Multimedia ontology population through semantic analysis and hierarchical deep features extraction techniques

Michela Muscetti1 · Antonio M. Rinaldi1 · Cristiano Russo1 · Cristian Tommasino1

Received: 5 July 2021 / Revised: 14 February 2022 / Accepted: 17 February 2022 / Published online: 25 April 2022 © The Author(s) 2022, corrected publication 2022

Abstract
The rapid increase of available data in different complex contexts needs automatic tasks to manage and process contents. Semantic Web technologies represent the silver bullet in the digital Internet ecosystem to allow human and machine cooperation in achieving these goals. Specific technologies as ontologies are standard conceptual representations of this view. It aims to transform data into an interoperability format providing a common vocabulary for a given domain and defining, with different levels of formality, the meaning of informative objects and their possible relationships. In this work, we focus our attention on Ontology Population in the multimedia realm. An automatic and multi-modality framework for images ontology population is proposed and implemented. It allows the enrichment of a multimedia ontology with new informative content. Our multi-modality approach combines textual and visual information through natural language processing techniques, and convolutional neural network used the features extraction task. It is based on a hierarchical methodology using images descriptors and semantic ontology levels. The results evaluation shows the effectiveness of our proposed approach.

Keywords Image classification · Semantic analysis · Multimedia ontologies · Deep neural networks · NoSQL

Michela Muscetti, Antonio M. Rinaldi, Cristiano Russo and Cristian Tommasino contributed equally to this work.

Antonio M. Rinaldi antoniomaria.rinaldi@unina.it
Michela Muscetti muscettimichela@gmail.com
Cristiano Russo cristiano.russo@unina.it
Cristian Tommasino cristian.tommasino@unina.it

1 Department of Electrical Engineering and Information Technology, University of Napoli Federico II, Via Claudio, 21, 80125 Naples, Italy
1 Introduction

Data has a central role in everyday life activities, and they represent one of the most valuable items in modern societies. Many available data produced every day in different formats by humans, and automatic agents need efficient and effective techniques to represent and manage them. Other dimensions have arisen from this plethora of data, and many efforts have been devoted to define and implement novel approaches to share and represent data using common formats [1].

In this context, ontologies and the related semantic web technologies represent a crucial aspect of allowing the transformation from human-readable to machine-readable data. They enable real advanced information processing and management techniques. The different definition has been proposed for years about the term ontology. In our vision, one of the most complete is presented in [2] where an ontology is a conceptual model of a specified reality; it has explicit definitions of its components; it is formally defined, and therefore it is “machine-readable”; it is one most accepted by the scientific community. Ontology Construction, enrichment, and adaptation tasks are a wide process called Ontology Learning [3]. In particular, ontology enrichment is the task of increasing an existing ontology with concepts and relations. On the other hand, the ontology population task can add new instances of concepts to the ontology. In this paper, we focus our attention on the Ontology Population task. The use of an ontology has several advantages in different fields [4], and it can be used to reduce the semantic gap [5, 6]. In particular, if we consider multimedia data, an ontology-based model can be designed to relate low-level features to high-level concepts [7]. Due to the high number of possible entities and properties involved in representing a knowledge domain, the ontology building process is very expensive and time-consuming. On the other hand, the updating task of an ontology instance requires frequent modifications. Therefore, one goal is to automate this process.

In our approach, the ontology evolution consists of extracting informative contents from different data sources and populating as a priori defined ontology schema. In our context, we consider a multimedia ontology schema and use low-level extraction techniques and semantic concept analysis to implement the ontology population process. Following the Ontology Learning, Layer Cake [8], the ontology learning process involves different tasks. In particular, the tasks concerning the ontology population process are related to object identification, through which an extracted object from a corpus can be associated with a single concept in the domain of interest. If we analyse textual data, this task consists in the recognition of terms or groups of words. On the other hand, if we analyse visual data, it consists of identifying particular areas of considered images that represent objects for the ontology population.

The object identification task is divided into the following sub-tasks: (i) Object Recognition used to find recognizable entities in a corpus or visual data; (ii) Object Classification related to the identification of specified categories (i.e. ontology concepts) assigned to the identified objects; (iii) Object Mapping to compute similarity between contents in different data sources; (iv) Synonyms Identification that refers in recognition of different representations of the same concepts and vice versa; (v) Concept Identification and Data Association which assigns a set of instances to a concept in the ontology (i.e. the ontology population step); (vi) Entity Disambiguation given by the process of identifying instances that refer to the different concepts. We explicit point out that we extend the model described so far considering multimodal representations (i.e. visual information). The Ontology Population deals with adding new instances of concepts to ontology without changing its structure. There-
fore, after the population process, neither the hierarchy of concepts nor the non-taxonomic relations change. The population process requires an existing ontology and an extraction engine that processes data, identifying objects associated with concepts. Different surveys have been proposed in the context of ontology learning and population and, due to its essential applications in various fields, it is a very prolific research field [9].

This paper proposes a novel approach for the ontology population using different visual descriptors and semantic analysis. The process is wholly automatized, and it has been implemented using hybrid big data technologies.

The paper is organised as follows: in Sect. 2, we introduce and discuss the existing literature; Sect. 3 is devoted to the presentation of the proposed approach and its implementation together with the description of the used multimedia ontology model and the used data sources; the experimental strategy and the obtained results are showed in Sect. 4 and, eventually, conclusions and future works are in Sect. 5.

2 Related works

This section presents and discusses some of the main works in the literature related to our approach from an application point of view (i.e. ontology population) and a methodological one (i.e. data sources and analysis techniques).

