Modality Classification for Medical Images Using Sparse Coded Affine-Invariant Descriptors

Viktor Gál[1], Illés Solt[2], Etienne Kerre[1], and Mike Nachtegael[1]

[1] Department of Applied Mathematics and Computer Science,
Ghent University, Belgium
{viktor.gal,etienne.kerre,mike.nachtegael}@ugent.be
[2] Department of Telecommunications and Media Informatics,
Budapest University of Technology and Economics, Hungary
solt@tmit.bme.hu

Abstract. Modality is a key facet in medical image retrieval, as a user is likely interested in only one of e.g. radiology images, flowcharts, and pathology photos. While assessing image modality is trivial for humans, reliable automatic methods are required to deal with large un-annotated image bases, such as figures taken from the millions of scientific publications. We present a multi-disciplinary approach to tackle the classification problem by combining image features, meta-data, textual and referential information. We test our system's accuracy on the ImageCLEF 2011 medical modality classification data set. We show that using a fully affine-invariant feature descriptor and sparse coding on these descriptors in the Bag-of-Words image representation significantly increases the classification accuracy. Our best method achieves 87.89% accuracy and outperforms the state of the art.

Keywords: image classification, image feature extraction, image modality, sparse coding, text mining.

1 Introduction

Imaging modality is an important aspect of the image for medical retrieval [1]. 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. However, this modality is typically extracted from the caption and is often not correct or present. Studies have shown that the modality can be extracted from the image itself using visual features [2,3,4]. Therefore, in this paper, we propose to use both visual and textual features for medical image representation.

Our main focus in this paper is on the impact of using a fully affine-invariant feature descriptor (ASIFT [5]) and (extension of) the Bag-of-Words (BoW) image representation [6]. In the classical BoW image representation vector quantisation is applied to encode descriptors (e.g SIFT, ASIFT) of local image patches. Prior to encoding, a codebook is learned via an unsupervised clustering algorithm (e.g. K-means), which summarizes the distribution of signals by a set of "visual words".

T. Washio and J. Luo (Eds.): PAKDD 2012 Workshops, LNAI 7769, pp. 1–10, 2013.

As a result, these vector quantised codes represent the image through the frequencies of these visual words.

As vector quantisation introduces a significant error in encoding a signal, to overcome this problem, sparse coding has attracted much attention in image classification [7,8]. In this paper we use Least angle regression algorithm [9] for sparse coding the extracted feature descriptors.

We show that by using ASIFT on the images and applying sparse coding on these features we achieved better performance as the state-of-art results for modality classification of medical images.

The proposed algorithm is evaluated in the context of the ImageCLEF 2011 Modality Classification task [10], which uses a data set of 988+1024 images taken from PubMed articles.

The rest of this paper is organised as follows. In Section 2, we describe in detail our experimental setting. In Section 3, we present and discuss the different experiments and we conclude in Section 4.

2 Methods

In this section, we describe in detail our experimental setting.

2.1 Feature Extraction

Caption Text. Figures in scientific publications often have descriptive captions that provide information on the modality of the image. "Contrast-enhanced axial computed tomographic scan", "HRCT showing extensive areas of consolidation with air bronchogram" are examples of captions of images assigned to the 'CT' modality class. However, the caption may be missing or may not hint at the modality, e.g. "E. coli that satisfy the similarity threshold values." As the examples suggest, the linguistic constructs expressing modality can have a high variation.

Considering these remarks, we extract binary features from caption texts as follows. We define a set of regular expressions to be matched against the caption text, a match results in a value of 1. Regular expressions were initially created for each word having a high information gain for any of the modality classes and were later manually refined to capture linguistic variations (e.g. `f?MRI?`) and multi-word phrases (e.g. `error bars?`).

