Tousch, S. Herbin, and J.-Y. Audibert. Semantic hierarchies for image annotation: A survey. *Pattern Recognition*, 45(1):333–345, 2012.

[10] D. Zhang, M. M. Islam, and G. Lu. A review on automatic image annotation techniques. *Pattern Recognition*, 45(1):346–362, 2012.