Report for Annotation task in ImageCLEFmed 2005

Bo QIU, Wei XIONG, Qi TIAN, Chang Sheng XU
Institute for Infocomm (I2R), Singapore, 119613
visqiu{wxiong, tian, xucs}@i2r.a-star.edu.sg

Abstract

In the medical image annotation task we have mainly explored ways to use different image features to achieve robust classification performance, including both global features and regional blob features. Experimental results show that using a combination of the blob region feature and three low resolution pixel maps (gray level, texture and contrast) can achieve the highest recognition accuracy. All these features are normalized and stacked to form a one-dimension feature vector as inputs of classifiers. In our experiments Supporting Vector Machines (SVM) with RBF (radial basis functions) kernels are used for the classification task, trained over a subset of 9000 given medical training images. Our proposed method has achieved a recognition rate of 89% over a subset of the training images which were not used in the SVM training. According to the evaluation result from the imageCLEF05 organizers, our method has achieved a recognition rate of about 80% over the 1000 testing images.

Categories and Subject Descriptors

H.3 [Information Storage and Retrieval]: H3.1 Content Analysis and Indexing; H3.3 Information Search and Retrieval;

General Terms

Measurement, Performance, Experimentation

Keywords: automatic medical image annotation, SVM, low resolution map, multi-class classification, unbalance, over-fitting

1. Introduction

With the fast development of modern medical devices, more and more medical images are generated, so that the demand becomes more and more urgent for automatically indexing, comparing, analyzing and annotating the huge volume of medical images. As a benchmark work, ImageCLEFmed gets more and more well-known owing to its open data platform.

Two tasks are published in ImageCLEFmed 2005: retrieval and annotation. We target at the automatic annotation task, which is the first time to be published. In this task, 9,000 radiographs are classified into 57 classes (see Figure 1), which can be taken as training datasets; and 1,000 unlabeled radiographs are taken as the test dataset. As the first stage, the annotation task is very simple: automatic labeling the 1,000 images is the whole object of the annotation system. Evaluation of the system will base on the 'error rate', which means the percent of how many images are wrongly classified.

Medical image annotation can be regarded as an interpretation to medical evidence, while in this research, evidences are images. Generally it is a doctor who uses the specialist vocabulary and natural language phrases to interpret those medical evidences, and relates them to some specific cases. For automatic machine-based reasoning based on the evidence gathered, additional interpretive semantics must be attached to the data. Some methods have been explored in special domains, like the diagnosis of breast cancer [1]. In the annotation task of ImageCLEFmed 2005, the annotation has been simplified to mark a class label for each medical image as one of 57 given classes. But in fact the images have been annotated with complete IRMA codes (both in English and in German), which are multi-axial codes for image annotation. In future the results of current annotation task will be used for further textual image retrieval tasks.

So by now, the annotation task can be taken as a multi-class classification problem, which is a great challenge for medical images with 57 classes. Compared with other classification problems, there are some particular difficulties for medical images:

- Great unbalance between 57 classes;
  See Figure 2. In 57 training sets, class 6 has more than 500 samples (images), class 12 has more than 2,500 samples, class 34 has near 1,000 samples, while all the others are much less. 20 classes occupies near 80% of the whole training sets. This unbalance causes many common classification methods unavailable.
- Visual similarities between some classes;
  See Figure 3. For people who are not experts of radiographs, it is impossible to find the differences between some classes visually.
- Variety in one class and difficulty in defining visual features.
  See Figure 4. Too many modalities vary in one class. To find a general visual feature for one class is often very difficult.

It’s more like an experimental work, where many features and methods have been tested based on simulation experiments.

In Section 2 feature extraction techniques are described; Section 3 overviews SVM; Section 4 gives the results of experiments; at last Section 5 gives the conclusion and future direction.

2. Feature sets

Feature extraction is a basic problem in image processing field. In [2] there are 56 CBIR (content-based image retrieval) systems reviewed, and a summary of low-level features are listed in 3 main categories: color, texture, and shape, plus 2 single features: layout, face detection. In [3][4] there are some similar classifications of features.

‘Color’ includes dominant colors, region histogram, color coherence vector, color moments, correlation histogram, global/ sub-image histogram, eigen image, etc.

‘Texture’ includes edge statistics (edge image histogram, edge orientation histogram), local binary patterns, random field decomposition, atomic texture features (contrast, anisotropy, density, randomness, directionality, softness, uniformity, often variations of Tamura features, and often derived from a concurrence matrix of pixel values). The results of wavelet decomposition, Gabor filter and Fourier filters, etc., are also taken as texture features in [2].

‘Shape’ includes elementary descriptors (centroid, area, orientation, length of major and minor axes, eccentricity, circularity, and features derived from algebraic moments), bounding box/ellipse, curvatures scale space, elastic models, Fourier descriptors, template matching, edge direction histogram, etc.

