Medical-Image Retrieval Based on Knowledge-Assisted Text and Image Indexing

Caroline Lacoste, Joo-Hwee Lim, Jean-Pierre Chevallet, and Diem Thi Hoang Le

Abstract—**Voluminous medical images are generated daily. They are critical assets for medical diagnosis, research, and teaching. To facilitate automatic indexing and retrieval of large medical-image databases, both images and associated texts are indexed using medical concepts from the Unified Medical Language System (UMLS) meta-thesaurus. We propose a structured learning framework based on support vector machines to facilitate modular design and learning of medical semantics from images. We present two complementary visual indexing approaches within this framework: a global indexing to access image modality and a local indexing to access semantic local features. Two fusion approaches are developed to improve textual retrieval using the UMLS-based image indexing. First, a simple fusion of the textual and visual retrieval approaches is proposed, improving significantly the retrieval results of both text and image retrieval. Second, a visual modality filtering is designed to remove visually aberrant images according to the query modality concept(s). Using the ImageCLEFmed database, we demonstrate the effectiveness of our framework which is superior when compared with the automatic runs evaluated in 2005 on the same medical-image retrieval task.**

Index Terms—**Content-based image retrieval, knowledge-based image indexing, Unified Medical Language System (UMLS), visual ontology.**

I. Introduction

CURRENT medical-image analysis research mainly focuses on image registration, quantification, and visualization. Although large amounts of medical images are produced in hospitals every day, there is relatively less research in medical content-based image retrieval (CBIR) [1]. Nevertheless, CBIR systems have a large potential in medical applications. The three main applications concern medical diagnosis, teaching, and research.

For the clinical decision-making process, it can be beneficial to find other images of the same modality, of the same anatomic region, and of the same disease [2]. For instance, for less experienced radiologists, a common practice is to use a reference text to find images that are similar to the query image [3]. Hence, medical CBIR systems can assist doctors in diagnosis by retrieving images with known pathologies that are similar to a patient's image(s). Although modality, anatomy, and pathology information are normally contained in the DICOM header—the standard medical image header—there are still some problems. Indeed, DICOM headers contain a high rate of errors: error rates of 16% have been reported in [4] for the field "anatomical region." Although the purely image-based query methods will not be able to replace text-based methods, they are a very good complement to text-based query methods. In teaching and research, visual retrieval methods could help researchers, lecturers, and student find relevant images from large repositories. Visual features not only allow the retrieval of cases with patients having similar diagnoses, but also cases with visual similarity but different diagnoses.

Current CBIR systems [5] generally use primitive features such as color, texture, or logical features such as object and their relationships to represent images. Because they do not use medical knowledge, such systems provide poor results in the medical domain. More specifically, the description of an image by a low-level image is not sufficient to capture the semantic content of a medical image. This loss of information is called the semantic gap.

In reality, pathology-bearing regions tend to be highly localized [3]. However, it has been recognized that pathology-bearing regions cannot be segmented out automatically for many medical domains [1]. As an alternative, a comprehensive set of 15 perceptual categories related to pathology bearing regions and their discriminative features are carefully designed and tuned for high-resolution CT lung images to achieve superior precision rates over a brute-force feature-selection approach [1]. Hence, it is desirable to have a medical CBIR system that represents images in terms of semantic features that can be learned from examples (rather than handcrafted with a lot of expert input) and do not rely on robust region segmentation.

The semantic gap can also be reduced by exploiting all sources of information. In particular, mixing text and image information generally increases the retrieval performance significantly. Recent evaluation of visual, textual, and mixed approaches within the cross language evaluation forum (CLEF) [6] shows that the mixed approaches outperformed the results of each single approach [7]. In [8], statistical methods are used for modeling the occurrence of document keywords and visual characteristics. The proposed system is sensitive to the quality of the segmentation of the images.

In this paper, we propose to use medical concepts from the National Library of Medicine's (NLM) [9] Unified Medical Language System (UMLS) meta-thesaurus to represent both image and text. The use of UMLS concepts allows our system to work at a higher semantic level and to standardize the semantic

Manuscript received July 14, 2006; revised December 14, 2006. This work was supported by A*STAR, CNRS, and French Ministry of Foreign Affairs under Grant 185-02 InterMed-IR. This paper was recommended by Associate Editor J. Zhang.

The authors are with the French-Singapore IPAL Joint Lab (UMI CNRS 2955, I2R, NUS, UJF), Institute for Infocomm Research, Singapore 119613 (e-mail: joohwee@i2r.a-star.edu.sg).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2007.897114

index of medical data, facilitating the communication between visual end textual indexing and retrieval.

To bridge the semantic gap between low-level images features and the semantic UMLS concepts, we propose a structured learning framework based on support vector machines (SVMs) [10]. This framework facilitates modular design and learning of medical semantics from images. Each image is then represented by visual percepts and UMLS concepts. We developed two complementary visual indexing approaches within this framework: a global indexing to access image modality, and a local indexing to access semantic local features. Indeed, it is important to extract local information from images as pathology-bearing regions tend to be highly localized [3]. This local indexing does not rely on region segmentation but builds upon a patch-based semantic detector [11].

We propose two fusion approaches to benefit from both images and associated text (e.g., DICOM headers and medical report). First, a simple fusion of the textual and visual retrieval approaches is proposed. Second, a visual-modality filtering is designed to remove visually aberrant images according to the query modality concept(s).

The main contributions of this paper are:

1) the use of the medical meta-thesaurus UMLS to standardize the semantic indexes of the textual and visual data;
2) a structured approach for designing and learning medical semantics—that are a combination of UMLS concepts and visual percepts—from images;
3) the fusion between global and local image indexing to capture both modality, anatomy, and pathology information from images;
4) the fusion between textual and visual information: 1) through a simple late fusion between textual and visual similarities to the query and 2) through a visual filtering according to UMLS Modality concepts;
5) the experimental results of the UMLS-based system on the ImageCLEFmed medical-image-retrieval benchmark.

The textual and visual UMLS-based indexing approaches are presented in Section II and III, respectively. Retrieval approaches derived from these UMLS indexing methods are presented in Section IV. We evaluate the UMLS-based retrieval system on ImageCLEFmed 2005 Medical Image Retrieval task in Section V, analyzing the potential of each approach. The mean average precision (MAP) over 25 query topics is compared with the best automatic runs in ImageCLEFmed 2005. Our proposed approach, based on the fusion between text and image with a visual modality filtering, achieves seven more MAP points (23% relative improvement) over the best automatic run in ImageCLEFmed 2005.

II. Text Indexing Using UMLS Concepts

Conventional text information-retrieval (IR) approaches extract words or terms from text and use them directly for indexing. Despite staying at the text "signal" level (syntactic) and the simplicity of word extraction, this method is relatively effective most of the time and is efficiently used in various applications including Web search engines. This success is probably due to the relatively higher semantic level of text compared with other media.

If the use of words for indexing is enough in general, we are confronted with imprecision and ambiguity in the case of precise technical domain. Precise technical domain means restricted domain knowledge, like medicine, where terms are much more important than words. By definition, a term belongs to a terminology: an exhaustive list of noun phrases that have unique meanings in a given domain. Medicine is a typical domain where new terms are forged by specialists to express new diseases or new treatments, for example.

A. Conceptual Indexing

Indexing using terms (e.g., "skin cancer") should improve precision as the index denotes a unique meaning. However, this can lead to a recall problem due to term variation and synonymy (e.g., "melanoma"). Indexing at the conceptual level solves this problem because concepts are abstraction of terms. Moreover, at this conceptual level, we are not language-dependent, and such an IR system becomes multilingual as only one unique set of concepts is used to index a document in any language.

However there are challenges in setting up a conceptual indexing. First, a domain knowledge resource built by specialists is mandatory. This resource should incorporate all useful terms and term variations of a domain, eventually in different languages, and each term should be properly associated with concepts. It is costly to manually build such a resource. Second, we need an automated tool to extract concepts from raw text. Concept extraction is difficult because of the inherent ambiguity and flexibility of the natural language. There are also many language phenomena such as elision that complicates the task of detecting concepts: some term variation refers to the previous expression in the full-length text (e.g., "this destruction is due to ... it is due to"). Moreover, by definition, concepts have unique meanings. Extracting concept means disambiguating the text, which is always a very difficult task. Finally, a flat set of concepts can lead to a sharp decline of recall if the system is not able to establish a link from general concepts in a query (e.g., "bone fracture"), and perhaps more precise concepts present in documents (e.g., "fracture of the femur"). Relation in the knowledge resource is hence mandatory.

