Computerized Medical Imaging and Graphics xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

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

Improved medical image modality classification using a combination of visual and textual features

Ivica Dimitrovski[a], Dragi Kocev[b], Ivan Kitanovski[a], Suzana Loskovska[a], Sašo Džeroski[b]

[a] Faculty of Computer Science and Engineering, University Ss. Cyril and Methodius, Skopje, Macedonia
[b] Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia

ARTICLE INFO

Article history:
Received 18 December 2013
Received in revised form 2 June 2014
Accepted 5 June 2014

Keywords:
Image modality classification
Visual image descriptors
Feature fusion

ABSTRACT

In this paper, we present the approach that we applied to the medical modality classification tasks at the ImageCLEF evaluation forum. More specifically, we used the modality classification databases from the ImageCLEF competitions in 2011, 2012 and 2013, described by four visual and one textual types of features, and combinations thereof. We used local binary patterns, color and edge directivity descriptors, fuzzy color and texture histogram and scale-invariant feature transform (and its variant opponentSIFT) as visual features and the standard bag-of-words textual representation coupled with TF-IDF weighting. The results from the extensive experimental evaluation identify the SIFT and opponentSIFT features as the best performing features for modality classification. Next, the low-level fusion of the visual features improves the predictive performance of the classifiers. This is because the different features are able to capture different aspects of an image, their combination offering a more complete representation of the visual content in an image. Moreover, adding textual features further increases the predictive performance. Finally, the results obtained with our approach are the best results reported on these databases so far.

1. Introduction

Large collections of medical images have become a valuable source of knowledge, taking an important role in education, medical research and clinical decision making. One of the main problems is that the size of the collections is in constant growth due to the increasing availability of imaging equipment in hospitals. Average-sized radiology departments now produce several tera-bytes of data annually. This generates huge repositories of valuable information, which in many cases is difficult to process and manage appropriately. This prompts for development of tools for efficient and effective access to this type of information.

Medical image databases are typically accessed via textual information through the standard Picture Archiving and Communication System (PACS) [1,2]. PACS integrates imaging modalities and interfaces with hospital and departmental information systems to manage storage and distribution of images to medical personnel, researchers, clinics, and imaging centers. The task of indexing and cataloging these collections has been traditionally performed manually. This is an expensive and time-consuming procedure, and it is also prone to errors. Consequently, there is a strong need for automated indexing of medical image collections in order to improve the ability to search for and retrieve relevant images [3].

Medical image retrieval systems have traditionally been text-based, relying on the annotation or captions associated with the images as the input to the retrieval system. In the last few decades, several advancements in the area of content-based image retrieval (CBIR) have been made [4,5]. CBIR systems have had some success in fairly constrained medical domains, such as pathology [6], head MRIs [7], lung CTs [8], and mammograms [9]. Furthermore, combining both textual and visual techniques improves the retrieval performance over using them individually [3,10]. The queries, in that case, consist of a textual part (i.e., textual sub-query) and/or sample images (i.e., visual sub-query). For example, the queries could contain information about patients' demographics, a limited set of symptoms and medical examination results including imaging studies.

Medical image databases used for retrieval or for teaching purposes often contain images of many different modalities, such as X-ray, CT scan, ultrasound, etc. An additional complication is that these images are typically taken under different conditions and the accuracy of their annotations is variable and inconsistent [11]. This is especially true for images found in various on-line resources, including those that access the on-line content of journals.

E-mail addresses: Ivica.Dimitrovski@finki.ukim.mk (I. Dimitrovski), Dragi.Kocev@ijs.si (D. Kocev), Ivan.Kitanovski@finki.ukim.mk (I. Kitanovski), Suzana.Loskovska@finki.ukim.mk (S. Loskovska), Saso.Dzeroski@ijs.si (S. Džeroski).

http://dx.doi.org/10.1016/j.compmedimag.2014.06.005
0895-6111/

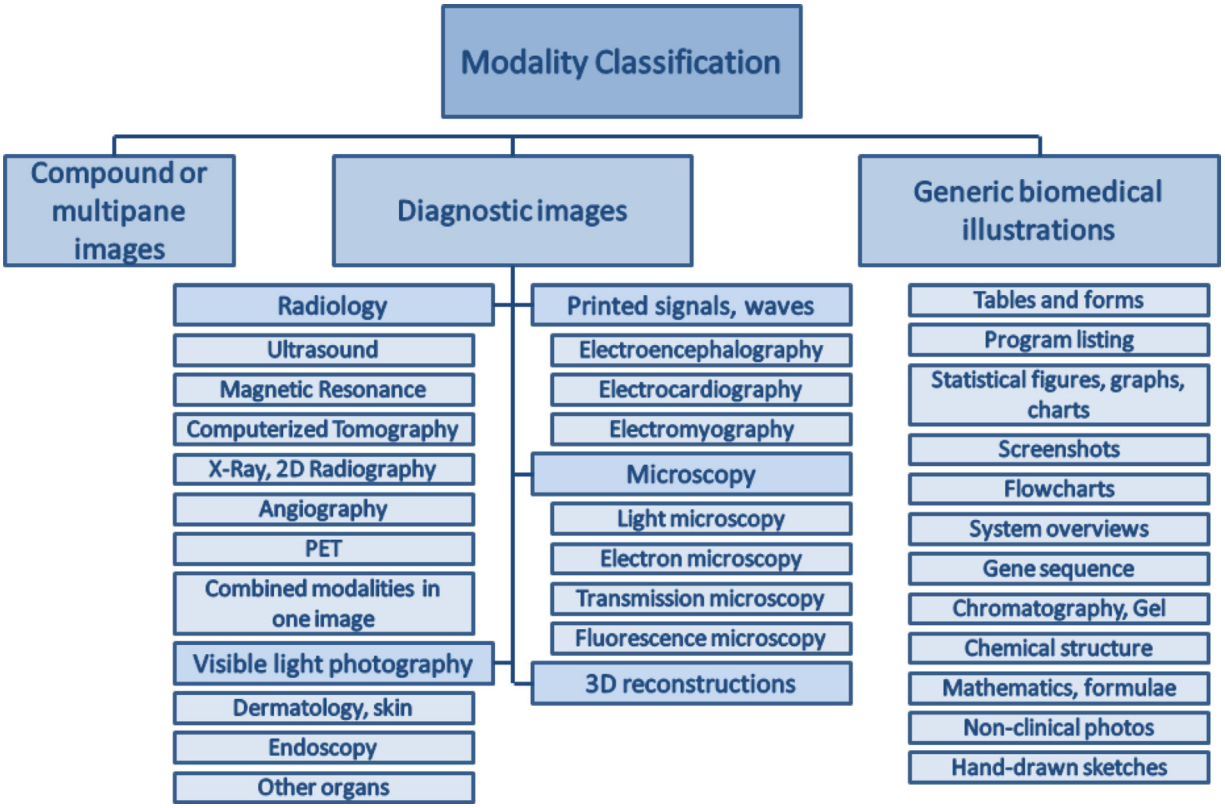

Fig. 1. The classification hierarchy used for the ImageCLEF 2012 and 2013 competitions. The image is taken from http://www.imageclef.org/2013/medical.

Image modality is a fundamental visual characteristic of an image and can be exploited for improving retrieval performance. However, the annotations or captions associated with images often do not capture information about the modality. The DICOM header contains tags to decode the body part examined, the patient position and the modality [12]. Some of the tags are automatically set by the digital imaging system according to the imaging protocol used to capture the pixel data. Other tags are set manually by the physicians or radiologists during routine documentation. This procedure cannot always be considered very reliable, since it frequently happens that some entries are either missing, false, or do not describe the anatomic region precisely [13].

The medical retrieval task in ImageCLEF has provided both a forum as well as test image collections to benchmark image retrieval techniques. Over the years, our group has participated in different subtasks of this medical task including the subtasks of automatic medical image annotation [14], medical modality classification, compound figure separation, ad-hoc image-based retrieval and case-based retrieval [3,10]. In this paper, we present in detail the results of applying our approach to modality classification for the competitions organized in 2011, 2012 and 2013.

In the modality classification task at the ImageCLEF competition, the examples are images from medical articles. The goal is to correctly classify the modality of the images using the visual information from the images and the text from the article where this image is encountered. In this work, we extensively compare the performance of 4 different techniques for feature extraction from images: local binary patterns (LBP) [15], the color and edge directivity descriptor (CEDD) [16], fuzzy color and texture histograms (FCTH) [17] and the scale-invariant feature transform (SIFT) with its variant opponentSIFT (OSIFT) [18,19]. Next, we evaluate the performance of textual features extracted from the text surrounding the images by using the standard bag-of-words representation together with the TF-IDF weighting [20].

The main focus of this work is to first explore which visual features extraction technique captures the most relevant information about the medical image modality. Second, we investigate whether the combination of the different visual features improves the predictive performance. Next, we compare the performance of visual and textual features in the context of medical image modality classification. Finally, we investigate whether combining visual and textual features improves the performance and yields state-of-the-art performance.

The reminder of this paper is organized as follows. Section 2 presents the task of modality classification at ImageCLEF competitions. The visual and textual feature extraction techniques are described in Section 3. The specific experimental setup used to evaluate the feature extraction techniques is outlined in Section 4. Section 5 discusses the results from the experiments and compares the different evaluation scenarios. Finally, Section 6 states our conclusions.

