A Knowledge-Driven Multimedia Retrieval System Based on Semantics and Deep Features

Antonio Maria Rinaldi *,†, Cristiano Russo †, and Cristian Tommasino †

Department of Electrical Engineering and Information Technologies, University of Naples Federico II, Via Claudio, 21, 80125 Napoli, Italy; cristiano.russo@unina.it (C.R.); cristian.tommasino@unina.it (C.T.)
* Correspondence: antoniomaria.rinaldi@unina.it; Tel.: +39-081-768-3911
† These authors contributed equally to this work.

Received: 18 September 2020; Accepted: 26 October 2020; Published: 28 October 2020

Abstract: In recent years the information user needs have been changed due to the heterogeneity of web contents which increasingly involve in multimedia contents. Although modern search engines provide visual queries, it is not easy to find systems that allow searching from a particular domain of interest and that perform such search by combining text and visual queries. Different approaches have been proposed during years and in the semantic research field many authors proposed techniques based on ontologies. On the other hand, in the context of image retrieval systems techniques based on deep learning have obtained excellent results. In this paper we presented novel approaches for image semantic retrieval and a possible combination for multimedia document analysis. Several results have been presented to show the performance of our approach compared with literature baselines.

Keywords: content base image retrieval; semantic information retrieval; deep features; multimedia document retrieval

1. Introduction

The main aim of a search engine is to satisfy user information needs [1] retrieving relevant information for the user [2]. Relevance may be divided in two main classes called objective (system-based) and subjective (user-based) relevance respectively [3–5]. The objective relevance takes into account the direct match between the topic of the retrieved document and the desired topic, according to a user query. Several studies on human relevance [6–8] show that several criteria are involved in output evaluation of information retrieval process. The subjective relevance is related to the concepts of aboutness and appropriateness of retrieved information, so it depends on the user and his/her judgment. Relevance can also been divided into five typologies [9]: an algorithmic relevance between the query and the set of retrieved information objects; a topicality-like type, associated with aboutness; cognitive relevance, related to the user information need; situational relevance, depending on the task interpretation; and motivational and effective, which is goal-oriented. Moreover, relevance has two main features—multidimensional, because different users can grasp and evaluate differently the acquired information and dynamic because their skills and knowledge about a specific information can change over time. The information must been represented to be analyzed. We can use symbols to represent information and these symbols are called signs [10]. A sign can be defined as, “something that stands for something else, to someone in some capacity” [11]. A sign cannot be limited to words, but also images, sounds, videos and more. Starting form these considerations, the design of a modern information retrieval system takes into account that information can be represented in different forms, in order to improve the efficiency and effectiveness of the whole retrieval process. The existence of different kinds of information representation increases the semantic gap between low-level features and high-level concepts. Moreover, in a general content based approach
the query is unknown to the system if it is not represented with low-level features. The main effect of the semantic gap is that a query expresses in terms of low-level features can return wrong results if the conceptual content is not given. For this reason, the representation of low-level features is a crucial task in information retrieval systems.

In the last years, the extracting of low-level features is an important topic highly investigated in literature. New techniques based on deep-learning approach have been presented form the scientific community. In particular, novel feature called deep feature are extracted as an output of deep neural networks (DNNs).

In our work we propose novel techniques to analyze and combine semantic information and visual features for multimedia web documents retrieval. Our techniques are the base of a whole framework for multimedia query posing and document analyzing. Our approach has been extensively tested and compared with well-known semantic and content-based retrieval techniques.

The paper is organized as follows—in Section 2 we present a literature overview putting in evidence the differences with our approach; Section 3 gives a top-level view of our system, while in Section 4 we discuss the used strategy for our experiments; eventually, discussion and conclusions are reported in Section 5.

2. Related Work

Retrieval of multimedia objects had a contribution by several fields of scientific communities such as artificial intelligence, pattern recognition, computational vision and deep learning. In this section, we present the topics of our work, introducing and analyzing several approaches and techniques for text and image retrieval.

A user tends to handle high-level concept representations, such as keyword and textual descriptors, to interpret images and measure their similarity [12]. The difference in knowledge representation between low-level features and user semantic knowledge is referred as semantic gap [13,14]. The authors of References [15,16] proposed an interesting survey on low-level features and high-level semantics for image analysis and retrieval. Moreover, an interesting approach is performed at the query level [17]. Authors represent a query by three levels: (i) low-level features, such as text, colour and shape; (ii) derived features with the same degree of logical inference; (iii) abstract features. In the authors’ opinion there are two ways to mitigate the semantic gap. The first consists in a visual and textual query posing performed by the user; on the other hand, the use of a multimedia Knowledge base to drive the multimedia analysis. In this paper we investigate the second approach and propose a framework based on semantic similarity to analyze the multimedia contents. In Reference [18], authors discuss semantic similarity measures and divide them into four types: (i) path-based; (ii) information content-based; (iii) features-based and (iv) hybrid methods. Path-based measures express the similarity between two concepts as a function of path length between concepts in a knowledge structure (e.g., taxonomies or semantic networks). The main idea of the information content-based measures is that each concept represent a well defined information in a knowledge structure and two or more concepts are similar if they share common information. The features-based measures are independent from knowledge structure and they derive similarity exploiting the properties of the knowledge structure itself. An hybrid measure combines the measure types described above and different relations between concepts such as is-a and part-of. In our work we use a path-based and information content-based semantic similarity measures.