To sum up, for a conceptual indexing, we need the following.

- A terminology: it is a list of terms (single or multiterm) from a given domain in a given language. Terms are coming from actual language usage in the domain. They are generally stable noun phrases (i.e., less linguistic variations than any other noun phrase), and they should have an unambiguous meaning in the restricted domain they are used.
- A set of concepts: in our context, a concept is just a language-independent meaning (with a definition), associated with at least one term of the terminology. This notion of concept is close to the notion of *acception* [12].
- A conceptual structure: Each term is associated with at least one concept. Concepts are also organized into several networks. Each network links concepts using a conceptual relation.
- A conceptual mapping algorithm: this is a method that selects a set of potential concepts from a sentence using the terminology and the conceptual structure.

We used to call the conceptual structure with the concept set and the terminology the domain knowledge resource. This point of view is a strong simplification of the complex reality, but this description is enough for an indexing usage: conceptual indexing is then the operation of transforming natural language document into an indexing structure of concepts (e.g., sets, vectors, and graphs) by the way of the conceptual mapping algorithm using the domain knowledge resource.

B. From Terms to Concepts

One may think that dealing with a precise domain may reduce some of the concept extraction problems, like ambiguity. This is partly true. Term ambiguity arises when different concepts may be mapped to a single term. In practice, ambiguity depends on the precision of the knowledge resource. If we reduce the domain, then we also reduce possible term exceptions, hence also ambiguity. For example, "X-ray" may refer to a wave in physics, but could only refer to an image modality in radiology. Unfortunately, when we have more precise concepts (and terms), we are confronted with another form of ambiguity: a structure ambiguity. This corresponds to several ways of concept extraction, where some are composition of others. For example, the term "right lobe pneumonia" can be associated with a single concept but can be split into two terms associated with other concepts: "right lobe" and "pneumonia." A common solution is to model "concepts structure equivalence." This consists of setting up a model that expresses concept composition and relations. Some of these relations can be equivalence or subsumption. A Terminological Logic can be used, but it is often neither simple nor possible to set up such a process for indexing purpose using a real large set of concepts, because concepts have to be expressed in the chosen formalism. Such a formalized concept resource is then an "ontology." A large ontology on medicine with concepts expressed in a logical format does not yet exist.

Another common difficulty for concept extraction is term variation. Despite the fact that terms should be stable noun phrases, there still exist in practice many variations in technical terms. It is the role of the terminology to list all term variations, but, in practice, some variations have to be processed by the conceptual mapping algorithm.

Despite these difficulties, conceptual indexing can produce a high-precision multilingual indexing and can solve every precise query. This solution is adapted for the medicine domain.

C. Using a Meta-Thesaurus

A meta-thesaurus is a merger of existing thesauri. A thesaurus in IR differs from a terminology by its usage and its structure. Terminology is used to describe possible or acceptable terms from a domain, for language normalization, official translation, and so on. A thesaurus is used for document indexing. It is often used for manual indexing, hence a thesaurus may include word entries that are not really *terms* because they are not used in actual text. For example, the entry `Technology and Food and Beverages` in the MESH thesaurus is not a term in medicine, but it is used to support the thesaurus hierarchy. This refers to another difference between terminology and a thesaurus: to be used as a (manual) indexing tool, a thesaurus needs to be structured, at least as a hierarchy.

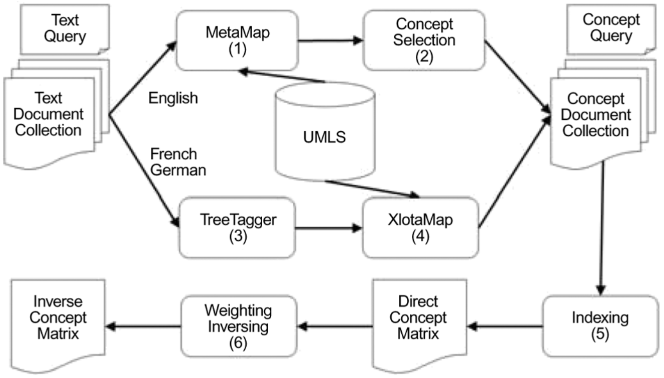

Fig. 1. Indexing path of the text.

Merging different thesauri produces a meta-thesaurus, and it does not lead to the ideal conceptual structure we briefly described in the previous section: not all entries are terms, so not all entries can be found in actual text. Moreover, different thesaurus structures (e.g., hierarchy) have to be merged in one structure. This is a difficult problem.

UMLS is a good candidate to approximate a domain knowledge resource for medical image and text indexing. First, UMLS has a large base that includes more than 5.5 million terms in 17 languages. It is maintained by specialists with two updates a year. Unfortunately, UMLS as a merger of different sources (i.e., thesaurus and terminology), is neither complete nor consistent. In particular, the links among concepts are not equally distributed. In UMLS, the notion of *concept* has been added to unify the merging of resources. The interconcept relationship (like hierarchies) is those from the source original thesaurus. Hence, there is a sort of redundancy as multiple similar paths can be found between two concepts using different sources. From the 5.5 million terms, UMLS identifies 1.1 million unique concepts.

In order to have a common categorization of this concept set, UMLS has a global high-level semantic category called *semantic types* and *semantic groups* [13] assigned manually and independently of all thesaurus hierarchies by the meta-thesaurus editors. This partially solves the problem of merging existing thesaurus hierarchy during the merging process.

In the following, we describe the way we have used this meta-thesaurus to build up and test an effective conceptual indexing.

D. Indexing Process

Fig. 1 depicts the indexing path. From the text document collection, it produces a document identifier matrix, usually called an inverse matrix, ready to be queried. The global first step (on the top) is the transformation from raw text documents to documents expressed as concepts. The treatment path is the same for documents and queries.

Despite the large set of terms and term variations available in UMLS, it still cannot cover all possible term variations. The concept mapping algorithm has to manage term variations. For English texts, we use MetaMap [14] (box (1) in Fig. 1) provided by NLM. By using lexical variants in the SPECIALIST lexicon and the database of synonyms supported by UMLS knowledge source, MetaMap provides good coverage on different variants for concept identification. The concept identification in MetaMap involves the following four steps where each of the first three steps produces output identified by a keyword.

Step 1) Text parsing (`phrase`): the goal is to identify noun phrases. The remaining steps are performed on noun phrases only.

Step 2) Generating variants and candidate selection (`candidates`): MetaMap computes noun phrases variants including acronyms, abbreviations, synonyms, derivational variants, inflectional and spelling variants, as well as meaningful combination of these variants. The candidate set is all meta-thesaurus strings containing at least one of the variants.

Step 3) Candidate evaluation (`ev` in `candidates` and `mappings`): this evaluation produces a confident score value based on four components: *centrality*, *variation*, *coverage* and *cohesiveness*. Centrality is equal to 1 if candidate term involves the head of the noun phrase. Variation computes the distance between the text and candidate term. This distance is related to the steps needed to produce the term variant. Coverage is related to number of words in the candidate term present in the text. Cohesiveness is similar to coverage but takes into account connected words.

Step 4) Final mapping proposition between concept and text (`mapping`): this step is the final MetaMap mapping proposition. The system combines the best candidate terms to form mapping between the noun phase and candidate terms.

We have developed a similar tool XIotaMap (box (4) in Fig. 1) for French and German documents. The tool receives the outcome of the part-of-speech parsing provided by TreeTagger [15]. This concept extraction tool is a simplified version of MetaMap. The steps are described as follows.

Step 1) Text parsing: it is provided by TreeTagger, which assigns part of speech (POS) tag and also a stemmed version for each word (box (3) in Fig. 1).

Step 2) Generating candidate variants: based on POS, noun phrases are selected. Only case and stemmed variants are examined for each word in the noun phrases.

