Grant Agreement Number: 257528

KHRESMOI

www.khresmoi.eu

Report on results of the WP2 first evaluation phase

Deliverable number	*D2.3*
Dissemination level	*Public*
Delivery data	*due 31.8.2012*
Status	*Final*
Authors	*Georg Langs, Joachim Ofner, Andreas Burner, René Donner, Henning Müller, Adrien Depeursinge, Dimitrios Markonis, Célia Boyer, Alexandre Masselot, Nolan Lawson*

This project is supported by the European Commission under the Information and Communication Technologies (ICT) Theme of the 7th Framework Programme for Research and Technological Development.

Executive Summary

This deliverable reports the results of the first evaluation phase of KHESMOI WP2 *Large Scale Biomedical Image Mining and Search*. The objective of the workpackage is to provide methods, algorithms, and tools to perform fast and accurate content-based image retrieval within the KHRESMOI prototype. The fundamental building blocks of the functionality required in KHRESMOI are descriptions of image content, and its context, that provide specificity, and generalizability sufficient for accurate retrieval. In this report we evaluate the core components of image retrieval in KHRESMOI. First we discuss the performance of the retrieval system for anatomical structures, and pathologies, that are the two most relevant functionalities in clinical radiology. We provide a detailed assessment of the image features used, based on a classification task, that illustrates the necessity of a feature extractor bank as opposed to a single *best* feature extractor. In the complementary second line of development, two components primarely relevant in the retrieval of publications, documents, and web-pages, are evaluated with regard to their ability to retrieve images based on image information, or based on image content, and contextual text information. At this point we have developed working initial prototypes, and understand strengths of the existing status, as well as limitations that will be the focus of coming research within the project.

Table of Contents

List of Figures

Notation

$\mathbf{I}_i$	Image or volume with index i. If it is 2D or 3D data will become clear from the context.
$\mathbf{I}_i \in \mathbb{R}^2$	2D data such as images.
$\mathbf{I}_i \in \mathbb{R}^3$	3D data such as volumes.
$\mathbf{I}_i(x)$	Value of image of volume at position x
$\mathbf{f}(x)$	Feature (vector) extracted at position x
$\mathbf{d}(\mathbf{I})$	Image descriptor (vector) for an entire image/volume

Abbreviations

LBP	Local Binary Patterns
PACS	Picture archiving and communication system
ImageCLEF	Image retrieval task in the Cross Language Evaluation Forum
GLCM	Gray Level Coocurrence Matrix
SIFT	Scale-Invariant Feature Transform
DoG	Difference of Gaussians
EMEA	European Medicines Agencey
Europarl	Europarl: A Parallel Corpus for Statistical Machine Translation
MeSH	Medical Subject Headings
BoC	Bags of Colors
BoVW	Bags of Visual Words
DICOM	Digital Imaging and Communications in Medicine
SVM	Support Vector Machine
CT	Computed Tomography
HRCT	High Resolution Computed Tomography
MRI	Magnetic Resonance (Imaging)
ROI	Region Of Interest

1 Introduction

Various scenarios in medicine are characterized by the availability of rich image content, but limited textual information. Examples are the reading of medical imaging data in clinical radiology, where the goal is to derive diagnostically relevant information from imaging data, or the search for comparable images in literature, where query features might be present in the images, but not in the immediately surrounding text. Even if textual information is present and usable, images might give additional *orthogonal* cues for search that can increase specificity of search results.

As a consequence, content-based image retrieval is a central functionality in KHRESMOI. Research is focussed on the identification of usable features that can serve as basis for effective retrieval. In contrast to many general computer vision or recognition tasks, where the goal is the identification of objects, medical imaging necessitates the identification, and matching of diagnostically relevant visual aspects, that are characterized by often subtle deviations of appearance in gray scale or texture.

In this report we evaluate methods for the retrieval of medical imaging data. We focus on the core aspect of image description, and the capability of different approaches in light of several specific tasks. These tasks are designed to cover scenarios relevant to the two usecases of KHRESMOI: medical information retrieval for (1) citizens and medical professionals, and (2) for clinical radiologists. The tasks are as follows:

- **retrieval of anatomical structures** based on a query image or volume;
- **retrieval of cases with similar pathology** based on a query image or volume together with a region of interest, that is indicated by the user;
- **pathology classification** based on local appearance;
- **retrieval of two-dimensional images** in web- and document search;
- **retrieval based on text information** associated with images.

The data on which the evaluation is based covers a set of relevant applications, and provides realistic examples ranging from imaging data typically present in a hospital PACS, to publication data bases, and web pages.

The key insights of the work so far are that (1) content-based image retrieval is feasible, it is central in clinical radiology and contributes significantly to retrieval where both text- and image information is available. (2) Due to the large range of appearance characteristics, and the often subtle differences that are crucial in identifying diagnostically relevant data, domain specific image features are necessary. Thus, algorithms to learn feature sets and descriptors from training data are central to image description. Instead of a priori decisions for specific *allegedly* optimal features we have to provide feature banks and methods to choose and compose optimal image descriptors in an algorithmic fashion. (3) We conclude that much work has yet to be done, to provide for performant descriptors, and effective indices, in the context of minimal- or absent annotation during the indexing phase. The core challenge of the work in the next project phase is the development of methods for the weakly- and un-supervised learning of structure in massive imaging data, based on the tools described in deliverable D2.2 and evaluated in this report.

2 Evaluation tasks

We evaluate the image analysis algorithm in the context of 5 tasks, to answer 5 specific questions, and to provide insight into specific behavior of the used methodology with regard to corresponding objectives.

1. **Anatomy retrieval** Is the retrieval of imaging data that shows corresponding anatomical regions based on a query image or volume feasible? What is the accuracy of the retrieval?

2. **Pathology retrieval** Is the retrieval of cases that exhibit the same and potentially local pathology based on a query case and an indicated region of interest possible? What is the accuracy?

3. **Pathology classification** How do local features, and image descriptors compare if the goal is to differentiate among pathologies in one specific anatomical region? Does one feature exist, or are different aspects of pathologies represented by multiple feature extractors? Can — and should — feature extractors be learned from the data to improve specificity?

4. **Two dimensional Image retrieval** Is two-dimensional medical image retrieval feasible? What is the accuracy of retrieval? Does color information improve medical image retrieval, and classification performance?

5. **Combined text and image based retrieval** What is the contribution of image- and textual information when both are available and used for retrieval? Do images contain information that can enhance image- and case-based retrieval performance using keywords?

6. **Text based image retrieval** Is image retrieval based on textual information in the vicinity of images feasible and accurate?

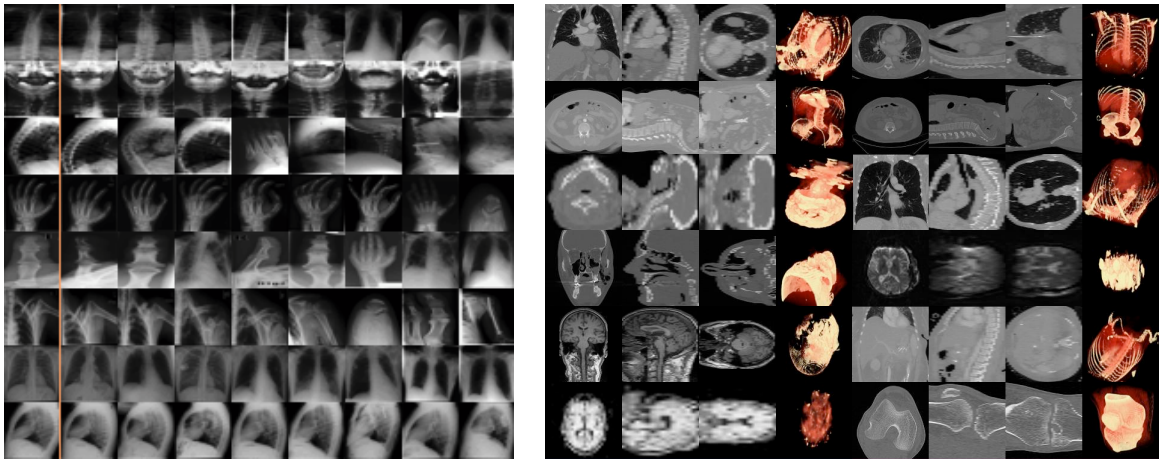
3 Evaluation of 2D/3D anatomy retrieval

The automatic localization of the anatomical region depicted in a certain image is highly important for working with large scale medical image data. It serves as a first step in retrieval systems and allows to invoke analysis stages that are tailored to the anatomical region in question.

The following approach was developed to satisfy these needs within the Khresmoi project. It operates on the large scale data set aquired in the project as well as a publicly available benchmarking data set and was published in [6].

3.1 Data

The proposed method was developed without restrictions on the type of image data used. The two data sets used for evaluation are the ImageCLEF benchmark dataset of radiographs and the the large scale set of CTs and MRs aquired from the PACS of the General Hospital Vienna / MUW. The data sets exhibit very different image characteristics and as such form a good testbed for an in depth evaluation.



(a) (b)

Figure 1: Data used for anatomy retrieval experiments: (a) The imageCLEF examples that are radiographs of different body regions, (b) the 3D MR and CT data collected at MUW.

ImageCLEF data The ImageCLEF 2009 classification challenge[1] provides a set of 12.677 (training) plus 1.733 (test) 2D radiographs of various body regions. Example images from this data set can be seen in Fig. 1a. All images come labeled with the IRMA code [16], which is a hierarchic multi-dimensional code providing information about modality, body orientation and anatomy.

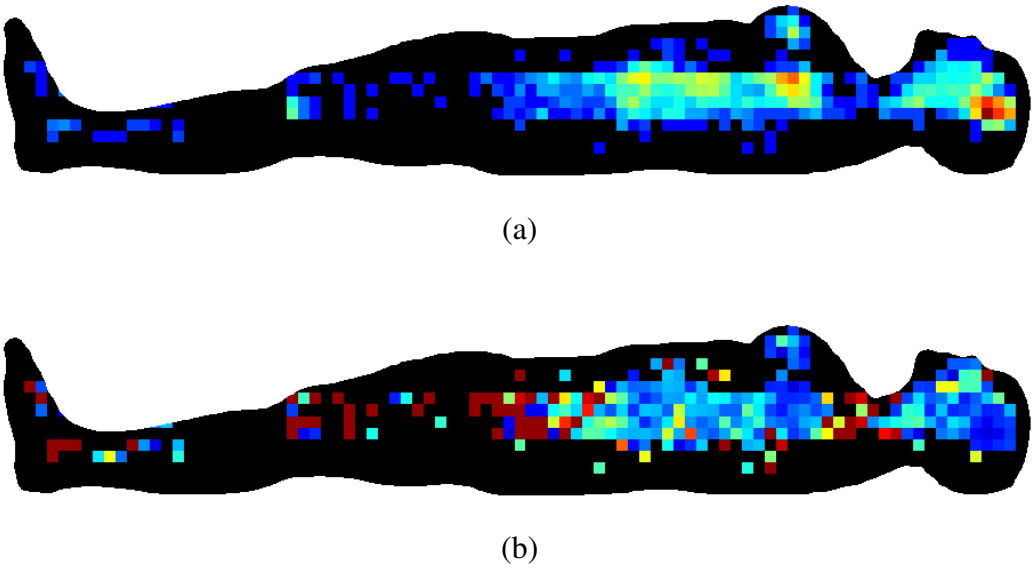
Clinical CT and MRI data This data set is a collection of 3876 3D-CTs and MRs extracted from the PACS of the General Hospital Vienna / MUW, obtained by querying for data sets acquired during a short period in 2011. The CTs and MRs originate from all different scanners present at the department of radiology. For each of the DICOMs the 3D position of its center in a full body atlas was annotated for evaluation purposes. An overview of the distribution of the centers of the DICOMs is depicted in Fig. 2. As can be seen the vast majority of acquisitions are performed with standardized protocols, for specific regions where performing CTs or MRs is of interest. This means that the image space of these regions is highly populated, whereas there are only 7 CTs of feet in the entire data set. A set of samples of the data set is depicted in Fig. 1b.

