IDIAP RESEARCH REPORT


# CLEF2008 IMAGE ANNOTATION TASK: AN SVM CONFIDENCE-BASED APPROACH

Tatiana Tommasi Francesco Orabona
Barbara Caputo

Idiap-RR-77-2008

DECEMBER 2008

**Centre du Parc**, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

# CLEF2008 Image Annotation Task: an SVM Confidence-Based Approach

Tatiana Tommasi, Francesco Orabona, and Barbara Caputo
Idiap Research Institute
Centre Du Parc, Rue Marconi 19
P. O. Box 592, CH-1920 Martigny, Switzerland
{ttommasi, forabona, bcaputo}@idiap.ch

**Abstract**

This paper presents the algorithms and results of our participation to the medical image annotation task of ImageCLEFmed 2008. Our previous experience in the same task in 2007 suggests that combining multiple cues with different SVM-based approaches is very effective in this domain. Moreover it points out that local features are the most discriminative cues for the problem at hand. On these basis we decided to integrate two different local structural and textural descriptors. Cues are combined through simple concatenation of the feature vectors and through the Multi-Cue Kernel. The trickiest part of the challenge this year was annotating images coming mainly from classes with only few examples in the training set. We tackled the problem on two fronts: (1) we introduced a further integration strategy using SVM as an opinion maker. It consists in combining the first two opinions on the basis of a technique to evaluate the confidence of the classifier's decisions. This approach produces class labels with "don't know" wildcards opportunely placed; (2) we enriched the poorly populated training classes adding virtual examples generated slightly modifying the original images. We submitted several runs considering different combination of the proposed techniques. Our team was called "idiap". The run using jointly the low cue-integration technique, the confidence-based opinion fusion and the virtual examples, scored 74.92 ranking first among all submissions.

## Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries; H.2.3 [**Database Managment**]: Languages—*Query Languages*

## General Terms

Measurement, Performance, Experimentation

## Keywords

Automatic Image Annotation, Cue Integration, Confidence Estimation, Virtual Examples, Support Vector Machines, Kernel Methods

# 1 Introduction

The rapid development of new medical image acquisition techniques and the widespread of computerized equipment to save transfer and store medical imagery in digital format have led to the need for new methods to manage and archive this data. Average-sized radiology departments produce nowadays several tera-bytes of data annually. Automatic image annotation systems turn out to be important tools to manage big databases, in avoiding manual classification errors and helping in image retrieval. The ImageCLEFmed challenge is an international event which gives the possibility to different research groups to benchmark their image annotation approaches. The aim is to find out how well current techniques can identify image modality, body orientation, body region and biological system examined based on the images.

In 2008 the ImageCLEFmed annotation task provided participants with 12076 x-ray images as training data spread across 197 classes. The task consisted in assigning the correct label to 1000 test images. To recognize these images, an automatic annotation system have to face two major problems: the intra-class variability vs inter-class similarity, and the data unbalance. The first problem arises from 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. The second one is connected to the natural disposition of organs in the human body and the frequency of diseases, which causes that some parts of the body are more likely to be object of image acquisition. The ImageCLEFmed organizers decided to focus on this second problem introducing in the training set 82 classes with a maximum of 6 images each and preparing a test set mainly with images from this low populated classes.

This paper describes the algorithms submitted by the "idiap" team as its second participation to the CLEF benchmark competition[1]. Last year we proposed different cue-integration approaches based on Support Vector Machine (SVM, [2]), using global and local features. They proved robust and able to tackle the inter-vs-intra class variability problem. Our run based on the use of the Multi-Cue Kernel (MCK, [18]) ranked first in 2007. After the competition we compared the results obtained by MCK with a scheme consistent in concatenating the different feature vectors. Results showed that the two methods do not produce significatively different results [18].

