Chapter 1

Medical image annotation [1]

Thanks to the rapid development of modern medical devices and the use of digital systems, more and more medical images are being generated. This has lead to an increase in the demand for automatic methods to index, compare, analyze and annotate them. Until 2005, automatic categorization of medical images was often restricted to a small number of classes. The ImageCLEF medical image annotation challenge was born in this scenario, proposing a task reflecting real life constraints of content based image classification in medical applications. In this chapter we report about our experience first as participants, then as co-organizers. This research activity started in 2007, supported by a 1-year IM2 fellowship. By leveraging over the initial IM2 support, in 2008 a 4-year project started (EMMA, Enhanced Multimodal Medical data Access), sponsored by the Halser foundation. Since 2009, B. Caputo has been an ImageCLEF task organizers, respectively for the medical annotation and robot vision tasks. Since 2013, she is main organizer of ImageCLEF.

1.1 Introduction

This chapter presents the algorithms and results of the Idiap team participation to the ImageCLEFmed annotation task in 2007, 2008 and 2009. The goal of the challenge was to develop an automatic image annotation system able to distinguish x-ray images on the basis of the body region, the biological system examined, the body orientation and the image modality. The idea is to exploit content based image analysis not using at all the textual information generally associated to medical images. A system which performs this task reliably would avoid the cost of manually annotating several terabytes of image data collected annually in radiology departments and it would help in image retrieval.

[1] This chapter was written by Barbara Caputo.

There are two main issues when working on big databases of medical images: intra-class variability vs inter-class similarity and data imbalance. The first problem is due to the fact that images belonging to the same visual class might look very different, while images that belong to different visual classes might look very similar. An example of this phenomenon is shown in Figs 1.1 and 1.2. In particular, Fig 1.1 shows some examples of visual variability within the class “foot, AP unspecified”: images of the same body region, with the same orientation, taken from different persons show high variability, because of differences in age or individual body structures. This problem can be solved with classification algorithms able to generalize well without compromising robustness. Fig 1.2 shows exemplar images of different classes which share some visual characteristics: “chest PA uspecified”, “chest, PA expiration”, “chest, AP inspiration”, and “chest, AP supine”. They must be classified differently because of clinical needs, but they present a strong visual similarity because they all contain the body part “chest”. Data imbalance is related to the natural statistics onset of diseases in the different parts of the body, thus it reflects the a priori probabilities of the routine diagnosis in a radiological clinic. To overcome both these problems an automatic annotation system should use the most discriminative information from the available data and it should be able to weigh properly the information coming from differently populated classes in the learning process.

For the CLEF challenge, the images were identified on the basis of the IRMA code (Lehmann et al., 2003). It is a multi-axial hierarchical scheme which adds a further difficulty in the annotation process asking for algorithms able to take advantage from the use of the unspecified character “0” or “*” when the confidence of the classifier's decision is not considered reliable.

In our experience as participants to the ImageCLEFmed challenge, we tackled all these problems and we proposed different discriminative solutions based on Support Vector Machines (SVM, Cristianini and Shawe-Taylor (2000)). In 2007 and 2008 our best run ranked first, while in 2009 the run reproducing the winning strategy of 2008 ranked second.

In the rest of the chapter we will focus on each of the described issues: Section 1.2 gives details about how we combined multiple cues to face the inter-class vs intra-class variability problem; Section 1.3 introduces our confidence-based approach to exploit the hierarchical structure of the data; Section 1.4 describes our strategy to overcome the data imbalance by creating virtual examples. Finally in Section 1.5 we describe our experimental setup and summarize our results. Conclusions are drawn in Section 1.6.

