IDIAP RESEARCH REPORT

Discriminative Cue Integration for Medical Image Annotation

Tatiana Tommasi [a] Francesco Orabona [a]
Barbara Caputo [a c]

IDIAP–RR 07-64

December 2007

[a] IDIAP Research Institute, P.O.Box 592, 1920, Martigny, Switzerland
[c] Ecole Polytechnique Fdrale de Lausanne (EPFL), Switzerland

IDIAP Research Institute **www.idiap.ch**
Av. des Prés–Beudin 20 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

Discriminative Cue Integration for Medical Image Annotation

Tatiana Tommasi Francesco Orabona Barbara Caputo

December 2007

Abstract. Automatic annotation of medical images is an increasingly important tool for physicians in their daily activity. Hospitals produce nowadays an increasing amount of data. Manual annotation is very costly and prone to human mistakes. This paper proposes a multi-cue approach to automatic medical image annotation. We represent images using global and local features. These cues are then combined together using three alternative approaches, a high-level, a mid-level and a low-level fusion scheme, all based on the Support Vector Machines (SVM) algorithm. We tested our methods on the IRMA database, and with the mid- and high-level integration scheme we did participate to the 2007 ImageCLEFmed benchmark evaluation, in the medical image annotation track. These algorithms ranked first and fifth respectively among all submission. Experiments using the low-level integration scheme also confirm the power of cue integration for this task.

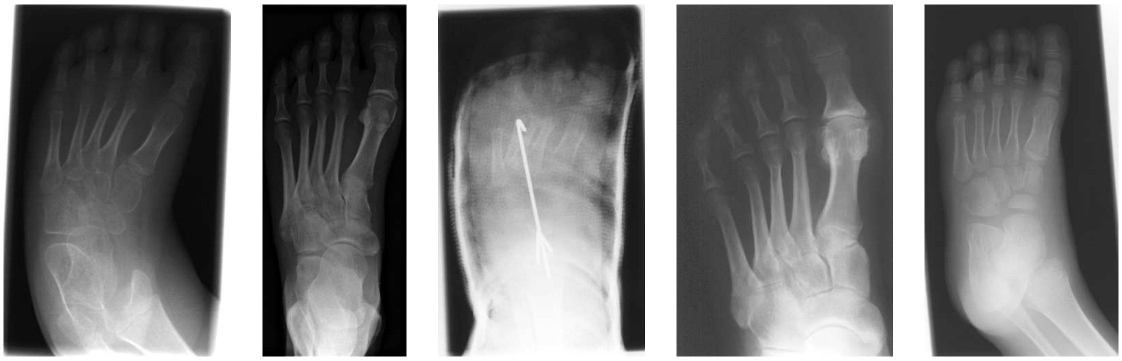
1 Introduction

The amount of medical image data produced nowadays is constantly growing, with average-sized radiology departments producing several tera-bites of data annually. The cost of manually annotating these images is very high; furthermore, manual classification induces errors in the tag assignment, which means that a part of the available knowledge is not accessible anymore to physicians [Gueld et al. (2002)]. This calls for automatic annotation algorithms able to perform the task reliably, and benchmark evaluations are thus extremely useful for boosting advances in the field. The ImageCLEFmed annotation task has been established in 2005, and in 2007 it has provided participants with 11000 training and development data, spread across 116 classes. The task consisted in assigning the correct label to 1000 test images. For further informations on the annotation task of ImageCLEF 2007 we refer the reader to [Mueller et al. (2007)].

An open challenge for automatic annotation of medical images is the intrinsically very high inter-class versus intra-class variability. Exemplar images of the same body part, taken from different individuals, might be quite different visually because of differences in age, or because the images were acquired in different hospitals, with slight variations in the image acquisition modality (see Figure 1 for some examples of visual variability within the body region class 'foot'). This problem is known in the literature as the *intra-class variability problem*, and it calls for classification algorithms able to generalize well while achieving robustness at the same time. Similarly, exemplar images of different classes might share some visual characteristics, so that they actually look similar. This is for instance the case for the classes 'chest unspecified', 'chest expiration', 'chest inspiration' and 'chest supine' (see Figure 2 for some exemplar images): they must be classified differently because of clinical needs, but they present a strong visual similarity because they all contain the body part 'chest'. This problem is known in the literature as the *inter-class variability problem*, and it calls for classification algorithms able to use the most discriminative information from the available data.

Several authors have tried to address this problem. State of the art approaches use global texture descriptors as features and discriminative algorithms, mainly SVMs, for the classification step. This approach has been adopted successfully by several groups [Mueller et al. (2006); W. Hersh (2006); Florea et al. (2006)]. Another popular approach is the use of local features combined to SVMs through ad-hoc kernels; notable examples are in [Mueller et al. (2006); Liu et al. (2006)] and they confirm the suitability of SVMs for the task, while showing the importance of using adhoc kernel functions. Local and global features, and more generally different types of descriptors, have been used separately or combined together in a multi-cue approach. The rationale behind the use of multiple cues is that these are expected to account for more and diverse data characteristics, thus they should lead to better performance. Several authors tried this strategy, proposing different integration schemes and testing various features [Mueller et al. (2006); Guld et al. (2006)]. Performance was lower than expected: the best result obtained by a multi-cue approach ranked 12th among 28 submissions in the ImageCLEF 2006 benchmark evaluation. Still, years of research on visual recognition in other domains have shown clearly that multiple-cue methods outperform single-feature approaches, provided that the features are complementary and that the integration scheme takes advantage from it.

