Ontology Matching for the Semantic Annotation of Images

Nicolas James and Konstantin Todorov and Céline Hudelot

Abstract—The linguistic description, i.e. 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, could be a suitable approach to solve several interoperability issues that coexist in semantic image annotation and retrieval. In this paper, we propose an original and generic instance-based ontology matching approach and a methodology to extract a minimal ontology defined as the common reference between different heterogeneous ontologies. Then, this approach is applied to two different semantic image retrieval issues: the bridging of the semantic gap by the matching of a multimedia ontology with a common-sense knowledge ontology and the matching of different multimedia ontologies to extract a common reference knowledge model dedicated to several multimedia applications.

I. 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 (as for instance medicine or geosciences) require new effective methods to search and retrieve in these collections. In particular, in order to enable and to improve the communication and the interface between humans and computers, it is necessary to understand the semantic content of images and to build linguistic descriptions of their content in an automatic way. Following decades of research on Content Based Image Retrieval (CBIR) [26], automatic image annotation is a very active research topic which aims at automatically assigning linguistic terms (semantic level) describing the visual information depicted in images (perceptual level) [4]. As opposed to the domain of analysis and indexing of textual documents, the visual domain has to face the important issue of matching human interpr etations of image information with the numerical image signature derivable by a computer. This issue is defined as the *Semantic gap* problem [26]. As suggested in [16] and [32], this problem is close to the symbol grounding problem [11] or anchoring problem [1], respectively addressed in A[rtificial] I[ntelligence] and in robotics. In most of the image annotation approaches, the computed linguistic description is often only related to perceptual manifestations of semantics. Nevertheless, as explained in [16], 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. As a consequence, explicit semantics, represented by ontologies, has been intensely used in the field of image retrieval recently. Indeed, ontologies are c onvenient not only to represent visual knowledge [23] but also to allow automatic processing over the represented knowledge [2].

Another benefit of ontologies in the context of shared digital image and video collections is that they are interoperable description schemes to represent, share and reason on visual knowledge. Indeed, a lot of interoperability issues exist in this context: (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, that we defined as the process of relating heterogeneous knowledge models, can be used to solve this kind of interoperability issues. Nevertheless, while ontology matching has been widely used for semantic web applications, it has been rarely used in the context of image sharing and retrieval.

In this paper, we propose an original and generic approach based on ontology matching and on the extraction of a minimal ontology to solve several interoperability-related problems in the visual domain. In particular, we address the following questions:

(1) Filling the semantic gap by matching ontologies at the semantic level with ontologies at the visual level.

(2) Matching multiple visual ontologies in order to extract a common reference, i.e. a common and consensual visual model for linguistic descriptions of images.

The paper is structured in the following manner. A generic approach for instance-based ontology matching and for the extraction of a minimal ontology out of a set of heterogeneous source ontologies is suggested in Section II. The two application-related issues, pointed above, are addressed, respectively, in Section III-A and Section III-B with some results showing the potential of the method. Section IV proposes a brief overview of related approaches concerning both the use of knowledge models for the linguistic description of images and ontology matching (general *and* specific to the problem of linguistic description for semantic annotation of images). Finally, Section V summarizes and outlines open ends.

II. Proposed Approach: Ontology Matching and Minimal Ontology Extraction

An ontology in an AI sense is understood as a collection of *concepts* and *relations* defined on these concepts, which represent the knowledge in a certain domain of interest

Nicolas James, Konstantin Todorov, Céline Hudelot are with the Applied Mathematics and Systems Laboratory (MAS), Ecole Centrale Paris, Grande Voie des Vignes, 92295 Châtenay-Malabry, France (phone: +33 141 131711; email: {nicolas.james,konstantin.todorov,celine.hudelot}@ecp.fr).

and provide reasoning and inference mechanisms. The *ontology matching* problem stems from the fact that different communities, independently from one another, are likely to adopt different ontologies, given a domain of interest. In consequence, multiple *heterogeneous* ontologies, describing similar or overlapping parts of the world are created. Heterogeneity may occur on syntactical, terminological, conceptual and other levels, not in isolation from one another; it can be observed among individuals or among groups of individuals. An ontology matching procedure aims at reducing this heterogeneity by yielding assertions on the relatedness of cross-ontology concepts, in an automatic or semi-automatic manner. To these ends, according to [7], one commonly relies on ex tensional (related to the concepts instances), structural (related to the inter-ontology concepts relations), terminological (language-related) or semantic (related to logical interpretation) information, separately or in combination.

In the current section, we present an ontology matching framework for aligning the concepts of multiple heterogeneous ontologies modeling intersecting domains of interest, in order to enable their interoperability and facilitate the interaction of human or artificial agents with these resources. Our goal has been to make explicit the relations that hold between the different cross-ontology concepts via a novel, *minimal ontology*, defined on and relevant to the set of source ontologies. A minimal ontology is understood in the sense of a structure limited to what is believed to have the same significance for the entire set of ontologies over a given instance-set. The novel ontology results from aligning similar concepts from the source ontologies, where the measure of concept similarity is defined by the help of their corresponding instance-sets. One may interpret the minimal ontology as a resource providing a common vocabulary for a set of heterogeneous vocabularies, w here the common reference is implicitly defined through a set of concept alignments.

