Medical Image Analysis 18 (2014) 1082–1100

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

# On combining image-based and ontological semantic dissimilarities for medical image retrieval applications

Camille Kurtz [a,b,*], Adrien Depeursinge [a], Sandy Napel [a], Christopher F. Beaulieu [a], Daniel L. Rubin [a]

[a] Department of Radiology, School of Medicine, Stanford University, USA
[b] LIPADE Laboratory (EA 2517), University Paris Descartes, France

ARTICLE INFO

*Article history:*
Received 5 December 2013
Received in revised form 18 June 2014
Accepted 23 June 2014
Available online 2 July 2014

*Keywords:*
Image retrieval
Riesz wavelets
Image annotation
Semantic dissimilarities
Computed tomographic (CT) images

ABSTRACT

Computer-assisted image retrieval applications can assist radiologists by identifying similar images in archives as a means to providing decision support. In the classical case, images are described using low-level features extracted from their contents, and an appropriate distance is used to find the best matches in the feature space. However, using low-level image features to fully capture the visual appearance of diseases is challenging and the semantic gap between these features and the high-level visual concepts in radiology may impair the system performance. To deal with this issue, the use of semantic terms to provide high-level descriptions of radiological image contents has recently been advocated. Nevertheless, most of the existing semantic image retrieval strategies are limited by two factors: they require manual annotation of the images using semantic terms and they ignore the intrinsic visual and semantic relationships between these annotations during the comparison of the images. Based on these considerations, we propose an image retrieval framework based on semantic features that relies on two main strategies: (1) automatic "soft" prediction of ontological terms that describe the image contents from multi-scale Riesz wavelets and (2) retrieval of similar images by evaluating the similarity between their annotations using a new term dissimilarity measure, which takes into account both image-based and ontological term relations. The combination of these strategies provides a means of accurately retrieving similar images in databases based on image annotations and can be considered as a potential solution to the semantic gap problem. We validated this approach in the context of the retrieval of liver lesions from computed tomographic (CT) images and annotated with semantic terms of the RadLex ontology. The relevance of the retrieval results was assessed using two protocols: evaluation relative to a dissimilarity reference standard defined for pairs of images on a 25-images dataset, and evaluation relative to the diagnoses of the retrieved images on a 72-images dataset. A normalized discounted cumulative gain (NDCG) score of more than 0.92 was obtained with the first protocol, while AUC scores of more than 0.77 were obtained with the second protocol. This automatical approach could provide real-time decision support to radiologists by showing them similar images with associated diagnoses and, where available, responses to therapies.



## 1. Introduction

Diagnostic radiologists need to maintain high interpretation accuracy while maximizing efficiency with increasing volumes of images. They are now confronted with the challenge of efficiently and accurately interpreting cross-sectional studies that often contain thousands of images (Rubin, 2000). Currently, this is largely an unassisted process (except for image visualization, volume measurement and for specific image analysis tasks such as lung nodule detection and mammography screening), and a radiologist's accuracy is established through training and experience. Despite this training, there is variation in interpretation among radiologists, and accuracy varies widely (Robinson, 1997). A promising approach to maintain interpretative accuracy in this "deluge" of data is to integrate computer-based assistance into the image interpretation process. Several general-purpose image retrieval systems have been proposed in the literature (Akgül et al., 2011). Among these systems, an emerging technique is content-based image retrieval (CBIR) that could assist users in finding visually similar images within large image collections. For medical

* Corresponding author at: LIPADE (EA 2517), Université Paris Descartes, 45 Rue des Saints-Pères, 75006 Paris, France. Tel.: +33183945807.
*E-mail addresses:* camille.kurtz@parisdescartes.fr (C. Kurtz), adepeurs@stanford.edu (A. Depeursinge), snapel@stanford.edu (S. Napel), beaulieu@stanford.edu (C.F. Beaulieu), dlrubin@stanford.edu (D.L. Rubin).

http://dx.doi.org/10.1016/j.media.2014.06.009
1361-8415/

purposes, the role of CBIR is powerful: in addition to enabling image similarity-based indexing, it could provide computer-aided diagnostic support based both on image content and on other meta-data associated with images.

The main idea of CBIR is to search for similar images based directly on their visual content. This is usually performed by example, where a query image is given as input and an appropriate distance is used to find the best matches in the corresponding feature space (Aigrain et al., 1996). In general, each image is represented by a set of quantitative features, extracted from regions of interest (ROIs) of the image (*e.g.*, lesions), that focus on their specific visual contents (*e.g.*, shape, texture). Although these features are powerful to describe the image in an automated fashion, they are often not specific enough to capture subtle radiological concepts in medical images. In addition, the performance of CBIR systems is often constrained by the low-level properties of these features because they cannot effectively model the user's high-level expectations (Mojsilovic and Rogowitz, 2001) (referred to as the "semantic gap" problem in the literature). Since this problem remains unsolved, current research in CBIR focuses on new methods to characterize the image content with higher levels of semantics, closer to that familiar to the user and more useful in retrieving similar medical cases (Liu et al., 2013).

In recent works on medical image retrieval that includes semantics, the images were characterized using a set of semantic terms (Demner-Fushman et al., 2009; Napel et al., 2010). These semantic terms can be used to describe a variety of information about the image content (*e.g.*, lesion shape or enhancement), and they are directly linked to the user's high-level understanding of image features. Semantic terms can improve diagnosis decision making by enabling radiologists to search image databases for cases that are similar in terms of shared high-level imaging features to the query case. Since they describe the image contents using terminologies used by radiologists to record their observations, they can be considered as robust features for CBIR systems. Thus, incorporating semantics into CBIR systems can be a promising attempt to bridge the semantic gap between the visual description of an image and its meaning (Ma et al., 2010).

Despite many efforts conducted to design innovative CBIR systems (Liu et al., 2007), two issues remain unsolved when using semantic descriptors to characterize medical images. A first issue is the automation of image annotation: usually the semantic descriptors are manually provided by radiologists. Although many approaches have been proposed to predict these semantic features from quantitative features extracted from the image content (Zhang et al., 2012), this automation remains challenging for complex lesions. A second issue is that most of the existing CBIR systems based on semantic features do not consider the intrinsic relations among the semantic terms for retrieving similar images (*e.g.*, relations between imaging observations). Consequently, we distinguish two unmet needs that we propose to address in this article. To this end we present a semantic framework that enables retrieval of similar database images based on their visual and semantic properties.

This article is organized as follows. Section 2 studies the state-of-the-art in semantic image retrieval and presents our research contributions. Section 3 describes our framework for the retrieval of images annotated with semantic terms. Section 4 gathers experiments enabling to assess its relevance. Conclusions will be found in Section 5.

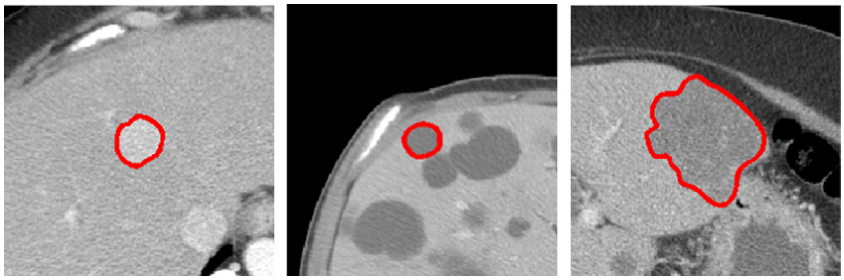
## 2. Image retrieval based on semantic features

### 2.1. Related work

The integration of semantics in medical image retrieval applications has been largely studied in the literature. Among the proposed approaches, "bag-of-words" (BOW) has become a popular representation to describe images with a higher level of semantics. BOW are generic models that have been successfully used in text information retrieval (Voorhees, 1999) to capture a summary of the semantic entities contained in textual documents based on word occurrence probabilities. As an extension of BOW, "bag-of-visual-words" (BOVW) models have been proposed to derive visual words from the image content (Van Gemert et al., 2010). BOVW methods represent an image as a distribution of local salient (or dense) patches, and the visual words are obtained by clustering the feature space populated with local image patches. Once the visual vocabulary has been built, an image is represented as a vector of visual word occurrences, similarly to the BOW approach. This representation is a compromise between low-level image features and high-level semantic features. Different approaches have been proposed to build the visual vocabulary from the image content. The authors of (Yang et al., 2012) have proposed a method for the retrieval of similar liver lesions in contrast-enhanced CT images at different injection phases, where the bags represent the lesion content. In the field of endomicroscopy, a method has been proposed in (Andre et al., 2012) to transform the low-level visual signatures of the images into semantic signatures. The latter reflects how much the presence of each semantic concept is expressed by the visual words describing the video frames. This approach relies on a linear discriminant analysis method that builds visual words based on semantic signatures. The main advantage of such approaches is that they do not require a manual annotation step to describe the image content. However, this advantage often leads to an underlying drawback: the visual vocabulary relies on low-level features and does not correspond to the user's high-level expectations about the semantic content of the image (Liu et al., 2012). In addition, BOVW models do not benefit from the domain knowledge included in semantics of image annotations (Rubin, 2012).

In medical imaging, an image can be characterized by a set of semantic terms, linked to the user's high-level understanding of image features, that accurately describe its visual content (Napel et al., 2010). However, the performance of CBIR systems based on semantic annotations relies on the choice of the terms being used to describe the content of the images: this choice is dependent of the application and the user experience. Consequently, it leads to different possible descriptions for a same image, thwarting good performance of CBIR systems based purely on semantic image descriptions. To deal with this issue, recent works used controlled vocabularies for describing the images (Deng et al., 2009). A controlled vocabulary provides a set of pre-defined terms with definitions that can facilitate the annotation of images since it provides standard terms for describing their visual features. Ontologies, which are related to controlled terminologies but also model explicit specification of relations among semantic terms, provide a formal way to model knowledge (Guarino, 1995). As they are machine-accessible and usually built from a consensus of domain experts, they represent a powerful way to structure semantic terms belonging to a particular knowledge source. In the context of medical imaging, numerous ontologies are being developed to organize biomedical terms in a comprehensive manner (*e.g.*, Medical Subject Headings (MeSH) (Lowe and Barnett, 1994), Unified Medical Language System (UMLS) (Bodenreider, 2004), Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT) (Stearns et al., 2001), Unified Language of Radiology Terms (RadLex) (Langlotz, 2006)).

Semantic terms can be leveraged to tackle the problems raised by the low-level properties of imaging features in CBIR systems. However, medical CBIR applications based on semantic terms are limited since they often require manual annotations of the images, which is a subjective time-consuming process. Different



(a) $\mathcal{I}_A$ (Ovoid). (b) $\mathcal{I}_B$ (Round). (c) $\mathcal{I}_C$ (Irregular).

**Fig. 1.** 3 CT images $\mathcal{I}_{A-C}$ of liver lesions annotated with 3 semantic terms (Ovoid, Round, Irregular) describing the lesion contour. These images could be represented in a 3D space as $\mathcal{I}_A(1,0,0), \mathcal{I}_B(0,1,0)$ and $\mathcal{I}_C(0,0,1)$ where each image is equidistant to every other image. However, we know intuitively that $\mathcal{I}_A$ and $\mathcal{I}_B$ are more similar than $\mathcal{I}_A$ (resp. $\mathcal{I}_B$) and $\mathcal{I}_C$ since the images annotated with the semantic terms ovoid and round are visually and semantically closer than the ones annotated with ovoid (resp. round) and irregular.

approaches have been proposed to automate the annotation process (Rasiwasia et al., 2007). The authors of (Gimenez et al., 2011) have proposed a machine learning strategy based on LASSO (Least Absolute Shrinkage and Selection Operator (Tibshirani, 1996)) regression models for the prediction of image annotations from low-level features (*i.e.*, shape, color) extracted from CT images. A limit of this approach is the difficulty of understanding the correlations between low-level imaging features and the predicted semantic annotations. In recent work (Depeursinge et al., 2014b), semantic terms were predicted from texture features based on Riesz wavelets, whose multi-scale properties are adapted to accurately model the visual signatures of the image annotations. An advantage offered by this approach is the possibility of visually representing the computed visual signatures, which enables understanding the intrinsic correlations between annotations and low-level imaging features.

A common problem raised by BOW and BOVW models is retrieving similar images given an input query image (Allampalli-Nagaraj and Bichindaritz, 2009). Under these models, images are often represented as vectors of values where each element represents the likelihood of appearance of a particular word, and the likeness between images is evaluated by computing the distance between these vectors (Gondra and Heisterkamp, 2008). The classical approaches only compare the contents of the corresponding elements of the vectors using various distance functions (*e.g.*, Manhattan or Euclidean distances). Their main drawback is ignoring the potential relationships between elements. Indeed, most of the classical approaches assume that every high-level term describing an image is independent of other features (no intrinsic relations between the words that are contained in a bag). To deal with this issue, different approaches based on the concept of Bag-of-Visual-Phrases have been recently proposed (Pedrosa and Traina, 2013; López-Monroy et al., 2013). In these approaches an image is represented as a combination of n-consecutive visual words that are constructed by considering the spatial information among visual words. Although these approaches provide encouraging results, they do not directly consider the visual and semantic relations between the visual words. In the case of characterization of medical images with semantic features, the visual and semantic relations among descriptive imaging terms are crucial (Deselaers and Ferrari, 2011) since these features usually have a strong correlation with each other (Fig. 1 presents an example using shape features).

The similarity between semantic features can be evaluated by considering their visual dissimilarity or their semantic dissimilarity (Kesorn et al., 2011). The visual dissimilarity between semantic terms is generally assessed using their image-based dissimilarity. Image-based dissimilarity consists of evaluating the dissimilarity between semantic terms by computing the distance between low-level quantitative features (*e.g.*, shape, color, texture) that characterize these terms in the image (Jiang and Ngo, 2009). Assessing semantic dissimilarity requires evaluating semantic proximity using external sources of linguistic knowledge. To this end, it is possible to use the ontological structure to quantify the semantic dissimilarity between the terms (Batet et al., 2011). Indeed, ontologies specify different kinds of taxonomic relations among the semantic terms (*e.g.*, subtype/supertype, homonyms, synonyms relationship) and can be seen as an oriented graph in which semantic terms are linked by taxonomic relations.

Although several efforts have been conducted to take into account the proximity between semantic features into CBIR applications, most of the proposed approaches focus either on the visual proximity or on the ontological relations between the semantic terms. Consequently they do not focus on the potential of combining these two relations. However, such a combination seems essential when evaluating the likeness between semantic terms: two lesions being similar in appearance may have been annotated by ontologically dissimilar terms while two lesions annotated by ontologically similar terms may look visually dissimilar.[1] During the interpretation of an image, the radiologist considers both visual and semantic dissimilarities. Consequently, there is an opportunity to improve CBIR applications by mimicking this process. To this end, our approach to improving CBIR is to consider both the image-based and semantic relations among the semantic terms when assessing the distance between pairs of images.