3.2 Evaluation set-up

3.2.1 Validation of retrieval on ImageCLEF Data Set

The task is to annotate each query image with an IRMA code. This assigns each test image not only to an anatomical region, but also indicates technical and pathological properties of the image. A kNN search with $k = 3$ was used to find the most similar training examples for each query image. The most often occurring code (or the one of the closest match in parity situations) was

[1]The ImageCLEFmed 2009 Data Set: http://www.irma-project.org/datasets.php?selected=0000900009.dataset



(a)

(b)

Figure 2: a) Sagittal projection of the distribution of the volumes in the 3D data set. For each volume the position of its center in a 3D atlas was annotated as ground truth. b) Mean error between predicted and actual position of the individual volumes using the HOG descriptor, with blue corresponding to 0cm error and red to ≥ 30cm. Note how the accuracy of the prediction strongly correlates with the number of examples available in the database for each location.

assigned to each query image. While the benchmark's rules allow to specify wildcards within the result codes this option was not used. Using the provided benchmark script[2] a cumulative error score for each proposed approach (miniature alignment as well as euclidean distance alone, Distribution Fields, Histograms of Gradients) was computed. To provide additional context a standard DOG-SIFT and a dense-SIFT implementation as detailed in [9] were included in the comparison.

3.2.2 Validation of retrieval on clinical CT and MRI Data Set

For a given query image, the task was to predict its position in the reference atlas. The coordinates of the most similar volume (kNN with $k = 1$) were assigned to the each query volume and the euclidean distance in centimeters between resulting position and ground truth were used as error measure. To exclude the influence of misclassification due to left/right similarities (i. e. acquisitions of single hands and feet) all annotations and subsequent evaluations were performed using absolute distances to the sagittal plane for the coordinate orthogonal to the sagittal plane.

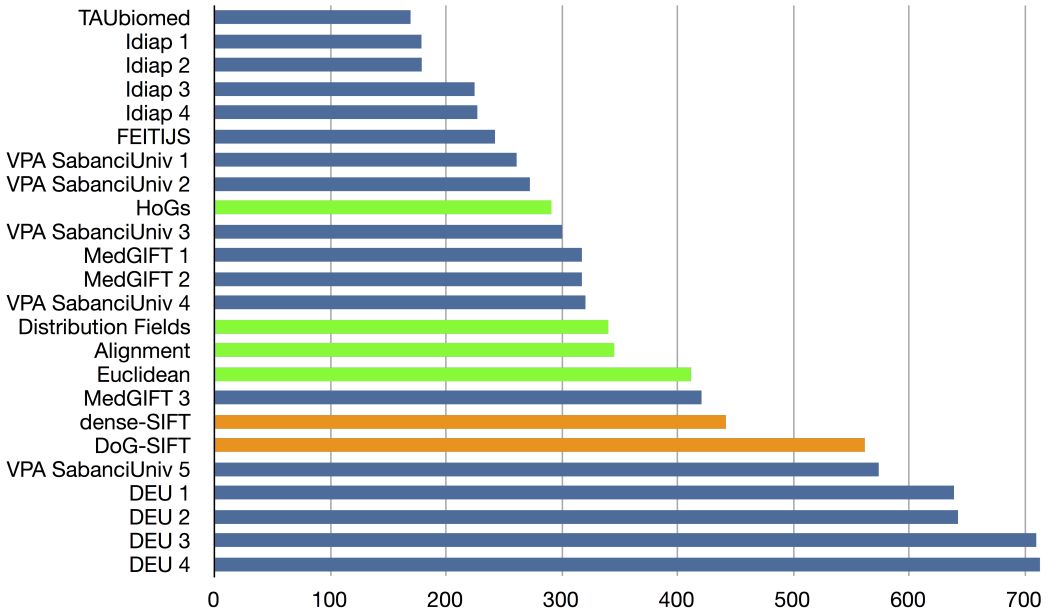


Figure 3: Error scores for the ImageCLEF 2009 Medical Image Annotation Task [27]. The published benchmark results are depicted in blue, the DOG-SIFT and dense-SIFT approaches in orange and the the methods investigated in this work in green. The error score is based on the evaluation scheme for the IRMA code classification set (Version *2008*), with a lower score indicating higher accuracy. Despite its simplicity, the HOG miniatures rank comparatively high, surpassing standard and more optimized BVWs approaches.

3.3 Results & Discussion

In the following the performance of the proposed approach on the two data sets is detailed.

3.3.1 ImageCLEF Data Set

Fig. 3 shows the performance comparison of the evaluated methods for the ImageCLEF 2009 data set. The published benchmark results are depicted in blue, the implementations of standard DOG-SIFT and dense-SIFT in orange and the methods proposed in this paper in green. The low performance of the standard BVW approaches in comparison to the top result, which also employ BVWs, shows the importance of the careful and data set dependent optimization of the interest point detector, the local descriptors and their integration into visual words. The performance of the published methods (blue) has been carefully optimized on this data set in the last years.

Also of note is, that the reference BVWs (orange) only employ a simple $k = 1$ NN using the χ^2-distance as opposed to more complex classifiers such as hierarchical kernel-SVMs. Also in relation to the miniature based approaches evaluated in this work it has to be kept in mind, that while using elaborated classification schemes increases the classification performance, the ability to yield a ranked list of the most similar training examples gets lost. This would have

[2] http://www.idiap.ch/clef2009/evaluation_tools/error_evaluation.pdf

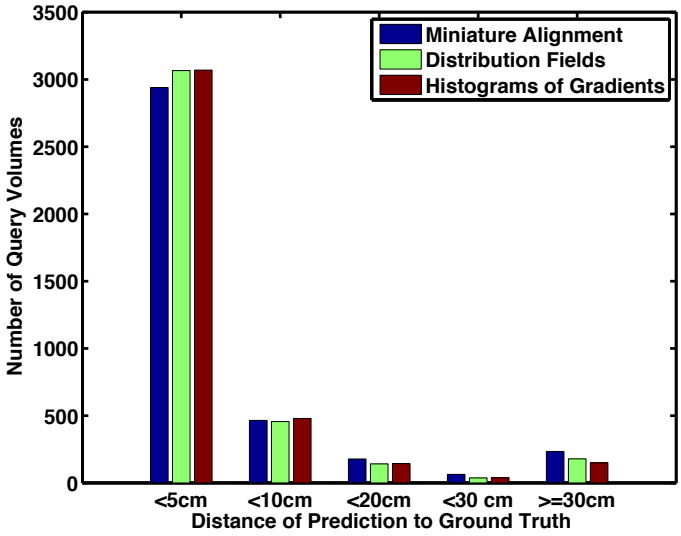


Figure 4: Histogram of the residual distances in centimeters for the 3D CT dataset. In a leave-one-out setting, the distance of the predicted position to the ground truth is recorded for each query volume.

to be performed, after obtaining the classification result, using the original features (often a combination of several different feature types), using a distance metric which has to be carefully selected.

Some of the image miniature descriptors, without optimizations for this particular data set, perform surprisingly well, given their simple construction and straightforward classification strategy. Retrieving the most similar miniatures using euclidean distance yields the worst results, and rigidly aligning the closest candidates from the *k*D-tree according to [28] improves the results only slightly. Almost the same performance is achieved using the distribution fields, which mimic the alignment process through the Gaussian smoothing of the individual bins.

Showing the best performance on the 2D data set, and surpassing the established medGIFT framework as well as five BVW approaches is the SIFT-like HOG descriptor, which deals best with the varying contrast and brightness as well as occlusions (implants) and spatial transformations. While it does not outperform more complex and considerably more optimized approaches, it provides an interesting baseline as to how far a simple model can yields useful results with extremely low computational complexity.

3.3.2 CT Data Set

For the 3D localization task we measured the distance in centimeters between the center of the query image and the center of the most similar image. Fig. 4 shows the histogram of the resulting distances when querying with all images in a leave-one-out fashion. 79.17% of the results for the HOG descriptor are within 5cm of the ground truth, 91.47% are within 10cm. The median residual for the HOG approach is 2.40cm, with the distribution fields and the miniature alignment performing similar at 2.42cm and 2.56cm. Due to the presence of outliers the mean distances are considerably larger: 5.72cm, 6.70cm and 7.53cm, respectively.

The source for the outliers becomes clear when looking at the patterns of their occurrence. The lower part of Fig. 2 shows the spatial distribution of the mean prediction error for the HOG descriptors. It exhibits a strong anti-correlation between density, i. e. number of images per region, in the data set, and prediction accuracy, with blue areas indicating an error of 0 and red representing $\geq 30cm$. Looking at the lower extremities, for which only few example are available in the data set, the results consist mainly of outliers. On the other hand, in the abdominal, thorax and head region, the most common areas for which CT is performed, localization accuracy is high.

We see this result as confirmation of our hypothesis, that given the highly constrained image space associated with medical images in clinical practice, the use of simple models estimated with the help of large data sets yields promising results and warrants closer investigation. Considering that the results presented in this work are based on only a few days worth of CT acquisitions, we expect the results to significantly improve in areas of lower density by expanding the dataset. We expect a data set with an even spatial distribution, weighted by the expected anatomical and pathological variance, to perform well while limiting the size of the data set.

4 Evaluation of 3D pathology retrieval

This section was partly published in [2]. Pathology retrieval based on imaging data is relevant in a clinical radiology environment. Based on a query case, for which imaging data is available and a manually indicated region of interest containing a particular structure or texture for which comparable examples are needed, similar cases are retrieved. In this section we present initial results for image retrieval on a pilot data set with different lung pathologies, collected from a hospital PACS. Retrieval is based on *texture bags* local image descriptors [2].

4.1 Data

The evaluation data contains 21 HRCT image series with a slice thickness of 3mm and an in-plane pixel spacing of 0.74mm. The cases of this study consist of 10 cases of lungs diagnosed as healthy and cases with two types of lung pathologies: emphysema (6 cases), lung metastasis (4 cases) and both (1 case). The diagnosis of each case was confirmed by two experienced radiologists of the contributing hospital. We focus on the retrieval of lungs suffering emphysema. A radiologist markes a query region in the query case. Query cases were emphysema cases. The system retrieves most similar regions in the remaining data set and ranks them based on an appearance distance [2].

As a basis for our tests, the contributing radiologists manually marked query regions in each lung where they detected patterns of emphysema. Typically these marked regions are small patches on three to five slices for each image series (see left images of figure 5).

