TAU MIPLAB at ImageClef 2008

U. Avni[1], J. Goldberger[2] and H. Greenspan[1]
[1]Tel-Aviv University
[2] Bar-Ilan University

Abstract

This paper describes the participation of Tel Aviv University Medical Image Processing Laboratory group at the ImageClef 2008 medical retrieval and medical annotation tasks. In both tasks we have used the bag-of-words approach for image representation. We submitted two purely visual automatic runs to the medical retrieval task, which used different normalization in the feature extraction stage. Images were converted to a histogram of visual words, and were compared using L1 distance. Our best run was ranked first among the automatic visual based retrieval systems, with MAP score of 0.042. For the medical annotation task we submitted four runs, all used support-vector-machines trained on the visual word histograms. The runs differ in image resolution, and in the way classifiers of two resolutions were combined. In this task our result was second best among the participating groups, with error scores between 105.75 and 117.17.

1 Introduction

In the last several years, "patch-based" representations and "bag-of-features" classification techniques have been proposed for general object recognition tasks [1 - 6]. In these approaches, a shift is made from the pixel entity to a "patch" – a small window centered on the pixel. In its most simplified form, raw pixel values (intensities) within the window are used as the components of the feature vector. It is possible to take the patch information as a collection of pixel values, or to shift the representation to a different set of features based on the pixels, such as SIFT features [7], and reduce the dimensionality of the representation via dimensionality reduction techniques, such as principle-component analysis (PCA) [8].

A very large set of patches are extracted from an image. Each small patch shows a localized "glimpse" at the image content; the collection of thousands and more such patches, randomly selected, have the capability to identify the entire image content (similar to a puzzle being formed from its pieces). A dictionary of words is learned over a large collection of patches, extracted from a large set of images. Once a global dictionary is learned, each image is represented as a collection of words (also known as a "bag of words", or "bag of features"), using an indexed histogram over the defined words. The matching between images, or between an image and an image class, can then be defined as a distance measure between the representative histograms. In categorizing an image as belonging to a certain image class, well-known classifiers, such as the k- nearest neighbor and support-vector machines (SVM) [9], are used.

Patch-based methods have evolved from texton methods in texture analysis [1, 2] and were motivated from the text processing world [3]. In the classical bag-of-features approach, spatial information and geometrical relationship between patches is lost. Recent works have shown that including the spatial information as additional features per patch may provide additional mage characterization strength. The patch-based, bag-of-features approach is simple, computationally efficient, and shows robustness to occlusions and spatial variations. Using this approach, a substantial increase in performance capabilities in general computer-vision object and scene classification tasks has been demonstrated [e.g., 4, 5]. Motivated by these works, and the by success of works based on similar approach in ImageClef2007 challenges [10, 11] we have developed a retrieval and classification system for large medical databases, and put it to the test in ImageClefMed 2008 tasks.

2 Medical Image Retrieval

In this task we are presented with 30 query topics, each with one or more example images, and a short textual description in several languages. Our objective is to return a ranked set of images from a database of over 66,000 images, sorted by their relevance to the presented queries. Some query examples are seen in Figure 4. We have submitted two purely visual automatic runs to this task: TAU_norm and TAU_orig. The two runs differ in the normalization process in the feature extraction stage.

2.1 Method

We model an image as a collection of local patches, where a patch is a small rectangular sub region of the image. Each patch is represented as a codeword index out of a finite vocabulary of visual codewords. Images are compared and classified based on this discrete and compact representation.

We selected a random set of 400 images from the database, and sampled patches of a fixed size of 9x9 pixels with a grid of 6 pixels spacing. We then computed a covariance matrix of this set of roughly 2,000,000 patches, and applied PCA to find its eigenvectors. The 6 vectors with the highest energy are shown in Figure 2. These eigenvectors are later used as a base for the rest of the patches in the database.

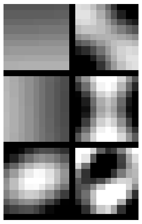


Fig. 2: *PCA components*

2.1.1 Feature extraction

In TAU_norm run, sampled patches are normalized to have 0 mean and 1 variance. This procedure gives some invariance to lighting and contrast. The normalized patch gray level is dimensionally reduced to 6 features using the basis vectors calculated previously. Patch mean gray level was lost in the normalization and PCA process; hence it is added as an additional seventh feature.

In TAU_orig, sampled patches undergo PCA dimensionality reduction to 7 features, without the initial normalization step. In this case the mean gray level is intrinsically included as the first PCA component.