The feature ‘layout’ is the absolute or relative spatial position of the pixels. It may include low-resolution-pixel-map (LRPM), which is used in our method.

In the annotation task of ImageCLEFmed 2005, all provided images are black and white. So texture seems to be a more powerful feature than color. To judge the influence of noises, texture maps are calculated on both initial images and filtered images. Moreover, texture histogram is calculated on those texture maps. In Table 1, three texture features are considered: contrast, anisotropy, and polarity, where ‘h’ means the height of an image and ‘w’ means its width. A~F gives different sets of features. In Figure 5, there is a small example of textures and LRPM.

As we can see from Table 1, if images’ sizes are different, the lengths of the feature vectors will also be different. In this case, LRPMs are used to unify all the images to the size 16x16. On the other hand, LRPMs can reduce the feature vectors’ length. We didn’t try other sizes except 16x16 because in [3] it shows that the sizes of LRPM have no obvious influence to the results.

Table 1. Texture feature sets

	contrast	anisotropy	polarity
Original Image (A)	h x w	h x w	h x w
Filtered Image (B)	h x w	h x w	h x w
Histogram of A (C)	256 x 1	256 x 1	256 x 1
Histogram of B (D)	256 x 1	256 x 1	256 x 1
Filtered C (E)	256 x 1	256 x 1	256 x 1
Filtered D (F)	256 x 1	256 x 1	256 x 1

Besides of texture, regional features are also considered, such as Blob [6]. A Blob's parameters include: color, texture, area, length of long axis and short axis, rotation angle, Fourier decomposition parameters, etc. It has been applied successfully in medical image retrieval in our past work [7].

Facing all kinds of features, it is really very difficult to find out which is more valuable than the others. And it is impossible to use all the features at the same time. So feature selection becomes a key problem. The 'best' features should be the most distinguishing features, and be invariant to different transformations of the input. To choose the best features is a very difficult theoretical problem. A practical way is to do simulation experiments, so as to select suitable features from the best result of classification.

Through simulation experiments, three kinds of LRPMs are finally chosen as features used in our work. One is from the initial images; the other two are from the maps of texture features: contrast and anisotropy.

3. Classification methods

According to [8], roughly speaking, classification methods can be divided into the parametric and the nonparametric. Parametric methods include Bayesian estimation (Maximum-Likelihood, Hidden Markov models, Expectation-Maximization, Fisher Linear Discriminant, Multiple Discriminant Analysis, etc), Linear Discriminant functions (Perceptron Criterion Function, Relaxation Procedures, Minimum Squared-Error Procedures, Principle Component Analysis, Support Vector Machines, Ho-Kashyap Procedures, etc), Multi-layer Neural Networks, Stochastic methods (Simulated Annealing, Boltzmann learning, Evolutionary methods, etc).

SVM is chosen in our program. It is a method widely used for statistical learning, and classifiers and regression models designing. Primarily SVM tackles the binary classification problem. The objective is to find an optimal separating hyper-plane (OSH) that correctly classifies feature data points as much as possible and separates the points of two classes as far as possible. The approach is to map the training data into a higher dimensional (possibly infinite) space and formulate a constrained quadratic programming for the optimization. Different mappings construct different SVMs [9].

SVM for multiple-classes classification is still under development. Generally there are two types of approaches. One type has been to incorporate multiple class labels directly into the quadratic solving algorithm. Another more popular type is to combine several binary classifiers. We used SVM$^{\text{Torch}}$, which belongs to the latter.

Kernel selection is a crucial issue for SVM. Kernels introduce different nonlinearities into the SVM problem by mapping input data into Hilbert space via a mapping function where it may then be linearly separable. Different kernels will accommodate different nonlinear mappings and the performance of the resulting SVM will often hinge on the appropriate choice of the kernel [10]. Generally it is impossible to judge in advance which kernel is the best for classification, and trial-and-error method is a common way to select kernels. There are 4 kernels in SVM$^{\text{Torch}}$: linear, polynomial, radial basis function (RBF), sigmoid tanh.

$$K(\mathbf{x},\mathbf{y}) = (\mathbf{x}\bullet\mathbf{y}+1)^{p} \tag{1}$$

$$K(\mathbf{x},\mathbf{y}) = e^{-\|\mathbf{x}-\mathbf{y}\|^{2}/2\sigma^{2}} \tag{2}$$

$$K(\mathbf{x},\mathbf{y}) = \tanh(\kappa\mathbf{x}\bullet\mathbf{y}-\delta) \tag{3}$$

Eq. (1) results in a classifier that is a polynomial of degree p in the data; Eq. (2) gives a Gaussian radial basis function classifier, and Eq. (3) gives a particular kind of two-layer sigmoidal neural network. For the RBF case, the number of centers, the centers themselves, the weights, and the threshold are all produced automatically by the SVM training and give excellent results compared to classical RBFs, for the case of Gaussian RBFs. For the neural network case, the first layer consists of N sets of weights, each set consisting of d_L (the dimension of the data) weights, and the second layer consists of N weights, so that an evaluation simply requires taking a weighted sum of sigmoids, themselves evaluated on dot products of the test data with the support vectors. Thus for the neural network case, the architecture (number of weights) is determined by SVM training. Note that the hyperbolic tangent kernel only satisfies Mercer's condition for certain values of its parameters [11].

