X-ray Categorization and Retrieval on the Organ and Pathology Level, Using Patch-Based Visual Words

Uri Avni, Hayit Greenspan*, Eli Konen, Michal Sharon, and Jacob Goldberger

***Abstract*—In this study we present an efficient image categorization and retrieval system applied to medical image databases, in particular large radiograph archives. The methodology is based on local patch representation of the image content, using a "bag of visual words" approach. We explore the effects of various parameters on system performance, and show best results using dense sampling of simple features with spatial content, and a nonlinear kernel-based support vector machine (SVM) classifier. In a recent international competition the system was ranked first in discriminating orientation and body regions in X-ray images. In addition to organ-level discrimination, we show an application to pathology-level categorization of chest X-ray data, the most popular examination in radiology. The system discriminates between healthy and pathological cases, and is also shown to successfully identify specific pathologies in a set of chest radiographs taken from a routine hospital examination. This is a first step towards similarity-based categorization, which has a major clinical implications for computer-assisted diagnostics.**

***Index Terms*—Chest radiography, computer-aided diagnosis (CAD), disease labeling, image categorization, image patches, image retrieval, visual words, X-ray.**

I. Introduction

In the last ten years there has been an explosion in the number of images that are acquired every day in any modern hospital, due to the increase in digital medical imaging techniques and patient image-screening protocols. The Geneva University Hospital radiology department alone produced 50 000 images per day in 2006 [36]. In the U.K. over 120 National Health Service Trusts have implemented Picture Archiving and Communication Systems (PACS), with over 640 million stored images as of March 2008 [28], and these numbers are rising fast. One outcome of this trend is an enormous increase in the number of images that must be reviewed by radiologists.

With the increase in number of images, there has been a concomitant need for computerized tools to aid radiologists in the diagnostic process. In the radiology workflow, after the image acquisition process, the images are archived in PACS [45]. The radiologist retrieves a case from the PACS system and then makes a diagnosis. Using PACS, the radiologist may want to browse through similar-content images in the archive to ensure an accurate diagnosis. Retrieving similar cases from a large archive is a very challenging task and is one of the key issues in the rapidly expanding domain of content-based medical image retrieval [11].

The field of content-based image retrieval (CBIR) deals with the analysis of image content and the development of tools to represent visual content in a way that can be efficiently searched and compared. Conventional databases allow for textual searches, in particular using the headers of the DICOM standard. Even though some of the important information is contained in the DICOM headers and many imaging devices are DICOM-compliant at this time, it remains suboptimal for a number of reasons: first, in recent studies DICOM headers have been shown to contain a fairly high rate of errors (error rates of up to 16% have been reported [25]). Second, a single image may contain a large number of regions-of-interest, each of which may be the focus of attention for the medical expert, depending on the diagnostic task at hand. A single chest image for example, contains the lungs, heart, rib cage, diaphragm, clavicle, shoulder blade, spine and blood vessels, any of which may be the focus of attention, and all of which should be readily accessible within an ideal retrieval system. Clinical decision support techniques such as case-based reasoning or evidence-based medicine can produce a strong need to retrieve images valuable for supporting certain diagnoses [20], [30]. For the clinical decision-making process it can be beneficial or even crucial to find other images of the same modality, the same anatomic region or the same disease. Computer-aided diagnostics for radiological practice, as presented at the Radiological Society of North America (RSNA) are on the rise and create a need for powerful data and meta-data management and retrieval [16]. Besides diagnostics, teaching and research can be greatly enhanced by visual access methods in existing large repositories.

A. Datasets and Medical Annotation Challenges

While rapidly developing, the field of medical content-retrieval is still in its infancy. Representing a medical image in a semantic space that captures the essence of the image is a key challenge. It involves determining an appropriate image representation and appropriate matching tools suitable for catego-

Manuscript received September 24, 2010; revised November 16, 2010; accepted November 17, 2010. Date of publication November 29, 2010; date of current version March 02, 2011.*Asterisk indicates corresponding author*

U. Avni is with the Department of Biomedical Engineering, Tel-Aviv University, 69978 Tel Aviv, Israel (e-mail: uriavni@gmail.com).

*H. Greenspan is with the Biomedical Engineering Department, Tel-Aviv University, Tel-Aviv 69978, Israel, currently on sabbatical at the Multimedia for Healthcare Group, IBM Almaden Research Center, San Jose, CA 95120 USA (e-mail: hayit@eng.tau.ac.il).

E. Konen and M. Sharon are with the Diagnostic Imaging Department, Sheba Medical Center, 52621 Tel Hashomer, Israel (e-mail: eli.konen@sheba.health.gov.il; michal.sharon@gmail.com)

J. Goldberger is with the School of Engineering, Bar-Ilan University, 52900 Ramat-Gan, Israel (e-mail: goldbej@eng.biu.ac.il).

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

Digital Object Identifier 10.1109/TMI.2010.2095026

rization and retrieval. The representation needs to be general enough to accommodate multiple modalities yet robust enough to handle the large variability of the data.

As research groups are increasingly attracted to medical image retrieval, international competitions are now emerging to assist in the benchmark of feature sets, retrieval and classification schemes. One such annual competition is known as ImageCLEF.[1]

Since 2004 the ImageCLEF competition has conducted text-based as well as image-based retrieval. As of 2005 it also includes a medical image annotation task. Competitions are mainly based on the IRMA project X-ray library [20], which consists of medical radiographs taken from clinical routine at the Department of Diagnostic Radiology, Aachen University Hospital, Germany. Images are classified by medical experts according to the imaging modality, the examined region, the image orientation with respect to the body and the biological system under evaluation.

The ImageCLEF competitions provide an important benchmark opportunity for the analysis of feature spaces (global versus local), similarity measures as well as classification schemes. Todate, the competitions focus on the identification of organs and organ-level characteristics (such as viewpoint). Additional challenges arise with the desire to shift to actual clinical settings and to provide clinical decision support tools. In these more realistic settings, the main challenge is how to shift from organ-level identification to pathology-level analysis.

B. Related Works

Initial systems, still mainly in the research community, have been developed for images of a specific modality or specific organ such as high-resolution CT lung images [13], mammography [29], [2], [15], and the spine [31]. A few systems focus on general medical categorizations (e.g., MedGIFT [34], [9], COBRA [14], and IRMA [20]). Overview papers on the current state of CBIR in medical applications include [35], [20], [1].

In [24] and [23] a continuous probabilistic localized image representation scheme was suggested, with information-theoretic matching tools to match and categorize X-ray images by body regions. The statistical framework suggested, termed Gaussian mixture model Kullback Leibler (GMM-KL), achieved good results on both generic and X-ray archives, due to the information-preserving representation and the strong matching measures. With archives of increasing size, such as the ones present in the ImageCLEF competition, it is necessary to return to more simplistic, discrete representations, and simple matching measures, to preserve computational efficiency. The paradigm of visual words, known as the bag-of-words (BoW) model, which has recently been adapted from the text retrieval domain to the visual analysis domain, provides the efficient means to address the CBIR challenge in large-size archives while maintaining solid classification rates.

The BoW model is commonly used in natural language processing and information retrieval for text documents [33]. In this model a document is statistically modeled as an instance of a multinomial word distribution and is represented as a frequency of occurrence word histogram. The representation as a frequency vector of word occurrences does not take grammar rules or word order into account. It does, however, preserve key information about the content of the document. This representation can be used to compare documents, and to identify document topics. The BoW representation is successfully used in document classification, clustering, and retrieval tasks and is the cornerstone of all Internet search engines.

To represent an image using the BoW model, the image must be treated as a document. Unlike the text world, there is no natural concept for a word or a dictionary. We thus need to find a way to break down the image into a list of visual elements, and a way to discretize the visual element space, since the number of possible visual elements in an image is enormous. In the visual BoW model, the image feature extraction step takes place in a procedure involving detection of points-of-interest, feature description, and codebook generation. The visual word model can thus take the form of a histogram representation of the image, based on a collection of its local features. Each bin in the histogram is a codeword index out of a finite vocabulary of visual codewords, generated in an unsupervised way from the data. Images are compared and classified based on this discrete and compact histogram representation.

In recent years the BoW approach has successfully been applied to general scene and object recognition tasks (see e.g., [43], [22], [37]). In [43] the idea of using the joint distribution of intensity values over compact neighborhoods for the task of texture classification was introduced. In [40] vector quantization of invariant local image descriptors were used to form clusters, referred to as visual "words." They then searched for objects throughout a movie sequence by analogy to text retrieval. Natural scene categories were learned using visual words in [22]. Local words were either grayscale patches or scale-invariant feature transform (SIFT) descriptors [32], sampled on a grid, randomly, or at interest points. They then learned a generative hierarchical model to describe the resulting visual word distribution. "Spatial pyramids" were introduced—a technique of partitioning the image into increasingly fine subregions, and computing histograms of local features within each subregion. They demonstrated significant performance improvement over orderless BoW in global scene classification and object recognition tasks. A large-scale evaluation of the visual words approach for texture classification and object recognition was presented in [46].

Approaches using patch-based, bag-of-visual-words concepts are gradually emerging in medical tasks. In [3] BoW is used as the representation of endomicroscopic images and achieves high accuracy in the tasks of classifying the images into neoplastic (pathological) and benign. In [8] an application to texture representation for mammography tissue classification and segmentation is presented. The use of BoW techniques for large scale radiograph archive categorization can be found in the ImageCLEF competition. In 2006 Deselaers *et al.* [12] displayed the best medical annotation results using the BoW approach, where the features were local patches of different sizes taken at every position and scaled to a common size. In that work, no dictionary was used; rather the feature space was

[1] http://imageclef.org

quantized uniformly in every dimension and the image was represented as a sparse histogram in the quantized space. The system described in [41] had the highest score in 2007 and 2008. In that work both global and local features were used. The global features were downscaled versions of the images (32×32). The local features were modified SIFT descriptors (128 values), sampled randomly. The set of local features was represented as a histogram over a dictionary, built using the K-means algorithm ($K = 500$). Four image quadrants were learned and represented separately. The final representation for a given image was thus the (32×32) pixel values of the global image along with 4 times the (500) histogram bins. Classification was done with SVM using different integration techniques for global and local features.

