Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Discriminative cue integration for medical image annotation ☆

Tatiana Tommasi, Francesco Orabona, Barbara Caputo *

IDIAP Research Institute, Centre Du Parc, Avenue des Pres-Beudin 20, P.O. Box 592, CH-1920 Martigny, Switzerland

ARTICLE INFO

Article history:
Available online 29 March 2008

Keywords:
Medical image annotation
Cue integration
Support vector machine

ABSTRACT

Automatic annotation of medical images is an increasingly important tool for physicians in their daily activity. Hospitals nowadays produce 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 using three alternative approaches, all based on the support vector machine algorithm. We tested our methods on the IRMA database, and with two of the three approaches proposed here we participated in the 2007 ImageCLEFmed benchmark evaluation, in the medical image annotation track. These algorithms ranked first and fifth, respectively among all submission. Experiments using the third approach 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-bytes 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 provided participants with 11,000 training and development images, spread across 116 classes. The task consisted in assigning the correct label to 1000 test images. For further information on the annotation task of ImageCLEF 2007 we refer the reader to Müller et al. (2007).

An open challenge for automatic annotation of medical images is that images that belong to the *same* visual class might look very *different*, while images that belong to *different* visual classes might look very *similar*. An example of this phenomenon is shown in Figs. 1 and 2. In particular, Fig. 1 shows some examples of visual variability within the class ‘foot, AP unspecified’: images of the same body region, with the same orientation, taken from different persons show high variability, because of differences in age or individual body structures. This problem can be solved with classification algorithms able to generalize well without compromising robustness. Fig. 2 shows exemplar images of different classes which share some visual characteristics: ‘chest, PA unspecified’, ‘chest, PA expiration’, ‘chest, AP inspiration’ and ‘chest, AP supine’. They must be classified differently because of clinical needs, but they present a strong visual similarity because they all contain the body part ‘chest’. This problem calls for classification algorithms able to use the most discriminative information from the available data. The challenge described above is well known in the medical image annotation literature, where it is usually referred to as the inter-class vs intra-class variability problem (Setia et al., 2006; Lehmann et al., 2004; Florea et al., 2006a). It appears also in other visual classification problems, such as face recognition and robotics (Sim and Zhang, 2004; Kim et al., 2007).

Several authors tried to address this problem using local and global features, and more generally different types of descriptors, separately or combined together in a multi-cue approach. Regarding the medical image annotation field, in 2006 three groups proposed cue integration methods for the ImageCLEFmed annotation task. Müller et al. (2006) combined different global and local features together. The annotation strategy was based on the GNU Image Finding Tool image retrieval engine. A similarity score is calculated based on the training data. In particular features that appear in images of the same class are weighted more than features appearing across many classes. The run was not submitted to the challenge, but given its results it would have ranked 26th among 28 real submissions. Güld et al. (2006) used the texture features proposed in Tamura et al. (1978) and Castelli et al. (1998), evaluating, respectively, the Jensen–Shannon divergence and the Mahalanobis distance to compare the query and the reference

☆ 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).

* Corresponding author. Fax: +41 277 217 712.
E-mail addresses: ttommasi@idiap.ch (T. Tommasi), forabona@idiap.ch (F. Orabona), bcaputo@idiap.ch (B. Caputo).

0167-8655/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2008.03.009

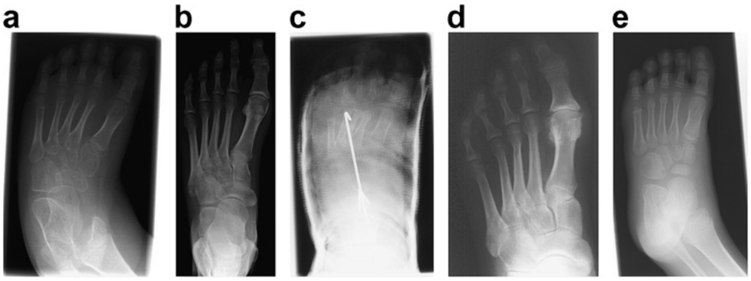

Fig. 1. Images from the IRMA database. Note the high visual variability within the images. They all belong to the same class annotated as: acquisition modality 'overview image'; body orientation 'AP unspecified'; body part 'foot'; biological system 'muscolosceletal'.