This year we decided to reuse both the above described methods changing the selected features into two different types of local descriptors: Scale Invariant Feature Transform (SIFT, [8]) and Local Binary Pattern (LBP, [11]). We also propose two strategies to tackle the unbalancing problem. On one hand we explore a technique to estimate the confidence of the classifier's decision and when it is not considered reliable, a soft decision is made using SVM as an opinion maker and combining its first two opinions to produce a less specific label. This approach was derived from the label hierarchical structure and the possibility to insert a "don't know" in some point in it. On the other hand we created examples for the classes with few images to enrich them. The new images were produced as slightly modified copies of the original ones through translation, rotation and brightness changes.

We submitted several runs, the results show that the classification performance increases passing from the use of a single cue (idiap-LBP score 128.58; idiap-SIFTnew score 100.27) to that of multiple cues (LOW_lbp_siftnew score 93.20), from the use of a hard decision (idiap-MCK_pix_sift score 313.01; LOW_lbp_siftnew score 93.20) to a soft decision through confidence based opinion fusion (idiap-MCK_pix_sift_2MARG score 227.82; LOW_2MARG score 83.79) and gets even better adding virtual examples in low populated classes (idiap-LOW_MULT_2MARG score 74.92).

The rest of the paper is organized as follows: section 2 describes the feature extracted from images and the two methods used to combine them. Section 3 gives details on the confidence based opinion fusion, while section 4 explains how we multiplied images to create virtual examples. Section 5 reports the experimental procedure adopted and the results obtained. Conclusions and outlook are given in Section 6.

[1] In 2007 the name was "BLOOM" due to our sponsors.

## 2 Cue Integration

The aim of the automatic image annotation task is to classify images into a set of classes based on the hierarchical IRMA code [7]. This code distinguishes images along the modality, body orientation, body region and biological system axis and errors in the annotation are counted depending on the level at which the mistake is made. Greater penalty is applied for incorrect classification then for a less specific one in the hierarchy. For each image the error ranges from 0 to 1 respectively if the image is correctly classified or if the predicted label is completely wrong. The error is normalized axis wise so that each axis contributes with a maximum of 0.25 to the score. It is also possible to assign a "don't know" label that counts half respect to an error.

We propose to extract a set of features from each image and to use then SVM to classify them. In the previous editions of the challenge, top-performing methods were based on the assumption that images consist of parts which can be modelled more or less independently. That methods used local features which thus seem to be the most discriminative cues for medical image annotation [16, 9]. Our past experience confirms this assumption, so this year we decided to explore two local approaches using SIFT and LBP based descriptors. We considered them separated and combined through two different integration schemes.

### 2.1 Feature Extraction

In 2007 for the medical annotation task we defined and used a modified version of the classical SIFT descriptor that we called modSIFT [18]. Patches were randomly sampled from images and the descriptor considered points at only one octave, discarding rotation invariance. We explored the "bag of words" approach for classification. This is based on the idea that it is possible to transform the images into a set of prespecified visual words, and to classify the images using the statistics of appearance of each word as feature vectors. We built the vocabulary randomly sampling 30 points from each training image and extracting a modSIFT in each point. The visual words were created using an unsupervised K-means clustering algorithm with K=500, that means considering a vocabulary with 500 elements. The feature vector for each image was then defined dividing the image in four parts, randomly extracting 1500 modSIFTs in each subimage, quantizing the resulting distribution of descriptors in the vocabulary and converting it into four histogram of votes that were then put in a row to build the feature vector of 2000 elements.

The two runs based on this feature ranked third and fourth in 2007, so we decided to reuse it doing only a slight modification, inspired by the approach in [6]. We added to the original feature vector the histogram obtained extracting modSIFT from the entire image. Actually we obtained this new part of the feature simply adding the four original histograms together and then concatenating the obtained values to the starting vector producing a vector of 2500 elements. We can say that doing this we are considering the image at two different space levels and in our preliminary tests this simple method brought a gain of about 2 score points.

Another approach that we explored was considering local texture features. As the x-ray images do not contain any color information, texture features play an important role for this task and in the past challenge editions they were used by several groups [3, 9]. We chose the Local Binary Pattern operator, a powerful method well known for its successes in face recognition and object classification [1, 20] and which recently achieved good results also in the medical area [19, 12]. The LBP basic idea is to build a a binary code that describes the local texture pattern in a circular region thresholding each neighborhood on the circle by the gray value of its center. After choosing the dimension of the radius $R$ and the number of points $P$ to be considered on each circle, the images are scanned with the LBP operator pixel by pixel and the outputs are accumulated into a discrete histogram. The operator is gray-scale invariant, moreover we used the rotational invariant LBP version which considers the uniform patterns ($\mathrm{LBP}^{riu2}_{P,R}$, see Figure 1).