Given a dissimilarity relation between semantic terms, the next step is to consider this relation in computing the distance between images described as BO (V) W. In the domain of histogram comparison, "cross-bin" distances have been proposed to compare corresponding vector elements as well as non-corresponding ones (Cha and Srihari, 2002). Based on this property, these distances can consider the relations between the elements of the vectors to be compared. Practically, this can be done by assigning, to each pair of vector elements, a weight (called "ground-distance") modeling their degree of dissimilarity (Niblack et al., 1993). Based on this paradigm, a distance called HSBD (Hierarchical Semantic-Based Distance) has been proposed, relying on a hierarchical merging strategy (Kurtz et al., 2013). Its computation relies on the iterative merging of the semantically closest vector elements to create coarser vectors of higher semantic levels. This hierarchical strategy is the cornerstone of this distance: it enables considering the multilevel semantic correlations among the vector elements and it

[1] For example, the ontological terms homogeneous and heterogeneous both describe texture properties (and thus are close in the RadLex ontology that represents a unified language of radiology terms), but they are visually antagonists.

induces a lower computational cost than quadratic costs required for cross-bin distances. In its original definition, the HSBD distance was proposed for comparing low-dimensional vectors, and the ground-distances (*i.e.*, the dissimilarities) between the elements of the vectors had to be manually defined. Recently, this distance has been extended to the comparison of images described with high-dimensional vectors of semantic terms (Kurtz et al., 2014). In this generic extension, the dissimilarity between semantic terms can be provided by any term dissimilarity measure. Furthermore, the extension of this distance, which has been applied to medical image retrieval, has provided encouraging results.

### 2.2. Research contributions

Based on these methodological considerations, we propose a novel semantic framework, devoted to the retrieval of similar medical database images described with high-level semantic annotations. The contribution of the current study is threefold:

1. We employ an original strategy proposed in (Depeursinge et al., 2014b) to automatically annotate the content of radiological images with semantic terms from a biomedical ontology. This strategy relies on learning, from a database of previously annotated images, specific term visual signatures based on Riesz wavelets that characterize semantic terms. Given a new query image, this strategy avoids the requirement of manual image annotation by the radiologists and it allows the end-users to search databases of images for cases that are similar in terms of high-level visual features;
2. We propose to consider the similarity between semantic terms during the retrieval of similar database images. To this end, we propose a new term dissimilarity measure enabling us to quantify both the image-based relations among the terms, which are provided by their visual signatures, and the semantic relations among the terms, which are automatically evaluated from the structure of the ontology. In this measure, the contributions of the image-based and the ontological dissimilarities are automatically evaluated from data examples using a learning strategy based on maximizing the agreement between the perceptual image dissimilarity provided by the radiologist and the global term dissimilarity value. This strategy provides a potential solution to the limitations of BOW approaches that assume that every term describing an image is independent of other features;
3. We use the HSBD distance proposed in (Kurtz et al., 2014) to enable comparing high-dimensional vectors of semantic features. The main advantage of this vector distance is to consider the multiscale dissimilarities among the semantic terms when comparing images characterized with BOW, with a lower computational cost than the ones induced by the classical approaches in CBIR. The retrieval strategy relies in this study on an extension of the HSBD distance to consider the proposed term dissimilarity measure during the comparison of vectors of semantic features.

The automatic image annotation strategy combined with the proposed radiological image retrieval step can be considered as a potential solution to the semantic gap problem. To show the value of considering the relations among semantic terms in the context of the retrieval of medical images, we apply it to the challenging task of retrieving similar computed tomographic (CT) images of the liver. We evaluate the effectiveness and the gain of considering this framework to retrieve relevant similar images in a database compared to state-of-the-art approaches.

## 3. Methodology

This section describes the proposed semantic framework, dedicated to the retrieval of medical images annotated with semantic terms. Section 3.1 introduces useful notations. The workflow of this framework is presented in Section 3.2. The offline steps composing this workflow are then detailed in Section 3.3 and the online steps are presented in Section 3.4.

### 3.1. Notations and definitions

An interval on $\mathbb{R}$, bounded by $a, b \in \mathbb{R}$, will be noted $[a, b]$ while an interval on $\mathbb{Z}$, bounded by $a, b \in \mathbb{Z}$, will be noted $[[a, b]]$. A set $\mathcal{S}$ of $k$ unordered elements $e_i$ with $i \in [[0, k-1]]$ is denoted by $\{e_0, e_1, \ldots, e_{k-1}\}$. A list (vector) $L$ of $k$ ordered elements $e_i$ with $i \in [[0, k-1]]$ is denoted by $\langle e_0, e_1, \ldots, e_{k-1}\rangle$. A semantic term is noted $x_i$ while its likelihood of presence in an image $\mathcal{I}_A$ is denoted $a_i \in [0, 1]$.

### 3.2. Workflow

The workflow of the proposed CBIR framework is divided into five steps that can be grouped in two phases:

- an *offline* phase (composed of two steps) used to build a visual model of the semantic terms employed to characterize the database images. The first step consists of learning, from the database images, a visual signature for each ontological term. These term signatures are used to (1) predict the image annotations from linear combinations of Riesz wavelets and (2) establish visual "image-based" dissimilarities between the semantic terms. The second step consists of pre-computing the global term dissimilarities using a combination of their image-based and ontological relations;
- an *online* phase (composed of three steps) used to retrieve similar images in the database given a query image. The first step consists of manually delineating an abnormality within the query image to capture the boundary of an ROI. The second step consists of automatic annotation of this image ROI by predicting semantic term likelihood values based on the visual term models built in the offline phase. These "soft" annotations are then summarized into a vector of semantic features modeling the image content. The third step consists of comparing the query image to previously annotated database images by computing the distance between their term likelihood vectors. The vectors are compared using the HSBD distance based on a term dissimilarity measure that leverages both the image-based and ontological term relations computed in the offline phase.

Fig. 4 provides a visual workflow of the offline (orange boxes) and online (blue boxes) phases. Each step of each phase is represented as a box whose content is detailed hereinafter.

### 3.3. Offline phase

#### 3.3.1. Learning of the visual term signatures

In this framework, we use an automatic strategy to predict semantic terms, belonging to an ontology $\Theta$, that characterize the lesion contents. In order to reduce the semantic search space we created pre-defined lists of semantic terms taken from the ontology $\Theta$. An experienced radiologist chose a set of semantic terms that are commonly used to describe the content of the images in a specific application. Among these terms, we selected those describing the margin and the internal texture of the lesions because they are difficult to describe in a reproducible manner

when compared to semantic terms describing overall lesion shape and lesion focality. We denote as $\mathcal{X} = \{x_0, x_1, \ldots, x_{k-1}\}$ with $x_i \in \Theta$ the resulting vocabulary.

Our strategy to predict semantic terms, originally proposed in (Depeursinge et al., 2014b), relies on the automatic learning of the visual term signatures from quantitative texture features derived from the image ROIs (Fig. 4①). Given a set of previously annotated image ROIs (*i.e.*, a training set), this approach automatically learns the image description of each semantic term using support vector machines (SVM) and Riesz wavelets (Depeursinge et al., 2012; Depeursinge et al., 2014a). The multi-directional and multi-scale properties of the Riesz wavelets allow for accurately modeling the visual signatures of the terms from the image content: they can model specific features of each semantic terms at multiple scales from sharp edges (finer scales) to larger nodules (coarse scale). Consequently, our assumption is that features based on Riesz wavelets are well suited to characterize subtle medical imaging concepts related to the description of lesion margin and inner texture in radiological images, which are key aspects of the visual appearance of lesions assessed by radiologists.

Practically, each annotated ROI is divided in a set of $12 \times 12$ image patches extracted from the margin and the internal texture of the lesion. Each patch is characterized by the energies of multi-scale Riesz wavelets and then represents an instance in the feature space. The learning step of this approach relies on linear SVMs, which are used to build visual term signatures in this feature space.[2] The direction vector of the maximal separating hyperplane in one-versus-all configurations defines the visual signature of the corresponding semantic term. Once the visual signatures have been learned, we obtain for each semantic term $x_i \in \mathcal{X}$ a model that characterizes a multi-scale visual description of the term into the image content. The visual signature $\mathbf{\Gamma}_i$ of a term $x_i \in \mathcal{X}$ can be modeled as the direction vector $\mathbf{\Gamma}_i = \langle \Gamma^i_0, \Gamma^i_1, \ldots, \Gamma^i_{U-1} \rangle$ where each $\Gamma^i_u$ represents the weight of the $u$-th Riesz template. The length $U$ of this vector depends on the order $N$ of the Riesz transform and the number $J$ of dyadic scales as: $U = J \cdot (N+1)$. The choice of $N$ and $J$ is presented in the experimental part (Section 4). Given a semantic term, its visual signature is a vector that contains the ability of each texture feature to visually express the presence of the semantic term in the images.

The visual model of a term can be represented by depicting the profiles of its signature related to the linear combination of Riesz templates using a color scale. This visualization strategy provides a means of analyzing comprehensively the correlations between the semantic annotations and their representation in terms of imaging features. Fig. 2 presents examples of visual signature models learned for terms related to the descriptions of liver lesions.

The visual term signature models are used both to predict the likelihood of the presence of the terms for new image ROI (Section 3.4.2) and to establish the image-based dissimilarity between semantic terms (Section 3.3.2).

### 3.3.2. Term dissimilarity assessment

The image retrieval step of our CBIR framework takes into account the term dissimilarity when comparing images described by vectors of semantic terms. We propose in this work to compute the term dissimilarities using both their image-based and ontological relations in an offline step described hereinafter (Fig. 4②).

[2] Our hypothesis is that the semantic terms employed to visually describe the image contents are not mutually exclusive. This means that one image ROI representing a lesion can be annotated by multiple semantic terms, and these terms may even refer to antagonist visual concepts related to the image content (*e.g.*, one lesion may be annotated simultaneously by the terms homogeneous and heterogeneous that describe local texture properties of the lesion). This assumption justifies the fact that the visual signature of each semantic term is learned independently without considering the visual signatures of the other terms composing the vocabulary.

To model the dissimilarity (from 0: equal, to 1: totally different) between the $k$ semantic terms of the considered vocabulary $\mathcal{X}$, we define a $k \times k$ symmetric term dissimilarity matrix $\mathcal{M}^{tdis}$ that contains the intrinsic relations between all the terms of the considered vocabulary $x \in X = \{x_0, x_1, \ldots, x_{k-1}\}$ as

$$\mathcal{M}^{tdis} = \begin{bmatrix} d(x_0,x_0) & \cdots & d(x_0,x_{k-1}) \\ \vdots & \ddots & \vdots \\ d(x_{k-1},x_0) & \cdots & d(x_{k-1},x_{k-1}) \end{bmatrix} \quad (1)$$

where $d \in [0,1]$ represents any function that computes the dissimilarity between two semantic terms $x_i$ and $x_j$. In this work, we use a dissimilarity function based on the combination of image-based dissimilarity provided by the visual signatures of the terms and semantic dissimilarity extracted from the structure of an ontology. These two measures and their combination are presented below.

*3.3.2.1. Image-based term dissimilarity.* The image-based dissimilarity between two semantic terms $x_i, x_j$ can be evaluated by computing the Euclidean distance between their visual signatures $\mathbf{\Gamma}_i, \mathbf{\Gamma}_j$ as

$$d_\Gamma(x_i,x_j) = \frac{\sqrt{\sum_{u=0}^{U-1} |\Gamma^i_u - \Gamma^j_u|^2}}{\omega^\Gamma_{norm}} \quad (2)$$

where $\omega^\Gamma_{norm} = \max_{\forall i,j}\left(\sqrt{\sum_{u=0}^{U-1} |\Gamma^i_u - \Gamma^j_u|^2}\right)$ is a normalization factor evaluating the maximal dissimilarity value between two terms. This dissimilarity models the proximity between the terms according to their visual expression into the image. The cornerstone of this measure is to capture the multi-scale organization of local orientations provided by the Riesz signatures. Fig. 2 presents the idea of evaluating the image-based dissimilarities between term signatures.

*3.3.2.2. Semantic term dissimilarity.* The semantic dissimilarity between two semantic terms $x_i, x_j$ belonging to an ontology can be evaluated by considering edge-based measures that consist of inferring the semantic dissimilarity between terms from the ontology structure (Wu and Palmer, 1994). These measures have been recognized as relevant to accurately quantify the semantic proximity from ontologies in the biomedical context (Lee et al., 2008).

Let $x_i$ and $x_j$ be two semantic terms. We define $path(x_i,x_j) = \{l_0, \ldots, l_{n-1}\}$ as a set of links connecting $x_i$ and $x_j$ in the ontology. Let $|path(x_i,x_j)| \geqslant 0$ be the length of this path. In order to quantify a semantic dissimilarity value between $x_i$ and $x_j$, an intuitive method has been originally proposed in (Al-Mubaid and Nguyen, 2006). It relies on a cluster-based strategy that combines the minimum path length between the semantic terms and the taxonomical depth of the considered branches. The underlying idea is that the longer is the path, the more semantically distant the terms are. Starting from the root of the hierarchy, this measure requires the creation of clusters for each main branch of the ontology (each branch is considered as a cluster of term nodes). The idea is then to assess the common specificity (CS) of two terms by subtracting the depth of their lowest common ancestor (LCA) from the depth $D_c$ of their common cluster (*i.e.*, branch). The common specificity is used to consider that lower level pairs of term nodes are more similar than higher level pairs. In a previous study (Kurtz et al., 2014), we have extended this definition to normalize it and to give an equal weight to the path length and the common specificity features. The proposed definition becomes:

$$d_\Theta(x_i,x_j) = \frac{\log\left(\sqrt{\min_{\forall p}|path_p(x_i,x_j)| \cdot CS(x_i,x_j)} + \gamma\right)}{\omega^\Theta_{norm}} \quad (3)$$

where $\omega^\Theta_{norm} = \max_{\forall i,j}\left(\log\left(\sqrt{\min_{\forall p}|path_p(x_i,x_j)| \cdot CS(x_i,x_j)} + \gamma\right)\right)$ is a normalization factor evaluating the maximal dissimilarity value

