Building Multimedia Ontologies using Linguistic Properties and Low-Level Visual Descriptors

Antonio M. Rinaldi
DIETI-Dipartimento di Ingegneria Elettrica e delle Tecnologie dell'Informazione-80125 Via Claudio, 21, Napoli, Italy
IKNOS-LAB-Intelligent and Knowledge Systems-LUPT-80134 Via Toledo, 402, Napoli, Italy
Università di Napoli Federico II
antoniomaria.rinaldi@unina.it

ABSTRACT

In the era of big data, the use of formal models and techniques to represent and manage information is a necessary task to implement efficient intelligent information systems. These models and techniques are basic components of such kind of systems and they should be used to analyze data and create useful informative contents. In this paper we propose a general and formal ontology model to represent knowledge using multimedia data and linguistic properties to bridging the gap between the target semantic classes and the available low-level multimedia descriptors. The multimedia features are automatically extracted using algorithms based on MPEG-7 descriptors.

1. INTRODUCTION

Every day, gigabytes of new multimedia information is being generated, stored, and transmitted and it is difficult to access this information unless it is organized and represented in a suitable way to allow efficient browsing, searching, and retrieval. In this context, representing and manage knowledge is one most important tasks in the information management process. In the literature there are a lot of algorithms designed to describe color, shape, and texture features, but they are far to adequately model image semantics and have many limitations when dealing with broad content image databases [11]. On the other hand, humans tend to use high-level features (concepts), such as keywords and text descriptors, to interpret and analyze multimedia object. More specifically, the discrepancy between the limited descriptive power of low-level multimedia features and the richness of user semantics, is referred to as the semantic gap [15]. An interesting survey on low-level features and high-level semantics for multimedia analysis and retrieval is in [7]. However, bridging the gap between the target semantic classes and the available low-level multimedia descriptors is an unsolved problem. Hence it is crucial to select an appropriate set of multimedia descriptors and to combine them in such a way that the results obtained with individual descriptors are improved together with high level concepts annotation. New techniques have been developed to solve those problems. Some of them are based on ontologies to delete or at least smooth conceptual or terminological mess and to have a common view of the same information [1, 13, 6, 9, 16]. In this paper we propose a general and formal ontology model to represent knowledge and build a general multimedia knowledge base. We argue that the main novelties of our model are that it is independent from the particular domain of interest; that our linguistic approach provides a simple and general way to represent knowledge; that the use of low-level multimedia features enables the representation of multimedia information; that the strong formalization of our model is useful to integrate and enrich general knowledge bases; and that the use of a standard language to represent our ontologies facilitates knowledge sharing and reusing.

2. THE PROPOSED ONTOLOGY MODEL

In this section the proposed model is presented with a description of all its single components and properties. This discussion starts with some notions about ontologies and the way to build them. Starting from some definitions of ontology [4, 12] we extend them using also visual data to denote a concept; these data are represented using visual low-level features defined in MPEG-7 standard. Thus an ontology can be seen as a set of "signs" and "relations" among them, denoting the concepts that are used in a knowledge domain. The proposed model is composed of a triple $\langle \boldsymbol{S},\boldsymbol{P},\boldsymbol{C} \rangle$ where:

$\boldsymbol{S}$ is a set of signs;

$\boldsymbol{P}$ is a set of properties used to link the signs in S;

$\boldsymbol{C}$ is a set of constraints on P.

In this context signs are words and visual data. The properties are linguistic relations, and the constraints are validity rules applied to linguistic properties with respect to the multimedia category considered. In the proposed approach, knowledge is represented by an ontology implemented with respect to a semantic network (SN). A semantic network can be seen as a graph where the nodes are concepts and the arcs are relations among concepts. A concept is a set of multimedia data which represent an abstract idea. In recent years, several languages have been proposed to represent ontologies and we choose to use OWL [2] due to its expressive power useful for our purposes and its extensive use in knowledge based systems. In our approach we use the DL version of OWL, because it is sufficiently effective to describe our model and its implementation. The DL version allows the declaration of disjoint classes, which may be used to assert that a word belongs to a syntactic category. Moreover, it allows the declaration of union classes used to specify domains and property ranges used to relate concepts and words belonging to different lexical categories. The ontology schema and corresponding semantic network representation is formally described using OWL. Every node (both concept and

multimedia) is an OWL individual. The connecting edges in the semantic network are represented as *ObjectProperties*. The considered linguistic properties are shown in table 1.