To date, ImageCLEF competitions are continuing with radiograph archives of increasing size, as well as new archives and tasks. The variability between the competing groups is mainly in the image representation space—using global versus local representations for the image, defining the patches, and the feature extraction per patch. The method described in this study was ranked first in the ImageCLEF 2009 medical annotation task. It uses dense sampling of simple features with spatial content, and a nonlinear kernel-based SVM classifier. The system detail is presented in Section II. An extensive set of experiments was conducted to optimize the set of components and respective parameters comprising the system. A summary of the optimization procedure is presented in Section III. We present an initial application to pathology-level categorization of chest X-ray data. Motivated by the success of chest and viewpoint identification, we extended the system to pathology-level discrimination. The input the radiologists provide is a global label for the entire image (healthy/pathology), and the categorization is conducted on the entire image, with no need for segmentation algorithms or any geometrical rules. The system presented here provides a new tool which can assist the radiologist in a variety of possible scenarios. For example, it can be utilized as a screening filter to support prioritization of cases for the medical expert. We explore chest pathology data in Section IV. A preliminary version of this study appeared in [6] and [5].

II. Visual Words Framework for Classification and Retrieval

In this section we describe the visual words framework. Fig. 1 displays a block-diagram of the image representation, which is based on a large set of image patches, and their respective representation via a learned dictionary. Following the representation phase, various similarity measures can be used for retrieval and well-known categorization schemes, such as SVM, can be used for classification. Classification and retrieval are defined in Section II-B.

There are three main datasets used in this work. 1) The IRMA archive [20], which is the basis for the ImageCLEF medical annotation competitions. It contains over 12 000 X-ray images which are categorized into 196 different, organ-level, categories. Sample IRMA images can be seen in Fig. 3. 2) A subset of the GoldMiner collection [27], used in ImageCLEF medical retrieval competitions. This is a collection of over 66 000 medical images, taken from leading radiology journals. 3) For our pathology experiments, we use chest X-rays obtained in the emergency room of Sheba Medical Center. We used 98 frontal chest images in DICOM format from the hospital PACS, taken during routine examinations. X-ray interpretations, made by two radiologists, served as the reference gold standard. The radiologists examined all of the images independently; they then discussed and reached a consensus regarding the label of every image. Sample images from the Sheba Medical Center are shown in Fig. 16. Additional details on the IRMA, GoldMiner and Sheba datasets are provided in Sections III-A, III-D, and IV, respectively.

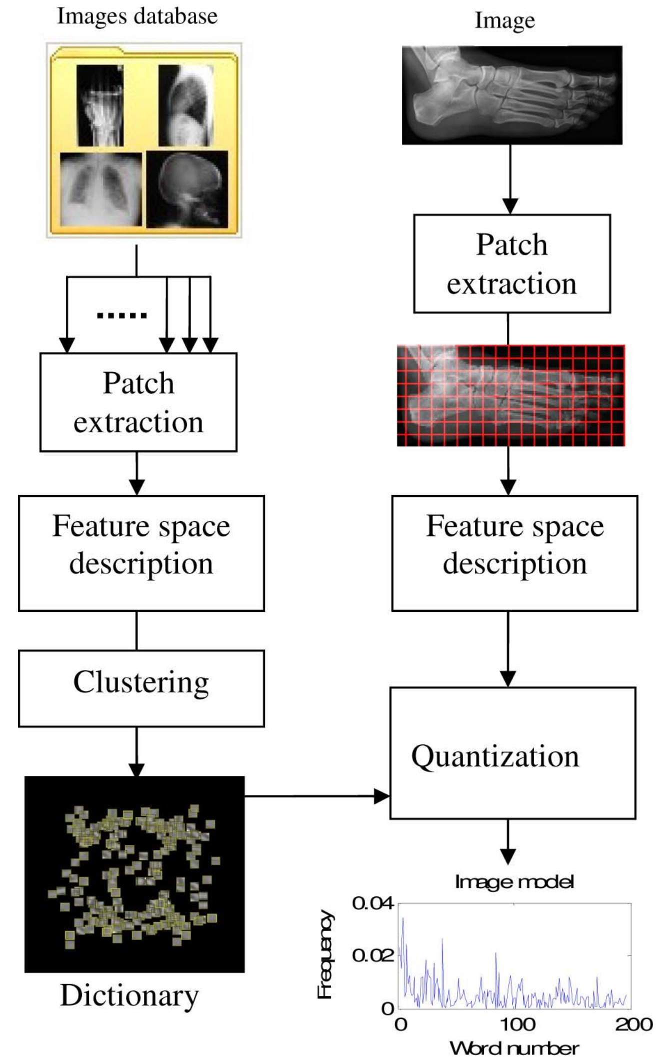


Fig. 1. Dictionary building and image representation flow chart.

A. Feature Extraction Step

Given an image, points of interest detection is used to extract several small local patches. Each small patch shows a localized view of the image content. These patches are considered as candidates for basic elements, or "words." The patch size needs to be larger than a few pixels across, in order to capture higher-level semantics such as edges or corners. At the same time, the patch size should not be too large if it is aimed to serve as a common building block for many images. In visual word image representation we are not directly using the image patch.

There is a quantization step by choosing a visual word from the dictionary that is most similar to the patch. If the patch size is large this quantization process is problematic since it is not likely that there exists a visual word that is similar enough to this patch. We chose a patch size of 9×9.

Common points of interest detection approaches include using a regular sampling grid, a random selection of points, or the selection of points with high information content using salient point detectors. We utilize all the information in the image, by sampling rectangular patches of fixed size around every pixel. This simple feature detection approach has been shown to be effective [37].

Following points of interest detection, the feature representation method involves representing the patches using feature descriptors. In this step, a large set of images is used (ignoring their labels). We extract patches using a regular grid, and normalize each patch by subtracting its mean gray level, and dividing it by its standard deviation. This step insures invariance to local changes in brightness, provides local contrast enhancement and augments the information within a patch. Patches that have a single intensity value are abundant in X-ray images (e.g., the brightness of the air surrounding the organ appears uniform especially in DICOM format).

These patches are common in all categories, much like stopwords in text documents [33]. These patches are ignored. We are left with a large collection of several million vectors. To reduce both the computational complexity of the algorithm and the level of noise, we apply a principal component analysis procedure (PCA) to this initial patch collection. The first few components of the PCA, which are the components with the largest eigenvalues, serve as a basis for the information description. A popular alternative approach to raw patches is the SIFT representation [32], a scale and rotation invariant description based on a local edge histogram. SIFT has been shown to be advantageous in scenery images [22], [46], where object scales can vary. We examine this option in the experiments defining the system parameter set.

In addition to patch content information represented either by PCA coefficients or SIFT descriptors, we add the patch center coordinates to the feature vector. This introduces spatial information into the image representation, without the need to explicitly model the spatial dependency between patches. Special care should be taken when combining features having different units, such as coordinates and PCA coefficients. The relative feature weights were tuned experimentally on a cross-validation set (see Section III).

The final step of the bag-of-words model is to convert vector-represented patches into *visual words* and generate a representative *dictionary*. A visual word can be considered to be a representative of several similar patches. A frequently-used method is to perform K-means clustering over the vectors of the initial collection. The vectors are then clustered them into K groups in the feature space. The resultant cluster centers serve as a vocabulary of K visual words. A sample dictionary of 1000 visual words generated by this process is shown in Fig. 1.

Due to the fact that we included spatial coordinates as part of the feature space, the visual words have a localization component in them, which is reflected as a spatial spread of the words in the image plane. Words are denser in areas with greater variability across images in the database.

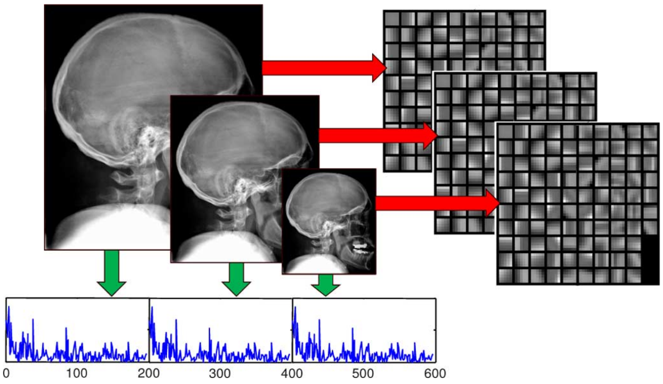


Fig. 2. Image representation at multiple scales.

A given (training or testing) image can now be represented by a unique distribution over the generated dictionary of words. In our implementation, patches are extracted from every pixel in the image. For a 512×512 image, there are several hundred thousand patches. The patches are projected into the selected feature space, and translated (quantized) to indices by looking up the most similar feature-vector in the generated dictionary. We store the dictionary words in a kd-tree, indexed by the spatial coordinates. A k-dimensional (kd) tree is a space-partitioning data structure for organizing points in a k-dimensional space. The nearest neighbor search can be done efficiently by using the tree properties to quickly eliminate large portions of the search space (see e.g., [10]). Using the spatial indexation of dictionary words, the dictionary lookup process is accelerated by comparing a new patch only to dictionary words at a certain radius from it. The dictionary generation process and the transformation from a given image to its representative histogram, are shown in Fig. 1, left column and right column, respectively. Note that as a result of including spatial features, both the local content and spatial layout of the image are preserved in the discrete histogram representation.

Multiscale image information may in some cases provide additional information that supports the required discrimination. To address this, we repeat the dictionary building process for scaled-down replications of the input image, using the same patch size. The image representation in this case is a 1-D concatenation of histograms from varying scales. This process, illustrated in Fig. 2, provides a richer image representation. It does not imply scale invariance, as in [12]. When detecting body parts and organ orientation, objects of interest in the radiographs appear at a roughly similar size-range across all images, thus invariance to scale is not a necessity.

B. Image Classification and Retrieval

1) Classification: Image classification is based on the ground truth of manually categorized images. We use a nonlinear multiclass SVM classifier. Several nonlinear kernels were examined, which are commonly used with histogram data.