In the targeted application domains, the minimal multimedia ontology could be interpreted as the set of predominant semantic concepts that appear in various image databases.

Although, throughout the exposition of the current section, it has been our goal to remain as abstract as possible, we will link the concepts to be further presented with concepts from the visual domain by small examples (where appropriate) in order to keep track of our eventual goal: indicate possible applications of the ontology matching paradigm for facilitating/enabling the semantic annotation of images.

A. Populated Ontologies

The definition below, a modified version of [30], relates structure to instances in the following manner.

Definition 1: A **populated ontology** is a tuple $O = \{C, \texttt{is\_a}, R, I, g\}$, where C is a set whose elements are called concepts, $\texttt{is\_a}$ is a partial order on C, R is set of other (binary) relations holding between the concepts from the set C, I is a set whose elements are called instances and $g : C \to 2^I$ is an injection from the set of concepts to the set of subsets of I.

In this formulation, a concept is *intensionally* defined by its relations to other concepts via the partial order and the set R, and *extensionally* by a set of instances via the mapping g. We note that the sets C and I are compulsorily non-empty, whereas R can be the empty set. In view of this remark, the definition above describes a *hierarchical ontology*: an ontology which, although not limited to subsumptional relations, necessarily contains a hierarchical backbone.

The set I is a set of concept instances – text documents, images or other (real world data) entities, representable in the form of real-valued vectors. The injection g associates a set of instances to every concept. By definition, the empty set can be associated to a concept as well, hence not every concept is expected or required to have instances. Whether g takes inheritance via subsumption into account in defining a concept's instance-set (hierarchical concept instantiation, assumed in our study) or not (non-hierarchical instantiation) is a semantics and design-related issue [17].

In the context of semantic image annotation, ImageNet [5] and LSCOM [27] 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, labeled by these concepts. Note that the set R is empty for the LSCOM ontology. In the case of ImageNet, R contains several useful (WordNet) relations like `is_a_member_of`, `is_a_part_of`, `opposes` etc.

B. Ontology Matching by Variable Selection

A concept similarity measure of some kind usually stands in the core of an ontology matching procedure. Previous work [31] introduces instance-based similarity measures, which use variable selection in order to represent concepts as sets of characteristic features. We will see (Section III) that this representation has multiple benefits for solving the problems pointed out in the introduction of this article. Therefore, we proceed to explain in some detail the main mechanisms of this approach.

Variable selection techniques in machine learning (reviewed in [10]) 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 variable selection procedure attaches to each variable a real value – *a score* – which indicates the variable's pertinence. This can be of help for dimensionality reduction tasks or, as in our case, for extracting latent input-output dependencies. Assuming that instances are represented as real-valued vectors, a variable selection procedure 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.

Let us consider two ontologies O_1 and O_2 together with their corresponding sets of instances $I_1 = \{\mathbf{i}_1^1, ..., \mathbf{i}_{m_1}^1\}$ and $I_2 = \{\mathbf{i}_1^2, ..., \mathbf{i}_{m_2}^2\}$, assuming that all instances from both

I_1 and I_2 live in the same n-dimensional space[1], m_1 and m_2 are integers. We recall that in the visual domain these instances are *images*, represented as feature vectors in one of the ways explained in Section III-A2. For a concept A from ontology O_1, we define a labeling $S^A = \{(\mathbf{i}_j^1, y_j^A)\}$, where $\mathbf{i}_j^1 \in \mathbb{R}^n$, y_j^A take values $+1$ when the corresponding instance $\mathbf{i}_j^1$ is assigned to A, and -1 otherwise, $j = 1, ..., m_1$. The labels split the instances of O_1 into those that belong to the concept A (positive instances), and thos e that do not (negative ones). Such a labeling can be acquired analogously and independently for any concept in both input ontologies.

For two concepts of interest, $A \in C_1$ and $B \in C_2$, we carry out a variable selection procedure independently on each of their corresponding sets $S^A = \{(\mathbf{i}_j^1, y_j^A)\}$, $j = 1, ..., m_1$ and $S^B = \{(\mathbf{i}_k^2, y_k^B)\}$, $k = 1, ..., m_2$, and score the variables by their importance for the respective class separation. In consequence, the concepts A and B can be represented by the lists of their corresponding variables scores:

$$L^A = (s_1^A, s_2^A, ..., s_n^A), \ L^B = (s_1^B, s_2^B, ..., s_n^B), \quad (1)$$

where s_i^A is the score assigned to the ith variable for the concept A.