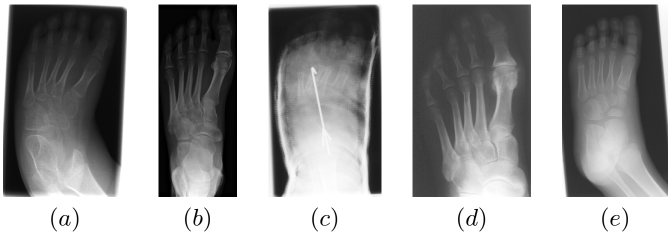


Figure 1.1: Images from the IRMA database used for the ImageCLEF challenge 2007 (Tommasi et al., 2011). Note the high visual variability within the images. They all belong to the same class annotated as: acquisition modality ‘overview image’; body orientation ‘AP unspecified’; body part ‘foot’; biological system ‘muscolosceletal’.

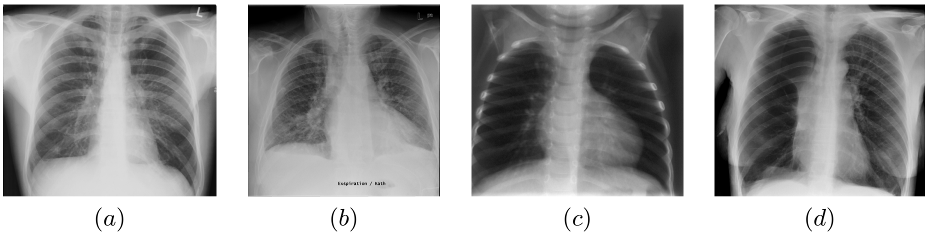


Figure 1.2: Images from the IRMA database used for the ImageCLEF challenge 2007 (Tommasi et al., 2011). Note the high visual similarity between the images. Each of them belongs to a different class. They all have as acquisition modality ‘high beam energy’, as body region ‘chest unspecified’, as biological system ‘unspecified’, but they differ for the body orientation: (*a*) ‘PA unspecified’, (*b*) ‘PA expiration’ (*c*) ‘AP inspiration’, (*d*) ‘AP supine’.

1.2 Multiple Cues for Image Annotation

Several authors tried to address the inter-class vs intra-class variability problem using local and global features, and more generally different types of descriptors, separately or combined together in a multiple cues approach (Müller et al., 2006; Güld et al., 2006; Florea et al., 2006). For some of these examples the performance was not very high. Still years of research on visual recognition in other domains have shown clearly that multiple cues methods outperform single-feature approaches (Matas et al., 1995; Orabona et al., 2012; Sun, 2003). To have the maximum advantage from cues integration, each feature should represent a different aspect of the data allowing for a more informed decision. Heterogeneous and complementary visual cues, bringing different information content, were successfully used in the past (Slater and Healey, 1995; Nilsback and Caputo, 2004; Gehler and Nowozin, 2009; Orabona et al., 2010, 2012). Regarding the integration techniques, they can all be reduced to one of these three approaches: *high-level*, *mid-level* and *low-level* integration (Sanderson and Paliwal, 2004; Polikar, 2006).

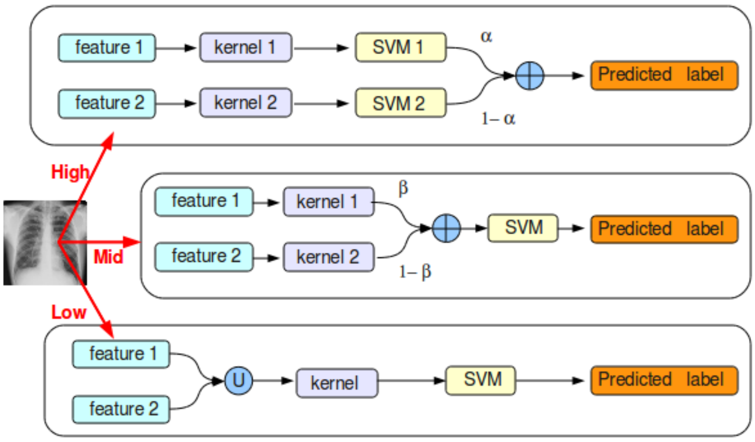


Figure 1.3: A schematic illustration of the high-level, mid-level and low-level cues integration approaches.

Figure 1.3 illustrates schematically the basic ideas behind these methods.