The segmentation of the lungs are performed semi-automatically, by a threshold algorithm, applying simple morphologic functions and a manual validation and correction step. Lung segmentation is not in the scope of this paper. Instead we focus on the characterization of anomalies within the anatomical structure.

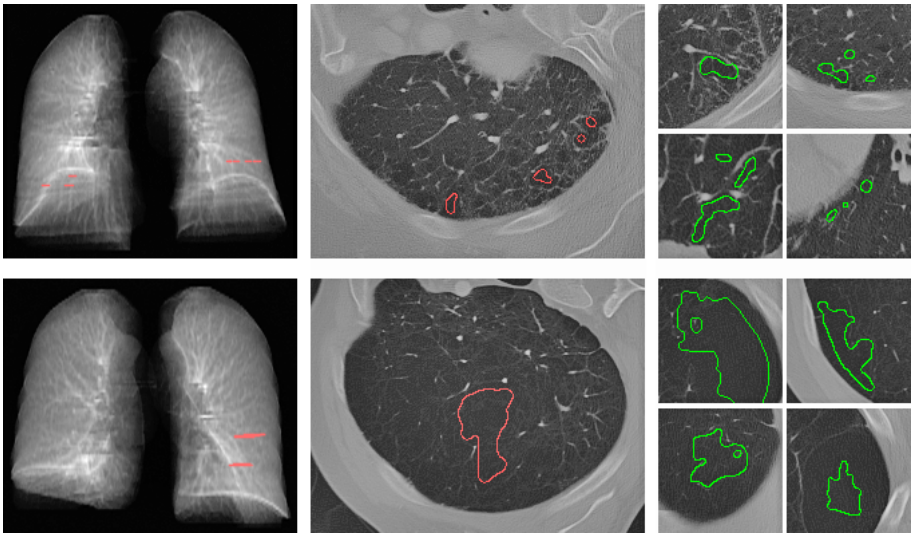


Figure 5: Retrieval ranking result of two distinct emphysemas with different tissue patterns (top: centrilobular emphysema, bottom: panlobular emphysema). The region highlighted in red on the left side shows the query region $\mathbf{R}_Q$ marked during search by a physician. On the right side, the green regions depict the four most similar regions $\mathbf{R}_{js}$ retrieved by our method.

4.2 Evaluation set-up

At the moment the pre-processing time to calculate three-dimensional, LBP-based descriptors **D** for all voxels in a volume of 512x512x150 voxels is less than one second on a quad core computer.

We chose fixed parameters for all runs: with a number of texture bag clusters of $k = 300$ and diffusion distance to compare across cluster histograms. The diffusion distance results in some degree of robustness regarding increasing k and ambiguity with regard to cluster membership of individual voxels. The performance is to some extent dependent on the precomputed oversegmentation of the volumes used for texture bag calculation. Therefore, the supervoxel algorithm is of importance: (1) where it computes the borders between regions, and (2) the number of supervoxels s per volume, which should be chosen dependent on the granularity of the pathology to be retrieved. After initial experiments, we fixed $s = 5000$ for all cases. For the descriptor **D** we chose the weights c_c and c_i to be 10 and use scale 1 to 4 for the similarity calculation of regions. Furthermore, we set the ranking threshold to 800.

We performed a 7-fold cross validation to evaluate the retrieval performance of the algorithm. In each run, a *query region* $\mathbf{R}_Q \subset \mathbf{I}_Q$ was marked in one of the emphysema cases and retrieval was performed on all other cases $\mathbf{I}_j, j \neq Q$. We validate the ranking by evaluating the ratio of cases with corresponding anomaly (emphysema) among the top ranked retrieved cases.

To validate the concept of *texture bag* ranking based on a local query region, the recognition rate of our method (table 2) was compared to a retrieval run based on the distance of *texture word* histograms of the entire query image $h(\mathbf{I}_Q)$ to texture word histograms of entire lung volumes of the data set $h(\mathbf{I}_j), j \neq Q$.

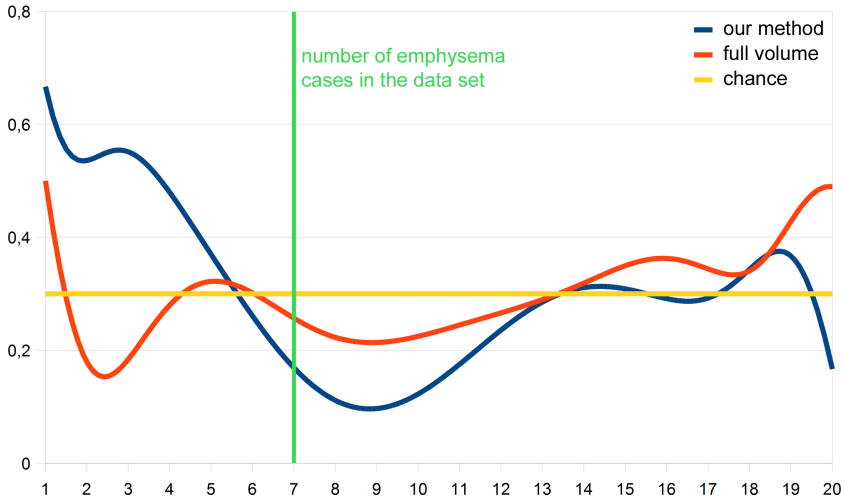


Figure 6: Comparison of retrieval results. Curves present the density of correct (emphysema) cases among the ranking indicated on the x-axis. Ideally there would be a step function with all emphysema cases ranked better than the cut-off line (green): (yellow) ratio of randomly picked image series is 30%. Local query regions based retrieval (blue) and full volume retrieval (red).

Table 1: Full image ranking: the result of seven runs (r1 to r7): the table shows the number of correctly retrieved pathologies by full volume histogram retrieval of the top 3, 5 and 7 image series.

ranking	r1	r2	r3	r4	r5	r6	r7	average
top 3	1	0	0	1	1	1	1	24%
top 5	2	0	1	2	2	2	1	29%
top 7	2	1	1	3	2	3	1	27%

Table 2: Texture bag ranking: the result of seven runs (r1 to r7): the table shows the number of correctly retrieved pathologies by our retrieval method of the top 3, 5 and 7 image series.

ranking	r1	r2	r3	r4	r5	r6	r7	average
top 3	2	2	2	2	2	1	3	67%
top 5	2	3	4	3	2	2	3	54%
top 7	2	3	4	4	3	3	3	45%

4.3 Results

Table 1 shows anomaly retrieval results for global image similarity, table 2 demonstrates the retrieval results by the proposed method that relies on localized similarity based on a query region. The ratio of correctly top ranked cases (i.e., cases with the correct query anomaly) for randomly picked image series is 30% (6 of 20). The full volume histogram retrieval's performance is in the range of 24%-29%. The likely reason is that the anomaly regions that

are relevant for retrieval often cover only a fraction of the overall volume. Therefore, the most frequent tissue types are typically not anomalies and thus not helpful during retrieval. The average recognition rate of correctly retrieved pathologies of all runs for the proposed method is 67% for the top rated three, 54% for the top rated five and 45% for the top rated seven image series. Figure 6 plots the percentage of correctly retrieved anomalies against the number of image series returned. Interestingly, there is a peak of the retrieval rate at the end. Our data shows that this peak is caused by two distinct types of emphysema that have two distinct texture characteristics.

4.4 Discussion

Results indicate that pathology retrieval is feasible. They highlight the importance of specific query regions, that capture a pathology of interest. Due to the local nature of many pathological imaging patterns - in the present case only part of the lung exhibits anomaly - retrieval based on the entire image is not feasible. Instead algorithms have to (1) use information regarding the anatomical structure to learn a corresponding feature vocabulary, (2) take local query regions into account, when ranking indexed cases, and (3) indicate not only the query result case, but also identify the region that are matches to the query region, to allow radiologists to evaluate and use the result.

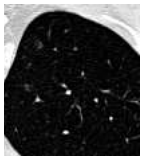
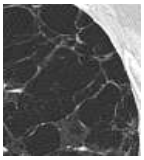
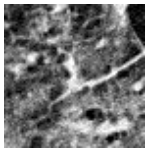
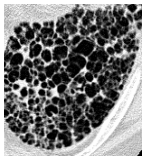
5 Evaluation of 3D pathology classification

In this section we provide an in-depth evaluation of local image features, and corresponding descriptors of image regions. The evaluation task is the classification of lung tissue into one of 5 different tissue categories: *healthy tissue, emphysema, ground glass, fibrosis, and micro-nodules*. We evaluate two related strategies: (1) the voxel-wise classification of volume data, and (2) the classification of areas in the volume data. The latter extends the descriptor range to vocabulary based descriptors, that learn statistic descriptors for the appearance of sets of voxels. This is particularly relevant to capture subtle texture characteristics.

5.1 Data

Part of this was published in [5]. A multimedia collection of cases with interstitial lung diseases (ILDs) created at the University Hospitals of Geneva was used for the evaluation. The dataset contains high–resolution computed tomography (HRCT) image series with three–dimensional annotated regions of pathological lung tissue along with clinical parameters from patients with pathologically proven diagnoses of ILDs. The library contains 128 patients affected with one of 13 histological diagnoses of ILDs, 108 image series with more than 41 liters of annotated lung tissue patterns as well as a comprehensive set of 99 clinical parameters related to ILDs [5]. A subset of 85 ILD cases with annotated HRCT images is used to evaluate our approach. Expert annotations were carried out in collaboration by two radiologists with 15 and 20 years of experience in CT imaging. The slice thickness is 1mm and the inter–slice distance 10mm. The images were acquired with two imaging devices at the Radiology Service of the University Hospitals of Geneva: a Philips Mx8000 IDT 16 CT Scanner and a General Electric HiSpeed CT. The

Table 3: Visual aspect and distribution of ROIs per class of lung tissue pattern. Note that a patient may have several types of lung disorders.

visual aspect					
tissue type	healthy	emphysema	ground glass	fibrosis	micronodules
hand–drawn ROIs	150	101	427	473	297
patients	7	6	32	37	16

five lung tissue classes encountered in most ILDs were chosen as lung texture classes: healthy, emphysema, ground glass, fibrosis and micronodules. The visual appearance of the lung texture classes and their distribution are detailed in Table. 3.

5.2 Evaluation set-up

We compare two categories of features: (1) voxel-wise features, and (2) features that capture properties of areas encompassing multiple voxels, we will refer to the latter as *texture bags*. The objective is to classify 5 lung-tissue classes, among which 4 are pathologic.

The texture samples (voxels or areas) are classified based on the features extracted from the imaging data. We use a Random Forest classifier [1] with 150 trees, and classifiy all 5 classes at the same time (in contrast to multiple one-vs-rest binary classifications). Random Forest classifiers are relatively robust with regard to noisy features. Thus the classification performance is a good indicator for the information encoded in the features, without being distorted by less informative features that might also be present in the training set. To evaluate, we perform 10-fold cross validation. That is, during each iteration we use 90% of the image data for training of the classifier, and use the remaining 10% for testing of the classification accuracy. The number of samples of the texture classes are balanced in the training set.

The performance of the runs is measured by the overall accuracy which indicates the percentage of correct predictions (in our case chance level is at 20%). In addition we report confusion matrices for selected examples to illustrate the class specific classification accuracy, and its feature extractor dependent behavior.