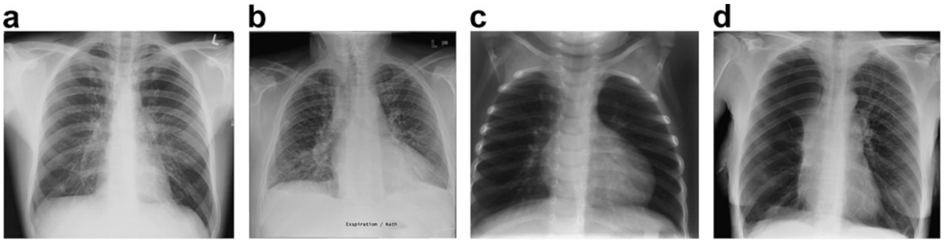

Fig. 2. Images from the IRMA database. Note the high visual similarity between the images. Each of them belongs to a different class. They all have as acquisition modality 'high beam energy', as body region 'chest unspecified', as biological system 'unspecified', but they differ for the body orientation: (a) 'PA unspecified', (b) 'PA expiration' (c) 'AP inspiration', and (d) 'AP supine'.

image. They also used a down-scaled representation of the original images evaluating the cross correlation function and the image distortion model value as distance measure. All these normalized distances are weighted and summed producing a single total distance which is considered as the basis of a nearest-neighbor decision function. The authors' run ranked 12th. Florea et al. (2006) used texture based features combined through concatenation with statistical gray level measures. Principal component analysis is then applied to reduce the number of feature elements. The obtained results corresponded to the 4th rank position. For some of these examples the performance was not very high. Still years of research on visual recognition in other domains have shown clearly that multiple cue methods outperform single feature approaches (Matas et al., 1995; Mel, 1997; Sun, 2003). Heterogeneous and complementary visual cues, bringing different information content, were successfully used in Nilsback and Caputo (2004), Mel (1997), Slater and Healey (1995). The usefulness of feature combination is even more evident when cues are extracted from different modalities, like vision and sound (Jie et al., 2008) or vision and lasers (Tapus and Siegwart, 2005).

In this paper we follow this route, and we propose to tackle the inter-class versus intra-class variability problem using a discriminative cue integration approach, based on support vector machines (SVM) (Cristianini and Shawe-Taylor, 2000). We extract local and global descriptors and we combine the two features using three integration schemes. The first is the discriminative accumulation scheme (DAS), 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 consists of designing a new Mercer kernel,[1] 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 (Vapnik, 1998). The third integration scheme 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 29.90 and 26.85, ranking fifth and first among all submissions. The third approach, developed after the submission deadline, achieved a score of 26.96, which would have corresponded to ranking second among all submissions. We also tested the approaches on a different task, namely object categorization on a subset of the Caltech database (Fergus et al., 2003), using various feature types. Results in both domains clearly prove the power of using multiple cues.

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 of the performance of each algorithm. Conclusions are drawn and potential avenues for future work are given in the last section.

2. Single-cue image annotation

The strategy we propose is to extract a set of features from each image and to use SVM for the classification task. We used a local approach, SIFT-based descriptors, and a global approach, raw pixels.

2.1. Local features

We adopted the framework of "bag of words" commonly used in many state of the art approaches in images classification (Nowak et al., 2006; Nister and Stewenius, 2006) and medical image annotation (Deselaers et al., 2006). In analogy to text classification, the

[1] A very similar approach was proposed simultaneously and independently in Bosch et al. (2007).

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" (BOVW, Bag Of Visual Words). Note that the ordering of the visual words is not important and only the frequency of appearance of each word is used to form the feature vectors. The main implementation choices are thus: (1) how to sample patches, (2) what visual patch descriptor to use, and (3) how to build the vocabulary.

Regarding point (1), we used random sampling. 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 Nowak et al. (2006) that a dense random sampling is always superior to any strategy based on interest points detectors.

Regarding point (2), we decided to use a modified version of the SIFT descriptor (Lowe, 1999). SIFTs are designed to describe an area of an image so to be robust to noise, illumination, scale, translation and rotation changes. Given the specific constraints of our classification task, we slightly modified the classical version of this descriptor. The SIFT rotation-invariance is not relevant for the ImageCLEFmed classification task, as the various structures 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 performance on a validation set, and we removed the rotation-invariance. We call the modified SIFT descriptor modSIFT.