In this paper we follow this route, and we propose to tackle the inter-class versus intra-class variability problem using a discriminative, SVM-based cue integration approach. We extract from images local and global descriptors, so to capture different kinds of information. The two feature types are combined together using three different SVM-based integration schemes. The first is the Discriminative Accumulation Scheme (DAS), a high-level strategy proposed first in [Nilsback and Caputo (2004)]. For each feature type, an SVM is trained and its output consists of the distance from the separating hyperplane. Then, the decision function is built as a linear combination of the distances, with weighting coefficients determined via cross validation. The second integration scheme is a mid-level strategy. It consists of designing a new Mercer kernel, able to take as input different feature types for each image data. We call it Multi Cue Kernel (MCK); the main advantage of this approach is that features are selected and weighted during the SVM training, thus the final solution is optimal as it minimizes the structural risk. The third integration scheme is a low-level approach. It



(*a*) (*b*) (*c*) (*d*) (*e*)

Figure 1: Images from the IRMA database. Intra-class variability within the class annotated as 1121-120-914-700, overview image, AP unspecified, foot, muscolosceletal system.

(*a*) (*b*) (*c*) (*d*)

Figure 2: All these images from the IRMA database 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".

creates a unique feature vector from the original two by concatenating them, and then uses an SVM for classification.

We tested our approaches on the IRMA database used for the ImageCLEFmed 2007 benchmark evaluation, in the medical image annotation track. DAS and MCK were submitted to the benchmark evaluation. They achieved respectively a score of 29.9 and 26.85, ranking fifth (DAS) and first (MCK) among all submissions. The low-level approach, developed after the submission deadline, achieved a score of 26.96, which would have corresponded to ranking second among all submissions. These results clearly prove the power of using multiple cues for the task, and underline the impact of the choice of the fusion strategy.

The rest of the paper is organized as follows: section 2 describes the two types of feature descriptors we used at the single cue stage, and gives a brief review of the theory behind SVMs. Section 3 gives details on the three alternative SVM-based cue integration approaches. Section 4 reports the experimental procedure adopted and the results obtained, with a detailed discussion on the performance of each algorithm. The paper concludes with a summary discussion.

2 Single Cue Image Annotation

The strategy we propose is to extract a set of features from each image and to use then a Support Vector Machine (SVM) to classify the images. We have explored a local approach, using SIFT descriptors, and a global approach, using the raw pixels.

2.1 Local Features

We explored the framework of "bag of words" for classification, a common concept in many state of the art approaches in images classification (e.g. [Nowak et al. (2006); Nister and Stewenius (2006)]) and also used in biomedical image classification [(Deselaers et al., 2006)]. In analogy to text classification, the basic idea is to sample image patches, following some specific criteria (e.g. an interest point detector), and to match these patches to a set of prespecified "visual words". Note that the ordering of the visual words is not used and only the statistics of appearance of each word is used to form the feature vectors. The main implementation choices are thus how to sample patches, what visual patch descriptor to use, and how to classify images based on the resulting global image descriptor.

Regarding the first issue, most of these systems are based on the use of the SIFT descriptor [Lowe (1999)]. The SIFT descriptors are designed to describe an area of an image in a way that is robust to noise, illumination, scale, translation and rotation changes. In particular the SIFT points are selected in the image as local maxima of the scale-space [Lowe (1999)], also fixing the strategy for the second issue. For the classification task we have chosen Support Vector Machines (SVMs), given their good performance and strong theoretical background, and their ability to work with features with almost any nature, through the use of the kernels.

Given the specific constraints of our classification task we decided to change in several ways the feature extraction part. The first thing is that due to the low contrast of the radiographs it would be difficult to use any interest point detector. Moreover it has been pointed out by different works and systematically verified by [Nowak et al. (2006)] that a dense random sampling is always superior to any strategy based on interest points detectors. The intuitive reason is that the criterion of the interest point detector is to find points easy to be tracked. On the other hand for classification task it could be important to have a dense set of points that capture more the appearance of the image than specific single points. So in our approach we randomly sampled each input image, extracting in each point a SIFT descriptor.