2. The task of modality classification

For medical retrieval purposes, imaging modality is an important aspect of the image. In user studies, clinicians have indicated that modality is one of the most important filters that they would like to be able to limit their search by [3]. The usage of modality information often significantly improves the retrieval results.

The ImageCLEF medical modality classification task is a standardized benchmark for systems that automatically classify medical image modality from PubMed journal articles. This task was first introduced in ImageCLEF 2010, when the total number of modalities was eight. In ImageCLEF 2011, the number of modalities was expanded to 18. The ImageCLEF 2012 and 2013 dataset have 31 classes (same number of classes and same classification hierarchy): However, in ImageCLEF 2013, a larger number of compound/multipane images (i.e., images that contain figures of several

Please cite this article in press as: Dimitrovski I, et al. Improved medical image modality classification using a combination of visual and textual features. Comput Med Imaging Graph (2014), http://dx.doi.org/10.1016/j.compmedimag.2014.06.005

Table 1
Properties of the image databases for evaluation of medical modality classification algorithms.

Image database	#Train images	#Test images	#Classes
2011	988	1024	18
2012	1001	1000	31
2013	2901	2582	31

types) is present. This makes the task significantly harder, but corresponds much more closely to the reality of biomedical journals [3].

In this work, we focus on the databases used for the competitions from 2011 until 2013. We made this choice primarily because of the data availability. Namely, our research groups participated at the competitions held in these years. In addition, the task for 2010 is relatively simple: the accuracy of the top ranked system was very high 94%. Nevertheless, the methods we present here can also be easily applied to the database from 2010.

The modality classes used for the competitions in 2012 [10] and 2013 [3] are given in Fig. 1. The competition in 2011 [21] contained in total 18 classes: 3*D* reconstruction, angiography, compound figure (more than one type of image), computed tomography, dermatology, drawing, electron microscopy, endoscopic imaging, fluoresence, gel, graphs, gross pathology, histopathology, magnetic resonance imaging, general photo, retinography, ultrasound and X-ray. In this work, we did not exploit the hierarchy of the classes, but rather treated the problem as a multi-class classification task.

The benchmark databases consist of different numbers of training images and different numbers of testing images, distributed differently across the modalities. These databases are briefly summarized in Table 1. Finally, Fig. 2 shows sample images from the ImageCLEF 2013 database used in our experiments.

3. Extraction of visual and textual features

Collections of medical images typically contain various images obtained using different imaging techniques. In order to properly represent the images, different feature extraction techniques that are able to capture the different aspects of an image (e.g., texture, shapes, color distribution...) need to be used [14]. Texture is especially important, because it is difficult to classify medical images using shape or gray level information. Effective representation of texture is needed to distinguish between images with equal modality and layout. Furthermore, local image characteristics are fundamental for image interpretation: while global features retain information on the whole image, the local features capture the details. They are thus more discriminative concerning the problem of inter and intra-class variability [22].

Our approach to medical modality classification uses different visual features combined with textual features extracted from the surrounding text content of the images. More specifically, as visual features, we used local binary patterns, color and edge directivity descriptors, fuzzy color and texture histograms and (opponent) scale-invariant feature transform descriptors. For the textual information, we used the text surrounding the image and its bag-of-words representation. We made the obtained descriptors publicly available for download at http://kt.ijs.si/DragiKocev/imageclef/descriptors.zip. In the remainder, we briefly describe these feature extraction techniques.

3.1. Local binary patterns

Local binary patterns (LBP) are one of the best representations of texture content in images [15]. They are invariant to monotonic changes in gray-scale images and fast to compute. Furthermore, they are able to detect different micro patterns, such as edges, points and constant areas.

The basic idea behind the LBP approach is to use the information about the texture from a local neighborhood. First, we define the radius R of the local neighborhood under consideration. The LBP operator then builds a binary code that describes the local texture pattern in the neighborhood set of P pixels. The binary code is obtained by applying the gray value of the neighborhood center as a threshold. The binary code is then converted to a decimal number

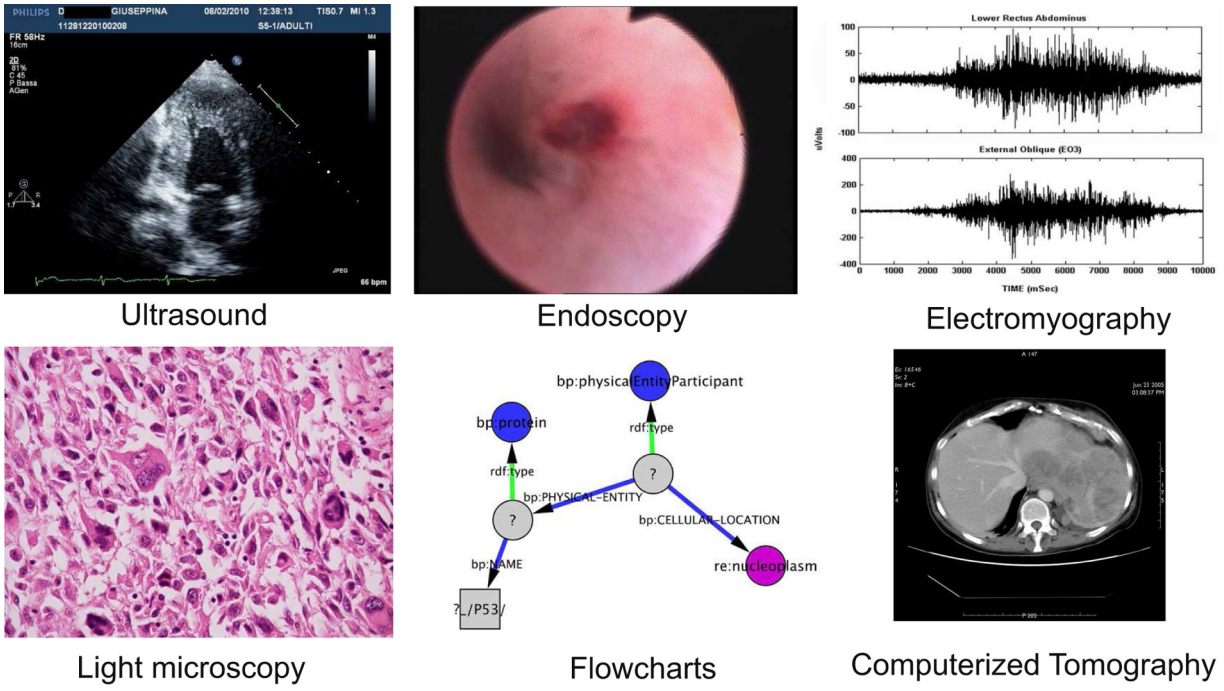

Fig. 2. Sample images from the ImageCLEF 2013 database.

Please cite this article in press as: Dimitrovski I, et al. Improved medical image modality classification using a combination of visual and textual features. Comput Med Imaging Graph (2014), http://dx.doi.org/10.1016/j.compmedimag.2014.06.005

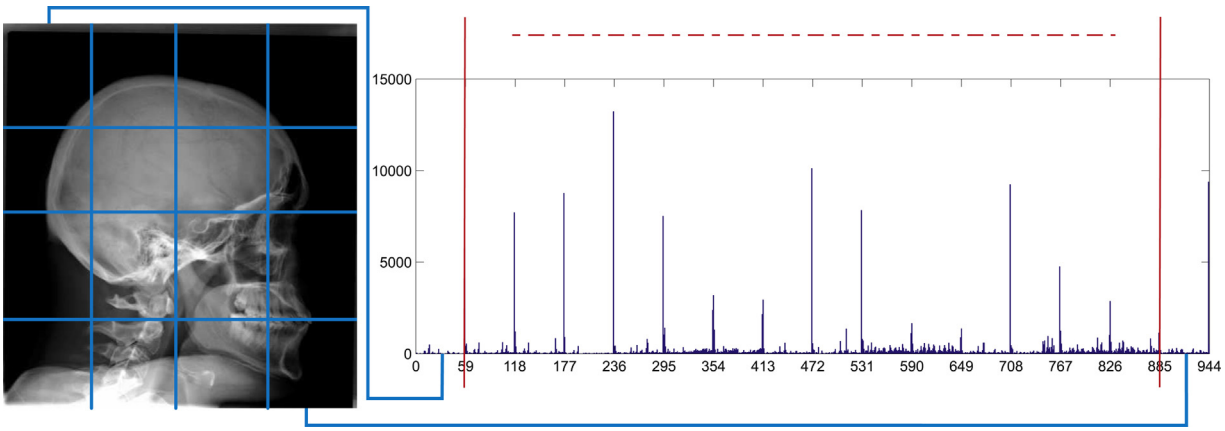

Fig. 3. The image is divided into 4 × 4 non-overlapping sub-images from which LBP histograms are extracted and concatenated into a single, spatially enhanced histogram.

which represents the LBP code. Formally, given a pixel at position (x_c, y_c) the resulting LBP code can be expressed as follows:

$$\mathrm{LPB}_{(P,R)}(x_c, y_c) = \sum_{n=0}^{P-1} S(i_n - i_c)2^n \quad (1)$$

where n ranges over the P neighbors of the central pixel (x_c, y_c), i_c and i_n are the gray-level values of the central pixel and the neighbor pixel, and $S(x)$ is defined as:

$$S(x) = \begin{cases} 1, & \text{if} x \geq 0 \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

The image is traversed with the LBP operator pixel by pixel and the outputs are accumulated into a discrete histogram.