Participating to the ImageCLEF challenge we proposed a discriminative approach for integration of cues by defining three strategies, one for each of the possible level of cue integration. The methods used are described in detail in the following sections.

1.2.1 High-Level Integration

High-level cue integration methods start from the output of two or more classifiers dealing with complementary information. Each of them produces an individual hypothesis about the object to be classified. All those hypotheses are then combined together to achieve a consensus decision. We applied this integration strategy using the Discriminative Accumulation Scheme (DAS) proposed first in (Nilsback and Caputo, 2004). It is based on a weak coupling method called accumulation, which does not neglect any cue contribution. Its main idea is that information from different cues can be summed together.

1.2.2 Mid-Level Integration

Combining cues at the mid-level means that the different feature descriptors are kept separated, but they are integrated in a single classifier generating the final hypothesis. To implement this approach we developed a scheme based on multi-class SVM with a Multi Cue Kernel, K_{MC}. This new kernel combines different features ($T_p(I)$) extracted from the images (I):

$$K_{MC}(\{T_p(I_i)\}_p, \{T_p(I)\}_p) = \sum_{p=1}^{P} a_p K_p(T_p(I_i), T_p(I)), \quad \sum_{p=1}^{P} a_p = 1 \ . \tag{1.1}$$

1.2.3 Low-Level Integration

To combine cues it is also possible to use a low-level fusion strategy, starting from the descriptors and combining them in a new representation. In this way the cue integration does not directly involve the classification step. Here we used feature concatenation: two feature vectors f_i and c_i are combined into a single feature vector $v_i = (f_i, c_i)$ that is normalized to have its sum equal to one and is then used for classification. In this way the information related to each cue is mixed without a weighting factor that allows to control the influence of each information channel on the final recognition result. A general drawback of this method is that the dimension of the feature vector increases as the number of cues grows, implying longer learning and recognition times, greater memory requirements and possibly curse of dimensionality effects.

1.3 Exploiting the Hierarchical Structure of Data: Confidence Based Opinion Fusion

The evaluation scheme for the medical image annotation task addresses the hierarchical structure of the IRMA code considering the number of possible choices at each node and the position of each node in the hierarchy. So, wrong decisions in easy nodes were more penalized than wrong decisions in difficult nodes and mistakes at an early stage in the code were more costly than at a later stage. Moreover, the error evaluation method allowed the classifier to decide a “don’t know” at any level of the code, independently for each of the four axes: image modality, body orientation, body region and biological system (Lehmann et al., 2003).

Discriminative classifiers usually do not provide any out-of-the-box solution for estimating the confidence of the decision, but in some cases they can be transformed in opinion makers on the basis of the value of the discriminative function. In the case of SVM, it can be done considering the distances between the test samples and the classification hyperplane. This approach turns out to be very efficient due to the use of kernel functions and does not require additional processing in the training phase.

1.4 Facing the Class Imbalance Problem: Virtual Examples

Unbalanced datasets define a challenging problem in machine learning. Classifiers generally perform poorly on unevenly distributed datasets because they are designed to generalize from sample data and output the simplest hypothesis that best fit them. In a binary problem with negative instances which heavily outnumber the positive ones, this means to classify almost all instances as negative. On the other hand, making the classifier too specific may make it sensitive to noise and prone to overfitting. There are two known approaches to solve this problem. One is to bias the classifier so that it pays more attention to samples from poorly populated classes. This can be done, for instance, by increasing the penalty associated with misclassifying the class with few data with respect to the others. The second approach is to preprocess the data by resampling methods (Akbani et al., 2004). A possible alternative to resample consists in exploiting the known invariances of the data to generate new synthetic minority instances and rebalance the dataset. We adopted this solution.