5.3 Feature extractors compared

In this section the texture features and the used parameters for the evaluation are described. Features are either extracted on a *voxel level*, i.e., each voxel in a volume is assigned a corresponding feature vector, or on a *area-*, or *region level*, i.e., sets of voxels are assigned a joint feature vector.

5.3.1 Voxel features

We compare a number of texture feature extractors. They are organized in feature sets describing method, parameters and length of the resulting feature vector. As texture exists only for more than one voxel, each of them includes information of a local neighborhood to calculate texture features. In the following we list the feature evaluated:

Intensity Histogram The intensity histogram features describe the gray value distribution of a window around a voxel. Gray values outside the interval [-1024;600] are excluded. The local neighborhood is defined as a 33×33 pixel window around the voxel. A histogram with 22 bins is calculated and normalized by the number of voxels in the window.

Haralick Features One set of Haralick Features [12] is defined by offset distance, window size, gray value boundaries and the number of gray levels to which the gray levels are reduced to. Internally four matrices with fixed offset directions of $0°, 45°, 90°$ and $135°$ and tunable distance are calculated. Ordering of the pixel pairs does not matter, i.e. a matrix is symmetric accross its diagonal. 13 Haralick Features are calculated from the mean matrix of the four directed matrices. A single Haralick feature set is defined by window size and gray level reductions. Gray value boundaries are set to [-1024;600] for all sets. One set combines offset distances of 1 to 5 pixels resulting in $13 \times 5 = 65$ features per set. For evaluation 9 Haralick feature sets with permutation of window sizes of 11×11 pixels, 17×17 pixels and 33×33 pixels and gray level reductions of 8, 16 and 32 are calculated.

Gradient histograms Gratient histograms analogous to the descriptor used in Scale invariant feature transform (SIFT) [17]. Length of feature vector: 128

Riesz wavelets The Riesz transform is an extension of the Hilbert transform, which maps any function $\mathbf{I}(x)$ to its harmonic conjugate. In this case it yield a steerable filterbank [30] allowing to analyze textures in any direction. See deliverable D2.2 for details. Length of feature vector: 20

Riesz wavelets with intensity histograms As a variant we also explore a Riesz wavelets augmented with local intensity histogram. Length of feature vector: 20 + 22 = 42

Local Binary Patterns Local Binary Patterns [25] are calculated by sampling the gray values in the neighborhood of a pixel in a circle with radius r and n samplepoints. The sampled gray values are thresholded with the center pixel and encoded in a binary code of length n. In addition to this basic formulation of LBPs, all codes can be mapped to a subset of representative codes. Beside no mapping, the two mappings to rotation invariance and rotation invariant uniform patterns are investigated. In addition to the gray-scale invariant LBPs, local contrast is measured by local variance of the pixel values for every parameter pair (r,n). A LBP feature set used in this evaluation combines LBPs with radii 1,2 and 3 pixels with 8, 16 and 24 samplepoints respectively and contains the LBP and local variance measures resulting in $3 \times 2 = 6$ features per set. Three sets with no mapping (NOMAP), rotation invariance (RI) and rotation invariant uniform patterns (RIU2) are defined.

Texton filterbank This featureset consists of the filter responses of a filterbank with three parts [18]. The first part is the gaussian second derivative $fb_{even}(x,y) = G''_{\sigma_1}(y)G_{\sigma_2(x)}$. Where $G_{\sigma(x)}$ is a Gaussian with standard deviation of σ. The ratio $\sigma_2 : \sigma_1$ represents the elongation of the filter. The second part of the filterset is calculated by the Hilbert transform of

$fb_{even} : fb_{odd}(x,y) = Hilbert(fb_{even}(x,y)$. The first two parts are directed filters. They are calculated in 6 orientations and 3 scales ($\sigma_{start} = 1$, scale factor: $\sqrt{(2)}$). The elongation is set to 3. The third part is a Difference-of-Gaussians center surround filter. All filters are zero-mean and L_1 normalized. The length of the featurevector of this featureset is 39.

Feature selection In addition to the individual evaluations, we performed feature selection. Feature importance calculations were performed on a *feature selection testset*. It is the concatenation of following feature sets with the parameters described above:

- intensity histogram;
- Haralick features, Windowsize 11×11, reduced to 16 gray levels;
- Haralick features, Windowsize 17×17, reduced to 16 gray levels;
- Haralick features, Windowsize 33×33, reduced to 16 gray levels;
- LBPs, rotation-invariant mapping;
- Riesz wavelets;
- Texton filterbank.

This feature importance testset has a length of 282 features. 100.000 samples are randomly selected out of all voxels in the set. The Gini Importance [15] for all features is determinded by training a Random Forest with 600 trees. The features are ranked according to their Gini Importance. Then Feature Forward Selection (FFS) is performed by classifying voxels in the dataset with increasing amount of importance-ranked features. In every iteration of the FFS the amount of the class samples is balanced and the maximum of the samples is set to 30.000, by randomly sampling out of the whole set. The samples are classified with a 10-fold CV with Random Forests with 300 trees. In the first iteration 5% of the most important features are used. In every iteration 5% are added, until all 282 features are used in the last iteration.

5.3.2 Region features: texture bags

Texture bags [2] learn the structure of features present in a training data set, to obtain a vocabulary of features that are optimal for describing the variability present in a specific region. The features extracted from training data are quantized by performing clustering on the feature descriptor sets F_i of a subset of voxels randomly sampled from all annotated voxels. The k clusters that formulate the texture vocabulary are reffered to as texture words W_k. Each voxel is represented with its closest texture word $\mathbf{W}_k^s$, i.e., with the index of the closest cluster center. Features for an area are *texture bags*. A texture bag of an area $\mathbf{A}$ is the histogram $h(\mathbf{R})$ of *texture words* $\mathbf{W}_k$ it contains. This is analogous to the *bag of visual word* paradigm of Sivic et al. [22]. The histogram describes an area in terms of its textural structure. The texture histograms are normalized to be independent of the size of the area.

For the calculation of texture words kMeans is used as clustering algorithm. The maximum number of iterations was set to 1000. The size of the randomly sampled subset for clustering was fixed with 250000. The size of the feature vector is at the same time the number of histogram bins of the texture bag, or also referred to as number of texture words. Values of 100, 200 and 300 were compared for all calculated texture bags.

Table 4: Classification accuracy for voxel-wise classification of all voxel feature sets. The length of the feature vector is listed in the rightmost column. A graphical overview can be found in Figure 7.

Set ID	Feature Set	Overall Accuracy [%]	# of features
1	Intensity Histogram	61.44	22
2	LBP NOMAP	41.03	6
3	LBP RI	40.94	6
4	LBP RIU2	40.91	6
5	SIFT	30.05	128
6	Riesz Wavelets	57.62	20
7	Riesz and Intensity Hist.	66.66	42
8	Texton Filterbank	48.99	39
9	Haralick 11×11, 8 gls	59.06	65
10	Haralick 11×11, 16 gls	61.12	65
11	Haralick 11×11, 32 gls	60.59	65
12	Haralick 17×17, 8 gls	64.14	65
13	Haralick 17×17, 16 gls	66.36	65
14	Haralick 17×17, 32 gls	64.97	65
15	Haralick 33×33, 8 gls	68.81	65
16	Haralick 33×33, 16 gls	69.59	65
17	Haralick 33×33, 32 gls	68.81	65

For the evaluation, texture bags are calculated for three kinds of area sets. First, every annotated region of interest (ROI) of the dataset is considered an area for classification. The other two area sets are calculated by over-segmenting the image data with a super-pixel algorithm of Wildenauer et al. [31]. All resulting super-pixel areas that lie with over 75% inside an annotated ROI are included in the area sets. The Wildenauer algorithm allows to set the number of super-pixels that are produced per slice. By setting this number to 250 (larger areas) or 500 (smaller areas), two super-pixel area sets are created. To ensure that the histogram features are meaningful, i.e., that the area contains a sufficient amount of features to sample the histogram, all areas that are smaller than 300 pixels are excluded from the area sets.

Texture bags are calculated for all voxel feature sets except of SIFT. An additional feature set is defined by using 40% of the most important features of the feature selection test set.

5.4 Results

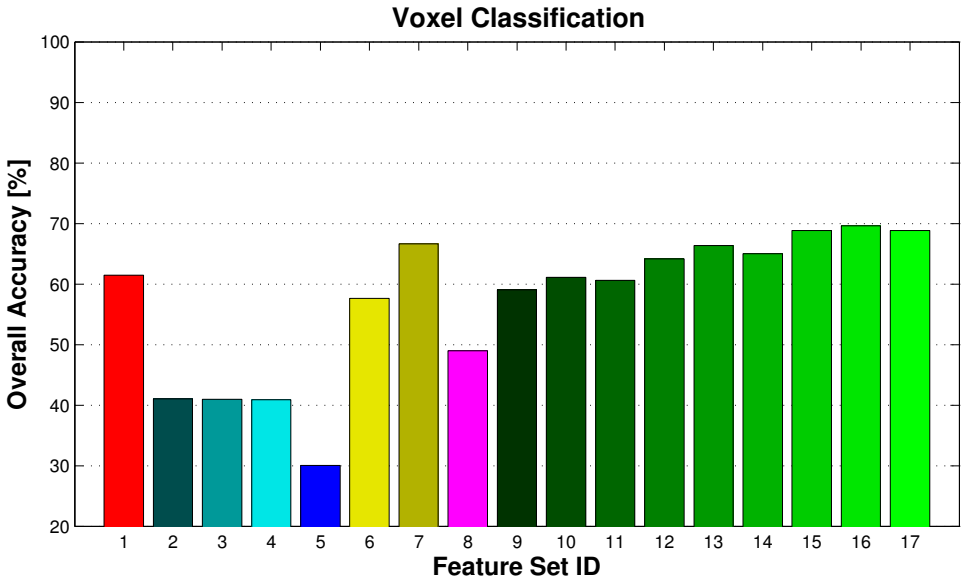


Figure 7: Classification accuracy for voxel-wise classification. Feature set IDs listed in table 4.

5.4.1 Voxel classification

In Table 4 we report the classification accuracy, and the length of the feature vector for 17 evaluate local feature extractors describing texture on a voxel-level. Figure 7 offers a plot of the voxel-wise classification accuracy.

Results achieved with gray-level co-ocurrence matrix descriptors show that in general features that draw information from a large neighborhood perform better (for example Haralick with windowsize 33×33) compared to descriptors, that capture only small neighorhoods. LBP features perform particularly poor. This is interesting since it is in stark contrast to the performance of texture bag descriptors (evaluated in the next section) based on LBP features (see Section 5.5 for a discussion).