One of the main projects in our field of interest is BOEMIE [10]. BOEMIE adopts a synergistic approach that combines multimedia extraction and ontology evolution in a bootstrapping process. Its aim is the improvement of the acquisition process of multimedia content using ontologies. This process is used to enrich ontologies to drive a multimedia information extraction task. The system uses a semi-automatic approach to control the annotation task. An approach to populate an image ontology based on textual representation is presented in [11]. It combines a segmentation process and a ranking function considering a colour distribution. A web search engine fetches the images considered for the population process. A waterfall segmentation algorithm is applied to object recognition. The textual information is associated with a single region detection. The obtained regions are clustered through an unsupervised algorithm using texture and colour features. The concept identification is performed by assigning a score to each cluster. An approach for ontology population using linked-open data is proposed in [12]. It is based on domain-specific ontology and the DBpedia ontology structure. The system is based on three steps related to image fetching for the ontology population retrieved from the Web and textual information. Each image is segmented, and the textual information is associated with each region. This operation is done through the Labelme annotation tool [13]. It associates concept to image using colour and textural information. Moreover, concepts description are fetched from DBpedia, and low-level features are extracted from images. Each image and its candidate concept is shown to a group of users. Each user gives a score to each image. The concept with the highest score is assigned to the considered image. PROPheT [14] is a software tool for the population and enrichment of a local ontology model with different instances fetched from several Linked Data sources served by SPARQL endpoints as DBPedia. In the realm of big data, the Bishop project [15] presents methods for ontology population-based on big data-driven and large-scale self-learning support. A conceptual framework is recognised from the user requirements to capture the integration schema for self-learning methods in the proposed approach. Different big data frameworks are taken into account to measure the performance in the data analysis step. The results comparisons are set using test scenarios.
A system to extract information from heterogeneous sources is presented in [16]. It populates an ontological knowledge base using a rule-based information extraction approach to recognise named entities. They are added into semantic structures using declarative mapping rules. OntoPRiMa [17] is a system for semi-automatic ontology population based on text analysis combining NLP techniques, semantic approaches and web technologies. It uses an automatic weighting schema to evaluate and support the decision process in the ontology population. Specific knowledge domains are also investigated for the ontology population. In [18] a method for ontology population in the domain of cultural heritage is proposed. It is based on an automatic approach to extract instances from semi-structured corpora with the support of extraction patterns manually defined. In [19] a tool for extracting semantic information from CAD drawings and populate an ontology is presented. The drawing primitives are considered to perform pattern matching, and classification algorithms extract the semantic information. The resulting information is mapped to the corresponding ontology classes, and related individuals are created to populate the ontology. The authors in [20] present a system able to extract web document contents to populate an ontology related to the tourism domain. Their methodology is based on a two-stage approach where the system obtains a set of semantic annotations considered as possible candidates of ontology instances. In the second step, semantic ambiguities are detected and solved to relate the annotations to the right concept in the ontology.

Our proposed approach differs from the ones discussed so far in terms of automation degree, and we combine different data sources using both textual and visual information. The aim of the implemented system is the ontology population using multiple informative sources analysed through semantic and deep learning techniques. Different techniques have been proposed in the literature [21–26] and, over years, the use of deep convolutional neural network as feature extractor exceeds the state of the art [27–29]. The deeper levels of a CNN underline lower-level features while the last level points out higher-level concepts [30]. This suggests that lower-level features are used for fine-grained tasks while higher-level features are used for semantic similarity. The lower-levels features suffer from two fundamental problems: semantic ambiguity [31] and background-clutter [32]. In literature, some approaches suggest combining the features extracted from multiple levels. In this way, the strength of complementarity is exploited and, semantic and structural features are combined. However, it has been shown that direct concatenation between the features coming from several layers not permit positive results, not only because of high dimensionality but also because the strong influence of high-level features on low-level features weakens the effectiveness of the latter. To address these issues, in [33] is proposed an approach for the semantic category identification through descriptors from the last convolutional layer and, after this step, it makes images ranking through features extraction from lower and intermediate levels. In [34], instead, the concept of hypercolumn is explored. It considers the output, at a given location, of all the units that are above a given level. The reason is that all pieces of information are distributed on all network layers. The combination of multiple levels, as happens in the approaches just presented, leads to an excessively large size descriptor. In this case, it would be necessary to apply an embedding technique and to avoid overfitting [35]. An important aspect is the choice of aggregation methods for locals descriptors. In the literature, several aggregation methodologies have been presented. One of the most used is based on Bag Visual Words [36]. The idea is to quantise local invariant descriptors into a set of visual words. The frequency vector of the visual words represents the image, and an inverted index file is used for efficient comparison of such Bag of Visual Words. Another methodology as the Fisher Vector [37] transforms an incoming variable-size set of independent samples into a fixed-size vector representation, assuming that the samples follow a parametric generative
model estimated on a training set. This description vector is the gradient of the sample like-
lihood concerning the parameters of this distribution, scaled by the inverse square root of the
Fisher information matrix. The \textit{VLAD} representation \cite{38} can be seen as a simplification of
the Fisher kernel. \textit{VLAD} aggregates descriptors based on a locality criterion in the feature
space. Another aggregation method is \textit{Triangular embedding}, and it is based on an anchor
graph and realises triangulation operation \cite{39}. Another novelty of our approach is the use
of more fine-grained tasks in the ontology population process because the ontology is built
hierarchically, starting from high semantic level concepts to specialisations. Our goal is to
identify, for each input image, the related semantic category for concept validation and its
subcategory for concept identification.

3 The proposed approach

In this section, we describe all aspects of our approach for ontology population. The proposed
methodology is based on Natural Languages Processing (NLP) and deep learning techniques.
The section describes the system architecture highlighting the main task of each module. This
is summarized in Fig. 1.

The ontology population starts with an existing ontology. The ontology that we want to
populate is ImageNet \cite{40} and it represents our knowledge base. It organizes into different
classes of images in a densely populated semantic hierarchy. ImageNet follows the same
structure as WordNet. Each node represents a synset. Synsets are a group of synonymous
words that express the same concept. Representative images are associated with each synset.
In our approach, Imagenet as been represented as a multimedia ontology represented by OWL
following an ontology-based model \cite{41}. ImageNet has been imported into two different
NoSQL Databases following a methodology and model proposed in \cite{7, 42–44}. We choose
Neo4j graph DB to store the ontology structure and Mongo document-oriented DB to store
images as feature vectors. The main problem of ImageNet is that not all synsets are populated.
For this reason, ImageNet can be seen as a thinly populated multimedia ontology, and we
can use it as a real example to realize our population framework. We perform an analysis of
ImageNet to find not populated synsets. For each root node of each ImageNet sub-tree, we
use ImageNet API to fetch the number of images for each synset. In this way, we can know
which nodes are empty and which semi-populated. The first step is data sources identification.