On the basis of the concept representations (1), one can define various measures of similarity between A and B. The k-TF measure takes the sets of variables corresponding to the k largest elements of L^A and L^B and measures their intersection on set theoretic bases. Alternatively, parameter-free measures of statistical correlation, which act as measures of similarity, can be computed over the ranks (integers corresponding to the scores) or directly over the scores associated to the variables. In the experimental part of the paper, we have used Spearman's measure of correlation, given by:

$$sim_\rho = 1 - 6 \frac{\sum_i d_i^2}{n(n^2 - 1)}, \quad (2)$$

where d_i is the difference of the ranks calculated for the ith variable for the two classes. Definitions and comparison of the performance of several other measures on textual instances are found in [31]. The concept representation (1) and the ensuing similarity measures can be successfully applied on images and we will provide methodological and experimental support of that claim in Section III.

We assume that $sim : C_1 \times C_2 \rightarrow \mathbb{R}$ is a measure that uses the similarity criteria discussed above (abstracting ourselves from a particular choice) applied on two concept sets taken from two different source ontologies. A match-and-merge procedure, based on the measure sim, will be defined as a procedure $\mathcal{M}$, which takes two ontologies O_1 and O_2 and produces a third ontology, O, by aligning the concepts of the smaller of the two (assuming O_1) to the concepts of the bigger one (O_2). In fact, the resulting merged ontology O is the ontology O_2, enriched with links to the concepts from

---

[1]In the sequel, when using the term *instance*, we will be referring to the instance's representation in that space.
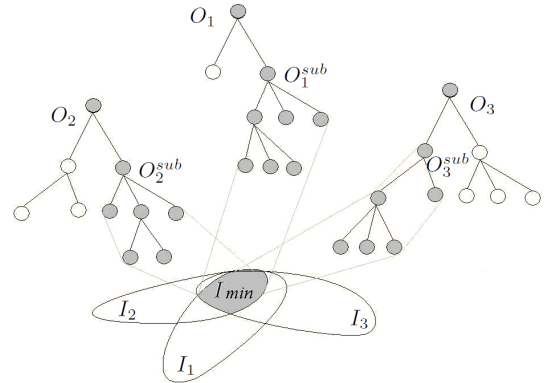

Fig. 1. An example with three source ontologies.

O_1. The concepts of O use as names the set of names of the concepts from the ontologies O_1 and O_2, which form them. Thus, the links between the two source ontologies via the merged ontology are made explicit.

To indicate that the procedure described above is applied on two input ontologies O_1 and O_2 resulting in an output ontology O, we will use the denotation $\mathcal{M}(O_1, O_2) = O$.

C. Extraction of a Minimal Ontology

We will focus on the problem of defining a common reference for a set of heterogeneous ontologies, which are assumed to be extensionally similar, to a certain extent.

Let $\Omega = \{O_1, ..., O_N\}$ be a set of source ontologies. Following definition 1, every ontology $O_i \in \Omega$, $i = 1, ..., N$, is defined as a pentuple $O_i = \{C_i, \texttt{is\_a}, R_i, I_i, g_i\}$. We assume that there exists a non-empty intersection of the instance-sets of the source ontologies , i.e. $\bigcap_{i=1}^N I_i \neq \emptyset$ and let $\bigcap_{i=1}^N I_i = I_{min}$. The term *intersection* is not understood in the classical sense of a strict intersection, but rather in the sense of a definition, specific for the type of instances (in the case of images, we consider the definition introduced in Section III-B). We will define a sub-ontology of each member of Ω which is based on the instance set I_{min}.

Definition 2: A **minimal set of concepts** for an ontology O_i corresponding to the instance-set I_{min} is defined as the set $C_i^{min} = \{A_j | A_j \in C_i, g_i(A_j) \cap I_{min} \neq \emptyset, j = 1, ..., M_i\}$, where M_i is the cardinality of the concept set of O_i. **A sub-ontology of** O_i, based on I_{min} will be defined as

$$O_i^{sub} = \{C_i^{min}, \texttt{is\_a}, R_i^{min}, g_i^*, I_{min}\},$$

where $\texttt{is\_a}$ is a partial order on C_i^{min}, $R_i^{min} = \{(A_k, A_l) \in R_i | A_k, A_l \in C_i^{min}, \forall k, l = 1, ..., M_i\}$ and g_i^* maps a concepts A_j to the intersection of its instance set with I_{min}, i.e. $g_i^*(A_j) = g_i(A_j) \cap I_{min}$.

We emphasize two implications of the definition above: (1) a concept from O_i will be included in the ontology O_i^{sub} not only when all of its instances are from I_{min}, but when at least some of them are; (2) the concepts from O_i^{sub} are extensionally redefined (as compared to the same concepts in O_i) by removing from their extensions all the instances that have failed to belong to I_{min}. The latter is an important
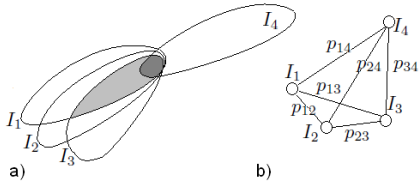
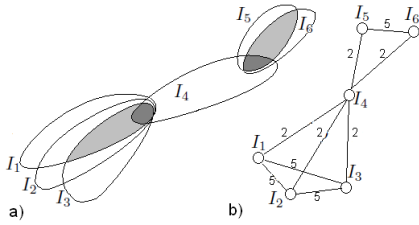Fig. 2. Intersection of 4 sets (a) and their corresponding WIG (b).

