Multimedia ontology matching by using visual and textual modalities

Nicolas James, Céline Hudelot, Konstantin Todorov

Combining Visual and Textual Modalities for Multimedia Ontology Matching

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. Multimedia search and retrieval are considerably improved by providing explicit meaning to visual content by the help of ontologies. Several multimedia ontologies have been proposed recently as suitable knowledge models to narrow the well known semantic gap and to enable the semantic interpretation of images. Since these ontologies have been created in different application contexts, establishing links between them, a task known as ontology matching, promises to fully unlock their potential in support of multimedia search and retrieval. This paper proposes and compares empirically two extensional ontology matching techniques applied to an important semantic image retrieval issue: automatically associating common-sense knowledge to multimedia concepts. First, we extend a previously introduced matching approach to use both textual and visual knowledge. In addition, a novel matching technique based on a multimodal graph is proposed. We argue that the textual and visual modalities are to be seen as complementary rather than as exclusive means to improve the efficiency of the application of an ontology matching procedure in the multimedia domain. An experimental evaluation is included.

1 Introduction

In recent years, many research efforts have been directed towards the problem of improving search and retrieval in large image collections by providing semantic annotations in a fully automatic manner. Ideally, semantic image annotation results in a linguistic description of an image, which, in the current state of affairs, is often only related to perceptual manifestations of semantics. Indeed, most of the existing approaches are based on the automatic detection of semantic concepts from low level features with machine learning techniques. Nevertheless, as explained in [10], the image semantics cannot be considered as being included explicitly and exclusively in the image itself. It rather depends on prior knowledge and on the context of use of the visual information. In consequence, explicit formal knowledge bodies (ontologies) have been growingly used to relate semantics and images. Their application in the multimedia domain aims at improving image search and retrieval by providing high-level semantics to visual content, thus facilitating the interface between human and artificial agents and narrowing the well-known *semantic gap* between *low-level visual features* and *high-level meaning* [19].

However, differences in the scopes and purposes of these ontologies (reviewed in the following section) as well as in their application contexts tend to result in various heterogeneities on terminological, conceptual and / or semantic level. Therefore, relating these knowledge resources, a process termed as *ontology matching* [5], is crucial in order to fully unlock their potential in support of multimedia search and retrieval - a field in which ontology matching has found little application in contrast to its use in various semantic web applications. To accomplish an ontology matching task one could rely on the instances contained in the ontologies concepts (*extensional* or *instance-based* matching), make use of the relations that hold between the different concepts (*structural* matching), measure the similarities of the concept names and their lexical definitions (*terminological* matching), etc. In the case of multimedia ontologies, which often come equipped with sets of annotated images, extensional matching is a suitable paradigm since it enables the benefit from both the visual and the textual knowledge.

This paper considers two generic instance-based ontology matching techniques - one based on variable selection (developed in a previous study of the same authors [12], [23]) and another, novel approach, exploring the benefits of discovering correlations in a multimodal graph. We apply and compare these approaches in the context of an important semantic image retrieval problem: associating common-sense knowledge to multimedia concepts. In particular, the paper proposes to narrow the semantic gap by matching a common sense ontology (WordNet [15] associated to the image database LabelMe[18]) with a specific multimedia ontology (LSCOM [20] associated to the TRECVID2005 development data set).

Since our matching approaches rely on extensional information, it is important to explore and make use of all possible instance-knowledge that can be made available. In extensional terms, these two resources can be considered as bi-modal, each possessing a *visual* and a *textual* modality. On one hand, the concepts of these ontologies serve to annotate a given set of images which can be considered as instances of these concepts. On the other hand, every image can be assigned a text document by taking the concepts that it is annotated by and their corresponding textual definitions (LSCOM definitions or WordNet glosses). In order to apply the suggested matching approaches, one can rely on either of the two modalities and we will refer to the two resulting types of matching as, respectively, *visual matching* and *textual matching*. What the paper investigates more closely, are the benefits of using both in combination, instead of each of them in isolation. The variable selection model is able to work with only one modality at a time and an integration of the results have to be performed *post factum*. Since we are primarily interested in obtaining concept correspondences based on the visual characteristics of the images in the two datasets, we will rely on the visual modality to produce a baseline matching, which will be later adjusted and refined by the help of the textual modality. A potential advantage of the graph-based model is that it allows the simultaneous, *built-in* application of the two modalities in the matching process.

The rest of the article is structured as follows. Next section reviews related work. Section 3 provides a summary of the generic instance-based ontology matching techniques that we use. The application of these methods on visual and textual instances is described in detail in Section 4: we report experimental results of these matchings and discuss the benefits of their integrated interpretation. We conclude in Section 5.

2 Related Work

Despite many recent efforts to provide approaches for automatic annotation of images with high-level concepts [11], the semantic gap problem is still an issue for the understanding of the *meaning* of multimedia documents. In this context, many knowledge models or ontologies have been proposed to improve multimedia retrieval and interpretation by the explicit modeling of the different relationships between semantic concepts.

In particular, many generic large scale multimedia ontologies or multimedia concept lexicons together with image collections have been proposed to provide an effective representation and interpretation of multimedia concepts [21, 20, 3]. We propose to classify these ontologies in four major groups: (1) semantic web multimedia ontologies often based on MPEG-7 (a review can be found in [3]) (2) visual concept hierarchies (or networks) inferred from inter-concept visual similarity contexts (among which VCNet based on Flickr Distance [26] and the Topic Network of Fan [6]), (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 concept of multimedia documents (LSCOM [20], multimedia thesauri [22], [21]) and (4) generic ontologies based on existing semantic concept hierarchy such as WordNet and populated with annotated images or multimedia documents (ImageNet [4], LabelMe [18]).