MeSH Terms. Scientific articles indexed by Medline/PubMed are tagged with MeSH terms (medical subject headings) by field experts. MeSH terms can be seen as a thesaurus for the life sciences containing entries like 'Human', 'Liver Neoplasms' and 'Magnetic Resonance Imaging', entries can be further qualified by e.g. 'methods', 'pathology'. We hypothesise that the article's MeSH terms and its figures' modality are correlated, and hence define features corresponding to individual MeSH terms and qualifiers. A unique identifier for the article (e.g. PMID or DOI) is required to retrieve its MeSH annotations, however, such

identifiers can be absent. As the number of MeSH terms, qualifiers and their combinations far exceeds the number of modality labels, we perform feature selection by keeping only those that are present for at least a predefined number of articles in the training set.

Colour Histogram. Using colour histograms in content-based image retrieval system has been successfully applied in the past, for a detailed review see [11]. Based on these studies we have chosen to use HSV colour-space based histogram, and quantised the *hue* and the *saturation* to three and the *value* to four levels.

Based on this we defined $\boldsymbol{f}_{hist}$ feature vector, where each element of the vector represents the normalised number of pixels in a given histogram bin.

Mean of Pixels. Through manually supervised error analysis on the training set, we identified that the images in `Graphic` 1st-level group are mainly having a white background. Hence, we have defined a simple feature $f_{mean} = \overline{\boldsymbol{I}_j}$, that represents the mean value of the pixels in an image. By simply thresholding these values one could identify the images that belong to the `Graphic` group with a very high accuracy.

Axis Recognition. The previously mentioned mean of pixels method gave a strong support for recognising images in the `Graphic` top-level group, but as it consists of two sub-groups, `Graphs` and `Drawing`, thus a new feature was required to differentiate the images belonging to one or the other category. By manually observing the images in these two categories one can easily point out the main difference by using a simple edge detector: the images belonging to the `Graphs` category are mainly consisting of horizontal and vertical lines (i.e. the x-y axis of a graph), whereas the images in `Drawing` category are mostly diagrams, where the orientation of the lines is random.

Based on this idea we have defined the following feature. Let $L_{\boldsymbol{I}_j}$ be the set of all the detected lines and $GL_{\boldsymbol{I}_j}$ be the set of *good lines* in an arbitrary image $\boldsymbol{I}_j$, where a given line is a *good line* if its orientation is horizontal or vertical and it is within a given margin of the picture's border. The latter condition is to eliminate the borders of an image as *good lines*.

Using these two sets we defined a feature

$$f_{lines}(\boldsymbol{I}_j) = \frac{|GL_{\boldsymbol{I}_j}|}{|L_{\boldsymbol{I}_j}|} \tag{1}$$

In order to detect the lines and their orientation in an image we used a simple Hough transform [12].

Skin Detection. The images in the `Dermatology` category was one of the most difficult to recognise. As not only it was the least represented category in the whole training set, i.e. there are only seven examples (see Table 1) for this category, but the images in this set are simple photographs (of various skin abnormalities) thus they have very similar characteristics to the `general photo` labeled images.

Hence, most of the previously defined features failed to distinguish the images in `Dermatology` set from the others.

Using a simple skin detector algorithm [13] we defined a new feature $f_{skin}(\boldsymbol{I}_j)$ for an image $\boldsymbol{I}_j$

$$f_{skin}(\boldsymbol{I}_j) = \overline{SD(\boldsymbol{I}_j)} \tag{2}$$

where the function $SD(\cdot)$ calculates the skin-segmented binary image of an input image, and $\overline{\boldsymbol{I}_k}$—as previously defined—is the mean value of image $\boldsymbol{I}_k$.

Radiopaedia. Radiopaedia (`http://radiopaedia.org`) is a community wiki for radiology images and patient cases. Images are tagged by users with the body system (e.g. Heart, Musculoskeletal) depicted, but unfortunately for us, not with the type of radiology method used to create the image. Leveraging the mutual information between body systems and radiology methods, we derived features for modality classification by taking the output probabilities of a classifier trained to predict body systems shown in the image.