The features used to represent images have a low dimension compared with the original raw data, they are generally composed by a series of numerical values and can be represented through different data formats (e.g., a vector). The authors of Reference [19] propose a comprehensive review on feature extraction for content-based image retrieval systems (CBIRs) and in Reference [20] is presented a review on SIFT and different features extracted from convolutional neural networks (CNN) for image retrieval. In general, we can divide features in global, local and deep features. The global features aim to represent an image as a whole considering, for example, colour, shape and texture.
Local features describe the key-points of an image object, generally used in object detection and recognition. Deep features are a new approach based on DNNs. In some cases, deep features are the output of the last level of the DNN or other levels with additional operators [20]. Many authors consider the second to last level of the DNN as a deep feature applying an aggregation layer based on Global Max Pooling or Global Average Pooling.

There are several descriptors presented in literature and the remainder of this section. We introduce different baselines used in our work, as discussed in the following sections. PHOG (Pyramid of Histograms of Orientation Gradients) is described in detail in [21], where the authors propose a descriptor based on HOG (Histograms of Orientation Gradients). Its goal is the representation of an image through its local shape and their spatial arrangement. A local feature is an image pattern which differs from its immediate neighbourhood. It is usually related to a change of an image property or several properties simultaneously [22]. Example of local features are SIFT [23] and SURF [24]. ORB [25] was proposed as a valid alternative to mitigate the high computational cost of SIFT and SURF; it is a fusion of FAST keypoint detection and BRIEF descriptor. In this paper, we use PHOG as a global descriptor due to its good results reported in the discussed literature. In addition, we use ORB as a local feature because it has a comparable accuracy with SIFT and SURF but it presents a fast computation [26]. In the last years, the progress of Computer Vision-based on Deep Learning achieved impressive performances in terms of accuracy and image understanding [27]. In this context, different DNNs have been proposed, and specific architectures called Convolution Neural Network presents the best results. A Convolution Neural Network (CNN/ConvNet) is a type of artificial neural feed-forward network inspired by the organization of the animal visual cortex. A CNN is organized in different layers: (i) input layer; (ii) convolution; (iii) normalisation; (iv) pooling; (v) full connection [27].

Over the years, the scientific community proposed different CNN architectures. In this section we briefly introduce the ones used in our work. VGGNet [28], developed by VGG (Visual Geometry Group) of the University of Oxford and presented at ImageNet LSVRC (Large Scale Visual Recognition Competition) in 2014 in the classification task, it is one of the CNNs more used in research works becoming one of the most cited in the literature. ResNet [29] won the first place in the ILSVRC 2015 classification competition with a top 5 error rate of 3.57%. ResNet is a residual network. Residual networks solve some learning problems when many levels are added to the convolutional network in some points because in this condition, the previous CNNs architecture the performance degrades quickly. In 2014, Google researchers introduced the Inception network [30], which ranked first in the ILSVRC 2014 competition. The issues addressed by the authors to design Inception are mainly related to the difficulty of choosing the kernel size, the creation of too deep networks which generate overfitting and reduce the computational cost. The authors present different versions of Inception, and in each version, they apply a set of optimizations to improve accuracy and decrease the computational complexity. MobileNet [31,32] is a neural network proposed by Google researchers. This neural network can be used on mobile devices or in cases where processing power is limited.

In our system, we use the CNNs reported above for feature extraction. In References [33,34] the authors propose a literature review of deep learning in Content-based Information Retrieval Systems (CIBR) and, according to them we use deep descriptors extracted from the second last layer of a CNN and apply max or average pooling.

Over the years, in literature can find many frameworks for CBIR and Information retrieval systems based on ontology and multimedia features. The authors of Reference [35] proposed an approach to support multi-modal image retrieval based on the Bayes point machine to associate words and images. In Reference [36], authors use the latent semantic indexing together with both textual and visual features to extract the underlying semantic structure of a web page. The authors of Reference [37] propose an iterative similarity propagation approach to explore the inter-relationships between web images and their textual annotations for image retrieval. In Reference [38], it is introduced a semantic combination technique to efficiently fuse text and image retrieval systems in multimedia information retrieval. In Reference [39], the authors report a description of an ontology model, the integrates
domain specific features, and processing algorithms focused on the domain specified by the user. YaSemIR [40] is a free and open-source semantic IR semantic system based on Lucene, which uses conceptual labels to annotate documents and questions. In Reference [41], authors discuss and present a study about different multimedia retrieval techniques based on ontologies in the semantic web. They compare these techniques to highlight the advantages of text, image, video and audio-based retrieval systems. The authors of Reference [42] propose a recommendation system for e-business applications. The recommendation strategy aimed is based on a hybrid approach, combining intrinsic characteristics of objects, past behaviours of users in terms of usage patterns and user interest expressed by ontologies to computes customized recommendations. In Reference [37], the authors present two methods to improve the Image recovery system performance. The first proposed method defined the most efficient way to GLCM texture for the recovery process. It also increased the recovery precision, combining the most efficient GLCM structure with DWT decomposition. The second proposal combined colour and texture characteristics to improve the method recovery services. This method combined the HSV colour with the most efficient GLCM texture features and with the GLCM and DWT texture features.