Step 3) Concept selection: either the largest or all candidate variants that match an UMLS entry are selected. We have also reduced the thesaurus list to exclude concepts that are not in the medical domain. For example, this enables the identification of "x-ray" as radiography and not as the physical phenomenon (the wave) which seldom appears in our documents.

Like MetaMap, XIotaMap does not provide any disambiguation. But selecting concepts associated to the largest terms tend to reduce ambiguity. Concepts extraction is also limited to noun phrases (i.e., verbs are not treated). Also, we do not solve the structure ambiguity: a partial solution is to include in the index all concepts potentially extracted from texts.

The extracted concepts are then organized in conceptual vectors, using a conventional vector space model (VSM) [16] in IR. This is the role of the indexing treatment [box (5)] in the indexing path. The documents in each different language follows a parallel treatment path. The indexing also merges concepts extracted from these different sources.

Finally, in the case of documents, the XML matrix is inverted and a weighting scheme is applied. A concept identifier is associated with the list of all documents where it appears with the corresponding weight. This is a classic file in VSM indexing, except that the format is XML and vector dimensions are concept identifiers. We use the weighting scheme provided by our XIOTA indexing system [17]. This solution enables us to reuse classical standard indexing as retrieval tools. XIOTA is a VSM-based experimental IR system which handles and produces data in XML format. Documents and queries represented by UMLS concepts under XML format are firstly used to build document vectors and query vectors of concepts with their frequencies. All document vectors are then combined and reweighted to form a unique direct matrix of all concepts with their $tf \cdot idf$ (i.e., product of term frequency and inverse document frequency) weighting values.

Query vectors are treated the same way. The querying process is performed to yield the ranking list of relevant documents. This process is in fact a matrix product between the direct matrix of query $(\mathrm{queryID} \times \mathrm{conceptID})$ and the inverse document matrix $(\mathrm{conceptID} \times \mathrm{documentID})$. Dimension filtering and reweighting, as well as fusion with image retrieval, are performed subsequently on the document ranking list. All results are presented in the experimental evaluation section (Section V).

The choice of standard VSM for concepts is debatable, as we cannot guarantee independence of vector dimensions. In practice (see the results in Section V), it does not affect the results. Conceptualization of documents and queries is just filtering noun phrases and replacing them with a concept identifier: from a statistical distribution point of view, the change should be small. This explains why the use of classical IR word weighting still tends to be effective with concepts. We know that this aspect should be more carefully investigated. In this study, we mostly experience the effectiveness of dimension filtering and reweighing that boost results, more than the effect of conceptual indexing per se (see Section V-A). Also, this enables us to build a unique multilingual index instead of an index for each language. In this way, the document base can be queried using any of the 17 languages of UMLS. Finally, using concepts instead of terms enables a "standard" description that is used in the image indexing part. We can then have a common "inter-media" conceptual index between multilingual text document and images.

III. Semantic Medical-Image Indexing

We aim to bridge the semantic gap between low-level visual features (e.g., texture and color) and medical concepts (e.g., brain, computed tomography, and fracture) for content-based indexing and retrieval. More precisely, our aim is to associate with each image or with each region a semantic label that corresponds to a combination of UMLS concepts and visual percepts. In this way, we have a common language for both images and associated textual data (e.g., DICOM headers or medical reports). We define three types of UMLS concepts that could be associated to one image or one region:

- modality concepts that belong to the UMLS semantic type: "Diagnostic Procedure";

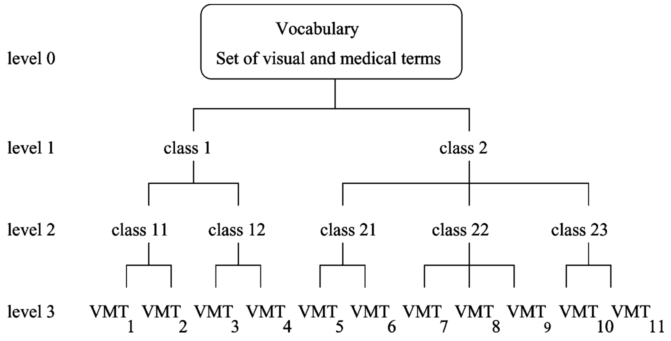

Fig. 2. Typical tree structure for a VMT classifier.

- anatomy concepts that belong to the UMLS semantic types: "Body Part, Organ, or Organ Component," "Body Location or Region," "Body Space or Junction," or "Tissue";
- pathology concepts that belong to the UMLS semantic types: "Acquired Abnormality," "Disease or Syndrome," or "Injury or Poisoning."

We propose a structured learning framework based on SVMs to facilitate modular design and learning of medical semantics from images. We developed two complementary indexing approaches within this statistical learning framework:

- a global indexing to access image modality (e.g., chest X-ray, gross photography of an organ, or microscopy);
- a local indexing to access semantic local features that are related to modality, anatomy, and pathology concepts.

After a brief presentation of our common learning framework in Section III-A, we detail both approaches in Sections III-B and III-C.

A. Common Statistical Learning Framework

First, a set of disjoint semantic tokens with visual appearance in medical images is selected to define a Visual and Medical (VisMed) vocabulary with reference to the UMLS concepts. This notion of using a visual and semantic vocabulary to represent and index images has been applied to consumer images in [18]. Here, we use UMLS concepts to represent each token in the medical domain. Second, low-level features are extracted from image-region instances to represent each token in terms of color, texture, and shape, for example. Third, these low-level features are used as training examples to build a semantic classifier according the VisMed vocabulary. We use a hierarchical classification scheme based on SVMs.

A tree whose leaves are the VisMed Terms (VMTs) is designed and constructed in a top-down manner, guided by the possible hierarchy of the associated terms in UMLS and manual inspection on the visual similarities among the VMTs. The upper levels of the tree consist of auxiliary classes that group similar terms with respect to their visual appearances. A schematic example of the tree structure is given in Fig. 2. A learning process is performed at each node in the following way. If a node corresponding to a class $\mathcal{C}$ has $N_\mathcal{C}$ direct children, $N_\mathcal{C}$ SVM classifiers are learned to classify a class against the other $N_\mathcal{C} - 1$ classes. The positive and negative examples for a class $c \in \mathcal{C}$ are given by the instances of the term(s) associated with the class c and the instances of the terms associated to the $N_\mathcal{C} - 1$ other classes, respectively.

TABLE I
MODALITY-RELATED GVM INDEXING TERMS (ITALIC: COLOR) AND NUMBERS OF IMAGES

GVM index terms	#	GVM index terms	#
Angiography	140	*Photograph-gross-organ*	468
CT-abdomen	58	*Photograph-patient*	534
CT-brain	112	*Presentation-slides*	136
CT-chest	184	Scintigraphy	239
Doppler	174	*SPECT-color*	11
Electrocardiography	106	SPECT-grey	7
Endoscopy	34	Spectroscopy	22
Fluoroscopy	17	*Thermography*	35
Mammography	80	Ultrasound	24
Microscopy	290	X-ray-bones	332
MRI-brain-axial	250	X-ray-hand	29
MRI-brain-frontal	54	X-ray-knee	82
MRI-brain-sagittal	128	X-ray-neck	122
Ophtalmoscopy-color	39	X-ray-pelvis	104
Ophtalmoscopy-grey	15	X-ray-skull	52
PET	34	X-ray-thoracic	135

The classifier for the VisMed vocabulary is finally designed from the tree of SVM conditional classifiers in the following way. The conditional probability that an example z belong to a class c given that the class belong to its superclass $\mathcal{C}$ is first computed using the softmax function [19]

$$P(c|z,\mathcal{C}) = \frac{\exp^{\mathcal{D}_c(z)}}{\sum_{j\in\mathcal{C}} \exp^{\mathcal{D}_j(z)}} \tag{1}$$

where $\mathcal{D}_c$ is the signed distance to the SVM hyperplane that separates class c from the other classes under the same superclass $\mathcal{C}$. The probability of a VisMed Term VMT_i (i.e., a leave of the tree) for an example z is finally given by

$$P(\mathrm{VMT}_i|z) = \prod_{l=1}^{L} P\left(\mathcal{C}^l(\mathrm{VMT}_i)|z, \mathcal{C}^{l-1}(\mathrm{VMT}_i)\right) \tag{2}$$

where L is the number of hierarchical levels, $\mathcal{C}^L(\mathrm{VMT}_i)$ is equal to VMT_i, $\mathcal{C}^{l-1}(\mathrm{VMT}_i)$ denotes the superclass to which $\mathcal{C}^l(\mathrm{VMT}_i)$ belongs, $\mathcal{C}^0(\mathrm{VMT}_i)$ is the root class containing all of the vocabulary (it corresponds to the tree root), and $P(\mathcal{C}^l(\mathrm{VMT}_i)|z, \mathcal{C}^{l-1}(\mathrm{VMT}_i)$ is given by (1). For example, the probability of the term VMT_1 of Fig. 2 is given by

$$P(\mathrm{VMT}_1|z) = P(\mathrm{VMT}_1|z, \mathcal{C}_{11})P(\mathcal{C}_{11}|z, \mathcal{C}_1)P(\mathcal{C}_1|z) \tag{3}$$

where $\mathcal{C}_i$ denotes the class i in Fig. 2.