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 $\mathrm{LBP}^{riu2}_{8,8}$ together with $\mathrm{LBP}^{riu2}_{16,12}$ and concatenating it with the two dimensional histogram made by $\mathrm{LBP}^{riu2}_{16,18}$ together with $\mathrm{LBP}^{riu2}_{24,22}$. In this way a feature vector of 648 elements is obtained. Each image is divided in

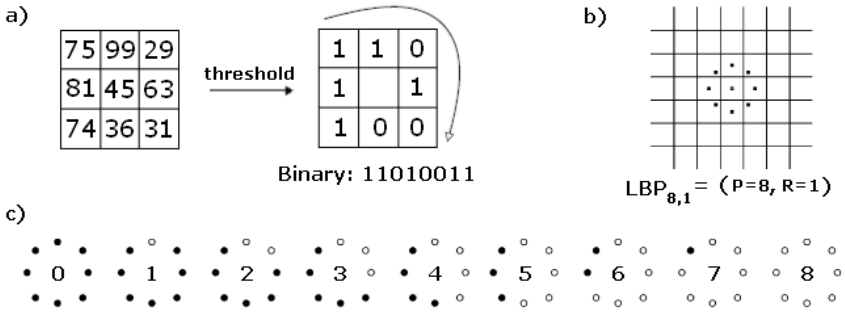


Figure 1: a) The basic LBP operator. b) Circularly symmetric neighbor set with radius of 1 pixel and 8 points on the circle. Samples that do not exactly match the pixel grid are obtained via interpolation. c) The rotation invariant (ri) binary patterns that can occur in the circular symmetric neighbor set of $LBP^{ri}_{8,1}$ are 36. Here we show just 9 of them corresponding to the uniform patterns with 2 spatial transition i.e. bitwise 0/1 changes (riu2). In the figure black and white circles correspond to the bit values of 0 and 1 in the 8-bit output of the LBP operator [11].

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 2).

## 2.2 Low and Mid Level Integration Schemes

In the computer vision and pattern recognition literature some authors have suggested different methods to combine information derived from different cues. They can all be reconducted to one of these three approaches: high-level, mid-level and low-level integration [13]. In the low-level integration scheme, image data or the corresponding features are combined together before classification; in the mid-level integration the different feature descriptors are kept separated but they are integrated in a single classifier generating the final hypothesis; finally a high-level cue integration starts from the output of two or more classifiers dealing with complementary information. The hypothesis are then combined together to achieve a consensus decision.

Considered our results in the ImageCLEF 2007 [18], we decided to use again the Multi-Cue Kernel as mid-level integration scheme and the concatenation of feature vectors as low-level integration.

The Multi-Cue Kernel is a linear combination of kernels each dealing with a single feature.

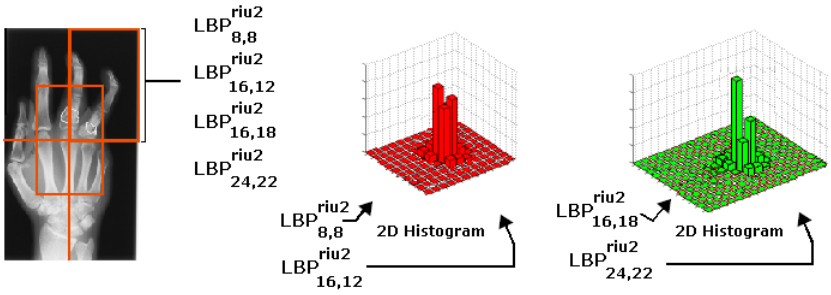


Figure 2: A schematic drawing which shows how we built the texture feature vector combining the 1-dimensional histograms produced by the LBP operators in 2-dimensional histograms.