The reasoning power of ontological models has also been used for semantic image interpretation. In [2],[9] and [17], 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 [28]. While these different formal models are highly integrated in multimedia processing, their main drawback is that they are specific to the application domain.

All 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 among which the interoperability issue between visual concepts and high level concepts. To solve this issue, some ontology-based infrastructures have also been proposed to guide image annotation [1]. These infrastructures are mainly based on different ontologies (multimedia ontologies, application domain ontologies and top ontologies for interoperability purposes) and the link between the different ontologies is often done manually. In [21], the authors also propose

to build a multimedia thesaurus by linking manually 101 multimedia concepts with WordNet synsets.

Due to the fact that these large scale multimedia ontologies are often dedicated to (or initially built for) particular needs or a particular application, they often tend to exhibit a certain heterogeneity which allows their use as complementary knowledge sources. For instance LSCOM was built for video news annotation purposes, while the scope of WordNet/LabelMe is rather general and common-sense. Hence, these ontologies differ both in their conceptual content (number, granularity and genericity of the concepts) and in their usage (LSCOM is dedicated to multimedia annotation and therefore the extensional and terminological knowledge that it assigns to each concept is defined by the visual occurrences of this concept). While studies have been done to analyze the different inter-ontology concept similarities in different multimedia ontologies [13], to the best of our knowledge, there are no approaches in the state of the art which propose a cross analysis and a joint use of these different and complementary resources.

3 Instance-based Ontology Matching

We propose a methodology to narrow the semantic gap by matching two complementary resources: a *visual* ontology and a *semantic* thesaurus. Contrarily to [21], we suggest to accomplish this matching in an automatic manner. We apply a generic extensional ontology matching approach based on discovering cross-ontology concept similarities via variable selection, which has been previously introduced for matching textually populated ontologies [23]. In [12], we propose a first extension of this approach based on visual extensional knowledge. In the framework of this paper, the approach has been extended to use both *textual* and *visual* knowledge with the objective to combine both in the concept alignment process. In addition, we suggest a novel matching technique, based on a multi-model graph and a random walk with restart (RWR). In the sequel, we will describe the main elements of these approaches in a generic manner, by referring to an abstract notion of *instance*, without specifying whether it comprises a text or a multimedia document and how precisely it is represented. We only assume that each instance is indeed representable as a real-valued vector. We start by giving several assumptions and definitions.

An ontology is based on a set of *concepts* and *relations* defined on these concepts, which describe in an explicit and formal manner the knowledge about a given domain of interest. In this paper, we are particularly interested in ontologies, whose concepts come equipped with a set of associated instances, referred to as **populated ontologies** and defined as tuples of the kind $O = \{C, \texttt{is\_a}, R, I, g\}$, where C is a set whose elements are called concepts, $\texttt{is\_a}$ is a partial order on C, R is a set of other 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. In this way, a concept is *intensionally* modeled by its relations to other concepts,

and *extensionally* by a set of instances assigned to it via the mapping g. By assumption, every instance can be represented as an n dimensional real-valued vector, defined by n input **variables** of some kind (the same for all the instances in I).

To build a procedure for ontology matching, we need to be able to measure the **pair-wise similarity of concepts**. The measures used in the current study are based on *variable selection* (Section 3.1) and on correlations discovered by a random walk in a *mixed multimedia graph* (Section 3.2).

3.1 Variable Selection-Based Method (VSBM)

Variable selection [7] is defined as a procedure for assigning *ranks* to the input variables with respect to their importance for the output, a ranking criterion provided. On these bases, we propose to evaluate concept similarity by comparing the ranks assigned to the input variables w.r.t. two given concepts.

We define a binary 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 (i.e. evaluating the importance of the input variables w.r.t. the concept c), we obtain a representation of the concept c as a list

$$L(c) = (r_1^c, r_2^c, ..., r_n^c), \tag{1}$$

where r_i^c is the rank associated to the ith variable. To compute a rank per variable and per concept, we apply a standard *Point-wise Mutual Information* criterion approximated for a variable v_i and a concept c by

$$r_i^c = PMI(v_i, c) = \log \frac{A \times |I|}{(A + C) \times (A + B)}, \tag{2}$$

where A is the number of co-occurrences of v_i and c, B is the number of times v_i occurs without c, C is the number of times c occurs without v_i and $|\cdot|$ stands for set cardinality [27].

Given two source ontologies O and O', a representation as the one in (1) is made available by following the described procedure for every concept of each of these ontologies. The similarity of two concepts, $c \in O$ and $c' \in O'$ is then measured in terms of their corresponding representations $L(c)$ and $L(c')$. Several choices of a similarity measure based on these representations are proposed and compared in [23]. In the experimental work contained in this paper, we have used Spearman's measure of correlation and the n'-TF similarity measure. Spearman's coefficient is given by

$$s_{Spear}(c, c') = 1 - 6\frac{\sum_i (r_i^c - r_i^{c'})^2}{n(n^2 - 1)}. \tag{3}$$

The n'-TF (n' Top Features) simply measures the size of the intersection of the subsets of the $n' < n$ top variables (i.e. the ones with highest ranks) according

to the lists $L^{vars}(c)$ and $L^{vars}(c')$:

$$s_{n'TF}(c,c') = \frac{|\{v_{i_1},...,v_{i_{n'}}\} \cap \{v_{j_1},...,v_{j_{n'}}\}|}{n'}, \tag{4}$$

where $v^c_{i_p}$ stands for the variable which has a rank $r^c_{i_p}$ and $v^c_{j_q}$ - for the variable which has a rank $r^{c'}_{j_q}$