B. Global UMLS Indexing

The global UMLS indexing is based on a two-level hierarchical classifier according to the main modality concepts. This modality classifier is learned from about 4000 images separated into 32 classes: 22 gray-level modalities and 10 color modalities. All of these global visual and modality (GVM) indexing terms and the numbers of image samples are given in Table I, where the GVM terms in italic font refer to color modalities.

Except for the indexing term "Presentation-slides," each indexing term is characterized by a UMLS modality concept and, sometimes, an anatomy concept (e.g., neck or pelvis), a spatial concept (e.g., axial or frontal), or a color percept (color, gray). The training images come from the CLEF [6] database (about 2500 examples), from the IRMA [20] database (about 300 examples), and from the Web (about 1200 examples). Potential training images from the ImageCLEFmed database was first automatically selected based on the presence of the modality concepts found in the associated medical reports. A manual inspection step was then performed on these potential training images to remove irrelevant examples. For instance, text description such as "A CT scan is recommended for the patient" in the medical report while the associated image is actually an X-ray image and the term "X-ray" may not appear in the medical report itself. We plan to automate this filtering in the near future.

The first level of the classifier corresponds to a classification for gray-level versus color images. Indeed, some ambiguity can appear due to the presence of colored images or the slightly blue or green appearance of X-ray images. This first classifier uses the first three moments in the HSV color space computed on the entire image. The second level corresponds to the classification of Modality UMLS concepts given that the image is in the gray or the color class. For the gray-level class, we use gray-level histogram (32 bins), texture features (mean and variance of Gabor coefficients for five scales and six orientations), and thumbnails (gray values of 16 × 16 resized image). For the color class, we have adopted the HSV histogram (125 bins), Gabor texture features, and thumbnails. Zero-mean normalization [21] was applied to each feature. For each SVM classifier, we adopted an RBF kernel

$$\exp\left(-\gamma \|x-y\|^2\right) \tag{4}$$

where $\gamma = 1/2\sigma^2$ and with a modified city-block distance:

$$\|x-y\| = \frac{1}{F}\sum_{f=1}^{F}\frac{d(x_f,y_f)}{N_f} \tag{5}$$

where $x = \{x_1,\ldots,x_F\}$ and $y = \{y_1,\ldots,y_F\}$ are feature vectors, x_f, y_f are feature vectors of type f, $d(x_f,y_f) = \sum_i |x_{fi} - y_{fi}|$, N_f is the feature vector dimension, and F is the number of feature types: $F = 1$ for the gray versus color classifier and $F = 3$ for the conditional modality classifiers: color, texture, and thumbnails. This just-in-time feature fusion within the kernel combines the contribution of color, texture, and spatial features equally [22]. It is simpler and more effective than other feature fusion methods we have attempted.

The classifier has been first trained—using SVM-Light software [10], [23], [24]—on half of the dataset (training set) to evaluate its performance on the other half of the dataset (validation set) to determine the optimal SVM parameter γ that gives the lowest error rate averaged over all classes. For the RBF kernels, we have derived $\gamma = 1$. The average error rate on the validation set is 18%, with recall and precision rates higher than 70% for most of the classes. The classification is quite good given the high intraclass variability of some classes and high interclass similarity among some classes. For example, to differentiate a brain MRI image and a brain CT image is a difficult task, even for a human operator.

The probability of a modality MOD_i for an image z is given by (2). More precisely, we have

$$P(\mathrm{MOD}_i|z) = \begin{cases} P(\mathrm{MOD}_i|z,C)P(C|z), & \text{if } \mathrm{MOD}_i \in C \\ P(\mathrm{MOD}_i|z,G)P(G|z), & \text{if } \mathrm{MOD}_i \in G \end{cases} \tag{6}$$

where C and G denote the color and the gray-level classes, respectively.

A modality concept label L can thus be assigned to an image z using the following formula:

$$L(z) = \mathrm{argmax}_i P(\mathrm{MOD}_i|z). \tag{7}$$

After learning (using the entire dataset in order to have more training samples), each database image z is indexed according to modality given its low-level features z_f. The indexes are the probability values given by (6).

C. Local UMLS Indexing

To better capture the medical image content, we propose to extend the global modeling and classification with local patch classification of local visual and semantic (LVM) terms. Each LVM indexing term is expressed as a combination of UMLS concepts from Modality, Anatomy, and Pathology semantic types. In these experiments, we have adopted color and texture features from patches (i.e., small image blocks) and a classifier based on SVMs and the softmax function [19] given by (1). A semantic patch-based detector based on SVM, similar to the GVM classifers described above but now on local image patches, was designed to classify a patch according to the 64 LVM terms given in Table II.

The color features are the three first moments of the Hue, the Saturation, and the Value of the patch. The texture features are the mean and variance of Gabor coefficients using five scales and six orientations. Zero-mean normalization [21] is applied to both the color and texture features. We adopted an RBF kernel with a modified city-block distance given by (5). The training dataset is composed of 3631 patches extracted from images mostly coming from the Web (921 images coming from the Web and 112 images from the ImageCLEFmed collection ~0.2%). The classifier has been first trained—using SVM-Light software—on the first half of the dataset to be evaluated on a second half. The error rate of this classifier is about 30%.

Since the resolutions of the training images are different, we first normalized the images, respecting their aspect ratios, to resolutions with a maximum of 360 pixels on the longer side. As most of the LVM terms are visible within areas of 40 × 40 pixels (i.e., 1/9 along horizontal or vertical dimension), we have decided to build our LVM classifiers based on 40 × 40 image patches.

After learning, the LVM indexing terms are detected during image indexing from image patches without region segmentation to form semantic local histograms. Essentially, an image is tessellated into overlapping image blocks of size 40 × 40 pixels after image-size normalization, similar to the preprocsing step

TABLE II
LVM INDEXING TERMS AND NUMBERS OF SAMPLES

LVM indexing terms	#	LVM indexing terms	#
Angio-artery	42	Photo-gross-background	73
Angio-vessels	40	Photo-gross-brain	61
Black-background	8	Photo-gross-head	93
CT-abdomen-bone-rib	52	Photo-gross-kidney	64
CT-abdomen-bone-spine	47	Photo-gross-lung	95
CT-abdomen-liver	31	Photo-gross-lung-pneumonia	22
CT-chest-bone	42	Photo-gross-stomach-endoscopic	66
CT-chest-lung-nodules	33	Photo-gross-stomach-macroscopic	59
CT-chest-lung-pneumonia	34	Photo-hand-finger-osteoarthritis	24
CT-head-center	30	Photo-infected-wound	47
CT-head-skull	44	Photo-skin-lesion	114
Electrocardiogram	27	Photo-skin-normal	49
Hand-drawn-illustrations	152	Photo-teeth-gum	37
Micro-cerebral-Alzheimer-plaque	58	Photo-tumor	31
Micro-blood	42	Scintigraphy-person-head	49
Micro-cerebellum	26	Scintigraphy-person-limb	83
Micro-cerebral-Alzheimer-tangle	75	Scintigraphy-person-torso	157
Micro-confocal-cerebellum	28	Ultrasound-Doppler	51
Micro-kidney	305	Ultrasound-fetus	114
Micro-multi-nucleated-giant-cell	28	Ultrasound-gallstone	31
Micro-muscle	57	Ultrasound-grey	72
Micro-polymorphonuclear-neutrophils	73	White-background	6
Micro-parvovirus-infection	32	X-ray-bone-femur	37
MRI-head-face	62	X-ray-bone-fracture	45
MRI-head-brain	70	X-ray-bone-hand-finger-osteoarthritis	41
MRI-head-skull	57	X-ray-bone-hand-wrist	40
Photo-face-eye	26	X-ray-bone-implant	113
Photo-face-mouth	32	X-ray-bone-joint	50
Photo-face-nose	39	X-ray-bone-pelvis	68
Photo-fetus-head	25	X-ray-chest-heart	36
Photo-fetus-limb	45	X-ray-chest-lung-clouded	58
Photo-fetus-torso	25	X-ray-vertebral	58