Histogram intersection kernel [7]:

$$K(x,y) = e^{-\sum_i \min(x_i,y_i)}.$$

Radial basis function kernel:

$$K(x,y) = e^{-\gamma\|x-y\|^2}.$$

χ^2 kernel:

$$K(x,y) = e^{-\gamma \sum_i \frac{|x_i - y_i|^2}{|x_i + y_i|}}.$$

Note that histogram intersection has no free kernel parameters, which makes it convenient for fast parameter evaluation. The two other kernels have a free tradeoff parameter γ, and require careful optimization. In order to classify multiple categories, we use the one-versus-one extension of the binary classifier, where binary classifiers are trained for all pairs of categories in the dataset. Whenever an unknown image is classified with a binary classifier it casts one vote for its preferred class; the final result is the class with the most votes. Since each binary classifier runs independently, parallelization of both training and testing phases of the SVM is straightforward. It is implemented as a parallel enhancement of the LIBSVM library.[2]

2) Retrieval: Image retrieval requires a way to measure similarity between images. Using the image representation described in the previous section, the distance between images is defined as the sum of the bin-to-bin distance of the representing histograms. For query image I and target image J, the distance is $d(I,J) = \sum_i d(I_i, J_i)$, where i runs on the bins. A popular choice for the image-to-image distance is the L_1 distance: $L_1(I,J) = \sum_i |I_i - J_i|$.

3) Region-of-Interest Retrieval: For image archives that are noisy or of non-consistent quality, full image matching may result in retrieval results that are much less consistent and informative to the user. A step ahead in medical image retrieval is the concept of region-of-interest (ROI) retrieval. The task in this scenario is defined as follows. The human expert marks an ROI in a given image. This can be a certain anatomical region within the image or a pathology region of interest. The ROI can also be automatically detected by computer, e.g., using CAD algorithms. The system then prioritizes the retrieval results such that a high-confidence matching is required within the ROI and a low priority (or "don't-care" score) is given to the nonmarked regions. The task of the ROI query and retrieval is a challenging one as it requires new ways to represent a region within an image, and new ways to compare a region representation to a full-image representation within the archive [39].

In our image representation, visual words have spatial coordinates. This property can be utilized for ROI based retrieval by adjusting the distance function to be more sensitive to differences in words within the ROI. The histogram distance between a query image I and an image in the database J is adjusted to

$$d^R(I,J) = \sum_i w_i \cdot d(I_i, J_i)$$

where w_i is a weight parameter that reflects the importance of similarity in the ROI. For words outside the ROI $w_i = 1$, and $w_i > 1$ for words inside the ROI. This modification enables comparison of content in specific image regions, without performing additional processing on the database.

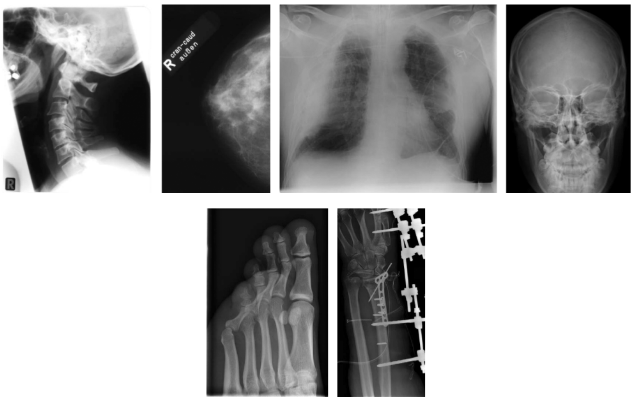


Fig. 3. Sample images from the IRMA database.

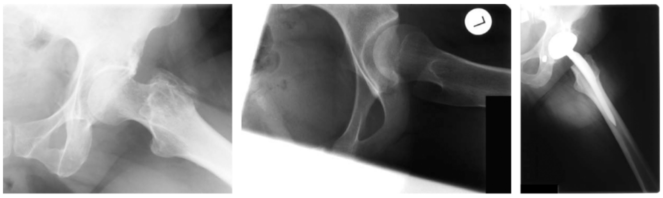


Fig. 4. Images from IRMA category: "Overview image, Mediolateral, Left hip, Musculosceletal system." Large intra-class variability can be seen.

III. Experimental ImageCLEF Results

A key component in using the BoW paradigm in a categorization task is the tuning of the system parameters. An optimization step is thus required for a given task and image archive. We focus on three components of the system: finding the optimal set of local features, finding the optimal dictionary size, and optimizing the classifier parameters. We use a large generic archive of radiographs (IRMA) [20] to tune the system parameters. We then show comparative results of automated organ and orientation detection and visual image retrieval in the ImageCLEF competition.

A. Database

The IRMA database [20] has served algorithm development teams for many years, and in the past several years has been a source for the ImageCLEF medical annotation competition. Images in the IRMA database consist of scanned X-ray images, gray scale, 512 pixels long. A sample of IRMA images is shown in Fig. 3. The X-ray images are noisy with irregular brightness and contrast, and may contain dominant visual artifacts such as artificial limbs and X-ray frame borders. Images in the archive are labeled according to the IRMA coding system [19], with each category described by four axes: 1) A technical axis that describes the image modality; 2) a directional axis that defines body orientation, 3) an anatomical axis that describes the body region examined, and 4) a biological axis that describes the biological system being examined. The axes have a hierarchical description. For example, the complete category description of the image shown in Fig. 4 is: Technical axis: X-ray, plain radiology, analog; Directional axis: Sagittal, mediolateral; Anatomical axis: Lower extremity (leg), hip, left hip; Biological axis: Musculoskeletal system.

[2] http://www.csie.ntu.edu.tw/~cjlin/libsvm

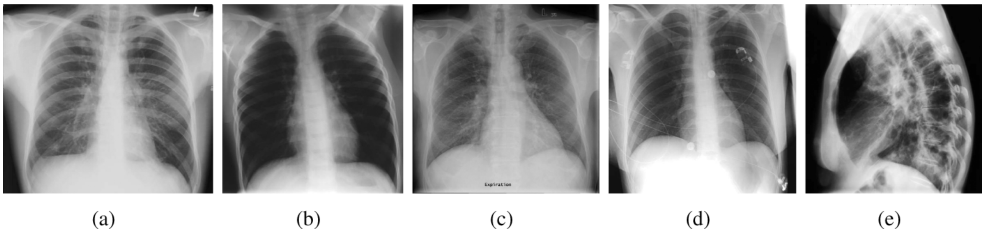


Fig. 5. Different chest categories in the IRMA database. (a) “High beam energy, Posteroanterior.” (b) “Child filter, Anteroposterior–inspiration. (c) “High beam energy, Posteroanterior—expiration.” (d) “High beam energy, Anteroposterior—supine.” (e) “High beam energy, Sagittal—lateral right–left inspiration.”

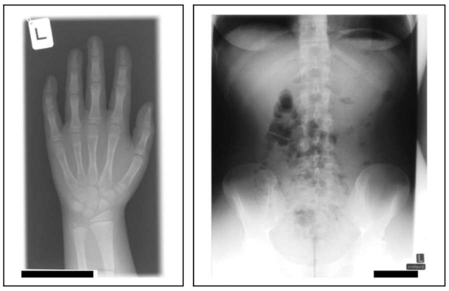


Fig. 6. Sample images with artifacts near the borders, such as misaligned X-ray frame, blacked out bars and various labels.

The database has grown in size and in number of categories over the years. In 2009 it had close to 15 000 images from 196 distinct categories, with category labels consisting of the four axes defined above. Some classes have large intra-variability, as seen for example in Fig. 4, while images from different classes may be visually similar, as seen in Fig. 5. These properties make the automatic classification task challenging.

B. Optimization of System Parameters

We optimized the system parameters by classifying subsets of the database, using several cross-validation experiments. The optimization is performed independently in three steps: finding the optimal set of local features, finding optimal dictionary size, and optimizing classifier parameters. In the following experiments, unless otherwise noted, 10 cross-validation experiments were run per case, with 10 667 images used for training and 2000 images used for testing. In all the experiments there is a clear separation between train and test datasets which are disjoint sets.

We examined three feature extraction strategies: raw patches, raw patches with normalized variance, and SIFT descriptors. In all cases we added the patch center coordinates to the feature vector as described in Section II-A. We used dense extraction of features around every pixel in the image. There are often strong artifacts near the image border that are not relevant to the image category. This is evident in the IRMA dataset, as seen in Fig. 6. In this case we chose to ignore a 5% margin from the image border. As a result of the dense sampling, a single image (following the border removal step) yields a large feature set of between 100 000 to 200 000 features.

It is our experience that X-ray images from the same category usually appear with a similar scale and orientation in a given archive. In this task the invariance of the SIFT features to scale and orientation is thus unnecessary. We used SIFT descriptors taken at a single scale, with no orientation alignment [41].

TABLE I
COMPARISON OF DIFFERENT FEATURES, USING SVM CLASSIFIER WITH HISTOGRAM INTERSECTION KERNEL

Features	Average %	Standard Deviation
Raw Patches	88.43	0.32
SIFT	90.80	0.41
Normalized	**91.29**	0.56

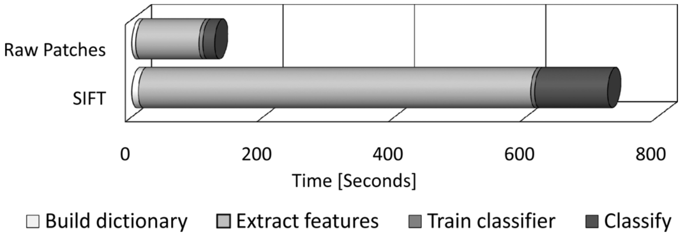


Fig. 7. Running time using SIFT descriptors and normalized raw patches.

The three feature sets tested: raw patches, normalized patches and the 128 dimensional SIFT descriptors, were reduced in dimension using PCA. Classification was done using an SVM classifier with the histogram intersection kernel. Images were classified into one of the IRMA categories as defined above. Table I summarizes the percentage of correct classification of the three feature sets, averaged over 10 runs. In each run we used randomly chosen 10667 images for training and 2000 images for testing. Normalizing patch variance improved the classification rate significantly compared to raw patches (t-test p-value less than 0.0001). The gain can be attributed to the local contrast invariance achieved in this step. Using the normalized raw patches proved marginally preferable to the SIFT descriptors in this task, in terms of classification accuracy (t-test p-value of 0.0385).

