ImageCLEF 2010 Working Notes on the Modality Classification subtask

Olivier Pauly, Diana Mateus, and Nassir Navab

Computer Aided Medical Procedures,
Technische Universität München, Germany
pauly@cs.tum.edu, mateus@cs.tum.edu, navab@cs.tum.edu

Abstract. The goal of this work is to investigate the performance of classical methods for feature description and classification, and to identify the difficulties of the ImageCLEF 2010 modality classification subtask. In this paper, we describe different approaches based on visual information for classifying medical images into 8 different modality classes. Since within the same class, images depict very different objects, we focus on global descriptors such as histograms extracted from scale-space, log-Gabor and phase congruency feature images. We also investigated different classification approaches based on support vector machines and random forests. A grid-search associated to a 10 folds cross-validation has been performed on a balanced set of 2390 images to find the best hyperparameters for the different models we propose. All experiments have been conducted with MATLAB on a Workstation with Intel Duo Core 3.16 Ghz and 4Gb of RAM. Our approach based on simple SVM and random forests give best performance and achieve respectively an overall f-measure of 74.13% and 73.59%.

1 Introduction

In this paper, we investigate different classification approaches for the new ImageCLEF subtask modality classification [1]. The problem brings new challenges as in the database given as training set (see Fig.1), images show a very high variability and very different kinds of anatomy appear within the same class. Furthermore, the problem is different from classical object recognition, in which features are especially designed for recognizing an object subject to different imaging conditions. In the present case, we aim at recognizing the imaging modality. Since we can not rely on the different structures appearing in images to discriminate them, we propose to focus on statistics based on global texture information. This is equivalent to making the assumption that each imaging modality shows different texture patterns. To characterize statistical texture information, we propose descriptors based on different feature images such as scale-space, log-Gabor and phase congruency feature images. Once a representation has been chosen to describe global textural statistics, a multi-class classification scheme is applied. Therefore, we investigated 3 different multi-class approaches: a simple multi-class SVM, a multi-kernel multi-class SVM and random forests.

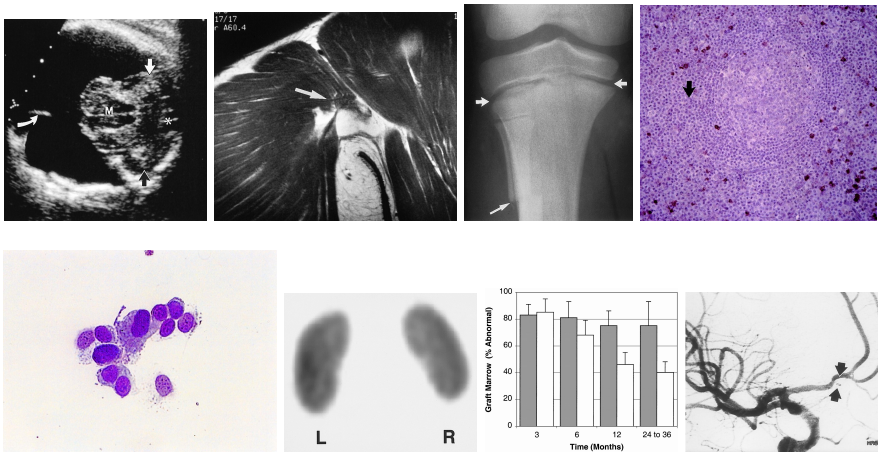

Fig. 1. Examples of images taken from the training set showing a very high variability and different kinds of anatomy

The remaining of this paper is organized as follows: Section 2 formulates the modality classification problem, Section 3 describes the different descriptors and classification methods investigated, Section 4 reports our experiments on the training dataset provided during for the modality classification subtask, and Section 5 concludes this paper and gives an outlook on our future work.

2 Problem Statement

The modality classification problem is an instance of a multi-class classification problem. We denote $I : \mathbb{R}^2 \rightarrow \mathbb{R}$ an image we want to classify into one of the 8 different classes $\{\mathcal{C}_k\}_{k\in\{1,\cdots,8\}}$, namely computerized tompography, graphics, magnetic resonance imaging, nuclear medicine, positron emission tomography, optical imaging, ultrasound and X-ray. I is described by a feature vector we denote $\mathbf{x}$, and y its corresponding class label. The goal of multi-class classification is then to train a decision function Ψ such that:

$$\Psi(\mathbf{x}_n) = y_n, \quad \forall n \in \{1, \cdots, N\} \tag{1}$$

where $\{\mathbf{x}_n, y_n\}_{n\in\{1,\cdots,N\}}$ is the training set composed of image descriptors and their corresponding class labels. In the following section, we will show how to compute the different components of $\mathbf{x}$.

