A Structured Learning Approach for Medical Image Indexing and Retrieval

Joo-Hwee Lim[1] and Jean-Pierre Chevallet[2]
[1]Institute for Infocomm Research and [2]IPAL-CNRS
21 Heng Mui Keng Terrace, Singapore 119613
joohwee@i2r.a-star.edu.sg, Jean-Pierre.Chevallet@imag.fr

Abstract

Medical images are critical assets for medical diagnosis, research, and teaching. To facilitate automatic indexing and retrieval of large medical image databases, we propose a structured framework for designing and learning vocabularies of meaningful medical terms with associated visual appearance from image samples. These VisMed terms span a new feature space to represent medical image contents. After a multi-scale detection process, a medical image is indexed as compact spatial distributions of VisMed terms.

When queries are in the form of example images, both a query image and a database image can be matched based on their distributions of VisMed terms, much like the matching of feature-based histograms though the bins refer to semantic medical terms. In addition, a flexible tiling (FlexiTile) matching scheme has been proposed to compare the similarity between two medical images of arbitrary aspect ratios. This matching scheme supports similarity-based retrieval with visual queries. The ranked list of such retrieval is denoted as “i2r-vk-sim.txt” in our submission to ImageCLEF 2005.

When a query is expressed as a text description that involves modality, anatomy, and pathology etc, it can be translated into a visual query representation that chains the presences of VisMed terms with spatial significance via logical operators (AND, OR, NOT) and spatial quantifiers for automatic query processing based on the VisMed image indexes. This query formulation and processing scheme allows semantics-based retrieval with text queries. The ranked list of such retrieval is denoted as “i2r-vk-sem.txt” in our submission to ImageCLEF 2005.

By fusing the ranked lists from both the similarity-based and semantics-based retrievals, we can leverage on the information expressed in both visual and text queries respectively. The ranked list of such retrieval is denoted as “i2r-vk-avg.txt” in our submission to ImageCLEF 2005.

We apply the VisMed approach on the Medical Image Retrieval task of the ImageCLEF track under CLEF 2005. Based on 0.3% (i.e. 158 images) of the $50,026$ images from 4 collections plus 96 images obtained from the web, we cropped 1460 image regions to train and validate 39 VisMed terms using support vector machines. The Mean Average Precisions (MAP) over 25 query topics for the submissions “i2r-vk-sim.txt”, “i2r-vk-sem.txt”, and “i2r-vk-avg.txt” are 0.0721, 0.06, and 0.0921 respectively, according to the evaluation results released by the ImageCLEF 2005 organizers. The submission “i2r-vk-avg.txt” is also combined with text-only submissions “IPALI2R_Tn” and “IPALI2R_T” to form submissions for mixed retrieval. The best MAP among these submissions for mixed retrieval is 0.2821 from submission “IPALI2R_TIan”.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries; H.2 [**Database Managment**]: H.2.3 Languages—*Query Languages*; I.4 [**Image Processing and Computer Vision**]: I.4.10 Image Representation—*Statistical*; J.3 [**Life and Medical Sciences**]: *Medical Information Systems*

General Terms

Algorithms, Design, Experimentation, Languages, Performance

Keywords

Medical Images, Visual Ontology, Similarity-Based Retrieval, Semantics-Based Retrieval

1 Introduction

Medical images are an integral part in medical diagnosis, research, and teaching. Medical image analysis research has focused on image registration, measurement, and visualization. Although large amounts of medical images are produced in hospitals every day, there is relatively less research in medical content-based image retrieval (CBIR) [16]. Besides being valuable for medical research and training, medical CBIR systems also have a role to play in clinical diagnosis [13]. For instance, for less experienced radiologists, a common practice is to use a reference text to find images that are similar to the query image [3]. Hence, medical CBIR systems can assist doctors in diagnosis by retrieving images with known pathologies that are similar to a patient's image(s).

Among the limited research efforts of medical CBIR, classification or clustering driven feature selection and weighting has received much attention as general visual cues often fail to be discriminative enough to deal with more subtle, domain-specific differences and more objective ground truth in the form of disease categories is usually available [3, 11].

In reality, pathology bearing regions tend to be highly localized [3]. Hence, local features such as those extracted from segmented dominant image regions approximated by best fitting ellipses have been proposed [6]. A hierarchical graph-based representation and matching scheme has been suggested to deal with multi-scale image decomposition and their spatial relationships [6]. However, it has been recognized that pathology bearing regions cannot be segmented out automatically for many medical domains [16]. As an alternative, a comprehensive set of 15 perceptual categories related to pathology bearing regions and their discriminative features are carefully designed and tuned for high-resolution CT lung images to achieve superior precision rates over a brute-force feature selection approach [16].

