Ontology of Visual Objects

Svetla Koeva
Institute for Bulgarian Language, Bulgarian Academy of Sciences
svetla@ddcl.bas.bg

Abstract

The focus of the paper is the **Ontology of Visual Objects** based on WordNet noun hierarchies. In particular, we present a methodology for bidirectional ontology engineering, which integrates the pre-existing knowledge resources and the selection of visual objects within the images representing particular thematic domains. The Ontology of Visual Objects organizes concepts labeled by corresponding classes (dominant classes, classes that are attributes to dominant classes, and classes that serve only as parents to dominant classes), relations between concepts and axioms defining the properties of the relations. The Ontology contains 851 classes (706 dominant and attribute classes), 15 relations and a number of axioms built upon them. The definition of relations between dominant and attribute classes and formulations of axioms based on the properties of the relations offers a reliable means for automatic object or image classification and description.

1 Introduction

The recent trends in Computer vision are directed towards the robust combination of deep learning techniques and image processing methods to solve problems, such as image and video understanding, robot vision and processing of multimodal and multilingual content. Despite this, much effort is still directed to specific domain knowledge or even to specific object instance recognition, and the significant progress in the field as a whole does not mean that particular tasks have been solved satisfactorily.

The concept of Cognitive vision (Vernon, 2021) was introduced quite a long time ago (Auer et al., 2005): "A cognitive vision system can achieve the four levels of generic computer vision functionality of detection, localization, recognition, and understanding. It can engage in purposive goal-directed behaviour, adapting to unforeseen changes of the visual environment, and it can anticipate the occurrence of objects or events" (Vernon, 2006).

Such understanding of Cognitive vision systems involves the application of ontology-based representations in modern Computer vision systems in order to add real world relations between static objects and video feed (Xie et al., 2020; Chaisiriprasert et al., 2021). Ontology-based applications might be powerful tools for diverse Computer vision tasks: application of semantics according to the function of an object (Agostini et al., 2015), ontology-based object recognition in robotics (Riazuelo et al., 2015), and so on.

The focus of the paper is the **Ontology of Visual Objects**.[1] In particular, we present a methodology for bidirectional ontology engineering, which integrates the pre-existing knowledge resource (WordNet) and the selection of visual objects within the images representing particular thematic domains.

We show how the presented Ontology benefits from WordNet (Miller et al., 1990; Fellbaum, 1999): providing ontological representation of visual objects based on WordNet noun hierarchies, and building interconnectivity of classes by means of the WordNet. On the other hand, we present how the Ontology of Visual Objects builds on the WordNet by adding new concepts corresponding to concrete objects, and formulating new relations that express the objects' function, purpose, location, etc.

We begin with a brief overview of the current state in the art in Section 2. In Section 3 we present the principles of Ontology-based image annotation. Section 4 is dedicated to the main components of an ontology and the description of the Ontology of Visual Objects. Finally, evaluation (section 5), conclusions and future directions of our work (section 6) are presented.

[1] https://doi.org/10.57771/a0w5-8480

2 Related Work

In this section, we briefly present some of the most prominent knowledge representation resources, the image datasets, which involve (in different ways) ontologies in the process of their building, and the few existing examples of ontologies specially dedicated to image descriptions.

2.1 Ontology-based Semantic Resources

The taxonomic organization of nouns in **WordNet** allows for using more abstract and fine-grained categories when describing objects. WordNet[2] is a semantic network, whose nodes host synonyms denoting different concepts, and whose arcs, connecting the nodes, encode different types of relations (semantic: genus-kind, part-whole, etc.; extralinguistic: membership in a thematic domain; interlanguage: translation equivalents).

The idea for organizing the lexicon of a given language into a (lexico-)semantic network was first executed in the Princeton WordNet (Miller et al., 1990). Some of the fundamental ideas on which the WordNet is based encompass: a) the use of a semantic network which embraces taxonomies, meronomies and non-hierarchical relations with clearly defined properties, which allow for quick and easy automatic processing; b) a different organization of the lexicon in comparison with the traditional dictionaries where words are ordered alphabetically and the links among semantically related words (such as between sister hyponyms, between a whole and its parts, etc.) are not explicitly presented (Miller, 1986).

WordNet is connected to a generic ontology based on **DOLCE**.[3] A set of heuristics for mapping all WordNet nouns, verbs and adjectives to the ontology were developed, which also allows to represent predicates in a uniform and interoperable way, regardless of the way they are expressed in the text and in which language (Laparra et al., 2012). Together with the ontology, the WordNet mappings provide powerful basis for semantic processing of text in different domains.