in learning as described above. Each patch is then classified into one of the 64 LVM terms using the Semantic Patch Classifier. An image containing P overlapping patches is then characterized by the set of P LVM histograms and their respective location in the image. An histogram aggregation per block gives the final image index: $M \times N$ LVM histograms. Each bin of a given block B corresponds to the probability of a LVM term presence in this block. This probability is computed as follows:

$$P(\mathrm{VMT}_i|B) = \frac{\sum_z area(z \cap B) P(\mathrm{VMT}_i|z)}{\sum_z area(z \cap B)} \qquad (8)$$

where B is a block of a given image, z denotes a patch of the same image, $area(z \cap B)$ is the area of the intersection between z and B (i.e., the number of pixels common in both image patch z and image block B), and $P(\mathrm{VMT}_i|z)$ is given by (2).

Note that we require the design of each tessellation block B to cover or overlap with at least some patch z so that $area(z \cap B)$ is nonzero. This requirement can be easily satisfied as long as the sampling of image patches is dense enough to cover the entire image. Also, we have adopted block-based tessellation instead of region segmentation as robust segmentation is still an open problem and the segmentation outcome is erroneous in our preliminary experimentation.

To facilitate spatial aggregation and matching of image with different aspect ratios ρ, we design five tiling templates, namely $M \times N = 3 \times 1, 3 \times 2, 3 \times 3, 2 \times 3$, and 1×3 grids resulting in three, six, nine, six, and three probability vectors per image, respectively. This design choice was based on the observation on the most common aspect ratios of the images in the database so as to balance the tradeoff between computational complexity in image matching and distortion in comparing images with different aspect ratios.

IV. IR BASED ON SEMANTIC MEDICAL INDEXING

Here, we consider the problem of retrieving images that are relevant to a textual query (free text) and/or to a visual query (one or several images) from a large medical database. This database is supposed to be constituted of cases, each case containing a medical report and one or several images. We developed a retrieval system enabling three types of retrieval methods: a textual retrieval method that matches textual query and the textual indexes of the medical reports, a visual retrieval method that matches query image(s) and the visual indexes of the database images, and a mixed retrieval that combines visual and textual indexes. This system is presented in Fig. 3, and each proposed retrieval method is described in the next sections.

A. *Textual Retrieval*

Textual retrieval consists of performing the matching between the weighted concept vector of the query and the Inverse Concept Matrix produced. We have tested three text-retrieval approaches based on conceptual indexing using UMLS concepts extracted as described in Section II. The first conceptual text retrieval approach (denoted as (T_1)) uses a VSM [16] for representing each document and a cosine similarity measure to compare the query index to the database medical report. The $tf \cdot idf$ measure is used to weight the concepts.

One major criticism we have against VSM is the lack of structure of the query. VSM is known to perform well using long textual queries but ignoring query structure. The ImageCLEFmed 2005 queries we tested are rather short. Moreover, it seems obvious to us that it is the *complete* query that should be solved and not only part of it. After query examination, we found out that queries are implicitly structured according to some semantic types (e.g., anatomy, pathology, or modality). We call this the "semantic dimensions" of the query. Omitting a correct answer to any of these dimensions may lead to incorrect answers. Unfortunately, VSM does not provide a way to ensure answers to each dimension.

To solve this problem, we decided to add a semantic dimension filtering step to the VSM in order to explicitly taking into account the query dimension structure. This extra filtering step retains answers that incorporate at least one dimension. We use

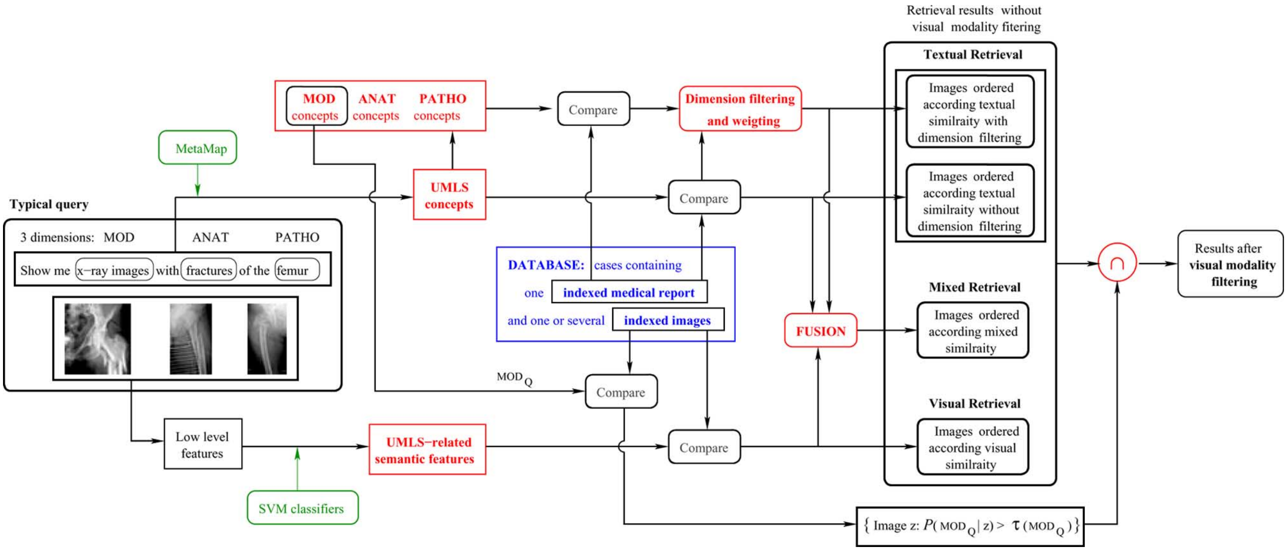
Retrieval results without visual modality fitering
Textual Retrieval
Images ordered according textual similraity with dimension filtering
Images ordered according textual similraity without dimension filtering
Mixed Retrieval
Images ordered according mixed similraity
Visual Retrieval
Images ordered according visual similraity
Results after visual modality filtering
∩
MOD concepts
ANAT concepts
PATHO concepts
Compare
Dimension filtering and weigting
MetaMap
UMLS concepts
Typical query
3 dimensions: MOD ANAT PATHO
Show me x-ray images with fractures of the femur
DATABASE: cases containing
one indexed medical report
and one or several indexed images
FUSION
MOD Q
Low level features
UMLS-related semantic features
SVM classifiers
{ Image z: P(MOD Q | z) > τ (MOD Q)}

Fig. 3. Retrieval system based on UMLS-based image and text indexing.