Suppose that for each image $I_i$, we extract a set of $P$ different cues, $T_p(I_i)$, $p = 1 \dots P$. Hence we have $P$ different training sets and a corresponding set of $P$ kernels $K_p$, $p = 1 \dots P$. The Multi-Cue Kernel between two images, $I_i$ and $I_j$, is defined as

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

where $a_p \in \Re^+$ are weighting factors found through cross validation while determining the optimal separating hyperplane.

On the other hand, in the low-level scheme, the single features vectors are combined in a unique vector, that is normalized to have sum equal to one.

### 2.3 Classification

For the classification step we used an SVM with an exponential $\chi^2$ as kernel, for both the local structural and textural approaches and the cue-integration methods:

$$K(X, Y) = \exp\left(-\gamma \sum_{i=1}^{N} \frac{(X_i - Y_i)^2}{X_i + Y_i}\right) \ . \tag{2}$$

The parameter $\gamma$ was tuned through cross-validation. This kernel has been successfully applied for histogram comparison and it has been demonstrated to be positive definite [4], thus it is a valid kernel. In our experiments we used also the linear, RBF and histogram intersection kernel but all of them gave worse results than the $\chi^2$.

Even if the labels are hierarchical, we have chosen to use the standard multi-class approaches. This choice is motivated by the finding that, with our features, the error score was higher using an axis-wise classification.

## 3 Confidence Based Opinion Fusion

As previously described, the evaluation scheme for the medical image annotation task addresses the hierarchical structure of the IRMA code by allowing the classifier to decide a "don't know" at any level of the code, independently for each of the four axes. To effectively support this scheme, models which estimate the classifier's confidence in its decision could be useful, a fortiori if we consider the high unbalancing of the classes in the training set.

Discriminative classifiers usually do not provide any out-of-the-box solution for estimating confidence of the decision, but in some cases they can be transformed in opinion makers on the basis of the value of the used discriminative function. This gives the possibility to derive confidence information and hypothesis ranking from the produced opinions. In case of SVM, it can be done considering the distances between the test samples and the hyperplanes. The evaluation results very efficient due to the use of kernel functions and does not require additional processing in the training phase.

In the One-vs-All multiclass extension of SVM, if $M$ is the number of classes, $M$ SVMs are trained each separating a single class from all remaining ones. The decision is then based on the distances of the test sample, $\boldsymbol{x}$, to the $M$ hyperplanes, $D_j(\boldsymbol{x})$, $j = 1 \dots M$. The final output is the class corresponding to the hyperplane for which the distance is largest:

$$j^* = \underset{j=1\dots M}{\operatorname{argmax}} D_j(\boldsymbol{x}) \ . \tag{3}$$

If now we think of the confidence as a measure of unambiguity of the decision, we can define it as the difference between the maximal and the next largest distance:

$$C(\boldsymbol{x}) = D_{j^*}(\boldsymbol{x}) - \max_{j=1\dots M, j\neq j^*} D_j(\boldsymbol{x}) \ . \tag{4}$$

The value $C(\boldsymbol{x})$ can be thresholded for obtaining a binary confidence information. Confidence is then assumed if $C(\boldsymbol{x}) > \tau$ for threshold $\tau$.

Hence what we did was considering the first two margins produced by SVM corresponding to the distances of the test samples from the two closest hyperplanes. If the decision is not confident, that is $C(\boldsymbol{x}) < \tau$, then the label corresponding to the first two opinions are compared and where they differ we put a "don't know" term. We looked for the best threshold considering the results obtained in the preliminary validation phase and we adopt that for the subsequent experiments.

# 4 Adding Virtual Examples

To achieve good results in machine learning based classification, it is important to use training data which are sufficient not only in quality but also in quantity. For an SVM, working with classes very sparsely populated means that during the training phase, it is forced to individuate the best hyperplane which separate classes with few examples, to all the rest of the training set. Obviously the work done by the classifier in that condition can't be considered really reliable. To improve the reliability of the classification, we thought to enrich the poorly populated classes. Using virtual examples, i.e. artificially created images, is a well known method to expand the training data in an automatic way on the basis of a prior knowledge [10, 14, 15].