1.5 Experiments

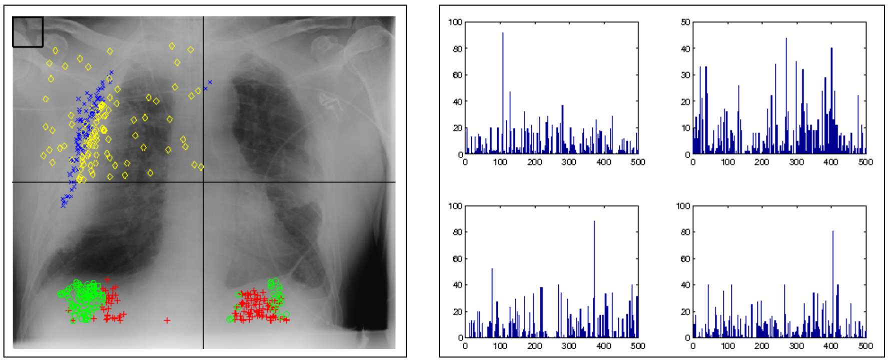
All the techniques described above were optimized on the released training set and applied on the unlabeled test set of the last three editions of the CLEF challenge. In the following subsections we summarize the specific choices done in running the experiments and the results obtained.

1.5.1 Features

To extract different and complementary information from the images, we chose two types of features that were then combined with the high, mid and low level integration strategies. In 2007 we combined a local (modSIFT) and a global feature (Raw Pixels), while in 2008 and 2009 we considered two different local cues (modSIFT and LBP).

ModSIFT. Scale Invariant Feature Transform (SIFT, (Lowe, 1999)) is a well known algorithm in computer vision used to detect and describe local features in images. We decided to use it adopting a bag of words approach: in analogy to text classification, the basic idea is to sample image patches and to match them to a set of prespecified “visual words”. The main implementation choices are (1) how to sample patches, (2) what visual patch descriptor to use, and (3) how to build the vocabulary.

Regarding point (1), we used random sampling. Regarding point (2), we decided to use a modified version of the SIFT descriptor, i.e. we extracted the points at only one octave, the one that gave us the best classification performance on a validation set, and we removed the rotation-invariance. We call the modified SIFT descriptor modSIFT. Regarding point (3), we built the vocabulary randomly sampling 30 points from each input image and extracting a modSIFT feature in each point. The visual words are created using an unsupervised K-means clustering algorithm. Note that in this phase both training and test images could be used, because the process does not need the labels. We chose K template modSIFTs with K equal to 500 so we defined a vocabulary with 500 words.



(*a*) (*b*)

Figure 1.4: (*a*) The four most present visual words in the image are drawn, each with a different color (better viewed in color). The square in the upper left corner represents the size of the patch used for computing the modSIFT descriptor. (*b*) Total counts of the visual words in the 4 subimages.

Finally, the feature vector for an image is defined by extracting a random collection of points from the images. The resulting distribution of descriptors in the feature space is then quantized in the visual words of the vocabulary and converted into a frequency histogram. To add some spatial information we decided to divide the images in four parts, collecting the histograms separately. In this way the dimension of the input space is multiplied by four (feature vector with $500 \times 4 = 2000$ elements) but in our tests we gained about 3% in classification performance. We extracted 1500 modSIFTs in each subimage: such dense sampling adds robustness to the process. Figure 1.4 shows an example of the extracted local features.

In 2008 and 2009 we slightly modified the modSIFT feature inspired by the approach in (Lazebnik et al., 2006). We added to the original vector the histogram obtained extracting the feature from the whole image producing a final vector of 2500 elements.

LBP. Local Binary Patterns (LBP, (Ojala et al., 2002)) have been used extensively in face recognition, object classification (Ahonen et al., 2006; Zhang et al., 2007) and also in the medical area (Unay et al., 2007; Oliver et al., 2007). Our preliminary results on a validation set showed that the best way to use LBP on the medical image database at hand was combining in a two dimensional histogram $\mathrm{LBP}_{8,8}^{riu2}$ together with $\mathrm{LBP}_{16,12}^{riu2}$ and concatenating it with the two dimensional histogram made by $\mathrm{LBP}_{16,18}^{riu2}$ together with $\mathrm{LBP}_{24,22}^{riu2}$. In this way a feature vector of 648 elements is obtained. Each image is divided in four parts, one vector is extracted from each subimage and from the central area and then they are concatenated producing a vector of 3240 elements (see Figure 1.5).