RBF is chosen in our program, and the standard variance σ is its parameter. Another parameter c is the trade-off between training error and the margin.

PCA (Principle Component Analysis) [5] is also tried in our experiments. But owing to the serious unbalance of the training dataset, it can not overcome the over-fitting problem. In Section 4 more details about this will be shown.

4. Experiments and result analysis

Starting from 9,000 labeled training images (see Figure 2 and Table 2), the goal is to classify the 1,000 given unlabeled images. In this process, there are 3 key points: feature selection, method selection, parameter selection. All of these can be decided by simulation experiments.

- **What is the simulation experiment?**

In 'simulation experiments', both the training images and test images are coming from the given training dataset. In this case, the ground-truth of the test dataset is known. Then it is possible to evaluate the result of an experiment. From the feedback of simulation experiments, we can choose the best features, suitable method, and best parameters.

Of course different partitions of the 9,000 training images will cause different results. Thus the simulation experiments should be preceded under various partitions, various features, various methods, and various parameters. Some ratios are assigned to the partitions. Each ratio is applied in each of 57 training sets separately. For example, if defined the ratio to 4:1, in each of 57 training sets, 80% images will be taken as training data and the left 20% will be regarded as test data. Ensuring each class has some training data and corresponding test data, is obviously better than dividing the whole training dataset randomly, in which way, the effect of the classification method cannot be seen clearly because there may be some classes disappearing at the training stage.

- **Simulation experiments with PCA**

First of all, the PCA method is tested in simulation experiments. This is owing to in our past work [5], PCA plus Blob features made a good result in image retrieval. In Figure 6, it is shown the best result of the simulated experiment on PCA plus Blob. The average accuracy (AA) of the classification is 0.5026.

The AA mentioned here is equal to correctness rate, which is calculated according to the following formula:

$$AA = \frac{number\_of\_images\_classified\_correctly}{size\_of\_test\_dataset} \tag{4}$$

For example, in our case, the size of test dataset is 3,512, and there are 1,765 images classified correctly, so the AA is 0.5026.

Figure 7 shows the best result of the simulation experiment on PCA plus texture features. The AA is 0.6977. But using either Blob or texture features, PCA results in almost all the images trapped into a few 'attractive' classes, like 6, 12, 34. This is owing to their big sizes of corresponding training dataset, which are 576, 2563, 880 separately (see Table 2), and the 3 highest peaks in Figure 2. They occupy 44.65% of all the 9,000 training images. This obvious unbalanced distribution between different classes in training dataset can explain why the AA is not so low even in such a bad classification distribution (nearly all of the images are classified into the 3 classes).

If purely from the viewpoint of AA to evaluate the result of PCA for annotation, it's not too bad. But in Figure 6 and Figure 7, PCA is proved to be inoperative in solving such a seriously unbalanced problem, because there are too many empty classes. Like in Figure 7, only 3 classes have some classified samples but all the other classes have none.

- **Simulation experiment with SVM**

Too much training data existing in class 6, 12, 34 cause the risk of over-fitting. However, this risk can be controlled at least for simple noise models, e.g. models with constant noise levels, using soft margin SVM with specific sequences of regularization parameters [12]. In SVM$^{\text{Torch}}$, the parameter for the soft margin is C, and the parameter for the standard variance is std.

In our experiment, SVM plus different features is tested:

SVM+ Blob (Figure 8)
SVM+ LRPM (Figure 9)
SVM+ texture (Figure 10)
SVM+ LRPM + texture (Figure 11)
SVM+ Blob + LRPM + texture (Figure 12)

As shown in the above figures as well as Figure 14, 'SVM + all (=Blob+LRPM+texture)' reaches the highest AA (0.8890).

- **Comparison between methods:**

From those figures we can see that SVM are better than PCA because SVM can 'recognize' more classes. Further more, SVM can reach higher AA compared with PCA.

- **Comparison between features:**

Both by SVM, LRPM features (AA: 0.7725) seems to be better than Blob features (AA: 0.6311). This shows that in this task low-level features are more powerful than middle-level features, although Blob plays well in image retrieval field. Besides texture feature seems to be the best among the three with its AA 0.8318. Further more, the combination of the three kinds of features can reach a highest AA=0.8890. This is owing to different features reflecting different attributes of images, so that their combination can make a better result.