In both runs, in order to preserve spatial information, the patch center coordinates were added as two additional features.

2.1.2 Dictionary

A dictionary of visual words is built using k-means clustering algorithm on the set of sampled patches. The k-means is performed in the feature space, using the L2 distance. Clusters centers initialized by an iterative process, that selects the patch which is furthest from previously selected centers as a new cluster center. Figure 3 shows the centers retuned from the clustering process, after the patches were converted back to image space, and placed in their x,y coordinates.

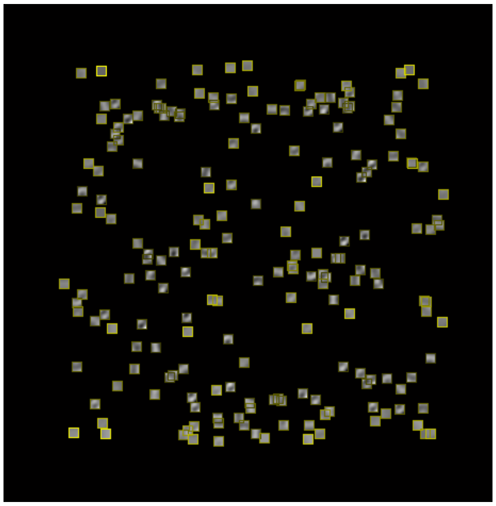


Fig. 3: *Dictionary layout*

Once a dictionary is ready, we sample every image in the database, and represent the image as an indexed histogram of visual words. In this step images are sampled with a denser grid, using a spacing of 2 pixels.

2.1.3 Query process

For image comparison distance measures between the representative histograms are used. Retrieved images are ranked by the distance between a target histogram and the histogram of the query image. When there are several query images we use the minimal distance between the target and the query set. Experiments on the ImageClef medical annotation challenge database indicated that L1 is a simple and effective distance measure.

2.2 Experiments and Results

The runs were ranked according to mean average precision (MAP), which is the arithmetic mean of average precision (AP) values over the 30 individual topics. AP is calculated by averaging the precision in the top *n* retrieved images, where the values of *n* are taken after each relevant image is returned. In addition bpref score [12] is calculated, and precision scores using 5, 10, 15, 20 and 30 retrieved images.

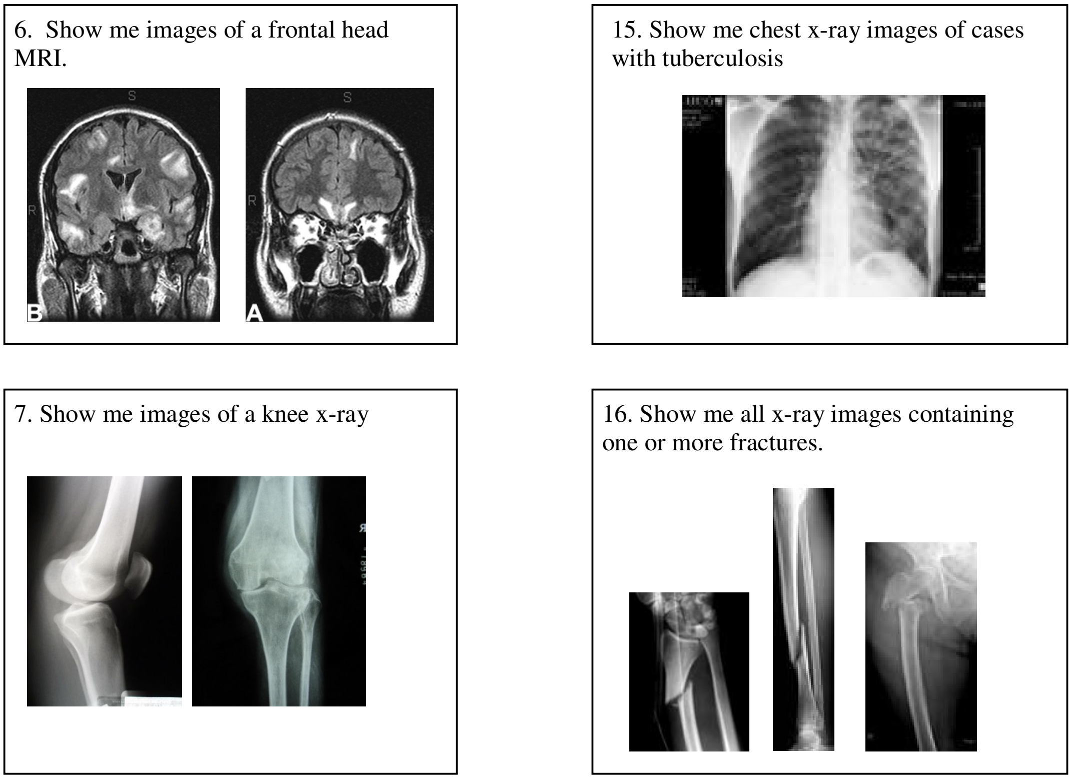
Results of our two submitted runs are shown in Table 1. TAU_norm run ranked first among the automatic visual retrieval runs, according to MAP, bpref, P5, P10, P15 and P20 scores, and second according to P30 score. TAU_orig run ranked 4th according to MAP score, indicating that normalizing the patch variance improves the retrieved results.

