Cue Integration for Medical Image Annotation

Tatiana Tommasi, Francesco Orabona, and Barbara Caputo

IDIAP Research Institute,
Centre Du Parc, Av. des Pres-Beudin 20,
P. O. Box 592, CH-1920 Martigny, Switzerland
{ttommasi,forabona,bcaputo}@idiap.ch

Abstract. This paper presents the algorithms and results of our participation to the image annotation task of ImageCLEFmed 2007. We proposed a multi-cue approach where images are represented both by global and local descriptors. These cues are combined following two SVM-based strategies. The first algorithm, called Discriminative Accumulation Scheme (DAS), trains an SVM for each feature, and considers as output of each classifier the distance from the separating hyperplane. The final decision is taken on a linear combination of these distances. The second algorithm, that we call Multi Cue Kernel (MCK), uses a new Mercer kernel which can accept as input different features while keeping them separated. The DAS algorithm obtained a score of 29.9, which ranked fifth among all submissions. The MCK algorithm with the one-vs-all and with the one-vs-one multiclass extensions of SVM scored respectively 26.85 and 27.54. These runs ranked first and second among all submissions.

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 and, when done manually, prone to errors [1]. This calls for automatic annotation algorithms able to perform the task reliably. The ImageCLEFmed annotation task in 2007 has provided participants with 11000 training and development data, spread across 116 classes [2]. State of the art approaches used texture-based descriptors as features and discriminative algorithms, mainly SVMs, for the classification step [3,4]. Local and global features, have been used separately or combined together in multi-cue approaches with disappointing results [3,5]. Still, years of research on visual recognition showed clearly that multiple-cue methods outperform single-feature approaches, provided that the features are complementary.

This paper describes a multi-cue strategy for biomedical image classification. We used raw pixels as global descriptors and SIFT features as local descriptors. The two feature types were combined together using two different SVM-based integration schemes. The first is the Discriminative Accumulation Scheme (DAS), proposed first in [6]. For each feature type, an SVM is trained and its output

C. Peters et al. (Eds.): CLEF 2007, LNCS 5152, pp. 577–584, 2008.

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. We submitted a run using this method that ranked fifth among all submissions. The second integration scheme consists in 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. We submitted two runs using this algorithm, the first using the one-vs-all multiclass extension of SVM; the second using instead the one-vs-one extension. These two runs ranked first and second among all submissions. These results overall confirm the effectiveness of using multiple cues for automatic image annotation.

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. Section 3 gives details on the two 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 aim of the automatic image annotation task is to classify images into a set of classes, according to the IRMA code [7]. The labels are hierarchical therefore, errors in the annotation are counted depending on the level at which the error is done and on the number of possible choices. For each image the error ranges from 0 to 1, respectively if the image is correctly classified or if the predicted label is completely wrong. The strategy we propose is to extract a set of features from each image (section 2.1) and to use then a Support Vector Machine (SVM) to classify the images (section 2.2).

2.1 Feature Extraction

We chose two types of features, local and global, with the aim to extract different informations.

Local Features. We explored the idea of “bag of words”, a common concept in many state of the art approaches to visual recognition. The basic idea is that it is possible to transform the images into a set of prespecified visual words, and to classify the images using the statistics of appearance of each word as feature vectors. To build the visual vocabulary, we used SIFT features [8], computed around interest points detected via random sampling [9]. With respect to the classic SIFT implementation, we removed the rotational invariance and the scale invariance by extracting the SIFT at only one orientation and at one octave, the one that obtained the best classification performance. To keep the complexity of the description of each image low and at the same time retain as much information as possible, we matched each extracted SIFT with a number of template

SIFTs. These template SIFTs form our vocabulary of visual words. It is built using a standard K-means algorithm, with K equal to 500, on a random collection of SIFTs extracted from the training images. Various sizes of vocabulary were tested with no significant differences, so we have chosen the smaller one with good recognition performances. Note that in this phase also testing images can be used, because the process is not using the labels and it is unsupervised. At this point each image could be described with the raw counts of each visual word. To add some kind of spatial information to our features we divided the images in four subimages, collecting the histograms separately for each subimage. In this way the dimension of the input space is multiplied by four, but in our tests we gained about 3% of classification performances. We have extracted 1500 SIFT in each subimage: such dense sampling adds robustness to the histograms. See Figures 1 for an example.

Global Features. We chose the simplest possible global description method: the raw pixels. The images were resized to 32x32 pixels, regardless of the original dimension, and normalized to have sum equal to one, then the 1024 raw pixels values were 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 SIFT based classifier (see section 2.2).

2.2 Classification

For the classification step we used an SVM with an exponential χ^2 as kernel, for both the local and global approaches:

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

The parameter γ was tuned through cross-validation (see section 4). This kernel has been successfully applied for histogram comparison and it has been demonstrated to be positive definite [10], thus it is a valid kernel.