Bag of visual-words The state-of-the-art content based image retrieval systems has been significantly improved by the introduction of scale-invariant feature transform (SIFT) [14] features and the bag-of-words image representation [15,16,17,18].

The bag-of-visual-words image representation is based on the bag of words (BoW) model in natural language processing (NLP). BoW in NLP is a popular method for representing documents. In this model a document is simply represented by the number of different words that are in the document. The idea behind this is, that documents on the same topic have similar words with similar number of occurrences in them (see LDA [19]).

In case of an image, the basic idea of bag-of-words model is that a set of local image patches is sampled using some method—e.g. densely or using a keypoint detector—and a vector of visual descriptors is evaluated on each patch independently.

In this paper we used two variants of the well known SIFT descriptor on each patch:

- **SIFT.** The SIFT descriptor computes a gradient orientation histogram within the support region. For each of eight orientation planes, the gradient image is sampled over a four by four grid of locations, hence resulting in a 128-dimensional feature vector for each region. In order to make the descriptor less sensitive to small changes in the position of the support region and put more emphasis on the gradients that are near the centre of the region a Gaussian window function is used to assign a weight to the magnitude of each sample point.
- **Affine-SIFT.** (ASIFT) [5] The SIFT detector normalizes rotations and translations and simulates all zooms out of the query and of the search images. Because of this feature, it is the only fully scale-invariant method. ASIFT simulates with enough accuracy all distortions caused by a variation of the camera optical axis direction. Then it applies the SIFT method. In

other words, ASIFT simulates three parameters: the scale, the camera longitude angle, and the latitude angle (which is equivalent to the tilt) and normalizes the other three (translation and rotation). The mathematical proof that ASIFT is fully affine invariant is given in [5]. The key observation is that, although a tilt distortion is irreversible due to its non-commutation with the blur, it can be compensated up to a scale change by digitally simulating a tilt of same amount in the orthogonal direction. As opposed to the normalization methods that suffer from this non-commutation, ASIFT simulates and thus achieves the full affine invariance.

After acquiring the feature descriptors for all the images in the data set, first we created a visual-word dictionary $\mathbf{D}$ (analogy to a word dictionary) by performing a K-means clustering algorithm over all the vectors. This dictionary is used to map similar visual patches into one, or more visual-words of the acquired dictionary. The mapping can be done by simple vector quantisation [20], where each visual patch is mapped to the nearest visual-word in the dictionary or by using sparse coding, where the visual patch is a linear combination of a small number of the visual-words.

The sparse coding of the visual patches was achieved by using least angle regression algorithm [9] for solving the Lasso. Given a matrix of signals $\mathbf{X} = [\boldsymbol{x_1}, \ldots, \boldsymbol{x_p}] \in \Re^{m \times p}$ and a dictionary $\mathbf{D} = [\boldsymbol{d_1}, \ldots, \boldsymbol{d_n}] \in \Re^{m \times n}$, the algorithm computes a matrix $\mathbf{A} = [\boldsymbol{\alpha_1}, \ldots, \boldsymbol{\alpha_p}] \in \Re^{n \times p}$, where for each column $\boldsymbol{x}$ of $\mathbf{X}$, it returns a coefficient vector $\boldsymbol{\alpha}$ which is a solution of

$$\min_{\alpha \in \Re^n} ||\boldsymbol{x} - \mathbf{D}\alpha||_2^2 \ \mathtt{s.t.}\ ||\alpha||_1 \leq \lambda \tag{3}$$

In our bag-of-visual-words model we used the *tf-idf* weighting [21] scheme, that has proven to be a very successful approach for image retrieval as well. The *tf* part of the weighting scheme represents the number of features described by a given visual word. The frequency of a visual word in the image provides useful information about repeated structures and textures. The *idf* part captures the informativeness of visual words, the ones that appear in many different images are less informative than those that appear rarely. This weighting scheme is generally applied only to integer counts of visual-words in images. Thus, in case of sparse coding the scheme had to be modified to handle the weight vector $\boldsymbol{\alpha}$. We found the same approach to be the best solution as in [22]. I.e. for the term frequency we simply used the normalized weight value for each visual word. For the inverse document feature measure, we found that counting an occurrence of a visual word as one, no matter how small its weight, gave the best results.