Feature selection In addition to learning a model for prediction of class labels based on feature vectors, Random Forests estimate the actual multivariate information contributed by each feature. The so-called *Gini importance* is a score that ranks features corresponding to their contribution, or differentiating power with regard to the class labels. When combining all features used for voxel wise classification, we obtain a ranking as depicted in Figure 8. Classification accuracy when using the top-ranked X% of available features is depicted in Figure 9. The combined feature-set "feature selection testset" outperforms other individual feature sets already with a fraction of 30% top-ranked features. However, the relatively slow increase shows that information is still added, even if we include more the 50% of all featerus, indicating that the information is distributed across features. Details of the top- and bottom-ranked features are provided in table 5. Features with larger neighborhood and having intensity information are ranked higher. The ranking does not exclude redundant features, and the top-ranked features in both intensity histograms, and the following GLCM Haralick features capture relatively sim-

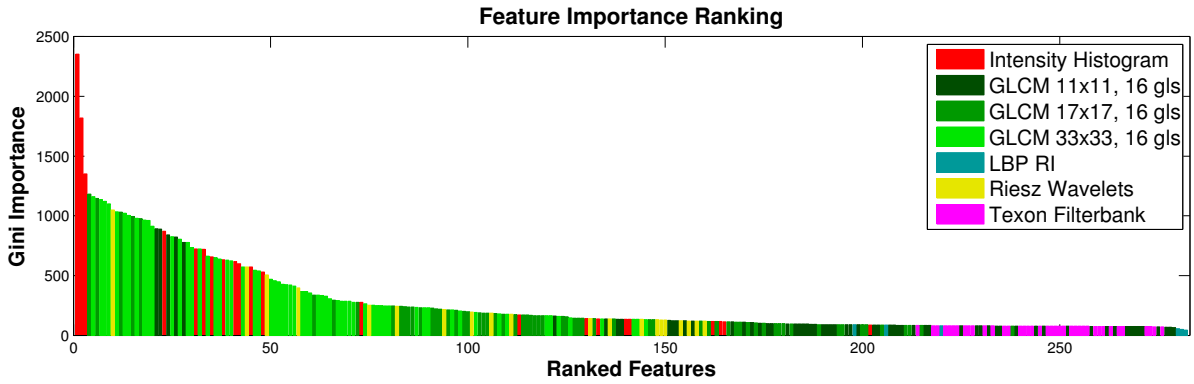


Figure 8: Featureset containing the features depicted in the legend ranked according to thier Gini importance. Details about the 10 most and least important features can be found in Table 5.

ilar information. On a voxel-basis LBPs have the lowest ranking, because they provide less information about the neighborhood, and discarded intensity information.

5.4.2 Area classification

Analogous to the voxel wise classification region-level featureset IDs, and corresponding classification accuracies are listed in Table 6. A corresponding graphical overview can be found in Figure 10. Among the three compared kinds of areas (annotated regions of interest (ROI), large supervoxels (spx n=250), small supervoxels (spx n=250)) performance for larger areas is better. An interesting observation i that LBP based texture bags outperform most other descriptors, despite having poor classification performance on a voxel-level.

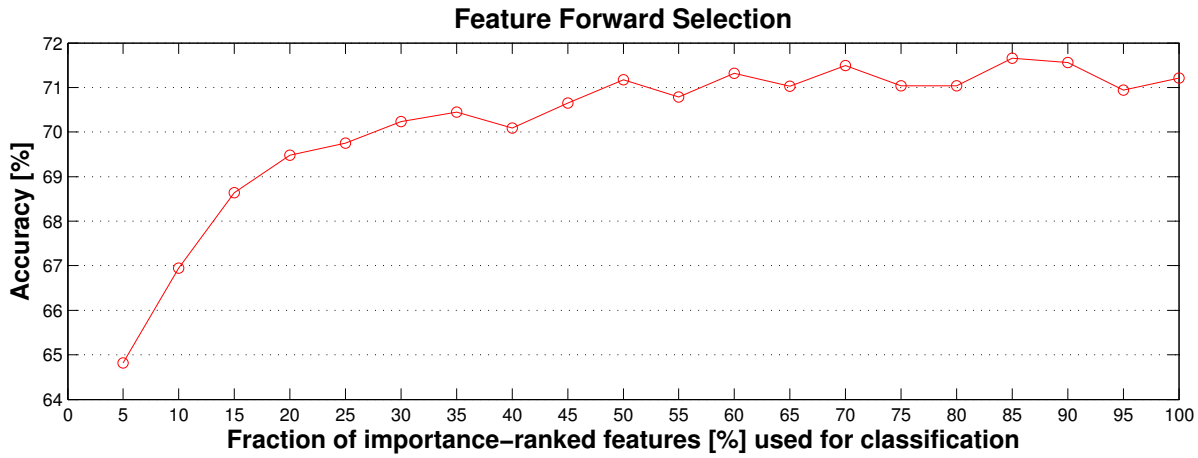


Figure 9: Feature Forward Selection of the feature selection testset. Indicating the overall accuracy of voxel classification including increasing fractions of importace-ranked features of the feature selection testset.

Rank	**Feature Description**
1	Intensity Histogram. Bin 1: gray value interval [-1024,-950].
2	Intensity Histogram. Bin 2: gray value interval [-949,-876].
3	Intensity Histogram. Bin 3: gray value interval [-875,-802].
4	Haralick Feature: Sum average. 17×17, 16 gls, distance 2.
5	Haralick Feature: Sum average. 33×33, 16 gls, distance 4.
6	Haralick Feature: Sum average. 17×17, 16 gls, distance 4.
7	Haralick Feature: Sum average. 33×33, 16 gls, distance 5.
8	Haralick Feature: Sum average. 33×33, 16 gls, distance 2.
9	Haralick Feature: Contrast. 33×33, 16 gls, distance 3.
10	Riesz wavelets. Feature 3.
...	
273	Texton filterbank. Feature 25.
274	Texton filterbank. Feature 26.
275	Haralick Feature: Information measure of correlation 2. 11×11, 16 gls, distance 5.
276	Texton filterbank. Feature 13.
277	Haralick Feature: Sum Entropy. 11×11, 16 gls, distance 4.
278	Haralick Feature: Entropy. 11×11, 16 gls, distance 1.
279	Haralick Feature: Sum Entropy. 11×11, 16 gls, distance 2.
280	LBP RI. r=3, n=24.
281	LBP RI. r=2, n=16.
282	LBP RI. r=1, n=8.

Table 5: Details of the 10 top ranked features and the 10 lowest ranked features for voxel-wise classification.

Table 6: Classification accuracy of the classification of areas for varying areas and number of texture words. The best perfomance is highlighted for every feature set. Graphical overview in Figure 10.

Area description		ROI			Spx n=250			Spx n=500		
Number of texture words		100	200	300	100	200	300	100	200	300
Set ID	Set description									
1	Intensity Histogram	63.56	60.99	61.42	61.54	64.35	**65.64**	62.13	63.40	65.07
2	LBP NOMAP	76.09	76.74	**77.45**	66.47	68.13	69.81	66.43	67.12	68.41
3	LBP RI	76.23	**77.45**	76.74	68.67	68.98	69.05	69.10	69.10	68.72
4	LBP RIU2	72.08	75.02	**75.09**	62.82	65.87	64.99	63.54	63.47	62.55
5	Riesz Wavelets	62.28	**64.28**	62.78	62.13	61.86	62.75	61.39	61.88	61.60
6	Riesz and Intensity Hist.	62.42	62.56	**63.92**	61.45	61.14	63.67	61.00	62.22	61.65
7	Texton filterbank	72.30	**75.23**	74.59	65.58	67.46	68.08	66.38	66.11	66.57
8	Haralick 11 $\times$ 11, 8 gls	72.80	71.94	**72.94**	69.07	71.69	71.69	67.61	69.11	70.21
9	Haralick 11 $\times$ 11, 16 gls	72.30	72.87	**73.73**	70.86	70.42	71.20	69.77	69.84	70.03
10	Haralick 11 $\times$ 11, 32 gls	69.36	72.94	71.44	70.71	72.30	**73.04**	68.20	72.20	72.52
11	Haralick 17 $\times$ 17, 8 gls	68.93	68.58	**70.22**	65.78	67.88	68.55	64.33	67.70	68.83
12	Haralick 17 $\times$ 17, 16 gls	70.15	**72.01**	70.15	68.08	70.17	69.70	68.50	69.13	69.89
13	Haralick 17 $\times$ 17, 32 gls	65.21	70.22	70.79	66.20	70.37	**72.32**	65.09	70.75	71.78
14	Haralick 33 $\times$ 33, 8 gls	59.48	60.27	62.13	61.95	61.34	63.00	61.88	62.14	**63.66**
15	Haralick 33 $\times$ 33, 16 gls	59.77	62.20	63.99	62.42	65.35	**66.65**	61.97	63.66	65.60
16	Haralick 33 $\times$ 33, 32 gls	63.71	62.06	64.14	64.26	63.25	65.22	64.24	64.93	**65.25**
17	40% (113) top-ranked features	60.77	60.77	**62.92**	52.56	54.38	55.19	51.23	52.85	54.91

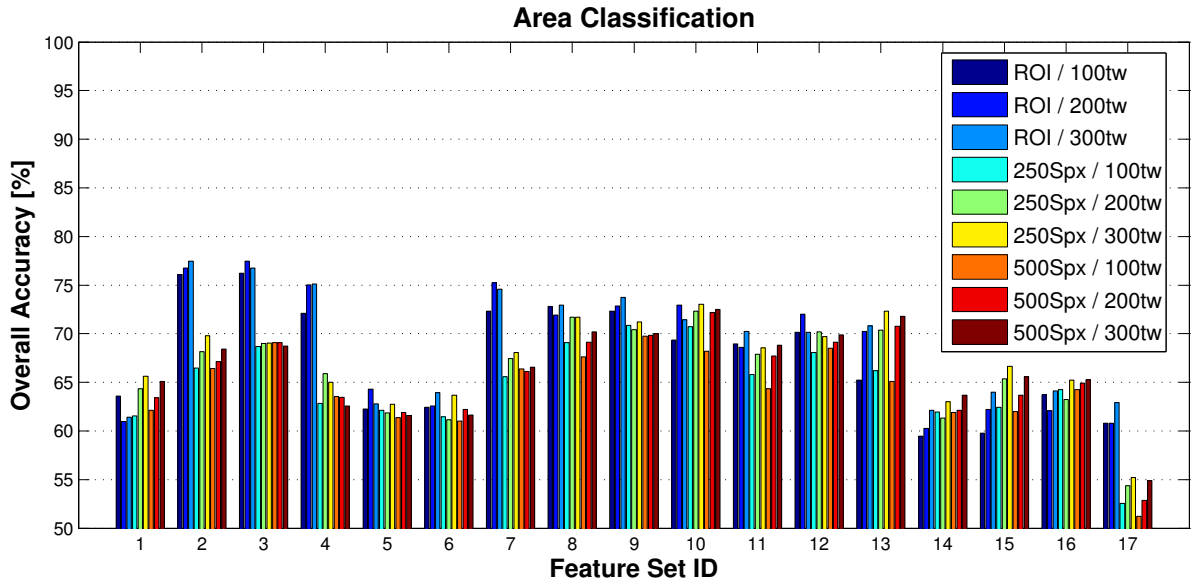


Figure 10: Classification accuracy for region-level classification. The bars represent accuracy for 3 types of regions (ROI, large- and small super-pixels), three vocabulary sizes (100-, 200-, and 300 words), and 17 different feature set. Feature sets are defined in Table 6.

5.4.3 Pattern specific performance

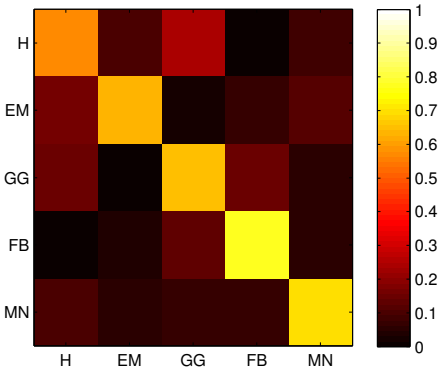
To understand the pattern specific performance of individual descriptors we provide confusion matrices of particular feature sets in Fig. 11. They illustrate the misclassification among pairs of tissue classes. Even if the overall classification accuracy might be comparable among the depicted examples, the class specific accuracy and confusion among pairs of classes exhibits substantial variability.

