Multi-disciplinary modality classification for medical images

Viktor Gál[1,3], Illés Solt[2], Tom Gedeon[3], Mike Nachtegael[1]

[1] Department of Applied Mathematics and Computer Science,
Ghent University, Belgium
{viktor.gal,mike.nachtegael}@ugent.be
[2] Department of Telecommunications and Media Informatics,
Budapest University of Technology and Economics, Hungary
solt@tmit.bme.hu
[3] School of Computer Science,
The Australian National University, Australia
tom.gedeon@anu.edu.au

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. Our system achieved an accuracy of 96.86 % in cross-validation on the ImageCLEF 2011 training dataset having 18 imbalanced modality classes, and an accuracy of 90.2 % on the ImageCLEF 2010 dataset having 8 well-balanced modality classes. We evaluate the importance of the individual feature sets in detail, and provide an error analysis pointing at weaknesses of our method and obstacles in the classification task. For the benefit of the image classification community, we make the results of our feature extraction methods publicly available at http://categorizer.tmit.bme.hu/~illes/imageclef2011modality.

Keywords: image classification, image feature extraction, image modality, text mining

1 Introduction

Imaging modality is an important aspect of the image for medical retrieval [6]. 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 [13, 10, 7]. Therefore, In this paper, we propose to use both visual and textual features for medical image representation, and combine the different features using normalised kernel function in SVM.

The proposed algorithm is evaluated in the context of the ImageCLEF 2011 Modality Classification task[9], which uses a dataset of 988+1024 images taken from PubMed articles.

The rest of this paper is organised as follows. In Section 1, we describe in detail our experimental setting. In Section 3, we present and compare different runs we submitted. We discuss the submitted runs and the results in Section 4 and we conclude in Section 5.

2 Methods

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

2.1 Evaluation setting

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

The ImageCLEF 2011 Modality Classification task used split-validation measuring the accuracy of the systems. On the training dataset, we performed stratified 10-fold cross-validation to evaluate feature sets and classifiers.

2.2 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 [16]. 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 number of *good lines* in an arbitrary image $\boldsymbol{I}_j$, where a given line is a *good line* if it's orientation is horizontal or vertical and it is within a given margin of the picture's border. The latter condition is for not to count 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 [4].

Skin detection The images in the `Dermatology` category was one of the most difficult 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[2] we defined a new feature $f_{skin}(\boldsymbol{I}_j)$ for and 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 $\boldsymbol{I}_k$.

Meta-data We determine whether an image post-processing software was used by analysing meta-data stored in JPEG files' EXIF section. For this, we analyze the 'Comment' field, to find mentions of commonly used image manipulation software (e.g. Adobe Photoshop, MS Paint). We also extract from the EXIF whether the image is stored as gray-scale only.

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 SIFT[11] features and the bag-of-words image representation [12, 8, 3, 14].

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[1]).

In case of and 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 key-point detector–and a vector of visual descriptors is evaluated on each patch independently. In this paper we used the well known SIFT descriptor on each patch. 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 y 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.

After acquiring these SIFT features for all the images in the dataset, the final step is to convert vector represented patches to ”codewords” (analogy to words in text documents), which also produces a ”codebook” (analogy to a word dictionary). A codeword can be considered as a representative of several similar patches. In our case we performed k-means clustering over all the vectors. Codewords are then defined as the centres of the learned clusters. Thus, each patch in an image is mapped to a certain codeword through the clustering process and the image can be represented by the histogram of the codewords.

In our bag-of-visual-words model we used the the *tf-idf* weighting[15] scheme, that has proven to be a very successful approach for image retrieval. The *tf* part of the weighting scheme represents the number of features described by a given visual word. The frequency of visual word in the image provides useful information about repeated structures and textures. While, the *idf* part captures the informativeness of visual words–visual words that appear in many different images are less informative than those that appear rarely.

Other systems The challenge organisers generously supplied participants with predictions of their in-house system. This classification was automatic for the test set, but confusingly enough, the ground truth labels were used for the train set. In order to exploit this valuable resource, we used it as an input to our classifier by introducing artificial smoothing to avoid overfitting on this particular otherwise noise free indicator variable. Also note that while split evaluation is sound in this setting, the cross-validation evaluation of those two runs is flawed (being over-optimistic) due to information leakage.

2.3 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 [5], 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 dataset. We used SMO with default settings for the rest of the experiments unless stated otherwise.