In the sequel, we will be interested in applying the measures above in order to represent a concept c from ontology O by a list of pairs of the kind (s_i, c'_i), where s_i (a shortcut for $s_i(c, c'_i)$) is the similarity score (issued from either measure (3) or (4)) of the concept c and the concept $c'_i \in O'$, $i = 1, ..., k$ with k the cardinality of the concept set of O'. We will denote the list of such pairs corresponding to the concept $c \in O$ by

$$L^{sim}(c) = \{(s_1, c'_1), ..., (s_k, c'_k)\}. \tag{5}$$

Provided the choice of a threshold $k' \leq k$, we will define for each concept $c \in O$ the *matching* $L^{sim}_{k'}(c)$ by keeping only those k' concepts from O' which have the highest similarity scores with respect to c. An *alignment* of O to O' will be defined as the set of matchings $A(O, O') = \{L^{sim}_{k'}(c_j)\}^l_{j=1}$, where l is the cardinality of the concept set of O.

3.2 Graph-based Method (GBM)

Graph-based procedures are well-known approaches for evaluating the similarity between objects, like our concepts. These approaches have been used in several domains: ranking algorithms for information retrieval [8], automatic image annotation [16], [25], data analysis and word sense disambiguation [14]. The idea is to exploit the relationships between objects and the different aspects of these objects. In our instance-based ontology matching framework, we have objects of different kinds: (1) concepts, (2) concept instances (i.e. images), and (3) features relevant to the images. We use a method based on the Mixed-Multimedia Graph (*MMG*) and the Random Walks with Restarts algorithm proposed in [16]. Fig. 1 represents a special case of a MMG in the scope of our concept matching procedure.

The MMG graph is well adapted for multimedia document processing because it allows to mix heterogeneous kinds of information, like illustrated in Fig. 1. For each instance we have: (1) a concept node, (2) a textual representation of the instance and (3) a visual representation of the instance. For our experiments, the textual representation is based on a bag-of-words model built from the textual definition of all the concepts associated to an instance (the instances are multi-annotated), and the visual representation is based on histograms of visual words computed over the instances. The kind of the instance representation is seen as a modality, the graph is modular and a modality can be easily used or not. As we will see in the experiments in Section 4.1, we have used both the uni-modal and the bi-modal versions of the graph. Regardless to the chosen modality, the graph is completed with Nearest Neighbor (NN) links between the nodes of each

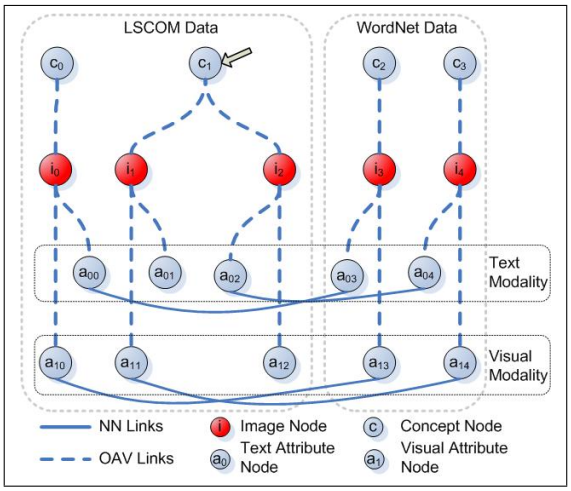

Fig. 1. An MMG Graph

modality. The similarity function needed to compute these links depends on the modality type, therefore we need a function for computing the similarity between two textual representations and another function for computing the similarity between two visual representations.

The process of discovering concept similarities consists in finding correlations between a specific concept of the ontology O and the concepts of the ontology O'. We use the RWR algorithm as described in [16]. The random walk starts at a fixed concept node - the one for which we search similar concepts in the ontology O' (see the concept c_1 in Fig. 1). At each step, the walker can either choose a link in the set of associated links to the node on which it is, or go to its starting point with a probability p (experimentally set; in [16] $p \in [0.8, 0.9]$ brings good results). A precise description with implementation details of the algorithm can be found in [24].

The probability that the walker is at node c', called the *steady state probability*, $\mu_{c_1}(c')$, can be interpreted like an affinity measure between the node c_1 and c'. Therefore, if we consider the results only for concept nodes, a high similarity between $c_1 \in O$, and $c' \in O'$ is observed when the probability $\mu_{c_1}(c')$ is high.

4 Aligning Two Multimedia Resources

The ontology matching techniques described above can be applied for any two ontologies whose concepts are used to label a set of real-world instances of some kind. Based on these techniques, we will align two complementary multimedia knowledge resources by using and integrating the visual and textual modalities of their extensions.

We chose, on one hand, LSCOM [20] – an ontology dedicated to multimedia annotation. It was initially built in the framework of TRECVID[1] with the criteria of concept usefulness, concept observability and feasibility of concept automatic detection. LSCOM is populated with the development set of TRECVID 2005 videos. On the other hand, we used WordNet [15] populated with the LabelMe dataset [18].

4.1 Experimental Setting

In our experimental work, we have used a part of the LSCOM ontology, LSCOM_Annotation_v1.0[2], 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 chosen LSCOM subpart is particularly adapted to annotate this kind of content, thus contains abstract and specific concepts (e.g. SCIENCE_TECHNOLOGY, INTERVIEW_ON_LOCATION). To the contrary, our subontology 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 and contains concepts such as CAR, COMPUTER, PERSON, etc.