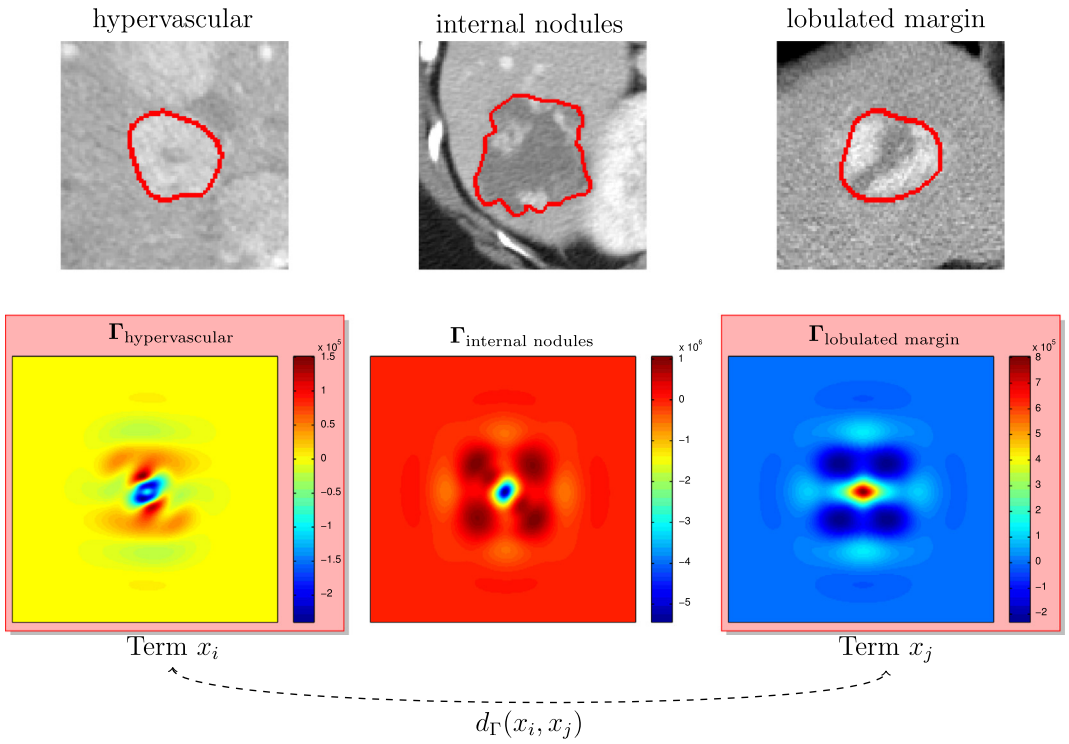



**Fig. 2.** Visual signatures learned for three semantic terms as linear combinations of multi-scale Riesz wavelets. The signatures are learned from a training set of previously annotated image ROIs that is described in Section 4. The color scale shows the profiles of the signatures $\mathbf{\Gamma}_i$ obtained as a weighted sum of Riesz basis templates. Each signature is presented with a representative image ROI (red boundaries) where the term modeled by the signature has a high probability of appearance. This figure also illustrates the idea of computing the image-based dissimilarity $d_\Gamma$ between two semantic terms $x_i$ and $x_j$ to evaluate their visual proximity. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

between two terms and $CS(x_i, x_j) = D_c - depth(LCA(x_i, x_j))$ is the common specificity of the terms. In practice, we set $\gamma = 1$ to force the proposed measure to be positive. Since this measure considers both the common specificity and the relative path length between the semantic terms, it presents powerful properties to evaluate the semantic dissimilarity between two ontological terms. Fig. 3 presents an extract of the RadLex ontology and the idea of evaluating the term dissimilarity from its hierarchical structure.

*3.3.2.3. Combination of image-based and ontological dissimilarities.* To combine both the image-based and the ontological dissimilarities, we define a weighted sum of $d_\Gamma$ and $d_\Theta$ as

$$d^{\rho}_{\Gamma*\Theta}(x_i, x_j) = \rho_{x_i,x_j} \cdot d_\Gamma(x_i, x_j) + (1 - \rho_{x_i,x_j}) \cdot d_\Theta(x_i, x_j) \quad (4)$$

where $\rho_{x_i,x_j} \in [0, 1]$ is a scalar weight factor that can vary for each couple of terms $(x_i, x_j)$. This formulation allows to adapt the contributions of the image-based and ontological dissimilarities for each pair of semantic terms $x_i$ and $x_j$.

To automate the choice of the $\rho_{x_i,x_j}$ values, a learning strategy has been employed to directly extract the optimal weights from the training set. The premise of this strategy relies on the fact that during the interpretation of an image, the radiologist considers both visual and semantic image features. Consequently, our assumption is that when assessing the dissimilarities between image pairs, the opinion of the radiologist will reflect a "perceptual" combination of visual and semantic image dissimilarities. Our strategy to learn the optimal weights is based on mimicking this process by extracting these perceptual combinations. This optional step requires the definition of a reference standard that models the perceptual dissimilarity between pairs of images of the training set. This reference standard can be defined visually by a radiologist by providing examples of reference dissimilarity values between pairs of images, without considering their annotations. Given this reference standard, the underlying idea is to optimize the values of the weights $\rho_{x_i,x_j}$ by maximizing the agreement between the perceptual image dissimilarity value provided by the radiologist and the global term dissimilarity values $d^{\rho}_{\Gamma*\Theta}(x_i, x_j)$. In this way, the optimal perceptual combination of the image-based $d_\Gamma(x_i, x_j)$ and semantic $d_\Theta(x_i, x_j)$ term dissimilarities can be determined.

Let us consider two images $\mathcal{I}_A, \mathcal{I}_B$ from the training set, which have been manually annotated with semantic terms $x_i \in \mathcal{X}$. These two images are modeled by two sets of terms $A, B$. We denote as $D_{\text{percep}}(\mathcal{I}_A, \mathcal{I}_B)$ the perceptual image reference dissimilarity provided by the radiologist between $\mathcal{I}_A$ and $\mathcal{I}_B$. We consider that the perceptual dissimilarity between the pair of images $\mathcal{I}_A$ and $\mathcal{I}_B$ can be used to evaluate the dissimilarity between the sets of terms $A, B$ that describe the two images leading to $D_{\text{percep}}(A, B) = D_{\text{percep}}(\mathcal{I}_A, \mathcal{I}_B)$. By pursuing this reasoning, we make the assumption that the perceptual dissimilarity between the two sets of terms $A, B$ can be reported at the term level, between each couple of terms employed to annotate $\mathcal{I}_A$ and $\mathcal{I}_B$, leading to $d_{\text{percep}}(x_i, x_j) = D_{\text{percep}}(A, B), \forall (x_i, x_j) \in A \times B$ (where $\times$ is the Cartesian product between the two sets of terms $A$ and $B$). The reference standard of dissimilarities between pairs of images $D_{\text{percep}}(\mathcal{I}_A, \mathcal{I}_B)$ can be used to initialize a system of linear equations, where each equation can be defined as $d_{\text{percep}}(x_i, x_j) = d^{\rho}_{\Gamma*\Theta}(x_i, x_j), \forall (x_i, x_j) \in A \times B$. By substituting $d^{\rho}_{\Gamma*\Theta}(x_i, x_j)$ by its formulation (Eq. (4)), we obtain for each equation a reference image dissimilarity value equals to a linear combination of the weight $\rho_{x_i,x_j}$ associated to the image-based and ontological dissimilarity values between couple of terms $x_i$ and $x_j$ belonging to $A$ and $B$:

$$\begin{aligned} d_{\text{percep}}(x_i, x_j) &= d^{\rho}_{\Gamma*\Theta}(x_i, x_j), \forall (x_i, x_j) \in A \times B \\ &= \rho_{x_i,x_j} \cdot d_\Gamma(x_i, x_j) + (1 - \rho_{x_i,x_j}) \cdot d_\Theta(x_i, x_j), \forall (x_i, x_j) \in A \times B \end{aligned} \quad (5)$$

where the weights $\rho_{x_i,x_j}$ are the unknown variables. The weights $\rho_{x_i,x_j}$ can then be obtained as

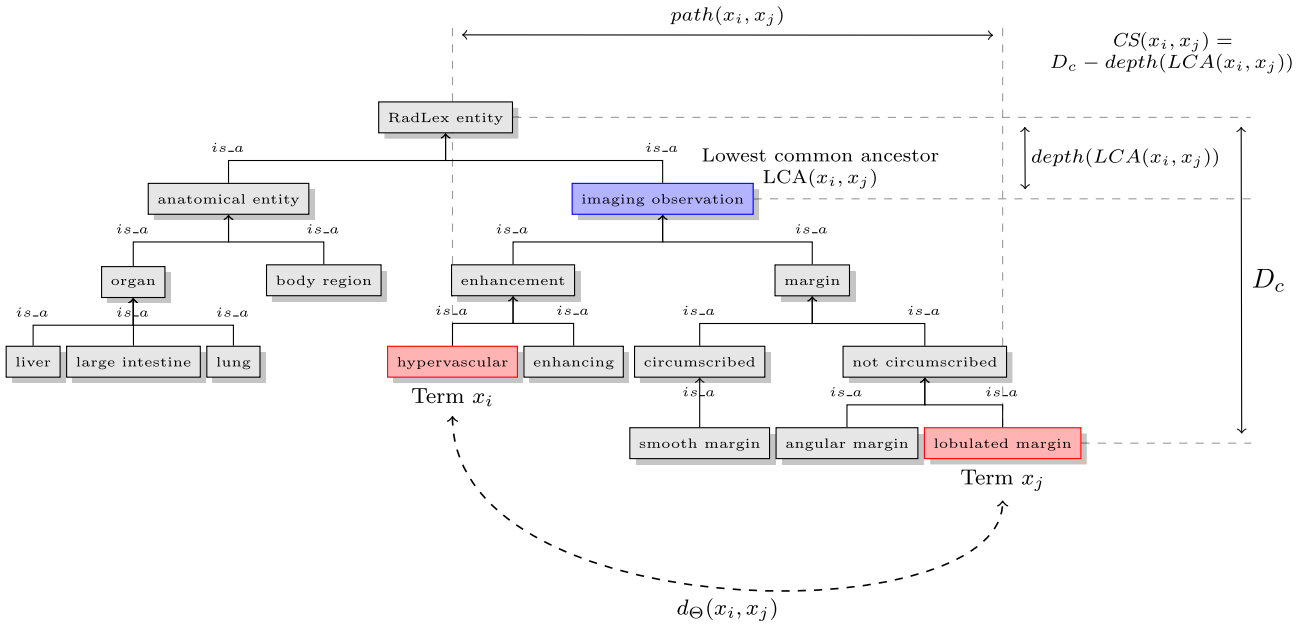



**Fig. 3.** An extract of the RadLex ontology that is a controlled terminology for radiology reporting. This figure also illustrates the idea of using the ontology structure to evaluate the semantic proximity $d_\Theta$ between two terms $x_i$ and $x_j$.

$$\rho_{x_i,x_j} = \begin{cases} \frac{d_{\text{percep}}(x_i,x_j) - d_\Theta(x_i,x_j)}{d_\Gamma(x_i,x_j) - d_\Theta(x_i,x_j)} & \text{if } \min(d_\Gamma(x_i,x_j), d_\Theta(x_i,x_j)) \leqslant d_{\text{percep}}(x_i,x_j) \leqslant \max(d_\Gamma(x_i,x_j), d_\Theta(x_i,x_j)) \\ 1 & \text{else if } |d_{\text{percep}}(x_i,x_j) - d_\Gamma(x_i,x_j)| \leqslant |d_{\text{percep}}(x_i,x_j) - d_\Theta(x_i,x_j)| \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

The two last cases of the equation above are used to constrain the weights $\rho_{x_i,x_j} \in [0,1]$. As the analytical system solution depends directly on the reference dissimilarity values $D_{\text{percep}}(\mathcal{I}_A, \mathcal{I}_B)$ picked from the reference standard to solve the system, the following strategy was then repeated 200 times:

1. a subset of the available image reference dissimilarity values – with a size sufficient to make the solving of the system possible – is randomly selected;
2. a solution of the system is determined using this random subset and the Eq. (6);
3. the weights $\rho_{x_i,x_j}$ are extracted from this solution and stored.

Finally, the mean weights $\overline{\rho_{x_i,x_j}}$ were computed for each pair of semantic terms $(x_i, x_j)$.

### 3.4. Online phase

Once the offline phase has been initialized, this framework enables automatical annotation of a query image and retrieval of similar database images, previously annotated with semantic terms.

#### 3.4.1. Delineation of the abnormality

Let $\mathcal{I}_A$ be a query image. An abnormality in the query image $\mathcal{I}_A$ is first manually identified and delineated to capture the boundary of an ROI (Fig. 4③). The next step is to automatically characterize the visual content of the ROI in terms of respective likelihoods of semantic terms belonging to an ontology $\Theta$.

#### 3.4.2. Automatic annotation of a query image

*3.4.2.1. Soft prediction of the terms from visual signatures.* The visual signatures $\mathbf{\Gamma}_i$ learned during the offline phase (Section 3.3) for each semantic term $x_i \in \mathcal{X}$ are used to automatically annotate the content of the input ROI of the query image $\mathcal{I}_A$ (Fig. 4④). The ROI instance is expressed in terms of the energies $E_u$ of the multi-scale $u$-th Riesz templates as: $\mathbf{\Gamma}_{ROI} = \langle E_0, E_1, \ldots, E_{U-1} \rangle$. The likelihood value $a_i$ of each term $x_i$ is computed as the dot product between the ROI instance $\mathbf{\Gamma}_{ROI}$ and the respective visual signatures $\mathbf{\Gamma}_i$ as

$$a_i = \mathbf{\Gamma}_{ROI} \cdot \mathbf{\Gamma}_i = E_0 \Gamma_0^i + E_1 \Gamma_1^i + \cdots + E_{U-1} \Gamma_{U-1}^i \quad (7)$$

*3.4.2.2. Image characterization with a vector of semantic features.* Once the query image $\mathcal{I}_A$ has been "softly" annotated (Eq. (7)), vector of semantic features can be built as: $A = \langle a_0, a_1, \ldots, a_{k-1} \rangle$, where each element $a_i$ is a numeric value representing the likelihood of the appearance of the semantic term $x_i \in \mathcal{X}$ in the ROI. The $a_i$ values are normalized in $[0,1]$ for all semantic terms. The vector $A$ constitutes a synthetic representation of the query image, which forms the feature clue for retrieval purpose.

#### 3.4.3. Image retrieval with term dissimilarities

Once the query image $\mathcal{I}_A$ has been characterized with a vector of semantic features, this image description can be used to retrieve similar images in the database based on their vector distances (Fig. 4⑤). To this end, the hierarchical semantic-based distance (HSBD) (Kurtz et al., 2013) was extended in (Kurtz et al., 2014) to enable the comparison of vectors of semantic features based on term dissimilarities. The computation of HSBD relies on the iterative merging of the semantically closest vector elements to create coarser vectors of higher semantic levels.