Regarding point (3), we built the vocabulary randomly sampling 30 points from each input image and extracting modSIFT in each point. The visual words are created using an unsupervised K-means clustering algorithm. Note that, in this phase all the 12,000 images could be used, because the process does not need the labels. We chose K template modSIFTs with K equal to 500, so we defined a vocabulary with 500 words. Various sizes of vocabulary were tested ($K = 500$, 1000, and 2000). Preliminary results on a validation set showed no significant differences in performance between these three vocabulary sizes. We chose therefore $K = 500$, the smallest, for computational reasons.

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

2.2. Global features

As global descriptor we used the simplest possible one: the raw pixels. Preliminary results on a validation set showed that downscaling images to 32×32 pixels did not produce any significant difference than downscaling to 48×48, but the classification performance was better than that obtained on 16×16 images. So the images were resized to 32×32, regardless of the original dimension, and normalized to have sum equal to 1 to use the χ^2 kernel. The obtained 1024 values were then used as input features. This approach is at the same time a baseline for the classification system and a useful "companion" method to boost the performance of the modSIFT-based classifier (see Section 3). Fig. 4 shows how we built the raw pixel representation for each image.

2.3. Support vector machines

SVM are a class of learning algorithms based on statistical learning theory (Vapnik, 1998). Born as a linear classifier, SVM can be easily extended to non-linear domains through the use of *kernel* functions. The kernels implicitly map the input space to 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 heavily affects the performance of the SVM. We used an exponential χ^2 kernel for both feature types (Fowlkes et al., 2004):

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

We used this kernel because: (1) it has been demonstrated to be positive definite by Fowlkes et al. (2004), thus it is a valid kernel (Cristianini and Shawe-Taylor, 2000); (2) in our experiments we tested also linear kernel and the RBF kernel, but all of them gave worse results than the χ^2. The parameter γ was tuned through cross validation together with the SVM-cost parameter C (see Section 4).

We used both one-vs-one (_oo) and one-vs-all (_oa) multi-class extensions for SVM (Cristianini and Shawe-Taylor, 2000). Even, if the labels are hierarchical, we used these standard multi-class

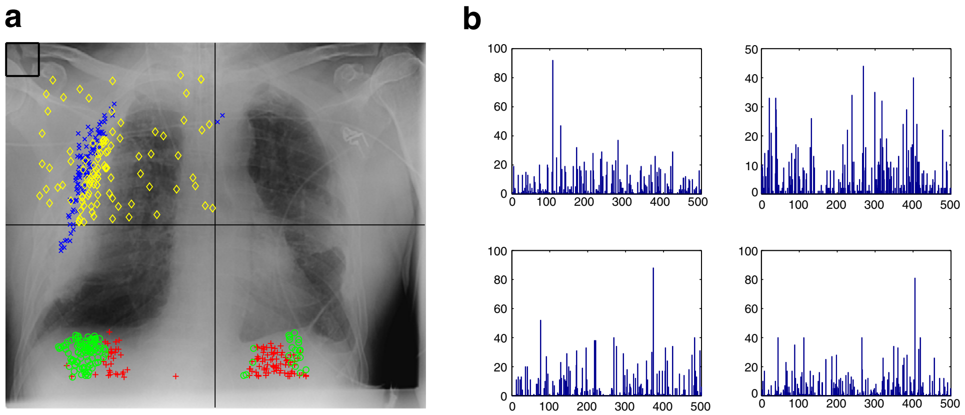

Fig. 3. (a) The four most present visual words in the image are drawn, each with a different color (better viewed in color). The square in the upper left corner represents the size of the patch used for computing the modSIFT descriptor. (b) Total counts of the visual words in the four subimages.