Hence, it is desirable to have a medical CBIR system that represents images in terms of semantic local features, that can be learned from examples (rather than handcrafted with a lot of expert input) and do not rely on robust region segmentation. In order to manage large and complex set of visual entities (i.e. high content diversity) in the medical domain, we propose a structured learning framework to facilitate modular design and extraction of medical visual semantics, *VisMed* terms, in building content-based medical image retrieval systems (Section 2). VisMed terms are image regions that exhibit semantic meanings to medical practitioners and that can be learned statistically to span a new indexing space (Section 2.1). During image indexing, they are detected in image content, reconciled across multiple resolutions, and aggregated spatially to form local semantic histograms (Section 2.2).

The resulting compact and abstract VisMed image indexes can support both similarity-based query and semantics-based query efficiently, as we will describe how they are applied to ImageCLEF 2005 datasets in Section 3. When queries are in the form of example images, both a query image and a database image can be matched based on their distributions of VisMed terms, much like the matching of feature-based histograms though the bins refer to semantic medical terms. In addition, a flexible tiling (FlexiTile) matching scheme has been proposed to compare the similarity between two medical images of arbitrary aspect ratios (Section 3.1).

When a query is expressed as a text description that involves modality, anatomy, and pathology etc, they can be translated into a visual query representation that chains the presences of VisMed terms with spatial significance via logical operators (AND, OR, NOT) and spatial quantifiers for automatic query processing based on the VisMed image indexes. This query formulation and processing scheme allows semantics-based retrieval with text queries (Section 3.2). By fusing the ranked lists from both the similarity-based and semantics-based retrievals, we can leverage on the information expressed in both visual and text queries respectively (Section 3.3). The relevant ImageCLEF 2005 evaluation results will be discussed (Section 3.4) before conclusion.

2 Learning VisMed Terms for Image Indexing

2.1 Learning of VisMed Terms

VisMed terms are typical semantic tokens with visual appearance in medical images (e.g. Xray-bone-fracture, CT-abdomen-liver, MRI-head-brain, photo-skin). They are defined using image region instances cropped from sample images and modeled and built based on statistical learning. In this paper, we have adopted color and texture features as well as support vector machines (SVMs) [18] for VisMed term representation and learning respectively though the framework is not dependent on a particular feature and classifier. The notion of using a visual vocabulary to represent and index image contents for more effective (i.e. semantic) query and retrieval has been proposed and applied to consumer images [7, 10].

To compute VisMed terms from training instances, we use SVMs on color and texture features for an image region and denote this feature vector as z. A SVM $\mathcal{S}_k$ is a detector for VisMed term k on z. The classification vector T for region z is computed via the softmax function [1] as

$$T_k(z) = \frac{\exp^{\mathcal{S}_k(z)}}{\sum_j \exp^{\mathcal{S}_j(z)}}. \tag{1}$$

i.e. $T_k(z)$ corresponds to a VisMed entry in the 39-dimensional vector T adopted in this paper.

In our experiments, we use the YIQ color space over other color spaces (e.g. RGB, HSV, LUV) as it performed better in our experiments. For the texture feature, we adopted the Gabor coefficients which have been shown to provide excellent pattern retrieval results [12].

A feature vector z has two parts, namely, a color feature vector z^c and a texture feature vector z^t. We compute the mean and standard deviation of each YIQ color channel and the Gabor coefficients (5 scales, 6 orientations) respectively [10]. Hence the color feature vector z^c has 6 dimensions and the texture feature vector z^t has 60 dimensions. Zero-mean normalization [15] was applied to both the color and texture features. In our evaluation described below, we adopted RBF kernels with modified city-block distance between feature vectors y and z,

$$|y - z| = \frac{1}{2}(\frac{|y^c - z^c|}{N_c} + \frac{|y^t - z^t|}{N_t}) \tag{2}$$

where N_c and N_t are the numbers of dimensions of the color and texture feature vectors (i.e. 6 and 60) respectively. This just-in-time feature fusion within the kernel combines the contribution of color and texture features equally. It is simpler and more effective than other feature fusion methods that we have attempted.