Before actually computing the distance between two vectors $A$ and $B$ that characterize $\mathcal{I}_A$ and $\mathcal{I}_B$, it is necessary to define a way to hierarchically merge the different semantic terms into clusters (*i.e.*, instances of higher semantic levels). To this end, it is possible

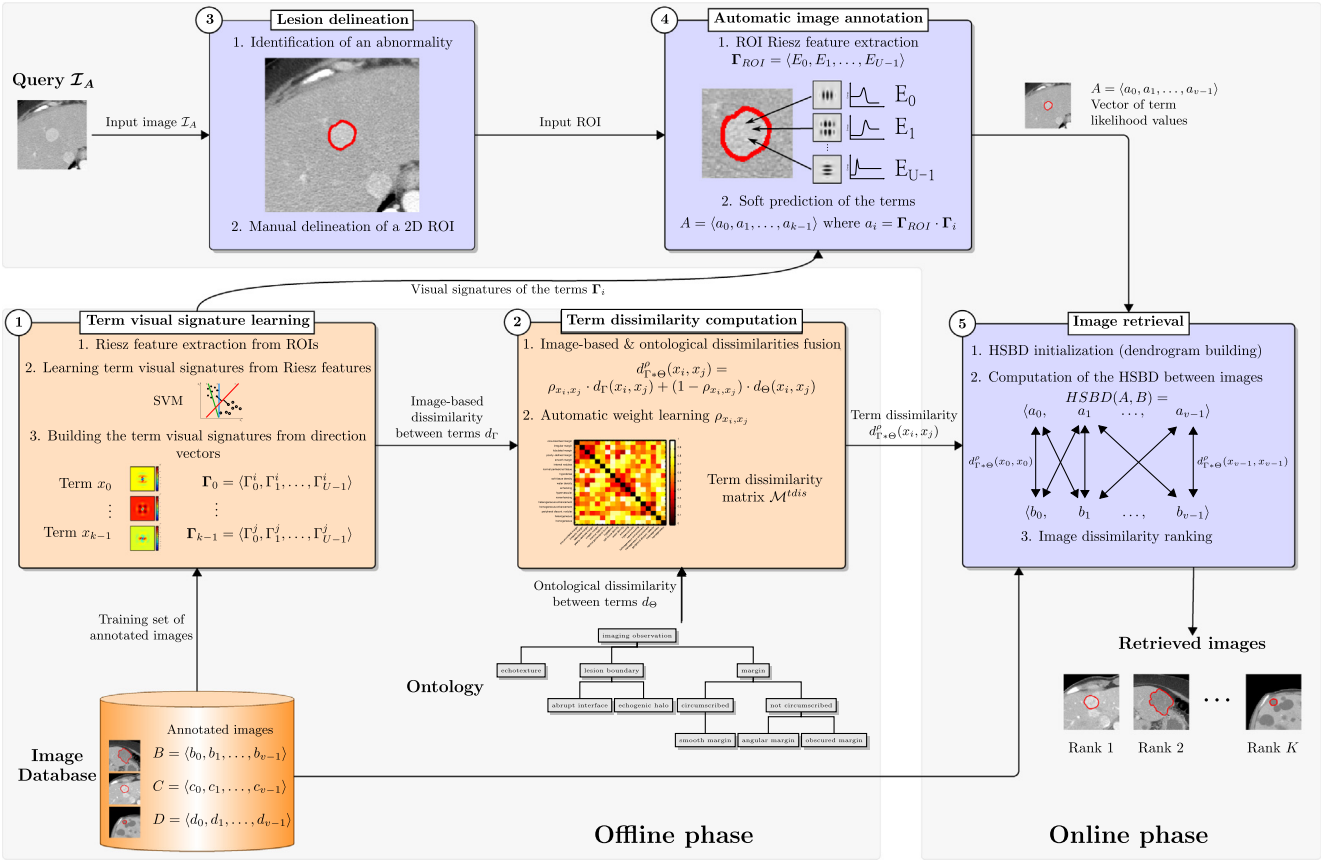



**Fig. 4.** Workflow of the proposed semantic framework for image retrieval. Orange boxes represent offline steps while blue boxes represent online steps. The content of each box is detailed in Section 3. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

to build a dendrogram $\mathfrak{D}$ (modeling this merging hierarchy) induced by $\mathcal{M}^{tdis}$ (Eq. (1), computed during the offline phase with the $d^{\rho}_{\Gamma * \Theta}$ visual-ontological term dissimilarity values). This dendrogram is obtained using the ascendant hierarchical clustering (AHC) algorithm (Ward, 1963). The AHC algorithm hierarchically builds clusters of semantic terms while minimizing their intra-group inertia. Each $s$ stage of $\mathfrak{D}$ corresponds to a particular semantic level. Such a hierarchy models an order of fusions between the semantic terms in a multilevel fashion, relatively to their image-based and ontological dissimilarities.

Once the dendrogram $\mathfrak{D}$ has been built, HSBD can be computed. This computation is organized in two steps:

- *Step 1. Hierarchical element-to-element sub-distances computation:* During an iterative merging process (scanning each stage of the dendrogram from the leaves to the root), the vectors $A^v$ and $B^v$ associated to $A$ and $B$, which are induced by the merging of the semantic terms composing each cluster of the stage $S_v$, are built. After each iteration, the classical Manhattan "sub-distance" $D_{L_1}$ is then computed between the pair of (coarser) vectors $A^v$ and $B^v$ created previously. The resulting series of element-to-element sub-distances $\mathbf{D}^0_{L_1}, \ldots, \mathbf{D}^{s-1}_{L_1}$ enables assignment of vector distances at different semantic levels.
- *Step 2. Element-to-element sub-distances fusion:* The sub-distances $\mathbf{D}^v_{L_1}$ computed for all the stages of $\mathfrak{D}$, and the "energy of merging" required to go from one stage to the next $h_{\mathfrak{D}}(v)$ (the dissimilarity value computed by AHC), are then fused into a function $\mathbf{D}^{inter}_{L_1}$ which is integrated to provide the HSBD distance. Practically, this distance can be obtained by computing the area $\mathcal{A}_{\mathbf{D}^{inter}_{L_1}}$ under the curve representing the function $\mathbf{D}^{inter}_{L_1}$.

The HSBD distance between two vectors $A$ and $B$ is then defined as:

$$\mathrm{HSBD}(A,B) = \frac{1}{2}\sum_{v=0}^{s-2}\left[\left(\mathbf{D}^{v+1}_{L_1} + \mathbf{D}^{v}_{L_1}\right)\left(h_{\mathfrak{D}}(v+1) - h_{\mathfrak{D}}(v)\right)\right] \quad (8)$$

Details about the computation of the HSBD distance were recently described in (Kurtz et al., 2013). In this paper, we demonstrate that the computational complexity of HSBD is directly linked to the number $s$ of stages in the dendrogram $\mathfrak{D}$. If $\mathfrak{D}$ is a "flat"(*i.e.*, a 2-stage) dendrogram ($s = 2$) or a totally balanced dendrogram ($s = \log_2(k)$), the complexity of HSBD becomes $\Theta(k)$. If $\mathfrak{D}$ is a totally unbalanced dendrogram ($s = k$), then the complexity of HSBD becomes $\Theta(k^2)$. Consequently, the complexity of HSBD is lower in average than $\Theta(k^2)$, namely the complexity of most state-of-the-art "cross-bin" distances.

## 4. Experimental study

### 4.1. Experiments and user workflow

To assess our semantic framework, we applied it in a system for retrieving liver lesions from a database of CT images. Liver lesions stem from a variety of diseases, each with different (though sometimes overlapping) visual manifestations. Some liver lesions are benign while others may be malignant. The ability to differentiate

these lesions efficiently and accurately is important to patient treatment and outcome.

Given a new CT image corresponding to a new imaging case/patient, the first step for the end-user is to select the best slice showing the lesion and then to delineate (manually or automatically with a segmentation software) an ROI around the lesion. The user can then enter this image ROI into the system, which will automatically predict the semantic terms that are the best suited to visually describe the lesion content. Given these semantic image annotations, the proposed semantic framework will retrieve, in a database of previously annotated images, and rank by similarities, the images that are the most similar (semantically and visually) to the query image. This task is crucial since it can help radiologists to query the database to find similar medical cases and patient cohorts based on the visual appearance of the images.

### 4.2. Material

We received institutional review board approval for retrospective analysis of deidentified images. We used 72 CT images of liver lesions (with one lesion per targeted image) in the portal venous phase acquired from 44 patients, including six types of lesion diagnoses (Table 1) that were used in a previous research study (Korenblum et al., 2011). These types of lesions are common and span a range of visual appearances in CT. Scans were acquired during the time period February 2007 and August 2008 and used the following range of parameters: 120 kVp, 140–400 mAs, and 5 mm slice thickness. For each scan, the axial slice with the largest lesion area was selected for analysis. Fig. 5 shows representative examples of this dataset with different lesion diagnoses.

Our approach requires that lesions on CT images be delineated by a 2D ROI. In this experimental study, a radiologist (C.F.B., 15 years of abdominal CT experience) drew an ROI around the lesion on these images (Fig. 5). This task led to 72 individual ROIs that were used as input to our semantic image retrieval framework.

### 4.3. Offline phase: experiments and results

#### 4.3.1. Learning the visual term signatures

Once the lesions have been circumscribed by ROIs, the next step is to annotate the ROIs using semantic terms belonging to an ontology. The considered approach to learn the visual term signatures $\mathbf{\Gamma}_i$ from Riesz visual features required the creation of a training set of manually annotated ROIs. To build the training set, each lesion was annotated by a radiologist with an initial set of semantic terms from the RadLex ontology (Langlotz, 2006). We consider that the semantic terms are not mutually exclusive. Among these semantic terms, we selected and conserved those describing the margin and the internal texture of the lesions. The resulting vocabulary was composed of 18 terms that are presented in Table 2. The radiologist used the electronic physician's annotation device (ePAD) system for annotating the images (Rubin et al., 2008). Each annotated lesion was used to feed our system that builds a visual signature from Riesz visual features for each one of the 18 semantic terms.

**Table 1**
The types of diagnoses and the number of lesions.

| Diagnosis type | # Of lesions |
|---|---|
| Cyst | 21 |
| Metastasis | 24 |
| Hemangioma | 13 |
| Hepatocellular carcinoma | 6 |
| Focal nodular hyperplasia | 5 |
| Abscess | 3 |
| Total | 72 |

In practice, the number of scales was chosen as $J = \lfloor \log(12) \rfloor = 3$ to cover the full spatial spectrum of $12 \times 12$ patches. The order of the Riesz transform $N = 8$ was used, which a previous study showed this provides an excellent tradeoff between computational complexity and the degrees of freedom of the filterbanks (Depeursinge et al., 2014a). To avoid any bias during the learning of the visual term signatures, a specific cross-validation strategy was considered. This strategy is detailed hereinafter in Section 4.4.1.

Fig. 2 depicts three visual signature models that have been learned for three particular semantic terms (hypervascular, internal nodules, lobulated margin).

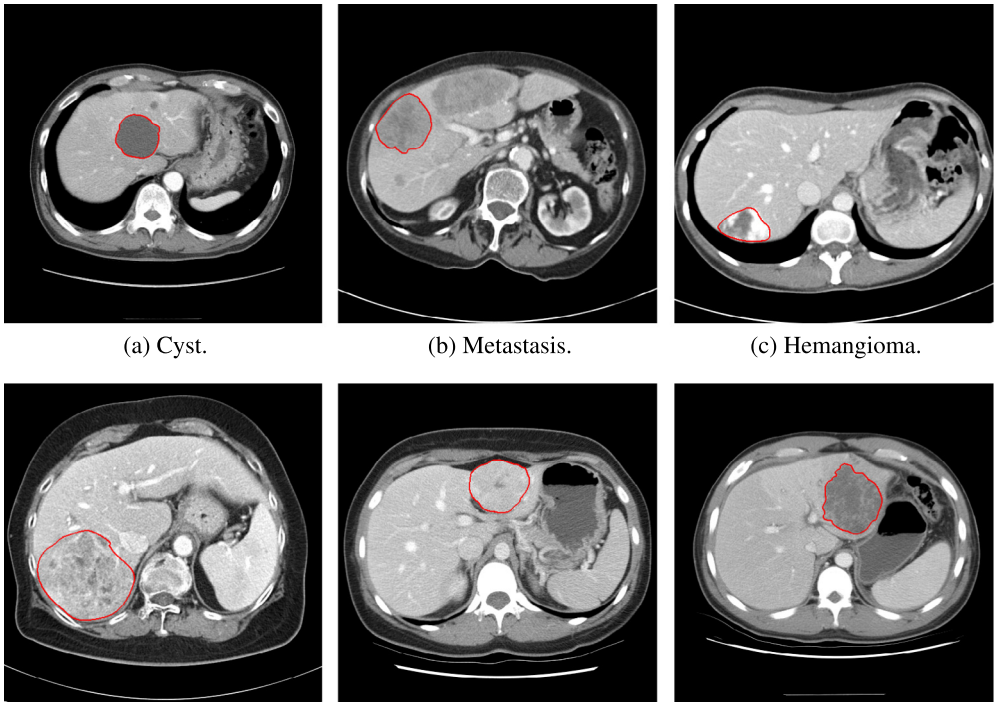
#### 4.3.2. Term dissimilarity computation

*4.3.2.1. Dissimilarity measures.* To assess independently the contributions of the image-based and the semantic term dissimilarities in the proposed term dissimilarity measure, four experiments were carried out (Table 3). In a first experiment, we only considered the image-based term dissimilarity $d_\Gamma$ provided by the difference between the visual signatures that characterized the terms in the images (Eq. (2)). In a second experiment, we only considered the semantic term dissimilarity $d_\Theta$ computed using the RadLex ontological relations (Eq. (3)). In a third experiment, we used the combination of the image-based and the semantic term dissimilarities $d^{\rho}_{\Gamma*\Theta}$ (Eq. (4)). In this experiment we assigned an equal weight to the two term dissimilarity measures ($\rho = \frac{1}{2}$). We denote the resulting term dissimilarity measure as $d^{1/2}_{\Gamma*\Theta}$. In the last experiment we used the local weighted version of $d^{\rho}_{\Gamma*\Theta}$. We explained hereinafter how the weight learning approach has been initialized.

*4.3.2.2. Learning the weights of the term dissimilarity measure.* We used the learning strategy (Section 3.3.2) to automatically determine the weights $\rho$ of the term dissimilarity measure defined in Eq. (4). This learning strategy requires a reference standard that models the perceptual dissimilarity between pairs of images belonging to the dataset. The reference standard used in these experiments is described in Section 4.4.4. To avoid any overfitting problem, we used a leave-one-out cross-validation strategy. We withheld one of the $M$ images considered in the reference standard and we used the learning strategy described previously with the remaining $M-1$ images and computed optimal values for the weights to be used in the term dissimilarity measure. This experiment was repeated for all the images composing the dataset. Given the small size of the reference standard (a subset of $M = 25$ CT images described previously), it appears that some of the terms belonging to the vocabulary $\mathcal{X}$ have not been used to annotate the images belonging to the reference standard. This makes it impossible to use the learning strategy to determine the weights of the term dissimilarity measure for some particular couples of terms. In these particular cases (21% of the possible couples of terms), we assigned an equal weight ($\rho = \frac{1}{2}$) to the image-based and the semantic term dissimilarity measures.