TABLE III
COMPARATIVE RESULTS ON THE MEDICAL-IMAGE RETRIEVAL TASK OF ImageCLEFmed 2005

Method	Visual	Textual	MAP
(V_1) Global UMLS image indexing	X		10.38%
(V_2) Local UMLS image indexing	X		6.56%
(V_3) Fusion between (V_1) and (V_2)	X		12.11%
(BV) Best automatic visual run in 2005 (GIFT)	X		9.42%
(T_1) UMLS text indexing		X	16.08%
(T_2) UMLS text indexing with DF		X	18.94%
(T_3) UMLS text indexing with DWF		X	22.01%
(BT) Best automatic textual run in 2005 (DFMT)		X	20.84%
(F_1) Fusion between (V_3) and (T_3)	X	X	27.96%
(F_2) Visual modality filtering on (T_3)	X	X	29.25%
(F_3) Visual modality filtering on (F_1)	X	X	34.60%
(BM) Best automatic mixed run with DFMT in 2005	X	X	28.21%
(BM') Best automatic mixed run without DFMT in 2005	X	X	23.89%

a semantic structure on concepts provided by UMLS. Semantic dimension of a concept is defined by its UMLS semantic type, grouped into semantic groups: Anatomy, Pathology, and Modality. Only a conceptual indexing and a structured meta-thesaurus like UMLS enable us to do such a semantic dimension filtering (DF). This filtering discards noisy answers regarding to the dimension query semantic structure. We shall denote this VSM with a DF approach as (T_2).

Another solution to take into account query semantic structure is to reweight answers according to dimensions. Here, Relevance Status Value output from VSM is multiplied by the number of concepts matched with the query according to the dimensions. This simple reweighting scheme strongly emphasizes the presence of a maximum number of concepts related to semantic dimensions. This reweighting step is followed by the previous DF step. We use (T_3) to indicate this dimension weighting and filtering (DWF) approach. According to our results in Table III, this produces the best results for the ImageCLEFmed 2005 collection with 22% of MAP. This result outperforms any other classical textual indexing reported in ImageCLEFmed 2005. Hence, we have shown here the potential of conceptual indexing.

B. Visual Retrieval

For query by example(s), we propose three retrieval methods based on the two visual indexing presented in Section III. When several images are given in the query, the similarity between a database image z with the query is given by the maximum value among the similarities between z and each query image.

The first method, denoted as (V_1), is based on the global indexing scheme according to the modality. An image is represented by a semantic histogram, each bin corresponding to a modality probability. Even if the anatomy and the pathology are not directly taken into account in such an index, it captures implicitly more information than the modality. Indeed, the image is not represented by a single modality but by its projection in a semantic space. The distance between two images is given by the Manhattan distance (i.e., a city-block distance) between the two semantic histograms. The similarity between a query image q and a database image z is then given by

$$\lambda(q,z) = 1 - \frac{1}{2}\sum_k |P(\mathrm{MOD}_k|q) - P(\mathrm{MOD}_k|z)| \qquad (9)$$

where $P(\mathrm{MOD}_k|.)$ is given by (6).

The second method, denoted as (V_2), is based on local UMLS visual indexing. An image is then represented by $M \times N$ semantic histograms. Given two images represented as different grid patterns, we propose a flexible tiling (FlexiTile) matching scheme to cover all possible matches. For instance, given a query image q of 3 × 1 grid and an image z of 3 × 3 grid, intuitively q should be compared with each of the three columns in z, and the highest similarity will be treated as the final matching score.

The FlexiTile matching scheme is formalized as follows. Suppose a query image q and a database image z are represented as $M_1 \times N_1$ and $M_2 \times N_2$ grids, respectively. The overlapping grid $M \times N$, where $M = \min(M_1, M_2)$ and $N = \min(N_1, N_2)$ is the maximal matching area. The similarity λ between q and z is the mean of the similarity of all possible $M \times N$ tilings

$$\lambda(q, z) = \sum_{\substack{m_1=1, n_1=1, \\ m_2=1, n_2=1}}^{\substack{m_1=u_1, n_1=v_1, \\ m_2=u_2, n_2=v_2}} \frac{\lambda\left(q_{m_1,n_1}, z_{m_2,n_2}\right)}{u_1 v_1 u_2 v_2} \tag{10}$$

where $u_1 = M_1 - M + 1$, $v_1 = N_1 - N + 1$, $u_2 = M_2 - M + 1$, $v_2 = N_2 - N + 1$, and the similarity for each tiling $\lambda(q_{m_1,n_1}, z_{m_2,n_2})$ is defined as the average similarity over $M \times N$ blocks as

$$\lambda\left(q_{m_1,n_1}, z_{m_2,n_2}\right) = \frac{\sum_i \sum_j \lambda_{ij}\left(q_{m_1,n_1}, z_{m_2,n_2}\right)}{M \times N} \tag{11}$$

and, finally, the similarity $\lambda_{ij}(q_{m_1,n_1}, z_{m_2,n_2})$ between two image blocks is computed based on L_1 distance measure (city-block distance) as

$$\begin{aligned} &\lambda_{ij}\left(q_{m_1,n_1}, z_{m_2,n_2}\right) \\ &\quad = 1 - \frac{\sum_k \left|P\left(\mathrm{VMT}_k | q_{p_1,q_1}\right) - P\left(\mathrm{VMT}_k | z_{p_2,q_2}\right)\right|}{2} \end{aligned} \tag{12}$$

where $p_1 = m_1 + i, q_1 = n_1 + j, p_2 = m_2 + i, q_2 = n_2 + j$ and $P(\mathrm{VMT}_k|B)$ is the probability of a LVM term presence in block B given by (8). There is a tradeoff between content symmetry and spatial specificity. If we want images of similar semantics with different spatial arrangement (e.g., mirror images) to be treated as similar, we can have larger tessellated blocks (i.e., the extreme case is a global histogram). However, in applications such as medical images where there is usually very small variance in views and spatial locations are considered differentiating across images, local histograms will provide good sensitivity to spatial specificity. Furthermore, we can attach different weights to the blocks to emphasize the focus of attention (e.g., center) if necessary. In this paper, we report experimental results without block weighting.

The last visual retrieval method, denoted as (V_3), is the fusion of the two first approaches, i.e., (V_1) and (V_2). This approach thus combines two complementary sources of information, the first concerning the general aspect of the image (global indexing according to modality), and the second concerning semantic local features with spatial information (local UMLS indexing). The similarity to a query is given by the mean of the similarity to a query according to each index. This simple fusion was more effective than other fusion methods we have tested [25].

C. Mixed Retrieval

The last module of our retrieval system concerns the fusion between text and image retrieval.

The first fusion method, denoted as (F_1), is a late fusion of visual and textual similarity measures, obtained from approaches (V_3) and (T_3), respectively. The similarity between a mixed query $Q = (Q_I, Q_T)$ (Q_I : image(s), Q_T : text) and a couple composed of an image and the associated medical report (I, R) is then given by

$$\lambda(Q, I, R) = \alpha \frac{\lambda_V(Q_I, I)}{\max\limits_{z \in \mathcal{D}_I} \lambda_V(Q_I, z)} + (1 - \alpha) \frac{\lambda_T(Q_T, R)}{\max\limits_{z \in \mathcal{D}_T} \lambda_T(Q_T, z)} \tag{13}$$

where $\lambda_V(Q_I, I)$ denotes the maximum of the visual similarity between I and an image of Q_I, $\lambda_T(Q_T, R)$ denotes the textual similarity between the textual query Q_T and the medical report $R, \mathcal{D}_I$ denotes the image database, and $\mathcal{D}_T$ denotes the text database. After systematic experimentations with α ranges from 0 to 1 at 0.1 equal intervals, we choose $\alpha = 0.7$, which gives the best retrieval performance. The factor α allows the control of the weight of the textual similarity with respect to the image similarity. In order to compare similarities in the same range, each similarity is divided by the corresponding maximal similarity value on the entire database.