Fig. 3. Intersection of 6 sets (a) and their corresponding WIG (b) (the weights are chosen arbitrarily).

step towards removing "noisy" instances which, if kept, may lead to flawed similarity values in the matching procedure described in the previous section (applied in the sequel for the extraction of the minimal ontology).

Given these assumptions, we will discuss a possible approach to define an ontology which is extensionally based on the instance-set I_{min}, and intensionally on the set of concepts from the ontologies Ω', called a *minimal ontology* of Ω, denoted $O_{min}(\Omega)$ (see Fig. 1 for an illustration).

In order to construct an ontology out of a set of existing ontologies, we need to identify the similarities between their elements. In that sense, the problem amounts to a multiple-ontology matching task. Given the set Ω', containing *a priori* different ontologies, all based on one and the same set of instances, the resulting minimal ontology should be *extensionally* based entirely and exhaustively on this set. In *intensional* terms, our aim will be to preserve as much of the overall *conceptual richness* of the set Ω' as possible. Therefore, a leading design principle will be to keep the concept-set cardinality of the new ontology not lower than the cardinality of the largest ontology from Ω'. Hence our objective to create a minimal ontology with regard to the set Ω and maximal with regard to the set Ω'.

We make the convention that by ontology cardinality we will mean the cardinality of the ontology's concept-set and let the cardinality of $O_i^{sub} \in \Omega'$ be M_i^{sub}. Let $\{O_{\xi(1)}^{sub}, O_{\xi(2)}^{sub}, ..., O_{\xi(N)}^{sub}\}$ be the set of sub-ontologies, ordered by their respective cardinalities, where ξ is a permutation on the set of ontology indexes such that $M_{\xi(1)}^{sub} > M_{\xi(2)}^{sub} > ... > M_{\xi(N)}^{sub}$.

The procedure consists in performing multiple binary matches by recursively *aligning and merging* the two largest ontologies from the set Ω'. In algorithmic terms, that means to assign $O_{min} := O_{\sigma(1)}^{sub}$ and perform $O_{min} := \mathcal{M}(O_{min}, O_{\sigma(j)}^{sub}), \forall j = 2, ..., N$, where $\mathcal{M}$ is a match and merge procedure defined in Section II-B. In N match-and-merge steps, we end up with the ontology O_{min}. As specified

in Section II-B, every node of O_{min} contains references to the names of all corresponding concepts from the source ontologies which have been merged into it, in the form of a list of concept names. The hierarchical links inside O_{min} indicate how these sets of corresponding concepts are related. In terms of application, the formulation above is intended to provide explicit links between different (types of) image annotations taking into account conceptual inheritance in a manner specified in detail in Section III.

D. Deciding on I_{min} by the Help of a WIG

As we have seen in the previous sub-section, the quantity I_{min} plays a key role in the definition of the minimal ontology of Ω. In the discussion above and as shown in Fig. 1, we have assumed that the sets I_i are of approximately the same cardinalities and I_{min} approximately equals the pair-wise intersection of any two of the sets. This describes the unlikely case when the "contribution" of each of the sets to I_{min} is equal. To handle more general cases, like the one depicted in Figure 2a, we introduce the notion of a weighted intersection graph.

Definition 3: A **weighted intersection graph (WIG)** corresponding to the family of instance-sets $\mathbf{I} = \{I_1, ..., I_N\}$ is a graph $G(\mathbf{I}, E, p)$, where the set $\mathbf{I}$ forms the set of vertices of G, $E = \{\{I_i, I_j\}|I_i \cap I_j \neq \emptyset\}, \forall i, j = 1, ..., N$ is a set of arcs, and $p : E \to \mathbb{R}$ is a function which assigns to each arc $\{I_i, I_j\}$ a weight, p_{ij}.

An example with four sets is shown in Fig. 2b. The weights p_{ij} are intended to quantify the closeness of two intersecting sets of instances. This can be done in terms of (1) the distance of the two populations in a semantic or an euclidean space, $p_{ij} = dist(I_i, I_j)$; or (2) the size of their intersection[2], $p_{ij} = \frac{|I_i \cap I_j|}{|I_i \cup I_j|}$.

The formulation above allows the user to navigate through the WIG corresponding to the instance sets of her ontologies and decide which sets and respective ontologies should be taken into account in the definition of the set I_{min}, the basis for the construction of the minimal ontology. In a certain sense, the WIG representation ensures that no important concepts from the source ontologies are felt out from the definition of the minimal ontology. The advantage of using a WIG, is that it allows to detect and use the structure of the family of instance-sets and the contribution of each set to a potential choice of I_{min}. Several indicators of the degree of importance of an instance set I_i might be helpful as, for instance, the order of every WIG-node, defined as the number of arcs stemming from that node, the sum of the weights of these arcs, or a weighted sum of the two, allowing to give more importance to either the number of ontologies to which a certain set is connected, or the strength of these connections.