Not all LBP codes are informative. Certain LBP codes capture fundamental properties of the texture and are called uniform patterns because they constitute the vast majority, sometimes over 90 percent, of all patterns present in the observed textures [15]. These patterns have one thing in common, namely, a uniform circular structure that contains very few spatial transitions. They function as templates for micro-structures such as bright spots, flat areas or dark spots.

To spatially enhance the descriptors and improve the performance, it has been suggested to repeatedly sample predefined sub-regions/sub-images of an image (e.g., 1 × 1, 2 × 2, 4 × 4, etc.) [3]. The descriptors from the different sub-images are then aggregated/concatenated into one spatially enhanced descriptor. Following this suggestion, we divide the images into a number of non-overlapping sub-images and concatenate the LBP histograms extracted for each sub-image into a single, spatially enhanced feature histogram. This approach aims at obtaining more local description of the images. Fig. 3 shows how we build the LBP histogram with 16 × 243 = 3888 bins in total for each image. The optimal parameters of the LBP approach, such as neighborhood size, neighborhood radius and spatial pyramids, were determined with a set of extensive experiments. The results of the experiments are outlined in Section 4.2.1.

3.2. Color and edge directivity descriptor

The color and edge directivity descriptor (CEDD) includes color and texture information into a single histogram [16]. The CEDD histogram consists of 6 regions/bins, determined by the texture information. Each region further contains 24 individual regions/bins, emanating from the color information. Consequently, the final histogram includes 6 × 24 = 144 regions/bins.

In order to compute the histogram, the image is divided into image blocks [16], and for each image block the texture and color information is calculated as follows. The texture information is extracted by using the five digital filters (i.e., edges present in a given region) proposed by the MPEG-7 edge histogram descriptor (EHD) [23,24]: vertical, horizontal, 45-degree diagonal, 135-degree diagonal and non-directional edges. Hence, the texture information is represented with a histogram of six bins, five of them corresponding to the types of edges found in the image plus one for no edges of any type found.

The color information is extracted by processing the image blocks in the HSV color space. First, a fuzzy system produces a fuzzy linking histogram that takes the three HSV channels as inputs and creates a histogram with 10 bins as output. Each of the bins represents a preset color: black, gray, white, red, orange, yellow, green, cyan, blue and magenta.

Next, the histogram is expanded into a color histogram with 24 bins by using coordinate logic filters for vertical edge detection as follows. The Hue channel is divided into 8 regions: red to orange, orange, yellow, green, cyan, blue, magenta, and blue to red; Saturation is divided into two fuzzy regions defining the shade of a color based on white; The value channel is divided into three areas: one defines when the block will be black and the other two, in combination with Saturation, when it will be gray. Finally, considering these divisions, a set of 4 fuzzy-like rules are applied transforming the previous 10 color histograms into a 24 color bin histogram comprising black, gray, white, dark red, red, light red, dark orange, orange, light orange, dark yellow, yellow, light yellow, dark green, green, light green, dark cyan, cyan, light cyan, dark blue, blue, light blue, dark magenta, magenta, and light magenta.

Each image block interacts successively with all the fuzzy systems. Defining the bin produced by the texture information fuzzy system as n and the bin produced by the 24-bin fuzzy-linking system as m, each image block is placed in the bin position: $n \times 24 + m$.

3.3. Fuzzy color and texture histogram

Fuzzy color and texture histogram (FCTH) includes the texture information produced in the eight-bin histogram of the fuzzy system that uses the high frequency bands of the Haar wavelet transform [17]. For color information, the descriptor uses the 24-bin color histogram produced by the 24-bin fuzzy-linking system [17]. Overall, the final histogram includes 8 × 24 = 192 regions/bins.

The image is first segmented into a predefined number of image blocks. Each image block is transformed into the YIQ color space, and then transformed with the Haar Wavelet transform. The f_{LH}, f_{HL} and f_{HH} values are calculated and with the use of the fuzzy system

Please cite this article in press as: Dimitrovski I, et al. Improved medical image modality classification using a combination of visual and textual features. Comput Med Imaging Graph (2014), http://dx.doi.org/10.1016/j.compmedimag.2014.06.005

that classifies the f coefficients, this image block is classified into one of the eight output bins.

Next, the same image block is transformed into the HSV color space and the mean H, S and V block values are calculated. These values become inputs to the fuzzy system that forms the ten-bin fuzzy color histogram. Then, the next fuzzy system uses the mean values of S and V as well as the position number of the bin (or bins) resulting from the previous fuzzy ten-bin unit, to calculate the hue of the color and create the fuzzy 24-bin histogram. The process is repeated for all the blocks of the image. This descriptor is quite similar to the CEDD, but it captures the texture information through the Haar Wavelet transform.

3.4. Scale-invariant feature transform descriptors

The scale-invariant feature transform (SIFT) image descriptors employ the bag of features approach commonly used in many state-of-the-art approaches in image classification [3,25]. The basic idea of this approach is to sample a set of local image patches using some method (densely, randomly or using a key-point detector) and calculate a visual descriptor on each patch (SIFT descriptor, normalized pixel values). The resulting set of descriptors is then matched against a pre-specified visual codebook, which converts it to a histogram. The main issues that need to be considered when applying this approach are: sampling of patches, selection of visual patch descriptors and building a visual codebook.

We use dense sampling of the patches, which samples an image grid in a uniform fashion using a fixed pixel interval between patches. We use an interval distance of 6 pixels and sample at multiple scales ($\sigma = 1.2$ and $\sigma = 2.0$ for the Gaussian filter [19]). Due to the low contrast of some of the medical images (e.g., radiographs), it would be difficult to use any detector for points of interest. We calculate SIFT and opponentSIFT (OSIFT) descriptors for each image patch [18,19,26]. OpponentSIFT describes all the channels in the opponent color space using SIFT descriptors. The information in the O_3 channel is equal to the intensity information, while the other channels describe the color information in the image. These other channels do contain some intensity information, but due to the normalization of the SIFT descriptor they are invariant to changes in light intensity [19].

The crucial aspects of a codebook representation are the codebook construction and assignment. An extensive comparison of codebook representation variables is given by van Gemert et al. [27]. We employ k-means clustering (a custom implementation in the C programming language was used) on 250K randomly chosen descriptors from the set of images available for training. k-means partitions the visual feature space by minimizing the variance between a predefined number of k clusters. Here, we set k to 1000 and thus define a codebook with 1000 codewords [22].

3.5. Textual features

The images from the ImageCLEF 2013 database are taken from medical articles and can be indexed using the surrounding text content (article title, article abstract, full text or image captions). The text representation adopted in this work includes information from the title of the article and the image caption, which can be found in the XML representation of the corresponding article. This constitutes the text corpus for the image collection. Next, standard text processing operations are applied, including tokenization, stemming, and stop-word removal using Terrier IR [20], which is a high performance and scalable information retrieval platform. We calculate the weight for each term in each medical article using the standard TF-IDF weighting model. The calculated weights were $L2$ normalized and used as textual features for the corresponding images.

4. Experimental setup

In this section, we present the experimental setup we used to evaluate the proposed feature extraction methods for medical image modality classification. First, we present the learning algorithm/classifier that was used in our experiments. Next, we give the procedure for the selection of the optimal spatial resolution of the visual descriptors. We then define the feature fusion schemes used to improve the predictive performance of the image descriptions. Finally, we state the experimental questions that we investigate in this study.

4.1. Classifier setup

For classification, we used the libSVM implementation of support vector machines (SVMs) [28] with probabilistic output [29]. To solve the multi-class classification problems, we employ the *one-vs-all* approach. Namely, we build a binary classifier for each modality/class: the examples associated with that class are labeled positive and the remaining examples are labeled negative. This results in an imbalanced ratio of positive versus negative training examples. We resolve this issue by adjusting the weights of the positive and negative class [19]. In particular, we set the weight of the positive class to $((\#pos + \#neg)/\#pos)$ and the weight of the negative class to $((\#pos + \#neg)/\#neg)$, with $\#pos$ the number of positive instances and $\#neg$ the number of negative instances in the train set.

The nature of the visual descriptors and the textual descriptors differs significantly: The textual features are very sparse. This requires that a different classifier is used to compensate for this. Consequently, SVMs for the visual features as examples were trained with a χ^2 kernel. For the textual features, we used SVMs with a precomputed kernel obtained using the cosine similarity as a distance measure over the $L2$ normalized tf-idf weights. We optimize the cost parameter C of the SVMs using an automated parameter search procedure. For the parameter optimization, we separate 20% of the training set and use it as validation set. After finding the optimal value for C, the SVM is finally trained on the whole set of training images and evaluated on the test images.

To assess the performance of the classifiers, we use the overall recognition rate/accuracy. This is a very common and widely used evaluation measure. It is calculated as the fraction of the test images whose class/modality was predicted correctly.

4.2. Parametrization of the feature extraction methods

Preliminary experiments were performed to select the optimal parameters for the feature extraction methods. These include the parameters P (number of neighbors) and R (radius of neighborhood) in LBPs and number of subimages/spatial pyramids for the other descriptors. The experiments were performed in the following fashion.

