Building Ontologies from Annotated Image Databases

Angelo Chianese, Vincenzo Moscato, Antonio Penta and Antonio Picariello

Department of Computer Science, University Federico II of Naples, Naples, Italy

Abstract. Defining and building Ontologies within the multimedia domain still remain a challenging task, due to the complexity of multimedia data and the related associated knowledge. In this paper, we investigate automatic construction of ontologies using the Flickr image databases, that contains images, tags, keywords and sometimes useful annotation describing both the content of an image and personal interesting information describing the scene. We then describe an example of automatic ontology construction in a specific semantic domain.

1 Introduction

In the last few years, an increasing number of multimedia data has been produced and stored in distributed repositories and data bases. Despite the vast amount of work on multimedia processing and analysis, multimedia databases and knowledge representation, there is no commonly accepted solution to the problem of how to represent, organize and manage multimedia data and the related semantics by means of a formal framework. Several formal models based on ontologies have been proposed. In particular, a great emphasis has been given to the *extensional aspects* of image, video and audio ontologies, usually containing information on rough data, format and annotation [24]: anyway, traditional domain ontologies, are substantially inadequate to support complete annotation and retrieval by content of image documents. There is still a great work to do with respect to the *intensional aspects* of multimedia ontology: starting from the very beginning, it is still not at all clear wether a multimedia ontology is simply a taxonomy, or a semantic network, what is the role of concrete data (if any) or wether it is a simple organization of metadata. In addition, the semantics of multimedia data itself is very hard to define and to capture and once defined a suitable formal framework, still remains opened the problem of how to build in an automatic way the extensional ontologies.

In this paper, we first propose a novel multimedia ontology framework, in particular related to the image domain; thus, we describe a technique for building ontologies, that operates on large corpora of human annotated repositories, namely the Flickr [16] database, integrating both low level image processing strategies and NLP techniques to keywords and annotations produced by humans when they store the produced data. In particular, the key points of the proposed technique are the following:

1. A low level image analysis based on active vision is performed in order to retrieve an image feature vector.

Chianese A., Moscato V., Penta A. and Picariello A. (2009).
Building Ontologies from Annotated Image Databases.
In *Proceedings of the 1st International Workshop on Ontology for e-Technologies OET 2009*, pages 41-51
DOI: 10.5220/0002223400410051

2. A text categorization process is applied to image tags and metadata extracted from the Flickr and associated to a given image in order to retrieve the most relevant topics
3. Rough data, features, topics and inferred semantic description are combined and stored in an appropriateMultimedia Ontology.

We provide an algorithm for creating image ontology in a specific domain gathering together all this different information. We then provide an example of automatic construction of image ontology and a discussion of the encountered problems and the provided solutions.We concluded that the framework is promising and sufficiently scalable to different domain.

2 Multimedia Ontology Requirements

If we look at the main definition about ontology we could defined it as "an explicit specification of a conceptualization"which is, in turn, "the objects, concepts, and other entities that are presumed to exist in some area of interest and the relationships that hold among them " [11]. Gruber, in the previous definition, claims that "while the terms specification and conceptualization have caused much debate, the essential points are the following:

– "an ontology defines (specifies) the concepts, relationships, and other distinctions that are relevant for modeling a domain".
– "the specification takes the form of the definitions of representational vocabulary (classes, relations, and so forth), which provide meanings for the vocabulary and formal constraints on its coherent use."

Gruber stresses the conceptual nature of the ontology as a *theory* that can be used to represent relevant notion about domain modeling. Domain that is classified in terms of concepts, relationships and constraint on them. Nothing is said about what we mean for the conceptualization of our domain. Typically when we deal with the problem of the knowledge representation, it is not clear *what* is the knowledge to represent and *how* this could be aware from any context. This is a key point because we know that the knowledge modeling is an expensive operation and we would assure that the final result could be shared. This is possible only if it is valid *for all the users*. This is one of the issue that Guarino [12] expressed in his well known ontology definition: "ontology is a logical theory accounting for the intended meaning of a formal vocabulary, i.e. its ontological commitment to a particular conceptualization of the world. The intended models of a logical language using such a vocabulary are constrained by its ontological commitment. "

We note that in both definitions the ontological domain is considered sufficiently abstract in order to manage a general knowledge; instead, if we consider multimedia domain objects, we have to deal with a number of complex issues, and it is not simple to capture and represent its related semantics.

