Enabling Interoperability between Multimedia Resources: An Ontology Matching Perspective

Nicolas James, Konstantin Todorov, and Céline Hudelot
{nicolas.james,konstantin.todorov,celine.hudelot}@ecp.fr

MAS Laboratory, École Centrale Paris, F-92 295 Châtenay-Malabry, France

Abstract. The semantic annotation of images can benefit from representations of useful concepts and the links between them as ontologies. Recently, several multimedia ontologies have been proposed in the literature as suitable knowledge models to bridge the well known semantic gap between low level features of image content and its high level conceptual meaning. Nevertheless, these multimedia ontologies are often dedicated to (or initially built for) particular needs or a particular application. Ontology matching, defined as the process of relating different heterogeneous models, we will argue, is a suitable approach to solve interoperability issues in semantic image annotation and retrieval. We propose a generic instance-based ontology matching approach, applied to an important semantic image retrieval issue: the bridging of the semantic gap by matching a multimedia ontology against a common-sense knowledge resource.

1 Introduction

The fast growth of shared digital image and video collections together with the intensive use of visual information for decision making in many domains (medicine, geosciences, etc) require new effective methods for search and retrieval in these collections. In order to enable and improve the communication and the interface between humans and computers, it is necessary to understand the semantic content of images and to built linguistic descriptions of their content in an automatic way. Following decades of research on Content Based Image Retrieval (CBIR), automatic image annotation is nowadays an active research topic which aims at bridging the semantic and the perceptual levels of abstraction, known as the *Semantic gap* problem [11]. In most of the image annotation approaches, the computed linguistic description is often only related to perceptual manifestations of semantics. Nevertheless, as explained in [5], the image semantics cannot be considered as being included explicitly in the image itself. It rather depends on prior knowledge and on the context of use of the visual information. In consequence, explicit semantics, represented by ontologies, has been intensely used in the field of image retrieval recently.

With the growth of the application of ontology-based solutions in the multimedia domain, a lot of interoperability issues have arisen: (a) *At the semantic level* – between different representations of the same domain knowledge; (b) *At*

the visual level – between different multimedia ontologies; (c) *Between the visual level and the semantic level*, i.e. the semantic gap problem. Ontology matching, widely used for semantic web applications and rarely in the context of image sharing and retrieval, that we defined as the process of relating heterogeneous knowledge models, can be used to solve these kinds of interoperability issues. This paper proposes a generic approach to address the question of filling the semantic gap by matching an ontology at the semantic level (Wordnet[1] associated to the image database LabelMe[10]) with an ontology at the visual level (LSCOM [12]).

Next section is a short review of existing multimedia ontologies and related approaches. Section 3 describes the ontology matching framework which forms the methodological background of our approach, presented in turn in Section 4. Results of our preliminary experiments are discussed in Section 5; Section 6 concludes.

2 Related Work

In the past few years, *concept-based multimedia retrieval* has been a very active research field with a major effort in the automatic detection of semantic concepts from low level features with machine learning approaches. Despite these efforts, the semantic gap problem is still an issue for the semantic understanding of multimedia documents. Recently, many knowledge models have been proposed to improve multimedia retrieval and interpretation by the explicit modeling of the different relationships between semantic concepts. Indeed, many generic large scale multimedia ontologies or multimedia concept lexicons together with image collections have been proposed to improve multimedia search and retrieval by providing an effective representation and interpretation of multimedia concepts [13,12,1]. We propose to classify these ontologies in four major groups: (1) semantic web multimedia ontologies often based on MPEG-7, reviewed in [1] (2) visual concept hierarchies (or networks) inferred from inter-concept visual similarity contexts (among which VCNet based on Flickr Distance [15] and the Topic Network of Fan [3]), (3) specific multimedia lexicons often composed of a hierarchy of semantic concepts with associated visual concept detectors used to describe and to detect automatically the semantic concepts of multimedia documents (LSCOM [12], multimedia thesauri [13]) and (4) generic ontologies based on existing semantic concept hierarchies such as WordNet populated with annotated images or multimedia documents (ImageNet [2], LabelMe [10]). These ontologies have proved to be very useful mainly in the context of semantic concept detection and automatic multimedia annotation but many problems still remain unsolved among which enabling the interoperability between visual concepts and high level concepts. Although there exist attempts to solve these problems by manual concept mappings [13], little effort has been directed towards performing them in an automatic manner. Moreover, these ontologies are often dedicated to (or built for) particular needs or a particular application and are