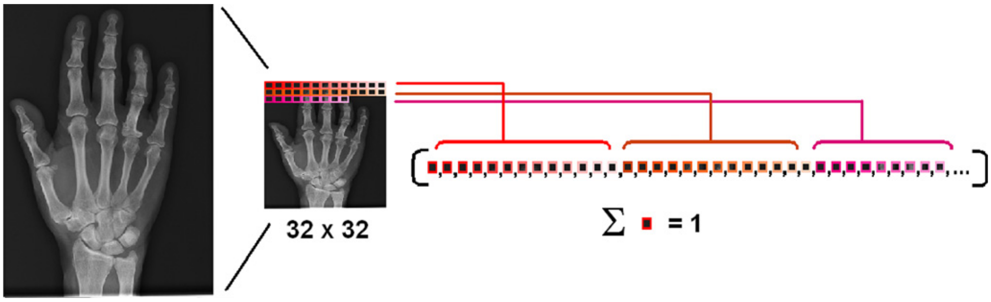

Fig. 4. An example showing the raw pixel representation.

approaches because, 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.

3. Multi-cue image annotation

Psychophysical evidence suggests 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 (Polikar, 2006; Sanderson and Paliwal, 2004). Fig. 5 illustrates schematically the basic ideas behind these methods. In this paper, we tried three different SVM-based integration schemes, one for each of these different methods. They are described in details in Sections 3.1 and 3.3. Experiments showing the effectiveness of these techniques 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, to achieve a consensus decision. In this paper, we applied this integration strategy using DAS. It is based on a weak coupling method called accumulation, which does not neglect any cue contribution. Its main idea is that information from different cues can be summed together.

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 $T_p(I_i^j), p = 1\ldots P$, so that for an object j we have P new training sets. 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 $M \geqslant 2$, for each single-cue SVM we compute the distance from the separating hyperplane $D_j(p),\ p = 1,\ldots P$. After collecting all the distances $\{D_j(p)\}_{p=1}^{P}$ for all the M objects and the P cues, we classify the image $\hat{I}$ using the linear combination:

$$j^* = \operatorname*{argmax}_{j=1}^{M} \left\{ \sum_{p=1}^{P} a_p D_{j(p)} \right\}, \qquad \sum_{p=1}^{P} a_p = 1. \tag{2}$$

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

3.2. Mid-level cue integration

Combining cues at the mid-level means that the different feature descriptors are kept separated, but they are integrated in a

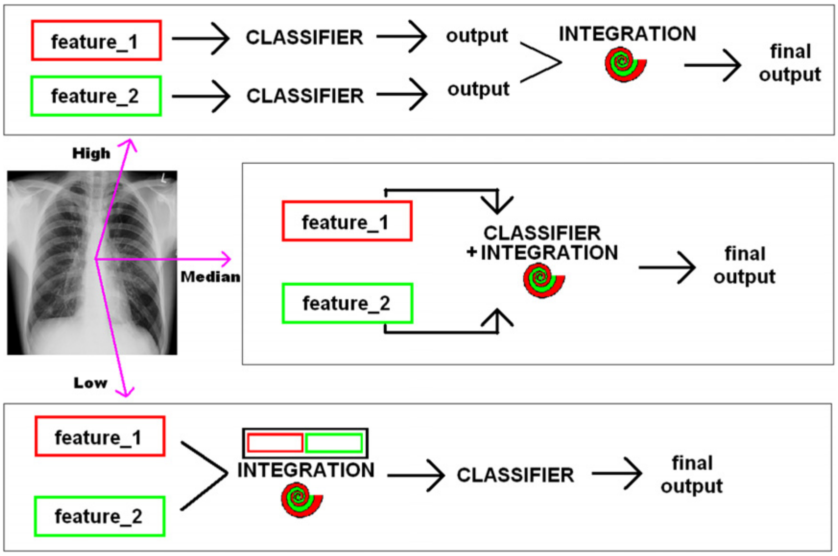

Fig. 5. A schematic illustration of the high-level, mid-level and low-level cue integration approaches.

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):

$$K_{MC}(\{T_p(I_i)\}_p, \{T_p(I)\}_p) = \sum_{p=1}^{P} a_p K_p(T_p(I_i), T_p(I)), \quad \sum_{p=1}^{P} a_p = 1. \qquad (3)$$

In this way, it is possible to perform only one classification step, identifying the best weighting factors $a_p \in \Re^+$ through cross validation while determining the optimal separating hyperplane. This means that the coefficients a_p are guaranteed to be optimal. An advantage of this approach is that it makes, it possible to work both with one-vs-all and one-vs-one SVM extensions to the multiclass problem.