Let us consider the image domain. Given an image I, a human decodes its knowledge content after different cognitive steps. Each step is related to a human perceptive process and some of these steps are iterated in order to derive more complex concepts. Several steps are image processing blocks that approximate the human vision process on the whole image or on parts of an image. Psychological theories propose two different approaches for recognizing concepts related to an image: the holistic and the constructive one [3]. According to the holistic theory, and image is processed and recognized by humans considering the whole image. In contrast, the constructive theory considers image understanding as a progressive process: a human first recognizes an interesting part of an image and infers the knowledge of the whole image from the knowledge of its parts, in a recursive fashion.We follow this latter approach. In addition, we also need a further environmental knowledge that describes all the necessary knowledge as evidences by the classical "meaning triangle" [23]: in a given media, we detect symbols, objects and concepts; in a certain image we have a region of pixels (symbol) related to a portion of multimedia data; this region is an instance (object) of the certain concept. In other words, we can detect concepts but we are not able to disambiguate among the instances without some specific knowledge. A simplified version of the described vision process will consider only three main levels: *Low, Medium and High*. In fact, the knowledge associated to an image is described at three different levels:

– *Low level*: raw images, computationally and abstractly thought of as discrete functions of two variables defined on some bounded and usually regular regions of a plane, i.e a pixel map used to structure the image perceptual organization and in filtering processes in order to obtain new maps;
– *Intermediate level*: an aggregation of data related to the use of spatial features - including points, lines, rectangles, regions, surfaces, volumes - color features, textures and shape features, for example colors are usually described by means of *color histograms* and several features have been proposed for texture and shapes, all exploiting spatial relations between a number of low level features (pixels) in a certain region;
– *High level*: this layer is related to axioms that involved concepts and their relations conveyed by an image; looking at Figure 1 we could use these sentences to define her high level: "*An elephant on a green savannah with a sky background*".

The features associated to these layers should be characterized in terms of a *fuzzy value*, representing a certain *degree of uncertainty* that each image processing algorithm produces, i.e we might say the shape is "highly" trapeze, or that it is "a little bit" rectangular. Expressions such as *highly*, *a little bit*, and so on, recall this notion of fuzziness implicitly related to the *similarity of visual stimuli*. We can associate this fuzziness to some regions inside the image related to colors, shapes and textures. Considering the running example image, the derived features are the following: Colors:{<Green, 0.8>, <White, 0.89> <Black, 0.7>, <Brown, 0.75>}, Shapes: {<Rectangle, 0.6>,<Trapeze, 0.9>}.Textures: {<Animal, 0.8>, <Nature, 0.6>}.

Starting from these considerations, it's the author's opinion that a multimedia ontology should take into account these specific characteristic of multimedia data and in particular of images: each image could be, in fact, decomposed into a set of regions properly characterized (for example in terms of texture, color and shape) and in addition, some of these regions can be associated to the instances of some concepts as derived from image analysis algorithms. Eventually, one could infer new kinds of concepts. More into details, an intelligent system, using a classifier, might associate some elementary concepts to the extracted multimedia feature, related to the image itself, e.g.{<person, horse, grass, sand>}.