- **Influence of training dataset**

Generally speaking, different training datasets will cause different classification results, and the larger the training dataset is, the better the result will be. But in our case, larger training dataset may make the result worse owing to over-fitting problem. E.g., Figure 12 uses only 600 samples out of 2563 samples from the training dataset of class 12, and its AA is 0.8890; if using 1000 samples to train the classifier, its AA only can reach 0.8727; when using 1500 samples, the AA lowers down to 0.8413.

- **The real experiments and results**

With the conclusions from above simulation experiments, the 'SVM+Blob+LRPM +texture' method is chosen at last, which seems to be the best combination of method and features. But unfortunately, in our submission result to ImageCLEFmed 2005, Blob hasn't been included, which resulted in an AA 0.7940.

Our latest result is 0.8070 (Figure 13), in which, AAs of 11 classes are higher than 0.9, containing 505 correctly classified images in all 515 images; AAs of 23 classes are from 0.5 to 0.9, containing 270 right classified images in all 356 images; in the last 23 classes with the AAs below 0.5, there are 129 images in total, and only 32 images are right classified. This shows that, with the features above, SVM is not good at classifying small classes with few samples.

In the following we discuss some factors influencing the AAs.

- **Threshold of sizes of training datasets**

According to Figure 2 and Table 2, there is a serious unbalance among the training datasets. The largest class contains 2563 samples (class 12), while the smallest class has only 9 samples (class 51, 52). In many cases, too many training samples will cause the over-fitting problem. One of the solutions is to define a threshold to limit the numbers of training samples. For example, the threshold is set to 300. Then each of 57 training datasets is to be limited to 300 training samples.

But as shown in Figure 15, using soft margin SVM, when the threshold increases from 500, it will do little influence to AA. If using PCA, the threshold will do great influence to AA. So on this point, the performance of SVM is much better than PCA.

- **Parameters selection of SVM**

For SVM's kernel RBF, there are two parameters: variance std and margin C.

In Figure 16, (a) shows the influence of std, and (b) shows the influence of C. It can be seen that std does more influence to AA while C does little.

- **Error analysis**

In the information retrieval field, the most common used item to evaluate the retrieval results is PR curve (precision recall curve). Though in our case it is a classification problem, we keep on using them as evaluation figures. The difference is there will be no curves but graphs. A 'curve' is used to describe the effect of different parameters of a retrieval method, and one curve only corresponds to one class. A 'graph' can show all the classes' position in the same coordinates simultaneously, and describe the effect of the classification method. It is more suitable to illustrate the results of a multi-class classifier.

An example of this kind of graph is shown in Figure 17: G is the best region because the points in this region have high recall and precision; B is the worst region because its points have low recall and precision. For a multi-class problem, a convincible result should let its most classes fall in region G. As for ours, related to Figure 13, in Figure 17 there are more than half of the classes (around 53%) falling in G.

5. Conclusions and future work

SVM plus some texture features and blob features reaches an AA of 80.7%, while according to the published results of the annotation task in a link to ImageCLEF's website, the highest AA of ImageCLEFmed is 87.4%. It means 67 right-classified samples more than ours.

For SVM's parameter's tuning, it is proven in our experiments that only std makes sense for better result. Other kernels except RBF are tried but no better.

PR graph is helpful to judge the effects of classification algorithms.

As we can see in our work, features are the most important factor for image classification. In the future, new features should be mined. As for methods, neural network like HMM should be tried.

References

[1] B.Hu, S.Dasmahapatra, P.Lewis, and N.Shadbolt, Ontology-based Medical Image Annotation with Description Logics, *Proceedings of The 15th IEEE International Conference on Tools with Artificial Intelligence*, pp. 77-82, Sacramento, CA, USA.

[2] Remco C. Veltkamp, Mirela Tanase, Content-Based Image Retrieval Systems: A Survey, Technical Report UU-CS-2000-34, Oct. 2000, http://give-lab.cs.uu.nl/cbirsurvey/.

[3] T. Lehmann, M. Güld, T. Deselaers, D. Keysers, H. Schubert, K. Spitzer, H. Ney, and B. Wein. Automatic Categorization of Medical Images for Content-based Retrieval and Data Mining, *Computerized Medical Imaging and Graphics*, vol. 29, pp. 143-155, 2005.

[4] Björn Johansson, A Survey on: Contents Based Search in Image Databases, LiTH-ISY-R-2215, Technical Reports from the Computer Vision Laboratory, Dept. of Electrical Engineering, Linköping University, SWEDEN, Feb., 2000, http://www.cvl.isy.liu.se/ScOut/TechRep/.

[5] Wei Xiong, Bo Qiu, Qi Tian, Changsheng Xu, Sim Heng Ong, Kelvin Foong, Content-based Medical Image Retrieval Using Dynamically Optimized Regional Features, ICIP 2005.

[6] C. Carson, M. Thomas, S. Belongie, J. M. Hellerstein, and J. Malik, Blobworld: A system for region-based image indexingand retrieval, *Proceeding of Third International Conference Visual Information Systems*, 1999.