To provide a low-scale evaluation of the suggested approach, we chose five concepts from the LSCOM ontology (BUS, COMPUTER, PEDESTRIAN_ZONE, SPEAKER_AT_PODIUM, SPORT) populated with 2317 images, and thirteen concepts from the WordNet ontology (ARM, CAR, GRASS, HEAD, LEG, PERSON, PLANT, PLATE, ROAD, SIDEWALK, TORSO, TREE, WHEEL) populated with 4964 images. The choice of the selected concepts was made on the basis of several criteria: (1) the number of associated instances, (2) the lack of semantic ambiguity in our dataset for every selected concept, (3) for WordNet only: a high confidence (arbitrarily decided) in the discrimination of the concept using only perceptual information, (4) the presence of contextually bounded cross-ontology concepts (such as BUS and CAR) as well as contextually isolated concepts (i.e. dissimilar to all the other concepts such as PLATE).

We draw the readers attention to the fact that the similarities of the concepts should be interpreted strictly within the extensional nature of their definitions and not in terms of any possible intuitive or common sense definition. Our methods imply that two concepts are similar if their corresponding instances contain similar visual or textual characteristics (i.e. the objects that correspond to two similar concepts co-occur in the instances corresponding to these concepts). In some cases, these similarity scores are in agreement with the common sense, but they are not in other cases. In that line of thought, taking two concepts (one from each of the ontologies) with identical names (e.g. BUS in WordNet and BUS in LSCOM) is not relevant for evaluating the quality of the alignments.

In the remainder of the section, we will first present and discuss results from the visual and textual matchings of the selected sets of concepts separately. We

[1] http://www-nlpir.nist.gov/projects/tv2005/
[2] http://www.ee.columbia.edu/ln/dvmm/lscom/

will further propose a method to integrate the two matching types. As matching procedures we have used the VSBM method with two different concept similarity measures - Spearman's correlation and the n'-TF measure[3]. Additionally, we have tested the GBM approach by using either only the visual or only the textual modality and by using both modalities simultaneously. This results in three independent alignments per matching type which, to improve readability, are all gathered at the end of the paper.

4.2 Visual Matching

Instances and Representation To construct image features, we use a bag-of-words model with a visual codebook of size 900, built classically using the well known SIFT descriptor and a K-Means algorithm. The quantification of the extracted SIFT features was done over all the instances associated to the selected concepts (from both LSCOM and WordNet) by using only the distinct objects in each image instead of the entire image in order to extract the SIFT features. The variables which describe an image are then the bins of the histograms of codewords corresponding to this image.

Results and Discussion The results of matching the 5 LSCOM concepts against the 13 WordNet concepts by following the variable selection-based matching procedure described above are presented in Tables 1 and 2 and the results from the GBM method are shown in Table 3. As introduced in (5), we provide for every LSCOM concept (in the top row) a list of pairs (score, WordNet concept) in a descending order of their importance with respect to this concept. The scores in Table 1 and Table 2 are issued from applying the similarity measures (3) and (4), respectively, whereas the scores in Table 3 are correlations found by the help of the graph-based method.

As a general tendency, the WordNet concepts PERSON and HEAD tend to appear up in the lists, whereas the concept PLATE achieves mostly low scores. These results are coherent with the nature of our data, since the concept PLATE stands alone in our selection of concepts (does not have a match), whereas the concepts PERSON and HEAD are highly relevant for the TRECVID dataset, containing shots from news videos where often we have a presenter or a speaker. For Table 3, some remarks about the graph construction have to be taken in account. The similarity used to compute the nearest-neighbor links is a Minkowski distance. Due to the nature of the LSCOM/TRECVID data (the images are visually very close to one another within TRECVID), we have taken into account only NN links from ontology O to O' (or vice-versa) in order to get a well connected graph. Without this constraint, we would have a graph with two disconnected components. However, this explains the results in Table 3 where the top 5 concepts are the same in all lists (subject to a permutation).

We observe examples of a lack of coherence between intuitive interpretations and achieved matchings as discussed previously in this section. For example, SIDEWALK w.r.t. COMPUTER is intuitively an erroneous matching in contrast

[3] Pearson's measure, also discussed in [23] showed to compete closely with Spearman's.

to LEG and PERSON w.r.t. SPEAKER_AT_PODIUM which is intuitively coherent (Table 1 and 2).We note that this is a perceptually induced conceptual mismatch, i.e. a bias, which is due to co-occurrences of *visual* objects within the instances of both concepts. In our example, images of SIDEWALK tend to contain the object *person*, so do images of COMPUTER, although COMPUTER and SIDEWALK are unrelated. In order to account for this problem, we suggest that a post-processing of the obtained matchings has to be performed with the objective to re-rank the WordNet concepts w.r.t. their importance for the respective LSCOM concept. To these ends, we perform a textual matching with the objective to complement the results achieved by the visual matching and filter out undesired alignments.

4.3 Textual Matching

Instances and Representation We present the results of the matching of the two selected sets of concepts, by using this time as instances textual documents, relevant to these concepts. A text document has been generated for every image, by taking the names of all concepts that an image contains in its annotation, as well as the (textual) definitions of these concepts (the LSCOM definitions for TRECVID images or the WordNet glosses for LabelMe images). An example is shown in Fig. 2. After a phase of standard text-processing (lemmatization and stop-word filtering), a vocabulary of size 544 has been constructed for the corpus containing the documents generated as instances for the two ontologies. Every document is standardly represented as a *tf-idf* vector of dimension 544.

107_Standing One or more people standing up. 227_Bus Shots of a bus. 224_Outdoor Shots of Outdoor locations. 217_Person Shots depicting a person. The face may be partially visible. 202_Crowd Shots depicting a crowd. 181_Adult Shots showing a person over the age of 18. 104_Male_Person One or more male persons. 290_Daytime_Outdoor shots that take place outdoors during the day. 316_Group We defined a group as 3-10 people. 109_Windows An opening in the wall or roof of a building or vehicle fitted with glass or other transparent material.