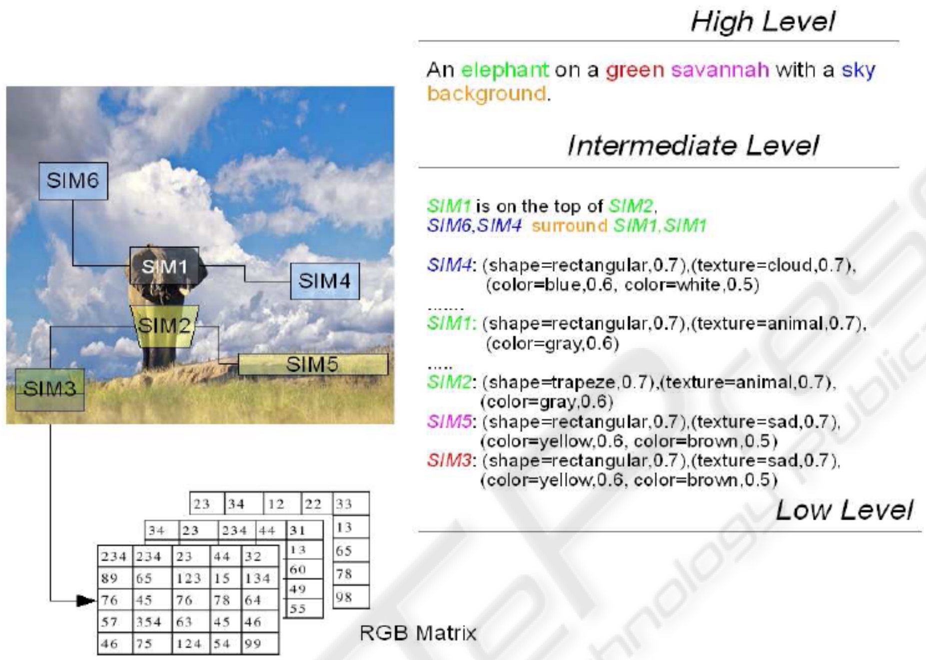


Fig. 1. The Levels Description of an image, being "SIM" a label for the SubIMage inside the given image.

We conclude that the representation requirements of image data and, more generally, of multimedia data can be improved if there is a model that is able to describemore complex concepts. In this context,we need a system that should allow specifications for:

- *Special Relationships* that exist between the different media entities characterizing a media object or an event - for example, geometrical, spatial, temporal relationships and so on;
- *Uncertainty* that is produced by Computer Vision systems when processing and recognizing multimedia contents - for example, object detection in an image or in a video is always associated to a certain membership degree;
- *Association between low-level properties and semantic properties of images* - for example, the semantics of an image can be enriched/learned/guessed by observing the relationships of its color and shape with real-world concepts; technically speaking, in order to associate a fuzzy membership among the elements of the previous concepts, a *reification pattern* should be used.
- An *associated reasoning service* which can use the available feature observations and concept description to infer other probable concepts in presence of uncertain relationships - for example, some shape, texture and color properties

might be used to infer that a certain media object is an instance of a given concept: e.g., colors=yellow with a grade μ_y, shape=circle with a grade μ_c may be associated with the concept of the sun with a grade $\min\{\mu_y, \mu_c\}$.

3 Building a Multimedia Ontology

3.1 Goals

The main aim here is to present an approach which could improve the accomplishment of a multimedia ontology building task by incorporating in the building process information derived from content features and text used to annotate the multimedia data themselves inside an annotated database.

In this paper, we propose an *image ontology* that can be seen as a particular instance of the defined multimedia ontology, taking into account only images data and hypernym/hyponymand synonym relationships among semantic concepts (by a lexical database), thus it can be considered as a particular *multimedia lexical ontology*.

To this goal, we used the *animate vision based algorithms*, that some of the authors have previously designed [4] in order to capture the visual content of the images; in addition the automatic annotation – i.e. the problem of associating images to a semantic descriptions – has been addressed, applying NLP techniques to keywords and annotations.

