An SVM Confidence-Based Approach to Medical Image Annotation

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 the "idiap" team participation to the ImageCLEFmed annotation task in 2008. On the basis of our successful experience in 2007 we decided to integrate two different local structural and textural descriptors. Cues are combined through concatenation of feature vectors and through the Multi-Cue Kernel. The challenge this year was to annotate images coming mainly from classes with only few training examples. We tackled the problem on two fronts: (1) we introduced a further integration strategy using SVM as an opinion maker; (2) we enriched the poorly populated classes adding virtual examples. We submitted several runs considering different combinations of the proposed techniques. The run jointly using the feature concatenation, the confidence-based opinion fusion and the virtual examples ranked first among all submissions.

1 Introduction

The rapid development of new medical image acquisition techniques and the widespread use 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. Automatic image annotation systems turn out to be important tools to manage big databases, in avoiding manual classification errors and helping in image retrieval. 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 has to face two major problems: intra-class variability vs inter-class similarity, and data imbalance. 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 these low populated classes.

** This work was supported by the EMMA project (B.C. and F.O.) thanks to the Hasler foundation (www.haslerstiftung.ch) and by the Blanceflor Boncompagni Ludovisi foundation (T.T., www.blanceflor.se).

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, [1]), 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, [2]) ranked first in 2007. After the competition we compared the results obtained by MCK with a scheme consistent in concatenating the different feature vectors. The benchmark showed that the two methods do not produce significatively different results [2]. 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, [3]) and Local Binary Pattern (LBP, [4]). We also propose two strategies to tackle the imbalancing problem. On one hand we explore a technique to estimate the confidence of the classifier's decision. 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. 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 one which combines feature concatenation with confidence based opinion fusion and introduction of virtual examples ranked first among all submissions.

2 Cue Integration

In the previous editions of the challenge, top-performing methods were based on local features which thus seem to be the most discriminative cues for medical image annotation [5, 6]. Our past experience confirms this assumption, so this year we decided to explore two local approaches. We considered them separated and combined through two different SVM-based integration schemes.

2.1 Feature Extraction

In 2007 for the medical annotation task we defined a modified version of the classical SIFT descriptor that we called modSIFT. We used it through a "bag of words" approach [2]. 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 [7]. We added to the original feature vector the histogram obtained extracting modSIFT from the entire image producing a vector of 2500 elements. In this way we are considering the image at two different space levels: in our preliminary tests this simple method brought a gain of approximately 2 score points.

As second local descriptor we chose the LBP operator, a powerful method well known in face recognition, object classification [8, 9] and also in the medical

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

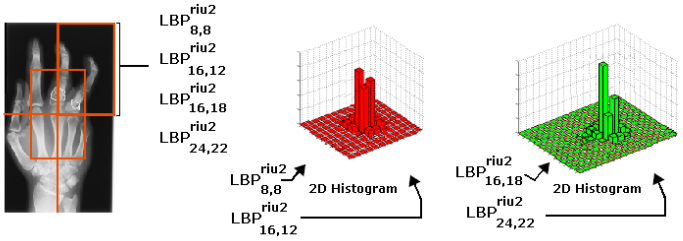

Fig. 1. A schematic drawing which shows how we built the texture feature vector combining the 1-dimensional histograms produced by the LBP operators in 2-dimensional histograms.

area [10, 11]. 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 [4]. The operator is gray-scale invariant, moreover we used the $riu2$ rotational invariant LBP version which considers the uniform patterns with two spatial transitions ($\mathrm{LBP}^{riu2}_{P,R}$, [4]).

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 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).

2.2 Low and Mid Level Integration Schemes

In the computer vision and pattern recognition literature some authors 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 [12]. Considered our results in the ImageCLEF 2007 [2], 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. 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, which is normalized to have sum equal to one.

2.3 Classification

For the classification step we used an SVM with an exponential χ^2 as kernel [13], 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 γ was tuned through cross-validation. In our experiments we used also the linear, RBF and histogram intersection kernel but all of them gave worse results than the χ^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

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 [14]. To effectively support this scheme, models which estimate the classifier's confidence in its decision could be useful. 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. In case of SVM, it can be done considering the distances between the test samples and the hyperplanes. 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. 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 at 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 to obtain a binary confidence information. Confidence is then assumed if $C(\boldsymbol{x}) > \tau$ for threshold τ. In the cases in which the decision is not confident, we decided to compare the labels corresponding to the first two margins and to put a "don't know" term in the points of the code in which they differ.