Some ontologies have been developed on top of the existing resources. The **YAGO** ontology[4] is a large knowledge base with general knowledge about people, cities, countries, movies, and organizations (Suchanek et al., 2007). YAGO contains both entities (such as *movies*, *people*, *cities*, *countries*, etc.) and relations between these entities (who played in which movie, which city is located in which country, etc.). The entities are arranged in classes: *Elvis Presley* belongs to the class of *people*, *Paris* belongs to the class of *cities*, and so on, which in their turn are arranged in a taxonomy: the class of *cities* is a subclass of the class of *populated places*, etc. YAGO combines Wikidata – the largest general-purpose knowledge base on the Semantic Web and schema.org (plus BioSchemas) – a standard ontology of classes and relations.

BabelNet[5] (as WordNet) combines features of multilingual encyclopaedic dictionary (with its wide lexicographic and encyclopaedic coverage of terms), and of semantic network or ontology, which links concepts and named entities in a very large network of semantic relations (about 20 million entries as of 2021) (Navigli et al., 2021). BabelNet brings together heterogeneous resources, such as WordNet, Wikipedia, OmegaWiki, Wikidata, Wiktionary, GeoNames, Open Multilingual WordNet and many others, and aims at providing as complete picture as possible of lexical and semantic knowledge available in many languages. BabelNet represents each meaning based on the WordNet notion of a synset. Analogously to WordNet, BabelNet can be viewed as a graph where synsets are nodes and edges are semantic relations between them.

2.2 Ontology-supported Image Datasets

There are several datasets that have been widely used as a benchmark for object detection, semantic segmentation and classification tasks. Only a few of them use ontologies or ontology-like resources for object classification.

Thousands of images, hundreds of thousands of polygon annotations and sequence frames with at least one tagged object are all included in the **LabelMe** dataset[6] (Russell et al., 2008). This collection is being created by users who can upload images, add categories and annotate images with these categories. However, depending on how each annotator chooses to use the annotation protocol, this choice can lead to some degree of inconsistency. By using the WordNet noun synsets, categories are expanded, inconsistent editing is avoided and user-provided descriptions are unified.

[2] http://wordnet.princeton.edu/

[3] http://www.loa.istc.cnr.it/dolce/overview.html

[4] https://yago-knowledge.org/

[5] https://babelnet.org/

[6] http://labelme.csail.mit.edu

One of the collections that sets standards in the increase of datasets sizes is **ImageNet**.[7] A dataset with roughly 50 million full-resolution images that have been accurately labelled has been set as a target (Deng et al., 2009). The WordNet noun hierarchies are used for image collection and labelling. ImageNet comprises 14,197,122 annotated images that are arranged according to the semantic hierarchy of WordNet and employs 21,841 synsets for focused image search (as of August 2014) (Russakovsky et al., 2015).

More than 328,000 images with carefully annotated object instances (2.5 million) can be found in the **COCO (Microsoft Common Objects in Context)** dataset[8] (Lin et al., 2014). Since 2014, the dataset has undergone a number of updates and covers object detection, segmentation, keypoint detection and captioning. The different parts of the dataset are annotated with bounding boxes (for object detection) and per instance segmentation masks with 80 object categories; natural language descriptions of the images; keypoints (17 possible key points, such as *left eye, nose*); per pixel segmentation masks with 91 stuff categories, such as *grass, wall*; full scene segmentation, with 80 thing categories (such as *person, bicycle, elephant*); dense pose – each labelled person is annotated with a mapping between image pixels and a template 3D model.

WordNet is typically utilized in current practice to generate text queries for building search-based image collections. Some of the datasets were developed using shallow ontologies (Griffin et al., 2007), and overall, the potential power of the ontological structure is not completely exploited.

2.3 Existing Ontologies of Visual Objects

The **LSCOM** ontology consists of 1,000 concepts and approximately 450 of them were used for the manual annotation of 80 hours of news video (Naphade et al., 2006). The taxonomy design organized concepts into six categories on a top level: *objects, activities/events, scenes/locations, people, graphics*, and *program* categories. These categories were further refined, such as by subdividing objects into *buildings, ground vehicles, flying objects*, etc.

Photo Tagging Ontology covering 100 concepts was issued with the ImageCLEF annotation task (Xioufis et al., 2011). The ontology restricts simultaneous assignment of some concepts (disjoint classes) and defines that one concept postulates the presence of other concepts. The purpose of the ontology is to allow integration of semantic knowledge in the algorithms for image annotations.

A **Visual Concept Ontology** organizes visual concepts (objects or abstract notions that are typically depicted in photos) (Botorek et al., 2014). For the construction of Visual Concept Ontology over 400 "significant" noun synsets (that have at least 300 hyponyms) were extracted from WordNet; then synsets with a very "general" meaning, such as *entity* or *thing*, were removed. This results in 14 top-level ontology classes, which are divided further into 90 more specific classes. On top of these, a final high-level generalization was performed, producing 4 super-classes: *nature, person, object* and *abstract concepts*. Semantically similar synsets are merged into a common class and additional links are established between semantically related synsets, such as *roof* and *house*. In other words, the ontology simplifies and flattens the WordNet hierarchy, removing concepts not relevant to the visual domain and adding semantic connections between interrelated WordNet subtrees. Relations are of two basic types – class-to-class and class-to-individual.