3 Results

In this section, we provide the final results of the five submitted runs for the modality classification tasks. Table 2 shows both the correctly classified percentage for the different features set compositions. Comparing the result of our best submitted run and the best submitted run of the modality classification task, one can see that there is very small (0.88%) difference between the two runs.

The performance of the runs broken down for the individual classes is show in Table 3 and in Figure 1.

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. The figures in parenthesis are the result of information leakage that only appears in the cross-validation setting, see Section 2.2 for details.

		Accuracy	
Run	Feature set	test	cross-val
#1	MeSH+BoW+RP+Cap+f_{hist}+f_{skin}+f_{mean}+f_{lines}	**86.03**	82.59
#2	MeSH+BoW+Cap+f_{hist}+f_{skin}+f_{mean}+f_{lines}	85.64	81.57
#3	BoW+Cap+f_{hist}+f_{skin}+f_{mean}	85.15	80.97
#4	Sys+MeSH+BoW+f_{hist}+f_{skin}+f_{mean}+f_{lines}	76.85	(94.83)
#5	Sys+MeSH+BoW+RP+Cap+f_{hist}+f_{skin}+f_{mean}+f_{lines}+f_{CEDD}	74.12	(96.86)
Best	*n/a*	86.91	*n/a*

4 Discussion

As can be seen on Figure 1, the systems performs well on higher support classes, while performance drops to zero for some more rare classes. This behaviour is tolerated by the challenge main evaluation metric accuracy, in contrast to a more pessimistic evaluation like F-measure. Table 2 shows, which features have been used in the different runs. It is important to see that omitting *Caption text* features results in almost about a ten percent accuracy loss, see the difference between the runs #3 and #4.

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.

	Ratio (%)		Run				
Modality class	train	test	#1	#2	#3	#4	#5
3D : 3D render	3.2	4.4	66.7	71.1	**73.3**	66.7	57.8
AN: Angiography	1.1	0.9	**88.9**	77.8	77.8	66.7	88.8
CM: Compound figure	1.7	2.0	0.0	**5.0**	5.0	5.0	5.0
CT : Computed tomography	7.1	8.1	**98.8**	97.6	95.2	91.6	89.2
DM: Dermatology	0.7	1.5	0.0	0.0	0.0	6.7	**13.3**
DR: Drawing	4.4	7.2	68.9	66.2	**70.3**	27.0	24.3
EM: Electronmicroscope	1.6	1.8	**55.6**	**55.6**	**55.6**	**55.6**	**55.6**
EN : Endoscope	1.0	1.1	**36.4**	36.4	27.3	36.4	27.3
FL : Fluorescence	4.5	2.7	96.4	96.4	**100**	**100**	**100**
GL : Gel	5.1	4.9	98.0	98.0	**100**	82.0	80.0
GR: Gross pathology	4.4	3.1	**46.9**	40.6	34.4	34.4	34.4
GX: Graphics	16.3	16.8	**97.1**	96.5	94.8	97.1	96.5
HX: Histopathology	21.1	19.0	**99.0**	99.0	99.0	95.4	95.9
MR: MRI	1.7	2.0	65.0	70.0	**75.0**	60.0	50.0
PX : Photo	16.7	13.8	**91.5**	90.1	88.7	73.8	66.7
RN: Retiongraph	0.5	0.3	**66.7**	66.7	66.7	0.0	33.3
US : Ultrasound	3.0	4.0	**95.1**	95.1	90.2	85.4	78.0
XR: X-ray	6.0	6.5	92.5	**94.0**	94.0	82.1	71.6

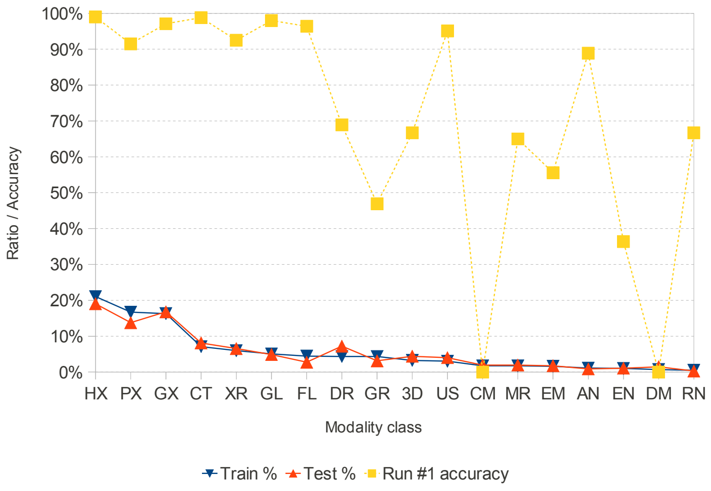

Fig. 1: Modality class distribution and best run performance. Modality classes are sorted by support in descending order. For the names of modality classes, see Table 3.