Table 1: Linguistic properties

Lexical properties	synonym, antonym, pertainym, nominalization, derived from adjective, participle of verb
Semantic properties	hypernyms,hyponyms, coordinate terms, holonym, meronym, hypernym, troponym, entailment, related nouns, similar to, coordinate terms, articiple of verb, root adjectives

These properties have constraints that depend on the syntactic category (noun, verb, adjective, adverb) or kind of semantic or lexical properties. For example, the hyponymy property can only relate nouns to nouns or verbs to verbs. In contrast, a semantic property links concepts to concepts, and a syntactic property relates word forms to word forms. Concept and multimedia are considered with *DatatypeProperties*, which relate individuals to pre-defined data types. Each multimedia is related to the concept it represents by the ObjectProperty *hasConcept*, whereas a concept is related to multimedia that represent it using the ObjectProperty *hasMM*. These are the only properties that can relate words to multimedia and vice versa; all of the other properties relate multimedia to multimedia and concepts to concepts. Concepts, multimedia and properties are arranged in a class hierarchy, resulting from the syntactic category for concepts and words, data type for visual descriptors and from the semantic or lexical for the properties. From a logical point of view, a visual representation can be related to all kind of concept. The two main classes are Concepts, in which all objects are defined as individuals, and MM, which represents all the "signs" in the ontology. These classes are not supposed to have common elements; therefore they are defined as disjoint. The class MM defines the logical model of the multimedia forms used to express a concept. On the other hand, the class Concept represents the meaning related to a multimedia form; the subclasses have been derived from related categories. There are some union classes that are useful for defining the properties of domain and codomain. Attributes have been defied for Concept and MM respectively; Concept has: *Name* that represents the concept name; *Description* that gives a short description of concept. On the other hand MM has *Name* as attribute that is the MM name and a set of features described in Table 2.

Table 2: Visual features

Data Type	Features
Visual	Dominant Color, Color Structure, Color Layout, Homogeneous Texture, Edge Histogram, Region-based Shape, Contour-based Shape

All elements have an *ID* within a unique identification number. The visual features are the low-level descriptors in MPEG-7 standard. Table 3 shows some of the properties considered and their domains and ranges of definition.

The use of domain and codomain reduces the property range application; however, the model as described so far does not exhibit perfect behavior in some cases. For example, the model does not know that a hyponymy property defined on sets of nouns and verbs would have 1) a range of nouns when applied to a set of nouns and 2) a range of verbs when applied to a set of verbs. Therefore, it is necessary to define several *constraints* to express the ways that the linguistic properties are used to relate concepts and/or MM. Table 4 shows some of the defined constraints specifying the classes to which they have been applied with respect to the properties considered. The table also shows the matching range.

Table 3: Property features

Property	Domain	Range
hasMM	Concept	MM
hasConcept	MM	Concept
hypernym	NounsAnd VerbsConcept	NounsAnd VerbsConcept
holonym	NounConcept	NounConcept
entailment	VerbWord	VerbWord
similar	AdjectiveConcept	AdjectiveConcept

Table 4: Model constraints

Costraint	Class	Property	Constraint range
AllValuesFrom	NounConcept	hyponym	NounConcept
AllValuesFrom	AdjectiveConcept	attribute	NounConcept
AllValuesFrom	NounWord	synonym	NounWord
AllValuesFrom	VerbWord	also_see	VerbWord

Sometimes, the existence of a property between two or more individuals entails the existence of other properties. For example, since the concept "dog" is a hyponym of "animal", animal is a hypernym of dog. These characteristics are represented in OWL by means of property features. Table 5 shows several of those properties and their features.

Table 5: Property features

Property	Features
hasMM	*inverse* of hasConcept
hasConcept	*inverse* of hasMM
hyponym	*inverse* of hypernym; *transitivity*
hypernym	*inverse* of hyponym; *transitivity*
cause	*transitivity*
verbGroup	*symmetry* and *transitivity*

The use of a linguistic approach allows an extension of linguistic properties also to visual data; e.g. different visual information related to the same concept are synonyms and in the same way hyperonym/hyponym or meronym properties entail a semantic relation among the multimedia representation of concepts.
The proposed model allows a high-level conceptual matching using different type of low-level representations. Moreover, an ontology built using this model can be used to infer information by means of formal representation of properties among multimedia data and concepts.