5.5 Discussion

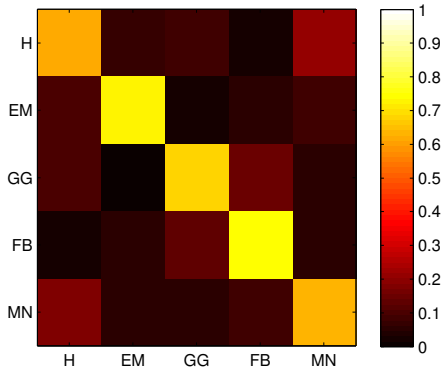
There are several observations worth discussing. The stark contrast between the poor performance of LBPs on a voxel-basis with high performance of the corresponding texture-bag features on an area level is particularly interesting. It suggests that (1) areas hold more information compared to very local patches. This is consistent with the overall improved performance of the bag features on area-level compared with voxel wise features, and the comparably better performance of large neighborhood voxel-wise features such as GLCM with large windows. However, in addition, it shows that learning these area descriptors — as done by texture bags — outperforms fixed features that have a comparable volume of influence (GLCM).

The confusion matrices demonstrate that classification accuracy is not a uniform property across all data. Instead specific characteristics of pathologies seem to be captured to a different extend by different descriptors. Again learning anatomy- and even task-specific feature extractors seems essential to achieve optimal accuracy.

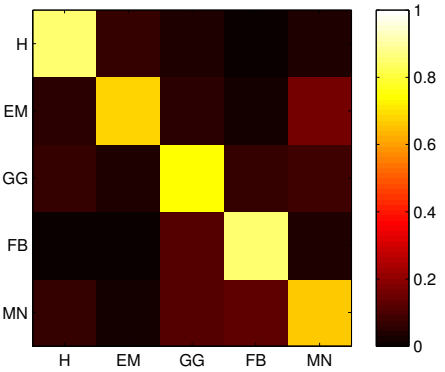
Automated feature selection obtains a sensible ranking of features on a voxel-level but — as expected — this is not predictive with regard to the performance of the corresponding texture



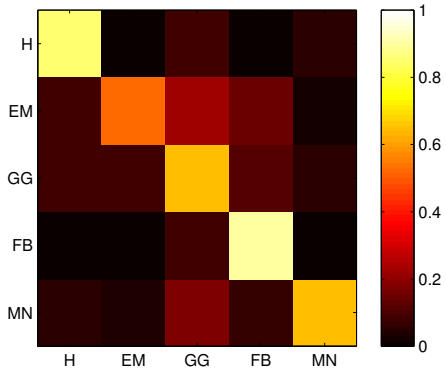
(a) Voxel Classification of Riesz Wavelets and Intensity Histogram.



(b) Voxel Classification of Haralick 17 $\times$ 17, 16 gls.



(c) Area Classification of LBP RI, 300TWs, ROI



(d) Area Classifcation of Haralick 11 $\times$ 11 16 gls, 300TWs, ROI

Figure 11: Confusion Matrices of featuresets with similar performance. Upper row: voxel classification. Lower row: area classification. The abbrevations stand for the tissue types healty, emhpysema, ground glass, fibrosis and micronodules.

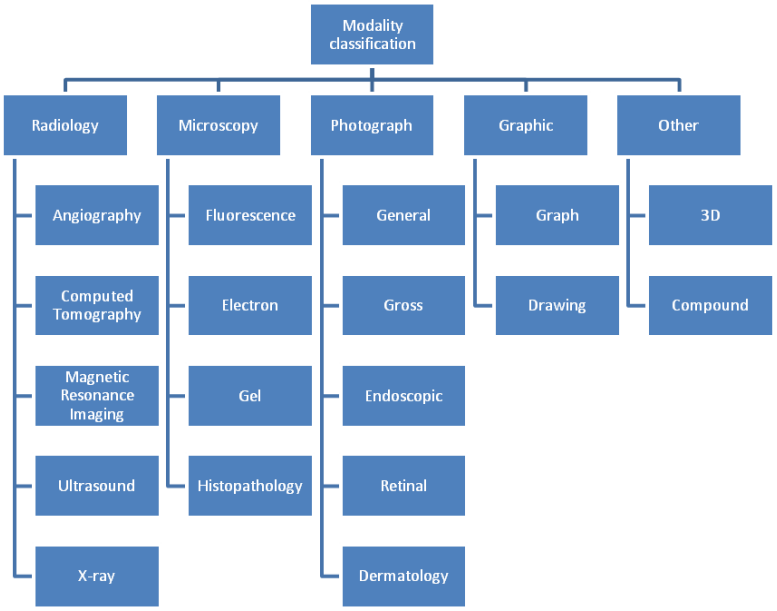


Figure 12: Modality categories of the ImageCLEF 2011 medical modality classification task.

bag features on area-level.

6 Evaluation of image retrieval and classification in 2D

6.1 Data

Part of this text was published in [20, 19]. The evaluation for the 2D medical image retrieval and modality recognition used a database created in the context of the ImageCLEF[3] 2011 benchmark [13] (Image retrieval task of CLEF, the Cross–Language Evaluation Forum). The images of the ImageCLEFmed 2011 collection constitute a realistic sample of medical literature that originate from a large variety of biomedical journals.

The database consists of over 230'000 images and 50'000 articles for image retrieval and case retrieval tasks. Two main challenges of the data set are that (1) there is a large variety of journals, not only radiology, meaning that rigor in figure legends is differing and the variety of images is large and that (2) the data set contains a majority of images that are not or little important for retrieval (such as tables, flow charts, graphs, etc.).

For the modality classification 1'000 training and 1'000 test images were made available with modality class labels and only this subset is being used. The training images are used to train the classifier while the test images are used to validate the classifier quality. Labels are one of 18 categories including graphs and several radiology modalities (see the hierarchy in Figure 12). The sample images presented in Figure 13 demonstrate the visual diversity of the classes of the data set.

[3] http://www.imageclef.org/

(a) Graph (b) Histopathology (c) Endoscopic (d) General Photo (e) Radiology

Figure 13: Sample images from ImageCLEF2011 medical data set

Table 7: Distribution of training and test images in classes

Modality	# of training images	# of test images
Angiography	11	9
Computed tomography	70	83
Magnetic resonance imaging	17	20
Ultrasound	30	41
X–Ray	59	67
Fluorescence	44	28
Electron microscopy	16	18
Gel	50	50
Histopathology	208	195
General Photo	165	141
Gross pathology	43	32
Endoscopic imaging	10	11
Retinograph	5	3
Dermatology	7	15
Graphs	161	172
Drawing	43	74
3D reconstruction	32	45
Compound figure	17	20

The number of images per class in the training and test sets varies from fewer than ten to several hundred (see Table 7 for the exact numbers). This uneven distribution can affect the training of the classifiers and the resulting performance. Participating groups in the ImageCLEF challenge [4] addressed this challenge and took action, for example to automatically expand the training set. This inclusion of additional data to increase the training set may improve the retrieval performance but can also worsen its query speed. In our tests, it was adopted to select a subset of 100 images uniformly distributed across the classes for the creation of the visual vocabulary.

6.2 Evaluation set-up: retrieval

In the ImageCLEF 2011 benchmark, 30 topics were given for evaluating medical image retrieval system performance [13]. Every topic includes one or more query image. Among others, the main evaluation metric used was the mean average precision (MAP). The same topics and evaluation metric were used for our evaluation. The features used was a state–of–the–art approach, called Bag–of–Visual–Words (BoVW) approach using local descriptors as visual words and the with Bag–Of–Colors (BoC) image descriptors. Both BoVW and BoC methods are described in KHRESMOI D2.2 deliverable report. Each image was represented as a list of histograms and the similarity between images was calculated by comparing their histograms using Histogram Intersection [24]. The results lists of the files were then combined into a single list using late fusion [23]. The 1000 most similar images were retrieved and were assessed against the manually created gold standard of the benchmark. For the fusion of the multiple query image results and image features several score–based and rank–based fusion rules [23] were used:

- combSUM;
- combMNZ;
- Borda count;
- reciprocal rank fusion;
- linear score-based fusion.

6.3 Evaluation set-up: classification

For the modality classification evaluation we used the state–of–the–art approach BoVW combined with Bag–Of–Colors (BoC) image descriptors. The evaluation proceduce is the following: late fusion was used to combine the results of BoVW and BoC. First we obtained similarity scores separately using BoVW and BoC descriptors. Then, these scores were fused by voting. The image was classified into one class by a k–NN weighted voting [11].

As accuracy measure, the percentage of correctly classified images of the entire test set of 1000 images was used. This procedure allowed for a fair comparison of the various schemes with and without the use of BoC.

6.4 Results

Table 8 presents the evaluation results of fusing the results of multiple queries and the two image descriptors (BoVW, BoC).

Table 8: Image retrieval Mean Average Precision (MAP) using BoVW / BoC / BoVW + BoVW for varying fusion rules

fusion rule	BoVW	BoC	BoVW+BoC
combSUM	0.0138	0.0120	0.0228
combMNZ	0.0137	0.0101	0.0207
Borda count	0.0109	0.0007	0.0141
Reciprocal	0.0138	0.0099	0.0173
Linear (0.6,0.4)	–	–	**0.0232**
Linear (0.7,0.3)	–	–	0.0214

In order to tune parameters for the modality classification task, the training data were split in half, with one half serving as training set and the other as test set. The results for the training data with varying k_c over BoC and k_{nn} are shown in Table 9.

The results for the test data with varying k_{nn} are shown in Table 10. Using only BoVW, the best accuracy is 62.5% and results are stable for varying k_{nn}. For BoC the best accuracy is 63.96%, also quite stable across varying k_{nn}. For each k_{nn}, the fusion of BoC and BoVW produces an improved accuracy. The best overall fused result is 72.46%.

6.5 Discussion

For the image retrieval task, it can be observed that all the fusion techniques improve the retrieval performance. Score–based fusion rules (combSUM, combMNZ, Linear fusion) outperform the rank–based ones (Borda count, Reciprocal rank fusion), with linear weighting (0.6 for BoVW, 0.4 for BoC) achieving the best MAP. Overall performance is still an order lower than the best textual run in ImageCLEF. However, as it will be shown in the next section, it retrieves different information than the textual retrieval and the combination of content–based image retrieval and retrieval by quering keywords outperforms both approaches.

We used several vocabulary sizes on the training data to obtain the optimal $k_c = 200$ that was applied on the test data. As seen in the confusion matrices in Figure 14, there are more misclassified color images using BoVW than BoC such as in histopathology (HX), general photos (PX) or fluorescence (FL). Using BoVW, there are fewer mistakes in radiology images (grey level) such as magnetic resonance imaging (MR), angiography (AN) or xray (XR). Figure 14(c) shows that the fusion of BoVW and BoC reduces the number of errors in both, color and grey level images.