The authors of Reference [43] propose an efficient “bag-of-words” model that uses deep local descriptors of the convolutional neural network. The selection of high-quality descriptors provides a simple and effective way to choose the most discriminating local descriptors, which significantly improves the accuracy of retrieval. They evaluate Different methods of pre-processing the descriptors, and they report that the RootCFM is to be the best. The model uses a large visual codebook combined with the inverted index for efficient storage and fast recovery. The authors of Reference [44] show a Semantic Event Retrieval System, that includes high-level concepts and uses concept selection based on semantic embeddings. In Reference [45], the authors propose a new multimedia embedding for few-example event recognition and translation.

In this article, we propose a novel framework for multimedia web document retrieval system combining semantic similarity measures based on a formal and semantic multimedia knowledge base and different image descriptors. In particular, deep descriptors have been computed using more performance aggregation functions, reducing their dimension and improving the accuracy of the result considerably.

3. The Proposed System

In this section, we present the architecture of the implemented system and describe in detail the modules. The Figure 1 shows the system at a glance.

The system consists of two main subsystems. The first one has in charge the creation and population of the database from a document collection; moreover, it extracts images and their descriptors and normalizing the text. The second one implements the retrieval process, using semantic measure and image retrieval techniques. Multimedia Web Documents Repository Processor creates the document collection used in the retrieval tasks. The Figure 2 shows an activity diagram of this subsystem. Web documents are processed to obtain a structured object, then is stored in a NoSQL document-based database. In particular, the subsystem extracts image and text from the document normalizing the text, while the images are processed to extract features. The result of the this tasks is then merged into a single structured object if the document has at least one image and the text is composed of at least 25 terms. We remove documents without images because we want evaluate a combination of textual and visual descriptors.

The Text normalization is a fundamental step in an information retrieval systems because it allows to obtain documents with only clean text without stopwords and other kind of “noise” as for example misspelling terms [1]. Moreover, we use stemming and lemmatization tasks to have each word in its basic form. As shown in the Figure 3 the Tags Removal module removes the HTML code from the text. The second module transforms the text into lowercase text and removes special chars (i.e., all chars except number, letter, _ and ’). Stopwords Removal processes the normalized
text removing words without any particular meaning (e.g., articles, adverbs, conjunctions, . . .).

*Stemming-Lemmatization* module analyzes the output of the previous blocks and performs text transformations. Stemming transforms words into their canonical form to achieve more effective matching. The problem in using stemming is that words with different meanings can be associated with the same root. To solve this issue, we use a more sophisticated algorithm which uses lemmas instead of the root. The use of lemmas is more efficient because it represents the canonical form of the word (e.g., for verbs it uses infinitive form). The last step before document storing is performed by the *Out of vocabulary words removal* module. It removes all words that do not have sense in English language (e.g., misspelling words).

The semantic similarity module is in charge of compute the similarity on the textual document information. In our work we use different kinds of similarity metrics as described in the previous section.

---

**Figure 1.** Architecture: Proposed Framework.

**Figure 2.** Activity diagram: Multimedia web Documents Repository Processor.
The semantic measures based on path are:

- **Shortest Path based Measure**: this measure only takes into account \( \text{len}(c_1, c_2) \). It assumes that the \( \text{sim}(c_1, c_2) \) depend on closeness of two concepts are in the taxonomy, and that a similarity between two terms is proportional to the number of edges between them.

\[
\text{sim}_{\text{path}}(c_1, c_2) = 2 \cdot \text{deep}_{\text{max}} - \text{len}(c_1, c_2). \tag{1}
\]

- **Wu & Palmer’s Measure**: it introduced a scaled measure. This similarity measure considers the position of concepts \( c_1 \) and \( c_2 \) in the taxonomy relatively to the position of the most specific common concept \( \text{ls}(c_1, c_2) \). It assumes that the similarity between two concepts is the function of path length and depth in path-based measures.

\[
\text{sim}_{\text{WP}}(c_1, c_2) = \frac{2 \cdot \text{depth}(\text{ls}(c_1, c_2))}{\text{len}(c_1, c_2) + 2 \cdot \text{depth}(\text{ls}(c_1, c_2))}. \tag{2}
\]

- **Leakcock & Chodorow’s Measure**: it uses the maximum depth of taxonomy from the considered terms.

\[
\text{sim}_{\text{LC}}(c_1, c_2) = -\log \frac{\text{len}(c_1, c_2)}{2 \cdot \text{deep}_{\text{max}}}. \tag{3}
\]

Semantic measures based on information content (IC) assume that each concept includes much information in a knowledge based (i.e., WordNet [46]). Similarity measures are based on the information content of each concept. The measures are:

- **Resnik’s Measure**: it assumes that for two given concepts, similarity is depended on the information content that subsumes them in the taxonomy

\[
\text{sim}_{\text{resnik}}(c_1, c_2) = -\log p(\text{ls}(c_1, c_2)) = \text{IC}(\text{ls}(c_1, c_2)). \tag{4}
\]

- **Jiang’s Measure**: Jiang’s Measure uses both the amount of information needed to state the shared information between the two concepts and the information needed to fully describe these terms. The value is a semantic distance between two concepts. Semantic similarity is the opposite of the semantic distance.

\[
\text{sim}_{\text{jiang}}(c_1, c_2) = \frac{2 \cdot \text{IC}(\text{ls}(c_1, c_2))}{\text{IC}(c_1) + \text{IC}(c_2)}. \tag{5}
\]