Fig. 2. The LSCOM concept BUS: a visual and a textual instance.

Results and Discussion To derive the textual similarity scores, we have applied the same procedures as those applied for the visual matching. For the VSBM matching, we first scored the variables by the help of a mutual information-based variable selection and then measured concept similarities by the help of Spearman's measure of correlation and n'-TF (equations (3) and (4)). We note that in this case the variables that define our instances are actual (lemmatized) terms that appear in the corpus with a certain (sufficiently high) frequency. The results of these matchings are presented in an analogous manner to the visual matching in Tables 4 and 5. The similarity scores achieved by applying the graph-based approach on our textual data are found in Table 6.

We observe that, while some of the correlations already found by the visual matching are confirmed (e.g. the low scores of the stand-alone concept PLATE), some of the WordNet concepts achieve different scores through the textual matching (e.g. the problematic SIDEWALK). This confirms the initial hypothesis that the two matching modalities are complementary and neither is to be applied self-dependently. The integration and the proper interpretation of the results of the two matching types is the subject of the following sub-section.

4.4 Integration of the Textual and the Visual Modalities

***A posteriori* pruning with VSBM** As a result of the matchings, every WordNet concept is assigned two scores w.r.t. each LSCOM concept - a visual similarity score s_v^i and a textual similarity score s_t^i.

In order to identify problematic matchings (e.g. (COMPUTER, SIDEWALK)), we propose an algorithm which serves to prune the list of most important (w.r.t. a given LSCOM concept) WordNet concepts. We compute for every LSCOM concept the quantities $s_\delta^i = |s_t^i - s_v^i|$, $\forall i = 1, ..., k$, (k is here the number of WordNet concepts) represented in Tables 7 and 8. The scores s_t^i and s_v^i are integers corresponding to the (real) similarity scores. When multiple consecutive concepts achieve identical scores (a likely case when applying the n'-TF measure) the same rank is attributed to each of these concepts. We take as a basis the matching achieved by using the visual modality and we fix a number k' of concepts to be kept. Our algorithm relies on the heuristics that the WordNet concepts c_i' for which the corresponding s_δ^i is too large (w.r.t. an experimentally set threshold) should be identified as subjective to removal. The list $L_{k'}^{sim}(c)$ is pruned by removing from it all WordNet concepts with too large a s_δ. By applying this algorithm on the results in Table 7 (fixing k' at 4), we are able to prune out some problematic concepts, such as the WordNet concept SIDEWALK w.r.t. the LSCOM concepts COMPUTER, SPEAKER_AT_PODIUM and SPORTS, the WordNet concept ROAD with respect to the LSCOM concept COMPUTER, or the WordNet concepts PLANT and WHEEL with respect to the LSCOM concept PEDESTRIAN_ZONE. Similar results are achieved based on the results obtained by the n'-TF measure (Table 8).

Built-in bi-modality matching with GBM Table 9 contains the results of the built-in bi-modality matching by the GBM approach in which the two modalities have been used as an integral part of the matching process. As we can see, the obtained results are in general coherent, although less performant as compared to the VSBM approach. The reasons for the flawed similarity values produced by this matching can be sought for, first of all, in the fact that we have too low a number of concepts resulting in too little number of nodes in the multi-modal graph, which in turn decreases the probability of discovering interesting matches. In addition, the graph has been constructed in a manner which does not allow that an image node is connected to more than one concept node (mono-annotation), which leads to a loss of co-occurrence information. The performance of the matching procedure can be significantly improved by lifting

this constraint and adding additional edges to the graph. An overall advantage of this method is the computational time of the RWR algorithm and the multi-modality which allows the concepts to be populated by documents of different types. These two points make this method very promising for a matching at a larger scale.

5 Conclusions

The problem of associating high level meaning to a set of visual concepts has been situated in an ontology matching framework. We have proposed and compared two generic matching techniques - one based on a variable selection method (VSBM) and one based on a random walk in a graph (GBM) by relying on instances of both *visual* and *textual* nature. We have demonstrated that these two extensional modalities are complementary and their combined use improves the achieved results. Although for the moment the VSBM outperforms the GBM approach, the full potential of the latter method is to be uncovered in a large-scale application which is a subject of near future work.

Due to the bias of the data and to the difficulty to extract the concrete semantics of a correlation, a quantitative measure of the efficiency of the approach is difficult to give. An evaluation of the approach can be envisaged within a concrete application context in an information access framework.

The achieved alignments allow for the semantic enrichment of concepts belonging to a multimedia ontology (LSCOM) with high level linguistic concepts from a general and common sense knowledge base (WordNet). This alignment could be used to build a linguistic description of the concepts of LSCOM and improve the retrieval process through: (a) query expansion and reformulation, i.e. retrieving documents annotated with concepts from an ontology O using a query composed of concepts of an ontology O', and (b) a better description of the documents in the indexing process.

References

1. T. Athanasiadis, V. Tzouvaras, K. Petridis, F. Precioso, Y. Avrithis, and Y. Kompatsiaris. Using a multimedia ontology infrastructure for semantic annotation of multimedia content. In *SemAnnot'05*, 2005.
2. S. Dasiopoulou, I. Kompatsiaris, and M. 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. Strintzis. Enquiring MPEG-7 based multimedia ontologies. *MM Tools and Appls*, pages 1–40, 2010.
4. J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. ImageNet: a large-scale hierarchical image database. In *CVPR*, pages 710–719, 2009.
5. J. Euzenat and P. Shvaiko. *Ontology Matching*. Springer-Verlag, 1 edition, 2007.
6. 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.