[1] `http://wordnet.princeton.edu/`

complementary knowledge sources. While studies have been done to analyze the different inter-concept similarities in different multimedia ontologies [8], to the best of our knowledge, there are no studies which propose a cross analysis and a joint use of these different and complementary ontologies.

This paper proposes to situate these problems in an O[ntology] M[atching] framework. The OM-approach presented in next section is much in line with the tradition of *extensional* matching. This comprises a set of techniques which base the similarity of concepts on characteristics of the instances that these concepts contain [6].

3 An Ontology Matching Approach

An ontology is based on a set of *concepts* and *relations* defined on these concepts, which altogether describe the knowledge in a given domain of interest. Due to the fact that different communities, independently from one another, tend to conceptualize differently the same domain of interest, a growing number of *heterogeneous* ontologies, describing similar or overlapping parts of the world are created. An OM procedure aims at reducing this heterogeneity by linking the correspondent elements of two ontologies in an automatic or semi-automatic manner.

Formally, a **populated ontology** will be defined by $O = \{C, \texttt{is\_a}, R, I, g\}$, where C is a set whose elements are called concepts, `is_a` is a partial order on C, R is a set of other (binary) relations holding between the concepts from the set C, I is a set whose elements are called instances and $g : C \rightarrow 2^I$ is an injection from the set of concepts to the set of subsets of I.

We note that the sets C and I are compulsorily non-empty, in contrast to R. Thus, the definition above describes an ontology which, although not limited to subsumptional relations, necessarily contains a hierarchical backbone, defined by the partial order. The set I may contain text documents, images or other (real world data) entities. By assumption, every instance can be represented as an n-dimensional real-valued vector, defined by n input *variables* of some kind which are the same for all instances in I.

In the context of semantic image annotation, WordNet together with the LabelMe database [10] and LSCOM [12] together with the TRECVID 2005 database are two examples of such populated ontologies. Concepts are the nodes of the WordNet hierarchy in ImageNet or the LSCOM categories, while instances are the images in the associated databases, which are labeled by these concepts. It is important to note that the set R is empty for the LSCOM ontology. In the case of WordNet, R contains several useful relations like `is_a_member_of`, `is_a_part_of`, `opposes`, etc.

Often the outcome of an OM-procedure is a set of cross-ontology concept alignments, issued from a **measure of concept similarity**. The measures used in the current study are based on *variable selection* and we will describe them in more detail.

Variable selection techniques (reviewed in [4]) serve to rank the input variables of a given problem (e.g. classification) by their importance for the output (the class affiliation of an instance), according to certain evaluation criteria. A real valued *score* which accounts for this importance is attached to every variable. In our case, this can be of help for uncovering latent input-output dependencies. Assuming that instances are represented as real-valued vectors, the computed scores would indicate which of the vector dimensions are most important for the separation of the instances (within a single ontology) into those that belong to a given concept and those that do not and thus best characterize this concept.

We define a binary classification training set S_O^c for each concept c from an ontology O by taking I, the entire set of instances assigned to O and labeling all instances from the set $g(c)$ as *positive* and all the rest $(I \backslash g(c))$ as *negative*. By the help of a variable selection procedure performed on S_O^c, we obtain a representation of the concept c as a list

$$L(c) = (s_1^c, s_2^c, ..., s_n^c), \tag{1}$$

where s_i^c is the score associated to the ith variable. To compute a score per variable and per concept, we apply the S[upport] V[ector] M[achine]-based variable selection technique introduced in [14]. A series of SVMs is learned on the training set S_O^c by subsequently removing a variable at a time. The ability of each variable to discriminate c from the other concepts in O is evaluated by measuring the sensitivity of the VC-dimension, an important SVM parameter, with respect to the variable in question.