It has been demonstrated that combining ontology knowledge with image recognition technologies can increase recognition precision, enhance high-level semantic recognition capability, decrease the need for a large number of training samples and improve the scalability of the image recognition systems (Ding et al., 2019).

In conclusion, it can be stated that the ontological representation of knowledge is not fully exploited in Computer vision: neither in the process of creating annotated datasets, nor in the implementation of algorithms and models for the recognition and classification of objects and images.

3 Ontology-based Image Annotation

The **Ontology of Visual Objects** was developed to serve for the annotation of the image objects in the Multilingual Image Corpus,[9] which provides pixel-level annotations, thus offering data to train models specialised in object detection, segmentation and classification in these domains (Koeva, 2021; Koeva et al., 2022).

Different ways of incorporating semantics to describe an image are discussed (Tousch et al., 2012).

[7] https://www.image-net.org

[8] https://cocodataset.org

[9] doi.org/10.57771/p2n7-f015

One possible level incorporates the relations between concrete and abstract objects, for example, a *crying person* vs. the notion of *pain*, which might be a subjective conclusion based on the knowledge of the semantic context. The other level describes generic vs. specific objects (individual instances), i.e., a *bridge* vs. *Golden gate bridge*. In our approach, we concentrate on visual (concrete) objects; however, specific instances of an object can be further related with it, and further inferences to abstract notions might be drawn as well.

We defined the following criteria for the development of the **Ontology of Visual Objects**:

• The specificity or generality of the concept (we include only specific concepts at a certain level of granularity: more concrete comparing to classes that are usually used in image datasets, for example *taxi* and *sedan* instead of a *car*, but not too concrete, in order for the annotators to be able to choose among the classes without employing specific knowledge for different thematic domains, (for example *sedan*, but not *Bentley* or *Dacia*).

• High frequency of occurrence of words denoting visual objects in everyday life and of respective objects depicted in images. The everyday use is based on the inclusion of the words in the so-called common vocabulary, which is evidenced by the Age of acquisition list of words (Brysbaert and Biemiller, 2017). The assumption is that words that are mastered at an early age belong to the basic vocabulary. The frequency of encounters of objects in the images is observed empirically, based on the collected over 750,000 images, of which about 21,000 were selected for annotation in the Multilingual Image Corpus. For example, although the object *baby rattle* is expected to meet frequently along with dominant objects, such as a *baby* and a *stroller*, empirical observations in images have shown a low frequency of encounters, and this visual object is not included in the Ontology.

• Coverage in ontologies (concepts already encoded within the WordNet and through WordNet in other ontologies).

• Covering gaps in existing ontologies, for example, some objects we observed in the collected images (such as *handball player*, *pole vaulter*, etc. have not been included in the Princeton WordNet so far).

The proposed **Ontology of Visual Objects** includes concepts that are characteristic for the thematic domains of **Sport**, **Transport**, **Arts**, and **Security**. The Multilingual Image Corpus contains 130 smaller datasets pertaining to different subdomains, each of which can be classified to one of the four main ones, for example, **Chess** and **Pole vaulting** are subdomains of **Sport**, while **Sedan** and **Double-decker** – to **Transport**, and so on. The choice of thematic domains and subdomains is motivated by two main factors:

(1) The images should contain objects that could be automatically recognized and labelled with upper-level classes (for example, *man* and *car*), which then could be sub-classified as *chess player*, *pole vaulter*, *sedan* and *taxi*;

(2) There should be a sufficient number of appropriate images available to illustrate objects from the selected thematic subdomains.

Ontologies are classified into three basic types: *top ontologies*, which contain a restricted set of general classes and are not related to a particular thematic domain; *top-domain ontologies*, which include essential classes that represent a particular thematic domain; and *domain ontologies*, which contain classes that comprehensively describe a particular thematic domain (Tan and Lambrix, 2009). From the point of view of this classification, the proposed ontology can be classified as a set of several domain ontologies.

The **Ontology of Visual Objects** provides options for extracting relationships between annotated objects, between diverse datasets with different levels of granularity of object classes, or between appropriate sets of images illustrating different thematic domains. Last but not least, the use of the Ontology of Visual Objects allows the expansion of the dataset depending on the specific needs of scientific or commercial projects.

The annotators' tasks were to create new polygons or approve or modify the automatic segmentation for objects in the images, and then classify the objects according to the specified Ontology's classes. The annotation adheres to the following conventions:

- An object displayed within an image is annotated if it represents an instance of a concept included in the Ontology.