Table 1: Ranking of submitted medical image retrieval runs

Rank *(purely visual)*	*Run*	*MAP*	*bpref*	*P5*	*P10*	*P15*	*P20*	*P30*
1	TAU_MIPLAB-TAU_norm	0.042	0.094	0.220	0.170	0.169	0.162	0.146
4	TAU_MIPLAB-TAU_orig	0.031	0.077	0.160	0.143	0.133	0.123	0.112

Using a system that is entirely visual based gives quantitative results which are overall much reduced as compared to text-based systems and mixed runs. This can be seen in the systems comparative table in which all visual based systems are ranked last (provided by the competition organizers). Using the MAP score, the proposed system above is ranked 95 out of 113 total runs. The system is computationally efficient, with average retrieval time of less than 400ms per query on a dual quad-core Intel Xeon 2.33 GHz.

It is interesting to note that on several query topics the proposed system proved highly accurate: On visual topics number 6 and 7 the TAU_norm system ranked 3rd and 17th, respectively. On the mixed topics number 15 and 16 it ranked 1st and 3rd, respectively, among the entire runs. These query topics are displayed in Fig. 4, and the retrieval result returned by TAU_norm for query 15 is displayed in Fig. 5.



6. Show me images of a frontal head MRI.

15. Show me chest x-ray images of cases with tuberculosis

7. Show me images of a knee x-ray

16. Show me all x-ray images containing one or more fractures.

Fig. 4: *Query topics on which our system was effective*

The topics in Figure 4 display body parts with distinct visual features. Our system performed well on these queries because its parameters were tuned using the ImageClef medical annotation challenge database, and as such it specialized in identifying body parts in x-ray images.