[7] Wei Xiong, Bo Qiu, Qi Tian, Henning Müller, Changsheng Xu, A novel content-based medical image retrieval method based on query topic dependent image features (QTDIF), *SPIE Medical Imaging,* San Diego, CA, USA, 2005.

[8] Richard O.Duda, Peter E.Hart, David G.Stork, Pattern Classification, A Wiley-Interscience Publication, 2nd ed., ISBN 0-471-05669-3, 2000.

[9] K.-S. Goh, E. Chang, K.-T. Cheng, Support vector machine pairwise classifiers with error reduction for image classification, *Proceedings of the ACM Multimedia Workshop on Multimedia Information Retrieval* (ACM MIR 2001), The Association for Computing Machinery, Ottawa, Canada, pp. 32-37, 2001.

[10] Tony Jebara, Multi-task feature and kernel selection for SVMs, *Proceedings of the twenty-first international conference on Machine learning ICML '04*, July 2004.

[11] C.J.C. Burges, A tutorial on support vector machines for pattern recognition, *Data Mining Knowledge Discovery*, 2:121—167, 1998.

[12] Ingo Steinwart, On the Influence of the Kernel on the Consistency of Support Vector Machines, *Journal of Machine Learning Research*, 2:67-93, 2001.

Appendix

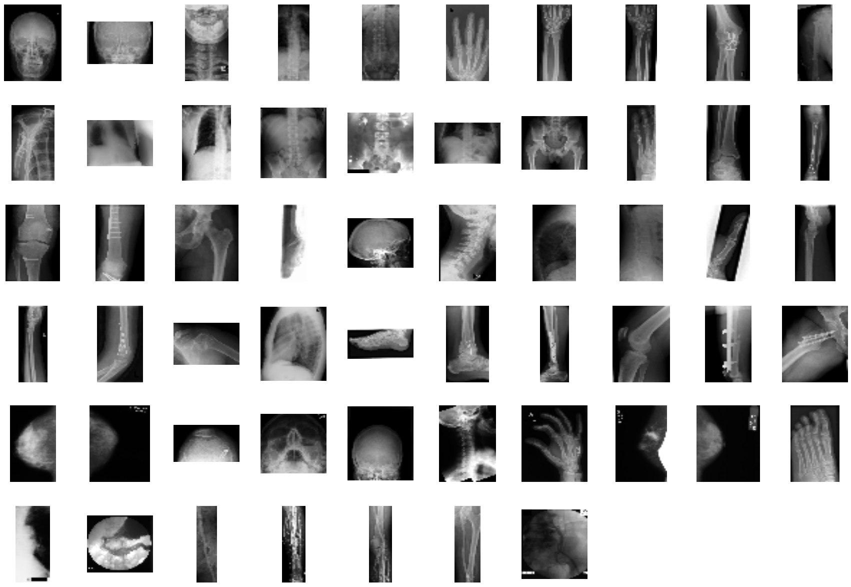


Figure 1. 57 Given classes in annotation task

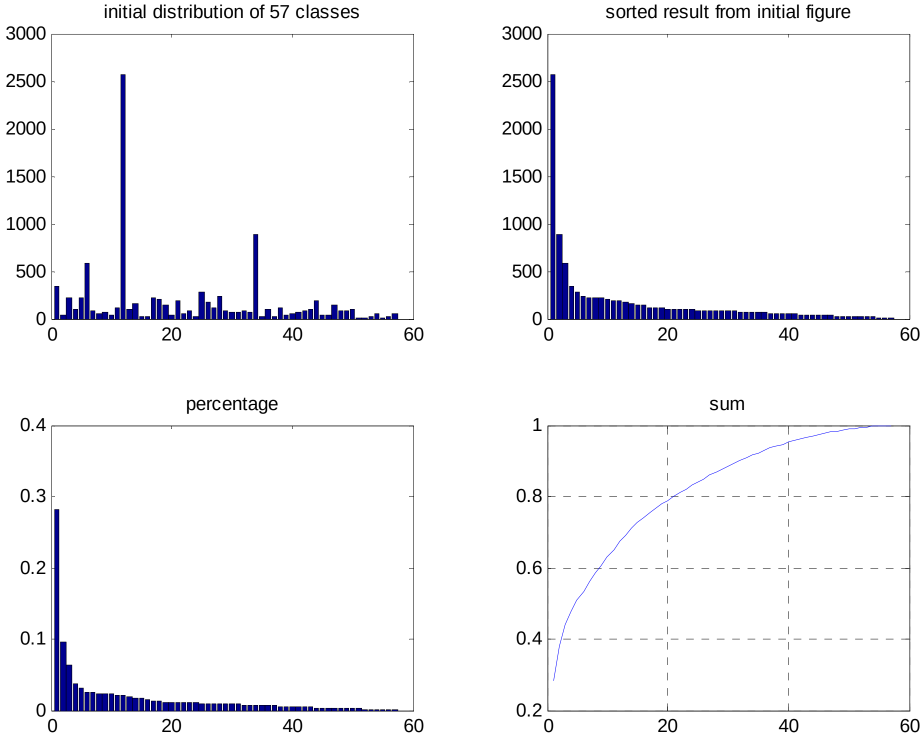


Figure 2. Great unbalance between 57 classes (training sets)