3. ONTOLOGY POPULATION STRATEGY AND EVALUATION

The proposed model has been implemented in a tool for create and manage ontologies. We use WordNet [10] as general knowledge base and an appropriate algorithm to extract from it a domain ontology (i.e. a semantic network) is defined and implemented; the semantic network provides a general representation of user domain of interest. Many information systems use a knowledge base to represent data in order to satisfy information requests and in the author's vision it is a good choice for having a common view of

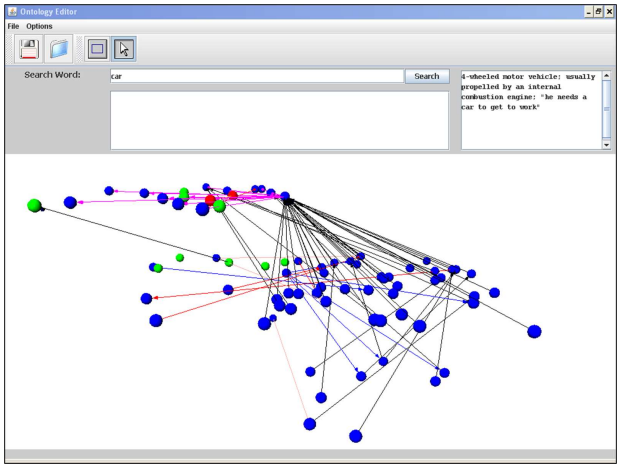

Figure 1: Editor Interface

the same general and specific knowledge domains. Moreover in the proposed framework WordNet can be a "starting point" for users because they can extract an initial general ontology from this knowledge base and expand it to have a specialized one. The semantic network is dynamically built using an ad hoc algorithm which takes into account the WordNet structure. WordNet organizes the several terms using their linguistic properties. Moreover, every domain keyword may have various meanings (senses) due the proprieties of polysemy, so a user can choose its proper sense of interest using the tool interface. Beyond the synonymy, other linguistic proprieties are considered and they are applied to the typology of the considered terms in order to have a strongly connected network. The network is built starting from a domain keyword that represents the context of interest for the user. After this step all the component synsets are considered to construct a hierarchy, only based on the hyponymy property; the last level of this hierarchy corresponds to the last level of WordNet one. Afterwards the hierarchy is enriched considering all the other kinds of relationships in WordNet (see Table 3). Based on these relations other terms are added in the hierarchy obtaining a highly connected semantic network. The algorithm to extract the semantic network from WordNet is described in pseudocode in Figure 2.

```
//------------------------------------------------------
// Semantic network extraction algorithm
//
// INPUT: Main_Synset: represents the synset chosen by user
//
// OUTPUT: Synset_List: the list returned from the function.
//         It contains all SN synsets
//------------------------------------------------------
Synset_List   CreateSN (Main_Synset)
{
  Add Main_Synset to a Synset_List
  Load from Wordnet the Category_terms of Main_Synset
  Add founded synsets to Synset_List
  While (Synset_List is not_empty)
  Do {
   Load from Wordnet all hyponyms of all synsets in Synset_List
   Add founded synsets to Synset_List
  }
  Start from head_list
  While(Synset_List is not_empty)
  Do {
   Load from Wordnet all synsets linked to all synsets in Synset_List using
    all other linguistic properties (count outhyponimy and hyperonimy)
  }
  return Synset_List
}
```