3.3. *Low-level cue integration*

To combine cues it is also possible to use a low-level (LL) fusion strategy, starting from the descriptors and combining them in a new representation. In this way, the cue integration does not directly involve the classification step. Here, we used feature concatenation: two feature vectors f_i and c_i are combined into a single feature vector $v_i = (f_i, c_i)$ that is normalized to one and is then used for classification. In this way the information related to each cue is mixed without a weighting factor that allows to control the influence of each information channel on the final recognition result. A general drawback of this method is that the dimension of the feature vector increases as the number of cues grows, implying longer learning and recognition times, greater memory requirements and possibly curse of dimensionality effects. Moreover, it is not always possible to use the LL 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 with vectors of the second kind. We refer the reader to Section 4.2 for an example.

4. Experiments

4.1. *Experiments on medical image annotation*

4.1.1. *The IRMA database:*

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 11,000 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 classification performance is judged according to an error count that takes into account the IRMA code (Deselaers et al., 2008). This 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).

4.1.2. *Experimental setup*

The original dataset was divided into two parts: training and validation. In order to obtain reliable results, we merged them and extracted five random and disjoint train/test splits of 10,000/1000 images. As a preliminary step we run experiments to find the best kernel parameter γ, the best SVM C parameter and the best weighting factors a_p for both MCK and DAS through cross validation.[2] The obtained results were then used to run our submission experiments on the 1000 unlabeled images of the challenge test set using all the 11,000 images of the original dataset as training. We considered as the best parameters the one giving the lower average score on the five splits. Note that, due to the score evaluation method, the best score does not correspond necessarily to the best recognition rate.

Table 1
Ranking of our runs, name, best feature's weights, percentage of support vectors respect to the total number of training vectors, score, gain with respect to the best run of other participants and recognition rate

Rank	Name	a_{sift}	a_{pixel}	#SV (%)	Score	Gain	Recognition rate (%)
1	MCK_oa	0.80	0.20	72.0	26.85	4.08	89.7
	LL_oa			73.6	26.96	3.96	89.1
	LL_oo			63.3	26.99	3.93	89.3
2	MCK_oo	0.90	0.10	64.0	27.54	3.38	89.0
3	modSIFT_oo			65.2	28.73	2.20	88.4
4	modSIFT_oa			70.0	29.46	1.47	88.5
5	DAS	0.76	0.24	82.6	29.90	1.03	88.9
6	RWTHi6-4RUN-MV3				30.93	0	
28	PIXEL_oa			75.7	68.21	−37.28	79.9
29	PIXEL_oo			67.1	72.41	−41.48	79.2

We first evaluated the performance of local and global features separately through single-cue annotation experiments. Then we adopted the same described experimental setup for DAS, MCK and the LL cue integration method. 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. Moreover, we counted the support vectors summing the ones from local and global features obtained with SVM one-vs-all multiclass extension, but considering only once the support vectors associated with the training images that resulted in common between the single-cues. 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 LL 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.

4.1.3. *Results and discussion*

Table 1 summarizes all the relevant information about our experiments: the challenge ranking, name, best feature's weights, number of support vectors as percentage of the total number of training vectors, score, gain with 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.

Besides obtaining the optimal kernel parameters, the single-cue annotation experiments showed that the BOVW with modSIFT features outperform the raw pixel ones. This is not unexpected, as 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 whole image as a source of context, while the local features capture the details, and thus they

[2] We varied the kernel parameter γ_{sift} in [0.02; 0.05; 0.1; 0.25; 0.5; 1], γ_{pixel} in [0.7; 1.5; 3; 5; 10], the SVM C parameter in [1; 5; 10; 20; 40; 80], the MCK a_p weighting factor between 0 an 1 with step 0.1 and the DAS a_p weighting factor between 0 and 1 with step 0.01.