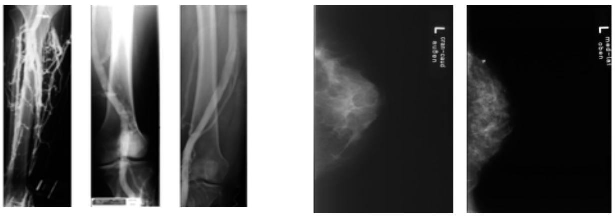


Figure 3. Visual similarities between some classes

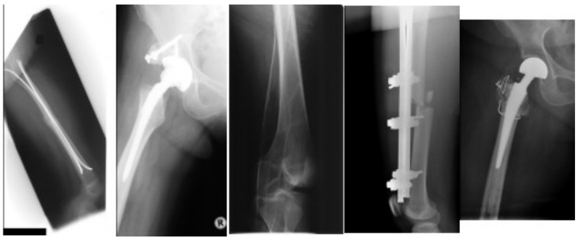
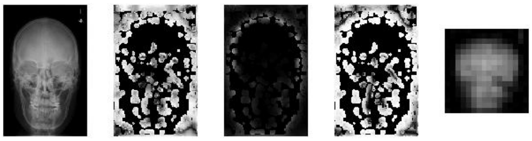


Figure 4. Variety in one class and difficulty to define visual features



Initial : contrast : anisotropy : polarity : LRPM

Figure 5. An image and its feature maps

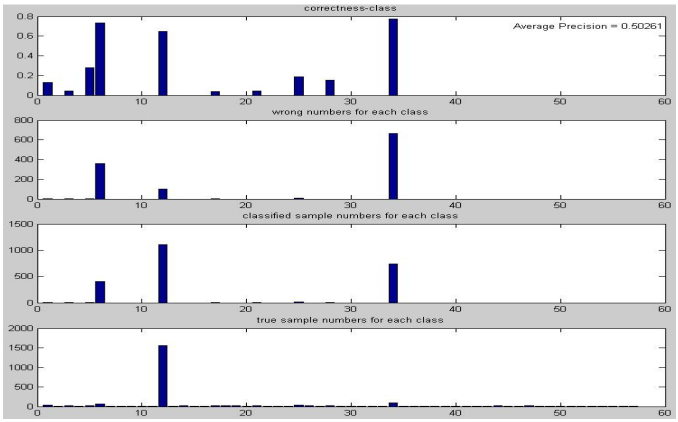


Figure 6. Classification result of PCA using Blob features

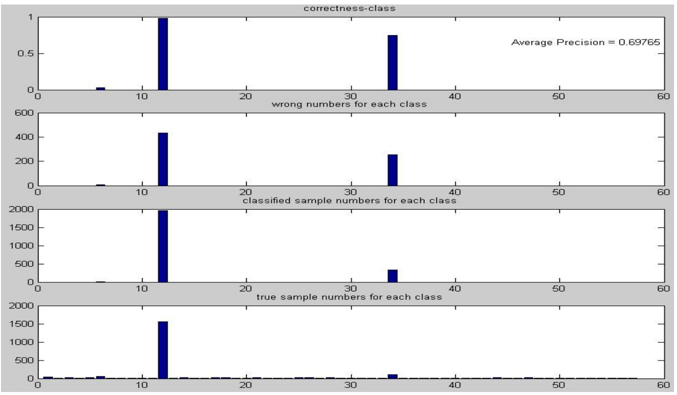


Figure 7. Classification result of PCA using texture features

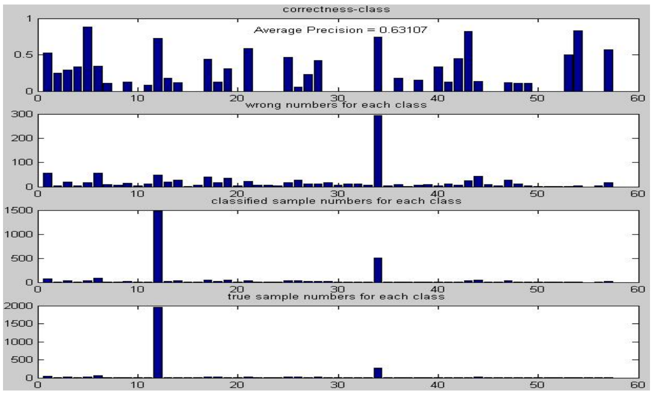


Figure 8. Classification result of SVM using Blob features

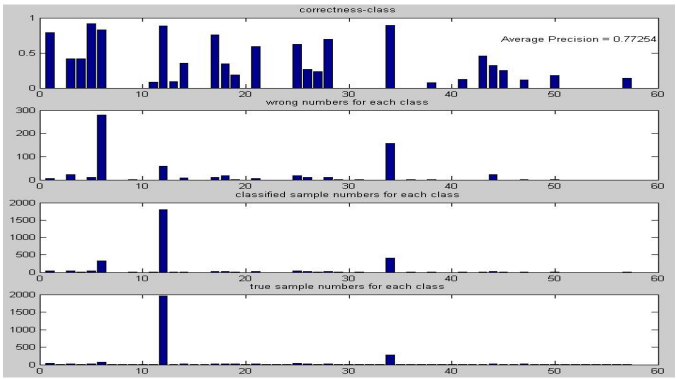


Figure 9. Classification result of SVM using texture features