As a final remark, we note that there need not necessarily exist a pair-wise intersection between any two instance-sets (see for example the sets and the corresponding WIG in Fig.

---

[2]We recall the reader that by *intersection* is meant an instance-specific quantity, defined for the goals of our application in Section III-B.

3). It might be the case that there exist several concentrations of intersections, or *modules of similarity*, of equal importance, which will be readily indicated by the WIG. This can lead to the definition of the set I_{min} as the union of several disjoint sets (what we had just called similarity modules). In order to create a minimal ontology on this minimal instance-set, we suggest to apply the algorithm described in Section II-C separately on each similarity module. The connections of the different modules are made available again through the WIG representation.

III. APPLICATION TO SEMANTIC IMAGE ANNOTATION

Over the last five years, many multimedia concept ontologies[3] (or more precisely lexica) have been proposed to assist multimedia search and retrieval by providing an effective representation and interpretation of multimedia concepts [28], [27]. These ontologies are also critical resources for the navigation and the exploration of large multimedia collections [8]. They can be classified into three major groups: (1) visual concept hierarchy (or network) inferred from inter-concept visual similarity contexts (e.g. VCNet based on Flickr Distance [36] and the Topic Network of Fan [9]), (2) 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 (e.g. LSCOM [27], multimedia thesaurus [28]) and (3) generic ontologies based on existing semantic concept hierarchies, such as WordNet, populated with annotated images or multimedia documents (e.g. ImageNet [5] and LabelMe [25]). In the following, we propose to use the ontology matching framework presented in Section II to:

- fill the semantic gap by the alignment of multimedia ontologies (second group) with ontologies with high level semantics (third group);
- extract a common reference and study the coherency and the specificity between the different multimedia ontologies.

A. Filling the Semantic Gap

1) Problem Statement: As introduced in Section I, the semantic gap problem implies challenging issues in the field of image retrieval and annotation, among which the building of robust high level concept detectors, of user oriented annotations, and, globally, image annotations with high level semantics.

In this section, we propose an approach to fill the semantic gap by mapping two complementary ontologies: a visual thesaurus and a semantic thesaurus. Contrary to [28] which proposed such a mapping in a manual way, our approach is automatic and generic (not dependent on the ontologies) and based on *visual knowledge* about the instances.

As a first ontology, we chose LSCOM [27] – a resource dedicated to multimedia annotation. It was initially built in

the framework of TRECVID[4], with the following criteria: concept usefulness, concept observability, and feasibility of concept automatic detection. A subset of the LSCOM hierarchy, containing 450 semantic concepts, is populated by the development set of TRECVID 2005 videos, and thus is coherent with the requirements of our approach (c.f. Section II-A).

As a second ontology, we use a subset of WordNet [24] populated with the LabelMe dataset [25] containing 3676 concepts.

Applying our matching approach to automatically link these two ontologies allows to tackle several problems in automatic semantic annotation of images, such as: (1) the semantic enrichment of concepts belonging to a multimedia ontology with high level linguistic concepts from a general and common sense knowledge base; (2) the evaluation of the quality of the baseline concept detectors by studying the coherency between concepts whose semantics is related to their perceptual manifestations and concepts whose semantics is related to common sense.

2) Methodology: In our setting, the instances that extensionally define a concept are images whose annotations contain the name associated to that concept. Choosing an appropriate image representation is, therefore, crucial for our approach. We consider the following options:

1) a bag-of-words vector representation (e.g. tags, metadata, keyword-based annotations);
2) a vector of descriptors (e.g. MPEG-7 descriptors, bag-of-features);
3) a vector built with the responses of each image to a set of baseline detectors (e.g. Columbia374 [37], Mediamill [29] and VIREO-374 [19]).

These image representations are not equal in expressiveness. The bag-of-words model does not allow to benefit from the visual content of images. For the second option, we can use a codebook built on a bag-of-features model and histograms of codewords – nowadays the best approach in the state-of-the-art [19]. The last option assumes we have a bank of visual detectors built for instance with the *bag-of-words* approach and can be seen as a combination of the first two. The detectors used to build the image signature are not mandatorily linked to real semantics (i.e. do not correspond to a concrete object), but, to be efficient, they have to be able to properly discriminate the concepts in O_1 and O_2.

Section II-B gives instructions on how to compute a similarity measure $sim : C_1 \times C_2 \to \mathbb{R}$ for two concepts $A \in O_1$ and $B \in O_2$ by using concept representations based on variables scores (see eq. (1)). To obtain these representations, we need to compute a score per variable and per concept. Considering as variables either words (representation 1), descriptors (representation 2), or detectors (representation 3) allows us to apply the generic variable selection approach described above. To score the variables for a concept A, an SVM is learned on the binary classification training set S^A

(defined in Section II-B), evaluating the capacity of every variable to discriminate the concept A from all the other concepts of its ontology. As an evaluation criterion we use the variation of the VC-dimension parameter per classifier proposed and tested in [31]. The advantage of using this approach is that it allows us to trace back the most important variables (words, codewords from a codebook model or detectors form a set of baseline detectors) that characterize a concept.