2.2 Classification

Based on the numerical and binary features of the images obtained through feature extraction, we perform vector space classification to predict modality classes of unseen images. Among the classification algorithms available in Weka [23], we found the support vector machine SMO to have the best standalone performance

over the full feature space in cross-validation on ImageCLEF 2011 training data set. We used SMO with default settings for the rest of the experiments.

2.3 Evaluation Setting

Our experiments are based on the ImageCLEF 2011 medical modality classification data set [10], where there are 988 images in training- and 1024 images in the testing data set, which were taken from PubMed articles. The data set defines 18 different modality classes.

Table 1 shows how imbalanced the distribution of the images within the various modality classes are.

Table 1. Modality labels at ImageCLEF 2011 and their distribution

	Modality label		Training	
Group	Code	Description	#	%
Radiology	AN	angiography	11	1.1
	CT	computed tomography	70	7.1
	MR	magnetic resonance imaging	17	1.7
	US	ultrasound	30	3.0
	XR	X-ray	59	6.0
Microscopy	FL	fluorescence	44	4.5
	EM	electronmicroscopy	16	1.6
	GL	gel	50	5.1
	HX	histopathology	208	21.1
Photograph	PX	general photo	165	16.7
	GR	gross pathology	43	4.4
	EN	endoscopic imaging	10	1.0
	RN	retinograph	5	0.5
	DM	dermatology	7	0.7
Graphic	GX	graphs	161	16.3
	DR	drawing	43	4.4
Other	3D	3D reconstruction	32	3.2
	CM	compound figure (> 1 type of image)	17	1.7
Total		18	988	100.0

3 Results and Discussion

In this section we provide the results of the four different experimental settings. Table 2 shows the correctly classified percentage for each case and we included the result of the best submitted run [24] of the challenge as well. In all of the four cases we used all the features that has been introduced in the previous section.

Table 2. Results of the runs for the medical modality classification task. For the reference we have included the best performing run of the competition.

Run	Method	Accuracy
#1	sparse coded Affine-SIFT	**87.89**
#2	hard vector quantised Affine-SIFT	84.66
#3	sparse coded SIFT	86.42
#4	hard vector quantised SIFT	86.03
Best of ImageClef '11	SLR with Fisher Vector	86.91

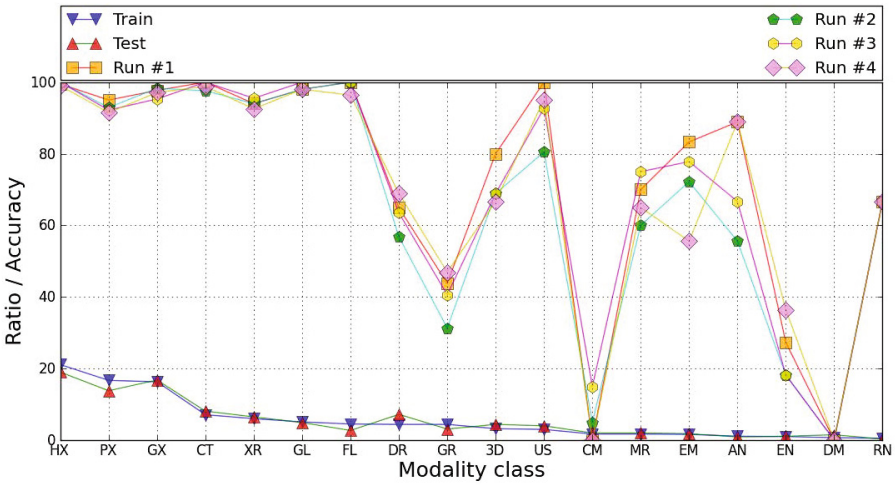

Fig. 1. Modality class distribution and the performance of different runs. Modality classes are sorted by support in descending order. For the names of modality classes, see Table 3.