![Fig. 1 System architecture](image-url)
We retrieve data from two different types of data sources, and we use both an image search engine (i.e. Google image) and an images dataset (i.e. COCO data set [45]). Our goal is to integrate data sources with different representations [46, 47] to accomplish a multi-modal image ontology population system using both visual and textual analysis combining different techniques.

After the data source identification, we have to validate the detected concept and associate images to concepts of our ontology. The concept validation and association is obtained through feature extraction module. Then, a thesaurus is used to support concept identification. So, we combine textual and visual information by performing semantic analysis. In the case of images retrieved by Google, the concept to be validated is the one used in the query formulation. Instead, in the case of COCO images, we realize an ontology alignment operation to associate COCO categories to ontology concepts. The association of visual object representations with concepts are computed with a features extraction process. After the concept association, we proceed with the ontology population. The feature extraction module is used to perform the concept association and validation. The features extraction engine is based on a hierarchical approach because our aim is to distinguish between a high-level category (i.e. root node) and a low level one represented by a child node for each sub-tree to populate. This approach is implemented in two steps. In the first step, the validation is performed by hyperonym identification through a global descriptor. The second one is used to associate concepts through hyponym identification. The association of low-level features to high-level concepts occurs through a semantic interpretation module based on reasoning techniques. At both levels, the extracted descriptors are aggregated in a global and compact feature vector representation. Features matching operation is performed through an image similarity measure. After the concept validation task, we finally carried out the Ontology population step. The following subsections will be explained in detail our framework.

3.1 Hierarchical deep feature extraction module

The proposed features extraction engine works hierarchically. The process is shown in Fig. 2. We consider hyperonym at the higher level and all hyponyms at the same hierarchical level of the considered synset. Therefore, we have a two-concepts levels hierarchy using the ImageNet structure. With the first level of the hierarchy, hypernym identification takes

![Feature extraction engine](image)

**Fig. 2** Feature extraction engine
place. We perform a concept validation verifying that the input image concept belongs to the correct semantic category. An example of hypernym identification is in Fig. 3, where the label node in the figure is composed of a lemma, part of speech, and a number that is a rank value obtained by the concept sorting by association frequency of it to lemma. The colour of the label is different by hyponym level. Instead, the size is by edge degree.

Both local and global descriptors are extracted using a convolutional neural networks due to the performance of these architectures [30, 48]. A global descriptor is extracted from the last convolutional layer of the DNN. Instead, a local descriptor is represented by features maps values concatenation on the same spatial position. The fine-grained similarity is obtained by applying a feature selection mechanism that allows an automatic object recognition. The selection of a representative subset of local convolutional features can remove a large number of redundant features improving the object recognition process. We use the Selective Convolutional Descriptor Aggregation for Fine-Grained Image Retrieval (SCDA) method [49] to extract the local feature. The method is based on a pre-trained VGG16 model to perform a fine-grained image retrieval task. VGG is a CNN (Convolution Neural Network) architecture
Selective convolutional descriptors aggregation proposed by [49] is the most common used is the VGG16 that consists in 16 convolution layers with $3 \times 3$ kernels. SCDA approach is summarized in Fig. 4 [49]. Following this approach we change the Mask Map using more accurate techniques as detailed in the following. For the sake of clarity we briefly describe the entry process. The SCDA process starts with the extraction of a $h \times w \times d$ 3-D tensor of an image from pool-5 and it is transformed in a $h \times w$ 2-D tensor, called aggregation map ($A$). The aggregation map is computed as $A = \sum_{n=1}^{d} S_n$, where $S_n$ is a feature map in pool-5. Then, the mask map $M$ is obtained as:

$$M_{i,j} = \begin{cases} 1 & \text{if } A_{i,j} > \bar{a} \\ 0 & \text{otherwise} \end{cases}$$

where $A$ and $M$ has same size, $(i, j)$ is a particular position in these $h \times w$ positions and $\bar{a}$ is the threshold computed as mean value of all positions in $A$. In order to improve the main object selection, an algorithm the collect the largest connected component of $M$ is applied. Therefore, the selected feature $F$ is:

$$F = \left\{ x_{(i,j)} | \tilde{M}_{(i,j)} = 1 \right\}$$

where $\tilde{M}$ is the selected valid and meaningful deep convolutional descriptor. Then, it applies on $F$ an algorithm to reduce further the noise in the way to isolate only the main object. Finally, it applies the dimension reduction method to obtain a 1-D feature, in particular, the authors choose average pooling, max pooling, Vector of Locally Aggregated Descriptors (VLAD) or Fisher Vector.

We modified the mask step using one of the following masks:

- **SIFT-mask** specifically, let

$$S = (x^i, y^i)_{i=1}^n$$

the SIFT feature locations extracted from an image with the size of $W_1 \times H_1$, each location on the spatial grid $W \times H$ is the location of a local deep convolutional feature. Based on the property that convolutional layers preserve the spatial information of the input image [49], we select a subset of locations on the spatial grid which correspond to locations of SIFT key-points. In this way, we discard features from background objects, and we consider the foreground ones.

$$M_{\text{SIFT}} = \left\{ x^{(i)}_{\text{SIFT}}, y^{(i)}_{\text{SIFT}} \right\}$$
where

\[ x^{(i)}_{\text{SIFT}} = \text{round} \left( \frac{x^i W}{W_1} \right) \quad \text{and} \quad y^{(i)}_{\text{SIFT}} = \text{round} \left( \frac{y^i H}{H_1} \right) \]

- **Max-mask** we define it as:

\[ M_{\text{MAX}} = \left\{ \left( x^{(k)}_{\text{MAX}}, y^{(k)}_{\text{MAX}} \right) \right\} \quad k = 1, \ldots, K \]

\[ \left( x^{(k)}_{\text{MAX}}, y^{(k)}_{\text{MAX}} \right) = \arg \max \left( x, y \right) F_{(x,y)}^k \]

We select a subset of local convolutional features which contain high activations values for all visual contents. The goal is to select the local features that capture the most prominent object structures in the input images. Specifically, we assess each feature map and select the location corresponding to the max activation value on that feature map.