3 Methods

3.1 Feature Representation

As mentioned in the introduction, we will focus on statistical texture features to describe our images. We propose to use 3 different types of descriptors based on scale-space, log-Gabor and phase-congruency feature images.

Scale-space statistical descriptors A basic approach to characterize the image variations at different band of the frequency spectrum is to perform its multi-scale analysis [2]. By convolving the input image with a smoothing kernel, structure can be analyzed at different scales. As smoothing operator, a Gaussian kernel with increasing values of variance σ provides coarser resolutions:

$$G_\sigma(i,j) = \frac{1}{2\pi\sigma}\exp\left(-\frac{(i^2+j^2)}{2\sigma}\right) \tag{2}$$

where $(i,j) \in [-s,s]^2$, $(2s+1)\times(2s+1)$ being the size of the filter mask. By convolving the input image I with the kernel in Eq.2, we obtain following feature images:

$$L_\sigma = I * G_\sigma \tag{3}$$

By doing this for a set of scales $\{\sigma_m\}_{m\in\{1,\cdots,M\}}$, we obtain feature images $\{L_{\sigma_m}\}_{m\in\{1,\cdots,M\}}$. For each of these feature images, we compute its corresponding intensity histograms which gives us a set of descriptors $\{H_m^{\text{gauss}}\}_{m\in\{1,\cdots,M\}}$.

log-Gabor statistical descriptors A common approach to characterize textures is to use a bank of Gabor filters which permit to analyze different part of the frequency spectrum for different orientations [3]. As suggested in [4], we use a logarithmic variant of these filters which have a Gaussian response when viewed on a logarithmic frequency scale. The transfer function of the log-Gabor filter is defined as:

$$\mathcal{G}(f) = \exp\left(-\frac{(\log(f/f_0))^2}{(\log(\kappa/f_0))^2}\right) \tag{4}$$

where f_0 is the center frequency of the sinusoid and κ is a scaling factor of the bandwidth. To cover the whole frequency spectrum, a range of different scales and orientations must be considered while designing the filter bank. After convolving the input image I with the corresponding even-symmetric and odd-symmetric filters for a certain scale σ and orientation θ we obtain pairs if response images $I^e_{\sigma,\theta}$ and $I^o_{\sigma,\theta}$. From these responses, we compute the corresponding amplitude and phase:

$$A_{\sigma,\theta} = \sqrt{(I^e_{\sigma,\theta})^2 + (I^o_{\sigma,\theta})^2} \tag{5}$$

$$\Phi_{\sigma,\theta} = \arctan\left(I^e_{\sigma,\theta}, I^o_{\sigma,\theta}\right) \tag{6}$$

By doing this for a set of scales $\sigma_{p\in\{1,\cdots,P\}}$ and orientations $\theta_{q\in\{1,\cdots,Q\}}$, we obtain a set of feature images $\left\{A_{\sigma_p,\theta_q}, \Phi_{\sigma_p,\theta_q}\right\}_{p,q}$. For each of them, we compute the corresponding intensity histogram which finally gives us the descriptors $\left\{H^{\text{gabor}}_{p,q}\right\}_{p,q}$.

Phase congruency statistical descriptors Phase congruency is an absolute measure of feature significance which is invariant to changes in illumination or contrast [5]. It is closely related to the concept of local energy model which postulates that significant features are perceived at points of maximum phase congruency. The phase congruency function is defined as:

$$PC = \frac{E}{\sum_\sigma A_\sigma} \tag{7}$$

where E is the local energy and A_σ the Fourier amplitudes. It is possible to approximate the function above by using quadrature filters such as the log-Gabor filters described in the previous part. Indeed, by using the response images $I^e_{\sigma,\theta}$ and $I^o_{\sigma,\theta}$, we can approximate the phase congruency for a certain orientation as follows:

$$PC_\theta = \frac{\sqrt{\left((\sum_\sigma I^e_{\sigma,\theta})^2 + (\sum_\sigma I^o_{\sigma,\theta})^2\right)}}{\sum_\sigma \sqrt{(I^e_{\sigma,\theta})^2 + (I^o_{\sigma,\theta})^2}} \tag{8}$$