We generated visual descriptors for the train and test images using different number of sub-images (grid layout) and different number of neighbors and different values for the radius of the neighborhood for LBPs. As in the other experiments, we optimized the C parameter in the SVMs by separating 20% of the training set and use it as validation set. After finding the optimal value for C, the SVM was finally trained on the whole set of training images and evaluated on the test images. The optimal number of sub-images for each visual descriptor was selected using the evaluation results. The descriptors with the selected optimal spatial pyramids were included in the concatenated visual descriptor. In the remainder,

Please cite this article in press as: Dimitrovski I, et al. Improved medical image modality classification using a combination of visual and textual features. Comput Med Imaging Graph (2014), http://dx.doi.org/10.1016/j.compmedimag.2014.06.005

Table 2
Predictive performance of the classifiers learned from *LBP* descriptors using different patterns for the ImageCLEF 2011, 2012 and 2013 databases. The computation time per image for the different patterns is presented in the last column. The time is calculated as average over 20 random images.

LBP pattern	2011	2012	2013	Comp. time per image (s)
$LBP^{u2}_{8,1}$	66.01	37.00	59.14	0.08
$LBP^{u2}_{16,2}$	69.53	45.20	62.31	0.17
$LBP^{u2}_{24,2}$	69.62	46.30	64.56	0.78
$LBP^{u2}_{24,3}$	70.41	47.10	64.32	0.79

we briefly outline the results of the preliminary experiments for each feature extraction method.

4.2.1. Parametrization of local binary patterns

In our experiments, we used the patterns $LBP^{u2}_{16,2}$, where the superscript $u2$ reflects the use of uniform patterns that have a U value of at most 2 on a neighborhood of size 16 and radius 2. The U value is the number of spatial transitions (bitwise 0/1 changes) in the pattern. The non-uniform patterns (patterns that have U value larger than 2) are grouped under one bin in the resulting histogram. With the $LBP^{u2}_{16,2}$ operator, the number of bins in the histogram is 243 (242 bins for uniform patterns and one bin for non-uniform/noisy patterns). The $LBP^{u2}_{16,2}$ patterns were selected based on preliminary experiments involving the use of several different types of LBP patterns. The results from these experiments are presented in Table 2.

Increasing the number of neighborhood pixels increases the performance of the LBP descriptor: For example, increasing the number from 8 to 16 increases the performance by 3% to 8% over the three different databases. Further increase of the number of local neighbors in the pattern to 24 increases the performance slightly (on average 1%) at the price of a significant increase of the dimensionality of the descriptors (up to 555). More importantly, the efficiency of the procedure for calculation of the descriptor is severely affected (the processing time per image is increased slightly more than 4 times). Because all of this, we have selected the $LBP^{u2}_{16,2}$ pattern for use in our further experiments.

To spatially enhance the descriptors and improve the performance, it has been suggested to repeatedly sample predefined sub-regions/sub-images of an image (e.g., 1×1, 2×2, 4×4, etc.). The optimal number of sub-images was found by extensive experiments using different numbers of sub-images. The results are presented in Table 3. From the results, we can see that increasing the number of sub-images over 16 results in a slight decrease of predictive performance. Considering all of this, we selected to use $LBP^{u2}_{16,2}$ pattern coupled with 4×4 spatial pyramid.

4.2.2. Parametrization of color and edge directivity descriptor

To further spatially enhance the descriptors and improve the performance, we applied the same technique as for the LBP descriptor: the images are divided into non-overlapping sub-images and the CEDDs for each sub-image are concatenated into a single feature histogram (see Fig. 3). To select the optimal number, we have conducted experiments with different numbers of sub-images, starting from one (entire image) and going to 4, 9, 16, 25, 36, 49, 64 sub-images. The results are presented in Table 4. The optimal number of sub-images is 36 (grid layout 6×6). The total number of bins is therefore $6 \times 6 \times 144 = 5184$.

Table 4
Predictive performance of the classifiers learned from *CEDD* descriptors using different numbers of sub-images for the ImageCLEF 2011, 2012 and 2013 databases.

#Sub-images	2011	2012	2013
1×1	68.84	38.90	58.28
2×2	69.33	44.70	61.69
3×3	72.16	50.10	68.51
4×4	72.36	50.30	68.00
5×5	73.73	53.20	68.12
6×6	73.92	52.60	68.86
7×7	74.12	53.70	68.74
8×8	74.60	53.60	69.51

4.2.3. Parametrization of fuzzy color and texture histogram

For this descriptor, the images are also divided into 6×6 non-overlapping sub-images to obtain a spatially enhanced feature histogram (as for LBP and CEDD, illustrated in Fig. 3). Hence, the total number of bins in this case is $6 \times 6 \times 192 = 6912$ (36 sub-images with 192 bins each). For this descriptor, we have also conducted extensive experiments with different numbers of sub-images in order to select the optimal number. The results are presented in Table 5.

4.2.4. Parametrization of scale-invariant feature transform descriptors

Dense sampling gives an equal weight to all key-points, irrespective of their spatial location in the image. To overcome this limitation, we follow the spatial pyramid approach proposed by Lazebnik et al. [30]. The image is repeatedly sampled into fixed sub-regions, e.g., 1×1, 2×2, 4×4, etc., and the different resolutions are aggregated into a so called spatial pyramid. Since every region is an image, the spatial pyramid can easily be used in combination with dense sampling. For the ideal spatial pyramid configuration, Lazebnik et al. [30] claim that 2×2 is sufficient, while Marszałek et al. [31] suggest to also include 1×3. We investigate multiple divisions of the image in our experiments. The results from the experiments are presented in Table 6. For the Image CLEF 2011 dataset, the best results are obtained by using the histogram from the entire image. Inclusion of spatial pyramids results in lower predictive performance for both the SIFT and OSIFT descriptors. For the Image CLEF 2012 and 2013 datasets, the inclusion of the spatial pyramids results in increased predictive performance of the classifiers for both local descriptors. Here, we have chosen to use the histogram that includes the spatial pyramids. The resulting vector in this case is with 8000 bins ($(1 \times 1 + 2 \times 2 + 1 \times 3) \times 1000$), and was obtained by concatenation of the eight histograms. Fig. 4 shows an

Table 3
Predictive performance of the classifiers learned from *LBP* descriptors using different numbers of sub-images for the ImageCLEF 2011, 2012 and 2013 databases.

#Sub-images	2011	2012	2013
1×1	69.53	45.20	62.31
2×2	68.65	48.70	65.91
3×3	69.92	50.50	69.86
4×4	70.01	51.70	69.44
5×5	69.43	51.40	69.86
6×6	68.65	52.00	69.86
7×7	68.65	51.40	69.86
8×8	67.73	51.40	69.86

Table 5
Predictive performance of the classifier learned from *FCTH* descriptors using different numbers of sub-images for the ImageCLEF 2011, 2012 and 2013 databases.

#Sub-images	2011	2012	2013
1×1	67.57	36.20	58.63
2×2	68.06	40.09	60.10
3×3	71.28	46.10	60.99
4×4	72.16	47.70	61.81
5×5	72.36	48.90	63.32
6×6	73.04	49.60	64.67
7×7	73.33	48.70	65.18
8×8	73.63	49.80	64.75

Please cite this article in press as: Dimitrovski I, et al. Improved medical image modality classification using a combination of visual and textual features. Comput Med Imaging Graph (2014), http://dx.doi.org/10.1016/j.compmedimag.2014.06.005

Table 6
Predictive performance of the classifiers learned from *SIFT* and *OSIFT* descriptors using different spatial pyramid configurations for the ImageCLEF 2011, 2012 and 2013 databases.

Spatial pyramid	2011		2012		2013	
	SIFT	OSIFT	SIFT	OSIFT	SIFT	OSIFT
1×1	81.15	83.69	66.30	66.60	75.43	76.21
2×2	79.98	81.25	64.60	65.90	75.31	77.21
$1 \times 1, 2 \times 2$	80.46	81.73	65.90	66.40	75.78	77.68
$1 \times 1, 2 \times 2, 3 \times 1$	80.95	82.00	66.50	67.30	76.36	78.10

example of the histograms extracted from an image for the spatial pyramids of 1×1, 2×2 and 1×3.

4.3. Feature fusion schemes

Several studies have shown that, taken together, various visual features bringing different information about the visual content of the images clearly outperform single feature approaches [32,22,14]. Furthermore, the textual features also bring information that can be exploited by the classifiers. Following these findings, we combine the different visual and textual features described above.

We use the two different feature fusion schemes depicted in Fig. 5: low level (LL) and high level (HL). For the low level feature fusion scheme, the descriptors are concatenated into a single feature vector and a classifier is trained on the joint feature vector. The high level fusion scheme averages the predictions from the individual classifiers trained on the separate descriptors.