2.2 Image Indexing based on VisMed Terms

After learning, the VisMed terms are detected during image indexing from multi-scale block-based image patches without region segmentation to form semantic local histograms as described below.

Conceptually, the indexing is realized in a three-layer visual information processing architecture (Figure 1). The bottom layer denotes the pixel-feature maps computed for feature extraction. In our experiments, there are 3 color maps (i.e. YIQ channels) and 30 texture maps (i.e. Gabor coefficients of 5 scales and 6 orientations). From these maps, feature vectors z^c and z^t compatible with those adopted for VisMed term learning (Equation (2)) are extracted.

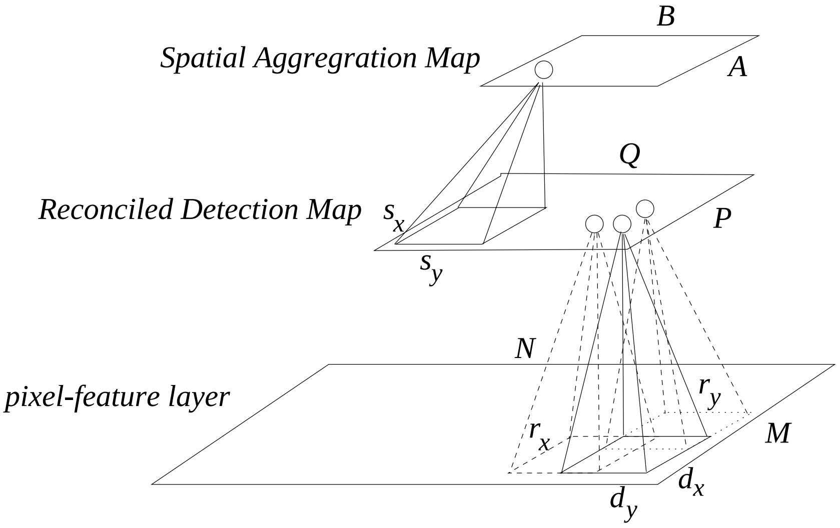

Figure 1: A 3-layer architecture for image indexing

To detect VisMed terms with translation and scale invariance in an image to be indexed, the image is scanned with windows of different scales, similar to the strategy in view-based object detection [17, 14]. More precisely, given an image I with resolution $M \times N$, the middle layer, Reconciled Detection Map (RDM), has a lower resolution of $P \times Q, P \leq M, Q \leq N$. Each pixel (p, q) in RDM corresponds to a two-dimensional region of size $r_x \times r_y$ in I. We further allow tessellation displacements $d_x, d_y > 0$ in X, Y directions respectively such that adjacent pixels in RDM along X direction (along Y direction) have receptive fields in I which are displaced by d_x pixels along X direction (d_y pixels along Y direction) in I. At the end of scanning an image, each pixel (p, q) that covers a region z in the pixel-feature layer will consolidate the classification vector $T_k(z)$ (Equation (1)).

In our experiments, we progressively increase the window size $r_x \times r_y$ from 20×20 to 60×60 at a displacement (d_x, d_y) of $(10, 10)$ pixels, on an image whose longer side is fixed at 360 pixels after a size normalization step that preserves the aspect ratio. After the detection step, we have 5 maps of detection of dimensions 23×35 to 19×31, which are reconciled into a common RDM as explained below.

To reconcile the detection maps across different resolutions onto a common basis, we adopt the following principle: If the most confident classification of a region at resolution r is less than that of a larger region (at resolution $r + 1$) that subsumes the region, then the classification output of the region should be replaced by those of the larger region at resolution $r + 1$. For instance, if the detection of a face is more confident than that of a building at the nose region (assuming that both face and building (but not nose) are in the visual vocabulary designed for a particular application), then the entire region covered by the face, which subsumes the nose region, should be labeled as face.

Using this principle, we compare detection maps of two consecutive resolutions at a time, in descending window sizes (i.e. from windows of 60×60 and 50×50 to windows of 30×30 and 20×20). After 4 cycles of reconciliation, the detection map that is based on the smallest scan window (20×20) would have consolidated the detection decisions obtained at other resolutions for further spatial aggregation.