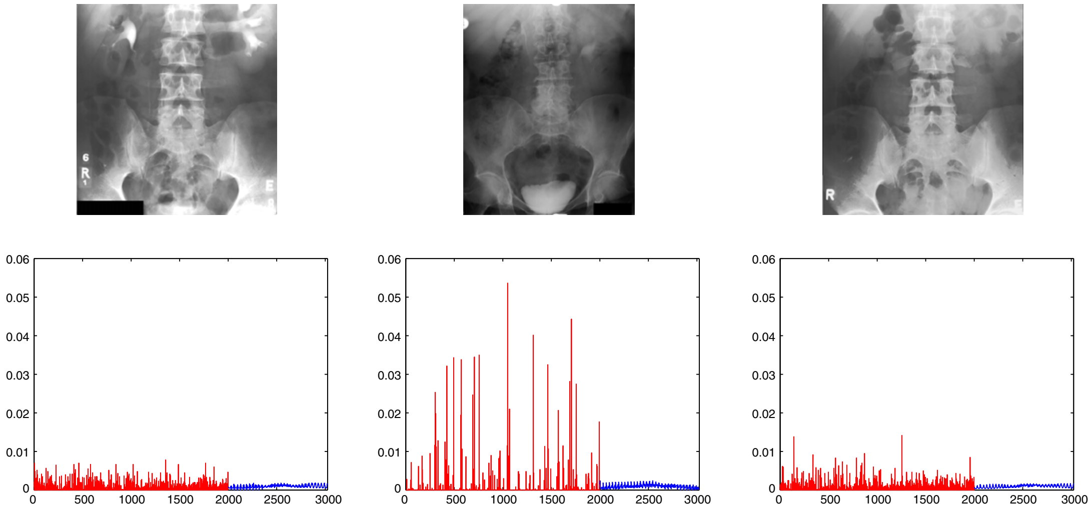

Fig. 6. Local (red) and global features (blue) corresponding to the images above (better viewed in color). The first and the last image belong to the same class, while the second comes from a different class. Local features are similar for images from the same class and different for images from different classes. On the other hand, global features appear quite alike. Remember that both features are normalized to have sum equal to 1. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

manage better the inter and intra-class variability. We can see an example of this in Fig. 6. The first and the third image belong to the same class, while the second one comes from a different class. The raw pixel features for the three images appear quite alike, while the BOVW with modSIFT features seem able to better capture the visual dissimilarities. The reader might wonder if the absolute number of features has an impact on these results, as it is 2000 for BOVW and 1024 for PIXELS. This is not the case, because we use the same kernel for both of them, with a range between 0 and 1. Note that even in case of kernels with different ranges, the a_p coefficients in DAS and MCK would rescale the features as needed.

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 modSIFT alone. It looks like DAS is not able to capture the hierarchical structure of the data, and this affects its score. Regarding the SVM multiclass extension, the one-vs-all overcome the one-vs-one.

For both the one-vs-one and the one-vs-all SVM multiclass extension, BOVW with modSIFT features need a lower number of support vectors (SV) than PIXEL. For the MCK run using one-vs-one multiclass SVM extension (MCK_oo) the number of SV is lower than that of both the single-cues modSIFT_oo and PIXEL_oo. This shows 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. The number of SV for DAS exceeds that obtained for both MCK_oa and MCK_oo showing a higher complexity of the classification problem.

In general we must notice that the number of support vectors is over 50% of the original training set. This arises from the structure of the database: on 116 classes, 50 have less than 30 images. Therefore, it is reasonable that the classifier retains almost all the training data, to cope with the unbalance data problem. On top of this, one should remember that the problem presents a high inter-class vs intra-class variability, which also pushes for storing a large percentage of training data as support vectors. For a more detailed discussion of the results, we refer the reader to Tommasi et al. (2007).

4.2. Experiments on object categorization

4.2.1. Experimental setup

We ran a second set of experiments on a subset of the Caltech database (Fergus et al., 2003), containing images of airplanes, cars, faces, motorbikes, and leaves (see Fig. 7). Training and testing sets consisted, respectively, of 15 and 60 images for each class. Our aim was to analyze how the cue integration approaches used on medical images behave when used for a different task, and with different features. We tested both the combination of raw pixels and SIFT features, and the combination of Composed Receptive Field Histograms (CRFH) (Linde and Lindeberg, 2004), and SIFT. Here, the raw pixel features are obtained resizing the original images to 16 × 16 pixels. The local features are the original SIFTs without any modification and not using the BOVW strategy. Note that in this case, the LL integration approach is not usable. For SIFT we get a different number of interest points per image, the feature elements are not sortable, so we need a specific kernel and we cannot mix this kind of feature with others (Wallraven et al., 2003). For the raw pixel and the CRFH we used the χ^2 kernel, while for the local features we used the matching kernel (Wallraven et al., 2003). We determined the best kernel parameters, the SVM C parameter and the weighting factors through five fold cross validation. Here, we considered as best the parameters which produced the highest recognition rate.