The rest of the methodology described in Section II-B is directly applied to get the resulting mappings between LSCOM concepts and WordNet concepts. These alignments can be used to produce a linguistic description of LSCOM concepts (dedicated to the multimedia document annotation) in the vocabulary of WordNet (a lexical ontology). This improves the retrieval process in several ways: (1) through query expansion and 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.

Contrary to [28], our mapping is done in an automatic and visual manner. An enrichment of the annotations (i.e. adding the linguistic description derived from the knowledge of *what is a LSCOM concept in the WordNet ontology*) is achieved via an ontology which has a rich semantic structure (we benefit from all the semantic relations in WordNet, like hypernymy, meronymy, antonymy). Furthermore, considering a step of annotation refinement, the annotation coherency assessment when considering the annotation in its integrity (rather than assessing the relevance of each concept of the annotation one by one) also benefits from the rich semantic structure of WordNet [18].

3) Preliminary Experimental Results: We provide a low-scale evaluation of the suggested matching approach. We chose three concepts from the LSCOM ontology and five concepts from WordNet, respecting several criteria. In the first place, we selected concepts with more than 500 associated instances. For WordNet only, we chose concepts for which we know that distinct visual features (in our case features from the bag-of-features model) are able to discriminate them. Finally, we verified that we have no semantic ambiguity among the chosen concepts in our dataset.

To construct image features, we use a bag-of-features model with a visual codebook (representation 2). The visual codebook is built classically with a K-Means algorithm. The quantification of the extracted SIFT features was investigated in two ways: the building of the codebook is done

1) over all the instances associated to the selected concepts (LSCOM and LabelMe),
2) only over the LabelMe images and a quantification per concept.

The latter is intuitively consistent, because the LabelMe dataset is a sufficiently generic collection of photographs which enables us to properly represent objects from both ontologies. The two experimentations gave very similar results; the results of the experiment based on the first codebook are

TABLE I
LSCOM/TRECVID2005 AGAINST WORDNET/LABELME: AUTOMATIC
CONCEPT MAPPINGS (ABOVE) VS. MANUAL ANNOTATIONS (BELOW).

Concept Names	Man	Car	Boat	TV	House
Natural Disasters	0.37	0.15	-0.33	0.12	0.44
US Flags	0.21	0.09	0.01	0.05	0.05
Single Family Homes	0.20	0.18	-0.36	0.13	0.41
Natural Disasters	103	51	4	0	73
US Flags	434	16	0	2	28
Single Family Homes	205	73	1	0	184

(a)   (b)

Fig. 4.   Two LSCOM images annotated by LSCOM:Natural_Disasters, that can be also effectively annotated by WordNet:Man and WordNet:House

resumed in Table I along with manual annotations of LSCOM instances by WordNet concepts.

We have tested all four similarity measures suggested in [31], all of them yielding competitive outcomes. The results presented here are achieved by using Spearman's correlation coefficient (see eq. (2)). The values in the first matrix, therefore, indicate high similarity for positive values and low similarity for non-positive ones. As we can see, the results are coherent with the data and the sense of the LSCOM concepts. For instance, 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.

In terms of improving the retrieval process, considering our results, we can say that the images annotated by LSCOM:Natural_Disasters could also be queried and annotated (after validation in the image) by the LabelMe concepts WordNet:Man and WordNet:House (Fig. 4).

B. Exploring the Dependencies and Specificities of Various Multimedia Ontologies through Ontology Matching

1) Problem statement: The development of multimedia thesauri and lexicons has generated many open issues related to concept usefulness and concept selection. Many experimental studies have been carried to partially answer these questions [12]. Recently, some authors proposed to build a

lexicon based on concepts with small semantic gaps [22]. Many approaches have also been proposed to study and to exploit inter-concept relationships but mainly in the same ontology [21], [35]. To the best of our knowledge, there is no work dealing with the relationships of concepts between different multimedia ontologies. We propose to study the potential of a WIG representation (introduced in Section II-D) both to build explicit *closeness* relationships between different multimedia ontologies and to extract a minimal multimedia ontology. This ontology can be defined as the set of core concepts for concept-based multimedia retrieval according to the considered collections.

Extracting a minimal ontology from Ω, a set of heterogeneous ontologies (which may contain ontologies of the group (1), (2) or (3) as categorized in the introduction of this section), has many benefits in multimedia indexing. On one hand, a minimal ontology offers the possibility to build an index over annotations written with concepts which share semantics from heterogeneous ontologies, and to make links between ontologies like in [28] or in [14] in an automatic way. On the other hand, the minimal ontology also allows to assess the generality and the specificity of the ontologies.

The method proposed in Section II-C can be directly applied to semantic image annotation; the one point which remains to explicate is the construction of I_{min}.