3 Multi Cue 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. Thus, we expect that by combining them through an integration scheme, we should achieve a higher classification performance and a higher robustness. In the rest of the section we describe the two alternative integration schemes we used. The first, the Discriminative Accumulation Scheme (DAS, [6]), is a high-level integration scheme, meaning that each single cue first generate 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. This method is described in section 3.1. The second, the Multi Cue Kernel (MCK), is a mid-level integration scheme, meaning that the different features descriptors are kept separated

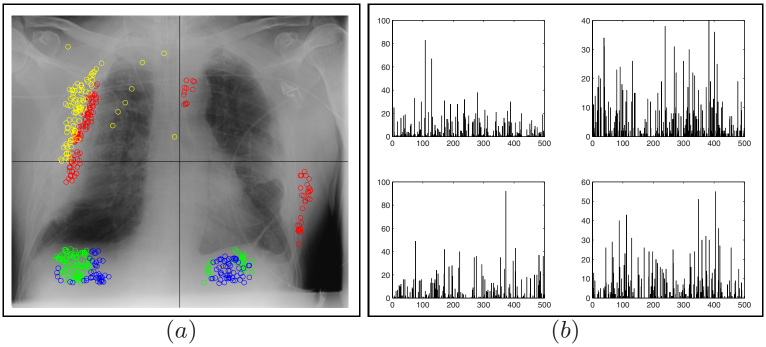


Fig. 1. (a) The four most present visual words in the image are drawn, each with a different color. and (b) total counts of the visual words in the 4 subimages.

but they are combined in a single classifier generating the final hypothesis. This algorithm is described in section 3.2.

3.1 Discriminative Accumulation Scheme

The Discriminative Accumulation Scheme is an integration scheme for multiple cues that 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 $\{I_i^j\}_{i=1}^{N_j}$, $j = 1, \dots M$. For each image, we extract a set of P different cues 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 $M \geq 2$, for each single-cue SVM we compute the distance from the separating hyperplane $D_j(p)$, $p = 1 \dots P$: 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:

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

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

3.2 Multi Cue Kernel

DAS can be defined a high-level integration scheme, as fusion is performed as a post-processing step after the single-cue classification stage. As an alternative, we developed a mid-level integrating scheme based on multi-class SVM with a Multi Cue Kernel K_{MC}. This new kernel combines different features extracted form images; it is a Mercer kernel, as positively weighted linear combination of Mercer kernels are Mercer kernels themselves [11]:

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

In this way it is possible to perform only one classification step, identifying the best weighting factors a_p while optimizing the other kernel parameters. Another 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.

4 Experiments

Our experiments started evaluating the performance of local and global features separately. Even if the original dataset was divided in training, validation and testing sets, we decided to merge them together and extract 5 random and disjoint train/test splits of 10000/1000 images using the cross validation technique for the parameters selection. We considered as the best parameters the ones giving the best average score on the 5 splits. Note that, according to the method used for the score evaluation, the best average score is not necessary the best recognition rate. Besides obtaining the optimal parameters, these experiments showed that the SIFT features outperform the raw pixel ones, as it was predictable.

Then we adopted the same experimental setup for DAS and MCK. In particular in DAS we used the best parameters of the previous step, so we only searched the best weights for cue integration. On the other hand, for MCK we looked for the best kernel parameters and the best feature's weights at the same time. Finally we used the results of the previous phases 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.

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. Our two runs based on the MCK algorithm ranked first and second among all submissions

Table 1. Ranking of our submitted runs, name, best parameters, percentage number of SVs, score, gain respect to the best run of the other participants and recognition rate

Rank	Name	a_{sift}	a_{pixel}	#SV(%)	Score	Gain	Rec. rate
1	MCK_oa	0.80	0.20	72.0	26.85	4.08	89.7%
2	MCK_oo	0.90	0.10	64.0	27.54	3.38	89.0%
3	SIFT_oo			65.2	28.73	2.20	88.4%
4	SIFT_oa			70.0	29.46	1.47	88.5%
5	DAS	0.76	0.24	82.6	29.90	1.03	88.9%
28	PIXEL_oa			75.7	68.21	−37.28	79.9%
29	PIXEL_oo			67.1	72.42	−41.48	79.2%

stating 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.

In the same table there is also a summary of the weighting parameters for the multi-cue approaches and the number of Support Vectors (SVs) obtained showed as percentage of the total number of training vectors. As we could expect, the best feature weight (see (2) and (3)) for SIFT results higher than that for raw pixels for all the integration methods. The number of SVs is a rough indicator of the difficulty of the problem. The percentage of SVs 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 percentage number of SVs 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. In general we must notice that the percentage number of support vectors is over 50%. This suggests that the classification task is challenging, and therefore the generalization properties of the method might not be optimal. For MCK_oa, the two classification problems with the highest number of SVs are class 1121-110-213-700 (overview image, coronal posteroanterior unspecified, nose area, muscolosceletal system) vs all, and class 1121-115-710-400 (overview image, coronal posteroanterior upright, abdomen unspecified, gastrointestinal system unspecified) vs all.