Fig. 6 presents a subset of the weights learned for 15 pairs of semantic terms. For pairs of purely descriptive terms, which characterize lesion internal texture (*e.g.*, heterogeneous vs. homogeneous, line 15), the learning algorithm gives a heavier weight to the image-based dissimilarity than to the ontological dissimilarity. On the contrary when the terms described more complex visual concepts including domain knowledge (*e.g.*, normal perilesional tissue vs. homogeneous enhancement, line 8), the learning algorithm gives a heavier weight to the ontological dissimilarity.

*4.3.2.3. Matrices of term dissimilarity values.* The four term dissimilarity measures $d_\Gamma, d_\Theta, d^{1/2}_{\Gamma*\Theta}$ and $d^{\rho}_{\Gamma*\Theta}$ (Table 3) were then used to fill four individual term dissimilarity matrices (Eq. (1)) containing the dissimilarity values between each possible pair of semantic terms among the 18 RadLex terms.



(a) Cyst. (b) Metastasis. (c) Hemangioma.

(d) Hepatocellular carcinoma. (e) Focal nodular hyperplasia. (f) Abscess.

**Fig. 5.** 6 CT images of liver lesions in the portal venous phase. The boundaries of the lesions are highlighted in red. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

**Table 2**
RadLex terms used to describe the appearance of the liver lesions from CT scans. The 18 semantic terms describing the margin and the internal textures of the lesions are marked in bold. The ontology tree associated with these semantic terms is available online at: http://bioportal.bioontology.org/ontologies/RADLEX.

| Category | Semantic term |
|---|---|
| Lesion margin and contour | **Circumscribed margin** |
| | **irregular margin** |
| | **lobulated margin** |
| | **poorly-defined margin** |
| | **smooth margin** |
| lesion substance | **internal nodules** |
| perilesional tissue | **normal perilesional tissue** |
| lesion focality | solitary lesion |
| | multiple lesions 2–5 |
| | multiple lesions 6–10 |
| | multiple lesions > 10 |
| lesion attenuation | **hypodense** |
| | **soft tissue density** |
| | **water density** |
| overall lesion enhancement | **enhancing** |
| | **hypervascular** |
| | **nonenhancing** |
| spatial pattern of enhancement | **heterogeneous enh.** |
| | **homogeneous enh.** |
| | **peripheral nodular enh.** |
| kinetics of enhancement | centripetal fillin |
| | homogeneous retention |
| | homogeneous fade |
| lesion uniformity | **heterogeneous** |
| | **homogeneous** |
| overall lesion shape | round |
| | ovoid |
| | lobular |
| | irregularly shaped |
| lesion effect on liver | **abuts capsule of liver** |

These matrices are depicted as color heatmaps in Fig. 7. From these heatmaps, one can note that the image-based dissimilarity provides a different (but complementary) information than the semantic dissimilarity: for instance, density terms (*e.g.*, hypodense, water density) and homogeneity terms (*e.g.*, homogeneous, heterogeneous) are ontologically similar while they are visually dissimilar. By studying the impact of the optimized weights learned for pairs of semantic terms (Fig. 6) on the term dissimilarity matrix computed with $d^{\rho}_{\Gamma_*\Theta}$ (Fig. 7(d)), we deduce some conclusions about the visual perception of the physicians. The value $\rho = 0.79$ for the pair {internal nodules, heterogeneous} with $d_\Gamma < d_\Theta$ may mean that two images annotated with the terms internal nodules and heterogeneous, respectively, are perceived as dissimilar by physicians. The value $\rho = 0.24$ for the pair {homogeneous enhancement, homogeneous} with $d_\Gamma < d_\Theta$ may mean that two images annotated with the terms homogeneous enhancement and homogeneous, respectively, are perceived as similar by physicians.

### 4.4. Online phase: experiments and results

Once the offline phase has been initialized, we evaluated the ability of our framework to automatically annotate query lesions and rank CT database images in order of visual similarity.

#### 4.4.1. Automatic annotation of the lesions

Once the term-specific signatures $\mathbf{\Gamma}_i$ were learned, they were used to predict the likelihoods of semantic terms describing the lesion content of query images. Given the small size of our database, our goal is to learn as much as possible from the data. Consequently, we use the same database for training and testing, but we take steps to avoid biasing the results. If we only perform a leave-one-out cross-validation, the hypothesis of independence is not respected, because several images are acquired on the same patient. In order to avoid producing bias, we perform a leave-one-patient-out cross-validation: The images acquired from one of the 42 patients were withheld. The Riesz features were extracted from the withheld lesions and the visual signature models learned with the remaining patient images were used to predict the

**Table 3**
Dissimilarity measures used for term comparisons. These combinations enable to assess independently the contributions of the image-based and the semantic dissimilarities between the terms during the image retrieval step.

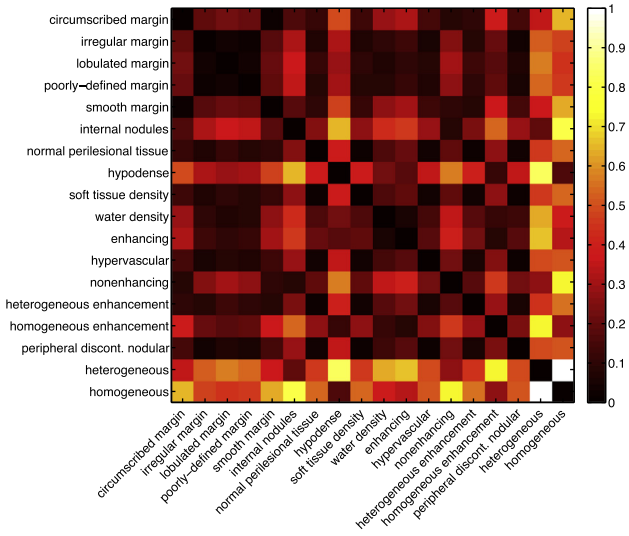
| Term dissimilarity measure | Equation | Definition |
|---|---|---|
| $d_\Gamma(x_i, x_j)$ | Eq. (2) | image-based |
| $d_\Theta(x_i, x_j)$ | Eq. (3) | ontological |
| $d^{1/2}_{\Gamma*\Theta}(x_i, x_j) = \frac{1}{2} \cdot d_\Gamma(x_i, x_j) + \frac{1}{2} \cdot d_\Theta(x_i, x_j)$ | Eq. (4) | image-based + ontological (equal contribution) |
| $d^{\rho}_{\Gamma*\Theta}(x_i, x_j) = \rho_{x_i x_j} \cdot d_\Gamma(x_i, x_j) + (1 - \rho_{x_i x_j}) \cdot d_\Theta(x_i, x_j)$ | Eq. (4) | image-based + ontological (weighted version) |


**Fig. 6.** Example of learned weights of the $d^{\rho}_{\Gamma*\Theta}$ term dissimilarity measure for 15 pairs of semantic terms. The weight $\rho$ is represented as a green dot on a $[0, 1]$ scale. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

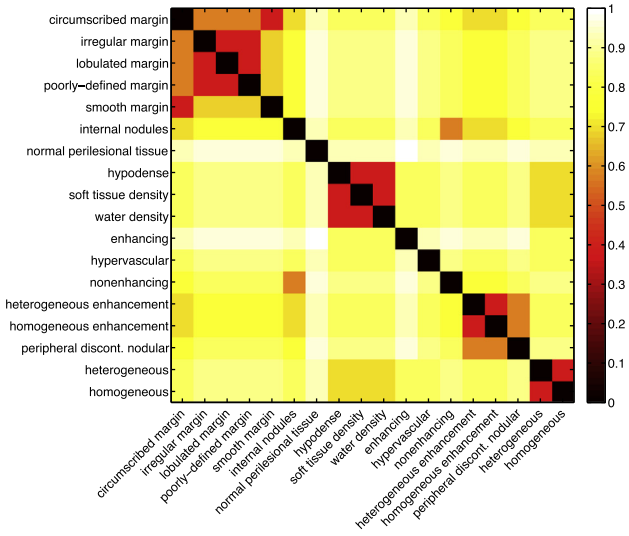
likelihood that a semantic term appears in the lesion content. This experiment was repeated for all the patient images in the dataset. For each image, a semantic feature vector of length 18 was created to indicate the likelihood of appearance of specific observations. The presence of a term did not imply the absence of all others because our system for the automatic image annotation relies on the assumption that the semantic terms employed to visually describe the image contents are not mutually exclusive.

The performances of the term prediction strategy were analyzed using the area under the receiving operator characteristic (AUC) curves. The global AUC for all 18 semantic terms is $0.76 \pm 0.004$. These results are detailed and interpreted in a recent specific study (Depeursinge et al., 2014b).

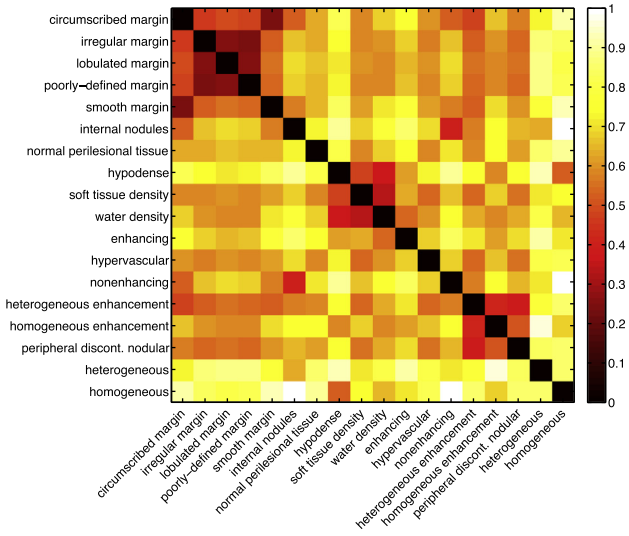
Fig. 8 presents the mean vectors of semantic term likelihood values obtained for the four most represented lesion diagnoses. From these diagrams, one can note that cysts are mainly characterized by homogeneous texture, circumscribed margin and nonenhancing properties while metastases are characterized by homogeneous texture, circumscribed margin and homogeneous enhancement. Hemangiomas are mainly characterized by internal nodules, irregular margin and heterogeneous texture. Hepatocellular carcinomas are characterized by soft tissue density, enhancing properties and internal nodules. We can observe that the hepatocellular carcinoma lesions are simultaneously annotated with the terms homogeneous and heterogeneous that describe the texture properties of the lesion. This may mean that these lesions present



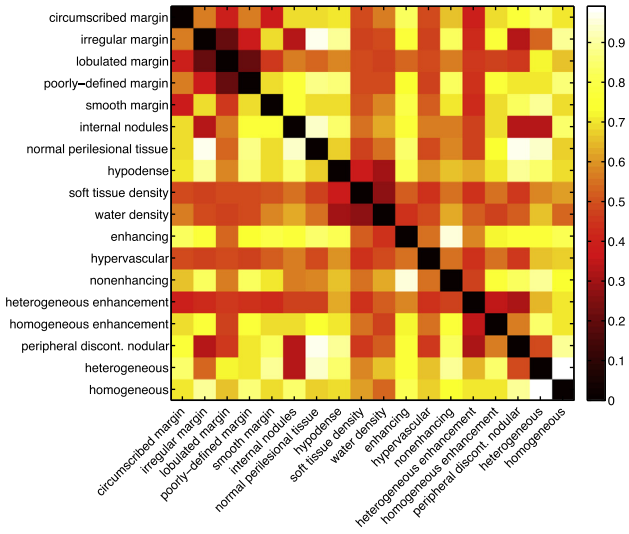

(a) Image-based term dissimilarities computed with $d_{\Gamma}$.




(b) Semantic term dissimilarities computed with $d_{\Theta}$.




(c) Combination of image-based and semantic dissimilarities computed with $d^{1/2}_{\Gamma*\Theta}$.




(d) Combination of image-based and semantic dissimilarities computed with $d^{\rho}_{\Gamma*\Theta}$.

**Fig. 7.** Matrices (heatmaps) modeling the term dissimilarities among the 18 RadLex terms used to characterize the lesions. The color scale varies from 0 (black – similar) to 1 (white – dissimilar).

local textural properties: some specific parts of the inner texture of the lesions may be homogeneous while some other parts may be heterogeneous.

*4.4.2. Initialization of HSBD*

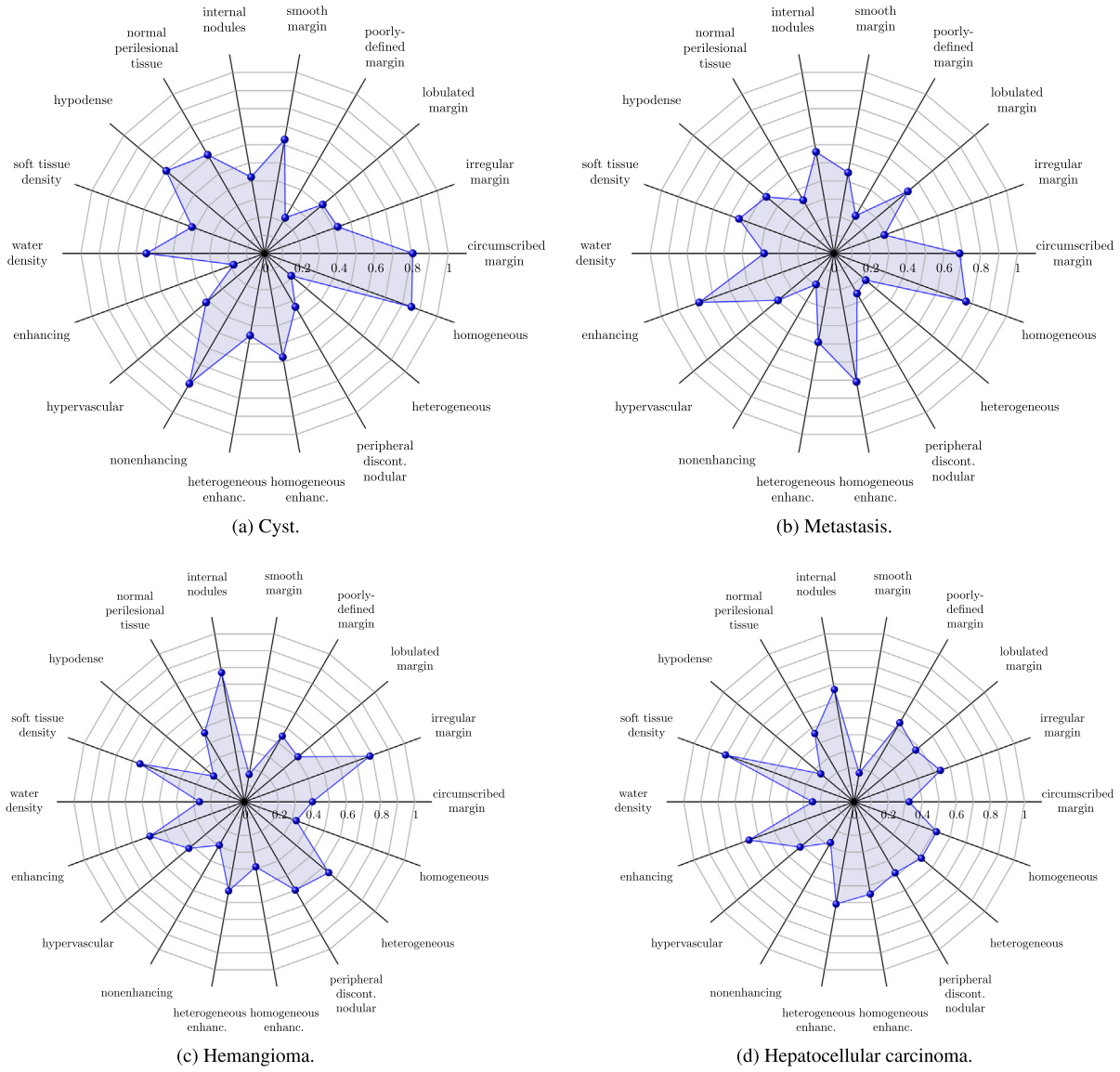
The computation of the HSBD distance between two vectors, requires the definition of a merging hierarchy which models the order of the fusions among the closest elements of the vectors to be compared. From the matrices presented in Fig. 7, four different dendrograms were automatically built leading to four different initializations of HSBD: $\text{HSBD}_{d_{\Gamma}}$, $\text{HSBD}_{d_{\Theta}}$, $\text{HSBD}_{d^{1/2}_{\Gamma*\Theta}}$ and $\text{HSBD}_{d^{\rho}_{\Gamma*\Theta}}$. These initializations enable to evaluate the interest of considering different term dissimilarity functions for assessing the term correlations.