By following the described procedure, given two source ontologies O_1 and O_2, a representation as the one in (1) is made available for every concept of each of these ontologies. The similarity of two concepts, $A \in O_1$ and $B \in O_2$ is then assessed in terms of their corresponding representations $L(A)$ and $L(B)$. Several choices of a similarity measure based on these representations are proposed and compared in [14]. In the experimental work contained in this paper, we have used Pearson's, Spearman's and Kendall's measures of correlation calculated on the variable scores or ranks (integers corresponding to the scores) given by

$$sim_{Pearson} = \frac{\sum_{i=1}^{n}(s_i^A - s_{mean}^A)(s_i^B - s_{mean}^B)}{\sqrt{\sum_{i=1}^{n}(s_i^A - s_{mean}^A)^2}\sqrt{\sum_{i=1}^{n}(s_i^B - s_{mean}^B)^2}}, \tag{2}$$

$$sim_{Spearman} = 1 - 6\frac{\sum_i d_i^2}{n(n^2-1)}, \quad sim_{Kendall} = \frac{n_c - n_d}{\frac{1}{2}n(n-1)}. \tag{3}$$

In the formulae above, s_{mean}^A and s_{mean}^B are the means of the scores over all input variables, d_i is the difference of the ranks calculated for the ith variable w.r.t. the two concepts, and n_c and n_d are the numbers of concordant and discordant pairs among the lists of scores $L(A)$ and $L(B)$.

4 Filling the Semantic Gap with Mapped Concepts

As noted in the introduction, many challenging issues in the field of image retrieval stem from the semantic gap problem. Two examples are the construction of robust high level concept detectors and the creation of user oriented annotations with high level semantics. In this section, we propose an attempt to fill the semantic gap by matching two complementary resources: a *visual* and a *semantic* thesaurus. Contrary to [13], our approach is *automatic*, *generic* (ontology independent) and makes use of the *visual knowledge* shared by the source ontologies.

On one hand, we chose LSCOM [12], an ontology dedicated to multimedia annotation. It was initially built in the framework of TRECVID[2] with the criteria of concept usefulness, concept observability and feasibility of concept automatic detection. LSCOM is populated by the development set of TRECVID 2005 videos (news broadcasting). On the other hand, we used WordNet [9] populated with the LabelMe dataset [10]. Many interoperability issues can be addressed for these two ontologies among which semantic interoperability and semantic gap interoperability. Aligning these resources allows for the semantic enrichment of concepts belonging to a multimedia ontology with high level linguistic concepts from a general and common sense knowledge base and the evaluation of the quality of the baseline concept detectors by studying the link between concepts whose semantics is related to their perceptual manifestations and concepts whose semantics is related to common sense.

In our setting, the instances that extensionally define a concept are images whose annotations contain the name associated to this concept. An image is represented as a vector of descriptors. We use a codebook built on a bag-of-features model and histograms of codewords which is, nowadays, the best approach in the state-of-the-art [7]. In that, the variables which describe the instances are the bins of these histograms. The generic variable selection approach described in Section 3 is applied directly on our data. In result, we obtain a concept representation as the one introduced in eq. (1) for every concept of our two source ontologies. As stated above (Section 3), there exist several plausible choices of a measure of similarity for two concepts represented in this manner. In our experiments, we have tested the three measures of correlation given in (2) and (3). Regardless of the particular choice, the similarity is always based on *visual criteria*, since the underlying concept representations are obtained by using visual characteristics of the instances (in the particular case of LSCOM and WordNet these are the sets of images of either TRECVID or LabelMe).