- **SUM-mask** based on the idea that a local convolutional feature is more informative if it has high values in more feature maps. Therefore, the sum of values of this kind of local feature is higher. In other words, if many channels are activated in the same image region, there is a high probability that an object of interest is in that region. We define SUM-Mask as follows:

\[ M_{\text{SUM}} = \left\{ \left( x, y \right) \mid \sum_{(x,y)} F \geq \alpha \right\} \quad (2) \]

where

\[ \sum_{(x,y)} F = \sum_{k=1}^{K} F_{(x,y)}^k \]

\[ \alpha = \text{median} \left( \sum_{(x,y)} F \right) \quad \text{or} \quad \alpha = \text{average} \left( \sum_{(x,y)} F \right) \]

In the evaluation section, we report results considering both values of \( \alpha \).

Similar to traditional local descriptors, an aggregation method is needed. In this paper, we use traditional aggregation methods for convolutional neural networks. Global pooling layers [51] can be used to reduce the dimensionality of the feature maps output. In our work, we have considered and evaluated Global max-pooling, Global average-pooling and Sum-pooling. After the use of one of the aggregation methods, we have a compact descriptor for each image. These descriptors will then be used with an image similarity measure. Image similarity computation is equivalent to solve a classification problem. We define \( M \) labels associated with images of our knowledge base. We measure the similarity using the cosine distance between vectors of descriptors representing images. The cosine similarity is defined as:

\[ \cos \theta = \frac{A \cdot B}{\| A \| \cdot \| B \|} = \frac{\sum_{i=1}^{n} A_i \cdot B_i}{\sqrt{\sum_{i=1}^{n} A_i^2 \cdot B_i^2}} \]

In our vision, the concept association process is composed of the concept validation and concept identification steps. We use an inter-class similarity for concept validation and an intra-class similarity for concept identification. So we use global descriptors from the last
convolutional layer for hyperonym identification and high-level semantic category validation by local descriptors selected by the mask for hyponym identification.

### 3.2 Ontology alignment

We populate ImageNet Ontology with images retrieved from the Google Search Images engine and the COCO data set. We realize an ontology alignment with COCO finding correspondences of its category with ontological concepts. In other words, an ontology alignment operation needs a measure of similarity between data sources that we are considering. Therefore, the purpose of this step is to find semantic links between COCO images and ImageNet concepts. In the literature there are different techniques to perform alignments [52–54]. They are based on the use of weights or/and thresholds. Some of them use the external resource as a thesaurus. We combine terminological and structural techniques to reach our goal.

According to [55], terminological techniques have as main aim to discover a similarity between terms related to a concept. These methods are based on the comparison of terms (i.e. strings of text). They can be divided into two sub-categories: (i) Intrinsic techniques based on similarity between terms that have morphological and syntactical variations (e.g. Porter Stemming Algorithm); (ii) Extrinsic techniques where external linguistic resources, such as dictionaries and thesaurus, are used in order to find a similarity between lexical variations of the same term. External techniques consider that there is a relationship equivalence between synonyms and a subsuming relationship between hyponyms.

On the other hand, structural techniques consider the similarity between two entities by exploiting structural information. The focus is on semantic or syntactic links, forming a hierarchy or a graph of entities. It compares the internal structure of the entities themselves or the shared relationships. Structural techniques are based on: (i) Internal structure, which compares internal characteristics of the entities, such as cardinality, transitivity, attributes and relationships; (ii) External structure are based on the similarity between entities by considering their position in the respective ontologies. They are based on the assumption that if two entities are similar, they have a similarity with their adjacent entities. These techniques tend to treat ontologies as graphs in which each node is a concept in the ontology, and each edge is a relationship between concepts.

We are now in the position to describe how we achieve alignments through the techniques described above. We point out that the COCO dataset has more objects of different sizes in the same image belong to different high-level categories. To address this issues, we consider an object as relevant if its surface is larger than 5% of entire image [56]. We use some information in the COCO .json file to have a semantic characterization and a more efficient use of its content for our purposes. The used information are listed in Table 1.

We retrieve categories related to images and their captions. If there are not categories, we use Natural Language Processing techniques (NLP) to derive them from image captions starting with pre-processing operations as (i) Uppercase to Lowercase conversion; (ii) Numerical characters elimination; (iii) Punctuational deletion; (iv) Stop words removal. After these operations, we obtain a list of terms, and we proceed with a Part of Speech Tagging (POSt) task. It may be defined as the process of assigning one of the parts of speech to a given term. We take into account only nouns because they are the preferable form to represent concepts [57]. After this step, we calculate the frequency of occurrence of each word and consider the most frequent couple of words. The mapping is obtained with:
• Intrinsic terminological techniques: for each word, we perform a stemming operation.
• Extrinsic terminological techniques: we use WordNet as a thesaurus to keep track of lexical variations of the same term.
• Structural techniques: we use hyponym relation.

Once the synsets associated with words have been obtained, we recognize the 'lowest common hypernym'. If it does not exist, we consider it the most frequent word to recognize the related synset. The recognized synset represents the image category. If an image has only one category, we apply terminological techniques, and we get the related synset through WordNet. In this case, we distinguish between two cases:

1. Images with one category but more instances of the same category: we crop the image exploiting the coordinates of the bounding boxes in the coco annotations file. In this way, we delete noisy background elements.
2. Images with one category and only one instance of the same category: the image is entirely considered.

If images have two categories and only one instance for each category, we make a crop of each object in the image. For images with more than two categories, our goal is to consider only foreground objects to recognize the relevant ones for the ImageNet population. We are interested in images or part of them that contains the entire object and in according to the previous assumption, and we define a threshold: \( \text{min}_\text{area} = \frac{1}{4} \text{TotalArea} \).

This threshold has been set by experiments and its value has a practical interpretation considering that an object is recognized as relevant if it fills a large part of the analyzed image compared to background and other objects in it.

We carry out the relative crop for objects that satisfy this property, and we associate a synset as previously explained. For images with objects that do not satisfy this property, we apply the above described NLP techniques but considering the categories associated with images. We measure the occurrence frequency of each category, and we consider the two most frequent words. If each category occurs only once, we compute semantic similarity between the most frequent couple of words using the measure presented in [58]. When the most semantic similar couple of words has been obtained, we get the lowest common hypernym to have a mapping between ImageNet synset and COCO categories, and we make a crop as in the previous cases. The whole ontology mapping process is summarized in Fig. 5.