The advantage of using normalized raw patches over the SIFT descriptors is even more significant when considering the computational cost of the process. Using raw patches, the feature extraction step was significantly faster than with SIFT descriptors, as seen in Fig. 7. Most of the running time was spent in the image representation step; this step took over three seconds per image with the SIFT features, but less than half a second with the simpler variance-normalized raw patches. Time was measured on a dual quad-core Intel Xeon 2.33 GHz.

Fig. 8 depicts the effect of using four to ten PCA components for variance-normalized raw patches. It can be seen that the number of components in this range has no effect on classification accuracy (ANOVA p-value equals 0.998). The addition

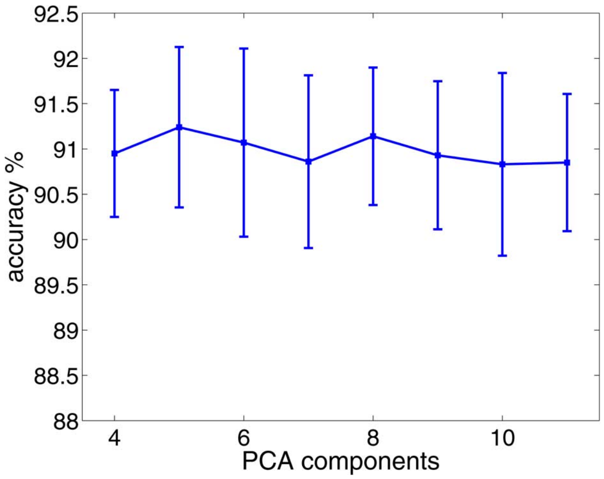


Fig. 8. Effect of the number of PCA components in a patch on classification accuracy.

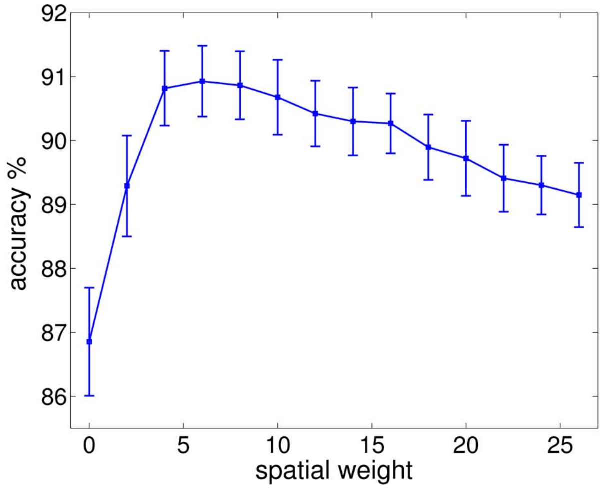


Fig. 9. Effect of spatial features: weight of spatial features (x-axis); classification accuracy (y-axis).

of spatial coordinates to the feature set, on the other hand, improved classification performance noticeably, as seen in Fig. 9. We found that when using seven PCA components, each having standard deviation of 1, the optimal weights for the x, y position of the patch center, $(0 < x, y < 1)$, is 6 as can be shown in Fig. 9. The bars in Fig. 9 show means and standard deviations from 20 cross validation experiments running on 1000 random test images. Using the optimal weight for the spatial coordinates increased the classification accuracy by 4% as compared to the case of not using spatial coordinates (t-test p-value less than 0.0001). Using higher than optimal values gradually decreases the performance. For example a weight of 10 decreases the performance by 0.25% (t-test p-value of 0.011).

As Fig. 10 shows, increasing the number of dictionary words proved useful up to 1000 words. For example, 1000 words was found to be better than 800 with a t-test p-value of 0.0146. Adding additional words after that point increased computational time with no evident improvement in the classification rate.

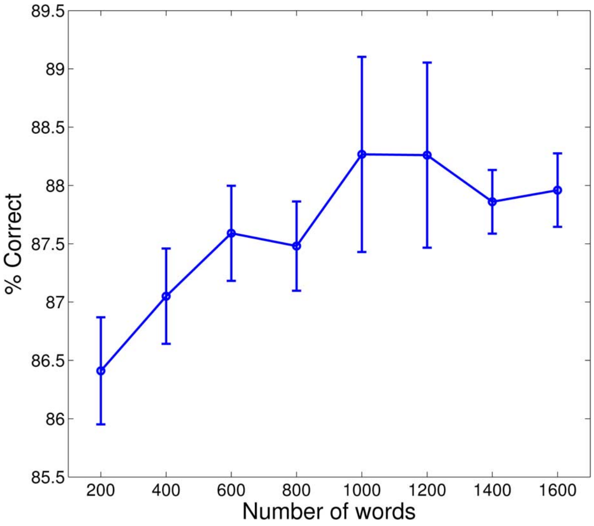


Fig. 10. Effect of dictionary size on classification accuracy.

TABLE II
COMPARISON OF SVM KERNEL TYPES, FOR 1-SCALE AND 3-SCALE MODELS

Kernel	Average % 1-scale	Average % 3-scales
Radial Basis	91.45	91.59
Histogram Intersection	91.29	91.89
χ^2	91.62	**91.95**

We used the SVM classifier with three possible kernels: the histogram intersection, the Radial Basis Function and the χ^2 kernels. We used the optimal features and dictionary size consistently across all experiments. The SVM cost parameter C, and free kernel parameter γ were scanned simultaneously over a grid to find the classifier's optimal working point. The histogram intersection kernel does not have a free kernel parameter, and the optimization is one dimensional over the SVM cost parameter. Table II summarizes the results using the best parameters for the different kernels. The χ^2 kernel is ranked first by a small margin with 91.62% accuracy, followed by the RBF kernel with 91.45%.

In the final experiment, we took information from multiple image scales into account by repeating the dictionary creation step on scaled-down versions of the original image. The image representation was thus a concatenation of histograms built on the single scale dictionaries. We used three scales: the original image, 1/2 size and 1/8 size. Using three scales further improved the accuracy for all kernels, as seen in the right-most column of Table II. The improvement was significant with the histogram intersection kernel (p-value = 0.0454), and insignificant with radial basis and chi-square kernels (p-value of 0.2140 and 0.3637, respectively). When using three scales the difference in performance between the kernels was not significant (ANOVA p-value of 0.267). The average classification accuracy with the χ^2 kernel was 91.95%.

C. ImageCLEF Classification Results

Based on the optimization procedure described above, the classification system has the following set of components and parameter settings (used throughout the rest of this paper). The

TABLE III
MEDICAL IMAGE ANNOTATION RESULTS FROM IMAGECLEF 2009 [21]. ERROR SCORE IS DEFINED IN [21], LOWER IS BETTER

Run & Error score	2005	2006	2007	2008	Sum
Our system	**356**	263	**64**	**169**	**852**
Idiap6	393	**260**	67	178	899
FEITIJS	549	433	128	242	1352
VPA-SabanciUniv	578	462	155	261	1456
MedGIFT	618	507	190	317	1633
IRMA	790	638	207	359	1994
VPA-SabanciUniv	587	1170	413	574	2744
DEU	1368	1183	487	642	3681

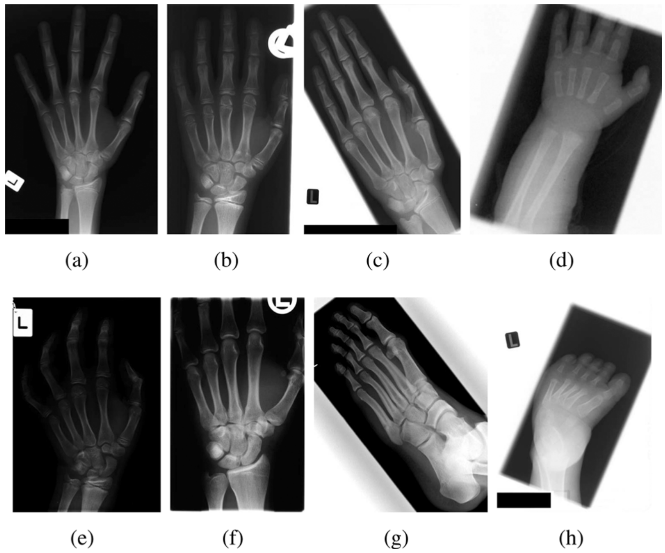


Fig. 11. Detecting category "posteroanterior, left hand:" (a)–(d) Correctly classified. (e) False negative, misclassified as "left anterior oblique, left hand." False positives come from categories: (f) anteroposterior, left carpal joint; (g) anteroposterior, left foot; (h) right anterior oblique, right foot.

system uses a set of densely extracted normalized raw patch features, with seven PCA components, spatial features with weight 6, and 1000 visual words. For classification we used the SVM algorithm with a χ^2 kernel.

Table III shows the accuracy of the classification system on four sets of data, taken from consecutive years of the ImageCLEF competition, as provided in the ImageCLEF 2009 medical annotation task. This task introduced a new test set of 1733 images, which were not included in the training. Results are compared to other submitted runs, with the best result in each column marked in bold. There were 19 submissions from seven research groups. Our system, presented in [4], was ranked first on three of the four sets, and first when using an overall error score [21]. The error score measures the classification error based on a predefined hierarchical structure of the classes. The error is zero when there is a complete match, and a positive number when there is a mismatch. The error score is defined such that confusion between similar classes is penalized less than a confusion between unrelated classes.

Fig. 11 illustrates the challenge in the IRMA archive categorization task, using the predefined IRMA categories. Correctly labeled images from the "Posteroanterior, Left hand" category are shown in Fig. 11(a)–(d). In this run there were 2000 random test images, with 57 images from the examined category, out of which 56 were correctly detected by the described system. A single image, (e), was falsely classified and was detected as a neighboring category—"Left anterior oblique, Left hand" (false negative). Three images from other categories, (f)–(h), were misclassified as "Posteroanterior, Left hand" (false positives). As can be observed, these images have a strong visual resemblance to the left hand category.