- All objects from the selected dominant class and attribute classes related with it are annotated (for example, the *tennis player* and the related objects *racket* and *tennis ball*; *chess player* and the related objects *chessman*, *chess board* and *clock*).

The following are some advantages of utilizing an ontology for object classification:

- Selection of mutually exclusive classes.
- Build-in interconnectivity of classes by means of formal relations.
- Easy extension of the proposed ontology with more concepts corresponding to visual objects.

4 Ontology of Visual Objects

It was pointed out that different knowledge representations share the following minimal set of components (Corcho et al., 2006): **concepts**, which represent sets or classes of entities in a thematic domain; **relations** between concepts; **instances**, which represent the actual entities (individuals); and **axioms**, which represent facts that are always true in the topic area of the ontology. We accepted the following definition (Bozsak et al., 2002): An ontology is a structure

$$O := (C, \leq_C, R, \leq_R)$$

consisting of (i) two disjoint sets C and R called concept identifiers and relation identifiers respectively, (ii) a partial order $\leq_C$ on C called concept hierarchy or taxonomy, (iii) a function $\sigma : R \rightarrow C \times C$ called signature and (iv) a partial order $\leq_R$ on R called relation hierarchy.

The **Ontology of Visual Objects** organizes concepts (represented by dominant classes, classes that are attributes to dominant classes and classes that serve only as parents to dominant classes), relations between concepts and axioms.

4.1 Classes

Classes correspond to (WordNet) concepts that can be represented by visual objects. Among the classes, we made a differentiation between dominant classes and attribute (contextual) classes.

Each thematic domain is represented by several **dominant classes**, which show the main "players" within this domain differentiated by their type or their function. For example, the dominant classes for the domain Security are: **policeman, soldier, fireman**, etc., altogether 15 dominant classes. For the definition of the dominant classes, we use the WordNet sister **hyponyms** at a certain level (the lowest level allowing classification without specific knowledge for the domain). So far, the selected dominant classes for all thematic domains in focus are 137.

For each dominant class a parent class is selected from the WordNet noun hierarchies and this procedure is repeated consecutively up to the final class that represents a visual object. For example, classes like *basketball player, acrobat, football player*, etc. are **hyponyms** of *athlete* 'a person trained to compete in sports'. *Athlete* in its turn is a **hyponym** of *contestant* 'a person who participates in competitions' which is a hyponym of *person*. However, the **hypernym** of *person* is *organism*, an abstract notion, which is not included in the ontology. As a result of this approach, thousands of annotations will be assigned to objects representing a small number of classes, while the annotations with more general classes will be inherited automatically. The WordNet hierarchical trees are very detailed, that is way only hypernyms, which are visual objects are selected with only one abstract notion on the top. For example, *jersey* is a *shirt*, which, in turn, is a *clothing*. From the hierarchy the node *garment* (an article of clothing) between *shirt* and *clothing* is excluded.

The Ontology design organized the 851 concepts into 11 categories on the top level, such as *person, animal, furniture, equipment* and so on (approximately half of the Ontology classes are contained in WordNet, 485 out of 851 classes)).

Following the strategy for category selection of the ImageNet, we applied the rule for no overlapping between the dominant classes and their attributes: "for any synsets i and j, i is not an ancestor of j" (Deng et al., 2009). Mutually exclusive classes are also defined for other well-known datasets, for example for the COCO thing and stuff classes (Caesar et al., 2018). As pointed out, the mutual inclusion might lead to some inconsistencies. An example was given with the PASCAL Context (Mottaghi et al., 2014) classes *bridge* and *footbridge*, which are in a parent-child relation (Caesar et al., 2018). The parent term can replace the child term in some context, but not vice versa; thus: if two images are annotated as *bridge* and *footbridge* respectively, it will not be known whether the parent concept can refer also to the child concept or not.

Attributes in the ontology are classes related with the dominant ones. The type of the dominant class and the type of attribute class determine the type of the relation between them, which expresses the specificity of property attribution: **wears, uses,**

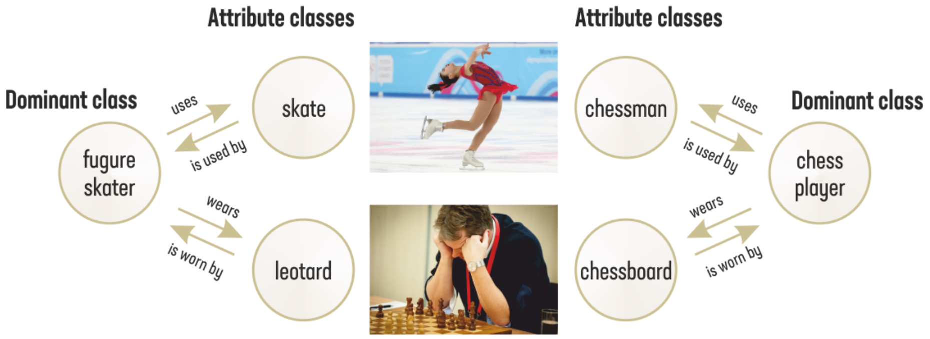

Figure 1: Attribute classes in the Ontology

has part, etc. For example, the attribute classes for *cricketer* are *cricket bat, cricket ball, cricket helmet, wicket* and *referee*, while for *climber – climbing helmet, chalk bag, claiming backpack*; the attribute classes for *chess player* are *chessman* and *chessboard*, and for the *figure skater – skate* and *leotard* (Figure 1), and so on.