In one of their publications [5], people who collected and organized the IRMA database suggest that reasonably small transformations of certain image objects do not affect the class membership. So we produced modified copies of the released images in the subsequent way:

- each side increased of 100 pixels;
- each side increased of 50 pixels;
- each side decreased of 50 pixels;
- left rotation of 40 degrees;
- right rotation of 40 degrees;
- left rotation of 20 degrees;
- right rotation of 20 degrees;
- left shift of 50 pixels;
- right shift of 50 pixels;
- up shift of 50 pixels;
- down shift of 50 pixels;
- left (50 pixels) + up (50 pixels) shift;
- left (50 pixels) + down (50 pixels) shift;
- right (50 pixels) + up (50 pixels) shift;
- right (50 pixels) + down (50 pixels) shift;
- brightness increased (gray scale enhanced adding 20 to the original gray level and putting to 255 values higher than 255);
- brightness decreased (gray scale lowered subtracting 20 to the original gray level and putting to 0 the obtained negative values).

Thus for each of the images belonging to poorly populated classes we produced 17 copies using matlab scripts.

# 5 Experiments

Before starting our validation experiments, we studied in-depth how to divide the released database to consider the high unbalancing between classes. We decided to separate the training images in:

- rich_set: images belonging to classes with more than 10 elements. A total of 11947 images divided in 115 classes;
- poor_set: images belonging to classes with less than 10 elements. A total of 129 images divided in 82 classes.

From the first group we built 5 disjoint sets, rich_train$_i$/rich_test$_i$, each with of 11372/575 images, where the test sets were created randomly extracting five images for each of the 115 classes. On the other hand, we used the whole poor_set as a second test set. In this way, although the classes with few images are not considered in the training phase, we can evaluate the performance of the classifier to assign to that images the corresponding nearest class in the hierarchy. So 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. 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 to produce the value of the score on the test set of the challenge hypothesizing that it would have been constituted half by images from the rich_set and half by images form the poor_set. The expected value of the score is then defined as the average of the scores obtained on the 5 splits. Each parameter in our methods was found optimizing this expected score.

Our validation experiments started considering the local structural and textural approaches separately, applying the proposed experimental setup. We used the above procedure to select the best kernel parameters for the single-cue SVMs, giving the lowest combined error score described above. We adopted the same procedure for our validation experiments with the two cue-integration schemes.

On top of these experiments we applied, as second step, the confidence based opinion fusion technique described in Section 3. Both the single-cue and the multiple-cue runs were executed using the One-vs-All SVM multiclass extension and saving a file containing for each test image the values of the distances from the separating hyperplanes. We considered these files related to the rich and poor test sets produced by the classification with the best parameters found in the previous phase. The first two higher margins for every test images were subtracted and the difference compared to the threshold $\tau$ varying in $[0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]$. The best threshold was considered that producing the lowest expected score, using the procedure described above.

To evaluate the effect of introducing virtual examples in the classes belonging to the poor set we divided it in two parts:

- poor_one: images belonging to classes with only one element. A total of 53 images from 53 classes;
- poor_more: images belonging to classes with more than one element. A total of 76 images from 29 classes.

We only considered the second group and created 6 poor_more_train$_i$/poor_more_test$_i$ splits of 29/47 images, where the train sets were defined extracting one image from each of the 29 classes. We also introduced virtual examples as described in Section 4 such that each poor_more_train set was enriched with 29*17=493 images. 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$_j$ with $i = 1, 2, 3, 4, 5$ and $j = 1, 2, 3, 4, 5, 6$. We executed experiments with this setup and the best kernel parameters obtained form the previous single and multiple-cue experiments. The described procedure, for each $i, j$ couple produced again two classification outputs. The error scores were normalized and combined as described above. We also repeated this group of experiments without

| Rank | Name | Score | Gain |
|---|---|---|---|
| 1 | idiap-LOW_MULT_2MARG | 74.92 | 30.83 |
| 2 | idiap-LOW_MULT | 83.45 | 22.30 |
| 3 | idiap-LOW_2MARG | 83.79 | 21.96 |
| 4 | idiap-MCK_MULT_2MARG | 85.91 | 19.84 |
| 5 | idiap-LOW_lbp_siftnew | 93.20 | 12.55 |
| 6 | idiap-SIFTnew | 100.27 | 5.48 |
| 7 | TAU-BIOMED-svm_full | 105.75 | 0 |
| 11 | idiap-LBP | 128.58 | −22.83 |
| 19 | idiap-MCK_pix_sift_2MARG | 227.82 | −122.07 |
| 24 | idiap-MCK_pix_sift | 313.01 | −207.26 |