4 Adding Virtual Examples

An SVM, working with classes very sparsely populated is not able to produce reliable results. To create the models it is forced to individuate the best hyperplane which separates classes with few examples, to all the rest of the training set. To improve the classification reliability, we enriched the poorly populated classes. In [15] the creators of the IRMA corpus describe that small transformation of the images do not alter the class membership. So we produced modified copies of the training images increasing and decreasing each side (100, 50 pixels); rotating them right and left (20,40 degrees); shifting right, left, up, down and in the four diagonal directions (50 pixels); increasing and decreasing brightness (add and subtract 20 to the original gray level). Thus for each of the images belonging to poorly populated classes we produced 17 different versions.

5 Experiments

Before starting our validation experiments, we studied in-depth how to divide the released database to consider the high imbalancing 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. From this 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. Note that in this way we are automatically considering a normalization on the classes.
- poor_set: images belonging to classes with less than 10 elements. A total of 129 images divided in 82 classes. 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. 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 as if it was 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.

Cross validation was done considering for LBP SVM_C=[50 **100** 150 200], γ=[0.5 1.5 **2.5** 3.5] and for modSIFT SVM_C=[80 120 **160** 200], γ=[0.01 0.025 **0.05** 0.1]. For the two cue integration schemes we used: low level feature concatenation SVM_C=[50 **100** 150 200], γ=[0.5 **1** 2 4]; MCK SVM_C=[50 **100** 150 200], γ_{LBP}=[0.25 0.5 **1** 1.5], $\gamma_{modSIFT}$=[0.25 **0.5** 1 1.5], a_{LBP}=[0.4 **0.3** 0.2 0.1], $a_{modSIFT}$=[0.6 **0.7** 0.8 0.9]. The best parameters are in bold.

On top of these preliminary experiments we applied, 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. The first two higher margins for every test images were subtracted and the difference compared to the threshold τ varying in $[0.1, 0.2, \ldots 0.9]$. The best threshold led to the lowest 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, it contained a total of 76 images from 29 classes. From it we created 6 poor_more_train$_j$/poor_more_test$_j$ 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$_j$. We run 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 introducing the virtual examples and the score resulted lower of approximately 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. Independently of the selected feature or combination of features, applying together our two proposed methods improved the score.

Even if the cross validation experiments required a preliminary effort in computational resources and time to select the best parameters, the subsequent confidence based opinion fusion, introduction of virtual examples and the combination of these two strategies turned out to be very fast.

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 [2]. 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

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 is the use of virtual examples. 2MARG stands for the combination of the first two SVM margins for the confidence based opinion fusion.

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

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 imbalanced database with great inter-vs-intra class variability. The first consists in combining cues through two different SVM approaches. 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. 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. 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. Future work will explore these directions.

References

1. Cristianini N., Shawe-Taylor J.: An Introduction to Support Vector Machines (and Other Kernel-Based Learning Methods). Cambridge University Press, (2004)
2. Tommasi T., Orabona F., Caputo B.: Discriminative cue integration for medical image annotation. PRL, in Press (2008)
3. D. G. Lowe: Object Recognition from Local Scale-Invariant Features. ICCV, (1999)
4. Ojala T., Pietikainen M., Maenpaa T.: Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns. PAMI, (2002)
5. Shyu C.R., Brodley C.E., Kak A.C., Kosaka A., Aisen A., Broderick L.: Local versus global features for content-based image retrieval. CBAIVL, (1998)
6. Müller H., Deselaers T., Deserno T.M., Clough P. Kim E., Hersh W.R.: Overview of the ImageCLEFmed 2006 Medical Retrieval and Medical Annotation Tasks. Proc. of CLEF, (2006)
7. Lazebnik S., Schmid C., Ponce J.: Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. CVPR, (2006)
8. Ahonen T., Hadid A., Pietikinen M.: Face description with local binary patterns: application to face recognition. PAMI, (2006)
9. Zhang L., Li S.Z., Yuan X.T., Xiang S.M.: Real-time Object Classification in Video Surveillance Based on Appearance Learning. CVPR, (2007)
10. Unay D., Ekin A., Cetin M., Jasinschi R., Ercil A.: Robustness of Local Binary Patterns in Brain MR Image Analysis. EMBS, (2007)
11. Oliver A., Lladó X., Freixenet J., Martí J.: False Positive Reduction in Mammographic Mass Detection Using Local Binary Patterns. MICCAI, (2007).
12. Sanderson C., Paliwal K. K.: Identity Verification Using Speech and Face Information. Digital Signal Processing 14, 449–480 (2004)
13. Fowlkes C., Belongie S., Chung F., Malik J.: Spectral Grouping Using the Nyström Method. PAMI, (2004)
14. Lehmann T.M., Schubert H., Keysers D., Kohnen M., Wein B. B.: The IRMA code for unique classification of medical images. SPIE, (2003)
15. Keysers D., Dahmen J., Ney H., Wein B.B., Lehmann T.M.: A Statistical Framework for Model-Based Image Retrieval in Medical Applications. J. Electronic Imaging 12, 59–68 (2003)