For the definition of attribute classes, we use some WordNet relations, such as meronymy. In most of the cases, such relations are not overtly established in WordNet and they are additionally defined in the Ontology.

4.2 Relations

The Ontology not only specifies the visual concepts, but also defines the relationship between concepts. Thus the **relations** used in the Ontology are relations between classes. The **is-a** relation is inherited from WordNet, where nouns build hierarchical structures based on the relations of **hypernymy** and **hyponymy**, assuming that WordNet contains representation for both members of the relation. When it comes to new concepts (not presented in WordNet), they are connected to the proper parent concept in WordNet.

Depending on their properties, the relations do or do not project hierarchical structures. Hierarchical relations (relations of inclusion) are of three basic types – taxonomic (classificatory, which associate an entity of a particular type with an entity of a more generic type), meronomic (expressing the relation of the whole to its parts) and proportional series (expressing proportions between values in a given series) (D. A. Cruse, 1996). Taxonomic relations are inverse and transitive (**is-a**) and meronomic relations are also inverse and could be transitive (**has part**). Non-hierarchical relations are inverse and non-transitive (most of the relations between dominant classes and their attribute classes), and symmetric, irreflexive and non-transitive (**is next to**).

Relation	Reverse R	Number
has hyponym	is hyponym of	827
wears	is worn by	241
has part	is part of	210
uses	is used by	119
is next to		34
plays with	is a devise for	23
is on	is a surface for	22
drives	is driven by	18
plays	is played by	17
is in	is around	15
operates	is operated by	14
propel	is propelled by	12
plays at	is where to play	10
creates	is created by	9
rides	is ridden by	9

Table 1: Types of relations and number of their occurrences

Relations between dominant and attribute classes are not hierarchical. For the linking of attribute classes, we use one WordNet relation – **has part** and 13 relations that are not overtly established in WordNet and are additionally created for the Ontology, for example, (**wears**, **is next to** and **plays with**). Altogether, 15 relations are used in the Ontology, with 827 instances of the **is a** relation; 241 instances of the **wears** relation, 210 instances of the **has part** relation, and so on. Table 1 shows the relations included in the Ontology of Visual Objects, their properties and number of occurrences.

4.3 Axioms

Axioms serve to model sentences that are always true (Gruber, 1995) and they can be used to infer new knowledge.

An axiom system for an ontology is a pair (AI, α) where (i) AI is a set whose elements are called axiom identifiers and (ii) α is a mapping. The elements of $A := \alpha(AI)$ are called axioms (Cimiano and Handschuh, 2003).

Axioms are assertions that are driven by the properties of the relations. In the **Ontology of Visual Objects** the axioms are:

If X is a hypernym of Y, then Y is a hyponym of X.

If X is a hypernym of Y, and Y is a hypernym of Z, then X is also a hypernym of Z.

If X is a holonym of Y, then Y is a meronym of X.

If X is a holonym of Y, and Y is a holonym of Z, then X is also a holonym of Z.

If X plays Y, then Y is played by X.

If X wears Y, then Y is worn by X.

If X uses Y, then Y is used by X.

If X plays at Y, then Y is a place where X plays.

If X plays with Y, then Y is a device with which X plays.

If X is on Y, then Y is a surface on which X is.

If X rides Y, then Y is ridden by X.

If X propel Y, then Y is propelled by X.

If X drives Y, then Y is driven by X.

If X creates Y, then Y is created by X.

If X is in Y, then Y is around X.

If X is next to Y, then Y is next to X

The set of non-hierarchical relations, which hold among target concepts, also holds among higher concepts, for example if a *soccer player* is next to a *referee*, then a *person* is next to a *person*.