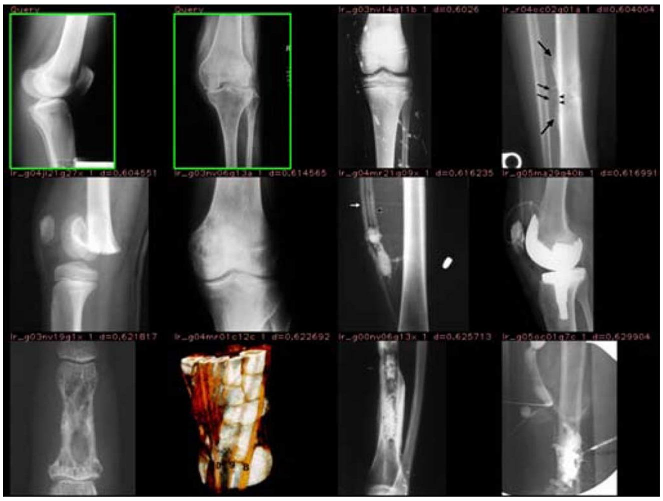

Query
Query
lr_g03nv14g11b 1 d=0.6026
lr_r04ec02g01a 1 d=0.604004
lr_g04ji21g27x 1 d=0.604551
lr_g03nv06g13a 1 d=0.614565
lr_g04mr21g09x 1 d=0.616235
lr_g05ma29g40b 1 d=0.616991
lr_g03nv19g1x 1 d=0.621817
lr_g04mr01c12c 1 d=0.622692
lr_g00nv06g13x 1 d=0.625713
lr_g05ec01g7c 1 d=0.629904

Fig. 12. Retrieval example: First two framed images are the query images; the following images (left to right, top to bottom) are retrieval results.

D. ImageCLEF Retrieval Results

The ImageCLEF 2008 [18] retrieval competition used a database of over 66 000 medical images which are part of the GoldMiner collection [27]. These images are taken from open-access content from five leading peer-reviewed radiology journals. Images come from different imaging modalities including radiography, CT, MRI, sonography, PET and others, as well as digital photos and graphic charts. Images were given in high resolution JPEG format, in either color or grayscale, depending on the modality.

The challenge was to answer 30 query topics, composed of one or more sample JPEG images and a short textual description in several languages. The objective was to return a ranked set of 1000 images from the complete database, sorted by their relevance to the presented queries. A sample query from this challenge and the first few returned images are shown in Fig. 12. The retrieved results were manually judged for relevance by medical experts. We applied a purely visual retrieval approach, disregarding the textual labels given in this task. There were no parameter adjustments suited to the specific queries or for the database of this task.

The system parameters were the same as those optimized using the database of X-ray images on the medical image annotation task, described above. For image comparison, L_1 distance measures between the representative histograms were used. Retrieved images were ranked by the distance between the target histogram and the histogram of the query image. When there were multiple query images, we used the minimal distance between the target and the query set. We tried two variants. In the first run we normalized the mean gray level and variance within a patch prior to performing PCA, for local brightness and

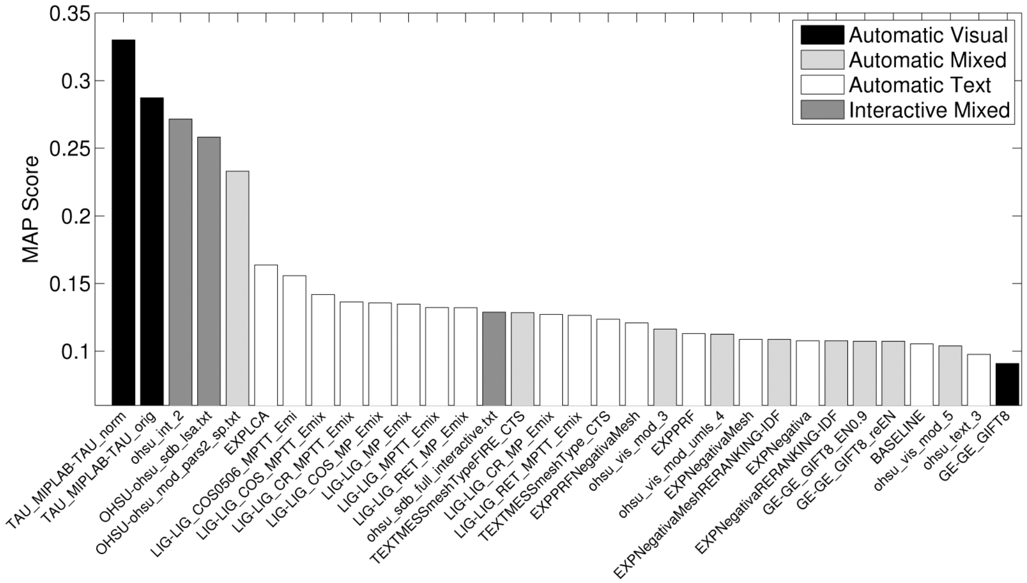


Fig. 13. MAP Score for a specific query: finding chest X-ray images of cases with tuberculosis from the ImageCLEF 2008 medical retrieval task. Our runs are the first two on the left.

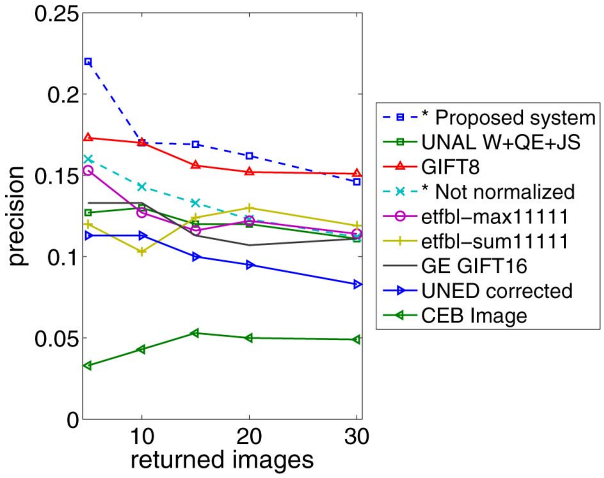


Fig. 14. Precision graph of visual retrieval systems; ImageCLEF 2008 medical database [18]. Average precision over 30 queries is shown for first 5, 10, 15, 20, and 30 returned images. Results of the proposed algorithm marked with dashed lines.

contrast invariance. In the second run we kept the original gray levels.

Fig. 14 shows the precision score averaged over answer 30 query topics of our submitted runs, marked with an asterisk $(*)$, along with visual retrieval algorithms submitted by additional groups [18]. In this figure, the run labeled "Proposed System" used patch normalization and the run labeled "Not Normalized" used the patches' original gray levels. The normalized patch approach system ranked first among the nine automatic purely visual competitors.

Purely visual methods traditionally rank low relative to text-based methods on medical image retrieval tasks [17], [18]. In absolute values, the precision of all purely visual methods are low. Some abstract queries are extremely difficult based only on visual similarity (for example: "Show me images of tumors"). Nevertheless, on three of the 30 queries our system was ranked in the top three of all 111 submitted runs, whether textual, visual or mixed systems, or using automatic, interactive or manual methods. In query number 15—finding X-ray images with Tuberculosis, our system was ranked first out of all systems, followed by an interactive visual+textual based system, far ahead of other automatic purely visual systems. The mean average precision (MAP) score [33] for the top 32 runs on this query appears in Fig. 13.

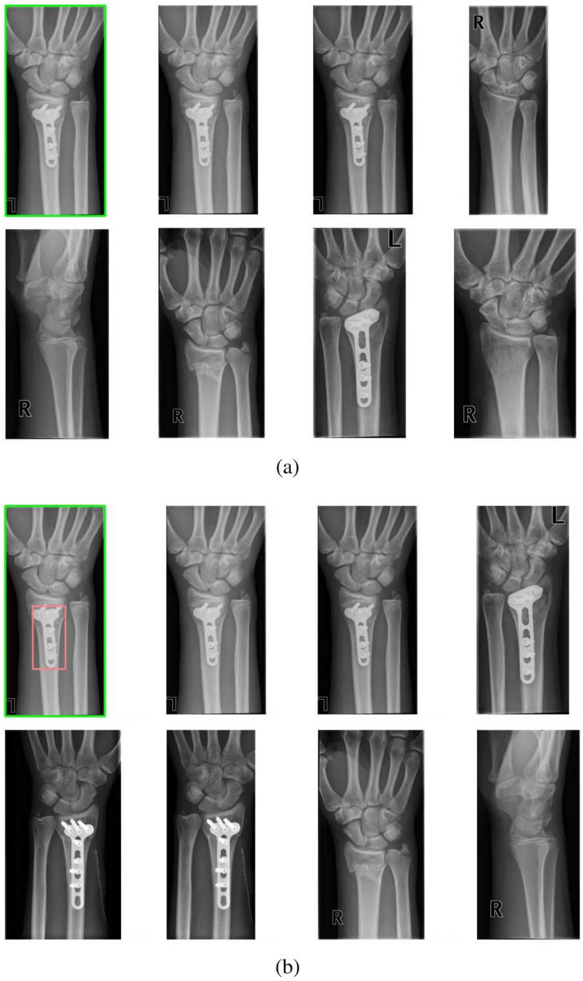
The retrieval system is also computationally efficient, with an average retrieval time of less than 400 ms per query.

1) ROI-Based Retrieval Results: A sample ROI query and retrieval are shown in Fig. 15(a). The query image (top left) is a left arm with a metal fixation device. Retrieved images are returned by order of similarity from left to right, top to bottom. Since the query image is part of the database, the first returned image is the query image itself. Except for the first image, images 2 and 5 have a similar metal fixation.

In Fig. 15(b) the user selects the metal fixation device as a region of interest. The difference of visual words in the ROI is multiplied by $w_i = 10$ inside the ROI, and $w_i = 1$ outside the ROI. The selection of an ROI in this case returned images with a fixation in the top five returned images. This exemplifies how a simple weighting of the distance function can be used to locate an interesting object in the database. Since the coordinates are part of the features, the similarity distance is not invariant to large translations of the ROI. This method is therefore limited to locating similar objects in the near vicinity of the ROI.