The low-level feature fusion scheme is used for the different visual features because it performs slightly better than high-level feature fusion for medical image annotation [14]. Because of the different nature of the visual and the textual descriptors, we learn classifiers for the concatenated visual features and the textual

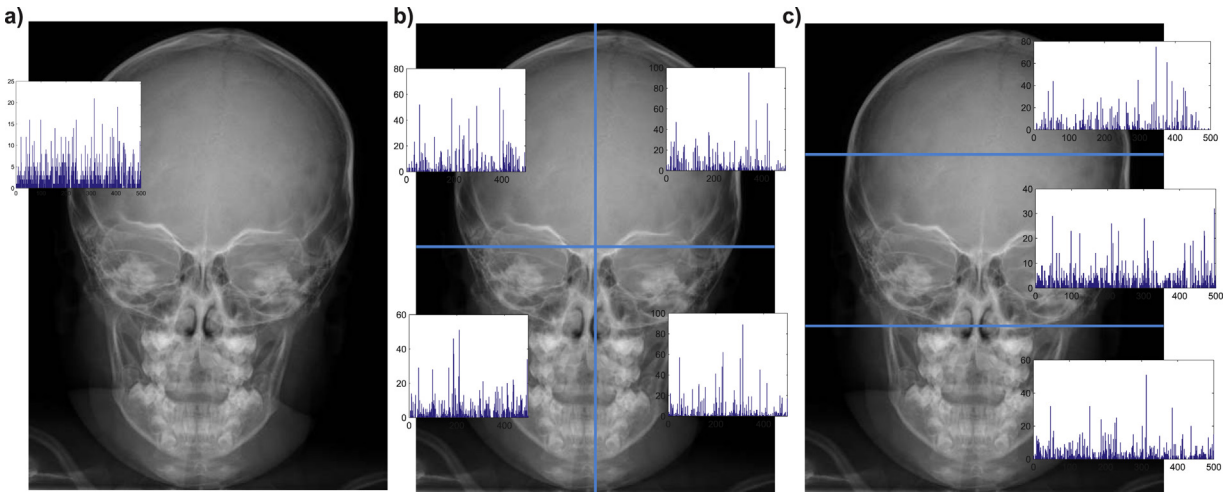

Fig. 4. Three different spatial pyramids used in our experiments, (a) 1×1, (b) 2×2 and (c) 1×3. The spatial pyramid constructs feature vectors for each of the specific part of the image.

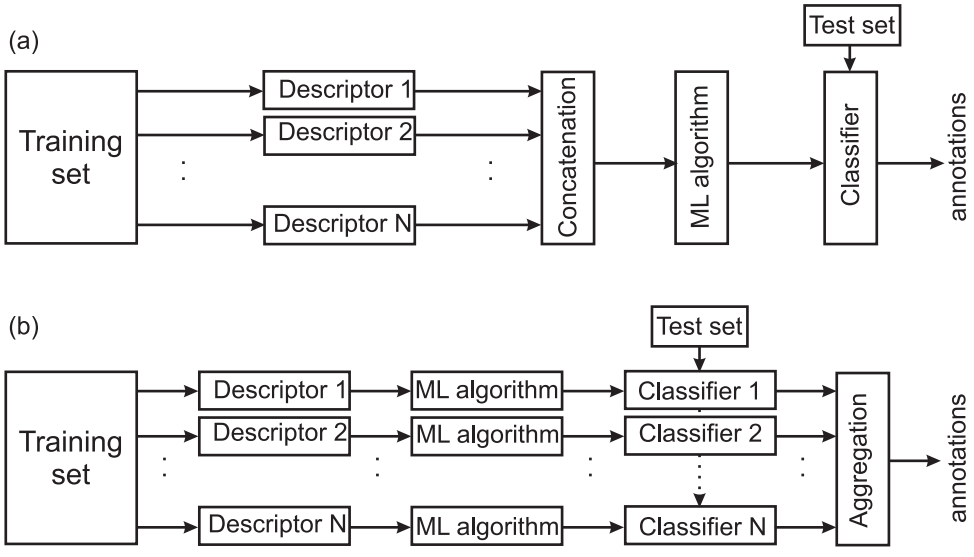

Fig. 5. Low level (a) and high level feature fusion schemes for different descriptors.

Please cite this article in press as: Dimitrovski I, et al. Improved medical image modality classification using a combination of visual and textual features. Comput Med Imaging Graph (2014), http://dx.doi.org/10.1016/j.compmedimag.2014.06.005

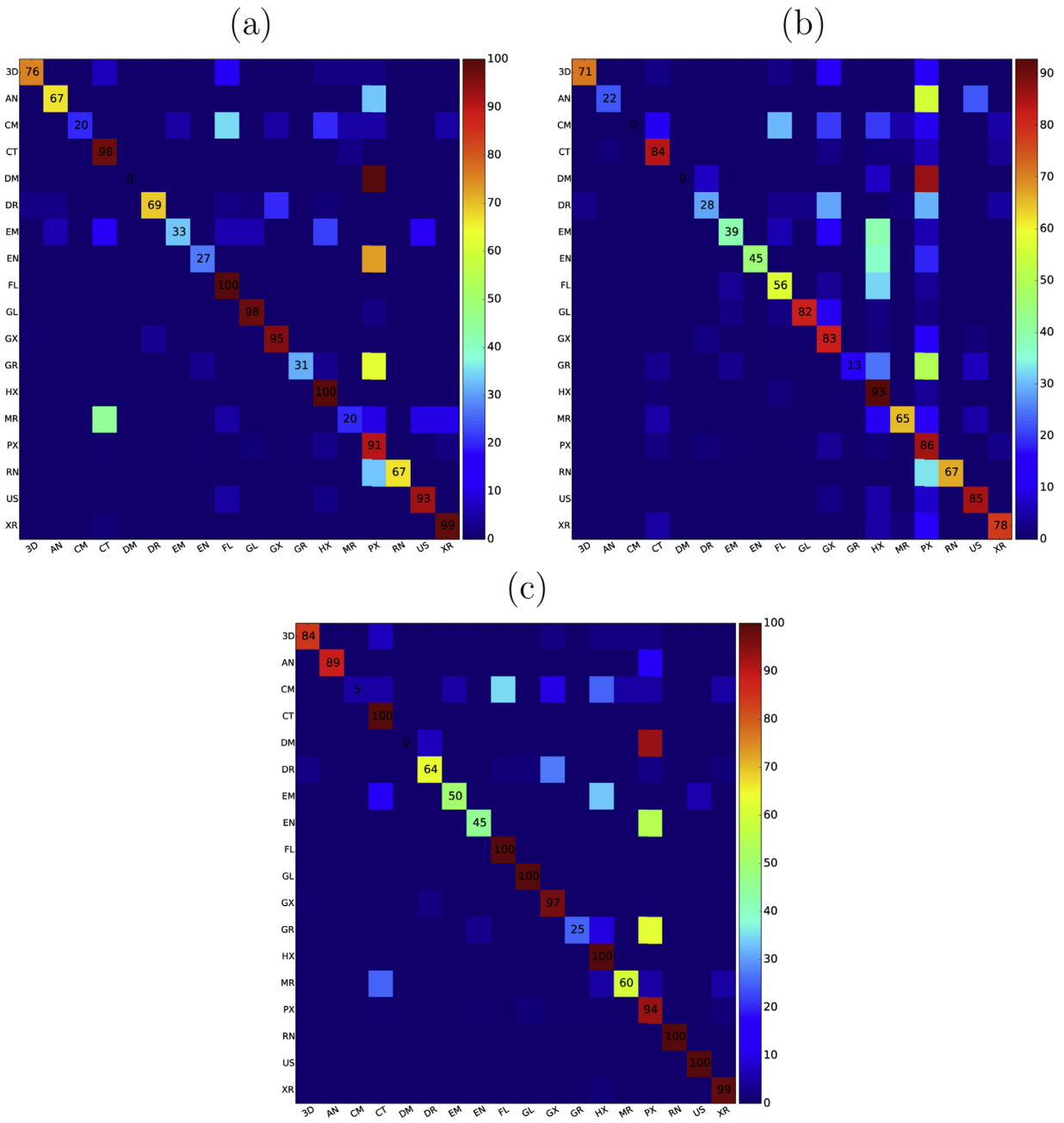

Fig. 6. Confusion matrix for the ImageCLEF 2011 dataset obtained using: (a) concatenated visual features, (b) textual features, and (c) a high level fusion scheme over the concatenated visual features and the textual features.

features separately, the final predictions are obtained by averaging the individual predictions from the two classifiers. The weight for the predictions of the classifier which is using the visual features was set to 0.5 and the weight for the predictions of the classifier learned with textual features was set also to 0.5. These weights were determined by using the optimization procedure described above.

4.4. Experimental questions

The goal of the experimental evaluation is to investigate the following research questions:

1. Which visual descriptor is most suitable for medical modality classification?
2. Does combining multiple visual features improve the predictive performance?
3. Which features are better for medical modality classification: visual features or text features?
4. Does the use of text features in fusion with visual features increase the predictive performance?