Raw Pixels. We used the raw pixels as simplest possible global descriptor. Preliminary results on a validation set showed that downscaling images

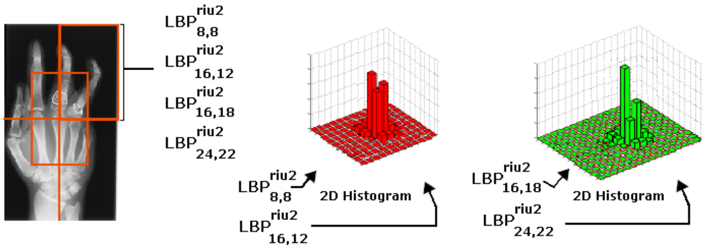


Figure 1.5: A schematic drawing which shows how we built the texture feature vector combining the 1-dimensional histograms produced by the LBP operators in 2-dimensional histograms.

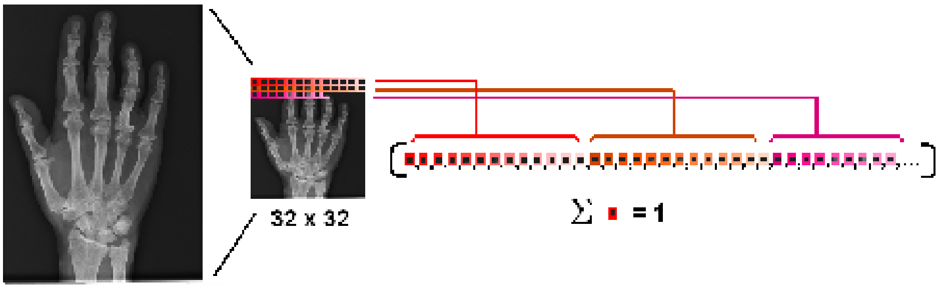


Figure 1.6: An example showing the raw pixel representation.

to 32x32 pixels didn't produce any significant difference than downscaling to 48x48, but the classification performance was better than that obtained on 16x16 images. So the images were resized to 32x32, regardless of the original dimension. The obtained 1024 pixel intensity values were then normalized to have sum equal to 1 and used as input features. Figure 1.6 shows how we built the raw pixel representation for each image.

1.5.2 Classifier

SVMs are a class of learning algorithms based on Statistical Learning Theory (Cristianini and Shawe-Taylor, 2000). Born as a linear classifier, SVM can be easily extended to non-linear domains through the use of kernel functions. The choice of the kernel heavily affects the performance of the SVM. We used an exponential χ^2 kernel for all the feature types and integration approaches. In our experiments we also tested the linear kernel and the RBF kernel, but all of them gave worse results than the χ^2. The parameter γ was tuned through cross-validation together with the SVM cost parameter C.

Even if the labels are hierarchical, we used the standard one-vs-all and one-vs-one multi-class approaches. We verified experimentally that with our features, the recognition rate was lower using an axis-wise classification.

This could be due to the fact that each super-class has a variability so high that our features are not able to model it, while they can very well model the small sub-classes.

1.5.3 Experimental Setup and Results

To obtain reliable results in the training phase we used all the images released from the CLEF organizers, not considering the distinction between training and validation when it was suggested. Our strategy was to create five disjoint train/test splits on which to optimize the learning parameters. The performance was evaluated on the basis of the error score, the same used in a second stage to rank the runs submitted to the challenge.

In 2008 and 2009, to take care of the class imbalance, the released database was divided into:

- rich_set: images belonging to classes with more than 10 elements. From this group we built 5 disjoint sets, rich_train$_i$/rich_test$_i$, where the test sets were created by randomly extracting five images for each of the classes. Note that in this way we automatically considered a normalization on the classes.

- poor_set: images belonging to classes with less than 10 elements. We used the whole poor_set as a second test set.