4.4 Ontology format

The concepts are represented by the respective WordNet ILI (Inter-Lingual-Index) number or an Ontology index (if the concepts are not represented in WordNet) and a unique label: either the most representative literal (synonym) from the WordNet synsets or a term picked as a more adequate to refer to the concept. The differentiation between dominant, attribute and only hypernym classes is explicitly stated. The relations between classes are also explicitly stated. In case of reverse relations, only the direct relation is encoded, and in case of symmetric relations only one record of the relation is encoded. The Ontology is defined in a JSON format. For example:

{
”HYPONYM_ID”: ”eng-30-09761310-n”,
”HYPONYM_LEMMA”: ”accordionist”,
”RELATION”: ”IS A”,
”HYPERNYM_ID”: ”eng-30-10340312-n”,
”HYPERNYM_LEMMA”: ”musician”
},

The Ontology is intended to be language-independent but the concepts are attached manually with labels in English and Bulgarian. All Ontology classes (used as annotation labels) have been presented in 25 languages: English (Princeton WordNet), Bulgarian, Albanian, Basque, Catalan, Croatian, Danish, Dutch, German, Greek, Finnish, French, Galician, Icelandic, Italian, Lithuanian, Polish, Portuguese, Romanian, Russian, Serbian, Slovak, Slovene, Spanish, Swedish. In providing translation equivalents to Ontology classes, priority is given to WordNet, employing openly available wordnets from the Extended Open Multilingual Wordnet project or official distribution webpages of particular wordnets. The synonyms of Ontology classes, the definitions of the concepts and some usage examples (if available) were extracted from the synsets in different languages. Where WordNet translations are not available, some additional sources of translations are employed: BabelNet and Machine translation (Koeva, 2021; Koeva et al., 2022).

5 Evaluation

A number of studies aimed at ontologies' evaluation are known (Hlomani and Stacey, 2014; Vrandečić, 2009; Raad and Cruz, 2015; Walisadeera et al., 2016; Khalilian, 2019; Wilson et al., 2021). On their basis several criteria for the evaluation of ontologies can be defined directed to confirm the ontology quality and correctness:

- Accuracy states if the definitions of classes are correct.
- Completeness measures if the domain of interest is appropriately covered.
- Conciseness states that the ontology does not include any unnecessary or useless definitions or explicit redundancies between definitions of terms do not exist.
- Adaptability measures if the ontology offers the conceptual foundation for a range of anticipated tasks.

- Clarity measures how effectively the ontology communicates the intended meaning of the defined concepts.
- Computational efficiency measures the ability of the used tools to work with the ontology.
- Consistency describes that the ontology does not include or allow for any contradictions.

We can define our approach for the evaluation as a corpus-based approach (Raad and Cruz, 2015). Instead of comparing an ontology with the content of a text corpus that covers significantly a given domain, we use the image annotation process to evaluate the Ontology of Visual Objects. At the beginning, we have identified 1,037 classes grouped in ten thematic domains: Sport, Medicine, Arts, Education, Food, Transport, Clothing, Security, Indoors, and Nature. For four of them (Sport, Transport, Arts and Security) an evaluation of the Ontology classes is performed during the annotation: whether a class is a visual object or not; whether all depicted objects in selected images can be described with the Ontology classes; and whether new classes can be added if necessary.

For the definition of classes we rely on the definition of concepts in WordNet; the definition of new classes is provided by means of finding their correct place within the WordNet taxonomy by linking them with already defined concepts. Finally, we made some evaluation tests for all selected classes with other sources providing lists with concrete objects, such as concreteness ratings (Brysbaert and Biemiller, 2017), word acquisition ratings (in our case of nouns) (Kuperman et al., 2012) and picture dictionaries (Parnwell, 2008).

6 Conclusion and Future Work

To improve object annotation and classification, several approaches based on ontologies have been proposed. However, image classification and annotation remain a challenging problem and one of the reasons is possible overlapping of selected classes. The use of a specially designed ontology improves the speed of object annotation as well as the accuracy of object classification.

Our contributions consist of the following:

(1) Definition of an Ontology of Visual Objects, whose classes are sufficient to annotate objects in 130 thematic subdomains related in four general domains;

(2) Introduction of attribute classes, which, in general, are related to the location, function and context of objects in focus (the dominant classes);

(3) Definition of relations between dominant and attribute classes and formulations of axioms based on the properties of the relations. This offers a reliable means for automatic object or image description, automatic assignment of image captions or classification of images and objects.

Using the **Ontology of Visual Objects** ensures the selection of mutually exclusive classes, built-in interconnectivity of classes via formal relations, and the ability to easily extend the proposed ontology with more concepts corresponding to visual objects.

Applying semantics can improve not only the performance of object recognition but also the performance and quality of individual tasks required for object recognition, such as image segmentation. Furthermore, the Ontology can be used to reduce the gap between human image comprehension and machine image interpretation, allowing for better automation in training neural networks (Bhandari and Kulikajevas, 2018).

A possible application of the Ontology of Visual Objects includes further use of the relations to compile bigger training datasets (for example, utilizing higher level concepts) or to construct contexts in which a particular object may or may not appear. The Ontology of Visual Objects provides options for extracting: relationships between annotated objects, diverse datasets with different levels of granularity of object classes and appropriate sets of images illustrating different thematic domains.