For different orientations $\theta_{t\in\{1,\cdots,T\}}$, we obtain the feature images $\{PC_{\theta_t}\}_{t\in\{1,\cdots,T\}}$. For each of them, we finally compute the corresponding intensity histogram which gives us the descriptors $\{H^{\text{pc}}_t\}_{t\in\{1,\cdots,T\}}$.

Global descriptors After computing all these descriptors for an input image I, a global descriptor $\mathbf{x}$ is created by concatenating all histograms:

$$\mathbf{x} = \left[\underbrace{\cdots\ H^{\text{gauss}}_m\ \cdots}_{\text{scale-space}}, \underbrace{\cdots\ H^{\text{gabor}}_{p,q}\ \cdots}_{\text{log-Gabor}}, \underbrace{\cdots\ H^{\text{pc}}_t\ \cdots}_{\text{phase-congruency}}\right]^T \tag{9}$$

3.2 Different multi-class classification approaches

We investigated several approaches to classify images according to the descriptors we described in the previous section.

Simple Multi-Class SVM From a set of images and their corresponding labels, we compute their descriptors and generate a training set $\{\mathbf{x}_n, y_n\}_{n\in\{1,\cdots,N\}}$. Recall the classifcation problem statement, we are looking for a function Ψ such that:

$$\Psi(\mathbf{x}_n) = y_n, \quad \forall n \in \{1,\cdots,N\} \tag{10}$$

Let us first consider a 2 class classification problem where $y_n \in \{-1, 1\}$. We model the function Ψ as:

$$\Psi(\mathbf{x}) = \langle w \ , \ \varphi(\mathbf{x}) \rangle + b \tag{11}$$

where φ is a non-linear mapping from the space spanned by our descriptors into a hidden feature space with higher dimensionality, w is a weighting vector and b a bias. w, φ and b define a hyperplane separating the different classes in the high-dimensional space. To find the optimal hyperplane, we maximize the margin between the classes and this, with a minimum of classification errors [6]. This can be written as the dual convex optimization problem:

$$\begin{aligned} &\text{minimize} \ \ \frac{1}{2} \left\| w \right\|^2 \\ &\text{subject to:} \ \ y_n(\langle w \ , \ \varphi(\mathbf{x}_n) \rangle + b) \geq 1 \end{aligned} \tag{12}$$

Since the problem may not be separable, we allow misclassification errors by introducing slack variables that induce a soft margin:

$$\begin{aligned} &\text{minimize} \ \ \frac{1}{2} \left\| w \right\|^2 + \mathrm{C} \sum_{n=1}^{N} \xi_n \\ &\text{subject to:} \ \ y_n(\langle w \ , \ \varphi(\mathbf{x}_n) \rangle + b) \geq 1 - \xi_n \end{aligned} \tag{13}$$

where C is a tradeoff parameter weighting the impact of the errors and thus the flexibility of the model. After solving the dual, it can be shown that a solution w_{opt} of this minimization is always a linear combination of the training vectors $\{x_n\}_n$ with weights $\{\alpha_n\}_{n \in \{1,\dots,N\}}$:

$$w_{opt} = \sum_{n=1}^{N} \alpha_n \ \varphi(\mathbf{x}_n) \tag{14}$$

which leads to the following model for Ψ:

$$\Psi(\mathbf{x}) = \sum_{n=1}^{N} \alpha_i \ \ \langle \varphi(\mathbf{x}_n) \cdot \varphi(\mathbf{x}) \rangle + b = \sum_{n=1}^{N} \alpha_i \ K(\mathbf{x}_n, \mathbf{x}) + b, \tag{15}$$

where K is the kernel associated to φ in the higher dimensional space. To handle complex non-linear relations between the descriptors and their labels, K is chosen as a RBF kernel, leading to the following decision function:

$$\Psi(\mathbf{x}) = \sum_{n=1}^{N} \alpha_n \ \exp\left(-\gamma \left\| \mathbf{x}_n - \mathbf{x} \right\|^2\right) + b. \tag{16}$$

Now that we have a decision function for a binary classification problem, we need to extend it to multi-class. Two common methods are to combine several binary SVM into a multi-class model:

- **One-vs-one**: binary classifiers are built for each combination of classes
- **One-vs-rest**: a binary classifier is built for each class against all the others

Since the one-vs-rest combination is less suitable in terms of class balancing, we choose a one-vs-one approach as suggested by [7].