2) Computing I_{min}: We can expect that, regardless of the image representation, we cannot get exactly the same descriptor values from different instances (that belong to different ontologies) although they share the same concept. Thus, computing I_{min} imposes to use a similarity function and to set a threshold parameter for deciding if the similarity between two instances $\mathbf{i}_1^1 \in O_1$ and $\mathbf{i}_1^2 \in O_2$ is high enough for considering $\{\mathbf{i}_1^1, \mathbf{i}_1^2\} \in I_1 \cap I_2$. Therefore, the presence of an instance $\mathbf{i}$ in I_{min} represents an agreement between the different ontologies, i.e. $\mathbf{i}$ is not strongly isolated in the image representation space, but rather *close to* instances belonging to the other ontologies.

> $I_{min} = \emptyset$
> **foreach** $k \in [1, ..., N]$ **do**
> $\quad$ **foreach** $(i,j)|i \in I^k, j \in I \setminus I^k, j \in kdtree(i, thr)$
> $\quad$ **do**
> $\quad\quad |\quad I_{min} \leftarrow I_{min} \cup \{i\} \cup \{j\}$
> $\quad$ **end**
> **end**

Algorithm 1: k iterates over all the considered ontologies, I^k is the set of instances of the ontology O_k. The procedure $kdtree(i, thr)$ retrieves all the instances contained in a hypersphere of radius thr (empirically set) centered in i.

Depending on the image representation, manipulating the concept instances can lead to working on observations in high dimensional spaces, not necessarily well sampled, where computing I_{min} could become prohibitive[5]. In our low-

---

[5]Considering the first image representation introduced in Section III-A2, computing I_{min} can be performed easily since instances are represented by bag-of-words.

scale application scenario, this is achieved by using instances indexed in a kd-tree and a nearest neighbor search, as presented in Alg. 1.

IV. 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 various efforts, the semantic gap problem is still an important open issue for the semantic understanding of multimedia documents. Recently, many knowledge models have been proposed to improve multimedia retrieval and to take into account the different relationships between semantic concepts in the processing. In [2] and [15], formal models of application domain knowledge are used, through fuzzy description logics, to help and to guide semantic image analysis. Prior knowledge on structured visual knowledge represented by an And-or graph (stochastic grammars) has been proved to be very useful in the context of image parsing or scene recognition in images [39]. While these different models ar e highly integrated in multimedia processing, their main drawback is that they are specific to the application domain. On the contrary, recently, many generic large scale multimedia ontologies or multimedia concept lexicons (see Section III), together with image collections have been proposed, mainly in the context of semantic concept detection and automatic multimedia annotation. These ontologies have proved to be very useful in this context but many problems still remain among which the automatic mapping of visual concepts to high level concepts, the selection strategies of the different concepts according to different criteria: their usefulness [13], their visual discriminative power [38], and the consideration of inter-concept relationships in the processing [34], [8].

The paper proposes to address these problems by an ontology matching method and an extraction of a minimal ontology. An important part of the existing ontology matching approaches, including ours, are characterized as *extensional*, i.e. grounded in the external world, relying on instances in order to judge concept similarity [6] [30] [17]. The procedure for minimal ontology extraction that we suggest is much in line with *module extraction* research, defined as the problem of finding, given a certain sub-vocabulary of an ontology, a minimal sub-structure *within that ontology* that provides the same description of the relationships holding between terms over the sub-vocabulary as the whole ontology [20], [33]. In contrast to related approaches, we base the extraction of the minimal structure on the existence of an *extensional* agreement of *multiple* source ontologies.

V. CONCLUSION AND OPEN ENDS

We have proposed directions for the application of ontology matching techniques to solve different interoperability issues in the area of semantic image annotation and retrieval. In particular, we have addressed two main problems: (1) bridging the semantic gap and (2) extracting a common

reference model for a set of multimedia ontologies. For solving problem (1), we have proposed to apply a generic instance-based ontology matching procedure (developed in a previous study) in order to produce concept-based annotations enriched with lexical descriptions on the concepts, also meant to improve indexing and retrieval. A novel multiple ontology alignment framework has been suggested to solve problem (2).

Many important problems still need to be addressed like, for instance, populating each ontology with existing or built image datasets[6], deciding on an appropriate representation of the instances, solving various complexity issues in the matching process related to the number of concepts and instances in the targeted ontologies. Although our preliminary experimental results are encouraging, the work of implementing and evaluating the propositions of this paper on a larger scale is still in progress.

REFERENCES

[1] S. Coradeschi and A. Saffiotti. An introduction to the anchoring problem. *Robotics and Autonomous Systems*, 43(2-3):85–96, 2003.

[2] S. Dasiopoulou, I. Kompatsiaris, and M.G. Strintzis. Using fuzzy dls to enhance semantic image analysis. In *Semantic Multimedia*, pages 31–46. Springer, 2008.

[3] S. Dasiopoulou, V. Tzouvaras, I. Kompatsiaris, and M.G. Strintzis. Enquiring MPEG-7 based multimedia ontologies. *Multimedia Tools and Applications*, pages 1–40, 2010.

[4] R. Datta, J. Li, and J.Z. Wang. Content-based image retrieval: approaches and trends of the new age. In *SIGMM MIR-Workshop*, pages 253–262. ACM, 2005.

[5] 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.