Image Extractor module fetches the images in each HTML document. The retrieved images are analyzed and filtered by size and format; in our case, we consider only JPEG images but we can easily consider other format. The same module extracts visual features from these images. In our framework, we consider three types of features—local, global and deep. We use ORB as local feature
due to its performance compared with other similar descriptors \cite{26}. The same consideration is made regrading the global descriptor PHOG \cite{47}. We consider four deep descriptors derived from CNNs—VGG16, ResNet50, InceptionV3, MobileNetV2 due to the novelty in the use of this kind of features. For each CNNs we have implemented a global average pooling and a global max pooling. In this work the CNNs are pre-trained on ImageNet. Multimedia Web Document Composer module builds the documents to store in our NoSQL database (i.e., MongoDB). In particular, we use a json document consists of the following fields:

- Name: web site name;
- Url: web site url;
- Images: array like structure, each element is a json like structure (see Figure 4):
  - Image name;
  - Image path: image storage path;
  - PHOG: PHOG feature;
  - ORB: 2-D array of ORB key point;
  - Deep Descriptor: array of deep descriptors;

```json
    _id: ObjectId("5da2e859a96f1812d0b1847")
    name: "113600"
    url: "http://www.cottagecraft.co.uk/
    topic: "Top/Sports/Equestrian/Equipment_and_Apparel"
    cleaned_text: "javascript seems disabled browser for best experience site sure turn j..."
    title: "Cottage Craft"
    summary: "Manufacturers of girths, headcollars, numnahs, and pads. Site offers p..."
    images: Array
      0: Object
        path: "/113600_1.jpg"
        size: 617880
        orb: Array
        deep_desc: Array
      1: Object
        path: "/113600_0.jpg"
        size: 617880
        orb: Array
        deep_desc: Array
      2: Object
        path: "/113600_7.jpg"
        size: 242248
        orb: Array
        deep_desc: Array
```

Figure 4. Document Example.

The Retrieval subsystem has been configured in three different search cases. The first case (Case A) is the text only query. In Figure 5 the activity diagram is shown. The system normalizes the text of the query posed by the user and considers the domain to extract the concept from the knowledge Base. It computes the semantic similarity between the concept and each term of the documents. In a previous step, we apply a word sense disambiguation algorithm, and for the assignment of the score to the single documents the Equation (6) is used.

\[
    \text{sim}_{doc}(tq, d) = \frac{\sum_{i=1}^{N} \text{sim}(tq, d_i)}{N}, \tag{6}
\]

where \(tq\) is a concept, \(d\) is a document, \(N\) is the number of tokens in the document and \(d_i\) is a document term.
The second case (Case B) is the visual query. In Figure 6 the activity diagram of this step is shown. The system extracts the feature from the image query and computes its similarity with regards to the document collection images. Semantic similarity score between query image and multimedia document is obtained as average of cosine similarity between query and each image contained in multimedia documents, as expressed in the Equation (7), and ranks the documents based on the results obtained.

\[
\text{sim}(v_q, d) = \frac{\sum_{i=1}^{N} \cos(v_q, d_i)}{N},
\]

where \(v_q\) is a descriptor of the query image, \(d\) represents all visual descriptors extracted from an image in the document, \(N\) is the number of images in the document and \(d_i\) is a visual descriptor of a single image.

The third case (Case C) is the combined visual and text query and the Figure 7 shows the activity diagram. In this case, the system performs an union of the two cases illustrated above. Then it adds them up to combine the results.

In this paper the knowledge base is implemented following a multimedia model proposed in References [48,49]. This formal representation uses signs as defined in Reference [11]. A concept can be represented in various multimedia forms. The structure of the model is composed of a triple \(<S, P, C>\), defined as:

- \(S\): the set of signs;
- \(P\): the set of properties used to relate signs to concepts;
- \(C\): the set of constraints on the set \(P\).
The knowledge base is an ontology logically represented by a semantic network (SN). We can see a SN as a graph where the nodes are the concepts and the arcs the semantic relations between them. The language chosen to describe this model is the DL version of the Web Ontology Language (OWL) [50], a standard language in the semantic web and we use a NoSql technology to implement a SN, in particular Neo4j graph DB. The hierarchies used to represent the objects of interest in our model are shown in Figure 8.

The knowledge base has been populated using ImageNet [51]. Imagenet is based on WordNet adding multimedia contents to a portion of it. The multimedia nodes added to the graph contain only the meta-data and the raw image data is stored in the document-based data structure previously described. The split of image metadata from its raw content is performed to improve the global performance of our knowledge base due to the problems related to graph db to store a manage large documents. Therefore we design and implement an hybrid technology solution based on document based and graph db. We explicit point out that this is a novelty of our work. Figure 9 shows a sketch of our knowledge base.
The **Word Sense Disambiguation (WSD)** is a fundamental process in semantic similarity computation because it associates each lemma of the document with the correct concept (i.e., the right sense). Our implementation uses the Lesk algorithm [52]. The assumption at the base of this algorithm is that words have neighborhoods which tend to share a topic. A simplified version of the algorithm has been proposed in Reference [53].

![Knowledge Base](image)

**Figure 9.** Knowledge Base.

The **Image Similarity Module** works with two models: in case of features extracted with PHOG or with a CNN, we use the cosine distance, while in case of ORB, we use the best matcher. This difference is due to the dimensionality of the descriptors. We apply **global max pooling** or **global average pooling** to obtain a feature expressed as a mono-dimensional array [54] to use the cosine similarity with features extracted from the CNNs. In the case of ORB, we use the best match as suggested in Reference [55].