Figure 2: The semantic network extraction algorithm

At present we have implemented low-level image features extraction with the related multimedia ontology management tool. The images are fetched using an image search engine (i.e. google image) by means of a query with the synset name in WordNet. In addiction, the user can use words from WordNet synset description or other ones manually added to refine her search. Once images have been fetched, they are automatically added to the consider concept; the user can also verify manually the accuracy of fetching and manage ontologies by ad hoc interfaces. All the ontologies can be exported in OWL following a schema model described in section 2. Multimedia features are extracted in according to MPEG-7 standard descriptions defined in the proposed model and the visual descriptor values are in the OWL file together with the linguistic properties among concepts. Clearly, even if a knowledge base could be large and detailed, it will never give us a high level of specialization for every existing knowledge domains; the proposed approach tries to give a solution to this problem. In fact users can interact with the system in order to create a first ontological knowledge representation or they can expand it or create a new one. In addiction a user can associate multimedia representations to concepts. A user can modify the ontology structure as a whole adding new MM and Concepts in the network, linking MM and Concepts using arrows (lexical and semantic properties), deleting nodes and arcs. The interaction with the semantic network is archived by means of Java 3D libraries. We also present several experimental results to show the accuracy of our model and techniques in the population of multimedia ontology. The experiments are conducted on visual MPEG-7 descriptors introduced so far and implemented in well-known similarity metrics which use single descriptors, their combination or some variations with improved features; these implementations are based on [8]. In order to have a reliable evaluation of our system and methods, we use Caltech 101 [3], a dataset of digital images developed by the California Institute of Technology. To evaluate the accuracy of the ontology population process, the outputs must be compared with a *ground truth*. The ground truth is determined by humans and, in our strategy, it is the categories set of the Caltech 101 dataset. The whole Caltech 101 dataset has been used during the experiments. We highlight that some categories are different from a conceptual point of view but they can be similar using low level visual features (e.g orange, sun, mars). We calculate what we call "local precision" simply considering the precision in the ontology population using each category. In this case, recall is the same of precision, because we consider each single category. The correspondence between the ground truth and association of images with synsets has been obtained by human manual inspection. Different classes of precision have been defined for the system analysis output. The *Right Classification* class is referred to the annotations fitting with the relevance assessment given to the Caltech 101 categories; in the *Wrong Classification* class are all the images with an erroneous analysis; with the label *General Classification* we suggest a too low accuracy to satisfy the user needs but with a right beginning root path; *None Classification* is the tag for those images with none association. The similarity has been calculated using images fetched by Google search image API following the strategy previously discussed and the Caltech 101 dataset. The following graphs show the results of our experimentation strategy. In Figure 3 is shown the precision of each metrics in the multimedia ontology population process calculated on all the Caltech categories. The best metric is JCD followed closely by FCTH; both these metrics combined some color and texture descriptors. We want also measure the precision in multimedia ontology population process using a combination of all MPEG-7 descriptors types. For this reason, we choose JCD (color and texture descriptors) and MSER (shape descriptors)

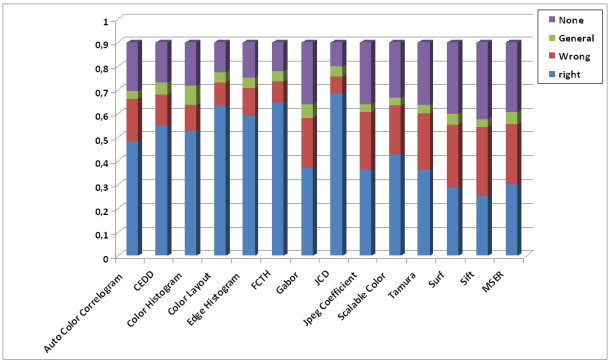

Figure 3: Multimedia Ontology Population Accuracy

combined by means of sum [5], OWA [17] and CombMNZ [14] functions. Figure 4 shows that OWA has the best performance. This combining function is based on a wighted average of the used

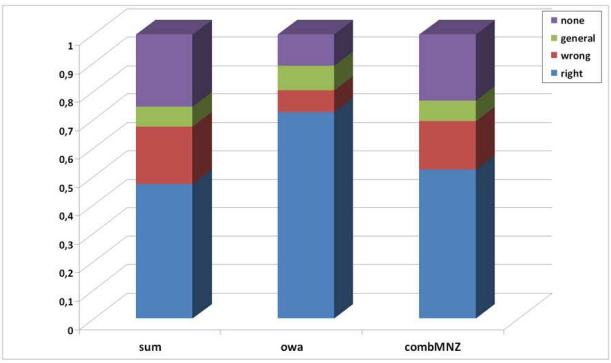

Figure 4: Combined Metrics Accuracy

classifiers. The wights are chosen by experimental results of each metric on the Caltech 101 dataset. We highlight that the overall precision using a combination of all MPEG-7 descriptors has an improvement of 10% with respect to the best single metric (JCD).