Multi-Kernel Multi-Class SVM In the previous section, the descriptors $\mathbf{x}$ are a concatenation of different feature histograms. Each of these feature types spans its own space and may live in a different subspace. For this reason, it is advantageous to use a kernel for each feature type that is especially adapted to the features. This leads to a set of P decision functions where P is the number of feature types considered:

$$\Psi^{(p)}(\mathbf{x}^{(p)}) = \sum_{n=1}^{N} \alpha_n^{(p)} \ \exp\left(-\gamma^{(p)} \left\|\mathbf{x}_n^{(p)} - \mathbf{x}^{(p)}\right\|^2\right) + b^{(p)}. \tag{17}$$

where $\mathbf{x}^{(p)}$ is the descriptors of type p. The global decision function would be then a combination of each "specialized" decision function. In this paper, we investigate two combination approaches:

$$\text{linear multi-SVM:} \ \ \Psi(\mathbf{x}) = \ \frac{1}{P}\sum_{p=1}^{P} \Psi^{(p)}(\mathbf{x}^{(p)}) \tag{18}$$

$$\text{product multi-SVM:} \ \ \Psi(\mathbf{x}) = \prod_{p=1}^{P} \Psi^{(p)}(\mathbf{x}^{(p)}) \tag{19}$$

which we respectively call linear and product Multi-SVM in the remaining of this paper.

Random Forests Random forests are basically a ensemble of T independent random trees we denote $\left\{\Theta^{(t)}\right\}_{t\in\{1,\cdots,T\}}$ which vote for the most popular class [8]. Each tree $\Theta^{(t)}$ can be seen as an ensemble of random split functions which partition the feature space and give an estimate of the posterior probability distribution $P(\mathcal{C}_k|\mathbf{x},\Theta^{(t)})$ for each class $\mathcal{C}_k$. The forest $\mathcal{F} = \left\{\Theta^{(1)},\cdots,\Theta^{(T)}\right\}$ combines the trees estimates as follows:

$$P(\mathcal{C}_k|\mathbf{x}) = \frac{1}{T}\sum_{t=1}^{T} P(\mathcal{C}_k|\mathbf{x},\Theta^{(t)}) \tag{20}$$

The class of an unseen point is then determined by using a maximum a posteriori criterion:

$$y = \underset{k}{\text{argmax}}(P(\mathcal{C}_k|\mathbf{x})) \tag{21}$$

A clear advantage of random forests over the SVM approaches is their inherent multi-class characteristic.

4 Experiments and Results

4.1 Experimental setup

In this section, we compare the different approaches we decribed above by performing a 10-folds cross-validation on a set of 2390 labeled images distributed over the classes as follows:

- **CT**: Computerized tomography (314 images)
- **GX**: Graphics, typically drawing and graphs, (355 images)
- **MR**: Magnetic resonance imaging (299 images)
- **NM**: Nuclear Medicine (204 images)
- **PET**: Positron emission tomography including PET/CT (285 images)
- **PX**: optical imaging including photographs, micrographs, gross pathology etc (330 images)
- **US**: ultrasound including (color) Doppler (307 images)
- **XR**: x-ray including x-ray angiography (296 images)

For each test, classes are rebalanced by using random subsampling. A grid-search has been performed to find the best hyperparameters as shown on Fig.2. For the SVM-based classifiers, the tradeoff parameter C and γ are optimized. Concerning the random forests, we are looking for the best number of trees and the best number of random tests at each node. Note that features are scaled so that each dimension is in the range $[0, 1]$.