The main aim is to improve the performance of the system ensuring the best accuracy of results and the best precision. Several studies have shown that the use of different combined techniques could achieve better results because one technique could compensate the lacks of the others. In Figure 10 is shown a combining process. In this work, we use a the sum as combining function [56].

![Activity diagram: sum combiner](image)

**Figure 10.** Activity diagram: sum combiner.
4. Experimental Results

In this section, we present the used document collections, the testing strategy and the obtained results.

We used three datasets—20 Newsgroups [57], PASCAL VOC2012 [58] and DMOZ [59]. The specific characteristic in terms of document features, contents and size allow us to evaluate in deep all the component of our framework. In particular, we use the first dataset to evaluate and select retrieval metrics based on semantic similarity, the second one to evaluate and select the deep features in CBIR context and the last one to perform the final tests in which we combine the different strategies.

20 Newsgroups is one of the most used document collections in literature. It is available in scikit-learn and managed by Python language. This version allows us to get the dataset by removing headers, footers and quotes. The 20 newsgroups collection contains twenty topics, as shown in Table 1.

| Topic                           | Documents |
|---------------------------------|-----------|
| alt.atheism                     | 799       |
| comp.graphics                   | 973       |
| comp.os.ms-windows.misc         | 985       |
| comp.sys.ibm.pc.hardware        | 982       |
| comp.sys.mac.hardware           | 961       |
| comp.windows.x                  | 980       |
| misc.forsale                    | 972       |
| rec.autos                       | 990       |
| rec.motorcycles                 | 994       |
| rec.sport.baseball              | 994       |
| rec.sport.hockey                | 999       |
| sci.crypt                       | 991       |
| sci.electronics                 | 981       |
| sci.med                         | 990       |
| sci.space                       | 987       |
| soc.religion.christian          | 997       |
| talk.politics.guns              | 910       |
| talk.politics.mideast           | 940       |
| talk.politics.misc              | 775       |
| talk.religion.misc              | 628       |
| **Total**                       | **18,828**|

The document collection has been pre-processed by the normalization module. The query set used for the evaluation of semantic similarity metrics is composed of ten queries, for each one has been recognized the set of relevant categories of 20Newsgroups. For example for the query “mars” as planet the corresponding category is “sci.space”.

The PASCAL VOC2017 is an image database composed by 20 objects as shown in Table 2. The query set used for the evaluation of the visual descriptors is composed of thirteen classes each consisting of five images. The classes have a direct correspondence with the classes pre-assigned in the dataset.

DMOZ has been one of the most popular and rich multilingual open-source web directories. The project initially called ODP—Open Directory Project was born in 1998. The purpose of DMOZ was to collect and index URLs to create a directory of hierarchically organized web contents. From the DMOZ dump both image and text has been extracted as described in Section 3. In Tables 3 and 4 are reported the statistics downstream of this process.

We use as evaluation metrics Precision-Recall curve and Mean Average Precision (MAP) [1].
The Precision is expressed as:

\[
\text{Precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}
\]  

(8)

Table 2. Number of images for each category VOC2017 dataset.

| Object       | Images |
|--------------|--------|
| Aeroplane    | 1340   |
| Bicycle      | 1104   |
| Bird         | 1530   |
| Boat         | 1016   |
| Bottle       | 1412   |
| Bus          | 842    |
| Car          | 2322   |
| Cat          | 2160   |
| Chair        | 2238   |
| Cow          | 606    |
| Diningtable  | 1176   |
| Dog          | 2572   |
| Horse        | 964    |
| Motorbike    | 1052   |
| Person       | 8174   |
| Pottedplant  | 1054   |
| Sheep        | 650    |
| Sofa         | 1014   |
| Train        | 1088   |
| Tvmonitor    | 1150   |
| **Total**    | **23,080** |

Table 3. Number of multimedia web document for each category in DMOZ collection.

| Top Category | Num. |
|--------------|------|
| Sports       | 1087 |
| Society      | 824  |
| Computers    | 758  |
| Shopping     | 1537 |
| Arts         | 785  |
| Business     | 1895 |
| Health       | 1025 |
| Games        | 385  |
| News         | 459  |
| Science      | 509  |
| **Total**    | **9264** |
Table 4. Number of multimedia web document for each level of DMOZ collection.

| Level | Num. |
|-------|------|
| I     | 2    |
| II    | 54   |
| III   | 1023 |
| IV    | 2353 |
| V     | 2485 |
| VI    | 1699 |
| VII   | 918  |
| VIII  | 641  |
| XI    | 81   |
| X     | 8    |
| Total | 9264 |

It is the percentage of retrieved relevant documents compared with all retrieved documents. Recall is expressed as:

\[
Recall = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}. \tag{9}
\]

It is the percentage of relevant retrieved documents compared with all relevant documents in the document collection.

The Precision-Recall curve is obtained as an interpolation of the Precision values for 11 standard Recall values ranging from 0 to 1 with step 0.1. Interpolation is estimated with the following criteria:

\[
P_{\text{interp}}(r) = \max_{r_i \geq r} p(r_i). \tag{10}
\]

We use The Precision-Recall curve to compare different retrieval algorithms considering the whole set of retrieved documents. Moreover, one of the most used measures in literature to measure the web information retrieval performances is the Mean Average Precision on the first top k results. We choose MAP, because in a web search engine the user is generally interested in the first k results. Mean Average Precision is expressed as:

\[
AveP = \frac{\sum_{k=1}^{Q} (P(k) \cdot \text{rel}(k))}{\text{number of relevant documents}} \tag{11a}
\]

\[
MAP = \frac{\sum_{q=1}^{Q} AveP(q)}{Q}. \tag{11b}
\]

In this work, we consider the MAP@10 because it is mostly used on web document retrieval considering the first 10 results.