4. CONCLUSIONS AND FUTURE WORKS

The design, implementation, and reuse of existing ontologies is a non-trivial task. Moreover, the complexity of multimedia data must be taken into account to have a complete representation of knowledge expressed using different "signs". In this paper, a global approach to define and develop multimedia ontologies has been presented. Our framework is based on a simple and general formal model for multimedia knowledge representation taking into account a linguistic approach considered as the natural communication way between human agents. The ontologies are represented using OWL. The evaluation in the process of multimedia ontology creation and show the efficiency of our approach and the expressive power of our model. The current research effort is based on the use of the proposed model together with integrated multimedia similarity metrics (i.e. textual and visual information) for document content based analysis. Moreover, we are implementing an extension of our system integrating peer-to-peer functionalities and web services to share ontologies in the web. Extensive experiments will carry on to show the effectiveness and the efficiency of the proposed framework compared with similar approach.

5. REFERENCES

[1] R. Arndt, R. Troncy, S. Staab, and L. Hardman. Comm: A Core Ontology For Multimedia Annotation. In S. Staab and R. Studer, editors, *Handbook on Ontologies*. Springer Verlag, second edition, 2009.

[2] M. Dean and G. Schreiber. OWL Web Ontology Language Reference. Technical Report http://www.w3.org/TR/2004/REC-owl-ref-20040210/, W3C, February 2004.

[3] L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(4):594–611, Apr. 2006.

[4] T. R. Gruber. A translation approach to portable ontology specifications. *Knowl. Acquis.*, 5(2):199–220, 1993.

[5] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas. On combining classifiers. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(3):226–239, 98.

[6] Q. Li, Z. Lu, Y. Yu, and L. Liang. Multimedia ontology modeling: An approach based on mpeg-7. In *Advanced Computer Control (ICACC), 2011 3rd International Conference on*, pages 351 –356, 2011.

[7] Y. Liu, D. Zhang, G. Lu, and W.-Y. Ma. A survey of content-based image retrieval with high-level semantics. *Pattern Recogn.*, 40(1):262–282, 2007.

[8] M. Lux and S. A. Chatzichristofis. Lire: lucene image retrieval: an extensible java cbir library. In *Proceedings of the 16th ACM international conference on Multimedia*, MM '08, pages 1085–1088, New York, NY, USA, 2008. ACM.

[9] A. Mallik and S. Chaudhury. Acquisition of multimedia ontology: an application in preservation of cultural heritage. *International Journal of Multimedia Information Retrieval*, 1(4):249–262, 2012.

[10] G. A. Miller. Wordnet: a lexical database for english. *Commun. ACM*, 38(11):39–41, 1995.

[11] A. Mojsilovic and B. Rogowitz. Capturing image semantics with low-level descriptors. In *Image Processing, 2001. Proceedings. 2001 International Conference on*, volume 1, pages 18–21 vol.1, 2001.

[12] R. Neches, R. Fikes, T. Finin, T. Gruber, R. Patil, T. Senator, and W. R. Swartout. Enabling technology for knowledge sharing. *AI Mag.*, 12(3):36–56, 1991.

[13] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy, and N. Vasconcelos. A new approach to cross-modal multimedia retrieval. In *Proceedings of the international conference on Multimedia*, MM '10, pages 251–260, New York, NY, USA, 2010. ACM.

[14] J. A. Shaw and E. A. Fox. Combination of multiple searches. In *The Second Text REtrieval Conference (TREC-2)*, pages 243–252, 1994.

[15] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(12):1349–1380, 2000.

[16] M. Suarez-Figueroa, G. Atemezing, and O. Corcho. The landscape of multimedia ontologies in the last decade. *Multimedia Tools and Applications*, 62(2):377–399, 2013.

[17] D. Wu and J. Mendel. Ordered fuzzy weighted averages and ordered linguistic weighted averages. In *Fuzzy Systems (FUZZ), 2010 IEEE International Conference on*, pages 1–7, 2010.