The purpose of spatial aggregation is to summarize the reconciled detection outcome in a larger spatial region. Suppose a region Z comprises of n small equal regions with feature vectors $z_1, z_2, \cdots, z_n$ respectively. To account for the size of detected VisMed terms in the spatial area Z, the classification vectors of the reconciled detection map are aggregated as

$$T_k(Z) = \frac{1}{n} \sum_i T_k(z_i). \tag{3}$$

This is the top layer in our three-layer visual information processing architecture where a Spatial Aggregation Map (SAM) further tessellates over RDM with $A \times B, A \leq P, B \leq Q$ pixels. This form of spatial aggregation does not encode spatial relation explicity. But the design flexibility of s_x, s_y in SAM on RDM (the equivalent of r_x, r_y in RDM on I) allows us to specify the location and extent in the content to be focused and indexed. We can choose to ignore unimportant areas (e.g. margins) and emphasize certain areas with overlapping tessellation. We can even have different weights attached to the areas during similarity matching.

To facilitate spatial aggregation and matching of image with different aspect ratios ρ, we design 5 tiling templates for Eq. (3), namely 3×1, 3×2, 3×3, 2×3, and 1×3 grids resulting in 3, 6, 9, 6, and 3 $T_k(Z)$ vectors per image respectively. Since the tiling templates have aspect ratios of 3, 1.5, and 1, the decision thresholds to assign a template for an image are set to their mid-points (2.25 and 1.25) as $\rho > 2.25$, $1.25 < \rho \leq 2.25$, and $\rho \leq 1.25$ respectively based on $\rho = \frac{L}{S}$ where L and S refer to the longer and shorter sides of an image respectively. For more details on detection-based indexing, readers are referred to [10].

3 Medical Image Retrieval using VisMed Terms

As part of the Cross Language Evaluation Forum (CLEF), the ImageCLEF 2005 track [2] that promotes cross language image retrieval has a Medical Image Retrieval (MedIR) task in 2005, organized by Henning Mueller and William Hersh. The test collection contains images from the Casimage, MIR, PEIR, and PathoPIC datasets with a total of $50,026$ images. The collection contains annotations in XML format. The majority of the annotations are in English but a significant number is also in French and German, with a few cases that do not contain any annotation at all. The 25 queries for the MedIR task have been formulated with example images and short textual descriptions. The organizers evaluate retrieval performance in terms of uninterpolated Mean Average Precision (MAP) computed across all topics using `trec_eval`.

We have applied the VisMed approach on the MedIR task. We set out to designed VisMed terms that correspond to typical semantic regions in the medical images. However due to time constraints, we only designed 39 VisMed terms relevant to the query topics. Table 1 lists the 39 VisMed terms (00-38) and Figure 2 illustrates one visual example each for the VisMed terms from top-left (00) to bottom-right (38) in row-wise order. The last two VisMed terms in Table 1, "image-region-bright" and "image-region-dark", refer to bright and dark patches in an image respectively. With a uniform VisMed framework, dark background in the scan images (e.g. CT, MRI) and bright (i.e. empty) areas in drawing etc are simply modeled as dummy terms instead of using image preprocessing to detect them separately.

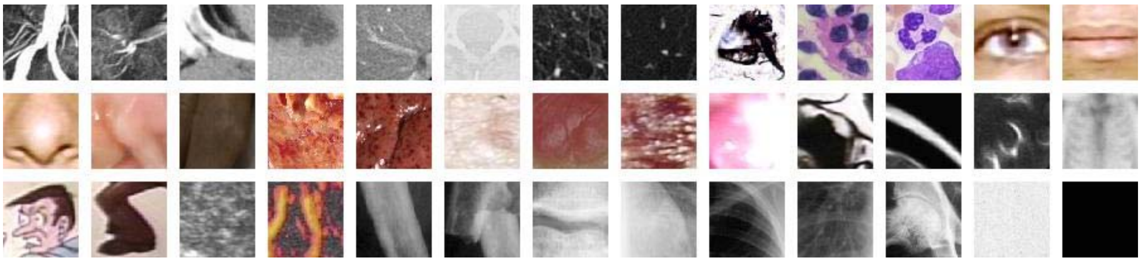

Figure 2: One visual example each for the VisMed terms

Based on 0.3% (i.e. 158 images) of the $50,026$ images from the 4 collections plus 96 images obtained from the web, we cropped 1460 image regions to train and validate 39 VisMed terms