For simplicity table 2 only shows the difference between the used features in the different runs.

Run #4 is our best best submitted run for the ImageCLEF 2011 medical image modality challenge. Although, sparse coding on the extracted SIFT features (run #3) does improve the classification accuracy, it is still lower than the best submitted run for the challenge.

It is important to note that using a fully affine-invariant feature descriptor will not necessarily improve the classification accuracy. On the contrary, as run #2 shows, the overall accuracy of the system significantly dropped by using ASIFT instead of SIFT. But as run #1 shows, if sparse coding is used instead of hard vector quantisation on ASIFT descriptors, the accuracy significantly improves and outperforms the state of the art. The performance of the runs broken down for the individual classes is shown in Table 3.

Table 3. Correctly classified images per category for the submitted runs. For each modality class, the result of the best performing run is typeset in bold.

Modality class	Ratio (%)		Run			
	train	test	#1	#2	#3	#4
3D : 3D render	3.2	4.4	**80.0**	68.9	68.9	66.7
AN : Angiography	1.1	0.9	**88.9**	55.6	66.7	**88.9**
CM : Compound figure	1.7	2.0	0.0	5.0	**15.0**	0.0
CT : Computed tomography	7.1	8.1	**100**	97.6	**100**	98.8
DM : Dermatology	0.7	1.5	0.0	0.0	0.0	0.0
DR : Drawing	4.4	7.2	64.9	56.8	63.5	**68.9**
EM : Electronmicroscope	1.6	1.8	**83.3**	72.2	77.8	55.6
EN : Endoscope	1.0	1.1	27.3	18.2	18.2	**36.4**
FL : Fluorescence	4.5	2.7	**100**	**100**	**100**	96.4
GL : Gel	5.1	4.9	98.0	98.0	**100**	98.0
GR : Gross pathology	4.4	3.1	43.8	31.3	40.6	**46.9**
GX : Graphics	16.3	16.8	97.7	**98.3**	95.3	97.1
HX : Histopathology	21.1	19.0	99.5	**100**	**100**	99.0
MR : MRI	1.7	2.0	70.0	60	**75.0**	65.0
PX : Photo	16.7	13.8	**95.0**	92.9	92.2	91.5
RN : Retiongraph	0.5	0.3	66.7	66.7	66.7	66.7
US : Ultrasound	3.0	4.0	**100**	80.5	92.7	95.1
XR : X-ray	6.0	6.5	94.0	94.0	95.5	92.5

4 Conclusion

In this paper, we proposed to extract different visual and textual features for medical image representation, and fusion the different extracted visual feature and textual feature for modality classification. To extract visual features from the images, we used some state-of-art methods like bag-of-visual words and some standard ones like colour histogram and introduced some heuristic representations of the images specialised for the ImageCLEF 2011 medical modality classification task.

We showed that using sparse coding instead of vector quantisation in the BoW representation for encoding the extracted affine-invariant feature descriptors with a given visual-word dictionary will increase the classification accuracy.

With the suggested feature extraction algorithms in this paper we have achieved a 87.89% accuracy that outperforms the state of the art.

Acknowledgements. Special thanks for Dr. Tikk Domonkos for his insightful comments and suggestions for the draft of the paper. Viktor Gál was supported by Marie Curie Initial Training Networks (ITN) Ref. 238819 (FP7-PEOPLE-ITN-2008).