4 Evaluation strategy

Our work aims to the population of an existing ontology represented by a multimedia knowledge graph using NoSql technologies. Therefore, it is essential to test and validate the
proposed approach and the implemented system. We present and discuss the strategy used to evaluate features extraction techniques and the accuracy of the ontology population task accuracy. The DNN used for the feature extraction is VGG16 [50] due to its performance.

We evaluated the system considering four distinct datasets: two with general informative contents and two fine-grain. We consider a general dataset composed of images belonging to different semantic categories, while with fine-grain, we refer to datasets having images belonging to specializations of the same semantic categories. Table 2 summarizes the datasets statistics used for top-level evaluation and, in Table 3 are shown the datasets statistics used for fine-grained evaluation. In detail, for general datasets, we reported the number of images, the number of images for each category and how we split the dataset for test and train steps. Instead, we report the number of categories, subcategories, images and used train/test images for the fine-grained dataset.

Table 2  General data sets statistics

| Dataset   | NTot img | NObj cat | NImg cat | NImg test | NImg train |
|-----------|----------|----------|----------|-----------|------------|
| Corel-10  | 10,000   | 10       | 100      | 1000      | 9000       |
| Caltech-101 | 8677    | 101      | 40–800   | 404       | 8273       |

Table 3  Fine-grained data sets statistics

| Dataset     | NCoarse cat | NSpecies cat | NTot img | NImg test | NImg train |
|-------------|-------------|--------------|----------|-----------|------------|
| Stanford-dog| 1           | 120          | 20.580   | 481       | 20.099     |
| Oxford-dog  | 2           | 25/12        | 10.000   | 148       | 9852       |
Table 4 Root node mean average precision (mAP)

| Pooling   | General dataset | Fine-grained dataset |
|-----------|-----------------|----------------------|
|           | Caltech-101     | Corel-10             | Avg | Stanford-dog | Oxford-pet | Avg |
| Max       | 0.621           | 0.888                | 0.7545 | 0.211 | 0.469 | 0.34 |
| Average   | 0.528           | 0.874                | 0.701 | 0.249 | 0.67 | 0.4595 |
| Avg and max | 0.52         | 0.833                | 0.6765 | 0.359 | 0.353 | 0.356 |
| Sum       | 0.211           | 0.33                 | 0.2705 | 0.299 | 0.2 | 0.2495 |

Table 5 Root node accuracy

| Pooling   | General dataset | Fine-grained dataset |
|-----------|-----------------|----------------------|
|           | Caltech-101     | Corel-10             | Avg | Stanford-dog | Oxford-pet | Avg |
| Max       | 0.888           | 0.966                | **0.927** | 0.643 | 0.744 | 0.6935 |
| Average   | 0.844           | 0.97                 | 0.907 | 0.67 | 0.764 | **0.717** |
| Avg and max | 0.856        | 0.96                 | 0.908 | 0.647 | 0.751 | 0.699 |
| Sum       | 0.332           | 0.522                | 0.427 | 0.21 | 0.212 | 0.211 |

Bold represents the maximum average value

For each dataset, the training set is the considered knowledge base, while the test set is the input query set to the system. Therefore, each test image is compared with each training image. For the input query, the predicted label is related to training images with the highest cosine similarity value.

In this paper, we choose as evaluation metrics the Mean Average Precision (mAP) and Accuracy due to its extensive use for our task of interest [59].

We consider as average precision (AP) is weighted sum of precision values: \( AP = \sum (R_n - R_{n-1}) p_n \). From this point of view, mean average precision (mAP) is the mean of AP. The accuracy is the fraction of correct predictions. We consider as correct prediction an image with the highest similarity value according to the test images: \( \text{Accuracy} = \frac{\text{Number of correct prediction}}{\text{Total number of predictions}} \).

### 4.1 Feature extraction evaluation

We evaluate each level of the feature extraction module considering all aggregation methods, and for each aggregation method, we evaluate all masks types as described in Sect. 3.1. We evaluate the root node with fine-grained datasets because we want to underline how the feature selection mechanism is essential when working with fine-grained images. For the Root Node, we extract features from the last convolutional layer of VGG16. The layer that we consider is \( \text{pool}_5 \). For each input image, we get a 3-d vector, \( 7 \times 7 \times 512 \). Each feature map has a \( 7 \times 7 \) size while there are 512 channels. We evaluate it for each aggregation method. The performance measures in terms of mAP and accuracy are, respectively, in Tables 4 and 5. The accuracy measure shows that the general dataset results are better than those on the fine-grained dataset. Furthermore, in the following, we showed as we had improved the performance in the worst case.
We show the accuracy mean values for each aggregation method. We consider general and fine-grained datasets separately. The results is shown in Fig. 6. The highest accuracy value is 0.927 with max-pooling. Therefore, the final configuration for root node is:

- pool5 layer features extraction;
- max-pooling aggregation method.

The highest accuracy value is 0.717. Starting to form these results obtained with standard methodologies, in the following of this section, we will show the accuracy improvement using our proposed approach using a features selection mechanism through mask-scheme.

We apply a feature-selection mechanism to the last convolutional layer (i.e. pool5). Tables 6 and 7 show the results obtained in terms of mAP and accuracy for the child node using pooling methods and mask types.

The average accuracy value grouped by aggregation methods for each mask type is shown in Fig. 7.

The highest accuracy value is 0.897 with SUM-mask with median value and Avg and Max pooling as aggregation method.

Figure 8 shows the performances comparisons with and without the features selection mechanism.

We have an improvement of about 20% using features selection mechanism.

### 4.2 Ontology population evaluation

We evaluate the accuracy of our proposed ontology population strategy considering the accuracy of ontology mapping through COCO and the Google Image Fetcher by query posing. The results are in Tables 8 and 9, respectively.