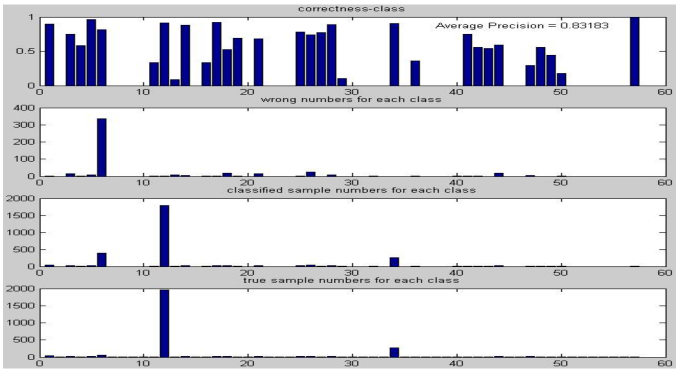


Figure 10. Classification result of SVM using LRPM features

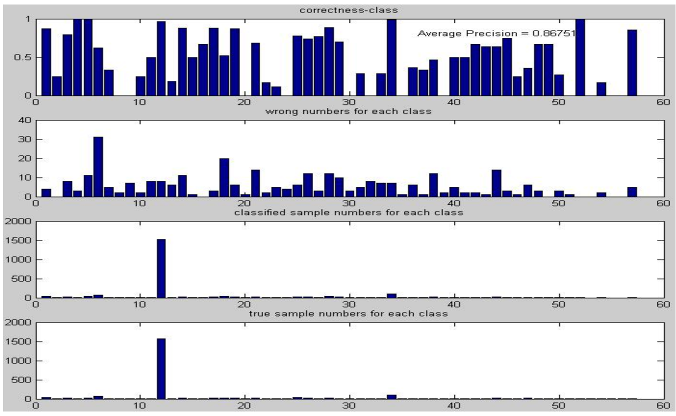


Figure 11. Classification result of SVM using LRPM + texture

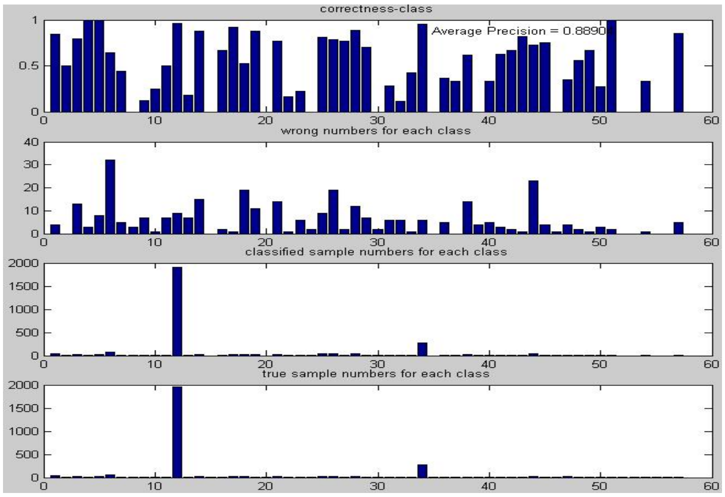


Figure 12. Classification result of SVM using LRPM + texture + Blob

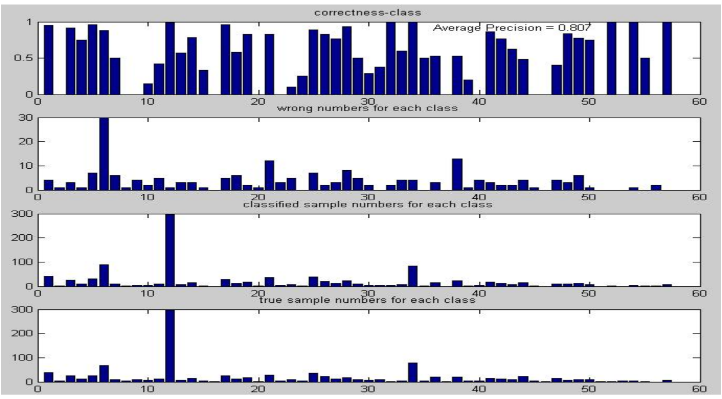


Figure 13. Last classification result for 1000 test images

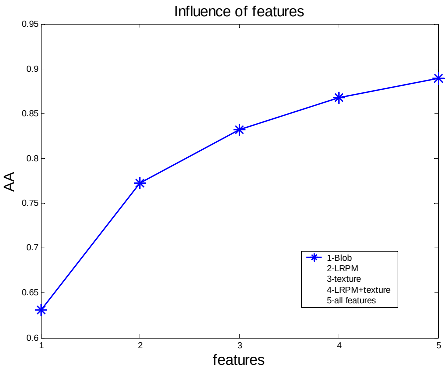


Figure 14. Different features vs. AA

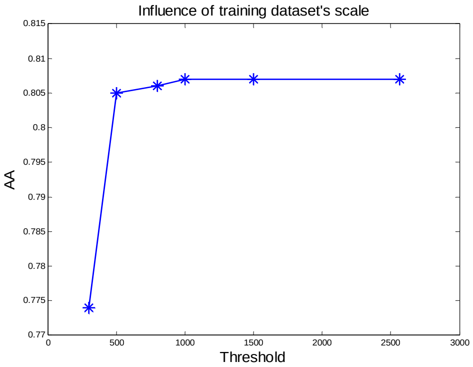
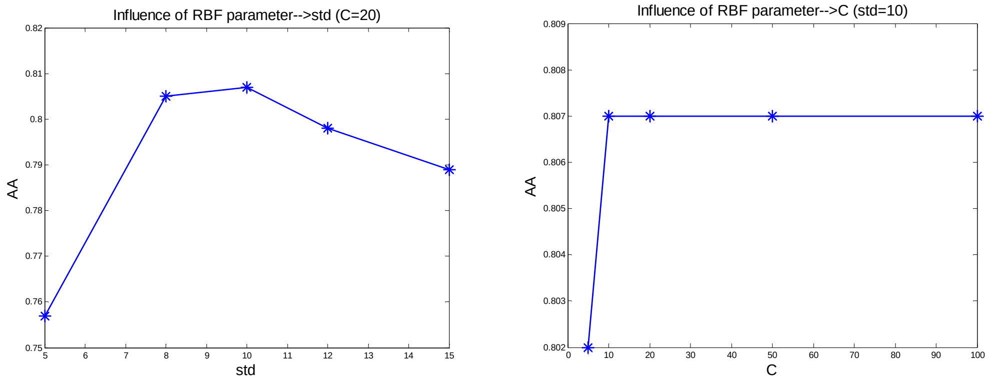


Figure 15. Threshold of training dataset size vs. AA



(a) (b)

Figure 16. Influence of SVM parameters to AA

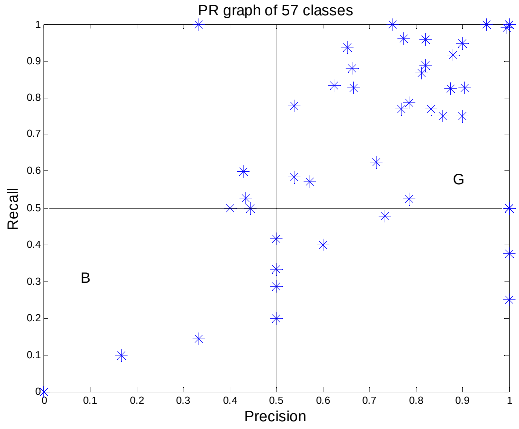


Figure 17. PR graph (G: good; B: bad)

Table 2. Sample numbers of 57 given classes (training datasets)

Class	1	2	3	4	5	6	7	8	9
number	**336**	**32**	**215**	**102**	**225**	**576**	**77**	**48**	**69**
Class	10	11	12	13	14	15	16	17	18
number	**32**	**108**	**2563**	**93**	**152**	**15**	**23**	**217**	**205**
Class	19	20	21	22	23	24	25	26	27
number	**137**	**31**	**194**	**48**	**79**	**17**	**284**	**170**	**109**
Class	28	29	30	31	32	33	34	35	36
number	**228**	**86**	**59**	**60**	**78**	**62**	**880**	**18**	**94**
Class	37	38	39	40	41	42	43	44	45
number	**22**	**116**	**38**	**51**	**65**	**74**	**98**	**193**	**35**
Class	46	47	48	49	50	51	52	53	54
number	**30**	**147**	**79**	**78**	**91**	**9**	**9**	**15**	**46**
Class	55	56	57						
number	**10**	**15**	**57**						