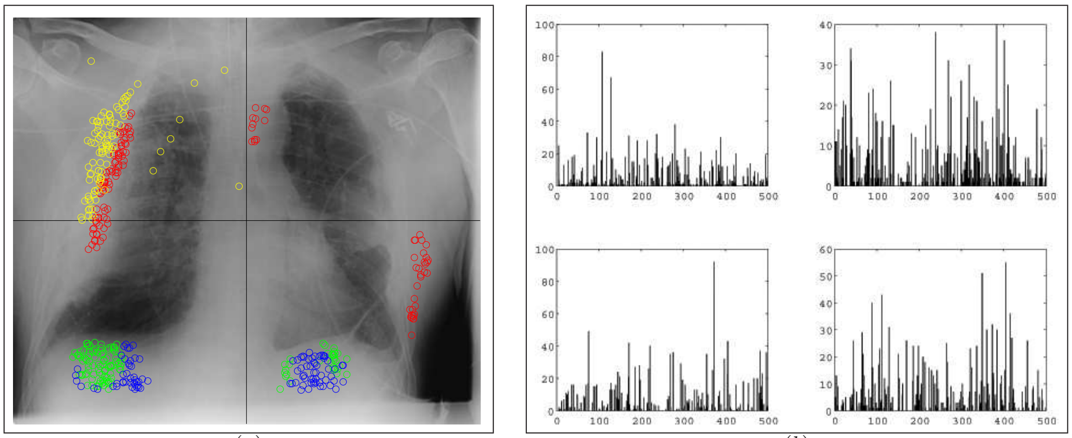
Another modification we made is based on the fact that the rotation invariance could be useless for the ImageCLEF classification task, as the various structures present in the radiographs are likely to appear always with the same orientation. Moreover the scale is not likely to change too much between images of the same class. Hence a rotation- and scale-invariant descriptor could discard useful information for the classification. So we extracted the points at only one octave, the one that gave us the best classification performances, and we removed the rotation-invariance. In this sense we have decoupled the extraction of a SIFT keypoint from the description of the point itself.

The resulting distribution of descriptors in the feature space is then quantized in visual words and converted into a histogram of votes. The visual words are created using an unsupervised K-means clustering algorithm, building a vocabulary of K template SIFTs. The input of the K-means is a random collection of SIFTs extracted from the training images. Note that in this phase also testing images could be used, because the process is not using the labels and it is unsupervised. Each image is then described with the raw counts of each visual word. We chose K equal to 500; various sizes of vocabulary were tested with no significant differences, so we chose the smaller one with good recognition performances.

To add some kind of spatial information, we divided the images in four subimages, collecting the histograms separately. In this way the dimension of the input space is multiplied by four, but in our tests we gained about 3% in classification performances. We extracted 1500 SIFTs in each subimage: such dense sampling adds robustness to the histograms. Figure 3 shows an example of the extracted local features.

2.2 Global Features

Another approach that we explored was the simplest possible global description: the raw pixels. The images were resized to 32x32 pixels, regardless of the original dimension, and normalized to have sum equal to 1, then the 1024 values were used as input features. This approach is at the same time a



(a) (b)

Figure 3: (a) The four most present visual words in the image are drawn, each with a different color (better viewed in color), and (b) total counts of the visual words in the 4 subimages.

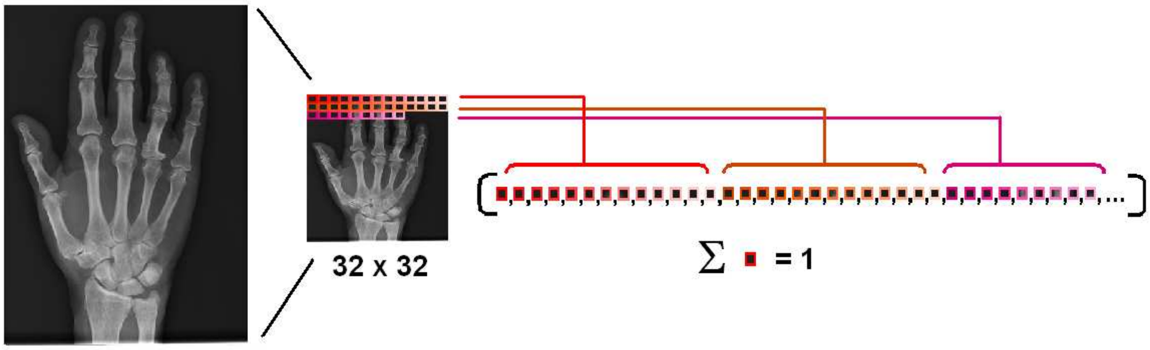


Figure 4: An example showing the raw pixel representation

baseline for the classification system and a useful "companion" method to boost the performance of the SIFT based classifier (see section 3). Figure 4 shows how we built the raw pixel representation for each image.

2.3 Support Vector Machines

Introduced in the early 90s by Boser, Guyon and Vapnik [Boser et al. (1992)], SVMs are a class of kernel-based learning algorithms deeply rooted in Statistical Learning Theory [Vapnik (1998)], now extensively used in, e.g., speech recognition, object classification and function approximation with good results [Cristianini and Shawe-Taylor (2000)]. Born as a linear classifier, SVM can be easily extended to non-linear domains through the use of the *kernel* functions. The kernels implicitly map the input space in a higher dimensional space, even with infinite dimension. At the same time the generalization power of the classifier is kept under control by a regularization term that avoid overfitting in such high dimensional spaces [Cristianini and Shawe-Taylor (2000)].

The choice of the kernel hence heavily affects the performance of the SVM: prior knowledge about the structure of the data can help in this choice. We have used an exponential χ^2 as kernel, for both the local and global approaches:

$$K(\mathbf{x}, \mathbf{y}) = \exp\left(-\gamma \sum_{i=1}^{N} \frac{(x_i - y_i)^2}{x_i + y_i}\right) \quad . \tag{1}$$

The parameter γ was tuned through cross-validation together with the SVM-cost parameter C (see section 4). The choice is motivated by the fact that the inputs are histograms and it is known that the χ^2 similarity measure works quite well for histograms comparison. Note that the global features also can be considered histogram-like inputs, given the normalization to sum equal to 1. Moreover it has been demonstrated to be positive definite by [Fowlkes et al. (2004)], thus it is a valid kernel [Cristianini and Shawe-Taylor (2000)]. We tested also linear kernel and the RBF kernel, but all of them gave worst results.