Table 1: Ranking of our submitted runs, name, score and gain respect to the best run of the other participants. The extension MULT stands for image multiplication, that isthe use of virtual examples. 2MARG stands for the combination of the first two SVM margins for the confidence based opinion fusion.

introducing the virtual examples and the score resulted lower of about 4 points on average showing that the addition of virtual elements is useful for the classification task.

Finally we applied the confidence based decision fusion on the output of the just presented experiments with the virtual examples in the training set. In this way we obtained our lowest expected score both for the single-cue and the multiple-cue approaches. Note that, independently of the selected feature or combination of features, applying together our two proposed methods always improves the score.

All the parameters of the validation phase were then used to run our submission experiments on the 1000 unlabelled images of the challenge test set using all the 12076 images of the original dataset as training. We submitted 9 runs. One of them (idiap-MCK_pix_sift) consisted simply in repeating our 2007 winner run, that is combining modSIFT and pixel features through MCK using exactly the same parameters of last year [17]. As expected, this run ranked last this year, due to the fact that the dataset varied a lot respect of 2007 and a new search for all the parameters was needed. It is interesting to note that simply applying the confidence based opinion fusion on the this run (idiap-MCK_pix_sift_2MARG) we have a gain in score of 85.19.

Considering that our validation results did not show great differences between the low-level and the mid-level integration scheme we decided to use just the low-level cue-integration scheme for sake of simplicity. We submitted only one MCK run using both the confidence based opinion fusion and the virtual examples. Hence the remaining runs consisted in:

- using the two new cues separately (idiap-SIFTnew, idiap-LBP);
- applying cue-integration (idiap-LOW_lbp_siftnew);
- combining cue-integration with the confidence based opinion fusion (idiap-LOW_2MARG);
- combining cue-integration with the introduction of virtual examples in the training set (idiap-LOW_MULT);
- combining cue-integration with the confidence based opinion fusion and the introduction of virtual examples in the training set (idiap-LOW_MULT_2MARG, idiap-MCK_MULT_2MARG).

The ranking, name and score of our submitted runs together with the score gain respect to the best run of other participants are listed in Table 1.

# 6 Conclusions

This paper presents a combination of three different strategies to face the medical image annotation in a highly unbalanced database with great inter-vs-intra class variability. The first consists in

combining cues through two different SVM approaches corresponding to a low-level and a mid-level integration scheme. The second allows to estimate the confidence of the classifier decision and, on this basis, to assign to a test image the class label corresponding to the hard decision of the classifier, or to a combination of the labels related to the first two produced opinions. The third consists in enlarging the training set through virtual examples defined as modified copies of the images in the less populated classes. The method obtained combining the low-level cue-integration scheme together with the confidence based opinion fusion and the introduction of virtual examples obtained a score of 74.92 ranking first among all submissions.

This work can be extended in many ways. First, it could be interesting to understand if the low-level cue-integration scheme results still better then the mid-level one when the number of combined cues grows. We could for example analyze what happens adding to the two presented local features a global one. Second, we would like to integrate the confidence estimation and the cue integration in a unique strategy. The classifier should measure its own level of confidence and, in case of uncertainty, to seek for extra information considering multiple cues, so to increase its own knowledge only when necessary. Third, here we have introduced virtual examples modifying the original images through translation, rotation and brightness changes. The results prove the effectiveness of this strategy, but we wonder if it is possible to avoid this passage. A solution could be to design different features which are able to capture the information coming from all the modified copies. This would make the classification accurate even working with few images. Finally, in this work we used the hierarchical structure of the data only to put in the image label some "don't know" terms. Moreover our preliminary results using the axis-wise classification did not produce good results. We want to study more deeply the hierarchical structure to understand if it is possible to exploit it to produce better classification performance. Future work will explore these directions.