Table 1: VisMed terms and numbers of region samples

VisMed Terms	#	VisMed Terms	#
00-angio-aorta-artery	30	01-angio-aorta-kidney	30
02-ct-abdomen-bone	40	03-ct-abdomen-liver	20
04-ct-abdomen-vessel	30	05-ct-chest-bone	30
06-ct-chest-emphysema	30	07-ct-chest-nodule	20
08-path-alzheimer	40	09-path-kidney	50
10-path-leukemia	30	11-photo-face-eye	60
12-photo-face-mouth	30	13-photo-face-nose	30
14-photo-fetus	50	15-photo-finger-osteo	60
16-photo-heart-attack	40	17-photo-kidney	30
18-photo-skin	60	19-photo-skin-benign	30
20-photo-skin-malignant	30	21-photo-stomach-ulcer	60
22-mri-head-face	50	23-mri-head-bone	40
24-mri-head-brain	50	25-sctg-body	60
26-sketch-human-head	40	27-sketch-human-limb	40
28-ultrasound-grey	30	29-ultrasound-color	20
30-xray-bone	40	31-xray-bone-fracture	60
32-xray-bone-joint	40	33-xray-chest-heart	20
34-xray-chest-lung-clear	30	35-xray-chest-lung-opaque	40
36-xray-pelvis	40	37-image-region-bright	20
38-image-region-dark	20		

using SVMs. As we would like to minimize the number of images selected from the test collection for VisMed term learning, we include relevant images available from the web. For a given VisMed term, the negative samples are the union of the positive samples of all the other 38 VisMed terms. We ensure that they do not contain any of the positive and negative query images given by the 25 query topics.

The odd and even entries of the cropped regions are used as training and validation sets respectively (i.e. 730 each) to optimize the RBF kernel parameter of support vector machines. The best generalization performance with mean error 1.01% on the validation set was obtained with $C = 100, \alpha = 1.0$ [4]. Both the training and validation sets are then combined to form a larger training set to retrain the 39 VisMed detectors. Both query and database images are indexed using the framework as described in the previous section (Eq. (1) to (3)).

3.1 Similarity-Based Retrieval with Visual Query

Given two images represented as different grid patterns, we propose a flexible tiling (FlexiTile) matching scheme to cover all possible matches. For instance, given a query image Q of 3×1 grid and an image Z of 3×3 grid, intuitively Q should be compared to each of the 3 columns in Z and the highest similarity will be treated as the final matching score. As another example, consider matching a 3×2 grid with 2×3 grid. The 4 possible tiling and matching choices are shown in Figure 3.

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

$$\lambda(Q, Z) = \max_{m_1=1,n_1=1}^{m_1=u_1,n_1=v_1} \max_{m_2=1,n_2=1}^{m_2=u_2,n_2=v_2} \lambda(Q_{m_1,n_1}, Z_{m_2,n_2}), \tag{4}$$

where $u_1 = M_1 - M + 1, v_1 = N_1 - N + 1, u_2 = M_2 - M + 1, v_2 = N_2 - N + 1$ and the similarity

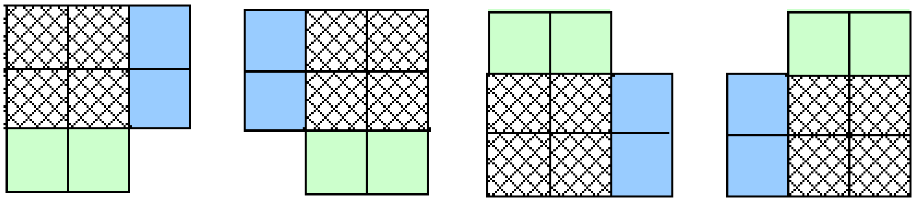

Figure 3: Example to illustrate FlexiTile matching

for each tiling $\lambda(Q_{m_1,n_1}, Z_{m_2,n_2})$ is defined as the average similarity over $M \times N$ blocks as

$$\lambda(Q_{m_1,n_1}, Z_{m_2,n_2}) = \frac{\sum_i \sum_j \lambda_{ij}(Q_{m_1,n_1}, Z_{m_2,n_2})}{M \times N}, \tag{5}$$

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

$$\lambda_{ij}(Q_{m_1,n_1}, Z_{m_2,n_2}) = 1 - \frac{1}{2}\sum_k |T_k(Q_{p_1,q_1}) - T_k(Z_{p_2,q_2})| \tag{6}$$

where $p_1 = m_1 + i, q_1 = n_1 + j, p_2 = m_2 + i, q_2 = n_2 + j$ and it is equivalent to color histogram intersection except that the bins have semantic interpretation as VisMed terms.