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 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. In particular we have tested both one-vs-one and one-vs-all multi-class extensions for SVM paying attention to the number of support vectors needed for the classification to have an idea of the difficulty of the task.

3 Multi Cue Image Annotation

Due to the fundamental difference in how local and global features are computed, it is reasonable to suppose that the two representations provide different kinds of information. Indeed, psychophysical evidence suggest that natural vision-based classification tasks are performed better when multiple visual cues can be combined to reduce ambiguity [Tanaka et al. (1991)]. Thus, we expect that by combining multiple cues through an integration scheme, we will achieve a better performance, namely higher classification performance and higher robustness.

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. Figure 5 illustrates schematically the basic ideas behind these three approaches; for a review on the topic we refer the reader to [Polikar (2006)]. We tested an SVM-based high-level integration scheme on the task at hand, namely the Discriminative Accumulation Scheme (DAS, [Nilsback and Caputo (2004)]). In this method each single cue first generates a set of hypotheses on the correct label of the test image, and then those hypotheses are combined together so to obtain a final output. The algorithm is revised in section 3.1. Another possible strategy is mid-level integration, where the features are merged during the classification step. To this end, we designed a new class of kernels, the Multi-Cue Kernel (MCK), that accepts as input different cues while building a unique optimal separating hyperplane. This new kernel is described in details in section 3.2. Finally we decided to use a low level integration scheme. This kind of approach is based on concatenating existing feature vectors in a new one, so in a sense it builds a new representation. It is questionable if this approach can solve the robustness problem because if one of the cues gives misleading information it is possible that the new feature vector will be adversely affected. A description of the chosen strategy is given in section 3.3. Experiments testing the effectiveness of these methods are then reported in section 4.

3.1 High-level Cue 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, so to achieve a consensus decision. In this paper we applied this integration strategy using the Discriminative Accumulation Scheme [DAS,

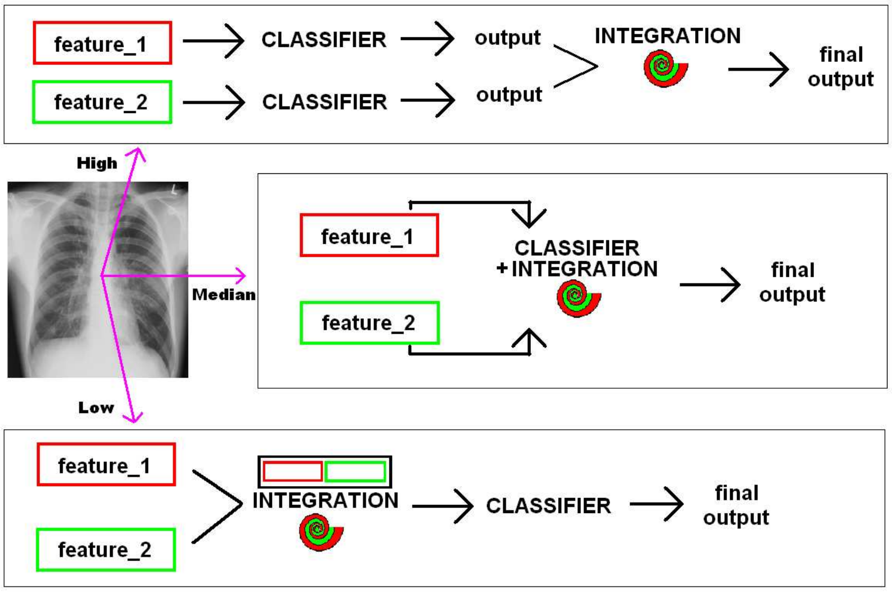


Figure 5: A schematic illustration of the high-level, mid-level and low-level cue integration approaches.

Nilsback and Caputo (2004)]. It is based on a weak coupling method called accumulation, which does not neglect any cue contribution. The DAS main idea is that information from different cues can be summed together.

Suppose we are given M object classes and for each class, a set of N_j training images $\{I_i^j\}_{i=1}^{N_j}$, $j = 1, \ldots M$. For each image, we extract a set of P different cues:

$$T_p = T_p(I_i^j), \quad p = 1 \ldots P \tag{2}$$

so that for an object j we have P new training sets $\{T_p(I_i^j)\}_{i=1}^{N_j}, j = 1, \ldots M, p = 1 \ldots P$. For each we train an SVM. Kernel functions may differ from cue to cue and model parameters can be estimated during the training step via cross validation. Given a test image $\hat{I}$ and assuming $M \geq 2$, for each single-cue SVM we compute the distance from the separating hyperplane:

$$D_j(p) = \sum_{i=1}^{m_j^p} \alpha_{ij}^p y_{ij} K_p\left(T_p(I_i^j), T_p(\hat{I})\right) + b_j^p. \tag{3}$$

After collecting all the distances $\{D_j(p)\}_{p=1}^P$ for all the j objects $j = 1, \ldots, M$ and the p cues $p = 1, \ldots, P$, we classify the image $\hat{I}$ using the linear combination:

$$j^* = \underset{j=1}{\overset{M}{\operatorname{argmax}}}\{\sum_{p=1}^{P} a_p D_{j(p)}\}, \quad a_p \in \Re^+. \tag{4}$$

The coefficients $\{a_p\}_{p=1}^P$ are evaluated via cross validation during the training step.