[6] A. Doan, J. Madhavan, P. Domingos, and A. Halevy. Learning to map between ontologies on the semantic web. In *WWW'02*, pages 662–673. ACM Press, 2002.

[7] J. Euzenat and P. Shvaiko. *Ontology Matching*. Springer-Verlag, 1 edition, 2007.

[8] J. Fan, Y. Gao, and H. Luo. Integrating concept ontology and multitask learning to achieve more effective classifier training for multilevel image annotation. *IEEE Trans. on Im. Proc.*, 17(3):407 – 426, 2008.

[9] 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.

[10] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *JMLR*, 3(1):1157–1182, 2003.

[11] S. Harnad. The symbol grounding problem. *Physica d*, 42(1-3):335–346, 1990.

[12] A. Hauptmann, R. Yan, W. Lin, M. Christel, and H. Wactlar. Can high-level concepts fill the semantic gap in video retrieval? a case study with broadcast news. *IEEE Trans. on Multimedia*, 9(5):958 – 966, 2007.

[13] A. Hauptmann, R. Yan, and W.-H. Lin. How many high-level concepts will fill the semantic gap in news video retrieval? In *CIVR'07*, pages 627–634. ACM, 2007.

[14] L. Hollink, A.T. Schreiber, J. Wielemaker, and B. Wielinga. Semantic annotation of image collections. In *Knowledge Capture*, pages 41–48, 2003.

[15] C. Hudelot, J. Atif, and I. Bloch. Fuzzy Spatial Relation Ontology for Image Interpretation. *Fuzzy Sets and Systems*, 159:1929–1951, 2008.

[16] C. Hudelot, N. Maillot, and M. Thonnat. Symbol grounding for semantic image interpretation: from image data to semantics. In *SKCV-Workshop, ICCV, Beijing, China*, 2005.

[17] 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.

[18] N. James and C. Hudelot. Towards semantic image annotation with keyword disambiguation using semantic and visual knowledge. In *The IJCAI workshop CIAM 2009*, pages 35–40, 2009.

[19] 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.

[20] R. Kontchakov, L. Pulina, U. Sattler, T. Schneider, P. Selmer, F. Wolter, and M. Zakharyaschev. Minimal module extraction from dl-lite ontologies using qbf solvers. In *IJCAI*, pages 836–841, 2009.

[21] M. Koskela, A.F. Smeaton, and J. Laaksonen. Measuring concept similarities in multimedia ontologies: Analysis and evaluations. *IEEE Trans. on Multimedia*, 9(5):912 – 922, 2007.

[22] Y. Lu, L. Zhang, Q. Tian, and W.Y. Ma. What are the high-level concepts with small semantic gaps. In *IEEE Conf. on Comp. Vis. and Patt. Rec.*, pages 1 – 8, 2008.

[23] N. Maillot, M. Thonnat, and C. Hudelot. Ontology based object learning and recognition: Application to image retrieval. In *The 16th IEEE ICTAI*, pages 620–625, 2004.

[24] G.A. Miller. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.

[25] 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.

[26] 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.

[27] J.R. Smith and S.F. Chang. Large-scale concept ontology for multimedia. *IEEE Multimedia*, 13(3):86–91, 2006.

[28] 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.

[29] C.G.M. Snoek, M. Worring, J.C. Van Gemert, J.M. Geusebroek, and A.W.M. Smeulders. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *MM'06*, pages 421–430. ACM, 2006.

[30] G. Stumme and A. Maedche. Fca-merge: Bottom-up merging of ontologies. In *IJCAI*, pages 225–230, 2001.

[31] 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.

[32] C. Town. Ontological inference for image and video analysis. *Mach. Vis. and Appl.*, 17(2):94–115, 2006.

[33] N. N. Tun and J. S. Dong. Ontology generation through the fusion of partial reuse and relation extraction. In *KR*, pages 318–328, 2008.

[34] Xiao-Yong Wei, Yu-Gang Jiang, and Chong-Wah Ngo. Exploring inter-concept relationship with context space for semantic video indexing. In *CIVR'09*, pages 1–8. ACM, 2009.

[35] Xiao-Yong Wei and Chong-Wah Ngo. Fusing semantics, observability, reliability and diversity of concept detectors for video search. In *MM'08*, pages 81–90. ACM, 2008.

[36] Lei Wu, Xian-Sheng Hua, Nenghai Yu, Wei-Ying Ma, and Shipeng Li. Flickr distance. In *MM'08*, pages 31–40. ACM, 2008.

[37] A. Yanagawa, S.F. Chang, L. Kennedy, and W. Hsu. Columbia university baseline detectors for 374 lscom semantic visual concepts. Technical report, 2007.

[38] Keiji Yanai and Kobus Barnard. Image region entropy: a measure of "visualness" of web images associated with one concept. In *MULTIMEDIA'05*, pages 419–422. ACM, 2005.

[39] B. Yao, X. Yang, L. Lin, M.W. Lee, and S.C. Zhu. I2t: Image parsing to text description. *IEEE Proc. Special Issue on Internet Vision (To appear)*.

---

[6]Note that in the current state of affairs, only a small part of the ontologies is populated.