However, to avoid spurious matching between very different grids (e.g. 3×1 and 1×3), we set the similarity to zero if the difference in a grid dimension between two image indexes is more than one. That is, two images are considered dissimilar if they exhibit very different aspect ratios.

There is a trade-off between content symmetry and spatial specificity. If we want images of similar semantics with different spatial arrangement (e.g. mirror images) to be treated as similar, we can have larger tessellated block in SAM (i.e. the extreme case is a global histogram). However in applications such as medical images where there is usually very small variance in views and spatial locations are considered differentiating across images, local histograms will provide good sensitivity to spatial specificity. Furthermore, we can attach different weights to the blocks to emphasize the focus of attention (e.g. center) if necessary. In this paper, we report experimental results based on even weights as grid tessellation is used.

Now we extend the similarity matching for multiple query images. Let us denote $\mathcal{Q}^+ = \{Q_1^+, Q_2^+, \cdots, Q_p^+\}$ and $\mathcal{Q}^- = \{Q_1^-, Q_2^-, \cdots, Q_n^-\}$ as the sets of positive and negative query images respectively and $\mathcal{Q} = \mathcal{Q}^+ \cup \mathcal{Q}^-$. We define the similarity between a set of query images and a database image Z as the maximum similarity among similarities between each query image and Z i.e.

$$\lambda(\mathcal{Q}^+, Z) = \max_i \lambda(Q_i^+, Z), \tag{7}$$

$$\lambda(\mathcal{Q}^-, Z) = \max_i \lambda(Q_i^-, Z). \tag{8}$$

If $\mathcal{Q}^- = \emptyset$, then $\lambda(\mathcal{Q}, Z) = \lambda(\mathcal{Q}^+, Z)$. Conversely, if $\mathcal{Q}^+ = \emptyset$, then $\lambda(\mathcal{Q}, Z) = 1 - \lambda(\mathcal{Q}^-, Z)$. If Z is exactly one of the positive query images or negative query images, then $\lambda(\mathcal{Q}, Z)$ should be 1 or 0 respectively i.e. $\lambda(\mathcal{Q}^+, Z) = 1$ or $\lambda(\mathcal{Q}^-, Z) = 1$ respectively. Otherwise,

$$\lambda(\mathcal{Q}, Z) = \frac{1}{2}(\lambda(\mathcal{Q}^+, Z) + (1 - \lambda(\mathcal{Q}^-, Z))) \tag{9}$$

3.2 Semantics-Based Retrieval with Text Query

A new visual query language, Query by Spatial Icons (QBSI), has been proposed to combine pattern matching and logical inference [10]. A QBSI query is composed as a spatial arrangement of visual semantics. A Visual Query Term (VQT) P specifies a region R where a VisMed i should appear and a query formulus chains these terms up via logical operators. The truth value $\mu(P, Z)$ of a VQT P for any image Z is simply defined as

$$\mu(P, Z) = T_i(R) \tag{10}$$

where $T_i(R)$ is defined in Equation (3).

As described in Section 2.2, the medical images are indexed as 3×1, 3×2, 3×3, 2×3, and 1×3 grids, depending on their aspect ratios. When a query involves the presence of a VisMed term in a region larger than a single block in a grid and its semantics prefers a larger area of presence of the VisMed term to have a good match (e.g. entire kidney, skin lesion, chest x-ray images with tuberculosis), Equation (10) will become

$$\mu(P, Z) = \frac{\sum_{Z_j \in R} T_i(Z_j)}{|R|} \tag{11}$$

where Z_j are the blocks in a grid that cover R and $|R|$ denotes the number of such blocks. This corresponds to a spatial universal quantifier ($\forall$).

On the other hand, if a query only requires the presence of a VisMed term within a region regardless of the area of the presence (e.g. presence of a bone fracture, presence of micro nodules), then the semantics is equivalent to the spatial existential quantifier ($\exists$) and Equation (10) will be computed as

$$\mu(P, Z) = \max_{Z_j \in R} T_i(Z_j) \tag{12}$$

A QBSI query $\mathcal{P}$ can be specified as a disjunctive normal form of VQT (with or without negation),

$$\mathcal{P} = (P_{11} \wedge P_{12} \wedge \cdots) \vee \cdots \vee (P_{c1} \wedge P_{c2} \wedge \cdots) \tag{13}$$