IV. Chest X-ray Characterization

Chest radiographs are the most common examination in radiology. They are essential for the management of various diseases associated with high mortality and morbidity and display a wide range of potential information, many of which are subtle. According to a recent survey [42], most of research in computer-aided detection and diagnosis in chest radiography has focused on lung nodule detection. Although the target of most


(a)

(b)

Fig. 15. A sample ROI query and retrieval. (a) Full image query. (b) Query with selected ROI.

research attention, lung nodules are a relatively rare finding in the lungs. The most common findings in chest X-rays include lung infiltrates, catheters and abnormalities of the size or contour of the heart [42]. Distinguishing the various chest pathologies is a difficult task even to the human observer. Research is still needed to develop an appropriate set of computational tools to support this task.

We focus next on the analysis of chest radiographs. Our study starts with chest identification and viewpoint determination (on the organ level) within the IRMA archive. We conclude with an initial set of experiments on data obtained in a clinical setting, in which we deal with pathology screening as well as the identification of individual pathologies including right and left pleural effusion, enlarged heart and cases of enlarged mediastinum.

Fig. 5 shows sample images from chest-related categories in the IRMA archive. There are four IRMA categories that contain frontal chest views, Fig. 5(a)–(d), and one category of a lateral view, Fig. 5(e). We selected a random set of 2000 images from the archive. Out of the 2000, 1938 had category labels which we could use as our ground-truth. In the first experiment we tested our ability to discriminate between chest (725 images) and nonchest categories (1213 images). Once the chest images were detected, a follow-up experiment was conducted to test the system's ability to discriminate between chest viewpoints with 559 frontal views and 166 lateral views in the chest image set.

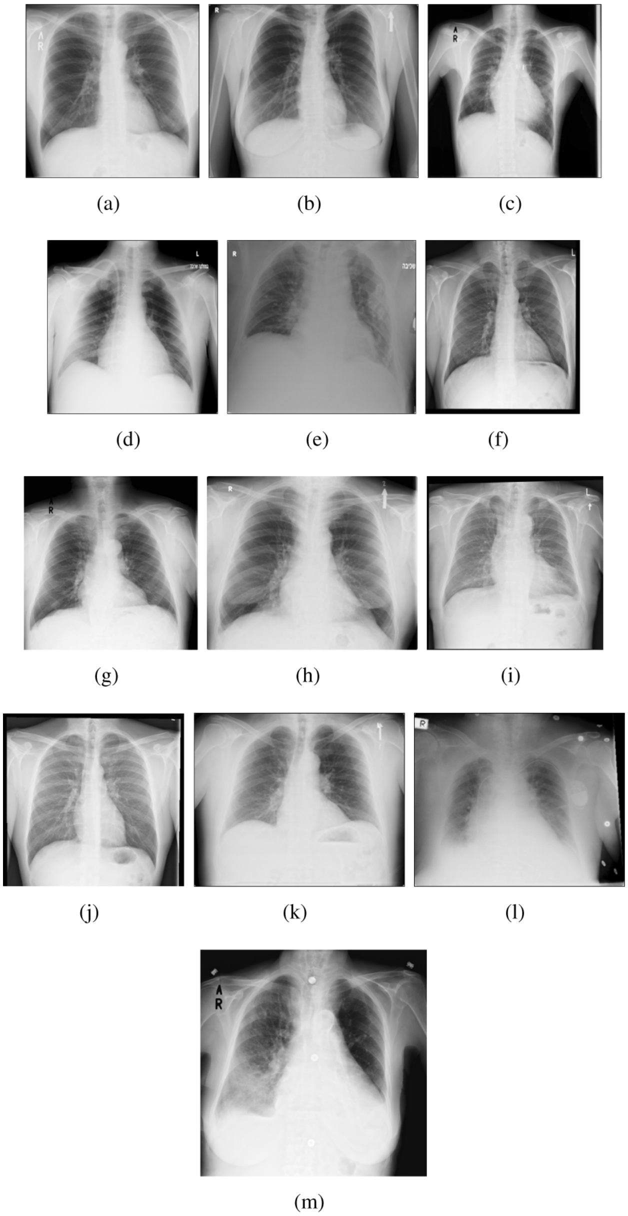
TABLE IV
CONFUSION MATRIX FOR CHEST CATEGORIES, FIRST FIVE COLUMNS ARE CATEGORIES SEEN IN FIG. 5

Detected \ True	(a)	(b)	(c)	(d)	(e)	Non-chest
Frontal (a)	141	0	17	6	0	0
Frontal (b)	0	17	3	0	0	0
Frontal (c)	12	1	355	2	0	0
Frontal (d)	0	0	1	4	0	0
Lateral (e)	0	0	0	0	166	1
Non-chest	0	0	0	0	0	1212

Table IV summarizes the experiments that were conducted. We used 10677 images from the IRMA dataset to train an SVM classifier and a separate set of 2000 images were used for testing. In detecting the chest images from assorted X-ray images of different body parts, sensitivity and specificity reached virtually 100% (sensitivity 100%, specificity 99.92%). In detecting chest viewpoints, 559 of 559 frontal chest images were detected, with no false positives. 166 of 166 lateral chest images were detected, with 1 false positive. The high accuracy in this case is expected. Detecting a chest from a non-chest, as well as detecting frontal versus lateral images are relatively easy tasks, since the categories are visually different. Discriminating between the frontal chest categories is more challenging. Still, there was less than 10% confusion between the visually similar frontal chest subcategories.

In our final investigation, we applied our system to chest X-rays obtained in the emergency room of Sheba Medical Center. We used 98 frontal chest images in DICOM format from the hospital PACS, taken during routine examinations. X-ray interpretations, made by two radiologists, served as the reference gold standard. The radiologists examined all of the images independently; they then discussed and reached a consensus regarding the label of every image. For each image and pathology type, a positive or negative label was assigned: 38 of the images were diagnosed as normal, 55 images had at least one pathology and the other five images were labeled as inconclusive. Fig. 16 shows a set of healthy (a)–(c) and pathological images (d)–(m). Pathology data include 24 images with enlarged heart shadow [three examples shown in Fig. 16(d)–(f)], 19 images with enlarged mediastinum, Fig. 16(g)–(i), 17 images with right pleural effusion and 21 images with left pleural effusion, Fig. 16(j)–(l). Some patients had multiple pathologies. For example, Fig. 16(m) exhibits all pathologies. We treated the multiple pathology detection as a set of binary classification tasks, where in each task we tried to detect an individual pathology.

The original high-resolution DICOM images were initially resized to a maximal image dimension of 1024 pixels, with aspect-ratio maintained. We follow the method described in Section II to extract features, build a visual dictionary, and represent an image as a histogram of visual words in multiple scales. We then detected each of the four pathologies using a binary SVM classifier, with a histogram intersection kernel. In


(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j) (k) (l)

(m)

Fig. 16. Frontal chest X-ray images, Sheba Medical Center: (a)–(c) healthy; (d)–(f) enlarged heart; (g)–(i) enlarged mediastinum; (j)–(l) left or right effusion; (m) multiple pathologies: enlarged heart and mediastinum, left and right effusion.

addition to individual pathology detection, we trained a classifier to distinguish between a healthy image versus a nonhealthy image (with any kind of pathology).

In order to avoid overfitting and to preserve the generalization ability of the classifiers, system parameters were used as found in the analysis described in Section III. Since the database was fairly small, a leave-one-out classification was implemented.

The data were unbalanced, usually with more healthy images than abnormal images. To address this problem we used an asymmetric penalty in the SVM training step, as in [38], with a higher cost to false negative errors. Modifying the relative cost of false negative errors determines the tradeoff point between sensitivity and specificity. This technique was used to produce the receiver operating characteristic (ROC) curves, shown in Fig. 17(a)–(e). The area under the curves (AUC), shown in Fig. 18, was calculated using trapezoidal approximation. Error bars show one standard deviation of the AUC, estimated using [26]. The software correctly identified abnormal mediastinum and effusions with an average AUC of around 80%, with standard deviation around 6%. Enlarged heart was detected with AUC of 88.19%(±4.72%). ROC curves were compared using the Mann–Whitney U-statistics [44]. Including spatial features improved the AUC when detecting right and left pleural effusions (p-value = 0.00017 and 0.0021, respectively), and enlarged heart (p-value = 0.00856). The improvement was not significant for abnormal mediastinum detection (p-value = 0.1189) and for non-healthy images detection (p-value = 0.93489). When including spatial features, there was no significant difference between raw patches and SIFT descriptors on any of the pathology detection tasks.

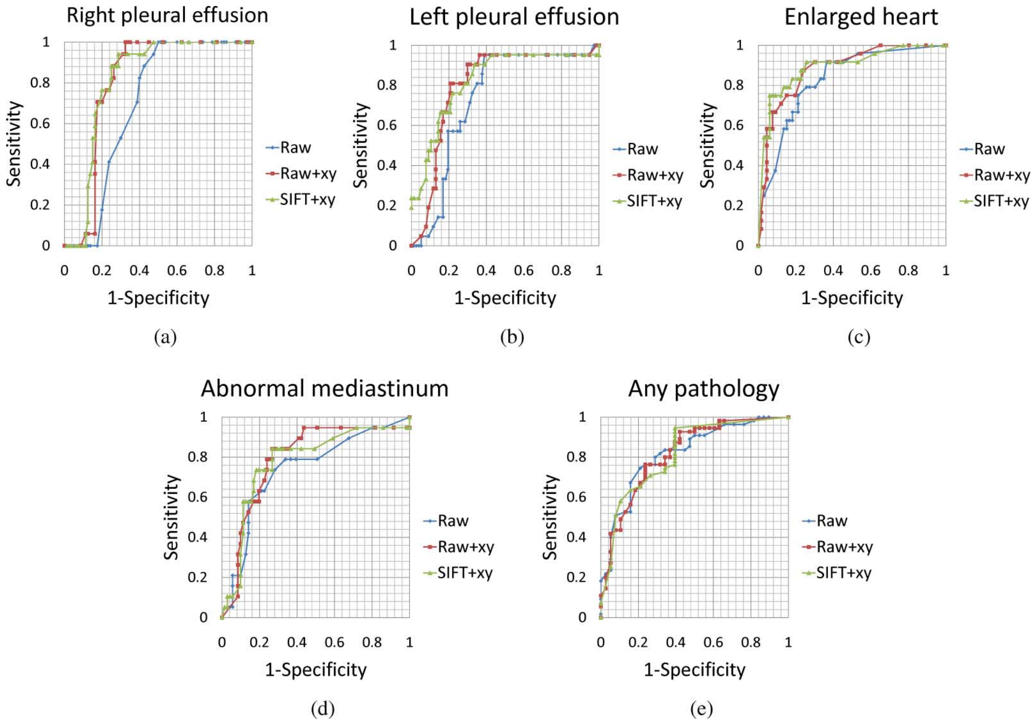
Retrieval examples are shown in Fig. 19. A sample query image is shown in the left column, with ranked results, ordered from left to right, shown as retrieval results on the right.