We trained the classifier on the rich_train$_i$ set and tested both on the rich_test$_i$ and on the poor_set, for each of the 5 splits. In this way, although the classes with few images were not considered in the training phase, we could evaluate the performance of the classifier to assign to those images the corresponding nearest class in the hierarchy. The error score was evaluated using the program released by the ImageCLEF organizers. The score values were normalized by the number of images in the corresponding test set, producing two average error scores. They were then multiplied by 500 and summed together supposing an ideal test set of 1000 samples constituted half by images from the rich_set and half by images form the poor_set. The average of the scores obtained on the 5 splits is an estimator of the expected value of the score. Each parameter in our methods was found by optimizing this expected score.

To evaluate the effect of introducing virtual examples in the poor_set we extracted from it only images belonging to classes with more than one element. We called this set poor_more. From it we created 6 poor_more_train$_j$/ poor_more_test$_j$ splits, where the train sets were defined extracting one image from each of the classes. Each poor_more_train set was enriched with the virtual examples as described in Section 1.4. Then we combined these sets joining rich_train$_i$ and poor_more_train$_j$ to build the training set and testing separately on rich_test$_i$ and poor_more_test$_j$.

Table 1.1: Ranking of our runs, name, score, gain with respect to the best run of other participants (RWTHi6-4RUN-MV3) in 2007. The Low level cues integration was used only after the challenge. "oa" and "oo" indicate respectively the one-vs-all and one-vs-one SVM multiclass extensions.

Rank	**Name**	**Score**	**Gain**
1	Mid_oa	26.85	4.08
	Low_oa	26.96	3.96
	Low_oo	26.99	3.93
2	Mid_oo	27.54	3.38
3	modSIFT_oo	28.73	2.20
4	modSIFT_oa	29.46	1.47
5	High	29.90	1.03
6	RWTHi6-4RUN-MV3	30.93	0
28	PIXEL_oa	68.21	−37.28
29	PIXEL_oo	72.41	−41.48

Tables 1.1, 1.2, 1.3 summarize all the results obtained by the Idiap team runs in 2007, 2008 and 2009 with the relative gain with respect to the best result from the other participating groups. In 2009 we took part to the ImageCLEFmed challenge organization, and we decided to simply reuse the best approaches proposed in 2008, submitting some baseline runs.

1.6 Conclusions

The Idiap team participated in the CLEF medical image annotation task from 2007 to 2009 proposing discriminative approaches coming from the image classification and recognition domain. The methods used are based on a combination of different local and global features and SVM as classifier, together with specific solutions to face the class imbalance problem and to exploit the hierarchical labeling structure of data. On the basis of the obtained results we can state that the adopted strategies are suited to solve the challenging issue of annotating a big medical image database.

Table 1.2: Ranking of our submitted runs, name, score and gain with respect to the best run of the other participants (TAU-BIOMED-svm_full) in 2008. The extension "virtual" stands for poor class enrichment by the use of virtual examples; "confidence" stands for the combination of the first two SVM margins for the confidence based opinion fusion. For all the runs we used the one-vs-all SVM multiclass extension.

Rank	**Name**	**Score**	**Gain**
1	Low_virtual_confidence	74.92	30.83
2	Low_virtual	83.45	22.30
3	Low_confidence	83.79	21.96
4	Mid_virtual_confidence	85.91	19.84
5	Low	93.20	12.55
6	modSIFT	100.27	5.48
7	TAU-BIOMED-svm_full	105.75	0
11	LBP	128.58	−22.83

Table 1.3: Ranking of our submitted runs, name, score and gain with respect to the best run of the other participants (TAUbiomed) in 2009. The extension "virtual" stands for poor class enrichment by the use of virtual examples; "confidence" stands for the combination of the first two SVM margins for the confidence based opinion fusion. For all the runs we used the one-vs-all SVM multiclass extension.