To find out which feature extraction technique is most suitable for medical modality classification (question 1), we compare the predictive performance figures of the classifiers constructed using each type of visual descriptor separately. We then investigate whether the combination of feature extraction techniques can increase the predictive performance of the constructed classifiers (question 2). For the third question, we compare the performance of the classifier constructed using the best visual descriptor to the

Please cite this article in press as: Dimitrovski I, et al. Improved medical image modality classification using a combination of visual and textual features. Comput Med Imaging Graph (2014), http://dx.doi.org/10.1016/j.compmedimag.2014.06.005

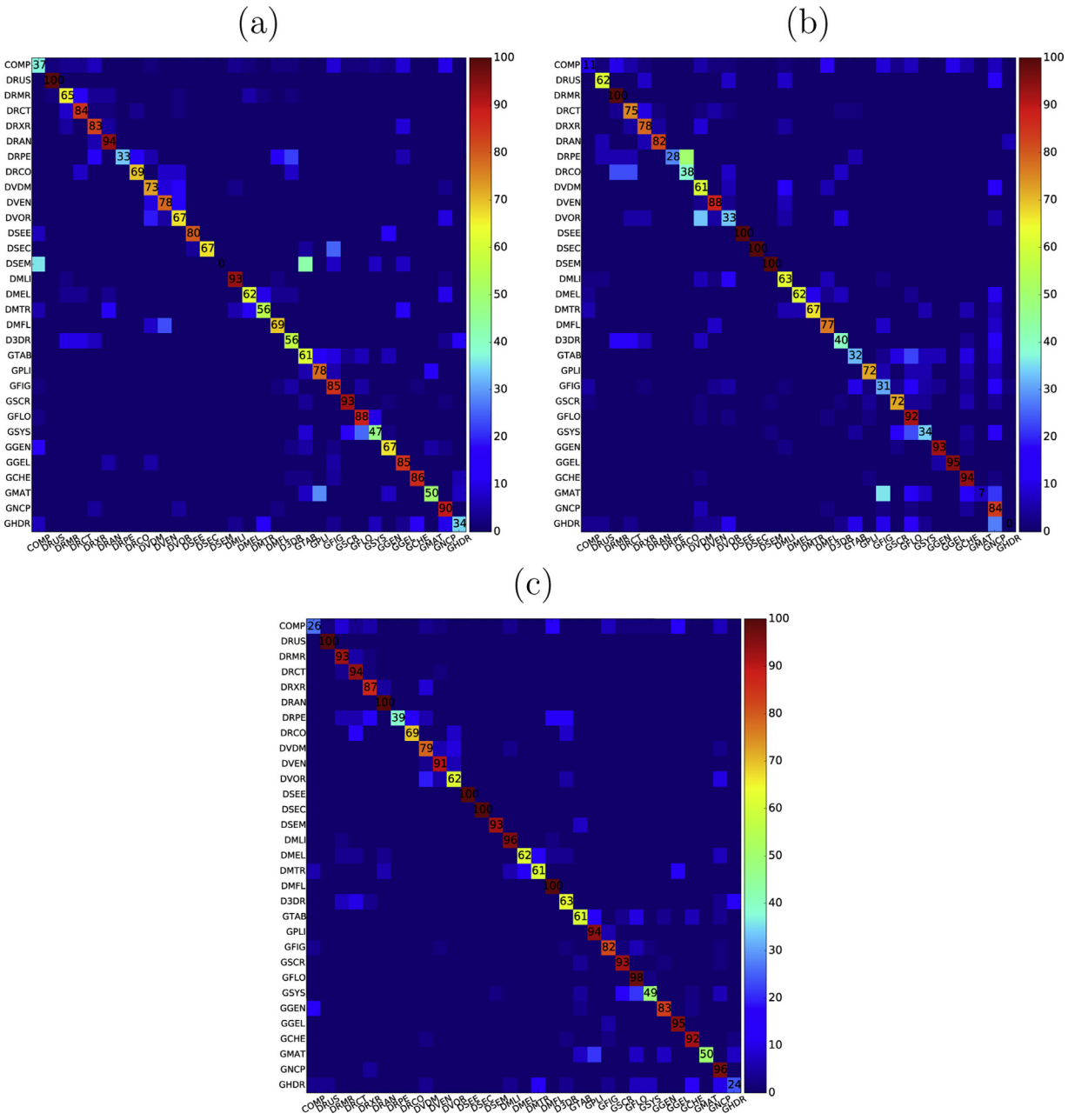

Fig. 7. Confusion matrix for the ImageCLEF 2012 dataset obtained using: (a) concatenated visual features, (b) textual features, and (c) a high level fusion scheme over the concatenated visual features and the textual features.

performance of the classifier constructed using the textual features. Finally, for the last question, we compare the predictive performance of the classifiers constructed without and with text features.

5. Results and discussion

In this section, we present and discuss the results from the experimental evaluation of the various feature extraction techniques in the context of medical image modality classification. We first explore the performance of the different descriptors across the three databases. Next, we compare the best performance obtained with our descriptors with the best results reported on this database.

The main results are summarized in Table 7. We begin by close inspection of the performance of the visual descriptors across the three image databases (first five rows in Table 7). We can note that the SIFT descriptors (SIFT and OSIFT) offer much better

Table 7
Predictive performance of the classifiers learned from descriptors produced by different feature extraction algorithms and their combinations for the ImageCLEF 2011, 2012 and 2013 databases. The visual features are concatenated using low level feature fusion (LLF), while visual+textual denotes the predictive performance of the high level feature fusion (HLF) approach applied to the fused visual descriptors and the textual features.

	2011	2012	2013
LBP	70.01	51.70	69.44
FCTH	73.04	49.60	64.67
CEDD	73.92	52.60	68.86
SIFT	81.15	66.50	76.36
OSIFT	83.69	67.30	78.10
Visual features (LLF)	84.86	71.20	80.31
Textual features	72.65	63.80	63.88
Mixed (visual+textual) (HLF)	87.10	78.60	82.25

Please cite this article in press as: Dimitrovski I, et al. Improved medical image modality classification using a combination of visual and textual features. Comput Med Imaging Graph (2014), http://dx.doi.org/10.1016/j.compmedimag.2014.06.005

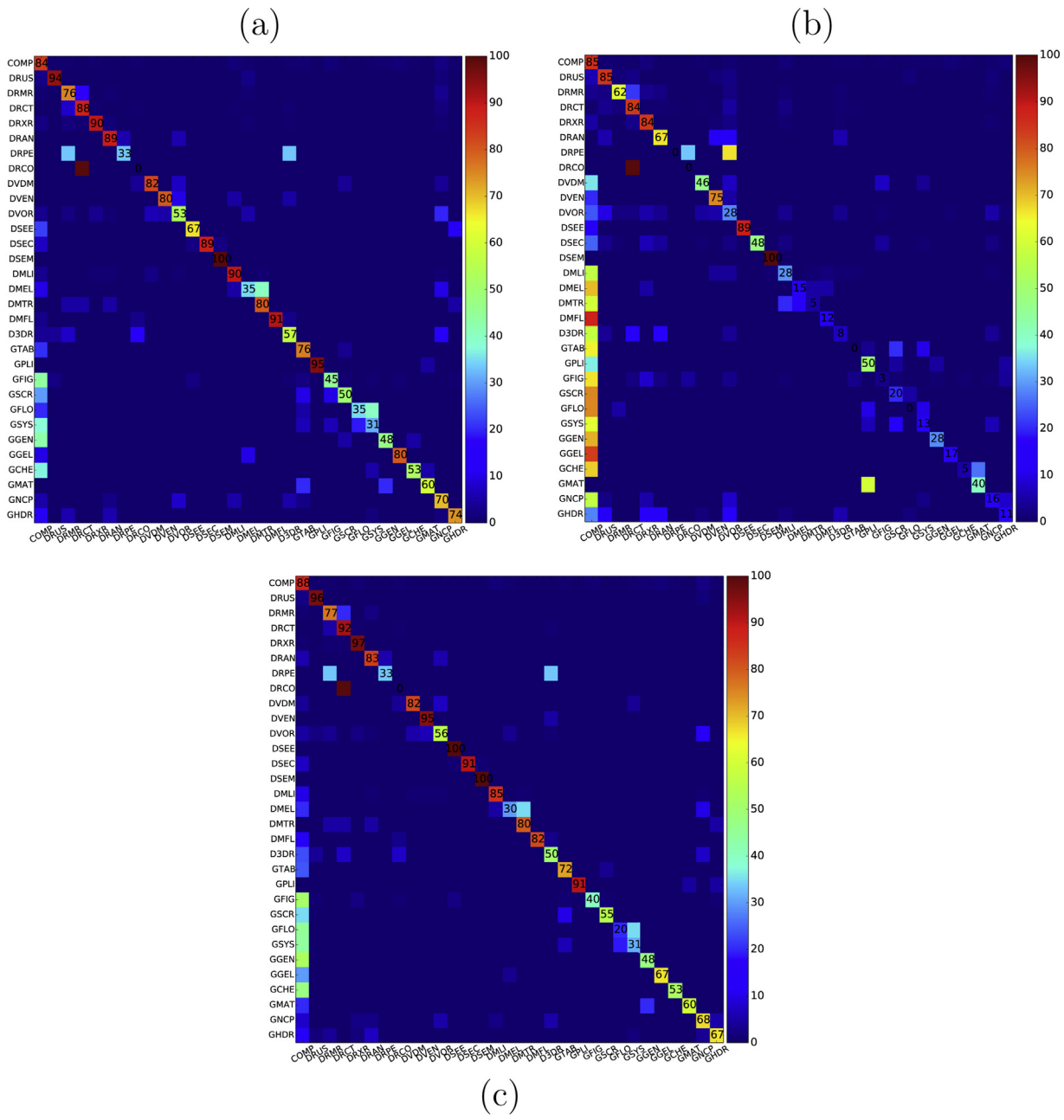

Fig. 8. Confusion matrix for the ImageCLEF 2013 dataset obtained using: (a) concatenated visual features, (b) textual features, and (c) a high level fusion scheme over the concatenated visual features and the textual features.

predictive performance than the other descriptors. Namely, the SIFT descriptors are able to capture specific details from the images and are robust to noise, illumination, scale, translation and rotation changes. In particular, the opponentSIFT descriptor is the one that exhibits the best predictive performance. The inclusion of color information in the calculation of the SIFT descriptor through the opponent color space consistently contributes a 1%-2% improvement in predictive performance as compared to the regular SIFT descriptor (computed over gray scale images).