Using *MeSH* and *Radiopaedia* features gained us about one percent in accuracy.

The in-house modality classifier of the challenge organisers proved to be superior in predicting the 'Dermatology' class (Table 3, however, its inferior performance on higher support classes prevented it from being benefitial in combination (Table 2).

4.1 Other experiments

Motivated by the grouping of modality labels by the challenge organisers, we experimented with hierarchical classification. In particular, we applied a hierarchical greedily ascending classifier scheme wrapping the baseline classifier. In this scheme, classification is first performed on the hierarchies uppermost level (here groups), then the most probable hierarchy node is selected where classification continues recursively. For hierarchical classification, cross-validation results were inferior to those obtained from the baseline (flat) classifier.

5 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 ImageCLEF2011 medical modality classification task.

With the suggested feature extraction algorithms in this paper and the SVM classifier we have achieved to 2nd place on the ImageCLEF2011 medical image modality classification task.

Acknowledgements

Viktor Gál was supported by Marie Curie Initial Training Networks (ITN) Ref. 238819 (FP7-PEOPLE-ITN-2008).

References

1. David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.
2. D Chai and K N Ngan. Face segmentation using skin-color map in videophone applications. *Circuits and Systems for Video Technology, IEEE Transactions on*, 9(4):551–564, 1999.
3. O Chum, J Philbin, J Sivic, and M Isard. Total Recall: Automatic query expansion with a generative feature model for object retrieval. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, October 2007.

4. RO Duda. Use of the Hough transformation to detect lines and curves in pictures. *Communications of the ACM*, 1972.
5. Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA data mining software: an update. *SIGKDD Explorations*, 11(1):10–18, 2009.
6. William R Hersh, Henning Müller, Jeffery R Jensen, Jianji Yang, Paul N Gorman, and Patrick Ruch. Advancing Biomedical Image Retrieval: Development and Analysis of a Test Collection. *Journal of the American Medical Informatics Association*, 13(5):488–496, 2006.
7. A Jain. Image retrieval using color and shape. *Pattern Recognition*, 29(8):1233–1244, August 1996.
8. H. Jegou, H. Harzallah, and C. Schmid. A contextual dissimilarity measure for accurate and efficient image search. In *Computer Vision and Pattern Recognition, 2007, IEEE Conference on, (CVPR '07)*, pages 1–8, 2007.
9. Jayashree Kalpathy-Cramer, Henning Müller, Steven Bedrick, Ivan Eggel, Alba Garcia Seco de Herrera, and Theodora Tsikrika. The CLEF 2011 medical image retrieval and classification tasks. In *CLEF 2011 working notes*, Amsterdam, The Netherlands, 2011.
10. Abolfazl Lakdashti and M. Moin. A New Content-Based Image Retrieval Approach Based on Pattern Orientation Histogram. In André Gagalowicz and Wilfried Philips, editors, *Computer Vision/Computer Graphics Collaboration Techniques*, pages 587–595. Springer Berlin / Heidelberg, Berlin, Heidelberg, 2007.
11. David G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the International Conference on Computer Vision-Volume 2 - Volume 2*, ICCV '99, pages 1150–, Washington, DC, USA, 1999. IEEE Computer Society.
12. D Nister and H Stewenius. Scalable Recognition with a Vocabulary Tree. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, pages 2161–2168, 2006.
13. A Pentland, R W Picard, and S Sclaroff. Photobook: Content-based manipulation of image databases. *International Journal of Computer Vision*, 18(3):233–254, 1996.
14. J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. 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.
15. Josef Sivic and Andrew Zisserman. Video Google: A Text Retrieval Approach to Object Matching in Videos. In *9th IEEE International Conference on Computer Vision (ICCV 2003)*, pages 1470–1477. IEEE Computer Society, 2003.
16. RC Veltkamp. A survey of content-based image retrieval systems. *Content-based image and video retrieval*, 2002.