7. I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *JMLR*, 3(1):1157–1182, 2003.
8. T. Haveliwala. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *IEEE transactions on knowledge and data engineering*, pages 784–796, 2003.
9. C. Hudelot, J. Atif, and I. Bloch. Fuzzy Spatial Relation Ontology for Image Interpretation. *Fuzzy Sets and Systems*, 159:1929–1951, 2008.
10. C. Hudelot, N. Maillot, and M. Thonnat. Symbol grounding for semantic image interpretation: from image data to semantics. In *SKCV-Workshop, ICCV*, 2005.
11. M. Inoue. On the need for annotation-based image retrieval. In *Proceedings of the Workshop on Information Retrieval in Context (IRiX), Sheffield, UK*, pages 44–46, 2004.
12. N. James, K. Todorov, and C. Hudelot. Ontology matching for the semantic annotation of images. In *FUZZ-IEEE*. IEEE Computer Society Press, 2010.
13. M. Koskela and A. Smeaton. An empirical study of inter-concept similarities in multimedia ontologies. In *CIVR'07*, pages 464–471. ACM, 2007.
14. R. Mihalcea, P. Tarau, and E. Figa. Pagerank on semantic networks, with application to word sense disambiguation. In *ICCL*, page 1126. Association for Computational Linguistics, 2004.
15. G. Miller. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.
16. J. Pan, H. Yang, C. Faloutsos, and P. Duygulu. Automatic multimedia cross-modal correlation discovery. In *ACM SIGKDD*, page 658. ACM, 2004.
17. I. S. E. Peraldi, A. Kaya, and R. Möller. Formalizing multimedia interpretation based on abduction over description logic aboxes. In *Description Logics*, 2009.
18. B. Russell, A. Torralba, K. Murphy, and W. Freeman. LabelMe: a database and web-based tool for image annotation. *IJCV*, 77(1):157–173, 2008.
19. A. 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.
20. J. Smith and S. Chang. Large-scale concept ontology for multimedia. *IEEE Multimedia*, 13(3):86–91, 2006.
21. C. 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.
22. R. Tansley. The multimedia thesaurus: An aid for multimedia information retrieval and navigation. Master's thesis, 1998.
23. 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.
24. H. Tong, C. Faloutsos, and J.-Y. Pan. Fast random walk with restart and its applications. In *ICDM '06*, pages 613–622, Washington, DC, USA, 2006. IEEE Computer Society.
25. C. Wang, F. Jing, L. Zhang, and H. Zhang. Image annotation refinement using random walk with restarts. In *ACM MM*, page 650, 2006.
26. L. Wu, X.-S. Hua, N. Yu, W.-Y. Ma, and S. Li. Flickr distance. In *MM'08*, pages 31–40. ACM, 2008.
27. Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In *Fourteenth ICML*, pages 412–420. Morgan Kaufmann Publishers, 1997.

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

Bus	**Computer**	**Ped. Zone**	**Speaker At Pod.**	**Sport**
0.602 head	0.646 person	0.726 plant	0.594 person	0.631 head
0.598 road	0.643 sidewalk	0.709 grass	0.532 head	0.613 sidewalk
0.588 car	0.636 head	0.694 wheel	0.522 grass	0.607 person
0.587 person	0.565 road	0.687 tree	0.521 plant	0.607 car
0.584 wheel	0.495 car	0.617 arm	0.503 sidewalk	0.580 road
0.581 arm	0.467 arm	0.575 leg	0.481 road	0.555 arm
0.570 tree	0.427 wheel	0.567 car	0.475 wheel	0.505 wheel
0.557 sidewalk	0.422 leg	0.478 road	0.468 tree	0.504 tree
0.552 grass	0.411 grass	0.477 sidewalk	0.383 arm	0.454 leg
0.542 plant	0.408 tree	0.467 torso	0.363 car	0.444 torso
0.509 leg	0.406 plant	0.440 person	0.341 leg	0.426 plant
0.460 torso	0.388 torso	0.413 head	0.233 torso	0.399 grass
0.336 plate	0.204 plate	0.117 plate	0.188 plate	0.320 plate

Table 1. Visual VSBM matching with Spearman's correlation measure (eq. (3)).

Bus	**Computer**	**Ped. Zone**	**Speaker At Pod.**	**Sport**
0.325 person	0.400 head	0.600 grass	0.456 person	0.306 road
0.318 grass	0.375 person	0.531 plant	0.400 grass	0.300 head
0.318 road	0.318 sidewalk	0.475 tree	0.368 head	0.300 person
0.293 head	0.306 road	0.456 wheel	0.350 plant	0.281 sidewalk
0.268 plant	0.231 torso	0.275 leg	0.325 road	0.250 car
0.268 tree	0.225 leg	0.250 plate	0.281 sidewalk	0.212 leg
0.243 sidewalk	0.193 grass	0.212 arm	0.225 tree	0.193 arm
0.237 wheel	0.193 plant	0.162 person	0.225 wheel	0.175 torso
0.218 torso	0.118 car	0.137 car	0.162 leg	0.168 plate
0.206 leg	0.112 wheel	0.137 torso	0.106 car	0.156 plant
0.150 arm	0.100 tree	0.112 head	0.087 arm	0.143 wheel
0.150 car	0.093 arm	0.112 road	0.068 plate	0.131 tree
0.143 plate	0.087 plate	0.112 sidewalk	0.062 torso	0.118 grass

Table 2. Visual VSBM matching with n'-TF measure (eq. (4)). $n' = 150$.