The ontological organization of object classes provides data for learning associations between objects in images, for identifying relations between objects and for aligning objects and relations with text fragments. Last but not least, using the Ontology of Visual Objects enables the dataset to be expanded based on the particular needs.

Acknowledgments

The Multilingual Image Corpus (MIC21) project was supported by the European Language Grid project through its open call for pilot projects. The European Language Grid project has received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement no. 825627 (ELG).

References

Alejandro Agostini, Mohamad Javad Aein, Sandor Szedmak, Eren Erdal Aksoy, Justus Piater, and Florentin Würgütter. 2015. Using structural bootstrapping for object substitution in robotic executions of human-like manipulation tasks. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6479–6486. IEEE.

Peter Auer, Isabelle Bloch, Hilary Buxton, Patrick Courtney, Sven Dickinson, Bob Fisher, Goesta Granlund, Walter Kropatsch, Giorgio Metta, Bernd Neumann, Axel Pinz, Giulio Sandini, Gerald Sommer, David Vernon, Aude Billard, Pia Boettcher, Henrik Christensen, Andrew Crookell, Christof Eberst, Wolfgang Forstnerand Vaclav Hlava cand Ales Leonardis, Hans-Hellmut Nagel, Heinrich Niemann, Fiora Pirri, Bernt Schiele, John Tsotsos, Markus Vincze, Horst Bischof, Heinrich Bulthof fand Tony Cohnand James Crowleyand Jan-Olof Eklundhand John Gilbyand Josef Kittler, Jim Little, Bernhard Nebel, Lucas Paletta, Gerhard Sagerer, Rebecca Simpson, and Monique Thonnat. 2005. *CA Research Roadmap of Cognitive Vision*. ECVision: The European Research Network for Cognitive Computer Vision Systems.

Sandeepak Bhandari and Audrius Kulikajevas. 2018. Ontology Based Image Recognition: A Review. In *Proceedings of the International Conference on Information Technologies*, pages 13–18.

Jan Botorek, Petra Budíková, and Pavel Zezula. 2014. Visual Concept Ontology for Image Annotations. *CoRR*, abs/1412.6082.

Erol Bozsak, Marc Ehrig, Siegfried Handschuh, Andreas Hotho, Alexander Maedche, Boris Motik, Daniel Oberle, Christoph Schmitz, Steffen Staab, Ljiljana Stojanovic, et al. 2002. KAON — towards a large scale Semantic Web. In *International Conference on Electronic Commerce and Web Technologies*, pages 304–313. Springer.

M. Brysbaert and A. Biemiller. 2017. Test-based age-of-acquisition norms for 44 thousand English word meanings. *Behavior research methods*, 49(5):1520–1520.

Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. 2018. COCO-stuff: Thing and stuff classes in context. In *Conference on Computer Vision and Pattern Recognition*, pages 1209–1218.

Parkpoom Chaisiriprasert, Karn Yongsiriwit, Matthew N. Dailey, and Chutiporn Anutariya. 2021. Ontology-based Framework for Cooperative Learning of 3D Object Recognition. *Applied Sciences*, 11(17).

Philipp Cimiano and Siegfried Handschuh. 2003. Ontology-based Linguistic Annotation. In *Proceedings of the ACL 2003 Workshop on Linguistic Annotation: Getting the Model Right*, pages 14–21, Sapporo, Japan. Association for Computational Linguistics.

Óscar Corcho, Mariano Fernández-López, and Asunción Gómez-Pérez. 2006. Ontological Engineering: Principles, Methods, Tools and Languages. In Coral Calero, Francisco Ruiz, and Mario Piattini, editors, *Ontologies for Software Engineering and Software Technology*, pages 1–48. Springer.

D. A. Cruse. 1996. *Lexical Semantics*. Cambridge University Press, Cambridge.

Jia Deng, Wei Dong, Socher Richard, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, page 248–255.

Zheyuan Ding, Li Yao, Bin Liu, and Junfeng Wu. 2019. Review of the Application of Ontology in the Field of Image Object Recognition. In *Proceedings of the 11th International Conference on Computer Modeling and Simulation, ICCMS 2019, North Rockhampton, QLD, Australia, January 16-19, 2019*, pages 142–146. ACM.

Christiane Fellbaum, editor. 1999. *WordNet: an Electronic Lexical Database*. MIT Press, Cambridge, MA.

Greg Griffin, Alex Holub, and Pietro Perona. 2007. Caltech-256 object category dataset. In *Technical Report 7694*, page 1–20, California Institute of Technology.

Thomas R. Gruber. 1995. Toward principles for the design of ontologies used for knowledge sharing. *Int. J. Hum.-Comput. Stud.*, 43:907–928.