In Table 10 are listen the number of new ImageNet populated node.
Table 6  Child node mAP

| Pooling         | Dataset      | Mask type       | MAX | SUM(Mean) | SUM(Median) | SIFT |
|-----------------|--------------|-----------------|-----|-----------|-------------|------|
| Max             | Stanford-dog | 0.656           | 0.682| 0.68      | 0.21        |      |
|                 | Oxford-pet   | 0.469           | 0.475| 0.474     | 0.27        |      |
| Average         | Stanford-dog | 0.678           | 0.714| 0.712     | 0.134       |      |
|                 | Oxford-pet   | 0.462           | 0.49 | 0.488     | 0.28        |      |
| Max and average | Stanford-dog | 0.658           | 0.688| 0.685     | 0.103       |      |
|                 | Oxford-pet   | 0.47            | 0.48 | 0.478     | 0.04        |      |
| Sum             | Stanford-dog | 0.678           | 0.714| 0.712     | 0.28        |      |
|                 | Oxford-pet   | 0.462           | 0.494| 0.488     | 0.11        |      |

Table 7  Child node accuracy

| Pooling         | Dataset      | Mask type       | MAX | SUM (Mean) | SUM (Median) | SIFT |
|-----------------|--------------|-----------------|-----|------------|--------------|------|
| Max             | Stanford-dog | 0.85            | 0.9 | 0.925      | 0.17        |      |
|                 | Oxford-pet   | 0.844           | 0.81 | 0.837     | 0.2         |      |
| Average         | Stanford-dog | 0.864           | 0.851| 0.837     | 0.12        |      |
|                 | Oxford-pet   | 0.925           | 0.925| 0.95      | 0.1         |      |
| Max and average | Stanford-dog | 0.85            | 0.81 | 0.844     | 0.028       |      |
|                 | Oxford-pet   | 0.85            | 0.9 | 0.95      | 0.05        |      |
| Sum             | Stanford-dog | 0.925           | 0.925| 0.95      | 0.05        |      |
|                 | Oxford-pet   | 0.864           | 0.851| 0.837     | 0.002       |      |

Fig. 7  Average accuracy grouped by mask type and compared by pooling methods
Fig. 8  Accuracy comparison between the using of last layer and using feature selection mechanism

| Table 8  COCO accuracy |
|---------------------------------------------------------------|
| SynsID | N7846 | N15388 | N17222 | N19128 | N21939 | N9287968 | Avg |
| Accuracy | 0.92 | 0.999 | 0.865 | 0.917 | 0.82 | 0.56 | **0.846** |
| Bold represents the maximum average value |

| Table 9  Google image fetcher accuracy |
|---------------------------------------------------------------|
| SynsID | N/15388 | N/523513 | N/1299868 | Avg |
| Accuracy | 0.641 | 0.998 | 0.858 | **0.832** |
| Bold represents the maximum average value |

| Table 10  Populated nodes |
|---------------------------------------------------------------|
| Synset | SynsID | NNodes |
| Sport | N00523513 | 18 |
| Fungus | N12992868 | 24 |
| Animal | N00015388 | 28 |
| Plant | N00017222 | 7 |
| Artifact | N00021939 | 187 |
| Natural object | n00019128 | 7 |
| Geological formation | n09287968 | 3 |
| Person | n00007846 | 2 |
| Number of nodes populated for each considered synset |
We explicit point out that we obtain images for 276 nodes where 43 nodes are entirely unpopulated.

5 Conclusion and future works

In this paper, we proposed a multi-modal approach to populate a multimedia ontology. We combine textual and visual content using different techniques. The proposed framework is fully automated, and it is based on semantic similarity techniques to implement an ontology mapping task. Compared with the approaches in the literature, no user intervention is required, and a feature selection mechanism is implemented to solve the issue related to descriptors dimensionality. Furthermore, we introduced a hierarchically approach to validate an object candidate to populate an ontology concept achieving a better precision in population task. Moreover, the exploitation of the semantic structure of ImageNet allows a performance improvement compared with traditional techniques with an 87% of accuracy. Different applications, from classification to retrieval tasks, could be improved by a more accurately population of ImageNet [60–63].

At the best of our knowledge, there are not systems that implement an full automatic ontology population approach in the literature. For this reason, a direct comparison is not possible. On the other hand, there is a lack of public results on standard datasets and available free software codes of other systems. Starting from these considerations and highlighting that the development and evaluation of semi-automated techniques is an expensive task, we could run the risk that our implementation of such systems could not be in line with the results of original version. From our side, we explicit point out that the use of a standard dataset to test the performance of the proposed techniques allow an effective exploitation of our results for a later comparison with novel similar automatic approaches.

Furthermore, the approach presents a high degree of modularity, and it can be extended considering other data sources. We are investigating different features works related to the use of more complex DNN architecture as Siamese Network or Auto-encoder. Moreover, for fine-grained recognition, it can be helpful to the implementation of a multiscale approach to take into account small images variations. In the future work, we will try to improve the local feature extraction using better methods and other CNNs architecture.