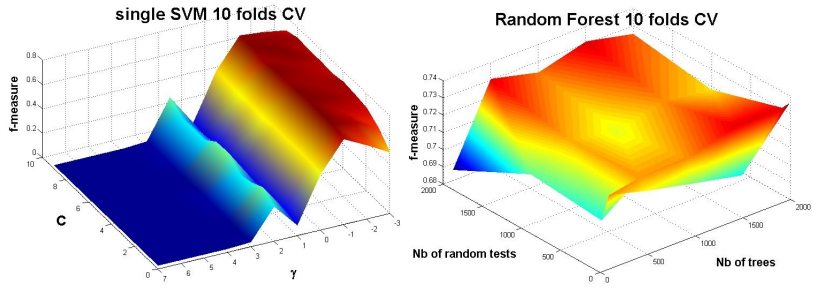

Fig. 2. Grid-search cross-validation to find the best hyperparameters for the single SVM (left) and the random forest (right)

4.2 Results and Discussion

To evaluate the different approaches, we compute several quality measures from the confusion matrix: overall accuracy, precision, recall, f-measure and error rate. Overall results summarized in Tab.1 show that the simple SVM and random forests perform better than the multi-SVM approaches. This is contrary to what could be expected from these methods designed to capture different aspects of the features. Note that in the multi-SVM approaches, each of the SVM was trained independently. A joint training of all kernels, called Multiple-Kernel Learning (MKL) [9, 10], may perform better.

As feature representation, we relied on global statistical descriptors extracted from the textural information contained in the images from the 8 different modalities. To extract this textural information we used classical filters which are not

Overall Classification Results in %					
	Accuracy	Precision	Recall	F-measure	Error rate
Simple SVM	93.53	74.13	74.7	74	25.88
Multi SVM linear	91.31	65.25	69.99	65.11	34.75
Multi SVM product	91.34	65.38	69.79	65.33	34.63
Random Forest	93.45	73.81	74.78	73.59	26.19

Table 1. Classification evaluation over all modality classes

especially adapted to the high variability we can observe in the different classes. Thus, the resulting texture information may not be discriminative enough. Indeed, some images such as charts or drawing present almost no textural information. Moreover, by computing global statistics on these features, a lot of information which might be usefull for classification is discarded. In the present work, we did not make use of any RGB information, this would have definitely helped to discriminate images from classes which do not contain any color images.

5 Conclusion

In the present work, our goal was to get a first insight into the ImageCLEF challenge by taking part to its new subtask modality classification. We presented different approaches for the classification of medical images into 8 imaging modalities. As a feature representation, we extracted global statistics computed from the textural information contained in the images. We then investigated four classification methods based on SVMs and random forests. Approaches based on simple SVM and random forests give the best performance and achieve respectively an overall f-measure of 74.13% and 73.59%. After analyzing our preliminary results, there is clearly a lot of room for improvement concerning our feature representation. In future work, we will focus on defining new sparse and more discriminative representations of the data which should lead to better classification results.

References

1. H. Müller, J. Kalpathy-Cramer, I. Eggel, S. Bedrick, C. E. Kahn Jr. and W. Hersh, *Overview of the CLEF 2010 medical image retrieval track*, Working Notes of CLEF 2010, Padova, Italy, 2010.
2. J. Babaud, A. Witkin, M. Baudin and R. Duda, *Uniqueness of the gaussian kernel for scale-space filtering*, IEEE Trans. Pattern Anal. machine Intell., 1986
3. AK. Jain, F. Farrokhnia, *Unsupervised texture segmentation using Gabor filters*, Pattern Recognition, 1991
4. D. J. Field., *Relations between the statistics of natural images and the response properties of cortical cells*, Journal of The Optical Society of America A, 1987

5. P. Kovesi, *Image Features From Phase Congruency*, Videre: A Journal of Computer Vision Research. MIT Press, 1999
6. C. Cortes and V. Vapnik, *Support Vector Networks*, Mach. Learn.,1995
7. C.W. Hsu and C.J. Lin, *A comparison of methods for multi-class support vector machines*, IEEE Transactions on Neural Networks, 2002.
8. L. Breiman, *Random Forests*, Machine Learning, 2001
9. F.R. Bach, G.R.G Lanckriet and M.I. Jordan, *Multiple kernel learning, conic duality, and the SMO algorithm*, International Conference on Machine Learning, 2004
10. G.R.G. Lanckriet, N. Cristianini, L.E. Ghaoui, P. Bartlett and M.I. Jordan, *Learning the kernel matrix with semidefinite programming*, J.Machine Learning Research, 2004