Then the query processing of query $\mathcal{P}$ for any image Z is to compute the truth value $\mu(\mathcal{P}, Z)$ using appropriate logical operators. As uncertainty values are involved in VisMed term detection and indexing, we adopt fuzzy operations [5] as follows:

$$\begin{aligned} \mu(\bar{P}, Z) &= 1 - \mu(P, Z), && (14) \\ \mu(P_i \wedge P_j, Z) &= \min(\mu(P_i, Z), \mu(P_j, Z)), && (15) \\ \mu(P_i \vee P_j, Z) &= \max(\mu(P_i, Z), \mu(P_j, Z)). && (16) \end{aligned}$$

For the query processing of the query topics in ImageCLEF 2005, a query text description is manually translated into a QBSI query with the help of a visual query interface [10] which outputs an XML format to state the VisMed terms, the spatial regions, the Boolean operators, and the spatial quantifiers. As an illustration, query 02 "Show me x-ray images with fractures of the femur" is translated as "$\forall$ xray-bone $\in$ whole $\wedge$ $\forall$ xray-pelvis $\in$ upper $\wedge$ $\exists$ xray-bone-fracture $\in$ whole
" where "whole" and "upper" refer to the whole image and upper part of an image respectively.

In fact, the VisMed terms can be further structured into an abstraction hierarchy, namely, IS-A hierarchy and Part-Whole hierarchy, to support more complex queries. Some possible examples of IS-A hierarchies are: a skin lesion can be either benign or malignant; different specific types of bone fracture belong to a common "bone fracture". A Part-Whole hierarchy allows us to detect (and query) a complex object in terms of its constituent parts. This is especially useful when a 3D object has no consistent shape representation in a 2D image. For more details about QBSI, please refer to [10].

3.3 Combining Similarity- and Semantics-Based Retrieval

If a query topic is represented with both query images and text description, we can combine the similarities resulting from query processing using Equations (4) to (9) and (10) to (16) respectively. A simple scheme would be a linear combination of $\lambda(\mathcal{Q}, Z)$ and $\mu(\mathcal{P}, Z)$ with $\omega \in [0, 1]$

$$\rho(\mathcal{Q}, \mathcal{P}, Z) = \omega \cdot \lambda(\mathcal{Q}, Z) + (1 - \omega) \cdot \mu(\mathcal{P}, Z) \tag{17}$$

where ρ is the overall similarity and the optimal ω can be determined empirically using even sampling at 0.1 intervals.

3.4 Evaluation from ImageCLEF 2005 Organizers

According to the ImageCLEF 2005 organizers, the MAP over 25 query topics for the submissions on similarity-based retrieval (Section 3.1, labeled as "i2r-vk-sim.txt"), semantics-based retrieval (Section 3.2, labeled as "i2r-vk-sem.txt"), and their fusion (Section 3.3, denoted as "i2r-vk-avg.txt") are 0.0721, 0.06, and 0.0921 respectively. The submission "i2r-vk-avg.txt" is also combined with text-only submissions "IPALI2R_Tn" and "IPALI2R_T" to form submissions for mixed retrieval. The best MAP among these submissions for mixed retrieval is 0.2821 by submission "IPALI2R_TIan". More details can be found at the website http://trec.ohsu.edu/image/.

The performance of the current VisMed implementation can be further improved. First of all, only two features, one each for color and texture, have been used to train the VisMed term detectors. More domain-specific features can be incorporated to enhance detection accuracies.

Secondly, some VisMed terms have high variations in visual appearances, it may be necessary to divide them into subclasses to ease the learning task. For example, 09-path-kidney may appear in different colors, 21-photo-stomach-ulcer has to cover both endoscopic and pathological images, etc.

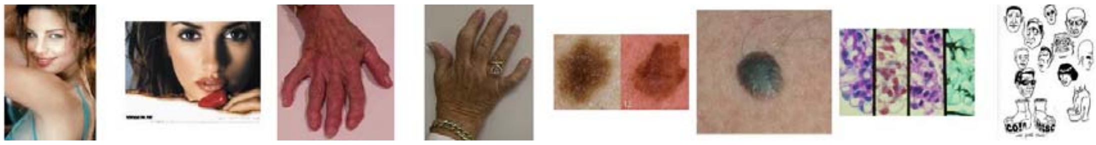

Figure 4: Some training images that unlikely or irrelevant for the ImageCLEF 2005 datasets

Lastly, more relevant training samples for the VisMed terms should be collected based on proper domain understanding to have better detection generalization. As we wanted to minimize the number of images from the test collection used for learning VisMed terms, we tried to look for additional images from the web. However, towards the end of the experiments of the VisMed approach, we realized that the web images, which were supposed to complement the very small training set selected from the test collection, consist of visual samples that are atypical (or even rather different) from those found in the medical test collection (i.e. over-generalization). For instance, as shown in Figure 4, the visual samples used to train VisMed terms related to face (11-13), hand osteoarthritis (15), skin and lesion (18-20), kidney pathologies (17), and sketch (26-27), are not easily found (if not irrelevant) in the given test collection.