Rank	**Name**	**Score**	**Gain**
1	TAUbiomed	852.8	0
2	Low_virtual_confidence	899.16	−46.36
3	Low_confidence:	899.4	−46.6
4	Low	1039.63	−186.83
5	Low_virtual	1042	−189.2

Bibliography

Ahonen, T., Hadid, A., and Pietikainen, M. (2006). Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12):2037–2041.

Akbani, R., Kwek, S., and Japkowicz, N. (2004). Applying support vector machines to imbalanced datasets. In *Machine Learning: ECML 2004*, volume 3201 of *Lecture Notes in Computer Science*, pages 39–50.

Cristianini, N. and Shawe-Taylor, J. (2000). *An introduction to support vector machines: and other kernel-based learning methods.* Cambridge University Press.

Florea, F., Rogozan, A., Cornea, V., Bensrhair, A., and Darmoni, S. (2006). MedIC/CISMeF at ImageCLEF 2006: image annotation and retrieval tasks. In *Working Notes of the 2006 CLEF Workshop.*

Gehler, P. and Nowozin, S. (2009). On feature combination for multiclass object classification. In *Proc. of the IEEE International Conference on Computer Vision (ICCV'09).*

Güld, M., Thies, C., Fischer, B., and Lehmann, T. (2006). Baseline results for the ImageCLEF 2006 medical automatic annotation task. In *CLEF 2006 Proceedings*, volume 4730 of *Lecture Notes in Computer Science*, pages 686–689.

Lazebnik, S., Schmid, C., and Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, 2:2169–2178.

Lehmann, T. M., Schubert, H., Keysers, D., Kohnen, M., and Wein, B. B. (2003). The irma code for unique classification of medical images. In *Proc. SPIE*, volume 5033, pages 109–117.

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Proc. of the IEEE International Conference on Computer Vision (ICCV'90)*, volume 2, page 1150.

Matas, J., Marik, R., and Kittler, J. (1995). On representation and matching of multi-coloured objects. *Proc. of the IEEE International Conference on Computer Vision (ICCV'95)*, page 726.

Müller, H., Gass, T., and Geissbuhler, A. (2006). Performing image classification with a frequency-based information retrieval schema for ImageCLEF 2006. In *Working Notes of the 2006 CLEF Workshop*.

Nilsback, M. and Caputo, B. (2004). Cue integration through discriminative accumulation. *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'04)*, 2:578–585.

Ojala, T., Pietikäinen, M., and Mäenpää, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987.

Oliver, A., Lladó, X., Freixenet, J., and Martí, J. (2007). False positive reduction in mammographic mass detection using local binary patterns. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2007*, volume 4791 of *Lecture Notes in Computer Science*, pages 286–293.

Orabona, F., Jie, L., and Caputo, B. (2010). Online-batch strongly convex multi kernel learning. *Proc. of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*.

Orabona, F., Jie, L., and Caputo, B. (2012). Multi kernel learning with online-batch optimization. *Journal of Machine Learning Research*, 13:165–191.

Polikar, R. (2006). Ensemble based system in decision making. *IEEE Circuits and Systems Magazine*, 6(3):21–45.

Sanderson, C. and Paliwal, K. K. (2004). Identity verification using speech and face information. In *Digital Signal Processing*, pages 449–480.

Slater, D. and Healey, G. (1995). Combining color and geometric information for the illumination invariant recognition of 3-d objects. *Proc. of the IEEE International Conference on Computer Vision (ICCV'95)*, page 563.

Sun, Z. (2003). Adaptation for multiple cue integration. *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'03)*, page 440.

Tommasi, T., Barbara Caputo, ., Welter, P., Guld, M., and Deserno, T. (2011). Overview of the clef 2009 medical image annotation track. *Multilingual Information Access Evaluation II. Multimedia Experiments*, pages 85–93.

Unay, D., Ekin, A., Cetin, M., Jasinschi, R., and Ercil, A. (2007). Robustness of local binary patterns in brain MR image analysis. *Proc. of the 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS 2007)*, pages 2098–2101.

Zhang, L., Li, S., Yuan, X., and Xiang, S. (2007). Real-time object classification in video surveillance based on appearance learning. In *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'07)*.