Next, the concatenation of the visual features further increases the predictive performance of the classifier by 2%-3%. In other words, the inclusion of more than one type of visual features in the classification process contributes to a better representation of the visual content of the images. We can note that the majority of the information was already represented by both of the SIFT descriptors, while the inclusion of the other descriptors helped to slightly improve the performance.

The textual descriptors have relatively good predictive performance. In some case (e.g., for the ImageCLEF 2012 database), they are even better than the CEDD, FCTH and LBP descriptors. While they are always worse than SIFT and opponent SIFT, the textual features carry important information about image modality, which could be exploited to achieve even better performance.

Furthermore, we inspect the performance of all descriptors combined (visual and textual), given in the last row of Table 7. These descriptors offer the best predictive performance. The increase in performance is particularly evident for the ImageCLEF 2012 database. Namely, taken together, all descriptors have an accuracy of 78.60%, which is better by 7.4 percentage points than the results using only the concatenated visual descriptors. For the other two

Please cite this article in press as: Dimitrovski I, et al. Improved medical image modality classification using a combination of visual and textual features. Comput Med Imaging Graph (2014), http://dx.doi.org/10.1016/j.compmedimag.2014.06.005

Table 8
Detailed performance per class for the 2011 dataset.

Description	Class code	Images		Accuracy		
		#Train	#Test	Visual	Text	Mixed
Electron microscopy	EM	16	18	33.33	38.89	50.00
Histopathology	HX	208	195	100.00	92.82	100.00
Dermatology	DM	7	15	0.00	0.00	0.00
Gross pathology	GR	43	32	31.25	12.50	25.00
Compound figure	CM	17	20	20.00	0.00	5.00
Fluoresence	FL	43	28	100.00	57.14	100.00
Graphs	GX	161	172	95.35	82.56	96.51
Ultrasound	US	30	41	92.68	85.37	100.00
Angiography	AN	11	9	66.67	22.22	88.89
Gel	GL	50	50	98.00	82.00	100.00
Endoscopic imaging	EN	10	11	27.27	45.45	45.45
Magnetic resonance imaging	MR	16	20	20.00	65.00	60.00
X-ray	XR	59	67	98.51	77.61	98.51
Retinography	RN	5	3	66.67	66.67	100.00
3D reconstruction	3D	32	45	75.56	71.11	84.44
Drawing	DR	43	74	68.92	28.38	63.51
General photo	PX	166	141	90.78	85.82	93.62
Computed tomography	CT	71	83	97.59	84.34	100.00

datasets, the performance increase is a bit smaller: 2.24 percentage points for the ImagelCLEF 2011 and 1.94 percentage points for the ImageCLEF 2013 database.

The largest increase in performance for the ImageCLEF 2012 database can be explained if we look at the confusion matrix from the classifier evaluation, shown in Figs. 6–8 and the detailed performance per modality given in Tables 8–10 for the ImageCLEF 2011, ImageCLEF 2012 and ImageCLEF 2013, respectively. Let us further focus on the confusion matrices for the visual features (Fig. 7(a)) and the textual features (Fig. 7(b)). We can note the interaction of the two types of features. First, for the classes denoted with DSEM, DSEC, DSEE and DRMR, the textual features clearly help to lift the overall predictive performance. Second, for the classes denoted with DMFL, GPLI, D3DR, GTAB, DRCT, DRCO and GNCP, the visual and the textual features offer complementary information to the classifier. This is also evident from the per-feature-type performances: the performance of the mixed feature set for most of the classes is better than any of the individual feature sets. Finally, for the classes denoted with GFIG, GMAT, DRUS, GHDR, DVOR and DMLI, the visual features offer much better information about the visual content to the classifier.

If we make a similar comparison for the databases from ImageCLEF 2011 and ImagleCLEF 2013, we can note that the number of classes that convey complementary information is smaller as

Table 9
Detailed performance per class for the 2012 dataset.

Description	Class code	Images		Accuracy		
		#Train	#Test	Visual	Text	Mixed
Tables and forms	GTAB	38	31	61.29	32.26	61.29
Fluorescence microscopy	DMFL	21	13	69.23	76.92	100.00
Chromatography, Gel	GGEL	49	20	85.00	95.00	95.00
Statistical figures, graphs and charts	GFIG	48	61	85.25	31.15	81.97
Other organs	DVOR	48	21	66.67	33.33	61.90
Chemical structure	GCHE	21	50	86.00	94.00	92.00
Light microscopy	DMLI	46	46	93.48	63.04	95.65
Angiography	DRAN	38	17	94.12	82.35	100.00
Screenshots	GSCR	40	54	92.59	72.22	92.59
Endoscopy	DVEN	32	32	78.13	87.50	90.63
Hand-drawn sketches	GHDR	17	29	34.48	0.00	24.14
Gene sequence	GGEN	47	42	66.67	92.86	83.33
System overviews	GSYS	48	47	46.81	34.04	48.94
Compound or multipane images	COMP	49	57	36.84	10.53	26.32
Ultrasound	DRUS	48	13	100.00	61.54	100.00
Combined modalities in one image	DRCO	12	13	69.23	38.46	69.23
Electromyography	DSEM	5	14	0.00	100.00	92.86
Program listing	GPLI	10	18	77.78	72.22	94.44
Electroencephalography	DSEE	6	15	80.00	100.00	100.00
Mathematics, formulae	GMAT	6	14	50.00	7.14	50.00
Electrocardiography	DSEC	5	24	66.67	100.00	100.00
X-ray, 2D radiography	DRXR	48	23	82.61	78.26	86.96
Transmission microscopy	DMTR	29	18	55.56	66.67	61.11
Flowcharts	GFLO	48	50	88.00	92.00	98.00
Dermatology, skin	DVDM	47	33	72.73	60.61	78.79
Electron microscopy	DMEL	22	29	62.07	62.07	62.07
Computerized tomography	DRCT	49	64	84.38	75.00	93.75
PET	DRPE	9	18	33.33	27.78	38.89
Non-clinical photos	GNCP	47	49	89.80	83.67	95.92
Magnetic resonance	DRMR	43	55	65.45	100.00	92.73
3D reconstructions	D3DR	25	30	56.67	40.00	63.33

Please cite this article in press as: Dimitrovski I, et al. Improved medical image modality classification using a combination of visual and textual features. Comput Med Imaging Graph (2014), http://dx.doi.org/10.1016/j.compmedimag.2014.06.005

Table 10
Detailed performance per class for the 2013 dataset.

Description	Class code	Images		Accuracy		
		#Train	#Test	Visual	Text	Mixed
Tables and forms	GTAB	65	29	75.86	0.00	72.41
Fluorescence microscopy	DMFL	33	33	90.91	12.12	81.82
Chromatography and Gel	GGEL	55	30	80.00	16.67	66.67
Statistical figures, graphs and charts	GFIG	102	102	45.10	2.94	40.20
Other organs	DVOR	70	92	53.26	28.26	56.52
Chemical structure	GCHE	62	19	52.63	5.26	52.63
Light microscopy	DMLI	91	121	90.08	28.93	85.12
Angiography	DRAN	54	18	88.89	66.67	83.33
Screenshots	GSCR	91	20	50.00	20.00	55.00
Endoscopy	DVEN	64	20	80.00	75.00	95.00
Hand-drawn sketches	GHDR	46	54	74.07	11.11	66.67
Gene sequence	GGEN	68	21	47.62	28.57	47.62
System overviews	GSYS	89	16	31.25	12.50	31.25
Compound or multipane images	COMP	1105	1014	84.22	85.21	87.67
Ultrasound	DRUS	60	85	94.12	84.71	96.47
Combined modalities in one image	DRCO	22	1	0.00	0.00	0.00
Electromyography	DSEM	18	1	100.00	100.00	100.00
Program listing	GPLI	28	22	95.45	50.00	90.91
Electroencephalography	DSEE	21	9	66.67	88.89	100.00
Mathematics and formulae	GMAT	20	5	60.00	40.00	60.00
Electrocardiography	DSEC	29	96	88.54	47.92	90.63
X-ray, 2D radiography	DRXR	70	344	90.12	84.01	96.80
Transmission microscopy	DMTR	46	20	80.00	5.00	80.00
Flowcharts	GFLO	94	20	35.00	0.00	20.00
Dermatology and skin	DVDM	79	28	82.14	46.43	82.14
Electron microscopy	DMEL	51	20	35.00	15.00	30.00
Computerized tomography	DRCT	113	186	87.63	83.87	92.47
PET	DRPE	16	3	33.33	0.00	33.33
Non-clinical photos	GNCP	96	37	70.27	16.22	67.57
Magnetic resonance	DRMR	97	90	75.56	62.22	76.67
3D reconstructions	D3DR	46	26	57.69	7.69	50.00

Table 11
The best performances on the medical modality classification task from ImageCLEF 2011, 2012, 2013. The competition entries are divided by type (visual only, textual only, mixed) and compared to our best results presented in this paper.