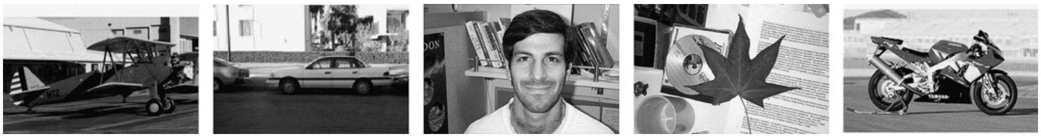

Fig. 7. Sample images from five classes of Caltech-101 database. They are taken, respectively from the class *Airplanes*, *Car side*, *Faces*, *Leaves* and *Motorbikes*.

Table 2
Recognition rate results on the five classes of the Caltech database

Name	Features	Recognition rate (%)
SVM_oa	SIFT	89.20 ± 1.85
	PIXEL	78.32 ± 3.99
	CRFH	84.24 ± 1.51
DAS	SIFT + PIXEL	92.64 ± 1.37
	SIFT + CRFH	93.60 ± 1.79
MCK_oa	SIFT + PIXEL	92.96 ± 1.85
	SIFT + CRFH	92.64 ± 0.67

4.2.2. Results and discussion

Results are reported in Table 2. We see that for the (SIFT, PIXEL) combination, MCK performs better than DAS on average, but the two algorithms achieve very close performance, therefore their results can be considered equivalent. For the (SIFT, CRFH) combination instead DAS performs better than MCK on average, but here also their results can be considered equivalent. On the other hand, the gain in performance is very high for both the cue integration methods with respect to the single-cue classifiers.

This result is in agreement with our findings on the IRMA database, and confirms that integrating multiple cues pays off as opposed to use one single feature type. Still it is not clear from our experiments, which cue integration strategy should be preferred. Indeed, our results seem to indicate that the task and the features used have a strong impact on which approach performs best.

5. Conclusions

In this paper, we presented a discriminative multi-cue approach to medical image annotation. We combined global and local information using three alternative fusion strategies: we used the discriminative accumulation scheme, the multi-cue kernel able to take as input different cues while keeping them separated during the optimization process, and we merged features together in a unique descriptor. The second method gave the best performance in the ImageCLEF 2007 benchmark evaluation, obtaining a score of 26.85, which ranked first among all submissions. Additional experiments in the domain of object categorization confirm the power of cue integration for visual classification, and emphasize the dependency from the task and the feature type in the selection of the optimal fusion strategy.

This work can be extended in many ways. We would like to use various types of local and global descriptors and add shape features as well, so to 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. Regarding the medical image annotation task, 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.

References

Bosch, A., Zisserman, A., Munoz, X., 2007. Representing shape with a spatial pyramid kernel. In: Proc. 6th ACM Internat. Conf. on Image and Video Retrieveal. ACM, New York, NY, USA, pp. 401–408.

Castelli, V., Bergman, L.D., Kontoyiannis, I., Li, C.-S., Robinson, J.T., Turek, J.J., 1998. Progressive search and retrieval in large image archives. IBM J. Res. Dev. 42 (2), 253–268.

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

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

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

Fergus, R., Perona, P., Zisserman, A., 2003. Object class recognition by unsupervised scale-invariant learning. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition.

Florea, F., Barbu, E., Rogozan, A., Bensrhair, A., Buzuloiu, V., 2006a. Medical image categorization using a texture based symbolic description. In: IEEE Internat. Conf. on Image Process., vol. 1, pp. 1489–1492.

Florea, F., Rogozan, A., Cornea, V., Bensrhair, A., Darmoni, S., 2006b. 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 Trans. Pattern Anal. Machine Intell. 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: Proc. SPIE Med. Imaging, vol. 4685, pp. 280–287.