Funding Open access funding provided by Università degli Studi di Napoli Federico II within the CRUI-CARE Agreement.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Rinaldi AM, Russo C (2020) Sharing knowledge in digital ecosystems using semantic multimedia big data. Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics) vol 12390 LNCS:109–131
2. Studer R, Benjamins VR, Fensel D (1998) Knowledge engineering: principles and methods. Data Knowl Eng 25(1–2):161–197
3. Maedche A, Staab S (2004) Ontology learning. Springer, Berlin, pp 173–190
4. Meenachi NM, Baba MS (2012) A survey on usage of ontology in different domains. Int J Appl Inf Syst 4(2):46–55
5. Rinaldi AM, Russo C, Tommasino C (2020) A knowledge-driven multimedia retrieval system based on semantics and deep features. Future Internet 12(11):183
6. Rinaldi AM, Russo C, Tommasino C (2021) Visual query posing in multimedia web document retrieval. In: 2021 IEEE 15th international conference on semantic computing (ICSC). IEEE, pp 415–420
7. Rinaldi AM, Russo C (2018) A semantic-based model to represent multimedia big data. In: Proceedings of the 10th international conference on management of digital ecosystems, pp 31–38
8. Cimiano P (2006) Ontology learning from text. In: Ontology learning and population from text: algorithms, evaluation and applications. Springer, Boston, MA, pp 19–34. https://doi.org/10.1007/978-0-387-39252-3
9. Asim MN, Wasim M, Khan MUG, Mahmood W, Abbasi HM (2018) A survey of ontology learning techniques and applications. Database 2018, pp 1–24. https://academic.oup.com/database/article/doi/10.1093/database/bay101/5116160?login=true
10. Spyropoulos CD, Paliouras G, Karkaletsis V, Kosmopoulos D, Pratikakis I, Perantonis S, Gatos B (2005) Boemie: bootstrapping ontology evolution with multimedia information extraction. In: The 2nd European workshop on the integration of knowledge, semantics and digital media technology, 2005. EWiMT 2005. (Ref. No. 2005/11099), pp 419–420
11. Millet C, Grefenstette G, Bloch I, Moëllic P-A, Hede P (2006) Automatically populating an image ontology and semantic color filtering. In: International workshop ontoimage. Citeseer, pp 34–39
12. Khalid YA, Noah S (2011) A framework for integrating dbpedia in a multi-modality ontology news image retrieval system. In: 2011 international conference on semantic technology and information retrieval. IEEE, pp 144–149
13. Russell BC, Torralba A, Murphy KP, Freeman WT (2008) Labelme: a database and web-based tool for image annotation. Int J Comput Vis 77(1–3):157–173
14. Kompatsiaris I (2018) Prophet–ontology population and semantic enrichment from linked data sources. In: Data analytics and management in data intensive domains: XIX international conference, DAM-DID/RCDL 2017, Moscow, Russia, October 10–13, 2017, Revised Selected Papers, vol 822. Springer, p 157
15. Knöll D, Atzmüller M, Rieder C, Scherer K-P (2016) Bishop-big data driven self-learning support for high-performance ontology population. In: LWDA, pp 157–164
16. Buitelaar P, Cimiano P, Racioppa S, Siegel M (2006) Ontology-based information extraction with soba. In: Proceedings of the international conference on language resources and evaluation (LREC)
17. Makki J (2017) Ontoprima: a prototype for automating ontology population. Int J Web Semant Technol: IJWESt 8:1–11. https://aircse.org/journal/ijwest/vol8.html
18. Navigli R, Velardi P (2006) Enriching a formal ontology with a thesaurus: an application in the cultural heritage domain. In: Proceedings of the 2nd workshop on ontology learning and population: bridging the gap between text and knowledge, pp 1–9
19. Häfner P, Häfner V, Wicaksono H, Ovtcharova J (2013) Semi-automated ontology population from building construction drawings. In: KEOD, pp 379–386
20. Ruiz-Martínez JM, Minarro-Giménez JA, Castellanos-Nieves D, García-Sánchez F, Valencia-Garcia R (2011) Ontology population: an application for the e-tourism domain. Int J Innov Comput Inf Control: IJICIC 7(11):6115–6134
21. Han J, Ma K-K (2002) Fuzzy color histogram and its use in color image retrieval. IEEE Trans Image Process 11(8):944–952
22. Ke Y, Sukthankar R (2004) Pca-sift: a more distinctive representation for local image descriptors. In: Proceedings of the 2004 IEEE computer society conference on computer vision and pattern recognition, 2004. CVPR 2004, vol 2. IEEE
23. Banerji S, Verma A, Liu C (2011) Novel color lbp descriptors for scene and image texture classification. In: 15th international conference on image processing, computer vision, and pattern recognition, Las Vegas, Nevada, pp 537–543
24. Zhang Y, Li S (2011) Gabor-lbp based region covariance descriptor for person re-identification. In: 2011 sixth international conference on image and graphics. IEEE, pp 368–371
25. Andrade FS, Almeida J, Pedrini H, Torres RdS (2012) Fusion of local and global descriptors for content-based image and video retrieval. In: Iberoamerican Congress on pattern recognition. Springer, pp 845–853
26. Rinaldi AM (2014) Using multimedia ontologies for automatic image annotation and classification. In: 2014 IEEE international congress on big data. IEEE, pp 242–249
27. Oquab M, Bottou L, Laptev I, Sivic J (2014) Learning and transferring mid-level image representations using convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1717–1724
28. Sharif Razavian A, Azizpour H, Sullivan J, Carlsson S (2014) Cnn features off-the-shelf: an astounding baseline for recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 806–813
29. Chatfield K, Simonany K, Vedaldi A, Zisserman A (2014) Return of the devil in the details: delving deep into convolutional nets. arXiv:1405.3531
30. Liu L, Shen C, van den Hengel A (2015) The treasure beneath convolutional layers: Cross-convolutional-layer pooling for image classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4749–4757
31. Poesio M (1995) Semantic ambiguity and perceived ambiguity. arXiv:cmp-lg/9505034
32. Moreno P, Williams CK, Nash C, Kohli P (2016) Overcoming occlusion with inverse graphics. In: European conference on computer vision. Springer, pp 170–185
33. Yu W, Yang K, Yao H, Sun X, Xu P (2017) Exploiting the complementary strengths of multi-layer cnn features for image retrieval. Neurocomputing 237:235–241
34. Hariharan B, Arbeláez P, Girshick R, Malik J (2015) Hypercolumns for object segmentation and fine-grained localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 447–456
35. Babenko A, Lempitsky V (2015) Aggregating deep convolutional features for image retrieval. arXiv:1510.07493
36. Jégou H, Douze M, Schmid C (2009) Packing bag-of-features. In: 2009 IEEE 12th international conference on computer vision. IEEE, pp 2357–2364
37. Sánchez J, Perronnin F, Mensink T, Verbeek J (2013) Image classification with the fisher vector: theory and practice. Int J Comput Vis 105(3):222–245
38. Jégou H, Douze M, Schmid C, Pérez P (2010) Aggregating local descriptors into a compact image representation. In: 2010 IEEE computer society conference on computer vision and pattern recognition. IEEE, pp 3304–3311
39. Jégou H, Zisserman A (2014) Triangulation embedding and democratic aggregation for image search. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3310–3317
40. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L (2009) Imagenet: a large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. IEEE, pp 248–255
41. Rinaldi AM (2014) A multimedia ontology model based on linguistic properties and audio-visual features. Inf Sci 277:234–246
42. Caldarola EG, Picariello A, Rinaldi AM (2016) Experiences in wordnet visualization with labeled graph databases. Commun Comput Inf Sci 631:80–99
43. Rinaldi AM, Russo C (2018) User-centered information retrieval using semantic multimedia big data. In: 2018 IEEE international conference on Big Data (Big Data). IEEE, pp 2304–2313
44. Caldarola EG, Picariello A, Rinaldi AM (2015) Big graph-based data visualization experiences: the wordnet case study. In: 2015 7th international joint conference on knowledge discovery, knowledge engineering and knowledge management (IC3K), vol 1. IEEE, pp 104–115
45. Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft coco: common objects in context. In: European conference on computer vision. Springer, pp 740–755
46. Rinaldi AM, Russo C, Madani K (2020) A semantic matching strategy for very large knowledge bases integration. Int J Inf Technol Web Eng: IJITWE 15(2):1–29
47. Madani K, Russo C, Rinaldi AM (2019) Merging large ontologies using bigdata graphdb. In: 2019 IEEE international conference on Big Data (Big Data). IEEE, pp 2383–2392
48. Razavian AS, Sullivan J, Carlsson S, Maki A (2016) Visual instance retrieval with deep convolutional networks. ITE Trans Media Technol Appl 4(3):251–258
49. Wei X-S, Luo J-H, Wu J, Zhou Z-H (2017) Selective convolutional descriptor aggregation for fine-grained image retrieval. IEEE Trans Image Process 26(6):2868–2881
50. Simonany K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556
51. Christlein V, Spranger L, Seuret M, Nicolaou A, Král P, Maier A (2019) Deep generalized max pooling. In: 2019 international conference on document analysis and recognition (ICDAR). IEEE, pp 1090–1096
52. Euzenat J, Shvaiko P et al (2007) Ontology matching, vol 18. Springer, Berlin
53. Faria C, Girardi R (2011) An information extraction process for semi-automatic ontology population. In: Soft computing models in industrial and environmental applications, 6th international conference SOCO 2011. Springer, pp 319–328
54. Etzioni O, Cafarella M, Downey D, Kok S, Popescu A-M, Shaked T, Soderland S, Weld DS, Yates A (2004) Web-scale information extraction in knowitall: (preliminary results). In: Proceedings of the 13th international conference on World Wide Web, pp 100–110
55. Granitzer M, Sabol V, Onn KW, Lukose D, Tochtermann K (2010) Ontology alignment—a survey with focus on visually supported semi-automatic techniques. Future Internet 2(3):238–258
56. Kim S-S, Son J-W, Park S-B, Park S-Y, Lee C, Wang J-H, Jang M-G, Park H-G (2008) Optima: an ontology population system. In: 3rd workshop on ontology learning and population (July 2008)
57. Rinaldi AM, Russo C (2021) Using a multimedia semantic graph for web document visualization and summarization. Multimed Tools Appl 80(3):3885–3925
58. Wu Z, Palmer M (1994) Verb semantics and lexical selection. In: 32nd annual meeting of the association for computational linguistics, pp 133–138
59. Baeza-Yates R, Ribeiro-Neto B (2011) Modern information retrieval: the concepts and technology behind search, 2nd edn. Addison-Wesley Publishing Company, USA
60. Chacko JS, Tulasi B (2018) Semantic image annotation using convolutional neural network and wordnet ontology. Int J Eng Technol 7(2.27):56–60
61. Zhang Y, Qu Y, Li C, Lei Y, Fan J (2019) Ontology-driven hierarchical sparse coding for large-scale image classification. Neurocomputing 360:209–219
62. Rinaldi AM, Russo C, Tommasino C (2021) Web document categorization using knowledge graph and semantic textual topic detection. In: International conference on computational science and its applications. Springer, pp 40–51
63. Rinaldi AM, Russo C, Tommasino C (2021) A semantic approach for document classification using deep neural networks and multimedia knowledge graph. Expert Syst Appl 169:114320