As previously stated, the measure of semantic similarity has been performed considering: path similarity, Leacock-Chodorow Similarity, Wu-Palmer Similarity, Resnik Similarity and Jiang-Conrath Similarity. We use the Lesk algorithm as disambiguation technique. The query set used consists of twelve text queries, which have polysemic meaning. An example of a polysemic query is “mars” which in WordNet has two meanings, “Mars” as god and “Mars” as a planet.

In this context we are interested in text analysis and the used document collection is 20 newsgroups. In the Figure 11 there is shown the Precision-Recall curve, where we can observe that Jiang-Conrath Similarity is the best measure, but in Figure 12 the path similarity where the best measure with respect to MAP@10.
According to results shown in Figures 11 and 12 we choice path similarity and Resnik similarity, the first one is path-based type. We consider two measures on semantic textual similarity due to the very small difference obtained obtained by experiments. In this way we can investigate more deeply in the use of a combination of these measures and multimedia descriptors.

The used visual descriptors are: PHOG, ORB and deep descriptors. In particular, the deep descriptors are:
- VGG-16 with global max pooling and global average pooling;
- ResNet-50 with global max pooling and global average pooling;
- Inception V3 with global max pooling and global average pooling;
- MobileNet V2 with global max pooling and global average pooling.

The used query set consists in sixty-five images, divided into ten classes. The used dataset is PASCAL VOC2012.
Figure 13 shows the Precision-Recall curve, where we can see that the descriptor extracted from ResNet-50 with global average pooling has the best result. Moreover, in the Figure 14 the MobileNet V2 with global average pooling is the best considering MAP@10.

The chosen descriptor, according to Figures 13 and 14, is the one extracted with MobileNet V2 with global average pooling. In this case we consider only one descriptor due to the high difference in accuracy obtained by the results comparison.

The evaluation strategy has been defined by different test cases performed on DMOZ data set to have a complete analysis of our framework in a real scenario. The test cases are:

- **Case A**: text query:
  - Path similarity;
  - Resnik similarity;
- **Case B**: visual query, using deep descriptor obtained with MobileNet V2 with global average pooling;
- **Case C**: visual and text query:
  - Path similarity and deep descriptor;
  - Resnik similarity and deep descriptor;

![Figure 13. Precision-Recall Curve for Image Descriptors.](image-url)
In the Figures 15 and 16 are reported in all experiments results on the DMOZ collection. Considering the results the best measure is the combination of the path similarity with the deep descriptor. The experiments show that on a real and general web document collection as DMOZ the path similarity is better than Resnick. We argue that the previous results depend on the specific used data sets. Moreover, the combination with the deep descriptor improves the results in both case.