Güld, M., Thies, C., Fischer, B., Lehmann, T., 2006. Baseline results for the ImageCLEF 2006 medical automatic annotation task. In: CLEF 2006. Lecture Notes in Computer Science, vol. 4730. Springer, pp. 686–689.

Jie, L., Caputo, B., Zweig, A., Bach, J.-H., Anemuller, J., 2008. Object category detection using audio-visual cues. In: Proc. Internat. Conf. on Computer Vision System, Santorini, Greece.

Kim, S., Kweon, I.S., Lee, C.-W., 2007. Visual categorization robust to large intra-class variations using entropy-guided codebook. In: IEEE Internat. Conf. on Robotics and Automation, pp. 3793–3798.

Lehmann, T., Fischer, B., Güld, M., Thies, C., Keysers, D., Deselaers, T., Schubert, H., Wein, B., Spitzer, K., 2004. The IRMA reference database and its use for content-based image retrieval in medical applications. In: Proc. Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie.

Linde, O., Lindeberg, T., 2004. Object recognition using composed receptive field histograms of higher dimensionality. In: Proc. 17th Internat. Conf. on Pattern Recognition, vol. 2, pp. 1–6.

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: Proc. Internat. Conf. on Computer Vision, vol. 2, pp. 1150–1157.

Matas, J., Marik, R., Kittler, J., 1995. On representation and matching of multi-coloured objects. In: Proc. 5th Internat. Conf. on Computer Vision. IEEE Computer Society, Washington, DC, USA.

Mel, B.W., 1997. SEEMORE: Combining color, shape and texture histogramming in a neurally-inspired approach to visual object recognition. Neural Comput. 9 (4), 777–804.

Müller, 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.

Müller, 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.

Nilsback, M., Caputo, B., 2004. Cue integration through discriminative accumulation. In: Proc. Internat. Conf. on Computer Vision and Pattern Recognition, vol. 2, pp. 578–585.

Nister, D., Stewenius, H., 2006. Scalable recognition with a vocabulary tree. In: Proc. 2006 IEEE Comput. Soc. Conf. 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: Proc. Eur. Conf. on Computer Vision, vol. 4, pp. 490–503.

Polikar, R., 2006. Ensemble based systems in decision making. IEEE Circ. System Mag. 6 (3), 21–45.

Sanderson, C., Paliwal, K.K., 2004. Identity verification using speech and face information. Digital Signal Process. 14 (5), 449–480.

Setia, L., Teynor, A., Halawani, A., Burkhardt, H., 2006. Grayscale radiograph annotation using local relational features. In: Evaluation Multilingual and Multi-modal Information Retrieval, Lecture Notes in Computer Science, pp. 644–651.

Sim, T., Zhang, S., 2004. Exploring face space. In: Conf. on Computer Vision and Pattern Recognition Workshop, vol. 5.

Slater, D., Healey, G., 1995. Combining color and geometric information for the illumination invariant recognition of 3D objects. In: Proc. 5th Internat. Conf. on Computer Vision. IEEE Computer Society, Washington, DC, USA.

Sun, Z., 2003. Adaptation for multiple cue integration. Proc. Conf. on Computer Vision and Pattern Recognition, vol. 01. IEEE Computer Society, Los Alamitos, CA, USA, pp. 440–445.

Tamura, H., Mori, S., Yamawaki, T., 1978. Textual features corresponding to visual perception. In: IEEE Trans. On Systems, Man, and Cybernet., vol. 8 (6), pp. 460–472.

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

Tapus, A., Siegwart, R., 2005. Incremental robot mapping with fingerprints of places. In: IEEE/RSJ Internat. Conf. on Intell. Robots and Systems. IEEE Computer Society, pp. 2429–2434.

Tommasi, T., Orabona, F., Caputo, B., 2007. Cue integration for medical image annotation. In: CLEF 2007 Proc., Lecture Notes in Computer Science.

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

Wallraven, C., Caputo, B., Graf, A., 2003. Recognition with local features: the kernel recipe. Proc. Internat. Conf. on Computer Vision, vol. 2. IEEE Press, pp. 257–264.