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Michela Muscetti received the M.Sc. degree in Computer Science Engineering from the Department of Electrical Engineering and Information Technologies, University of Naples Federico II, Napoli, Italy, in 2020. Her current research interests include machine learning, deep learning and ETL process developing.
Antonio Maria Rinaldi received the graduation degree in computer engineering and the Ph.D. degree in computer science and systems from the University of Napoli Federico II, Napoli, Italy, in 2005. His fields of interests are knowledge representation and retrieval, big data, multimedia database, deep learning and GIS. He is currently an Associate Professor with the Department of Electrical Engineering and Information Technology, University of Napoli Federico II, Napoli, Italy. Dr. Rinaldi is the Head of the Intelligent KNOWledge Systems Lab (IKNOS LAB) LUPT-UNINA, member of the CINI-ITEM Laboratory Board, the ARhemLab-DIETI Technical Board and MISlab-DIETI. He is an Associate Researcher at CNR IEOs “G. Salvatore” - National Research Council of Italy Institute of Experimental Endocrinology and Oncology “G. Salvatore”. He is member of several journal editorial boards, conference program and reviewer committees and he is author and co-author of numerous research papers in international journals and conference proceedings. In the last years, he has been involved in several research and industrial projects related to information retrieval, multimedia database integration, and geographic information systems.

Cristiano Russo received the PhD degree from both Università degli Studi Di Napoli Federico II and Université Paris-Est in 2020. He received the MSc degree from the Department of Electrical Engineering and Information Technologies, Università degli Studi Di Napoli Federico II, in 2017. His research interests include bioinformatics, information retrieval, ontologies, NoSQL databases, cognitive robotics, machine learning and data mining. He is currently studying the field related to gene regulatory networks at Consiglio Nazionale delle Ricerche (CNR) in Naples, Italy and he has also didactical commitments as contract lecturer with different universities. He has several publications in international conference and journals.

Cristian Tommasino received the M.Sc. degree in Computer Science Engineering from the Department of Electrical Engineering and Information Technologies, University of Naples Federico II, Napoli, Italy, in 2019. He is currently pursuing a PhD degree at the University of Naples Federico II. His current research interests include knowledge management, ontologies, information retrieval, deep learning, big data, and data mining.