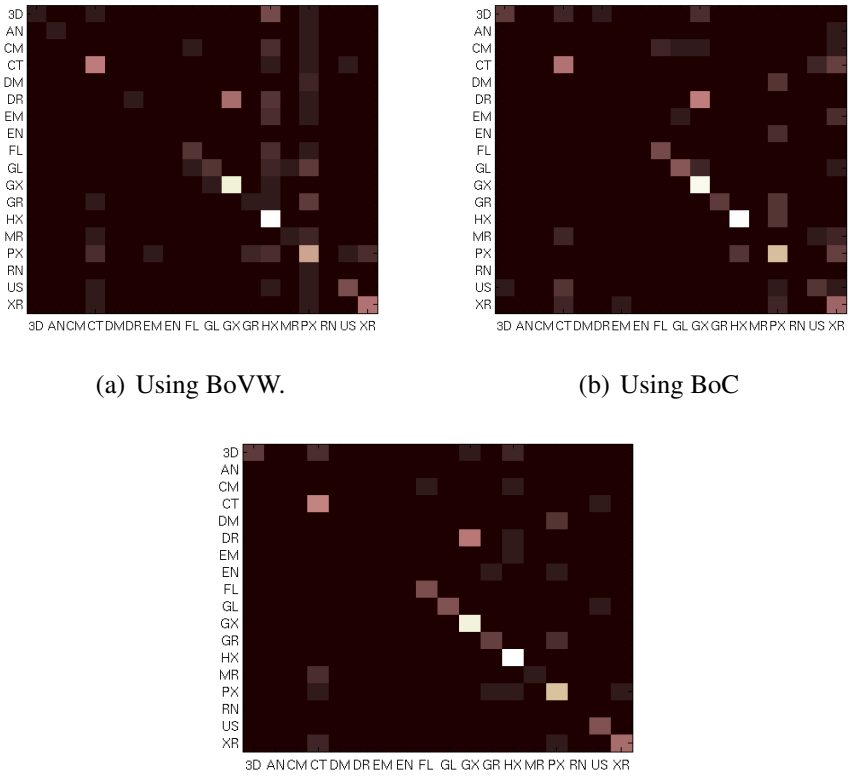
The best result in the modality classification task of ImageCLEF 2011 using visual methods [13] was obtained by Xerox Research with 83.59% accuracy. This result is not comparable with our technique as the improvement was mainly due to an increased training set using data

Table 9: Classification accuracy using BoC/BoVW/both with varying k_c and k_{nn} over the training data.

k_{nn}	BoVW	BoC & k_{10}	BoC & k_{20}	BoC & k_{200}	BoC & k_{500}	BoC & k_{1000}
2	27.27	14.834	24.40	24.83	24.84	24.40
3	28.06	17.80	27.58	26.81	26.81	27.03
4	28.56	16.15	28.13	28.46	28.46	28.24
5	28.85	17.80	27.80	28.90	29.23	28.46
6	29.15	17.14	28.57	**30.10**	29.45	30
7	28.95	19.01	28.57	29.78	29.34	28.57
8	28.85	19.01	29.78	**30.10**	29.01	28.68
9	**29.45**	18.57	29.67	28.68	28.79	28.79
10	29.35	18.35	29.45	29.12	29.23	28.35
11	29.05	18.57	28.79	29.12	29.12	29.23
12	29.15	18.79	28.90	29.01	29.89	29.01
13	28.75	18.79	29.34	29.12	28.68	29.01
14	28.95	18.57	29.23	29.45	28.79	28.68
15	29.35	18.79	29.45	28.79	28.46	28.35
16	29.25	18.02	28.90	28.35	28.68	28.35
17	**29.45**	18.35	28.79	27.91	28.57	28.79
18	29.25	18.57	29.45	27.80	28.46	27.91
19	29.15	18.35	29.56	27.91	28.46	27.80

Table 10: Classification accuracy using BoC/BoVW/both with varying k_{nn} over the test data.

k_{nn}	BoVW	BoC	BoVW+BoC	k_{nn}	BoVW	BoC	BoVW+BoC
2	59.77	54.20	63.96	11	61.23	63.28	**72.46**
3	60.94	59.47	69.14	12	60.94	63.09	71.58
4	62.01	59.96	70.61	13	61.23	62.40	72.17
5	62.21	62.60	71.39	14	61.52	63.18	71.48
6	**62.50**	62.99	70.61	15	61.04	63.18	70.61
7	62.40	62.60	71.19	16	60.55	**63.96**	70.70
8	**62.50**	63.48	71.58	17	61.04	63.67	70.51
9	61.82	62.89	71.29	18	60.64	63.38	70.41
10	61.62	63.48	70.61	19	60.16	63.67	70.12



(a) Using BoVW.

(b) Using BoC

(c) Using BoVW and BoC

Figure 14: Confusion Matrices obtained for the modality classification results using different features.

other than the original training data. Without the additional training set the obtained performance was at only 62.2% [3]. The best accuracy using visual methods without increasing the training data was 69.72%, obtained by the University of Campinas [7]. Using our fusion strategy of BoC and BoVW a better accuracy was obtained.

7 Evaluation of combined text and image based retrieval

7.1 Data

For the evaluation of the textual and visual information synergy the imageCLEF 2011 database was used also, as in in Section 6.1. The database, containing both medical articles and the images included in them, can be used in realistic scenarios of case retrieval.

7.2 Evaluation set-up

Topics from imageCLEF 2011 benchmark were used, again, for this evaluation. This time, the case retrieval task was used, containing 10 topics/cases from the teaching file Casimage [21]. The runs evaluated were visual, textual and mixed. The visual run used a combMNZ fusion of BoVW and BoC as described in Section 6.2. As these approaches retrieve images instead of articles, a mapper from images to the corresponding articles was used. When multiple images from the same article were retrieved, the score of the best ranked image was assigned to the article. The textual run was achieved using the Lucene[4] library, and included stop words and word stemming removal. This run achieved the second best performance in the ImageCLEF 2011 medical case retrieval challenge, obtaining 0.1293 against the 0.1297 MAP of the best run.

The mixed runs were created using late fusion of the visual and textual runs. Again, both score–based and rank–based fusion rules were used.

7.3 Results

The results are presented in table 11

7.4 Discussion

It is observed that only score–based fusion rules outperform the textual runs. This can explained by the gap in performance between textual and visual runs. The rank-based rules ignore the bad similarity scores of the visual run and weights equally the two systems. The scored–based rules however outperform the best text run in ImageCLEF 2011, even with this simple rule of mapping images to articles. This mapping doesn't rewards articles that contain more than one relevant images and this may hold back case retrieval performance. Small evaluation has also been done on the combination of text and image information for modality classification. The results on the modality classification task used in Section 6.2 display an accuracy of 83.00% over the 72.00% of the visual run and the 65.00% of the text search.

[4] http://lucene.apache.org/core/

Table 11: Case retrieval Mean Average Precision (MAP) of textual/ visual / mixed runs using the fusion rules described in Section 6.2

run	MAP
visual	0.0312
textual	0.1216
mixed-combSUM	0.1334
mixed-combMNZ	0.1379
mixed-Borda count	0.1090
mixed-Reciprocal	0.1132
mixed-Linear (0.6,0.4)	**0.1433**
mixed-Linear (0.7,0.3)	0.1349

8 Evaluation of text-based image retrieval

8.1 Description

The HON image retrieval system is a purely textual retrieval system, described in detail in [10], Section 4.2.4. Starting from general HTML pages crawled from the web, the HON system detects images referenced by the `img` tag of a web page, extracts textual metadata (such as the `alt` and `title` attributes) and identifies surrounding related text. Once indexed, this enables the retrieval of arbitrary web images based on their corresponding text.

8.2 Evaluation setup

A proper evaluation of a retrieval systems requires a document collection as a base for testing. For the evaluation of this system, the document collection of the ImageCLEF 2011 Wikipedia retrieval task[29] was used. Although this collection is not related to the medical domain, it is well-suited for the purpose of this evaluation, since it provides a means for testing the association of free-form text with images embedded in HTML pages.

The image search system developed by HON is based on crawling web pages and extracting images from those pages along with related text. It was determined that the medical image collection proposed for the CLEF2011 Medical Image Classification and Retrieval Tasks[14] is not suitable for the evaluation of such a system, since it consists of images extracted from journal articles, rather then those extracted from web pages.

This collection used in this evaluation provides 237,434 images, each associated with an annotation. Each annotation consists of an image caption in one or more of the following languages: English, French and German. It also provides information from the Wikipedia article from which the image was extracted. Listing 1 illustrates a sample of the image metadata.

Listing 1: Image metadata

```
<?xml version="1.0" encoding="UTF-8" ?>
<image id="1" file="images/1/1.jpg">
```

```
 <name>Eod2.jpg</name>
 <text xml:lang="en">
  <description />
  <comment />
  <caption article="text/en/1/309678">Inserting detonators into blocks of C−4 explosive</caption>
 </text>
 <text xml:lang="de">
  <description/>
  <comment />
  <caption />
 </text>
 <text xml:lang="fr">
  <description />
  <comment />
  <caption article="text/fr/3/523790">Preparation du C−4</caption>
 </text>
 <comment>(Lifted from [http://www.usmc.mil/marinelink/image1.nsf/lookup/2005230387?opendocument]
        Caption: "Pfc. Laura Mellinger, Headquarters and Headquarters Squadron aircraft rescue fire fighting crewman,
        inserts blasting caps into blocks of C−4 at Target Island)</comment>
 <license>Public Domain</license>
</image>
```

Because the text within the `caption` tag tends to concisely and accurately describe the image, and because it may be found on the original page, it was retained as a baseline for comparison. Next, the original Wikipedia articles were crawled based on the unique id encapsulated in the `article` attribute within the listing, which are elsewhere linked with enough data to reconstruct the original page URL. Only the relevant Wikipedia page revisions were crawled. These documents were then analysed using the HON system, which extracted images and their corresponding text.

In this way, the evaluation could compare the effectiveness of the HON system with that of the previously described baseline. To ensure that only the accuracy of the image extraction process itself was tested (i.e., the ability of the system to correctly associate text with images), all text fields were indexed using a Solr search server based on the Lucene Java search library[8]. This is the same index used in the user-facing HON Search system [10].

Two variations of the HON extraction algorithm were tested. The first, `honSearch-small`, limits extracted text to a window of 200 and 300 bytes for preceding and following text relative to the `img` tag, respectively. The second variation, `honSearch-large`, expands the window to within 3 HTML tags of the `img` tag, but only includes the text based on a nonzero term similarity with the `alt`, `title`, or filename.

The ImageCLEF 2011 Wikipedia retrieval task collection provides 50 queries for evaluation. An example of a typical query is given in Listing 2. In this evaluation, only the information encapsulated within the `title` was used for querying.

Listing 2: Topic example.

```
<?xml version="1.0" encoding="windows−1250"?>
<topic>
    <number> 71 </number>
    <title xml:lang="en"> colored Volkswagen beetles </title>
    <title xml:lang="de"> farbige Volkswagen Kfer </title>
```

<title xml:lang="fr"> Volkswagen Coccinelles colores </title>
<image> vwbeetle1.jpg </image>
<image> vwbeetle2.jpg </image>
<image> vwbeetle3.jpg </image>
<image> vwbeetle4.jpg </image>
<image> vwbeetle5.jpg </image>
<narrative>Photos of Volkswagen beetles in white or black are not relevant.
Photos of Volkswagen beetles in any other color, for example, red, blue, green or yellow are relevant.</narrative>
</topic>

8.3 Evaluation results

Table 12 presents the results obtained during the test runs. In this evaluation, two measures are used: Mean Reciprocal Rank and First Relevant Score[26].