More in details, image semantic content could be represented by a multimedia ontology that provides two kinds of information. The first one uses classical *database categories* that are general concepts such as animal, landscape, etc... that can be seen as the *semantic dimensions* of analysis; the second one exploits more refined concepts, called *labels* (e.g. cat, sunset), obtained by a discovering image label process and the related semantic relationships to associate to each image a specific meaning that could be very useful for retrieval and browsing aims. We have used Flickr [16] as a large repository of annotated images, in order to build a novel intelligent building strategy. Flickr is one of the most popular web-based tagging system, that allows human participants to annotate a particular resource, such as web pages, blogs, images, with a freely chosen set of keywords, or tags, together with a short description of the content. This kind of system has been recently termed *folksonomy* [15], i.e. a folk taxonomy of important and emerging concepts within user groups. The dynamic nature of these repositories assures the richness of the annotation; in addition, they are really accurate, because they are produced by humans that want to share their images and the experience they have had, using tags and an annotation process.

3.2 The Process

The purpose of the image ontology building process is to automatically perform a low and intermediate levels analysis by using the *feature extraction* module on each image and to automatically determine an high-level description by using the *Discovering*

Image Labels module and to organize the obtained knowledge in the shape of a multimedia ontology through the *Image Ontology Builder* module . We first describe the Feature Extraction step and then we outline the Discovering Image Labels and Image Ontology Building steps.

Feature Extraction. To obtain a low-level and intermediate description of the images, we could apply different computer vision algorithm to obtain several image features. We decide to use the salient points technique - based on the *Animate Vision* paradigm - that exploits color, texture and shape information associated with those regions of the image that are relevant to human attention (*Focus of Attention*), in order to obtain a compact characterization (namely *Information Path*) that could be also used to evaluate the similarity between images, and for indexing issues. An information path can be seen as a particular data structure $IP=hF_s(p_s;\tau_s),h_b(F_s),\Sigma_{F_s}i$ that contains, for each region $F(p_s;\tau_s)$ surrounding a given salient point (where p_s is the center of the region and τ_s is the the observation time spent by a human to detect the point), the color features in terms of HSV histogram $h_b(F_s)$, and the texture and shape features in terms of wavelet covariance signatures Σ_{F_s} (see [5] formore details).We could also use some confidence values obtained by the algorithm of *Animate Vision* to characterize these information by means of fuzzy values. In Figure 1 we can see an example of *Information Path* with the intermediate description related with it that is the results of this process.

Discovering Image Labels. Images in Flickr usually have two attached texts, namely a *content description* and a *title*, and/or a set of keywords, namely *tags*. Descriptions are very short and usually are not posted for retrieval purposes; they typically contain sentences concerning the context of the picture, or the opinion of the user. Tags are simple keywords users are asked (actually they may not insert any tag) to submit, that describe the context of the image (e.g. amongst the tags for a picture of a dog playing with a soccer ball, you will probably see the words 'dog', 'ball', etc.). Titles in the majority of the cases contain text that summarize the content of the images, while in other cases consist of automatically generated text that is not useful in the indexing process. The simple use of tags does not improve the efficiency of indexing and searching contents in our system. In fact, the absence of restrictions to the vocabulary from which tags are chosen can easily lead to the presence of *synonyms* (multiple tags for the same concept), *homonyms* (same tag used with different meaning), and *polysemies* (same tag with multiple related meanings). Also inaccurate or irrelevant tags result from the so called *'meta-noise'*, e.g. lack of stemming (normalization of word inflections), and from heterogeneity of users and contexts: hence an effective use of the tags requires these to be stemmed, disambiguated, and opportunely selected. A similar kind of pre-processing is also needed for descriptions and titles, which are to be analyzed by a suitable "Topic Detection algorithm" to extract a set of *relevant keywords* which represent their content, and could be treated the same way the tags are. Thus themain aim of the text categorization process is to automatically determine a set of *labels* which is a subset of the whole keywords coming from Flickr image title, description and tags, with an associated *confidence* value - that represents the relative *importance* of those keywords with respect to the other ones in the annotations. We can schematize such process as a function:

$$\varphi : A \rightarrow \{\langle \gamma_i, \varrho_i \rangle\}$$

being the couple $\langle \gamma_i, \varrho_i \rangle$ a generic label with the related confidence and A the set of tags, title an descriptions. We first compute a normalization process that has the aim to filter *emoticons* and http links in the text. Then, we extract from the text (titles, tags and descriptions) names, in particular names of people or organizations, geographical locations, and substitute them with the related entity (e.g. the sentence “Bob works in BMW” is transformed into “*Person* works in *Organization*”) iff they are non present in the WordNet database. This task is accomplished by using the *Annie* Named Entity Recognition (NER) module of *GATE* project [9]. Then, we perform the classical steps of *stemming* and *part of speech tagging* and *tokenization*, with this phase we can reject stop-words, adjective, verbs, that are not useful for our goal and then we build a token vector that contains the selected keywords and their frequencies, computed looking at the whole image text description retrieved from Flickr. Then, we disambiguate the word senses attaching to the each keywords belonging to the previous token vector the main meaning. We apply aWord Sense Disambiguation (WSD) algorithm to noun words. In particular we exploit the algorithm in [21], that disambiguates the words applying “minimum common hyperonim” considerations to the parsed and stemmed initial sentence. In the last step, a suitable Topic Extraction algorithm determines the set of labels that are more significant to the description of the content of the text [17]. In particular, for each label extracted by theWSD module, a confidence value is computed considering both the *semantic similarity* (by exploiting the same algorithm used for WSD) of the token to the other ones, and the related *frequency* of the token in the text. Finally, we selects the top-K of image labels by using a a confidence threshold τ determined in an experimental way. Table 1 reports an example of the results, that are a set of labels with confidence values, of this process after applying all these text processing steps for the image derived from Flickr with its annotations.

Table 1. Flickr images and their related extracted information.

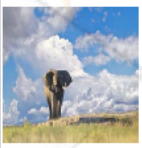
image	title	description	tags	$\langle \gamma_i, \epsilon_i \rangle$
	“Kruger Elephant”	“Bull elephant in Kruger National Park ”	{ Kruger National Park, Kruger, Elephant, bull, wildlife, Africa , South Africa, Portrait, landscape, BFGreatestHits }	“elephant,0.8”, “africa,0.68”, “landscape,0.76”

Image Ontology Building. In this section we describe how to perform the automatic image ontology building process.

The image ontology is generated in an incremental way and in correspondence of each images a pick-up operation performed by the fetching agent module from the Flickr repository is performed. The generic inputs of the construction procedure are: i) the set of WordNet synsets {π_i} related to the labels {γ_i} extracted by the Image Discovery Labels module and the *Information Path* of the input image coming from applying the Feature Extraction module on that input image. The building algorithm checks the synsets related to the different concepts (labels) relevant to the input image: if a node exists with the same synset in the current ontology, the image information path and the related description are associated to such at node, otherwise a new node with the related image information are instantiated and eventually connected to the most semantically similar nodes in the tree, i.e. the nodes correspondent to its *closest* hypernym and hyponym in theWordNet hierarchy. The aim here is on one hand, to automatically build a *taxonomy* concerning different semantic domains by exploiting the WordNet hyponym/hypernym relationships, on the other hand, to automatically discover the semantic concepts relevant to a given image, exploiting both users' high-level description and tags and the low and intermediate-level description. We state that it is possible to automatically create an Image Ontology for a whole image database without any precategorization; or a different tree for every pre-defined category in the database through a semi-automatic process (during the indexing operations, administrators are required to select the images from Flickr to be associated to each category). Finally, we want to remark that in our approach the only supervised step is the Discover Image Labels steps, in which human annotations from the Flickr semantic knowledge base are used to build an index that contains the concepts useful for the retrieval tasks.