References

1. Hersh, W.R., Müller, H., Jensen, J.R., Yang, J., Gorman, P.N., Ruch, P.: Advancing Biomedical Image Retrieval: Development and Analysis of a Test Collection. Journal of the American Medical Informatics Association 13(5), 488–496 (2006)
2. Pentland, A., Picard, R.W., Sclaroff, S.: Photobook: Content-based manipulation of image databases. International Journal of Computer Vision 18(3), 233–254 (1996)
3. Lakdashti, A., Moin, M.S.: A New Content-Based Image Retrieval Approach Based on Pattern Orientation Histogram. In: Gagalowicz, A., Philips, W. (eds.) MIRAGE 2007. LNCS, vol. 4418, pp. 587–595. Springer, Heidelberg (2007)
4. Jain, A.: Image retrieval using color and shape. Pattern Recognition 29(8), 1233–1244 (1996)
5. Morel, J.-M., Yu, G.: ASIFT: A New Framework for Fully Affine Invariant Image Comparison. SIAM Journal on Imaging Sciences 2(2) (April 2009)
6. Csurka, G., Dance, C., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: Workshop on Statistical Learning in Computer Vision, ECCV, vol. 1, p. 22 (2004)
7. Jarrett, K., Kavukcuoglu, K., Ranzato, M.A., LeCun, Y.: What is the best multi-stage architecture for object recognition? In: 2009 IEEE 12th International Conference on Computer Vision, pp. 2146–2153 (2009)
8. Yang, J., Yu, K., Gong, Y., Huang, T.: Linear spatial pyramid matching using sparse coding for image classification. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 1794–1801 (2009)
9. Efron, B., Hastie, T., Johnstone, I., Tibshirani, R.: Least Angle Regression. ArXiv Mathematics e-prints (June 2004)
10. Kalpathy-Cramer, J., Müller, H., Bedrick, S., Eggel, I., de Herrera, A.G.S., Tsikrika, T.: The CLEF 2011 medical image retrieval and classification tasks. In: CLEF 2011 Working Notes, Amsterdam, The Netherlands (2011)
11. Veltkamp, R.C.: A survey of content-based image retrieval systems. Content-based Image and Video Retrieval (2002)
12. Duda, R.O.: Use of the Hough transformation to detect lines and curves in pictures. Communications of the ACM (1972)
13. Chai, D., Ngan, K.N.: Face segmentation using skin-color map in videophone applications. IEEE Transactions on Circuits and Systems for Video Technology 9(4), 551–564 (1999)
14. Lowe, D.G.: Object recognition from local scale-invariant features. In: Proceedings of the International Conference on Computer Vision, ICCV 1999, pp. 1150–1157. IEEE Computer Society, Washington, DC (1999)
15. Nister, D., Stewenius, H.: Scalable Recognition with a Vocabulary Tree. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2161–2168 (2006)
16. Jegou, H., Harzallah, H., Schmid, C.: A contextual dissimilarity measure for accurate and efficient image search. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2007), pp. 1–8 (2007)
17. Chum, O., Philbin, J., Sivic, J., Isard, M.: Total Recall: Automatic query expansion with a generative feature model for object retrieval. In: 2007 IEEE 11th International Conference on Computer Vision, pp. 1–8. IEEE (October 2007)
18. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Lost in quantization: Improving particular object retrieval in large scale image databases. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2008)

19. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022 (2003)
20. Gal, V., Solt, I., Gedeon, T., Nachtegael, M., Kerre, E.: Multi-disciplinary modality classication for medical images. In: CLEF 2011 Working Notes, Amsterdam, The Netherlands (2011)
21. Sivic, J., Zisserman, A.: Video Google: A Text Retrieval Approach to Object Matching in Videos. In: 9th IEEE International Conference on Computer Vision (ICCV 2003), pp. 1470–1477. IEEE Computer Society (2003)
22. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Lost in quantization: Improving particular object retrieval in large scale image databases. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008, pp. 1–8 (2008)
23. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. SIGKDD Explorations 11(1), 10–18 (2009)
24. Csurka, G., Clinchant, S., Jacquet, G.: XRCE's Participation at Medical Image Modality Classification and Ad-hoc Retrieval Tasks of Image CLEF 2011. In: CLEF 2011 Working Notes, Amsterdam, The Netherlands (2011)