3.2 Mid-level Cue Integration

Combining two cues at a median level means that the different features 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 extracted from the images. The Multi Cue Kernel is a Mercer kernel, as positively weighted linear combination of Mercer kernels are Mercer kernels themselves [Cristianini and Shawe-Taylor (2000)]:

$$K_{MC}(\{T_p(I_i)\}_p, \{T_p(I)\}_p) = \sum_{p=1}^{P} a_p K_p(T_p(I_i), T_p(I)). \tag{5}$$

In this way it is possible to perform only one classification step, identifying the best weighting factors a_p while determining the optimal separating hyperplane; this means that the coefficients a_p are guaranteed to be optimal.

3.3 Low-level Cue Integration

To combine two or more image features it is possible to start from the descriptors, and to combine them together in a new representation. In this way the cue integration does not directly involve the classification step. This fusion strategy is called low-level. For the problem at hand we chose feature concatenation as the fusion approach: two feature vectors f_i and c_i are concatenated into a single feature vector $v_i = (f_i, c_i)$ that is normalized to one and is then used for classification. In this fusion strategy 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. In general terms a drawback of this method is that the dimension of the feature vector increases as the number of of cues grows, implying longer learning and recognition times, greater memory requirements and possibly curse of dimensionality effects. Moreover, it is not always possible to use the low-level integration approach: there are features that have a variable number of vector's elements per image, while some other have a defined number of them. Due to their intrinsic nature, the first ones ask for specialized classification algorithms and it is not possible to combine them together with vectors of the second kind.

4 Experiments

The database for the CLEF medical image annotation task was provided by the IRMA group from the University Hospital of Aachen, Germany. It consists of 11000 fully classified anonymous radiographs taken randomly from medical routine and 1000 radiographs for which the classification labels were not available to the ImageCLEF participants. The images' identity is defined through the IRMA code TTTT-DDD-AAA-BBB, a multiaxial classification method based on the acquisition modality (T=technique), the body orientation (D=direction), the body region (A=anatomy) and the biological system examined (B). This code is strictly hierarchical: each sub-code element to the right is connected to only one code element to the left. The way in which the classification performance is judged depends on the level of the hierarchy at which the error is done. Wrong decisions are more penalized if they are at an early stage than at a later stage in the code. The error count produces a score, the lower it is, the better the annotation is. For further details on the database and the ImageCLEF benchmark evaluation for the medical annotation task, we refer the reader to [Deselaers et al. (2008)].

The original dataset was divided in two parts: training and validation. In order to obtain reliable results, we merged them together and extracted 5 random and disjoint train/test splits of 10000/1000 images. As a preliminary step we run experiments to find the best kernel parameters through cross validation, the obtained results were then used to run our submission experiments on the 1000 unlabeled images of the challenge test set using all the 11000 images of the original dataset as training. We considered as the best parameters the one giving the lower average score on the 5 splits. Note that, due to the score evaluation method, the best score does not correspond necessarily to the best recognition rate.