4 Related Works

In the last few years, several papers have been presented about multimedia systems based on knowledge models, image ontologies, fuzzy extension of ontology theories.
In almost all the works, multimedia ontologies are effectively used to perform *semantic annotation* of the media content by manually associating the terms of the ontology with the individual elements of the image or of the video [22], [10], thus demonstrating that the use of ontologies can enhance classification precision and image retrieval performance. Instead of creating a new ontology from the scratch, other approaches [8] extendWordNet to image specific concepts, using the annotated image corpus as an intermediate step to compute similarity between example images and images in the image collection. For solving the uncertain reasoning problems, the theory of fuzzy ontologies is presented in several works, as an extension of the ontologies with crisp concepts as the papers [13] that presents a complete fuzzy framework for ontologies. In [19], the authors introduce a description logic framework for the interpretation of image contents. Very interesting are the multimedia semantic papers based on MPEG-7 [1], [6]. The MPEG-7 framework consists of Descriptors (Ds) and Descriptor Schemes (DSs) that represent features for

multimedia, and more complex structures grouping Ds and DSs, respectively. In particular, the MPEG-7 standard includes tags that describe visual features (e.g., color), audio features (e.g., timbre), structure (e.g.,moving regions and video segments), semantic (e.g., object and events), management (e.g., creator and format), collection organization (e.g., collections and models), summaries (e.g., hierarchies of key frames) and, even, user preferences (e.g., for search) of multimedia. In this way the standard includes descriptions of low-level media-specific features that can often be automatically extracted from media types. Unfortunately, MPEG-7 is not currently suitable for describing top-levelmultimedia features, because i) its XML Schema-based nature prevents the effective manipulation of descriptions and its use of URNs is cumbersome for the web; ii) it is not open to the web standards for representing knowledge. Some efforts was also done in order to translate the semantic of the standard in some knowledge representation languages [14], [18], [25]. All these methods perform a one to one translation of MPEG-7 types into OWL concepts and properties. Eventually, a very interesting work reported in [20], [2] was done in order to define a multimedia ontology. They try to define a new multimedia ontology that take into account the semantic of MPEG-7 standard. They started using some patterns derived from a foundational ontology DOLCE [7]. In particular they used two design patterns Descriptions & Situations (D & S) and Ontology of Information Objects (OIO), which are two of the main patterns provided by DOLCE. The ontology already covers a very large part of the standard, while their modeling approach has the aim to offer even more possibilities for multimedia annotation than MPEG-7 since it is truly interoperable with existing web ontology. This approach puts some constraints on the image semantic thought the use of foundational ontology even if this work is more focused on interoperability purposes.

5 Conclusions

In this paper we have addressed the problem of building a multimedia ontology in an automatic way using annotated image dabases. Our proposed work differs from the previous papers presented in the literature for different reasons. First, differently from the previous works we propose a notion of multimedia ontology, particularly suitable for capturing the complex semantics of images during several steps of the image analysis process; we do not propose any extension of the usual ontology theory and languages, but manage uncertainty implementing ternary properties by means of a reification process, thus taking advantages of the several existing reasoning systems; in addition, we obtain a dynamic generation of image ontologies without really asking any user to produce ad-hoc tagging or annotation: we just use tags and annotations that the user has produced in their social web network for themselves, when they really want to communicate something thus publishing pictures on the Flickr(or equivalent) system.