**Figure 14.** MAP@10 for Image Descriptors.

**Figure 15.** Precision-Recall Curve DMOZ.
5. Conclusions and Future Work

In this paper, we proposed a novel and complete framework for multimedia information retrieval using semantic retrieval techniques combined with content-based image descriptors. The paper presents several novelties. The use of a formal multimedia knowledge base allows us to have excellent results improving the precision of used techniques compared to literature. The documents retrieved to the user are more relevant for the query due to the possibility of discriminating the different meanings of the words used in the query. In this context, the implementation of an automatic WSD task in the information retrieval task improve considerably the performance of whole retrieval process. We also presented a system based on an hybrid big data technology that integrates graph-based knowledge representation and document based approach. Moreover, different kinds of image descriptors have been added in our knowledge based to improve the representation of concepts. A deep testing shows very promising results in the combination of textual semantic metrics and deep image descriptors. Finally, we implement and test a visual query with automatic concept extraction which simplifies the query posing process obtaining comparable results with other test cases. We explicitly point out that the modularity of the proposed framework allows an easily extension of our system with other functionalities. Our future works will focus in the definition and implementation of a novel multimedia semantic measure considering both textual and multimedia information stored in our knowledge and different combination strategies. Moreover, we will investigate on the automatic generation of stories starting from the retrieved relevant documents to improve the serendipity for the user and the integration of specific knowledge domains [60,61]. In future works we will investigate on efficient techniques to improve the number of multimedia contents in ImageNet. We think that this task could improve our results, having a more comprehensive knowledge base. Moreover, we will design and implement a novel WSD method based on the analysis of the relationships among concepts in a document using our knowledge base. In addition, the use of structured text as HTML fields can enhance the efficiency and effectiveness of our retrieval methodology. We will improve the evaluation of our novel approach and additional methods using a larger query set and a document collection with high polysemic document categories. Our approach will be compared with other similar baselines proposed in literature in order to prove the effectiveness of our framework.
Author Contributions: Conceptualization: A.M.R.; investigation: C.R. and C.T.; methodology: A.M.R. and C.T.; software: C.R. and C.T.; supervision.: A.M.R.; validation: A.M.R.; Writing–original draft, C.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Baeza-Yates, R.; Ribeiro-Neto, B. Modern Information Retrieval: The Concepts and Technology Behind Search, 2nd ed.; Addison-Wesley Publishing Company: Boston, MA, USA, 2011.
2. Rinaldi, A.M. An ontology-driven approach for semantic information retrieval on the web. ACM Trans. Internet Technol. (TOIT) 2009, 9, 10. [CrossRef]
3. Saracevic, T. Relevance: A review of and a framework for the thinking on the notion in information science. J. Am. Soc. Inf. Sci. 1975, 26, 321–343. [CrossRef]
4. Swanson, D.R. Subjective versus objective relevance in bibliographic retrieval systems. Libr. Q. 1986, 56, 389–398. [CrossRef]
5. Harter, S.P. Psychological relevance and information science. J. Am. Soc. Inf. Sci. 1992, 43, 602–615. [CrossRef]
6. Barry, C.L. Document representations and clues to document relevance. J. Am. Soc. Inf. Sci. 1998, 49, 1293–1303. [CrossRef]
7. Park, T.K. The nature of relevance in information retrieval: An empirical study. Libr. Q. 1993, 63, 318–351.
8. Vakkari, P.; Hakala, N. Changes in relevance criteria and problem stages in task performance. J. Doc. 2000, 56, 540–562. [CrossRef]
9. Rinaldi, A.M.; Russo, C. User-centered information retrieval using semantic multimedia big data. In Proceedings of the 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 10–13 December 2018; pp. 2304–2313.
10. Smeluders, A.W.; Worringer, M.; Santini, S.; Gupta, A.; Jain, R. Content-based image retrieval at the end of the early years. IEEE Trans. Pattern Anal. Mach. Intell. 2000, 22, 1349–1380. [CrossRef]
11. Chen, Y.; Wang, J.Z.; Krovetz, R. An unsupervised learning approach to content-based image retrieval. In Proceedings of the Seventh International Symposium on Signal Processing and Its Applications, Paris, France, 4 July 2003; Volume 1, pp. 197–200.
12. Rui, Y.; Huang, T.S.; Chang, S.F. Image retrieval: Current techniques, promising directions, and open issues. J. Vis. Commun. Image Represent. 1999, 10, 39–62. [CrossRef]
13. Liu, Y.; Zhang, D.; Lu, G.; Ma, W.Y. A survey of content-based image retrieval with high-level semantics. Pattern Recognit. 2007, 40, 262–282. [CrossRef]
14. Eakins, J.; Graham, M. Content-Based Image Retrieval. 1999. Available online: http://www.leeds.ac.uk/educol/documents/0001240.htm (accessed on 2 September 2020).
15. Meng, L.; Huang, R.; Gu, J. A review of semantic similarity measures in wordnet. Int. J. Hybrid Inf. Technol. 2013, 6, 1–12.
16. Wang, S.; Han, K.; Jin, J. Review of image low-level feature extraction methods for content-based image retrieval. Sens. Rev. 2019, 39, 783–809. [CrossRef]
17. Zheng, L.; Yang, Y.; Tian, Q. SIFT meets CNN: A decade survey of instance retrieval. IEEE Trans. Pattern Anal. Mach. Intell. 2017, 40, 1224–1244. [CrossRef]
18. Bosch, A.; Zisserman, A.; Munoz, X. Representing shape with a spatial pyramid kernel. In Proceedings of the 6th ACM International Conference on Image and Video Retrieval, Amsterdam, The Netherlands, 9–11 July 2007; pp. 401–408.
22. Mikolajczyk, K.; Tuytelaars, T. Local Image Features. In Encyclopedia of Biometrics; Li, S.Z., Jain, A.K., Eds.; Springer: Boston, MA, USA, 2015; pp. 1100–1105.

23. Introduction to SIFT (Scale-Invariant Feature Transform). Available online: https://docs.opencv.org/master/da/d55/tutorial_py_sift_intro.html (accessed on 1 September 2020).

24. Introduction to SURF (Speeded-Up Robust Features). Available online: https://opencv-python-tutroals.readthedocs.io/en/latest/py_tutorials/py_feature2d/py_surf_intro/py_surf_intro.html (accessed on 1 September 2020).

25. ORB (Oriented FAST and Rotated BRIEF). Available online: https://docs.opencv.org/3.4/d1/d89/tutorial_py_orb.html (accessed on 1 September 2020).

26. Karami, E.; Prasad, S.; Shehata, M. Image matching using SIFT, SURF, BRIEF and ORB: Performance comparison for distorted images. arXiv 2017, arXiv:1710.02726.

27. Alom, M.Z.; Taha, T.M.; Yakopcic, C.; Westberg, S.; Sidike, P.; Nasrin, M.S.; Hasan, M.; Van Essen, B.C.; Awwal, A.A.; Asari, V.K. A state-of-the-art survey on deep learning theory and architectures. Electronics 2019, 8, 292. [CrossRef]

28. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv 2014, arXiv:1409.1556.

29. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.

30. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.

31. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv 2017, arXiv:1704.04861.

32. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4510–4520.

33. Wang, X.J.; Ma, W.Y.; Xue, G.R.; Li, X. Multi-model similarity propagation and its application for web image retrieval. In Proceedings of the 12th Annual ACM International Conference on Multimedia, Mountain View, CA, USA, 18–19 June 2014; pp. 157–166.

34. Leng, C.; Zhang, H.; Li, B.; Cai, G.; Pei, Z.; He, L. Local Feature Descriptor for Image Matching: A Survey. IEEE Access 2019, 7, 6424–6434. [CrossRef]

35. Zhao, R.; Grosky, W.J. Narrowing the semantic gap-improved text-based web document retrieval using visual features. IEEE Trans. Multimedia. 2002, 4, 189–200. [CrossRef]

36. Wang, X.J.; Ma, W.Y.; Xue, G.R.; Li, X. Multi-model similarity propagation and its application for web image retrieval. In Proceedings of the 12th Annual ACM International Conference on Multimedia, New York, NY, USA, 10–16 October 2004; pp. 944–951.