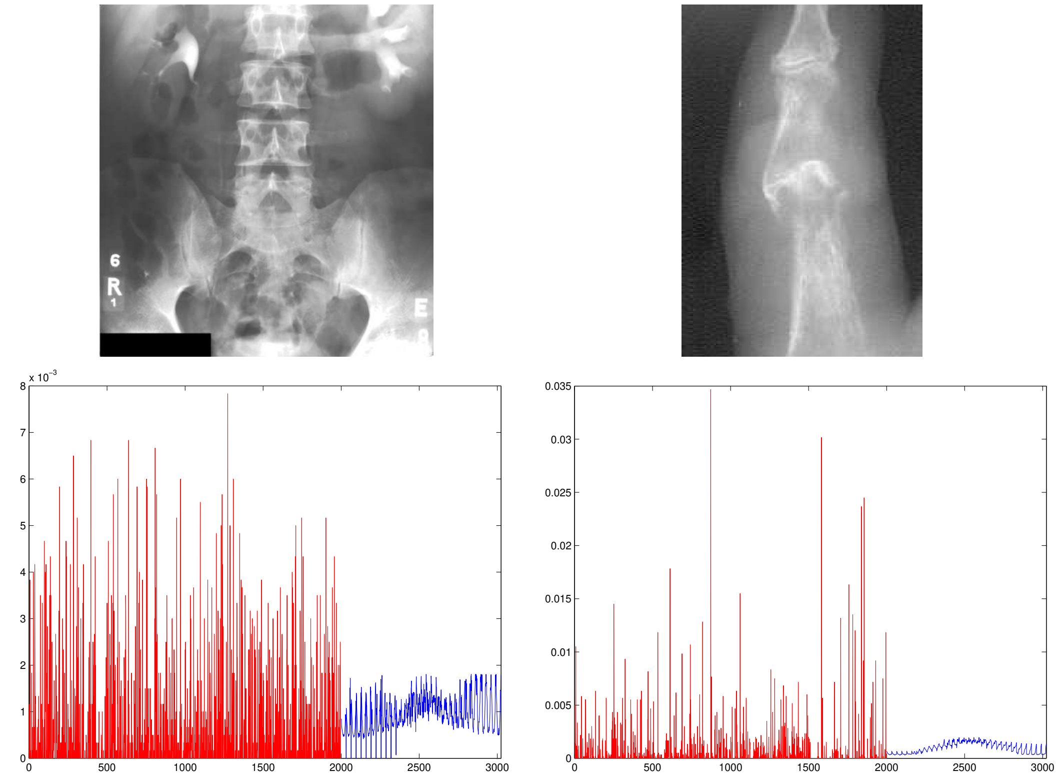


Figure 6: Local (in red) and global features (in blue) corresponding to the two images. The better discriminative power of the local features compared to local ones have an intuitive explanation in their behavior to different images (better seen in color).

4.1 Single Cue Annotation Experiments

Our experiments started evaluating the performance of local and global features separately before testing our integration methods. Besides obtaining the optimal kernel parameters, these experiments showed that the SIFT features outperform the raw pixel ones (see Table 1). This result could have been predicted, since the results of the ImageCLEF 2006 competition showed that local features are generally more informative than global features for the annotation task [Liu et al. (2006)]. We can say that global feature are able to retain information on the "gist" of the image as a source of context, while the local features capture the details, and thus they manage better the inter and intra-class variability. In our specific case, SIFT features related to two similar images look very different, while the correspondent raw pixel features are more similar, as shown in Figure 6. Considering the one-vs-one SVM multiclass extension, SIFT features need a lower number of support vectors than PIXEL and the same happens for the one-vs-all approach (see respectively SIFT_oo and SIFT_oa in Table 1).

4.2 Multi-Cue Annotation Experiments

We adopted the same described experimental setup for DAS, MCK and the Low Level cue integration method (LL). In particular for DAS we used the distances from the separating hyperplanes associated with the best results of the single-cue step, so the cross validation was used only to search the best weights for cue integration. On the other hand, for MCK the cross validation was applied to look for the best kernel parameters and the best feature's weights at the same time. In both cases weights could vary from 0 to 1. The Low Level integration method combines the global descriptors and the local descriptors in a unique feature vector. This gives rise to an experimental approach identical to that used for the single cue annotation: we applied cross validation to identify the optimal kernel parameters.

Table 1 shows that our two runs based on the MCK algorithm ranked first (score 26.85) and second (score 27.54) among all the challenge submissions, but considering all our experiments the mid-level cue integration approach shares the highest rank positions with the low-level integration scheme (LL_oa score 26.96, LL_oo score 26.99). These results state the effectiveness of using multiple cues for automatic image annotation. It is interesting to note that even if DAS has a higher recognition rate, its score is worse than that obtained using the feature SIFT alone. This could be due to the fact that when the label predicted by the global approach, the raw pixels, is wrong, the true label is far from the top of the decision ranking. Regarding the SVM multiclass extension, the one-vs-all overcome the one-vs-one.

4.3 Discussion

Table 1 summarizes all the relevant information about our experiments: the challenge ranking, name, best kernel parameters, best SVM-cost parameter C, best feature's weights, number of support vectors, score, gain respect to the best run of other participants and recognition rate. The LL integration approach was tested after the end of the ImageCLEF competition so our LL results are not part of the official rating.

As we could expect, for the MCK and DAS integration methods the best SIFT weights (see 4 and 5) turn out to be higher than the raw pixel's ones. In the LL integration approach the mixed features are not relatively weighted, nevertheless the obtained results tells that the two features cooperate well to classify images.

The number of support vectors for the MCK run using one-vs-one multiclass SVM extension (MCK_oa) is slightly higher than that used by the single cue SIFT_oa but lower than that used by PIXEL_oa. For the MCK run using one-vs-one multiclass SVM extension (MCK_oo) the number of support vectors is even lower than that of both the single cues SIFT_oo and PIXEL_oo. These results show that combining two features with the MCK algorithm can simplify the classification problem. Comparing the number of support vectors for the LL cue integration approach with that related to the single cue experiments leads to the same conclusions reached for MCK.

For DAS we counted the support vectors summing the ones from SIFT_oa and PIXEL_oa but considering only once the support vectors associated with the training images that resulted in common between the single cues. The number of support vector for DAS exceeds that obtained for both MCK_oa and MCK_oo showing a higher complexity of the classification problem.

Table 2 shows in details some examples of classification results. The second, third and fourth column contain examples of images misclassified by one of the two cues or by both, but correctly classified by DAS, LL_oa and MCK_oa. The fifth column shows an example of an image misclassified by both cues and by DAS but correctly classified by LL_oa and MCK_oa. The sixth column contains an image misclassified by the two cues, by DAS and LL_oa but recognized by MCK_oa. It is interesting to note that combining local and global features can be useful to recognize images even if they are extremely partial and compromised by the presence of artifacts that for medical images can be prosthesis or reference labels put on the acquisition screen.