Aligning LSCOM to WordNet allows to infer knowledge about the LSCOM concepts (dedicated to the multimedia document annotation) with regard to the concepts of WordNet and the alignment could be used to build a linguistic description of the concepts of LSCOM, or, in other words, to answer the question *"What is an LSCOM concept in WordNet?"* in an automatic manner. This improves the retrieval process in several ways: (1) through query expansion and

[2] http://www-nlpir.nist.gov/projects/tv2005/

reformulation, i.e. retrieving documents annotated with concepts from an ontology O_1 using a query composed of concepts of an ontology O_2, (2) through a better description of the documents in the indexing process. However, note that this relation is not symmetric: alignments in the other sense are prompt to fail to be of any help, since WordNet concepts are rather atomic (such as "car") as compared to the more complex LSCOM concepts (e.g. "Natural Disaster Scene").

5 Experimental Results

We use a part of the LSCOM ontology, LSCOM_Annotation_v1.0[3], which is a subset of 449 concepts from the initial LSCOM ontology, and is used for annotating 61,517 images from the TRECVID2005 development set. Since this set contains images from broadcast news videos, the choosen LSCOM subpart is particularly adapted to annotate this kind of content, thus contains abstract and specific concepts (e.g. `196_Science_Technology`, `330_Interview_On_Location`). To the contrary, our sub-ontology defined from WordNet populated with LabelMe (3676 concepts) is very general considering the nature of LabelMe, which is composed of photographs from the daily life.

In this way, to provide a preliminary evaluation of the suggested approach, we chose three concepts from the LSCOM ontology and five concepts from the WordNet ontology. The choice of the selected concepts was made on several criteria: (1) the number of associated instances, (2) for every selected concepts there is no semantic ambiguity in our dataset, (3) for WordNet only: a high confidence (arbitrarily decided) in the discrimination of the concept using only perceptual information.

Table 1. LSCOM/TRECVID2005 against WordNet/LabelMe from left to right rowwise: Variable selection-based concept similarities with Pearson's, Spearman's and Kendall's coefficients and a manual annotation of the TRECVID2005 images.

Concept Names	Man	Car	Boat	TV	House	Man	Car	Boat	TV	House
Natural Disasters	**0.28**	0.12	-0.08	-0.02	**0.17**	**0.37**	0.15	-0.33	0.12	**0.44**
US Flags	0.17	0.01	0.15	-0.15	0.063	**0.21**	0.09	0.01	0.05	0.05
Single Family Homes	**0.21**	0.13	-0.22	0.05	**0.23**	**0.20**	0.18	-0.36	0.13	**0.41**
Natural Disasters	**0.36**	0.19	-0.15	0.016	**0.37**	103	51	4	0	73
US Flags	0.15	0.13	0.09	0.012	0.11	434	16	0	2	28
Single Family Homes	**0.23**	0.19	-0.18	0.016	**0.35**	205	73	1	0	184

[3] http://www.ee.columbia.edu/ln/dvmm/lscom/

Fig. 1. An LSCOM image annotated by LSCOM:Single_Family_Homes, and four sample images from WordNet/LabelMe. The LSCOM image can be effectively annotated and retrieved using the WordNet concepts Man and House.

To construct image features, we use a bag-of-features model with a visual codebook, built classically using the well known SIFT descriptor and a K-Means algorithm. The quantification of the extracted SIFT features was investigated in two ways: (1) over all the instances associated to the selected concepts (LSCOM and WordNet), (2) only over the LabelMe images and quantification per concept. The two experimentations gave very similar results, and the results of the experiment based on the first codebook are resumed in Table 1.