Bus	**Computer**	**Ped. Zone**	**Speaker At Pod.**	**Sport**
4.2E-6 car	3.4E-6 car	2.3E-6 head	2.1E-6 head	3.0E-6 head
2.9E-6 head	3.3E-6 head	2.0E-6 car	1.7E-6 car	2.5E-6 car
2.2E-6 tree	2.1E-6 tree	1.5E-6 tree	1.2E-6 tree	2.0E-6 person
1.6E-6 road	1.4E-6 person	1.0E-6 road	1.0E-6 person	1.9E-6 tree
1.4E-6 person	1.4E-6 road	8.6E-7 person	7.9E-7 road	1.0E-6 road
9.6E-7 sidewalk	8.9E-7 sidewalk	7.3E-7 sidewalk	4.7E-7 sidewalk	9.9E-7 plant
7.3E-7 plant	7.8E-7 wheel	5.6E-7 arm	4.6E-7 grass	7.6E-7 sidewalk
6.2E-7 arm	6.6E-7 plant	5.5E-7 leg	3.3E-7 wheel	6.9E-7 grass
5.4E-7 grass	6.4E-7 grass	5.5E-7 torso	3.3E-7 plant	5.7E-7 wheel
5.3E-7 leg	5.8E-7 plate	4.8E-7 grass	2.9E-7 arm	4.7E-7 plate
5.1E-7 wheel	2.6E-7 arm	4.6E-7 wheel	2.4E-7 torso	3.4E-7 arm
5.1E-7 torso	2.4E-7 leg	2.4E-7 plant	2.4E-7 leg	3.4E-7 leg
3.1E-7 plate	2.4E-7 torso	1.6E-7 plate	2.3E-7 plate	3.3E-7 torso

Table 3. Visual GBM matching.

Bus	**Computer**	**Ped. Zone**	**Speaker At Pod.**	**Sport**
0.667 head	0.531 head	0.615 head	0.485 head	0.581 head
0.363 car	0.166 car	0.336 car	0.109 car	0.223 car
0.233 tree	0.059 person	0.234 tree	0.006 tree	0.170 tree
0.153 person	0.050 tree	0.157 person	0.011 person	0.118 person
0.130 road	0.013 torso	0.125 road	0.053 torso	0.070 torso
0.098 arm	0.005 arm	0.109 arm	0.068 arm	0.050 arm
0.092 torso	0.023 leg	0.104 grass	0.113 leg	0.042 grass
0.086 grass	0.094 plate	0.096 torso	0.170 road	0.023 road
0.069 leg	0.101 plant	0.068 leg	0.185 grass	0.017 leg
0.015 sidewalk	0.124 grass	0.017 sidewalk	0.186 plate	0.067 sidewalk
0.044 plant	0.137 road	0.066 plant	0.198 plant	0.108 plant
0.059 wheel	0.184 sidewalk	0.070 wheel	0.249 sidewalk	0.122 plate
0.091 plate	0.305 wheel	0.136 plate	0.380 wheel	0.194 wheel

Table 4. Textual VSBM matching with Spearman's correlation measure (eq. (3)).

Bus	**Computer**	**Ped. Zone**	**Speaker At Pod.**	**Sport**
0.257 road	0.257 person	0.185 sidewalk	0.200 person	0.228 grass
0.185 car	0.185 arm	0.185 road	0.157 plant	0.200 tree
0.171 wheel	0.171 torso	0.171 person	0.157 torso	0.171 road
0.171 sidewalk	0.171 leg	0.171 car	0.142 leg	0.157 plant
0.157 tree	0.157 tree	0.171 tree	0.100 arm	0.142 person
0.142 grass	0.142 plate	0.171 grass	0.100 grass	0.128 sidewalk
0.128 person	0.114 car	0.157 wheel	0.057 head	0.114 arm
0.114 plant	0.114 head	0.114 arm	0.057 plate	0.114 plate
0.100 head	0.100 road	0.114 head	0.057 sidewalk	0.114 torso
0.085 arm	0.100 plant	0.114 leg	0.057 tree	0.100 car
0.085 plate	0.085 grass	0.100 plant	0.042 car	0.100 head
0.085 torso	0.071 wheel	0.100 torso	0.042 road	0.100 leg
0.071 leg	0.028 sidewalk	0.028 plate	0.014 wheel	0.057 wheel

Table 5. Textual VSBM matching with the n'-TF measure (eq. (4)). $n' = 150$.

Bus	**Computer**	**Ped. Zone**	**Speaker At Pod.**	**Sport**
2.4E-6 road	1.0E-6 person	2.5E-6 road	5.6E-7 head	7.0E-7 grass
1.8E-6 tree	6.9E-7 head	2.2E-6 sidewalk	4.2E-7 person	6.0E-7 tree
1.5E-6 person	6.8E-7 arm	1.3E-6 tree	3.7E-7 tree	5.9E-7 person
1.4E-6 wheel	6.5E-7 leg	1.2E-6 car	3.4E-7 arm	5.6E-7 road
1.4E-6 plant	6.4E-7 torso	7.6E-7 wheel	3.3E-7 torso	4.3E-7 head
1.3E-6 car	5.4E-7 plant	6.2E-7 person	3.0E-7 leg	4.1E-7 torso
8.9E-7 sidewalk	2.5E-7 tree	4.4E-7 arm	2.8E-7 plant	4.1E-7 leg
8.3E-7 leg	2.1E-7 wheel	4.2E-7 head	1.6E-7 plate	4.0E-7 plant
7.3E-7 arm	1.9E-7 road	3.6E-7 leg	9.5E-8 grass	4.0E-7 arm
7.2E-7 head	1.6E-7 plate	3.5E-7 torso	9.4E-8 road	2.3E-7 sidewalk
6.7E-7 torso	1.3E-7 sidewalk	3.4E-7 plant	8.1E-8 wheel	1.8E-7 car
5.1E-7 grass	9.6E-8 grass	3.3E-7 grass	6.0E-8 sidewalk	1.7E-7 wheel
4.8E-7 plate	9.1E-8 car	1.2E-7 plate	5.9E-8 car	7.7E-8 plate

Table 6. Textual GBM matching.