The other fusion paradigm is based on modality visual filtering by exploiting the UMLS index of images directly. Indeed, it is based on a direct matching between the modality concept (MOD_Q) extracted from a given textual query and the estimated modality concept $P(\mathrm{MOD}_Q|I)$ precomputed (see Section III-B) as part of the conceptual image index for each image I. This direct matching is done automatically with the use of UMLS. More specifically, a comparison between the UMLS concept related to modality and the image modality index is done in order to remove all aberrant images. The decision rule is the following: an image I is admissible for a query modality MOD_Q only if

$$P(\mathrm{MOD}_Q|I) > \tau(\mathrm{MOD}_Q) \tag{14}$$

where $\tau(\mathrm{MOD}_Q)$ is a threshold defined for the modality MOD_Q determined empirically in our experiments for optimal retrieval performance. This decision rule defines a set of admissible images for a given modality $\mathrm{MOD}_Q : \{I \in \mathcal{D}_I : P(\mathrm{MOD}_Q|I) > \tau(\mathrm{MOD}_Q)\}$. The final result is then the intersection of this set and the ordered set of images retrieved by *any* retrieval method. This modality filter is particularly interesting for filtering textual retrieval results as several images of different modalities can be associated with the same medical report. The ambiguity is thus removed when using a visual modality filtering. We applied this modality filter to the best UMLS conceptual text indexing and retrieval method (i.e., (T_3) as described above) and denote this second fusion method as (F_2). Last but not least, we also applied the modality filter to the result of late fusion (F_1) as another fusion method (denoted as (F_3)).

V. Experimental Evaluation

We have tested each of our proposed UMLS-based indexing and retrieval approaches on the medical image collection of the ImageCLEFmed benchmark. This database consists of four public datasets (CASImage [26], MIR [27], PathoPic [28], and PEIR [29]) containing 50 026 medical images with the associated medical reports in three different languages.

In the ImageCLEF 2005 Medical Image Retrieval task (i.e., ImageCLEFmed), 25 topics were selected with the help of a radiologist. Each topic contains at least one of the following axes: anatomy (e.g., heart), modality (e.g., X-ray), pathology or disease (e.g., pneumonia), and abnormal visual observation (e.g., enlarged heart). Each topic is also associated with one or several query images to allow both textual and visual retrieval. An image pool was created for each topic by computing the union overlap of submissions and judged by three assessors to create several assessment sets. The relevance images are the ones judged as either relevant or partially relevant by at least two assessors. The relevance set of images is used to evaluate retrieval performance in terms of uninterpolated mean average precision (MAP) computed across all topics using `trec_eval` tool [30]. In 2005, 134 runs were evaluated on 25 queries with respect to the relevance sets.

We present here the results of nine approaches on these 25 queries to evaluate the benefit of using a UMLS indexing, especially in a fusion framework.

First, three UMLS image indexing were tested on the following visual queries.

(V_1) Global UMLS image indexing and retrieval based on Manhattan distance between two modality indexes;

(V_2) Local UMLS image indexing and retrieval based on Manhattan distance between two LVM histograms;

(V_3) Late fusion of the two visual indexing approaches (V_1) and (V_2).

Second, three purely textual approaches have been tested:

(T_1) UMLS conceptual text indexing and retrieval;

(T_2) UMLS conceptual text indexing and retrieval with Dimension Filtering (DF);

(T_3) UMLS conceptual text indexing and retrieval with DWF.

Finally, the benefit of combining the two sources of information is evaluated on three runs:

(F_1) Late fusion of the two best text and image indexing and retrieval approaches (V_3) and (T_3);

(F_2) Modality visual filtering on the best UMLS conceptual text indexing and retrieval (T_3);

(F_3) Modality visual filtering on the fusion between text and image UMLS indexing (F_1).

Comparative results are given in Table III. For the purely visual retrieval, two of our results (V_1, V_3) are better than the best 2005 (visual only) results (BV), especially when local and global UMLS indexes are mixed (V_3).

Our previous conceptual text-retrieval approaches with dimension filtering were ranked as the top two textual-retrieval methods presented in 2005 (BT : 20.84% and 20.75%, respectively) [7]. The approach is significantly better than the other textual approaches submitted to ImageCLEFmed 2005. Our previous approach used a textual dimension filtering on MeSH terms (DFMT) according to three dimensions: modality, anatomy, and pathology [31], [32]. The use of filtering according to modality, anatomy, and pathology dimensions improve the results significantly. The high average precision is principally due to this textual filtering. We note that the association between MeSH terms and a dimension had to be done manually.

With the use of the UMLS meta-thesaurus, we now have the advantage of accessing these dimensions using the semantic type associated to each UMLS concept. Using the same filtering automatically, we obtain a compatible result (T_2). A slight improvement of 1% with respect to the best result of 2005 (BT) is obtained with an additionnal weighting according the number of matched concept for the three dimensions (T_3).

We verify the benefit of the fusion between image and text: from 22% for the text only (T_3) and 12% (V_3) for the image only, we achieve 28% MAP with a simple fusion of the retrieval results (F_1). Also, the use of a visual filtering according to the modality concept of the textual query improves the results significantly (F_2).

Our best result is obtained using visual filtering on the result of the fusion between text and image retrievals (F_3), leading to an improvement of seven MAP points (i.e., from 28% to 35%) with respect to the best mixed result in 2005 (BM).

In summary, our image retrieval results on the ImageCLEFmed medical database have shown the real potential of the UMLS-based knowledge-assisted indexing approach with different fusion schemes for both images and text.

VI. Conclusion

Medical CBIR is an emerging and challenging research area. We have proposed a retrieval system that represents both texts and images at a very high semantic level using concepts from the UMLS. A structured framework was proposed for designing image semantics from statistical learning. This adaptive framework has been applied to ImageCLEDmed 2004 (CASImage database [26] and 26 queries with only image examples [11]) and 2006 (same image database and 30 queries with both text descriptions and image examples) with excellent performances. The framework is scalable to different image domains (for example, the consumer image domain as demonstrated in [18]) and embraces other design choices such as better visual features, learning algorithms, object detectors, spatial aggregation, and matching schemes when they become available. The flexible framework also allows fusion of global and local similarities to enhance IR results further, as demonstrated by the visual retrieval method (V_3) and other similar schemes for consumer images [25].

We have developed two fusion approaches to improve textual retrieval performance using this UMLS-based image indexes. First, a simple fusion of the two retrieval approaches was proposed, improving significantly the retrieval results of both text and image retrieval for a majority of topics. Second, a visual modality filtering was designed to remove visually aberrant images (i.e., the images whose modalities are different from the textual query modality). Using the ImageCLEFmed 2005 medical database and retrieval task, we have demonstrated the effectiveness of our framework that is very promising when compared with the current state-of-the-art methods.

In the near future, we plan to use the LVM terms from local indexing for semantics-based retrieval (i.e., cross-modal retrieval: processing textual query on LVM-based image indexes) and to complement the similarity-based retrieval [FlexiTile matching, (12)] [31]. We are currently investigating the potential of an early fusion scheme using appropriate clustering methods. A visual filtering based on local information could also be derived from the semantic local LVM indexing. For the visual indexing, currently we only use a two-scale approach: global image-level indexes for GVM terms and local fixed-size patch-level indexes for LVM terms. A natural extension would be to embed the UMLS-based visual indexing approach into a multiscale framework, using appropriate scales for various types of semantic features.

Acknowledgment

The authors would like to thank G. Hanlin and N. Vuillemenot for their contribution to data labeling and part of the implementation. The authors would also like to thank the ImageCLEFmed organizers for making available a large medical image dataset to evaluate our algorithms, and, especially, H. Mueller, W. Hersh, and T. Deselaers for the organization of image retrieval from medical collections. Finally, they would also like to thank T. Joachims for making his SVM^{light} software available.

References

[1] C.-R. Shyu, C. Pavlopoulou, A. C. Kak, C. E. Brodley, and L. S. Broderick, “Using human perceptual categories for contentbased retrieval from a medical image database,” *Comput. Vis. Image Understanding*, vol. 88, no. 3, pp. 119–151, 2002.

[2] H. Muller, N. Michoux, D. Bandon, and A. Geissbuhler, “A review of content-based image retrieval systems in medical applications—Clinical benefits and future directions,” *Int. J. Med. Informat.*, vol. 73, pp. 1–23, 2004.

[3] J. Dy, C. Brodley, A. Kak, L. Broderick, and A. Aisen, “Unsupervised feature selection applied to content-based retrieval of lung images,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 3, pp. 373–378, Mar. 2003.

[4] M. O. Guld, M. Kohnen, D. Keysers, H. Schubert, B. B. Wein, J. Bredno, and T. M. Lehmann, “Quality of dicom header information for image categorization,” in *Proc. Int. Symp. Med. Imaging*, San Diego, CA, 2002, vol. 4685, pp. 280–287, SPIE.