*4.4.3. Comparison to other distances*

To assess the relevance of the proposed semantic framework, and in particular the use of the HSBD distance for image comparison, we compared the results obtained with this distance to the results obtained by using other existing distances (Table 4) that are detailed hereinafter.

To highlight the benefits of considering the image-based and semantic relations between the features composing the vectors, our results were compared to the results obtained with the classical Manhattan $D_{L_1}$ and Euclidean $D_{L_2}$ distances that do no take into account the relations among the features.

To show the interest of using the HSBD distance instead of using an alternative distance that considers the feature relations, our results were also compared to the ones obtained with the earth mover's distance (EMD) (Rubner et al., 2000). Its principle is to estimate the cost of mapping two vectors. In the EMD, two features are considered as the "earth" and the "holes", respectively. Then, the distance measure problem is transformed into the earth moving problem, where the minimum cost of moving all the "earth" into the "holes" (relatively to the term proximity values) is calculated. The EMD distance was initialized by varying the term dissimilarity




(a) Cyst. (b) Metastasis.

(c) Hemangioma. (d) Hepatocellular carcinoma.

**Fig. 8.** Term likelihood values predicted by the automatic annotation step for the four most represented diagnoses of the dataset. Each diagram represents the mean vector of the term likelihood values for all lesions of a particular diagnosis. This visualization strategy has been proposed in (Andre et al., 2012).

measure used to consider the feature relations leading to four different initializations: $\mathrm{EMD}_{d_\Gamma}$, $\mathrm{EMD}_{d_\Theta}$, $\mathrm{EMD}_{d^{1/2}_{\Gamma*\Theta}}$ and $\mathrm{EMD}_{d^{\rho}_{\Gamma*\Theta}}$.

#### 4.4.4. CBIR evaluation and results

We considered two protocols to evaluate the retrieval performance: the first protocol consists of evaluating the ranking results relatively to a dissimilarity reference standard defined for image pairs while the second protocol consists of evaluating these results relatively to the diagnoses of the retrieved images.

*4.4.4.1. Evaluation with a reference standard.* As it was not conceivable in this study to generate a dissimilarity reference standard for the $M = 72$ considered images (would have required 2926 manual comparisons), we used a subset of the reference standard of image dissimilarity that has been proposed in (Napel et al., 2010), enabling us to evaluate image retrieval performances by using the semantic features. This reference standard is composed of a subset of $M = 25$ CT images described previously (8 cysts, 7 hemangiomas and 10 metastases). These 25 images were selected based on being the first set of cases collected. Two radiologists viewed each pair of images twice and reached a consensus opinion on a dissimilarity measure for the image pair (1, very similar; 2, similar; 3, not similar) by visually and globally addressing dissimilarity of texture, boundary shape, and sharpness. They did not consider size or location within the liver, nor did they consider any clinical data that might have implied a specific lesion type. They strived to base their evaluation purely on image appearance without considering the semantic annotations of the images.

We used normalized discounted cumulative gain (NDCG) (Järvelin and Kekäläinen, 2002) to evaluate performance. The NDCG index is a standard technique used to measure the effectiveness of information retrieval algorithms when ground truth is available, as represented by our three-point dissimilarity scale defined previously. NDCG is used to measure the usefulness (gain) on a scale of 0 to 1 of $K$ retrieved lesions on the basis of their positions in the ranked list compared with their dissimilarity to the query lesion according to a separate reference standard. The discounted cumulative gain (DCG) is evaluated with the weight of each retrieved lesion discounted at lower ranks. The DCG at a particular rank position $K$ is defined as:

**Table 4**
Distances used for comparison. EMD and HSBD are initialized with 4 term dissimilarity measures (Table 3).

| Symbol | Reference | Distance | Initialization | | | | Term relations | |
|---|---|---|---|---|---|---|---|---|
| $D_{L_1}$ | Cha and Srihari (2002) | Manhattan distance | $D_{L_1}(A,B)=\sum_{i=0}^{k-1}\lvert a_i-b_i\rvert$ | | | | } | **No** |
| $D_{L_2}$ | Cha and Srihari (2002) | Euclidean distance | $D_{L_2}(A,B)=\sqrt{\sum_{i=0}^{k-1}\lvert a_i-b_i\rvert^2}$ | | | | | |
| EMD | Rubner et al. (2000) | Earth Mover's Distance | $\text{EMD}_{d_\Gamma}$ | $\text{EMD}_{d_\Theta}$ | $\text{EMD}_{d^{1/2}_{\Gamma*\Theta}}$ | $\text{EMD}_{d^{\rho}_{\Gamma*\Theta}}$ | } | **Yes** |
| HSBD | Kurtz et al. (2013) | Hierarchical Semantic-Based Distance | $\text{HSBD}_{d_\Gamma}$ | $\text{HSBD}_{d_\Theta}$ | $\text{HSBD}_{d^{1/2}_{\Gamma*\Theta}}$ | $\text{HSBD}_{d^{\rho}_{\Gamma*\Theta}}$ | | |

$$\text{DCG}_K=\sum_{i=1}^{K}\frac{2^{rel_i}-1}{\log_2(i+1)} \tag{9}$$

where $rel_i$ is the graded dissimilarity of the result at position $i$ in the ranked list. Comparing a search engine's performance from one query to the next cannot be consistently achieved using DCG alone, so the cumulative gain at each position for a chosen value of $K$ should be normalized across queries. This is done by sorting retrieved images of a result list by relevance (relatively to the reference standard), producing the maximum possible DCG till position $K$, also called Ideal DCG (IDCG). For a query, the NDCG is then computed as:

$$\text{NDCG}_K=\frac{DCG_K}{IDCG_K} \tag{10}$$

Thus, for a given $K$, higher NDCG ($K$) means more lesions similar to the query image are ranked ahead of dissimilar ones, with NDCG ($K$) equal to 1 implying perfect retrieval of $K$ images.

*4.4.4.2. Results.* We withheld each image from the 25-images dataset and ranked the remaining 24 images according to HSBD and to the other distances used for comparison (Table 4). For each query image, the mean NDCG value was computed at each $K\in[[1,M-1]]$ ($M=25$). Fig. 9 shows the NDCG scores obtained for the different distances considered. From this graph, one can note that the Manhattan and Euclidean distances appeared to yield the worst results, with a NDCG score equals to 0.80 and lower than 0.84 for all values of $K$. The HSBD distance and its different initializations appeared to yield the best overall results, with a NDCG score higher than 0.84 for all values of $K$. Among the HSBD initializations, its instances initialized with the image-based dissimilarity $\text{HSBD}_{d_\Gamma}$ and the ontological one $\text{HSBD}_{d_\Theta}$ led to the lowest results with NDCG scores equal to 0.89. The combination of the image-based and the ontological dissimilarities $\text{HSBD}_{d^{1/2}_{\Gamma*\Theta}}$ yielded excellent retrieval results with a NDCG score equals to 0.89 and higher than 0.87 for all values of $K$. These retrieval results are higher than the ones obtained by considering either the image-based or the ontological relations. The optimized weighted version $\text{HSBD}_{d^{\rho}_{\Gamma*\Theta}}$ appeared to yield the best overall results with a NDCG score equals to 0.92 and higher than 0.87 for all values of $K$. For $K=5$, $\text{HSBD}_{d^{\rho}_{\Gamma*\Theta}}$ led to a NDCG score equals to 0.87 that can be considered as an excellent NDCG retrieval score. For $K=10$, the NDCG score was 0.92 for the $\text{HSBD}_{d^{\rho}_{\Gamma*\Theta}}$ distance, implying nearly perfect retrieval of 10 images. Results obtained with the EMD distance yielded intermediate overall results, with mean case retrieval accuracy greater than 0.87 and greater than 0.83 for all values of $K$. In terms of dissimilarity measures between the features, a similar behavior to HSBD has been observed.

The paired Wilcoxon (sign-rank) statistical test was used to test the null hypothesis that there is no significant difference in the accuracy scores obtained from highest score distances ($\text{HSBD}_{d^{1/2}_{\Gamma*\Theta}}$ and $\text{HSBD}_{d^{\rho}_{\Gamma*\Theta}}$). As we only compare two paired groups ($\text{HSBD}_{d^{1/2}_{\Gamma*\Theta}}$ and $\text{HSBD}_{d^{\rho}_{\Gamma*\Theta}}$), we do not need to use the Friedman test that is required for the comparison of three or more matched groups. Fig. 10 presents a plot of the Z-score comparing the NDCG values obtained with $\text{HSBD}_{d^{1/2}_{\Gamma*\Theta}}$ and $\text{HSBD}_{d^{\rho}_{\Gamma*\Theta}}$, by varying $K$ (where $K$ is the number of retrieved images). The horizontal line is at $Z=-1.98$, the value for a (two-sided) 0.05 p-value; the Z-score should be below the line for $\text{HSBD}_{d^{\rho}_{\Gamma*\Theta}}$ to be significantly better than $\text{HSBD}_{d^{1/2}_{\Gamma*\Theta}}$. For low values of $K$ ($K\leqslant 4$), $\text{HSBD}_{d^{\rho}_{\Gamma*\Theta}}$ is not better than the $\text{HSBD}_{d^{1/2}_{\Gamma*\Theta}}$ distance. However, for high values of $K$ ($K\in[[5,25]]$), $\text{HSBD}_{d^{\rho}_{\Gamma*\Theta}}$ is significantly stronger than $\text{HSBD}_{d^{1/2}_{\Gamma*\Theta}}$. These statistical results confirm the interest of using the proposed automatical weighting strategy to learn local weights when combining the image-based and the ontological term dissimilarity measures in the $d^{\rho}_{\Gamma*\Theta}$ definition.

The retrieval results obtained in this study were compared to the ones obtained by some of the authors in (Napel et al., 2010), where a comparable retrieval method has been applied on a subset of this database composed of 30 CT images. In this previous study, the images were manually annotated using a vocabulary composed of 72 semantic terms. During the retrieval step, the classical Manhattan distance was used to retrieve similar images without considering the relations among the terms. These experiments led to a NDCG score of 0.94 while the $\text{HSBD}_{d^{\rho}_{\Gamma*\Theta}}$ score is 0.92. Although our results are slightly lower than the ones obtained in this previous study, our framework is fully automated. The authors of (Napel et al., 2010) also evaluated the impact of only considering low-level visual features (*i.e.*, 46 texture features and 2 boundary features whose weights were learned from the data by using a modified version of a machine learning method known as adaptive boosting) extracted from the image content to retrieve similar images in the database. These experiments led to NDCG scores lower than 0.80 showing again the interest of considering semantic terms and their relations for the retrieval of similar radiological images.

*4.4.4.3. Evaluation with the diagnoses.* To evaluate the ability of our system to find similar images, we also tested the recall and fall-out of retrieving images of the same diagnosis in the database of $M=72$ lesions belonging to six types. We performed a leave-one-out test on the retrieval algorithm by querying each lesion against the remaining lesions in the database. We assessed the recall and fall-out for retrieving images having the same diagnosis as the query image. The recall and fall-out of the top $K$ retrieval results, with $K\in[[1,m-1]]$ (where $m$ is the total number of images in the database with the same diagnosis as the query image) were computed. For a fixed value of $K$, the recall was calculated by the number of identical diagnoses in the $K$ retrieval results divided by $m-1$. The fall-out is calculated by the number of non-identical diagnoses in the $K$ retrieval results divided by the total number of non-identical diagnoses in the database. The performance was analyzed using the area under the receiving operator characteristic (AUC) curves. The AUC indicates the potential effectiveness within the framework for retrieving lesions by diagnosis

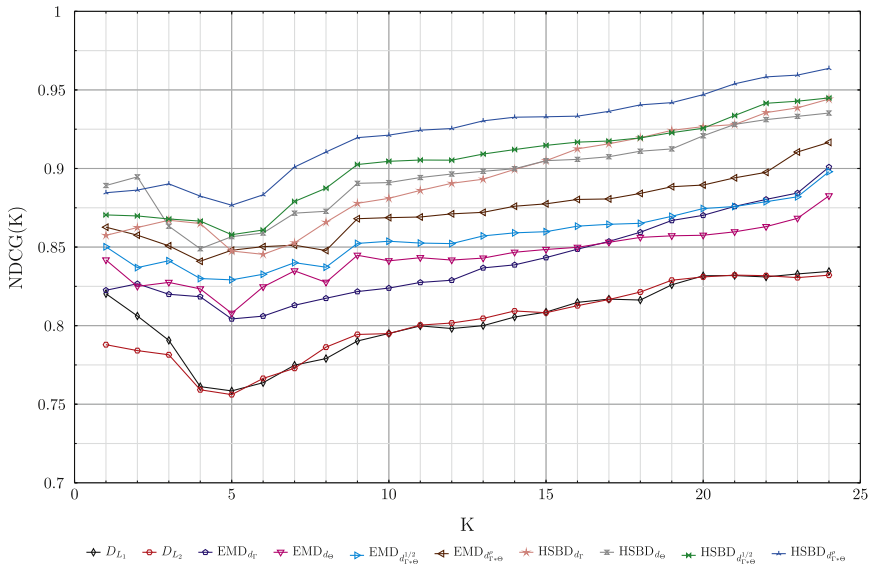



**Fig. 9.** NDCG ($K$) plots, where $K$ is the number of images retrieved in a set of 25 images (cysts, metastases and hemangiomas).