Bus	s_v s_t	s_δ	**Comp.**	s_v s_t	s_δ	**Ped.Zone**	s_v s_t	s_δ	**Sp.AtPod.**	s_v s_t	s_δ	**Sport**	s_v s_t	s_δ
head	1 1	0	person	1 3	2	plant	1 11	10	person	1 4	3	head	1 1	0
road	2 5	3	sidewalk	2 12	10	grass	2 7	5	head	2 1	1	sidewalk	2 10	8
car	3 2	1	head	3 1	2	wheel	3 12	9	grass	3 9	6	person	3 4	1
person	4 4	0	road	4 11	7	tree	4 3	1	plant	4 11	7	car	4 2	2
wheel	5 12	7	car	5 2	3	arm	5 6	1	sidewalk	5 12	7	road	5 8	3
arm	6 6	0	arm	6 6	0	leg	6 9	3	road	6 8	2	arm	6 6	0
tree	7 3	4	wheel	7 13	6	car	7 2	5	wheel	7 13	6	wheel	7 13	6
sidewalk	8 10	2	leg	8 7	1	road	8 5	3	tree	8 3	5	tree	8 3	5
grass	9 8	1	grass	9 10	1	sidewalk	9 10	1	arm	9 6	3	leg	9 9	0
plant	10 11	1	tree	10 4	6	torso	10 8	2	car	10 2	8	torso	10 5	5
leg	11 9	2	plant	11 9	2	person	11 4	7	leg	11 7	4	plant	11 11	0
torso	12 7	5	torso	12 5	7	head	12 1	11	torso	12 5	7	grass	12 7	5
plate	13 13	0	plate	13 8	5	plate	13 13	0	plate	13 10	3	plate	13 12	1

Table 7. Differences between the visual and the textual similarity scores issued from the VSBM matching with Spearman's correlation measure.

Bus	s_v s_t	s_δ	**Comp.**	s_v s_t	s_δ	**Ped.Zone**	s_v s_t	s_δ	**Sp.AtPod.**	s_v s_t	s_δ	**Sport**	s_v s_t	s_δ
person	1 6	4	head	1 6	4	grass	1 2	1	person	1 1	0	road	1 3	2
road	2 1	1	person	2 1	1	plant	2 5	3	grass	2 4	2	head	2 8	6
grass	2 5	3	sidewalk	3 10	7	tree	3 2	1	head	3 5	2	person	3 5	2
sidewalk	3 3	0	road	4 7	3	wheel	4 3	1	plant	4 2	2	sidewalk	4 6	2
tree	3 4	1	torso	5 3	2	leg	5 4	1	road	5 6	1	car	5 8	3
plant	3 7	4	leg	6 3	3	plate	6 6	0	sidewalk	6 5	1	leg	5 9	4
head	3 8	5	grass	7 8	1	arm	7 4	3	tree	7 5	2	arm	6 7	1
wheel	4 3	1	plant	7 7	0	person	8 2	6	wheel	8 7	1	torso	7 7	0
torso	5 9	4	car	8 6	2	car	9 2	7	leg	9 3	6	plate	8 7	1
leg	6 10	4	wheel	8 9	1	torso	9 5	4	car	10 6	4	plant	9 4	5
car	7 2	5	tree	9 4	5	head	10 4	6	arm	11 4	7	wheel	10 10	0
arm	7 9	2	arm	10 2	8	road	10 1	9	plate	12 5	8	tree	11 2	9
plate	8 9	1	plate	11 5	6	sidewalk	10 1	9	torso	13 2	11	grass	12 1	11

Table 8. Differences between the visual and the textual similarity scores issued from the VSBM matching with the n'-TF similarity measure. $n' = 150$.

Bus	**Computer**	**Ped. Zone**	**Speaker At Pod.**	**Sport**
3.8E-6 car	2.6E-6 head	2.7E-6 road	1.8E-6 head	2.2E-6 head
3.0E-6 road	2.2E-6 car	2.4E-6 sidewalk	1.2E-6 car	1.8E-6 tree
2.9E-6 tree	1.6E-6 tree	2.4E-6 car	1.1E-6 tree	1.8E-6 car
2.4E-6 head	1.6E-6 person	2.1E-6 tree	1.0E-6 person	1.7E-6 person
2.0E-6 person	1.1E-6 road	1.9E-6 head	6.7E-7 road	1.1E-6 road
1.5E-6 plant	8.9E-7 plant	1.0E-6 person	5.0E-7 arm	1.0E-6 grass
1.4E-6 wheel	7.7E-7 sidewalk	1.0E-6 wheel	4.7E-7 plant	1.0E-6 plant
1.4E-6 sidewalk	7.4E-7 arm	8.0E-7 arm	4.6E-7 torso	7.6E-7 sidewalk
1.0E-6 arm	7.0E-7 leg	7.4E-7 leg	4.4E-7 leg	5.9E-7 leg
1.0E-6 leg	6.9E-7 torso	7.4E-7 torso	4.2E-7 grass	5.9E-7 arm
9.4E-7 torso	6.8E-7 wheel	6.4E-7 grass	4.1E-7 sidewalk	5.8E-7 torso
7.9E-7 grass	5.4E-7 plate	4.7E-7 plant	3.1E-7 plate	5.5E-7 wheel
5.9E-7 plate	5.3E-7 grass	2.2E-7 plate	3.1E-7 wheel	4.1E-7 plate

Table 9. A built-in bi-modality matching with GBM.