V. Discussion and Conclusion

In this study we presented a visual words approach to medical image categorization and retrieval. We provided a comprehensive overview of the methodology and its application to ImageCLEF and in clinical settings. Statistical analysis of the results is shown on both the CLEF dataset, on the organ-level, and the Sheba chest X-ray dataset, on the pathology level. Retrieval is discussed in both domains, with initial discussion into ROI-based retrieval.

The transition from general imagery analysis to the medical image analysis and furthermore to applications in the clinical settings is not a trivial one. We investigated the effects of various parameters on overall classification, and tuned the system to achieve high accuracy in the classification of general X-ray images. We reported state-of-the-art results in the task of organ and orientation identification in the ImageCLEF 2009 medical annotation challenge, Table III, and top retrieval results among the purely visual based systems in the ImageCLEF 2008 medical retrieval challenge, as shown in Figs. 14 and 15. In the chest pathology categorization task, Figs. 17 and 18 indicate detection of left and right effusions with an AUC of 80%, abnormal mediastinum with an AUC of 79.2% and enlarged heart with an AUC of 88.2%. Abnormal images of any kind are detected with an AUC of 82%. These rates are compatible with many other medical tests, such as blood tests. We therefore view them as encouraging for further exploration towards future clinical use.

Two key characteristics that were evaluated throughout this work are the representation of the patch as normalized raw values versus SIFT, and the use of spatial features as part of the representation space. We found that using the raw (pixel) data, with minimal processing (versus SIFT as commonly used (e.g., [41]), gives good results, as long as a large amount of data are used. We propose using all the available data, and not subsampling it, as is common in the literature. A marginal advantage for using normalized raw data over SIFT descriptors is seen in Table I, for the ImageCLEF dataset. We show the percentage of correct classification averaged over 10 runs. In each run we used randomly chosen 10 667 images for training



(a) (b) (c)

(d) (e)

Fig. 17. ROC curves for pathology detection in frontal chest X-rays; Raw—normalized pixel-content; Raw+xy—content features with coordinates; SIFT+xy—SIFT features with coordinates.

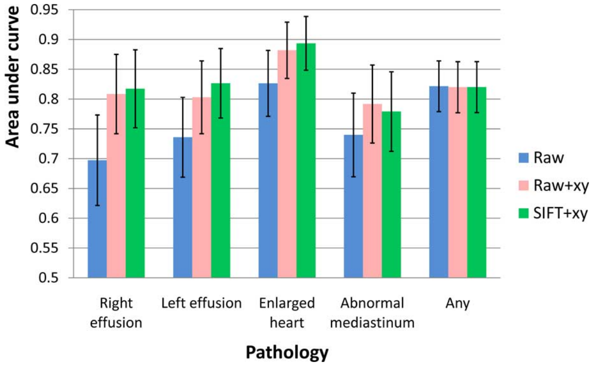


Fig. 18. Area under curves of the ROC curves.

and the rest 2000 images were used in the test step. In the clinical study, Fig. 17, we again see very close results between the two representations.

Instead of coarsely binning the multidimensional space (as in [12]) we use a dictionary, which provides for a data-tuned clustering of the space. We are able to tune the representation more closely to the data at-hand, without deterioration of the efficiency of the retrieval.

Incorporating the spatial information is shown to be advantageous in most of the scenarios. Fig. 9 shows a significant improvement when using spatial information for the ImageCLEF archive. Fig. 17 shows the spatial features improve the results in most of the individual chest pathology cases, but not in the overall pathology versus nonpathology case.

Using dense sampling while keeping the features simple makes the system both accurate as well as computationally efficient. The system achieves approximately half a second training and classification time per image. The use of the normalized raw pixel values as features together with spatial coordinates enables fast classification times: these (simple) features are about six times faster to calculate than SIFT descriptors. Indexing the dictionary by the spatial features provides for accelerated lookup. Based on the optimization experiment for the relative weighting of spatial features as compared to the PCA components, we obtained a variance of 4 for the spatial features, as compared to a variance of 1 for the PCA based features. This indicates that for optimal accuracy the spatial features have more energy than content (intensity-based) features. Therefore, the acceleration achieved by the indexed dictionary is substantial.

The initial retrieval results presented here, although with low precision, seem promising. For the full-image retrieval case we were ranked in first-place for visual-only retrieval, as shown in Fig. 14. In this task, the best average MAP score of all systems was 0.2908. the average MAP scores of purely visual systems was between 0.0094 (worst) and 0.0421 (ours). When looking at scores of individual queries, 86.5% of all 3390 submitted individual queries had a MAP score of less than 0.33. The MAP score of purely visual queries was between 0 and 0.3369. Overall, the score we achieved of 0.33 is considered high for the defined task. It is exceptionally high for a purely visual query.

Although encouraging, the mistaken images retrieved and the low precision values indicates the need for refining the retrieval process for the full image case, and even more so in the ROI procedure. One possible scheme to consider is a hierarchy of classification and retrieval, in which the query image is first classified

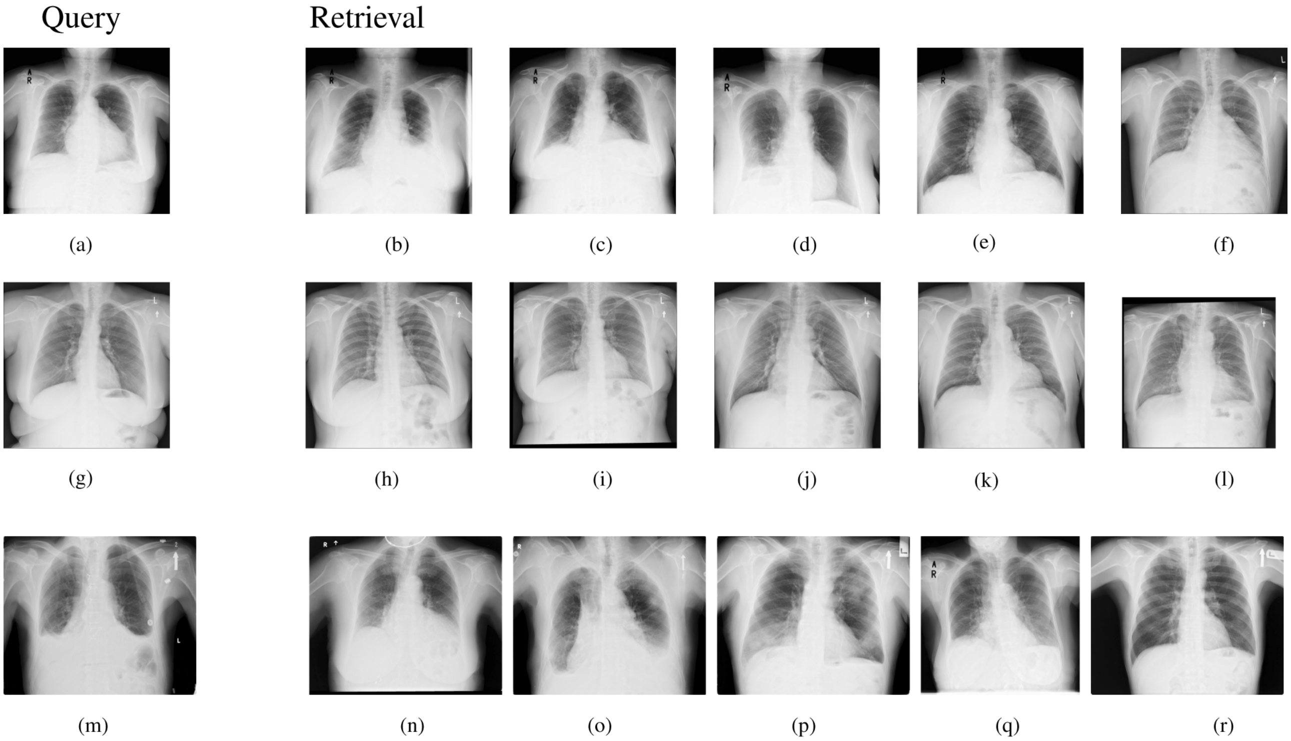


Fig. 19. Image retrieval of pathology examples from the Sheba dataset. A sample query image is shown in the left column, with ranked results, ordered from left to right, shown as retrieval results on the right. (a) Enlarged heart. (b) Enlarged heart. (c) Enlarged heart. (d) Enlarged heart. (e) Healthy. (f) Enlarged heart. (g) Healthy. (h) Healthy. (i) Healthy. (j) Healthy. (k) Healthy. (l) Left pleural effusion. (m) Right+Left pleural effusions. (n) Enlarged heart. (o) Right+Left pleural effusions. (p) Left pleural effusion. (q) Right pleural effusion. (r) Healthy.

and retrieval results are shown from that class only. In actual clinical settings it is important to consider the combined visual information with additional information such as the patient's demographics and the text extracted from clinical reports. Fusion across the modalities will facilitate an increase in retrieval performance. An additional limitation of the current approach is the fact that the visual word representation is a global representation of the image content. In cases where the category characterization is local and relatively small (e.g., lesions) a global image representation can miss-detect the pathology and result in a misclassification of the image.

Future work involves augmenting system capabilities by combining *a priori* task-specific knowledge such as a textual description with the visual words framework. In the clinical scenario we are currently working on increasing the collection of chest images and pathology types. We believe that the system presented here provides a new tool that can assist the radiologist in a variety of possible scenarios. It can provide a screening filter to support prioritization of cases for the medical expert. It can identify suspicious pathological X-rays and alert the referring clinicians to potential emergencies. Overall it is hoped that the development of such systems will contribute to the improvement of safety and quality of medical services.