The values in the first three matrices are correlations indicating high similarity for positive values (low for non-positive). As we can see, the concept WordNet:TV is weakly correlated to the chosen LSCOM concepts, and the concept WordNet:House is highly correlated with LSCOM:Natural_Disasters and LSCOM:Single_Familly_Homes but not with LSCOM:US_Flags. This is coherent with the TRECVID2005 data considering that the images annotated with LSCOM:US_Flags are mostly images from speeches of politicians during presidential elections. An example of an LSCOM image annotation that could be extended to WordNet by the help the concept mapping is given in Fig. 1.

6 Conclusion

The paper proposes an ontology matching technique to solve interoperability issues in the area of semantic image annotation and retrieval. In particular, we have addressed the problem of bridging the semantic gap by the help of a generic instance-based ontology matching approach which aims at automatically producing concept-based annotations enriched with a lexical description of the concepts. In preliminary experiments, we have tested a concept similarity measure on two small sets of concepts taken from the LSCOM ontology and WordNet/LabelMe. Our results are in good agreement with the nature of the

instances associated to the selected LSCOM concepts. However, the efficiency of the approach has to be tested on larger sets of concepts (currently in progress). A large-scale application would also allow us to benefit from all the semantic relations in WordNet, like hypernymy, meronymy, antonymy. In the future, we plan to investigate the qualities of our automatic approach in terms of retrieval efficiency as compared to approaches that solely rely on manual mappings.

Acknowledgments. This work is funded by the French National Research Agency (ANR) through the COSINUS program (project COLLAVIZ ANR-08-COSI-003) and by the region Île de France through the SEBASTIAN2 project (Cap Digital cluster).

References

1. S. Dasiopoulou, V. Tzouvaras, I. Kompatsiaris, and M.G. Strintzis. Enquiring MPEG-7 based multimedia ontologies. *Multimedia Tools and Applications*, pages 1–40, 2010.
2. J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, and L. Fei-Fei. ImageNet: a large-scale hierarchical image database. In *CVPR*, pages 710–719, 2009.
3. J. Fan, H. Luo, Y. Shen, and C. Yang. Integrating visual and semantic contexts for topic network generation and word sense disambiguation. *ACM CIVR'09*, pages 1–8, 2009.
4. I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *JMLR*, 3(1):1157–1182, 2003.
5. C. Hudelot, N. Maillot, and M. Thonnat. Symbol grounding for semantic image interpretation: from image data to semantics. In *SKCV-Workshop, ICCV*, 2005.
6. A. Isaac, L. van der Meij, S. Schlobach, and S. Wang. An empirical study of instance-based ontology matching. *The Semantic Web*, pages 253–266, 2008.
7. Y.G. Jiang, J. Yang, C.W. Ngo, and A.G. Hauptmann. Representations of keypoint-based semantic concept detection: A comprehensive study. *IEEE Trans. on Multimedia, in press*, 2010.
8. M. Koskela and A. Smeaton. An empirical study of inter-concept similarities in multimedia ontologies. In *CIVR'07*, pages 464–471. ACM, 2007.
9. G.A. Miller. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.
10. B.C. Russell, A. Torralba, K.P. Murphy, and W.T. Freeman. LabelMe: a database and web-based tool for image annotation. *IJCV*, 77(1):157–173, 2008.
11. 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. Patt. An. Mach. Intell.*, pages 1349–1380, 2000.
12. J.R. Smith and S.F. Chang. Large-scale concept ontology for multimedia. *IEEE Multimedia*, 13(3):86–91, 2006.
13. C.G.M. Snoek, B. Huurnink, L. Hollink, M. De Rijke, G. Schreiber, and M. Worring. Adding semantics to detectors for video retrieval. *IEEE Trans. on Mult.*, 9(5):975–986, 2007.
14. K. Todorov, P. Geibel, and K.-U. Kühnberger. Extensional ontology matching with variable selection for support vector machines. In *CISIS*, pages 962–968. IEEE Computer Society Press, 2010.
15. Lei Wu, Xian-Sheng Hua, Nenghai Yu, Wei-Ying Ma, and Shipeng Li. Flickr distance. In *MM'08*, pages 31–40. ACM, 2008.