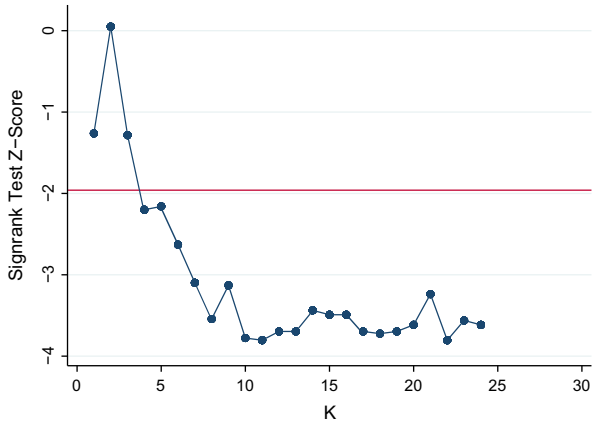



**Fig. 10.** Z-score from the Wilcoxon test comparing the NDCG scores obtained with $\text{HSBD}_{d^{1/2}_{\Gamma*\Theta}}$ and $\text{HSBD}_{d^{\rho}_{\Gamma*\Theta}}$, by varying the number $K$ of images retrieved.

based on their annotations, with the maximum area of 1 being optimal.

*4.4.4.4. Results.* Table 5 presents the different AUC values obtained under these experimental conditions. When all types of lesions were combined, all queries yielded mean AUC values greater than 0.50 for each distance, indicating that the retrieval results favor lesions with similar diagnoses. The Manhattan and Euclidean distances had a similar ranking behavior and appeared to yield the worst results with mean AUC values of 0.50 and 0.52. The HSBD distance and its initializations appeared to yield the best overall results, with mean ROC curve areas always higher than 0.58. Among the HSBD initializations, its instances initialized with the image-based dissimilarity measure $\text{HSBD}_{d_\Gamma}$ and the ontological one $\text{HSBD}_{d_\Theta}$ led to the lowest results with mean case retrieval accuracy equal to 0.58. The combination of the image-based and the ontological dissimilarity measures $\text{HSBD}_{d^{1/2}_{\Gamma*\Theta}}$ led to retrieval results with mean AUC equal to 0.65. The optimized version $\text{HSBD}_{d^{\rho}_{\Gamma*\Theta}}$ appeared to yield the best results with a mean AUC value equals to 0.77. The same behavior has been observed when evaluating independently the AUC for cyst and metastasis retrieval. Results obtained with the EMD distance yielded intermediate results, with mean AUC values always higher than 0.53. In terms of term dissimilarity measures, a similar behavior than for HSBD has been observed: $\text{EMD}_{d^{\rho}_{\Gamma*\Theta}}$ led to the best EMD overall results while $\text{EMD}_{d_\Gamma}$ and $\text{EMD}_{d_\Theta}$ led to the worst EMD results.

The paired Wilcoxon (sign-rank) statistical test was used to test the null hypothesis that there is no significant difference in terms of AUC between $\text{HSBD}_{d^{1/2}_{\Gamma*\Theta}}$ and $\text{HSBD}_{d^{\rho}_{\Gamma*\Theta}}$ for the three most

**Table 5**
Mean AUC values computed from the retrieval of the different types of lesions. The best AUC values for each diagnosis are shown in bold.

| Distance | AUC Cyst (# 21 lesions) | AUC Metastasis (# 24 lesions) | AUC Hemangioma (# 13 lesions) | AUC Hepatocellular carcinoma (# 6 lesions) | AUC Focal nodular hyperplasia (# 5 lesions) | AUC Abscess (# 3 lesions) | AUC All combined (# 72 lesions) |
|---|---|---|---|---|---|---|---|
| $D_{L_1}$ | 0.76 | 0.39 | 0.43 | 0.08 | 0.07 | 0.12 | 0.50 |
| $D_{L_2}$ | 0.78 | 0.40 | 0.44 | 0.09 | 0.13 | **0.26** | 0.52 |
| $\text{EMD}_{d_\Gamma}$ | 0.75 | 0.53 | 0.41 | 0.15 | 0.12 | 0.08 | 0.53 |
| $\text{EMD}_{d_\Theta}$ | 0.84 | 0.47 | 0.56 | 0.15 | **0.28** | 0.00 | 0.55 |
| $\text{EMD}_{d^{1/2}_{\Gamma*\Theta}}$ | 0.87 | 0.51 | 0.51 | 0.18 | 0.17 | 0.02 | 0.56 |
| $\text{EMD}_{d^{\rho}_{\Gamma*\Theta}}$ | 0.89 | 0.55 | 0.54 | 0.21 | 0.15 | 0.05 | 0.61 |
| $\text{HSBD}_{d_\Gamma}$ | 0.85 | 0.49 | 0.47 | 0.22 | 0.20 | 0.05 | 0.58 |
| $\text{HSBD}_{d_\Theta}$ | 0.88 | 0.51 | 0.57 | 0.12 | 0.17 | 0.18 | 0.58 |
| $\text{HSBD}_{d^{1/2}_{\Gamma*\Theta}}$ | 0.92 | 0.56 | 0.59 | 0.23 | 0.17 | 0.23 | 0.65 |
| $\text{HSBD}_{d^{\rho}_{\Gamma*\Theta}}$ | **0.93** | **0.64** | **0.68** | **0.27** | 0.24 | 0.14 | **0.77** |

represented diagnoses. This statistical test led to 3 sub-tests, with a Bonferroni significance level of $0.05/3 = 0.017$. So any comparison with a p-value lower than 0.017 can be considered significantly different. For the retrieval of cysts we obtained $p$-values equal to $p \leqslant 0.0001$ for $\mathrm{HSBD}_{d^{1/2}_{\Gamma*\Theta}}$ *versus* $\mathrm{HSBD}_{d^{\rho}_{\Gamma*\Theta}}$. For the retrieval of metastases we obtained $p$-values equal to $p \leqslant 0.005$ for $\mathrm{HSBD}_{d^{1/2}_{\Gamma*\Theta}}$ *versus* $\mathrm{HSBD}_{d^{\rho}_{\Gamma*\Theta}}$. Finally, for the retrieval of hemangiomas we obtained $p$-values equal to $p \leqslant 0.003$ for $\mathrm{HSBD}_{d^{1/2}_{\Gamma*\Theta}}$ *versus* $\mathrm{HSBD}_{d^{\rho}_{\Gamma*\Theta}}$. From these comparisons we deduce that the AUC scores obtained with the $\mathrm{HSBD}_{d^{\rho}_{\Gamma*\Theta}}$ distance are statistically better than those obtained with the $\mathrm{HSBD}_{d^{1/2}_{\Gamma*\Theta}}$ distance. These AUC comparison results confirm the relevance of our weight learning strategy in the definition of $d^{\rho}_{\Gamma*\Theta}$.

Fig. 11 illustrates retrieval results with the different EMD and HSBD initializations by using a hemangioma query. Perfect retrieval (in terms of the diagnoses) would result in a ranked list of images with only hemangioma lesions. Among the HSBD initializations, its instances initialized with the image-based dissimilarity measure $\mathrm{HSBD}_{d_\Gamma}$ and the ontological one $\mathrm{HSBD}_{d_\Theta}$ led to the lowest retrieval results while its instances initialized with the combination of the image-based and the ontological term dissimilarity measures ($\mathrm{HSBD}_{d^{1/2}_{\Gamma*\Theta}}$ and $\mathrm{HSBD}_{d^{\rho}_{\Gamma*\Theta}}$) led to the highest retrieval results. The same behavior has been observed with EMD.

### 4.5. Discussion

We have developed a complete framework for the retrieval of medical images described with high-level semantic annotations.

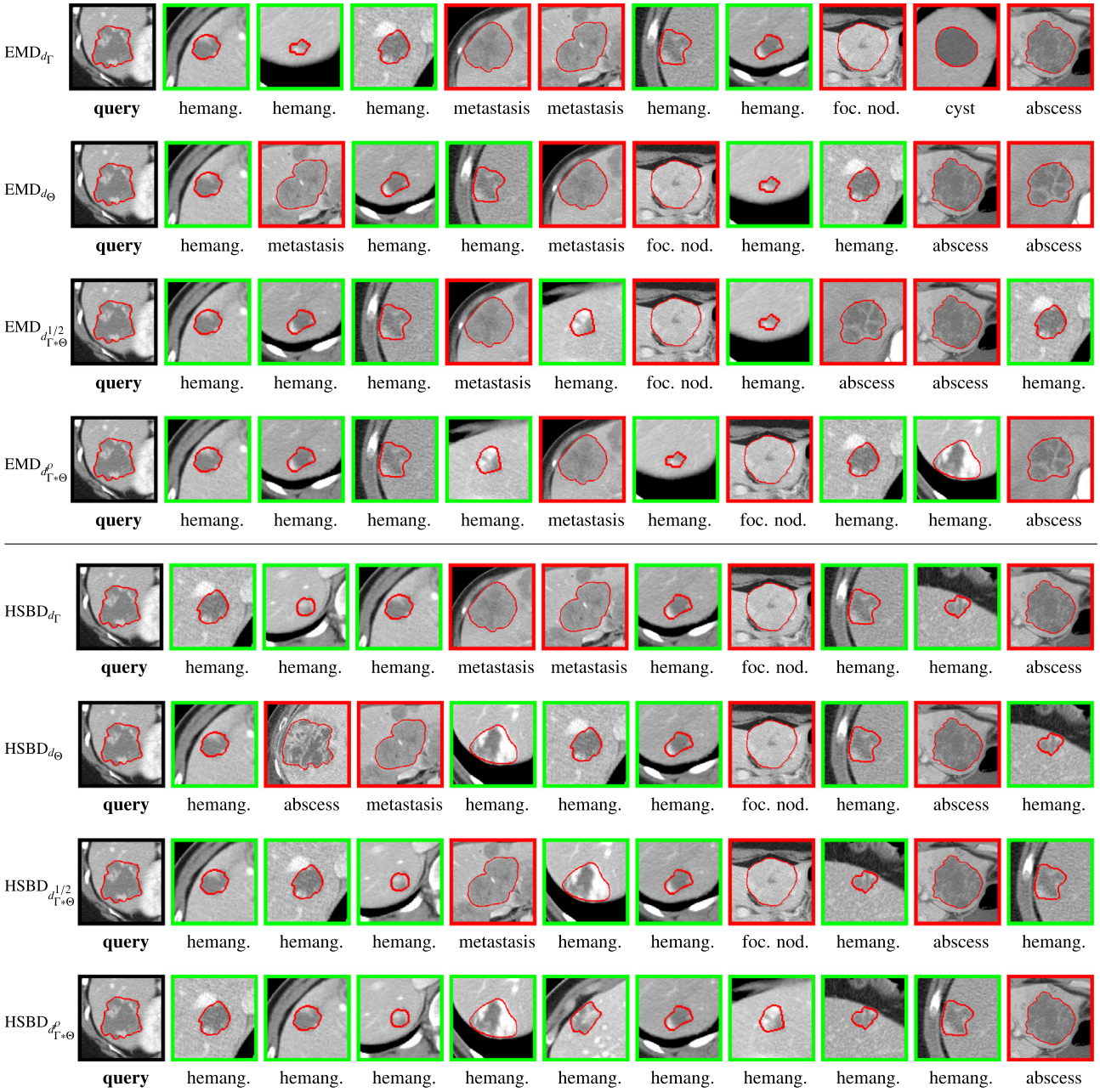



**Fig. 11.** Examples of image retrieval for a hemangioma query for the different initializations of EMD and HSBD. Only the top ten retrieved images are depicted. Dissimilarity rankings go from lowest (left) to highest (right). The query image is surrounded by a black frame. Retrieved images having the same diagnosis as the query image are surrounded by green frames while retrieved images having a different diagnosis are surrounded by red frames. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Given a new query image, this system automatically predicts the semantic terms that are the best suited to visually characterize the image content using the methodology proposed in (Depeursinge et al., 2014b). The prediction of the semantic terms relies on the learning, from a database of previously annotated images, of specific term visual signatures based on Riesz wavelets that model semantic terms through the image features. The retrieval of similar images in the database is then performed using the HSBD distance (Kurtz et al., 2014) that has been extended in this study to consider both the image-based and the ontological relations among the terms describing the images. We also propose an original strategy to determine the weights of the image-based and the ontological relations that are automatically evaluated from data examples using a learning strategy. This retrieval step enables retrieving radiological images described with terms that are not strictly equal but that are semantically correlated and that can be described in the same manner relative to the image content. Use of visual and semantic information associated with images is not new (Ruiz, 2006; Hsu et al., 2009; Xin et al., 2010). However the fusion in image retrieval systems of image-based and ontological dissimilarities in a multi-scale fashion is innovative. To the best of our knowledge, we are among the first ever to investigate this paradigm in medical imaging. In addition, due to the theoretical low computational cost of HSBD, this framework can be used to deal with large datasets.

We studied the impact of considering different kinds of relations among the semantic terms (image-based and/or ontological term relations) for retrieving similar database images. We observe that, in terms of dissimilarity measures between the semantic terms, using only the ontological semantic term relations when comparing pairs of images leads to better image retrieval results than when using the image-based term relations. The best results have been obtained with the optimized combination of the ontological term relations and the image-based term relations. These results confirm our two premises: (1) the incorporation of *a priori* high-level knowledge into the image retrieval methodology (by extracting term relations from the ontological structure) can improve the retrieval results and performed better than image retrieval strategies based purely on low-level imaging features, which are directly extracted from the image content (image-based term relations) and (2) the combination of high-level knowledge extracted from a biomedical ontology and visual information extracted directly from the image content enables taking advantages of the complementarity of both strategies to improve the image retrieval results. This combination reproduces the image interpretation process of the radiologists that consider both visual and semantic similarities for retrieving similar medical cases.