REFERENCES

[1] C. B. Akgul, D. L. Rubin, S. Napel, C. F. Beaulieu, H. Greenspan, and B. Acar, "Content based image retrieval in radiology: Current status and future directions," *J. Digital Imag.*, Jan. 2010.

[2] H. Alto, R. M. Rangayyan, and J. E. L. Desautels, "Content-based retrieval and analysis of mammographic masses," *J. Electron. Imag.*, vol. 14, no. 2, 2005.

[3] B. André, T. Vercauteren, A. Perchant, A. Buchner, M. Wallace, and N. Ayache, "Introducing space and time in local feature-based endomicroscopic image retrieval," in *Medical Content-Based Retrieval for Clinical Decision Support*, B. Caputo, H. Müller, T. Syeda-Mahmood, J. Duncan, F. Wang, and J. Kalpathy-Cramer, Eds. Berlin/Heidelberg: Springer, 2010, vol. 5853, Lecture Notes in Computer Science, pp. 18–30.

[4] U. Avni, J. Goldberger, and H. Greenspan, "Dense simple features for fast and accurate medical X-ray annotation," in *10th Workshop of the Cross-Language Evaluation Forum (CLEF 2009), LNCS*. New York: Springer, 2010, Lecture Notes in Computer Science.

[5] U. Avni, J. Goldberger, M. Sharon, E. Konen, and H. Greenspan, "Chest X-ray characterization: From the organ identification to the pathology categorization," presented at the 11th ACM SIGMM International Conference on Multimedia Information Retrieval (MIR-2010), Philadelphia, PA, Mar. 29–31, 2010.

[6] U. Avni, H. Greenspan, M. Sharon, E. Konen, and J. Goldberger, "X-ray image categorization and retrieval using patch-based visual words representation," in *ISBI'09: Proc. 6th IEEE Int. Conf. Symp. Biomed. Imag.*, Piscataway, NJ, 2009, pp. 350–353, IEEE Press.

[7] A. Barla, F. Odone, and A. Verri, "Histogram intersection kernel for image classification," in *Proc. ICIP*, 2003, vol. 3.

[8] A. Bosch, X. Muñoz, A. Oliver, and J. Martí, "Modeling and classifying breast tissue density in mammograms," in *Proc. CVPR*, 2006, pp. 1552–1558.

[9] W. W. Chu, A. F. Cárdenas, and R. K. Taira, "Kmed: A knowledge-based multimedia medical distributed database system," *Inf. Syst.*, vol. 20, no. 2, pp. 75–96, 1995.

[10] M. de Berg, M. van Kreveld, M. Overmars, and O. Schwarzkopf, *Computational Geometry (Algorithms and Applications)*. New York: Springer, 1998.

[11] A. Depeursinge, A. Vargas, A. Platon, A. Geissbühler, P. Poletti, and H. Müller, "3D case-based retrieval for interstitial lung diseases," *MCBR-CDS*, pp. 39–48, 2009.

[12] T. Deselaers, A. Hegerath, D. Keysers, and H. Ney, "Sparse patch-histograms for object classification in cluttered images," in *DAGM Symp.*, 2006, pp. 202–211.

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

[14] E. El-Kwae, H. Xu, and M. Kabuka, “Content-based retrieval in picture archiving and communication systems,” *J. Digital Imag.*, vol. 13, no. 2, pp. 70–81, Feb. 2000.

[15] I. El-Naqa, Y. Yongyi, N. P. Galatsanos, R. M. Nishikawa, and M. N. Wernick, “A similarity learning approach to content-based image retrieval: Application to digital mammography,” *IEEE Trans. Med. Imag.*, vol. 23, no. 10, pp. 1233–1244, Oct. 2004.

[16] Abe *et al.*, “Computer-aided diagnosis in chest radiography: Results of large-scale observer tests at the 1996–2001 RSNA scientific assemblies1,” *Radiographics*, vol. 23, no. 1, pp. 255–265, 2003.

[17] H. Müller *et al.*, “Overview of the ImageCLEFmed 2007 medical retrieval and medical annotation tasks,” in *CLEF*, 2007, pp. 472–491.

[18] H. Müller *et al.*, “Overview of the ImageCLEFmed 2008 medical image retrieval task,” in *CLEF*, 2008, pp. 512–522.

[19] T. M. Lehmann *et al.*, “The IRMA code for unique classification of medical images,” in *Proc. SPIE*, 2003, pp. 109–117.

[20] T. M. Lehmann *et al.*, “Content-based image retrieval in medical applications,” *Methods Inf. Medicine*, vol. 43, no. 4, pp. 354–361, Oct. 2004.

[21] T. Tommasi *et al.*, “Overview of the CLEF 2009 medical image annotation track,” CLEF Working Notes 2009 [Online]. Available: http://www.clef-campaign.org/2009/working_notes

[22] L. Fei-Fei and P. Perona, “A Bayesian hierarchical model for learning natural scene categories,” in *Proc. CVPR*, 2005, vol. 2, pp. 524–531.

[23] H. Greenspan, J. Goldberger, and L. Ridel, “A continuous probabilistic framework for image matching,” *Comput. Vision Image Understand.*, vol. 84, no. 3, pp. 384–406, 2001.

[24] H. Greenspan and A. T. Pinhas, “Medical image categorization and retrieval for pacs using the gmm-kl framework,” *IEEE Trans. Inf. Technol. Biomed.*, vol. 11, no. 2, pp. 190–202, Mar. 2007.

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

[26] J. A. Hanley and B. J. Mcneil, “The meaning and use of the area under a receiver operating characteristic ROC curve,” *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.

[27] C. E. Kahn and C. Thao, “Goldminer: A radiology image search engine,” *AJR Am. J. Roentgenol.*, vol. 188, no. 6, pp. 1475–1478, Jun. 2007.

[28] W. Khaliq, C. J. Blakeley, S. Maheshwaran, K. Hashemi1, and P. Redman, “Comparison of a pacs workstation with laser hard copies for detecting scaphoid fractures in the emergency department,” *J. Digital Imag.*, vol. 23, no. 1, pp. 100–103, 2008.

[29] P. Korn, N. Sidiropoulos, C. Faloutsos, E. Siegel, and Z. Protopapas, “Fast and effective retrieval of medical tumor shapes,” *IEEE Trans. Knowl. Data Eng.*, vol. 10, no. 6, pp. 889–904, Nov./Dec. 1998.

[30] C. Lebozec, M. Jaulent, E. Zapletal, and P. Degoulet, “Unified modeling language and design of a case-based retrieval system in medical imaging,” in *Proc. Annu. Symp. Am. Soc. Med. Informat. (AMIA)*, 1998, pp. 887–891.

[31] L. R. Long, S. Antani, D.-J. Lee, D. M. Krainak, and G. R. Thoma, “Biomedical information from a national collection of spine X-rays: Film to content-based retrieval,” in *Soc. Photo-Optical Instrumentat. Eng. (SPIE) Conf. Series, Soc. Photo-Optical Instrumentat. Eng. (SPIE) Conf. Series*, May 2003, vol. 5033, pp. 70–84.

[32] D. G. Lowe, “Object recognition from local scale-invariant features,” *Proc. ICCV*, vol. 2, pp. 1150–1157, 1999, vol. 2.

[33] C. D. Manning, P. Raghavan, and H. Schütze*, Introduction to Information Retrieval*, 1st ed. Cambridge, U.K.: Cambridge Univ. Press, Jul. 2008.

[34] H. Müller, C. Lovis, and A. Geissbuhler, “The medGIFT project on medical image retrieval,” *Med. Imag. Telemedicine*, 2005.

[35] H. Müller, N. Michoux, D. Bandon, and A. Geissbühler, “A review of content-based image retrieval systems in medical applications—Clinical benefits and future directions,” *I. J. Med. Informat.*, vol. 73, no. 1, pp. 1–23, 2004.

[36] H. Müller, X. Zhou, A. Depeursinge, M. J. Pitkänen, J. Iavindrasana, and A. Geissbühler, “Medical visual information retrieval: State of the art and challenges ahead,” in *Proc. ICME*, 2007, pp. 683–686.

[37] E. Nowak, F. Jurie, and B. Triggs, “Sampling strategies for bag-of-features image classification,” in *Proc. ECCV*, 2006, pp. 490–503.

[38] E. Osuna, R. Freund, and F. Girosi, Support vector machines: Training and applications Massachusetts Inst. Technol., Cambridge, MA, 1997, Tech. Rep..

[39] C. Shyu, C. Brodley, A. Kak, A. Kosaka, A. Aisen, and L. Broderick, “ASSERT: A physician-in-the-loop content-based retrieval system for HRCT image databases,” *Comput. Vis. Image Understand.*, pp. 111–132, 1999.

[40] J. Sivic and A. Zisserman, “Video google: A text retrieval approach to object matching in videos,” in *Proc. ICCV*, Apr. 2008, vol. 2, pp. 1470–1477.

[41] T. Tommasi, F. Orabona, and B. Caputo, “Discriminative cue integration for medical image annotation,” *Pattern Recognit. Lett.*, vol. 29, no. 15, pp. 1996–2002, 2008.

[42] B. van Ginneken, L. Hogeweg, and M. Prokop, “Computer-aided diagnosis in chest radiography: Beyond nodules,” *Eur. J. Radiol.*, vol. 72, no. 2, pp. 226–230, 2009, Digital Radiography.

[43] M. Varma and A. Zisserman, “Texture classification: Are filter banks necessary?,” in *Proc. CVPR*, 2003, vol. 2, pp. 691–698.

[44] I. Vergara, T. Norambuena, E. Ferrada, A. Slater, and F. Melo, “Star: A simple tool for the statistical comparison of ROC curves,” *BMC Bioinformatics*, vol. 9, no. 1, p. 265, 2008.

[45] A. Winter and R. Haux, “A three-level graph-based model for the management of hospital information systems,” *MethodsInfMed.*, vol. 34, pp. 378–396, 1995.

[46] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid, “Local features and kernels for classification of texture and object categories: A comprehensive study,” *Int. J. Comput. Vis.*, vol. 73, no. 2, pp. 213–238, 2007.