	Group	Type	Accuracy	Our best
2011	XRCE [21]	Mixed	86.91	87.10
	XRCE [21]	Visual	83.59	84.86
	IPL [21]	Textual	70.41	72.65
2012	medGIFT [10]	Mixed	66.20	78.60
	IBM Multimedia Analytics [10]	Visual	69.60	71.20
	ITI [10]	Textual	41.30	63.80
2013	IBM Multimedia Analytics [3]	Mixed	81.68	82.25
	IBM Multimedia Analytics [3]	Visual	80.79	80.31
	IBM Multimedia Analytics [3]	Textual	64.17	63.88

compared to the one for ImageCLEF 2012. Moreover, the extent of the increase of predictive performance by exploiting the complementary information is smaller for ImageCLEF 2011 and ImageCLEF 2013. Nonetheless, using these results we can further investigate the performance of our approach and identify possible directions for further improvements.

Finally, we compare the best performance obtained by using our approach with the best reported results for the ImageCLEF 2011, ImageCLEF 2012 and ImageCLEF 2013 competitions (given in Table 11). The main conclusion from the comparison is that the results obtained with our approach are better than the best results reported at the competitions. The largest difference is for the ImageCLEF 2012 database and it amounts to 12.4 percentage points in accuracy. The difference is mainly due to the more superior exploitation of the textual features: Our textual features achieve 63.80% accuracy, while the best result with textual features reported so far was only 41.3% [10]. This is probably because the competing results for the textual features were obtained by first performing dimensionality reduction and then training a different classifier than ours. We could further improve the performance of our approach by including even more different features, both visual and textual.

6. Conclusions

In this paper, we present in detail the approach that we used for the medical modality classification task at the ImageCLEF evaluation forum for the cross-language annotation and retrieval of images.

We evaluated several types of visual and textual features, and combinations thereof. More specifically, we used the modality classification databases from the ImageCLEF competitions in 2011, 2012 and 2013. We used LBP, FCTH, CEDD, SIFT and opponentSIFT as visual features and a standard bag-of-words textual representation coupled with TF-IDF weighting.

The results from the experiments reveal that the best performing features for modality classification are the SIFT and opponentSIFT features. Next, the low-level fusion of the visual features slightly improves the predictive performance of the classifiers. This is because the different features are able to capture

Please cite this article in press as: Dimitrovski I, et al. Improved medical image modality classification using a combination of visual and textual features. Comput Med Imaging Graph (2014), http://dx.doi.org/10.1016/j.compmedimag.2014.06.005

different aspects of an image, and thus, their combination offers a more complete representation of the visual content of an image. Furthermore, we investigate adding textual features: Due to the different nature of the visual and textual features and the sparsity of the textual features, these are combined with high-level fusion. This further increases the predictive performance by a significant amount. Finally, the results obtained with our approach are the best results reported thus far on these challenging databases.

Acknowledgements

We would like to acknowledge the support of the European Commission through the project MAESTRA - Learning from Massive, Incompletely annotated, and Structured Data (Grant number ICT-2013-612944).

References

[1] Choplin R, Boehme J, Maynard C. Picture archiving and communication systems: an overview. Radiographics 1992;12:127–9.
[2] Becker S, Arenson R. Costs and benefits of picture archiving and communication systems. J Am Med Inform Assoc 1994;1:361–71.
[3] de Herrera AGS, Kalpathy-Cramer J, Fushman DD, Antani S, Müller H. Overview of the ImageCLEF 2013 medical tasks. In: Working notes of CLEF. 2013. p. 1–15.
[4] Long LR, Antani S, Deserno TM, Thoma GR. Content-based image retrieval in medicine: Retrospective assessment, state of the art, and future directions. Int J Healthc Inf Syst Inform 2009;4:1–16.
[5] Müller H, Deserno T. Content-based medical image retrieval. In: Deserno TM, editor. Biomedical image processing, biological and medical physics, biomedical engineering. Berlin/Heidelberg: Springer; 2011. p. 471–94.
[6] Zheng L, Wetzel AW, Gilbertson J, Becich MJ. Design and analysis of a content-based pathology image retrieval system. IEEE Trans Inf Technol Biomed 2003;7:249–55.
[7] Simonyan K, Modat M, Ourselin S, Cash D, Criminisi A, Zisserman A. Immediate ROI search for 3-D medical images. In: MICCAI International Workshop on Content-Based Retrieval for Clinical Decision Support – LNCS 7723. 2012. p. 56–67.
[8] Shyu C-R, Brodley CE, Kak AC, Kosaka A, Aisen AM, Broderick LS. Assert: A physician-in-the-loop content-based retrieval system for {HRCT} image databases. Comput Vis Image Underst 1999;75:111–32.
[9] El-Naqa I, Yang Y, Galatsanos NP, Nishikawa RM, Wernick MN. A similarity learning approach to content-based image retrieval: application to digital mammography. IEEE Trans Med Imaging 2004;23:1233–44.
[10] Müller H, de Herrera AGS, Kalpathy-Cramer J, Demner-Fushman D, Antani S, Eggel I. Overview of the ImageCLEF 2012 medical image retrieval and classification tasks. In: CLEF (Online Working Notes/Labs/Workshop). 2012. p. 1–16.
[11] Kalpathy-Cramer J, Hersh W. Automatic image modality based classification and annotation to improve medical image retrieval. In: Proceedings of the 12th World Congress on Health (Medical) Informatics – Building Sustainable Health Systems – MedInfo. 2007. p. 1334–8.
[12] Association NEM. Digital imaging and communications in medicine – DICOM; 2009 http://dicom.nema.org/
[13] Guld MO, Kohnen M, Keysers D, Schubert H, Wein BB, Bredno J, et al. Quality of DICOM header information for image categorization. In: Medical Imaging 2002: PACS and Integrated Medical Information Systems: Design and Evaluation – SPIE, vol. 4685. 2002. p. 280–7.
[14] Dimitrovski I, Kocev D, Loskovska S, Dzeroski S. Hierarchical annotation of medical images. Pattern Recognit 2011;44:2436–49.
[15] Ojala T, Pietikainen M, Maenpaa T. Multiresolution gray scale and rotation invariant texture classification with local binary patterns. IEEE Trans Pattern Anal Mach Intell 2002;24:971–87.
[16] Chatzichristofis SA, Boutalis YS. EDD: color and edge directivity descriptor: a compact descriptor for image indexing and retrieval. In: ICVS'08 proceedings of the 6th international conference on computer vision systems. 2008. p. 312–22.
[17] Chatzichristofis S, Boutalis Y. Fcth: fuzzy color and texture histogram – a low level feature for accurate image retrieval. In: WIAMIS'08: 9th international workshop on image analysis for multimedia interactive services. 2008. p. 191–6.
[18] Lowe DG. Distinctive image features from scale-invariant keypoints. Int J Comput Vis 2004;60:91–110.
[19] van de Sande K, Gevers T, Snoek C. Evaluating color descriptors for object and scene recognition. IEEE Trans Pattern Anal Mach Intell 2010;32:1582–96.
[20] Ounis I, Amati G, Plachouras V, He B, Macdonald C, Johnson D. Terrier information retrieval platform. In: ECIR2005: advances in information retrieval – LNCS 3408. 2005. p. 517–9.
[21] Kalpathy-Cramer J, Müller H, Bedrick S, Eggel I, de Herrera AGS, Tsikrika T. Overview of the CLEF 2011 medical image classification and retrieval tasks. In: CLEF (notebook papers/labs/workshop). 2011. p. 1–15.
[22] Tommasi T, Orabona F, Caputo B. Discriminative cue integration for medical image annotation. Pattern Recognit Lett 2008;29:1996–2002.
[23] Ziou D, Tabbone S. Edge detection techniques an overview. Pattern Recognit Image Anal 1998;8:537–59.
[24] Park DK, Jeon YS, Won CS. Efficient use of local edge histogram descriptor. In: MULTIMEDIA00: proceedings of the 2000 ACM workshops on multimedia. 2000. p. 51–4.
[25] Zhang J, Marszalek M, Lazebnik S, Schmid C. Local features and kernels for classification of texture and object categories: A comprehensive study. Int J Comput Vis 2007;73:213–38.
[26] van de Sande KEA, Gevers T. University of Amsterdam at the visual concept detection and annotation tasks, vol. 32: ImageCLEF of the information retrieval series. Berlin Heidelberg: Springer-Verlag; 2010. p. 343–58.
[27] van Gemert JC, Veenman CJ, Smeulders AWM, Geusebroek JM. Visual word ambiguity. IEEE Trans Pattern Anal Mach Intell 2010;32:1271–83.
[28] Chang C-C, Lin C-J. LIBSVM: a library for support vector machines; 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm
[29] Lin H-T, Lin C-J, Weng RC. A note on Platt's probabilistic outputs for support vector machines. Mach Learn 2007;68:267–76.
[30] Lazebnik S, Schmid C, Ponce J. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: Proceedings of the 2006 IEEE conference on computer vision and pattern recognition. 2006. p. 2169–78.
[31] Marszałek M, Schmid C, Harzallah H, van de Weijer J. Learning object representations for visual object class recognition. In: Visual recognition challenge workshop, held in conjunction with ICCV2007. 2007.
[32] Tommasi T, Caputo B, Welter P, Güld M, Deserno T. Overview of the CLEF 2009 medical image annotation track. In: Multilingual information access evaluation II. Multimedia experiments – LNCS 6242. 2010. p. 85–93.

Please cite this article in press as: Dimitrovski I, et al. Improved medical image modality classification using a combination of visual and textual features. Comput Med Imaging Graph (2014), http://dx.doi.org/10.1016/j.compmedimag.2014.06.005