Hlomani Hlomani and Deborah Stacey. 2014. Approaches, methods, metrics, measures, and subjectivity in ontology evaluation: A survey. *Semantic Web Journal*, 1(5):1–11.

Saiede Khalilian. 2019. A Survey on Ontology Evaluation Methods. *Quarterly Knowledge and Information Management Journal*, 6(2):25–34.

Svetla Koeva. 2021. Multilingual Image Corpus: Annotation Protocol. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 701–707, Held Online. INCOMA Ltd.

Svetla Koeva, Ivelina Stoyanova, and Jordan Kralev. 2022. Multilingual Image Corpus – Towards a Multimodal and Multilingual Dataset. In *Proceedings of the Language Resources and Evaluation Conference*, pages 1509–1518, Marseille, France. European Language Resources Association.

Victor Kuperman, H. Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, 44:978–990.

Egoitz Laparra, German Rigau, and Piek Vossen. 2012. Mapping WordNet to the Kyoto ontology. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2584–2589, Istanbul, Turkey. European Language Resources Association (ELRA).

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2014. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision (ECCV)*, pages 740–755, Zürich.

George Miller. 1986. Dictionaries in the mind. *Language and Cognitive Processes*, 1:171–185.

George Miller, R. Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography*, 3:235–244.

Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. 2014. The role of context for object detection and semantic segmentation in the wild. In *Conference on Computer Vision and Pattern Recognition*, pages 891–898.

Milind R. Naphade, John R. Smith, Jelena Tesic, Shih-Fu Chang, Winston H. Hsu, Lyndon S. Kennedy, Alexander G. Hauptmann, and Jon Curtis. 2006. Large-scale Concept Ontology for Multimedia. *IEEE Multim.*, 13(3):86–91.

Roberto Navigli, Michele Bevilacqua, Simone Conia, Dario Montagnini, and Francesco Cecconi. 2021. Ten Years of BabelNet: A Survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4559–4567. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

E. C. Parnwell. 2008. *The New Oxford Picture Dictionary*. Oxford University Press, New York, Oxford.

Joe Raad and Christophe Cruz. 2015. A survey on ontology evaluation methods. In *Proceedings of the International Conference on Knowledge Engineering and Ontology Development, part of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*.

Luis Riazuelo, Moritz Tenorth, Daniel Di Marco, Marta Salas, Dorian Gálvez-López, Lorenz Mösenlechner, Lars Kunze, Michael Beetz, Juan D Tardós, Luis Montano, et al. 2015. RoboEarth semantic mapping: A cloud enabled knowledge-based approach. *IEEE Transactions on Automation Science and Engineering*, 12(2):432–443.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 116:157–173.

Bryan C. Russell, Antonio Torralba, Kevin P. Murphy, and William T. Freeman. 2008. LabelMe: a database and web-based tool for image annotation. *International Journal of Computer Vision*, 77:157–173.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: A Core of Semantic Knowledge. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 697–706, New York, NY, USA. ACM.

He Tan and Patrick Lambrix. 2009. Selecting an Ontology for Biomedical Text Mining. In *Proceedings of the BioNLP 2009 Workshop*, pages 55–62, Boulder, Colorado. Association for Computational Linguistics.

Anne-Marie Tousch, Stéphane Herbin, and Jean-Yves Audibert. 2012. Semantic hierarchies for image annotation: A survey. *Pattern Recognit.*, 45(1):333–345.

David Vernon. 2006. The Space of Cognitive Vision. In *Cognitive Vision Systems, Sampling the Spectrum of Approaches [based on a Dagstuhl seminar]*, pages 7–24.

David Vernon. 2021. *Cognitive Vision*, pages 164–167. Springer International Publishing, Cham.

Denny Vrandečić. 2009. Ontology evaluation. In *Handbook on ontologies*, pages 293–313. Springer.

Anusha Indika Walisadeera, Athula Ginige, and Gihan Nilendra Wikramanayake. 2016. Ontology evaluation approaches: a case study from agriculture domain. In *International Conference on Computational Science and Its Applications*, pages 318–333. Springer.

RSI Wilson, JS Goonetillake, WA Indika, and Athula Ginige. 2021. Analysis of Ontology Quality Dimensions, Criteria and Metrics. In *International Conference on Computational Science and Its Applications*, pages 320–337. Springer.

Xiao Xie, Xiran Zhou, Jingzhong Li, and Weijiang Dai. 2020. An Ontology-based Framework for Complex Urban Object Recognition through Integrating Visual Features and Interpretable Semantics. *Complexity*, 44:1–15.

Eleftherios Spyromitros Xioufis, Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis P. Vlahavas. 2011. MLKD's Participation at the CLEF 2011 Photo Annotation and Concept-based Retrieval Tasks. In *CLEF 2011 Labs and Workshop, Notebook Papers, 19-22 September 2011, Amsterdam, The Netherlands*.