4 Conclusion

Medical CBIR is an emerging and challenging research area. We have proposed a structured framework for designing image semantics from statistical learning. Our adaptive framework is scalable to different image domains [10, 8] and embraces other design choices such as better visual features, learning algorithms, object detectors, spatial aggregation and matching schemes when they become available.

We reckon that a limitation of the current VisMed approach is the need to design the VisMed terms manually with labeled image patches as training samples. We have begun some work in a semi-supervised approach to discover meaningful visual vocabularies from minimally labeled image samples [9]. In the near future, we would also explore the integration with inter-class semantics [8]. Last but not least, we would also work with medical experts to design a more comprehensive set of VisMed terms to cover all the essential semantics in medical images.

References

[1] C.M. Bishop. *Neural Networks for Pattern Recognition.* Clarendon Press, Oxford, 1995.

[2] P. Clough and H. Muller. The clef cross language image retrieval track (imageclef) 2005. http://ir.shef.ac.uk/imageclef2005/, 2005.

[3] J.G. Dy, C.E. Brodley, A.C. Kak, L.S. Broderick, and A.M. Aisen. Unsupervised feature selection applied to content-based retrieval of lung images. *IEEE Trans. on PAMI*, 25(3):373–378, 2003.

[4] T. Joachims. Making large-scale svm learning practical. In B. Scholkopf, C.J.C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, pages 169–184. MIT-Press, 1999.

[5] G.J. Klir and T.A. Folger. *Fuzzy Sets, Uncertainty, and Information.* Prentice Hall, 1992.

[6] T.M. Lehmann et al. Content-based image retrieval in medical applications. *Methods Inf Med*, 43:354–361, 2004.

[7] J.H. Lim. Building visual vocabulary for image indexation and query formulation. *Pattern Analysis and Applications*, 4(2):125–139, 2001.

[8] J.H. Lim and J.S. Jin. Combining intra-image and inter-class semantics for consumer image retrieval. *Pattern Recognition*, 38(6):847–864, 2005.

[9] J.H. Lim and J.S. Jin. Discovering recurrent image semantics from class discrimination. *EURASIP Journal of Applied Signal Processing*, 2005. to appear.

[10] J.H. Lim and J.S. Jin. A structured learning framework for content-based image indexing and visual query. *Multimedia Systems Journal*, 10(4):317–331, 2005.

[11] Y. Liu et al. Semantic based biomedical image indexing and retrieval. In L. Shapiro, H.P. Kriegel, and R. Veltkamp, editors, *Trends and Advances in Content-Based Image and Video Retrieval.* Springer, 2004.

[12] B.S. Manjunath and W.Y. Ma. Texture features for browsing and retrieval of image data. *IEEE Trans. on PAMI*, 18(8):837–842, 1996.

[13] H. Muller, N. Michoux, D. Bandon, and A. Geissbuhler. A review of content-based image retrieval systems in medical applications – clinical benefits and future directions. *Intl. J. of Medical Informatics*, 73(1):1–23, 2004.

[14] P.C. Papageorgiou, M. Oren, and T. Poggio. A general framework for object detection. In *Proc. of ICCV*, pages 555–562, 1998.

[15] Y. Rui, T.S. Huang, and S. Mehrotra. Content-based image retrieval with relevance feedback in mars. In *Proc. of IEEE ICIP*, pages 815–818, 1997.

[16] C.R. Shyu, C. Pavlopoulou, A.C. Kak, and C.E. Brodley. Using human perceptual categories for content-based retrieval from a medical image database. *Computer Vision and Image Understanding*, 88:119–151, 2002.

[17] K.K. Sung and T. Poggio. Example-based learning for view-based human face detection. *IEEE Trans. on PAMI*, 20(1):39–51, 1998.

[18] V.N. Vapnik. *Statistical Learning Theory.* Wiley, New York, 1998.