The difference between the single-cue approach and the multi-cue integration approach can be better evaluated considering the confusion matrices. They are shown as images in Figure 7. We used a colormap corresponding to the number of images varying from zero to five to let the misclassified images stand out. It is clear that single-cue and multi-cue methods differ principally for how the wrong images are labeled. The more matrices present sparse values out of the diagonal and far away from it, the worse the method is.

Finally, trying to do a differential analysis on the three cue integration methods used we can say that:

- the high level integration approach is not able to work at its best if the basic features produce very different hypothesis. We can suppose that this kind of behavior is not due to the Discriminative

Rank	Name	γ_{sift}	γ_{pixel}	C	a_{sift}	a_{pixel}	#SV	Score	Gain	Rec. rate
1	MCK_oa	0.5	5	5	0.80	0.20	7916	26.85	4.08	89.7%
	LL_oa	1		5			8095	26.96	3.96	89.1%
	LL_oo	0.03		60			6958	26.99	3.93	89.3%
2	MCK_oo	0.1	1.5	20	0.90	0.10	7037	27.54	3.38	89.0%
3	SIFT_oo	0.05		40			7173	28.73	2.20	88.4%
4	SIFT_oa	0.25		10			7704	29.46	1.47	88.5%
5	DAS	0.25	5	10	0.76	0.24	9090	29.90	1.03	88.9%
28	PIXEL_oa		5	10			8329	68.21	−37.28	79.9%
29	PIXEL_oo		3	20			7381	72.41	−41.48	79.2%

Table 1: Ranking of our runs, name, best kernel parameters, best SVM-cost parameter C, best feature's weights, number of support vectors, score, gain respect to the best run of other participants and recognition rate. LL stand for low-level cue combination: this approach has been developed after the CLEF competition. The extensions _oo and _oa refer respectively to the one-vs-one and one-vs-all SVM multiclass approaches.

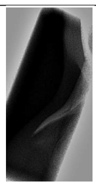
					
PIXEL_oa	11°	**1°**	12°	5°	6°
SIFT_oa	**1°**	2°	2°	5°	16°
DAS	**1°**	**1°**	**1°**	2°	6°
LL_oa	**1°**	**1°**	**1°**	**1°**	3°
MCK_oa	**1°**	**1°**	**1°**	**1°**	**1°**

Table 2: Example of images misclassified by one or both cues and correctly classified by DAS, LL or MCK.The values correspond to the decision rank.

Accumulation Scheme itself but in general it is a problem connected to combining the output of classifiers working on different cues. Some preliminary experiments using the voting scheme confirm this hypothesis;

- the low and the mid level integration methods combining the cues before and during the classification process seem able to better take advantage of the different information associated to the features. The mid-level approach is more refined and complex than the low-level approach thanks to the possibility to weight the features, and although using two cues its results are similar to that obtained by the low-level, MCK would probably gain in performance increasing the number of cues while the LL would suffer for the growth of the integral-feature dimension. Moreover with MCK it is possible to combine cues of an intrinsically different nature simply choosing for each the right kernel, while it could not always be possible to concatenate them in a single vector.

5 Conclusions

This paper presented a discriminative multi-cue approach to medical image annotation. We combined global and local information using three alternative fusion strategies: we combined features together in a unique descriptor, we used the discriminative accumulation scheme proposed first by [Nilsback and Caputo (2004)] and a new class of kernels, the multi cue kernel, able to take as input different cues while keeping them separated during the optimization process. This last method gave the best