[5] A. Smeulders *et al.*, “Content-based image retrieval at the end of the early years,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1349–1380, Dec. 2000.

[6] Cross Language Evaluation Forum. Nov. 2006 [Online]. Available: http://www.clef-campaign.org/

[7] P. Clough, H. Mller, T. Desealers, M. Grubinger, T. Lehmann, J. Jensen, and W. Hersh, “The CLEF 2005 cross-language image retrieval track,” in *Springer Lecture Notes in Computer Science*. Berlin, Germany: Springer-Verlag, 2005, LNCS 4022, pp. 535–557.

[8] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. Blei, and M. Jordan, “Matching words and pictures,” *J. Mach. Learning Res.*, vol. 3, pp. 1107–1135, 2003.

[9] National Library of Medicine. Jul. 2006 [Online]. Available: http://www.nlm.nih.gov/

[10] V. N. Vapnik, *The Nature of Statistical Learning Theory*. Berlin, Germany: Springer-Verlag, 1995.

[11] J. Lim and J.-P. Chevallet, “Vismed: A visual vocabulary approach for medical image indexing and retrieval,” in *Proc. AIRS*, 2005, pp. 84–96.

[12] G. Serasset, “Interlingual lexical organization for multilingual lexical databases in nadia,” in *Proc. 15th Conf. Computat. Linguistics*, Morristown, NJ, 1994, pp. 278–282.

[13] O. Bodenreider and A. T. Mccray, “Exploring semantic groups through visual approaches,” *J. Biomed. Informat.*, vol. 36, no. 6, pp. 414–432, 2003.

[14] A. Aronson, “Effective mapping of biomedical text to the UMLS metathesaurus: The MetaMap program,” in *Proc. Annu. Symp. Amer. Soc. Med. Informat.*, 2001, pp. 17–21.

[15] H. Schmid, “Probabilistic part-of-speech tagging using decision trees,” in *Proc. Int. Conf. New Methods in Language Process.*, Sep. 1994, pp. 44–49.

[16] G. Salton, A. Wong, and C. Yang, “A vector space model for automatic indexing,” *Commun. ACM*, vol. 18, pp. 613–620, 1975.

[17] J.-P. Chevallet, “X-iota: An open xml framework for ir experimentation application on multiple weighting scheme tests in a bilingual corpus,” in *Proc. AIRS*, Beijing, China, 2004, vol. 3211, pp. 263–280.

[18] J. Lim and J. Jin, “A structured learning framework for content-based image indexing and visual query,” *Multimedia Syst.*, vol. 10, pp. 317–331, 2005.

[19] C. Bishop, *Neural Networks for Pattern Recognition*. Oxford, U.K.: Clarendon, 1995.

[20] Image Retrieval in Medical Applications. Jun. 2006 [Online]. Available: http://phobos.imib.rwth-aachen.de/irma/index en.php

[21] T. Huang, Y. Rui, and S. Mehrotra, “Content-based image retrieval with relevance feedback in mars,” in *Proc. IEEE Int. Conf. Image Process.*, 1997, pp. 815–818.

[22] J. Lim and J. Jin, “Discovering recurrent image semantics from class discrimination,” *EURASIP J. Appl. Signal Process.*, vol. 21, pp. 1–11, 2006.

[23] Svm Light Software. May 2006 [Online]. Available: http://www.svm-light.joachims.org/

[24] T. Joachims, *Learning to Classify Text using Support Vector Machines*. Boston, MA: Kluwer, 2002.

[25] J. Lim and J. Jin, “Combining intra-image and inter-class semantics for consumer image retrieval,” *Pattern Recogn.*, vol. 38, no. 6, pp. 847–864, 2005.

[26] Casimage Medical Image Database. May 2006 [Online]. Available: http://www.casimage.com

[27] Mallinckrodt Institute of Radiology (mir) Nuclear Medicine Teaching File. May 2006 [Online]. Available: http://www.gamma.wustl.edu/home.html

[28] Pathology (Pathopic) Image Collection. May 2006 [Online]. Available: http://www.alf3.urz.unibas.ch/pathopic/intro.htm

[29] Pathology Education Instructional Resource (Peir) Database. May 2006 [Online]. Available: http://www.peir.path.uab.edu

[30] Trec Evaluation Tool. May 2006 [Online]. Available: http://www.trec.nist.gov/trec eval/

[31] J.-P. Chevallet, J. Lim, and S. Radhouani, “A structured visual learning approach mixed with ontology dimensions for medical queries,” in *Accessing Multilingual Inf. Repositories: 6th Workshop Cross-Language Evaluation Forum*, C. P. , Ed. *et al.*, Vienna, Austria, 2006, vol. 4022, LNCS, pp. 642–651.

[32] S. Radhouani, J. Lim, J.-P. Chevallet, and G. Falquet, “Combining textual and visual ontologies to solve medical multimodal queries,” in *Proc. IEEE ICME*, 2006, pp. 1853–1856.

Caroline Lacoste received the engineering degree in mathematical modeling and the M.S. degree in applied mathematics from the National Institute of Applied Sciences (INSA), Toulouse, France, in 2001, and the Ph.D. degree in signal and image processing from the University of Nice—Sophia Antipolis, France, in 2004.

In 2004 and 2005, she was with CREATIS French laboratory (CNRS/INSERM/INSA/UCBL) as a Research and Teaching Fellow. She joined the French-Singapore IPAL Joint Laboratory (CNRS/I2R/UJF/NUS), Singapore, in September 2005 as a Postdoctoral Fellow. Her research interests include image processing, content-based retrieval, stochastic geometry, and medical applications.

Joo-Hwee Lim received the B.Sc. (Hons I) and M.Sc. degrees in computer science from the National University of Singapore, Singapore, and the Ph.D. degree in computer science and engineering from the University of New South Wales.

He joined the Institute for Infocomm Research $(\mathrm{I}^2\mathrm{R})$, Singapore, in October 1990. He has conducted research in connectionist expert systems, neural-fuzzy systems, handwriting recognition, multi-agent systems, and content-based retrieval. He was a key researcher in two international research collaborations, namely the Real World Computing Partnership funded by METI, Japan and the Digital Image/Video Album project with CNRS, France, and School of Computing, National University of Singapore. He also contributed technical solutions to a few industrial projects involving pattern-based diagnostic tools for aircraft and battleship navigation systems and knowledge-based postprocessing for automatic fax/form recognition. He has published more than 80 refereed international journal and conference papers in his research areas including content-based processing, pattern recognition, and neural networks. He is currently the Principal Investigator of several projects in mobile image recognition and medical image retrieval as well as the co-Director of the French-Singapore IPAL Joint Laboratory, Singapore.

Jean-Pierre Chevallet received the B.Sc. and M.Sc. degrees in computer science from Grenoble University, Grenoble, France, the M.Sc. degree by research at the Grenoble Polytechnic Institute, and the Ph.D. degree in computer science from Grenoble University in 1992.

He has been an Associate Professor with Grenoble University France (U. Pierre Mendès-France) since 1993. Since September 2003, he has also been Director of the IPAL CNRS Mixed International Unit between $\mathrm{I^2R}$, and NUS based in Singapore. His research interests are in information retrieval (IR), including natural language processing for information indexing and retrieval, multilingual document indexing, logical model of IR, structured document indexing, and multimedia indexing and retrieval. He has participated in several European projects and working groups, to major information retrieval competitions including TREC, AMARYLLIS, and CLEF. He is the cofounder of the French Association and Conference for Information Retrieval (ARIA and CORIA) and also reviewer for several top IR international conferences. He has contributed more than 60 conference and journal papers in the field of information retrieval.

Diem Thi Hoang Le received the B.Sc. degree in information technology from the University of Natural Sciences, Ho Chi Minh City, Vietnam, in 2002, and the M.Sc. degree in computer science from the University of Joseph Fourier, Grenoble, France, in 2003. She is currently working toward the Ph.D. degree at the University of Joseph Fourier.

She is affiliated with both the CLIPS-IMAG CNRS laboratory and the French-Singapore IPAL Joint Laboratory, Singapore. Her current research area is text information retrieval with a focus on natural language processing and thesaurus application issues.