References

1. Mpeg-7. multimedia content description interface. standard no. iso/iec n15938, 2001.
2. Richard Arndt, Raphaël Troncy, Steffen Staab, Lynda Hardman, and Miroslav Vacura. Comm: Designing a well-founded multimedia ontology for the web. In *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC/ASWC*, volume 4825 of *Lecture Notes in Computer Science*, pages 30–43. Springer, 2007.
3. Irving Biederman. Human image understanding: recent research and a theory. In *Papers from the second workshop Vol. 13 on Human and Machine Vision II*, pages 13–57, San Diego, CA, USA, 1986. Academic Press Professional, Inc.
4. G. Boccignone, A. Chianese, V. Moscato, and A. Picariello. Foveated shot detection for video segmentation. *IEEE Trans. Circuits Syst. Video Techn.*, 15(3):365–377, 2005.
5. G. Boccignone, A. Chianese, V. Moscato, and A. Picariello. Context-sensitive queries for image retrieval in digital libraries. *Journal of Intelligent Information Systems*, on-line first, 2007.
6. Thomas Sikora B.S. Manjunath, Philippe Salembier. *Introduction to MPEG-7: Multimedia Content Description Interface*. 2002.
7. A. Gangemi N. Guarino A. Oltramari C.Masolo, S. Borgo and L. Schneider. The wonderweb library of foundational ontologies (wfol). Technical report, WonderWeb Deliverable 17, 2002.
8. Yih-Chen Chang and Hsin-Hsi Chen. Approaches of using aword-image ontology and an annotated image corpus as intermedia for cross-language image retrieval. In *Proceedings of Cross-Language Evaluation Forum, C. Peters et al. (Eds.), Lecture Notes in Computer Science*, 2006.
9. H. Cunningham. Information Extraction, Automatic. *Encyclopedia of Language and Linguistics, 2nd Edition*, 2005.
10. Stamou G, Van Ossenbruggen J, Pan J Z, Schreiber G, and Smith J R. Multimedia annotations on the semantic web. *Multimedia, IEEE*, 13:86 – 90, 2006.
11. Tom Gruber. Ontology. *Encyclopedia of Database Systems*, 2008.
12. Nicola Guarino. Formal ontology, conceptual analysis and knowledge representation. *Int. J. Hum.-Comput. Stud.*, 43(5-6):625–640, 1995.
13. Chang-Shing Lee; Zhi-Wei Jian; Lin-Kai Huang;. A fuzzy ontology and its application to news summarization. *Systems, Man and Cybernetics, Part B, IEEE Transactions on*, 35:859 – 880, 2005.
14. ane Hunter. Adding multimedia to the semantic web - building an mpeg-7 ontology. In *In International Semantic Web Working Symposium (SWWS*, pages 261–281, 2001.
15. L. Kennedy, M. Naaman, S. Ahern, R Nair, and T Rattenbury. How flickr helps us make sense of the world: context and content in community-contributed media collections. In *ACM Multimedia*, pages 631–640, 2007.
16. K. Lerman and L. Jones. Social browsing on flickr. *CoRR*, abs/cs/0612047, 2006.
17. Y. Li, Z. Bandar, and McLean D. An approach for measuring similarity between words using multiple information sources. *IEEE Trans. on Knowledge and Data Engineering*, 15(4):871–882, 2003.
18. Jacco Van Ossenbruggen, Frank Nack, and Lynda Hardman. That obscure object of desire: Multimedia metadata on the web, part 2. *IEEE Multimedia*, 12:54–63, 2005.
19. Bernd Neumann RalfMaller. Ontology-based reasoning techniques for multimedia interpretation and retrieval. In Springer, editor, *Semantic Multimedia and Ontologies*, pages 55–98. Springer London, 2008.
20. Steffen Staab Richard Arndt, Raphael Troncy and Lynda Hardman. Adding formal semantics to mpeg-7: Designing a well-founded multimedia ontology for the web. Technical report, Department of Computer Science. Technical Report of University of Koblenz., 2007.

21. P. Rosso, F. Masulli, and D. Buscaldi. Word sense disambiguation combining conceptual distance. In *Proc. Int. Conf. on Natural Language Processing and Knowledge Engineering*, pages 120–125, 2003.
22. Petridis K. Bloehdorn S. Saathoff C. Simou N. Dasiopoulou S. Tzouvaras V. Handschuh S. Avrithis Y. Kompatsiaris Y. Staab S. Knowledge representation and semantic annotation of multimedia contentvision. In *Image and Signal Processing, IEEE Proceedings*, volume 153, pages 255–262, 2006.
23. John F. Sowa. Knowledge representation: Logical, philosophical, and computational foundations. In *Brooks Cole Publishing Co., Pacific Grove, CA*, 2000.
24. V. S. Subrahmanian. *Principles of Multimedia Database Systems*. Morgan Kaufmann, 1998.
25. Jacco van Ossenbruggen, Frank Nack, and Lynda Hardman. That obscure object of desire: Multimedia metadata on the web, part 1. *IEEE MultiMedia*, 11(4):38–48, 2004.