Mean Reciprocal Rank (MRR) represents the mean of the inverse rank of the first relevant document, and is often used to evaluate web-based systems. This measure has two interesting advantages. First, its value is easily interpretable, taking into account only the first relevant document retrieved, and thus reflects users' preferences for finding relevant documents earlier in the search list. Secondly, it harshly penalises systems that do not provide a relevant document as the first result, awarding quite different scores for first and second place (i.e., 1 and 0.5).

First Relevant Score (FRS) is a similar measure which gives less importance to the first document position. Calculated as K^{1-r}, where r is the rank of the first relevant document and K is a constant fixed at 1.08, it awards values on a smoother scale with less severe penalties for small positional differences. For the first relevant document at the first position, the score is 1, as with MRR. After that, for the second position the score is 0.926 (compared to 0.5), for third 0.875 (compared to 0.333), etc., eventually arriving at a score of 0.5 for the tenth position.

Table 12: MRR and FRS results for the baseline and HON system.

Language	English		German		French	
Measure	MRR	FRS	MRR	FRS	MRR	FRS
baseline	0.6108	0.7513	**0.4561**	**0.6251**	**0.4474**	0.6469
honSearch-small	0.5826	0.8135	0.2451	0.4964	0.3437	0.6176
honSearch-large	**0.6412**	**0.8432**	0.3176	0.6064	0.3786	**0.6571**

8.4 Discussion

As it can be seen from the results, our system is capable of achieving, with `honSearch-large`, retrieval performance superior to the baseline for the English language. The baseline run, for this language, returns the first relevant image at the first position for 26 queries, while for the `honSearch-large` run this is the case for 24 queries. In the case of German and French documents, the performance is somewhat lower, even for the baseline. This can be explained by the fact that there are many fewer images within the collection containing an annotation in

these two languages; there are 124,342 annotated images for English, 107,373 for German, and 84,290 for French.

Comparing the results obtained for the German language, the performance difference between the baseline and `honSearch-large` is -30% for the MRR. If we compare the values obtained for FRS the difference in the performance is -3%. This indicates that the low MRR scores are mostly due to small positional differences at the beginning of the search results.

On the other hand, if one compares the MAP value of the HON system's best run (`honSearch -large` in English) of 0.0941 to the best performing textual run in the official CLEF2011 results (0.3141), the results obtained by the HON system can be judged as rather low. In this case, one has to take into account that the results presented here are all for monolingual runs, using only the topic title as the query, with no query feedback (FB) or query expansion (QE). Furthermore, in this collection only 10% of images are annotated in all three languages while 62% of images are annotated in only one language, thus penalising the monolingual retrieval results. This makes the results achived by the HON system not easily comparable to the official results, since most of the runs presented are multilingual runs using FB, QE or both.

9 Conclusion

In this deliverable we report evaluation results of the image description methods used for content-based image retrieval in KHRESMOI. The focus of the work at this stage is to build powerful image descriptors, and local appearance feature extractors, that represent image content and serve as basis for subsequent steps. We report results for 4 core functionalities in the image retrieval components of the KHRESMOI system. The retrieval of anatomical structures, pathologic structures relevant for diagnosis, 2D images in documents, and retrieval based on both text- and image information.

The results confirm that content-based medical image retrieval is feasible, and highlight specific characteristics of the data, and corresponding requirements, that deviate from general images. Chief among them is the necessity to learn feature extractors in order to obtain sufficient specificity for effective retrieval of often subtle details relevant in a general medical- or clinical radiology context.

The next steps of the project will build upon these results, and will use them as basis for the development of un- and weakly supervised learning schemes, that are necessary to structure, and index the medical imaging data for the final retrieval system.

10 References

[1] Leo Breiman. Random forests. In *Machine Learning*, pages 5–32, 2001.

[2] Andreas Burner, Rene Donner, Marius Mayerhoefer, Markus Holzer, Franz Kainberger, and Georg Langs. Texture bags: Anomaly retrieval in medical images based on local 3d-texture similarity. *Workshop on Medical Content-based Retrieval for Clinical Decision Support at MICCAI 2011*, September 2011.

[3] Gabriela Csurka, Stéphane Clinchant, and Guillaume Jacquet. XRCE's participation at medical image modality classification and ad–hoc retrieval task of ImageCLEFmed 2011. In *Working Notes of CLEF 2011*, 2011.

[4] Gabriela Csurka, Stephane Clinchant, and Guillaume Jacquet. Xrce's participation at medical image modality classification and ad-hoc retrieval tasks of imageclef 2011. In *Working Notes of CLEF 2011 (Cross Language Evaluation Forum)*, September 2011.

[5] Adrien Depeursinge, Alejandro Vargas, Alexandra Platon, Antoine Geissbuhler, Pierre-Alexandre Poletti, and Henning Müller. Building a reference multimedia database for interstitial lung diseases. *Computerized Medical Imaging and Graphics*, 36(3):227–238, April 2012.

[6] Rene Donner, Sebastian Haas, Andreas Burner, Markus Holzer, Horst Bischof, and Georg Langs. Evaluation of Fast 2D and 3D Medical Image Retrieval Approaches based on Image Miniatures. In *Proc. MICCAI Workshop on Medical Content-based Retrieval for Clinical Decision Support*, 2011.

[7] Fábio Augusto Faria, Rodrigo Tripoli Calumby, and Ricardo da Silva Torres. RECOD at ImageCLEF 2011: Medical modality classification using genetic programming. In *Working Notes of CLEF 2011*, 2011.

[8] Apache Software Foundation. Lucene. `http://lucene.apache.org`, 2011.

[9] Sebastian Haas, René Donner, Andreas Burner, Markus Holzer, and Georg Langs. SuperPixel-based Interest Points for Effective Bags of Visual Words Medical Image Retrieval. In Müller, Greespan, and Syeda-Mahmood, editors, *Proc. of MICCAI Medical Content-based Retrieval Workshop*, 2011.

[10] Allan Hanbury, William Belle, Nolan Lawson, Ljiljana Dolamic, Natalia Pletneva, and Matthias Samwald. D8.3: Prototype of a first search system for intensive tests. *Khresmoi project public deliverable*, 2012.

[11] David J. Hand, Heikki Mannila, and Padhraic Smyth. *Principles of Data Mining (Adaptive Computation and Machine Learning)*. The MIT Press, 2001.

[12] R. M. Haralick, Dinstein, and K. Shanmugam. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-3:610–621, 1973.

[13] Jayashree Kalpathy-Cramer, Henning Müller, Steven Bedrick, Ivan Eggel, Alba Seco de Herrera, and Theodora Tsikrika. The CLEF 2011 medical image retrieval and classification tasks. In *Working Notes of CLEF 2011 (Cross Language Evaluation Forum)*, September 2011.

[14] Jayashree KalpathyCramer, Henning Müller, Steven Bedricks, Alba G. Seco de Herrera Ivan Eggel, and Theodora Tsikrika. Overview of the clef 2011 medical image classification and retrieval tasks. *Working Notes of CLEF 2011*, 2011.

[15] Georg Langs, Bjoern H. Menze, Danial Lashkari, and Polina Golland. Detecting stable distributed patterns of brain activation using gini contrast. *NeuroImage*, 56(2):497 – 507, 2011.

[16] Thomas M. Lehmann, Henning Schubert, Daniel Keysers, Michael Kohnen, and Berthold B. Wein. The IRMA code for unique classification of medical images. In *Medical Imaging 2003: PACS and Integrated Medical Information Systems: Design and Evaluation (Proceedings Volume)*, volume 5033, pages 440–451. SPIE, 2003.

[17] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[18] Jitendra Malik, Serge Belongie, Thomas Leung, and Jianbo Shi. Contour and texture analysis for image segmentation. *IJCV*, 43(1):7–27, June 2001.

[19] Dimitrios Markonis, Ivan Eggel, Alba G.Seco de Herrera, and Henning Müller. The medGIFT group in ImageCLEFmed 2011. In *Working Notes of CLEF 2011*, 2011.

[20] Dimitrios Markonis, Alba Garcia Seco de Herrero, Ivan Eggel, and Henning Müller. Multi–scale visual words for hierarchical medical image categorisation. In *SPIE medical imaging: Advanced PACS-based Imaging Informatics and Therapeutic Application*, February 2012.

[21] Antoine Rosset, Henning Müller, Martina Martins, Natalia Dfouni, Jean-Paul Vallée, and Osman Ratib. Casimage project — a digital teaching files authoring environment. *Journal of Thoracic Imaging*, 19(2):1–6, 2004.

[22] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *Proceedings of the Ninth IEEE International Conference on Computer Vision - Volume 2*, volume 2, pages 1470–1477. IEEE Computer Society, 2003.

[23] Cees G. M. Snoek, Marcel Worring, and Arnold W. M. Smeulders. Early versus late fusion in semantic video analysis. In *MULTIMEDIA '05: Proceedings of the 13th annual ACM international conference on Multimedia*, pages 399–402, New York, NY, USA, November 2005. ACM.

[24] Michael J. Swain and Dana H. Ballard. Color indexing. *International Journal of Computer Vision*, 7(1):11–32, 1991.

[25] Ojala T., Pietikinen M, and Menp T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. 2002. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(7):971 - 987.

[26] Stephen Tomlinson. Bulgarian and Hungarian Experiments with Hummingbird SearchServer™ at CLEF 2005. In Carol Peters, Fredric C. Gey, Julio Gonzalo, Henning Müller, Gareth J. F. Jones, Michael Kluck, Bernardo Magnini, and Maarten de Rijke, editors, *CLEF*, volume 4022 of *Lecture Notes in Computer Science*, pages 194–203. Springer, 2005.

[27] Tatiana Tommasi, Barbara Caputo, Petra Welter, Mark Güld, and Thomas Deserno. Overview of the CLEF 2009 medical image annotation track. In Carol Peters, Barbara Caputo, Julio Gonzalo, Gareth Jones, Jayashree Kalpathy-Cramer, Henning Müller, and Theodora Tsikrika, editors, *Multilingual Information Access Evaluation II. Multimedia Experiments*, volume 6242 of *Lecture Notes in Computer Science*, pages 85–93. Springer Berlin / Heidelberg, 2010.

[28] A Torralba, R Fergus, and WT Freeman. 80 Million Tiny Images: A large Data Set for Nonparametric Object and Scene Recognition. *TPAMI*, 2008.

[29] Theodora Tsikrika, Adrian Popescu, and Jana Kludas. Overview of the wikipedia retrieval task at imageclef 2011. *Working Notes of CLEF 2011*, 2011.

[30] Michael Unser and Dimitri Van De Ville. Wavelet steerability and the higher–order Riesz transform. *IEEE Transactions on Image Processing*, 19(3):636–652, March 2010.

[31] H. Wildenauer, B. Micusk, and M. Vincze. Efficient texture representation using multi-scale regions. In *Proceedings of the 8th Asian conference on Computer vision - Volume Part I*, pages 65–74. Springer-Verlag, 2007.