Finally, our validations suggest the potential relevance and usefulness of this system for radiological image retrieval tasks. By combining the HSBD distance with the proposed term dissimilarity measure, we obtained a NDCG score of more than 0.92 on a 25-images dataset while AUC scores of more than 0.77 were obtained on a 72-images dataset. For comparison, the use of EMD, which is an alternative (but with higher-computational cost) distance metric that also considers the relations among the semantic terms, led to NDCG score of 0.87 and AUC results of 0.64. The use of classical distances (*e.g.*, Manhattan or Euclidean) that do not take into account relations among the semantic terms led to lower accuracy results. Our results are competitive with the state-of-the-art methods and highlight the interest of using the proposed semantic image retrieval framework. We provide potential methodological explanations to these results: the HSBD distance performs better than classical distances to retrieve similar images because of its ability to retrieve vectors of terms that are not strictly equal but are visually and semantically similar (*e.g.*, round and ovoid). In spite of the ability of the EMD distance to deal with outliers by matching terms that are not visually and semantically similar (*e.g.*, internal nodules and normal perilesional tissue) but that are occasionally used to describe images belonging to the same category, the HSBD distance performs better because of its hierarchical property to match terms by considering their inherent multilevel correlations.

Our work has some limitations. The dataset and the reference standard used in the first image ranking task were small and contained only 25 CT images with three diagnoses. We were unable to develop a larger dataset since it is time consuming for radiologists to assess and annotate images with semantic terms. In addition, the reference standard only models a consensus opinion based on a dissimilarity measure for the image pairs provided by two radiologists. This expert dissimilarity rating may not be sufficient to accurately represent the inter-reader agreement between radiologists (Faruque et al., 2013). In the future, we plan to build a larger reference standard by increasing the number of cases and the number of diagnoses, modeling the opinions of a higher number of radiologists avoiding potential experimental repeatability issues. Our work is also limited because the clinical protocol used to acquire our data was old (5 mm thickness slices); thinner slice thickness is now common in most CT scan acquisitions. We believe our methods are not likely to be affected by such differences and slice thickness, and we plan to use thinner CT slices for the development of a future dataset. Another limitation is that we do not evaluate the impact of the selection of the slice (used as query image for retrieval purpose) in the 3D CT volume, and the impact of the quality of the ROI delineation. We plan to address in a future study whose goal will be to evaluate the robustness of this image retrieval approach. Another limitation is that the current implementation of our framework only focuses on a subset of semantic annotations (Table 2) that can be predicted from texture features (*i.e.*, Riesz wavelets). These semantic terms cannot describe the global lesion content (*e.g.*, lesion shape, lesion focality) since they are more adapted to describe its inner texture. We plan to integrate supplementary quantitative imaging feature to automatically predict complementary semantic terms to enhance the lesion high-level annotations. An additional limitation is that the dissimilarity between the semantic terms partially relies on the quality of the ontology itself and the texture features used to characterize the terms. We plan to enhance the current framework by considering simultaneously different ontologies besides RadLex and complementary quantitative imaging descriptors. A final perspective is that the presented image-based term dissimilarity measure can be used to learn and/or update existing ontological structures. This could be studied in future work. Despite the presented limitations, we believe our results show the potential value of our methods for improving CBIR in the medical imaging context.

## 5. Conclusion

We present a new semantic framework that enables retrieving similar images based on high-level semantic image annotations. These annotations consist of semantic terms belonging to an ontology, automatically predicted from the image content, which ensure the performance reproducibility with radiologists having various levels of experience. In addition, such annotations could provide a robust alternative to quantitative features extracted from the image content that are often not sufficient to characterize complex lesions in an accurate and comprehensive fashion. Thanks to a hierarchical low-computational cost distance, this framework incorporates the high-level similarities among the semantic terms used to describe the images when retrieving similar images in a database. A unique aspect of our approach is the consideration of both image-based and semantic similarities between ontological terms that describe the image contents. These term similarities

are combined in a global term dissimilarity measure and their respective contributions are learned from data examples using a machine learning strategy based on the maximization of the agreement between the perceptual image dissimilarity and the global term dissimilarity value. To validate this framework, we applied it to the retrieval of medical images of the liver. The results obtained show reasonable accuracy scores compared with an independently constructed pairwise visual dissimilarity standard of liver lesions visible on portal venous CT images. The semantic framework we have developed is generalizable and can be easily adapted to other anatomic and diagnosis scenarios in which CT or other imaging modalities are used. Ultimately, this framework opens new opportunities for the development of computer-assisted image retrieval applications and softwares that can be used by Radiologists in combination with PACS systems.

## Acknowledgements

We thank Jarrett Rosenberg for his useful help on statistical evaluations. This project was funded in part by a Grant from National Cancer Institute, National Institutes of Health (# U01CA142555-01, # R01 CA160251), the Swiss National Science Foundation (# PBGEP2_142283), and by a Grant from GE Medical Systems.

## References

Aigrain, P., Zhang, H., Petkovic, D., 1996. Content-based representation and retrieval of visual media: a state-of-the-art review. Multimedia Tools Appl. 3, 179–202.
Akgül, C.B., Rubin, D.L., Napel, S., Beaulieu, C.F., Greenspan, H., Acar, B., 2011. Content-based image retrieval in radiology: current status and future directions. J. Digital Imag. 24, 208–222.
Al-Mubaid, H., Nguyen, H.A., 2006. A cluster-based approach for semantic similarity in the biomedical domain. In: Proceedings of the IEEE Symposium of the Engineering in Medicine and Biology Society, pp. 2713–2717.
Allampalli-Nagaraj, G., Bichindaritz, I., 2009. Automatic semantic indexing of medical images using a web ontology language for case-based image retrieval. Eng. Appl. Artif. Intell. 22, 18–25.
Andre, B., Vercauteren, T., Buchner, A.M., Wallace, M.B., Ayache, N., 2012. Learning semantic and visual similarity for endomicroscopy video retrieval. IEEE Trans. Med. Imag. 31, 1276–1288.
Batet, M., Sánchez, D., Valls, A., 2011. An ontology-based measure to compute semantic similarity in biomedicine. J. Biomed. Inf. 44, 118–125.
Bodenreider, O., 2004. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic Acids Res. 32, 267–270.
Cha, S.H., Srihari, S.N., 2002. On measuring the distance between histograms. Pattern Recogn. 35, 1355–1370.
Demner-Fushman, D., Antani, S., Simpson, M., Thoma, G.R., 2009. Annotation and retrieval of clinically relevant images. Int. J. Med. Inf. 78, 59–67.
Deng, J., Dong, W., Socher, R., Li, L., Li, K., Fei-fei, L., 2009. ImageNet: a large-scale hierarchical image database. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, pp. 248–255.
Depeursinge, A., Foncubierta-Rodriguez, A., Ville, D., Müller, H., 2012. Multiscale lung texture signature learning using the Riesz transform. In: Ayache, N., Delingette, H., Golland, P., Mori, K. (Eds.), Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, LNCS, vol. 7512. Springer, pp. 517–524.
Depeursinge, A., Foncubierta-Rodriguez, A., Van de Ville, D., Müller, H., 2014a. Rotation-covariant texture learning using steerable Riesz wavelets. IEEE Trans. Image Process. 23, 898–908.
Depeursinge, A., Kurtz, C., Beaulieu, C.F., Napel, S., Rubin, D.L., 2014b. Predicting visual semantic descriptive terms from radiological image data: preliminary results with liver lesions in CT. IEEE Trans. Med. Imag. J.. http://dx.doi.org/10.1109/TMI.2014.2321347 (in press).
Deselaers, T., Ferrari, V., 2011. Visual and semantic similarity in ImageNet. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1777–1784.
Faruque, J., Rubin, D.L., Beaulieu, C.F., Napel, S., 2013. Modeling perceptual similarity measures in CT images of focal liver lesions. J. Digital Imag. 26, 714–720.
Gimenez, F., Jiajing, X., Yi, L., Liu, T.T., Beaulieu, C.F., Rubin, D.L., Napel, S., 2011. On the feasibility of predicting radiological observations from computational imaging features of liver lesions in CT scans. In: Proceedings of IEEE International Conference on Healthcare Informatics, Imaging and Systems Biology, pp. 346–350.
Gondra, I., Heisterkamp, D.R., 2008. Content-based image retrieval with the normalized information distance. Comput. Vis. Image Understan. 111, 219–228.
Guarino, N., 1995. Formal ontology, conceptual analysis and knowledge representation. Int. J. Hum.–Comput. Stud. 43, 625–640.
Hsu, W., Antani, S., Long, L.R., Neve, L., Thoma, G.R., 2009. SPIRS: a Web-based image retrieval system for large biomedical databases. Int. J. Med. Informat. 78, 13–24.
Järvelin, K., Kekäläinen, J., 2002. Cumulated gain-based evaluation of IR techniques. ACM Trans. Inf. Syst. 20, 422–446.
Jiang, Y.G., Ngo, C.W., 2009. Visual word proximity and linguistics for semantic video indexing and near-duplicate retrieval. Comput. Vis. Image Understan. 113, 405–414.
Kesorn, K., Chimlek, S., Poslad, S., Piamsa-nga, P., 2011. Visual content representation using semantically similar visual words. Expert Syst. Appl. 38, 11472–11481.
Korenblum, D., Rubin, D.L., Napel, S., Rodriguez, C., Beaulieu, C.F., 2011. Managing biomedical image metadata for search and retrieval of similar images. J. Digital Imag. 24, 739–748.
Kurtz, C., Beaulieu, C.F., Napel, S., Rubin, D.L., 2014. A hierarchical knowledge-based approach for retrieving similar medical images described with semantic annotations. J. Biomed. Informat. 49, 227–244.
Kurtz, C., Gançarski, P., Passat, N., Puissant, A., 2013. A hierarchical semantic-based distance for nominal histogram comparison. Data Knowl. Eng. 87, 206–225.
Langlotz, C.P., 2006. RadLex: a new method for indexing online educational materials. Radiographics 26, 1595–1597.
Lee, W.N., Shah, N., Sundlass, K., Musen, M., 2008. Comparison of ontology-based semantic-similarity measures. In: Proceedings of the American Medical Informatics Association Annual Symposium, pp. 384–390.
Liu, J., Yang, Y., Saleemi, I., Shah, M., 2012. Learning semantic features for action recognition via diffusion maps. Comput. Vis. Image Understan. 116, 361–377.
Liu, S., Cai, W., Song, Y., Pujol, S., Kikinis, R., Feng, D., 2013. A bag of semantic words model for medical content-based retrieval. In: Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention – Workshop on Medical Content-Based Retrieval for Clinical Decision Support, pp. 125–131.
Liu, Y., Zhang, D., Lu, G., Ma, W.Y., 2007. A survey of content-based image retrieval with high-level semantics. Pattern Recogn. 40, 262–282.
López-Monroy, A.P., Montes-y Gómez, M., Escalante, H.J., Cruz-Roa, A., González, F.A., 2013. Bag-of-visual-ngrams for histopathology image classification. In: Proceedings of the SPIE International Seminar on Medical Information Processing and Analysis, pp. 1–12.
Lowe, H.J., Barnett, G.O., 1994. Understanding and using the medical subject headings (mesh) vocabulary to perform literature searches. J. Am. Med. Assoc. 271, 1103–1108.
Ma, H., Zhu, J., Lyu, M.R.T., King, I., 2010. Bridging the semantic gap between images and tags. IEEE Trans. Multimedia 12, 462–473.
Mojsilovic, A., Rogowitz, B., 2001. Capturing image semantics with low-level descriptors. In: Proceedings of the IEEE International Conference on Image Processing, pp. 18–21.
Napel, S.A., Beaulieu, C.F., Rodriguez, C., Cui, J., Xu, J., Gupta, A., Korenblum, D., Greenspan, H., Ma, Y., Rubin, D.L., 2010. Automated retrieval of ct images of liver lesions on the basis of image similarity: method and preliminary results. Radiology 256, 243–252.
Niblack, C.W., Barber, R., Equitz, W., Flickner, M.D., Glasman, E.H., Petkovic, D., Yanker, P., Faloutsos, C., Taubin, G., 1993. QBIC project: querying images by content, using color, texture and shape. In: Proceedings of the SPIE Conference on Storage and Retrieval for Image and Video Databases, pp. 173–187.
Pedrosa, G.V., Traina, A.J.M., 2013. From bag-of-visual-words to bag-of-visual-phrases using n-grams. In: Proceedings of the International Conference on Graphics, Patterns and Images, pp. 304–311.
Rasiwasia, N., Moreno, P.J., Vasconcelos, N., 2007. Bridging the gap: query by semantic example. IEEE Trans. Multimedia 9, 923–938.
Robinson, P.J., 1997. Radiology's Achilles' heel: error and variation in the interpretation of the röntgen image. British J. Radiol. 70, 1085–1098.
Rubin, D.L., 2012. Finding the meaning in images: annotation and image markup. Philos. Psych. Psychol. 18, 311–318.
Rubin, D.L., Rodriguez, C., Shah, P., Beaulieu, C., 2008. iPad: semantic annotation and markup of radiological images. In: Proceedings of the Symposium of the American Medical Informatics Association, pp. 626–635.
Rubin, G.D., 2000. Data explosion: the challenge of multidetector-row CT. Eur. J. Radiol. 36, 74–80.
Rubner, Y., Tomasi, C., Guibas, L.J., 2000. The Earth Mover's Distance as a metric for image retrieval. Int. J. Comput. Vis. 40, 99–121.
Ruiz, M.E., 2006. Combining image features, case descriptions and UMLS concepts to improve retrieval of medical images. In: Proceedings of the American Medical Informatics Association annual symposium, pp. 674–678.
Stearns, M.Q., Price, C., Spackman, K.A., Wang, A.Y., 2001. SNOMED clinical terms: overview of the development process and project status. In: Proceedings of the American Medical Informatics Association annual symposium, pp. 662–668.
Tibshirani, R., 1996. Regression shrinkage and selection via the LASSO. J. Roy. Stat. Soc. B, 267–288.
Van Gemert, J.C., Veenman, C.J., Smeulders, A.W.M., Geusebroek, J.M., 2010. Visual word ambiguity. IEEE Trans. Pattern Anal. Mach. Intell. 32, 1271–1283.
Voorhees, E.M., 1999. Natural language processing and information retrieval. In: Pazienza, M.T. (Ed.), Information Extraction, LNCS, vol. 1714. Springer, pp. 32–48.
Ward, J.H., 1963. Hierarchical grouping to optimize an objective function. J. Am. Stat. Assoc. 58, 236–244.

Wu, Z., Palmer, M., 1994. Verbs semantics and lexical selection. In: Proceedings of the Association for Computational Linguistics Symposium, pp. 133–138.
Xin, Z., Depeursinge, A., Müller, H., 2010. Information fusion for combining visual and textual image retrieval. In: Proceedings of the International Conference on Pattern Recognition, pp. 1590–1593.
Yang, W., Lu, Z., Yu, M., Huang, M., Feng, Q., Chen, W., 2012. Content-based retrieval of focal liver lesions using Bag-of-Visual-Words representations of single- and multiphase contrast-enhanced CT images. J. Digital Imag. 25, 708–719.
Zhang, D., Islam, M.M., Lu, G., 2012. A review on automatic image annotation techniques. Pattern Recogn. 45, 346–362.