## Acknowledgments

This work is part of the EMMA project and was supported by the Hasler Foundation and by the Blanceflor Boncompagni Ludovisi Foundation (www.haslerstiftung.ch, www.blanceflor.se). The support is gratefully acknowledged.

## References

[1] T. Ahonen, A. Hadid, and M. Pietikinen. Face description with local binary patterns: application to face recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28(12):2037–2041, 2006.

[2] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines (and Other Kernel-Based Learning Methods)*. Cambridge University Press, 2000.

[3] T. Deselaers, H. Müller, P. Clough, H. Ney, and T. Lehmann. The CLEF 2005 automatic medical image annotation task. *International Journal of Computer Vision*, 74(1), 2007.

[4] C. Fowlkes, S. Belongie, F. Chung, and J. Malik. Spectral grouping using the Nyström method. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 26(2):214–225, 2004.

[5] D. Keysers, J. Dahmen, H. Ney, B.B. Wein, and T.M. Lehmann. A statistical framework for model-based image retrieval in medical applications. *J. Electronic Imaging*, 12(1):59–68, 2003.

[6] S. Lazebnik, C. Schmid, Cordelia, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2169–2178, 2006.

[7] T.M. Lehmann, H. Schubert, D. Keysers, M. Kohnen, and B. B. Wein. The IRMA code for unique classification of medical images. In *Proc. of International Society for Optical Engineering*, pages 440–451, San Diego, CA, May 2003.

[8] D. G. Lowe. Object recognition from local scale-invariant features. In *Proc. of the International Conference on Computer Vision (ICCV)*, volume 2, pages 1150–1157, Washington, DC, USA, 1999. IEEE Computer Society.

[9] H. Müller, T. Deselaers, T.M. Deserno, P. Clough, E. Kim, and W.R. Hersh. Overview of the ImageCLEFmed 2006 medical retrieval and medical annotation tasks. In *Proc. of CLEF. Lecture Notes in Computer Science*, pages 595–608, 2006.

[10] P. Niyogi, F. Girosi, and T. Poggio. Incorporating prior information in machine learning by creating virtual examples. In *Proc. of IEEE*, volume 86, pages 2196–2207, 1998.

[11] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.

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

[13] C. Sanderson and K. K. Paliwal. Identity verification using speech and face information. *Digital Signal Processing*, 14(5):449–480, 2004.

[14] M. Sasano. Virtual examples for text classification with support vector machines. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*, 2003.

[15] B. Scholkopf, C. Burges, and V. Vapnik. Incorporating invariances in support vector learning machines. In *Proc. of Artificial Neural Network - ICANN*, volume 1112, pages 47–52, 1996.

[16] C.R. Shyu, C.E. Brodley, A.C. Kak, A. Kosaka, A. Aisen, and L. Broderick. Local versus global features for content-based image retrieval. In *Proc. of the IEEE Workshop on Content-Based Access of Image and Video Libraries*, pages 30–34, June 1998.

[17] T. Tommasi, F. Orabona, and B. Caputo. CLEF2007: Image annotation task: an svm-based cue integration approach. *Working Notes of the ImageCLEFmed 2007, Medical Image Annotation Task*, 2007.

[18] T. Tommasi, F. Orabona, and B. Caputo. Discriminative cue integration for medical image annotation. *Pattern Recognition Letters*, in Press, 2008.

[19] D. Unay, A. Ekin, M. Cetin, R. Jasinschi, and A. Ercil. Robustness of local binary patterns in brain MR image analysis. In *Proc. of the IEEE Engineering in Medicine and Biology Society*, pages 2098–2101, 2007.

[20] L. Zhang, S.Z. Li, X.T. Yuan, and S.M. Xiang. Real-time object classification in video surveillance based on appearance learning. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 17-22, pages 1–8, 2007.