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

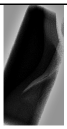
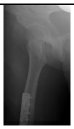
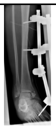
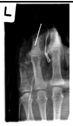
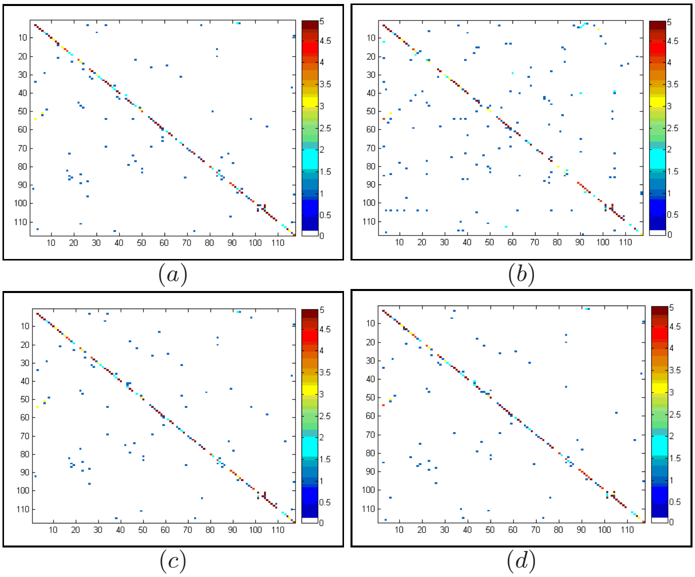
				
PIXEL_oa	11°	**1°**	12°	5°
SIFT_oa	**1°**	2°	2°	5°
DAS	**1°**	**1°**	**1°**	2°
MCK_oa	**1°**	**1°**	**1°**	**1°**

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

The confusion matrices corresponding to the single-cue, discriminative accumulation and multicue kernel approach are shown as images in Figure 2. It is



(*a*) (*b*)

(*c*) (*d*)

Fig. 2. These images represent the confusion matrices respectively for (*a*) SIFT_oa, (*b*) Pixel_oa, (*c*) DAS and (*d*) MCK_oa. To let the misclassified images stand out all the position in the matrices containing five or more images appear dark red.

clear that our methods differ principally for how the wrong images are labeled. The more the matrices present sparse values out of the diagonal and far away from it, the worse the method is. For the MCK_oa run the classes which contribute the most to the error score are 1123-127-500-000 confused with class 1123-110-500-000 (high beam energy, 127: coronal posteroanterior supine - 110: coronal posteroanterior unspecified chest unspecified) and class 1121-200-411-700 confused with class 1121-110-411-700 (overview image, 200: sagittal unspecified, upper extremity finger unspecified, muscolosceletal system). The class which obtains the higher benefit from the cue combination through MCK_oa is 1123-110-500-000, the number of correctly recognized images passes from 78 with SIFT_oa to 84 adding up the global (PIXEL_oa) information.

5 Conclusions

This paper presented a discriminative multi-cue approach to medical image annotation. We combined global and local information using two alternative fusion strategies, the Discriminative Accumulation Scheme [6] and the Multi Cue Kernel. This last method gave the best performance obtaining a score of 26.85, which ranked first among all submissions.

This work can be extended in many ways. First, we would like to use various types of local, global and shape descriptors, so to select the best features for the task. Second, our algorithm does not exploit at the moment the natural hierarchical structure of the data, but we believe that this information is crucial. 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

1. Güld, M.O., Kohnen, M., Keysers, D., Schubert, H., Wein, B.B., Bredno, J., Lehmann, T.M.: Quality of dicom header information for image categorization. In: Proc of SPIE Medical Imaging, vol. 4685, pp. 280–287 (2002)
2. Müller, H., Deselaers, T., Kim, E., Kalpathy-Cramer, J., Deserno, T.M., Clough, P., Hersh, W.: Overview of the ImageCLEFmed 2007 medical retrieval and annotation tasks. In: Working Notes of the 2007 CLEF Workshop (2007)
3. Müller, H., Gass, T., Geissbuhler, A.: Performing image classification with a frequency-based information retrieval schema for ImageCLEF 2006. In: Working Notes of the 2006 CLEF Workshop (2006)
4. Liu, J., Hu, Y., Li, M., Ma, W.Y.: Medical image annotation and retrieval using visual features. In: Working Notes of the 2006 CLEF Workshop (2006)
5. Güld, M., Thies, C., Fischer, B., Lehmann, T.: Baseline results for the imageclef 2006 medical automatic annotation task. In: Working Notes of the 2006 CLEF Workshop (2006)
6. Nilsback, M.E., Caputo, B.: Cue integration through discriminative accumulation. In: Proc of CVPR (2004)
7. Lehmann, T.M., Henning, S., Daniel, K., Michael, K., Bethold Wein, B.: The irma code for unique classification of medical images. In: Proc of SPIE Medical Imaging, vol. 5033, pp. 440–451 (2003)
8. Lowe, D.G.: Object recognition from local scale-invariant features. In: Proc of ICCV, vol. 2, pp. 1150–1157 (1999)
9. Nowak, E., Jurie, F., Triggs, B.: Sampling strategies for bag-of-features image classification. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3954. Springer, Heidelberg (2006)
10. Fowlkes, C., Belongie, S., Chung, F., Malik, J.: Spectral grouping using the nyström method. PAMI 26(2), 214–225 (2004)
11. Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines. Cambridge University Press, Cambridge (2000)