37. Clinchant, S.; Ah-Pine, J.; Csurka, G. Semantic combination of textual and visual information in multimedia retrieval. In Proceedings of the 1st ACM International Conference on Multimedia Retrieval, Trento, Italy, 18–20 April 2011; pp. 1–8.

38. Giordano, D.; Kavasidis, I.; Pino, C.; Spampinato, C. A semantic-based and adaptive architecture for automatic multimedia retrieval composition. In Proceedings of the 2011 9th International Workshop on Content-Based Multimedia Indexing (CBMI), Madrid, Spain, 13–15 June 2011; pp. 181–186.

39. Buscaldi, D.; Zargayouna, H. Yasemir: Yet another semantic information retrieval system. In Proceedings of the Sixth International Workshop on Exploiting Semantic Annotations in Information Retrieval, San Francisco, CA, USA, 28 October 2013; pp. 13–16.

40. Kannan, P.; Bala, P.S.; Aghila, G. A comparative study of multimedia retrieval using ontology for semantic web. In Proceedings of the IEEE-International Conference on Advances in Engineering, Science and Management (ICAESM-2012), Nagapattinam, Tamil Nadu, India, 30–31 March 2012; pp. 400–405.

41. Moscato, V.; Picariello, A.; Rinaldi, A.M. Towards a user based recommendation strategy for digital ecosystems. Knowl.-Based Syst. 2013, 37, 165–175. [CrossRef]
43. Cao, J.; Huang, Z.; Shen, H.T. Local deep descriptors in bag-of-words for image retrieval. In Proceedings of the on Thematic Workshops of ACM Multimedia, Mountain View, CA, USA, 23–27 October 2017; pp. 52–58.

44. Boer, M.H.D.; Lu, Y.J.; Zhang, H.; Schutte, K.; Ngo, C.W.; Kraaij, W. Semantic reasoning in zero example video event retrieval. *ACM Trans. Multimed. Comput. Commun. Appl. (TOMM)* 2017, 13, 1–17. [CrossRef]

45. Habibian, A.; Mensink, T.; Snoek, C.G. Videostory: A new multimedia embedding for few-example recognition and translation of events. In Proceedings of the 22nd ACM International Conference on Multimedia, Mountain View, CA, USA, 18–19 June 2014; pp. 17–26.

46. Miller, G.A. WordNet: A Lexical Database for English. *Commun. ACM* 1995, 38, 39–41.10.1145/219717.219748. [CrossRef]

47. Purificato, E.; Rinaldi, A.M. Multimedia and geographic data integration for cultural heritage information retrieval. *Multimed. Tools Appl.* 2018, 77, 27447–27469. [CrossRef]

48. Rinaldi, A. A multimedia ontology model based on linguistic properties and audio-visual features. *Inf. Sci.* 2014, 277, 234–246. [CrossRef]

49. Rinaldi, A.M.; Russo, C. A semantic-based model to represent multimedia big data. In Proceedings of the 10th International Conference on Management of Digital EcoSystems, Tokyo, Japan, 25–28 September 2018; pp. 31–38.

50. Web Ontology Language. Available online: https://www.w3.org/OWL/ (accessed on 1 September 2020).

51. ImageNet. Available online: http://www.image-net.org/ (accessed on 1 September 2020).

52. Lesk, M. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In Proceedings of the 5th Annual International Conference on Systems Documentation, Toronto, ON, Canada, 8–11 June 1986; pp. 24–26.

53. Vasilescu, F.; Langlais, P.; Lapalme, G. Evaluating Variants of the Lesk Approach for Disambiguating Words. Available online: http://www.iro.umontreal.ca/~felipe/Papers/paper-lrec-2004.pdf (accessed on 27 October 2020).

54. Tolias, G.; Sicre, R.; Jegou, H. Particular object retrieval with integral max-pooling of CNN activations. arXiv 2015, arXiv:1511.05879.

55. Rublee, E.; Rabaud, V.; Konolige, K.; Bradski, G. ORB: An efficient alternative to SIFT or SURF. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 2564–2571.

56. Kittler, J. Combining classifiers: A theoretical framework. *Pattern Anal. Appl.* 1998, 1, 18–27. [CrossRef]

57. 20 Newsgroups Scikit-Learn. Available online: https://scikit-learn.org/0.15/datasets/twenty_newsgroups.html (accessed on 1 September 2020).

58. Visual Object Classes Challenge 2012 (VOC2012). Available online: http://host.robots.ox.ac.uk/pascal/VOC/voc2012/ (accessed on 1 September 2020).

59. DMOZ Website. Available online: https://dmoz-odp.org/ (accessed on 1 September 2020).

60. Caldarola, E.; Rinaldi, A. A multi-strategy approach for ontology reuse through matching and integration techniques. *Adv. Intell. Syst. Comput.* 2018, 561, 63–90.

61. Rinaldi, A.M.; Russo, C. A matching framework for multimedia data integration using semantics and ontologies. In Proceedings of the 2018 IEEE 12th International Conference on Semantic Computing (ICSC), Laguna Hills, CA, USA, 31 January–2 February 2018; pp. 363–368.

**Publisher’s Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).