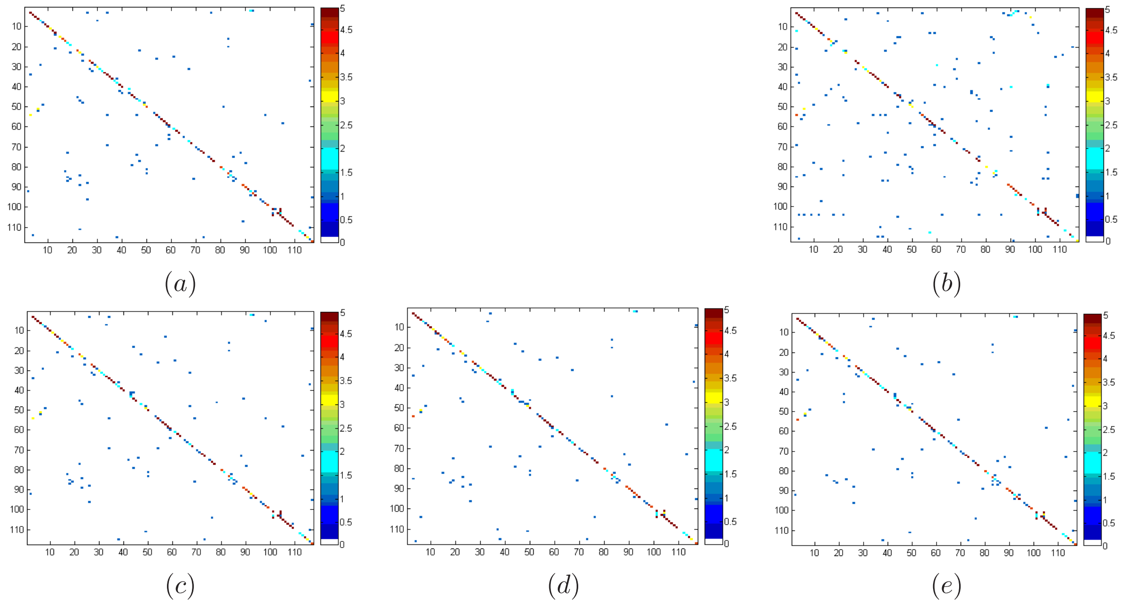


Figure 7: These images represent the confusion matrices respectively for (a) SIFT_oa, (b) Pixel_oa, (c) DAS and (d) MCK_oa (e) LL_oa. We used a colormap corresponding to the number of images varying from zero to five to let the misclassified images stand out. All the position in the matrices containing five or more images appear dark red (better seen in color).

performance in the ImageCLEF 2007 benchmark evaluation, obtaining a score of 26.85, which ranked first among all submissions. All our cue integration schemes achieved very high scored, proving the effectiveness of our approach.

This work can be extended in many ways. First, we would like to use various types of local and global descriptors and add shape features as well, so to select the best features for the task and test the performance of the three integration schemes when the number of cues grows. This, combined with a thorough theoretical and algorithmic analysis of the three methods, should make it possible to understand better their strengths and weaknesses, and to evaluate which of them should be preferred for this applications. Finally, our algorithm does not exploit at the moment the natural hierarchical structure of the data, but we believe that this information is crucial for achieving significant improvements in performance. Future work will explore these directions.

Acknowledgments

This work was supported by the ToMed.IM2 project (B. C. and F. O.), under the umbrella of the Swiss National Center of Competence in Research (NCCR) on Interactive Multimodal Information Management (IM2, www.im2.ch), and by the Blanceflor Boncompagni Ludovisi foundation (T. T., www.blanceflor.se).

References

Boser, B. E., Guyon, I. M., Vapnik, V. N., 1992. A training algorithm for optimal margin classifiers. In: Haussler, D. (Ed.), Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory. ACM press, pp. 144–152.

Cristianini, N., Shawe-Taylor, J., 2000. An Introduction to Support Vector Machines (and Other Kernel-Based Learning Methods). CUP.

Deselaers, T., Hegerath, A., Keysers, D., Ney, H., Sep. 2006. Sparse patch-histograms for object classification in cluttered images. In: DAGM 2006, Pattern Recognition, 27th DAGM Symposium. Vol. 4174 of Lecture Notes in Computer Science. Berlin, Germany, pp. 202–211.

Deselaers, T., Mueller, H., Deserno, T., 2008. Automatic medical image annotation in imageclef 2007: Overview, results, and discussion. Pattern Recognition Letters, Special Issue on Medical Image Annotation in ImageCLEF 2007.

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

Fowlkes, C., Belongie, S., Chung, F., Malik, J., 2004. Spectral grouping using the nyström method. IEEE Transactions on Pattern Analysis and Machine Intelligence 26 (2), 214–225.

Gueld, M., Kohnen, M., Keysers, D., Schubert, H., Wein, B., Bredno, J., Lehmann, T., 2002. Quality of dicom header information for image categorization. In: Proceedings of SPIE Medical Imaging. Vol. 4685. pp. 280–287.

Guld, M., Thies, C., Fischer, B., Lehmann, T., 2006. Baseline results for the imageclef 2006 medical automatic annotation task. In: Working Notes of the 2006 CLEF Workshop.

Liu, J., Hu, Y., Li, M., Ma, W.-Y., 2006. Medical image annotation and retrieval using visual features. In: Working Notes of the 2006 CLEF Workshop.

Lowe, D. G., 1999. Object recognition from local scale-invariant features. In: Proceedings of the International Conference on Computer Vision (ICCV). Vol. 2. pp. 1150–1157.

Mueller, H., Deselaers, T., Kim, E., Kalpathy-Cramer, J., Deserno, T. M., Clough, P., Hersh, W., 2007. Overview of the ImageCLEFmed 2007 medical retrieval and annotation tasks. In: Working Notes of the 2007 CLEF Workshop.

Mueller, H., Gass, T., 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., Caputo, B., 2004. Cue integration through discriminative accumulation. In: Proceedings of the International conference on Computer Vision and Pattern Recognition. Vol. 2. pp. 578–585.

Nister, D., Stewenius, H., 2006. Scalable recognition with a vocabulary tree. In: CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, Washington, DC, USA, pp. 2161–2168.

Nowak, E., Jurie, F., Triggs, B., 2006. Sampling strategies for bag-of-features image classification. In: Proceedings of the European Conference on Computer Vision. Vol. 4. pp. 490–503.

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

Tanaka, K., Saito, H., Fukada, Y., Moriya, M., 1991. Coding visual images of objects in the inferotemporal cortex of the macaque monkey. Journals of Neurophysiology 66 (1), 170–189.

Vapnik, V. N., 1998. Statistical Learning Theory. John Wiley and Sons, New York.

W. Hersh, J. Kalpathy-Cramer, J. J., 2006. Medical image retrieval and automated annotation: OHSU at ImageCLEF 2006. In: Working Notes of the 2006 CLEF Workshop.