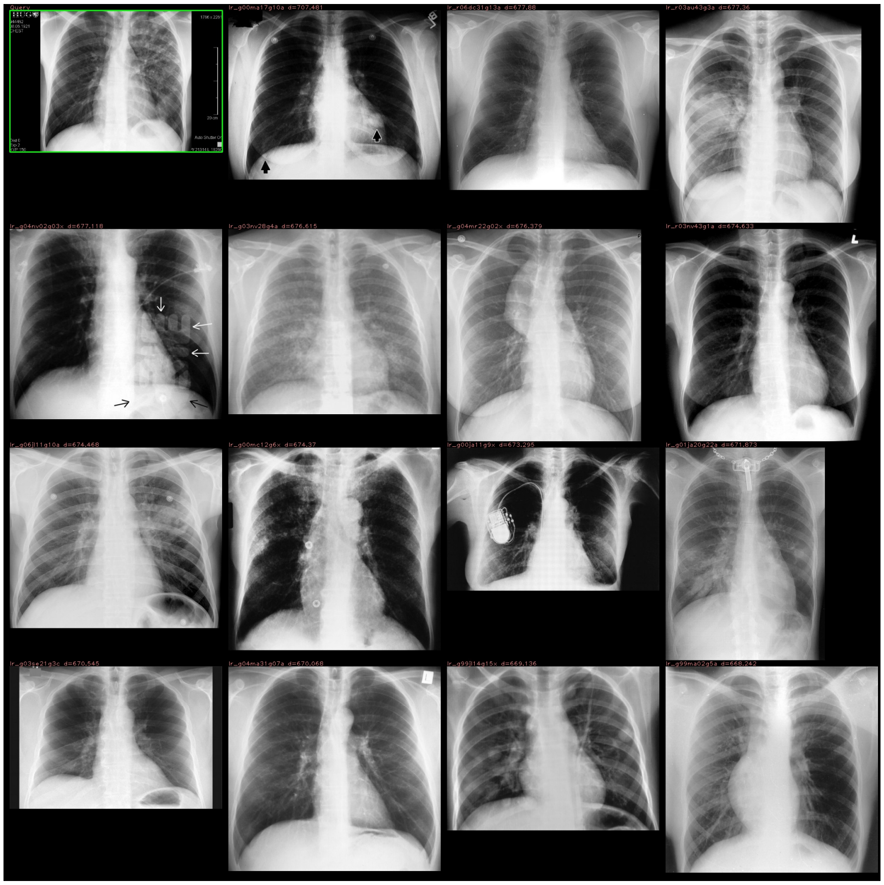


Fig. 5: *Top 15 retrieved images, for topic #15. Top left image is the query sample image. On this query topic TAU_norm run had the highest MAP score from the entire submissions, textual, visual, or mixed.*

3 Medical Image Annotation

In this task we are presented with 12,089 classified x-ray images, and our aim is to classify a set of 1000 previously unseen images, using the hierarchical IRMA code. We submitted four runs to this task, based on the same bag-of-words image representation presented in the previous section, with support-vector-machine classifiers.

3.1 Method

For the annotation challenge we used a dense sampling grid, a patch was extracted around every pixel. The dictionary size in this section was larger, with 700 visual words. Feature extraction, dictionary building and image representation was carried out as in the TAU_norn retrieval run, described in sections 2.1.1 and 2.1.2.

3.1.1 Run Descriptions

Four different runs were provided, as described next. In *svm_full* a support vector machine classifier is trained directly on the image word histograms, using one-vs-one technique for multi-class classification, with radial basis function kernel. Each IRMA code in the training set is treated as a different category label.

In *svm_small* the same method as in svm_full is applied to a 4 times scaled down version of the image, while the patch size remains 9x9 pixels.

In svm_vote the full scale and 1/4 scale classifiers are merged by summing up the votes for each category as was returned from the one-vs-one SVMs.

In *svm_prob* we calculate a probabilistic output of the SVM classifiers using [13], and multiply the probabilities from the full scale and 1/4 scale classifiers. The categories are then sorted by their combined probability. We finally return an IRMA code from the entire hierarchy that minimizes the expected error score on the 5 most probable categories. When the most probable categories have comparable probabilities, the IRMA code returned using this method is often partial, with ‘*’ replacing letters at some location along the IRMA code.

3.2 Experiments and Results

As in ImageCLEF 2007, the medical automatic annotation task rated the classification success according to a penalty score that takes into account the hierarchical structure of the IRMA code- the penalty is greater for errors made in higher levels of the hierarchy.

The results of our runs are shown in Table 2. Our four submissions ranked in places 7 to 10 out of the 24 submitted runs, with error scored between 105.75 and 117.17. This result trails only runs from the Idiap research institute group, which submitted the first 6 best runs, and had the best results last year [10].

Table 2: Ranking of submitted medical image annotation runs

Rank	*Run*	*Error score*
7	TAU-BIOMED-svm_full	105.75
8	TAU-BIOMED-svm_prob	105.86
9	TAU-BIOMED-svm_vote	109.37
10	TAU-BIOMED-svm_small	117.17

Our top ranking runs, svm_full and svm_prob, had a very close score, although they are defined quite differently. The errors in svm_full were due to misclassifications at some point in the IRMA code, where in svm_prob the errors where often due to partial classification.

In this challenge there was a slight advantage to svm_full submission over the rest of the runs, meaning that in our case combining the output of a lower resolution classifier didn’t improve the error score.

The total running time for the whole system, training and classification, was approximately 40 minutes on the full resolution images, and 3 minutes on the 1/4 scaled down images. Times were measured on dual quad-core Intel Xeon 2.33 GHz.

4 Summary

We presented an image retrieval and classification system for large medical databases, based on compact bag-of-features image representation. The system achieves comparatively good results in the ImageClefMed 2008 challenges, while maintaining efficient computation times. These qualities enable effective scaling to larger image collections.

References

1. Leung, T. & Malik, J. (2001). Representing and recognizing the visual appearance of materials using three-dimensional textons. *IJCV*, 43(1), 29–44.

2. Varma, M. & Zisserman, A. (2003) Texture classification: are filter banks necessary? In *CVPR03*, pages II: 691–698.

3. Sivic, J. & Zisserman, A. (2003) "Video Google: A Text Retrieval Approach to Object Matching in Videos," *Proc. Ninth Int'l Conf. Computer Vision*, pp. 1470-1478.

4. Fei-Fei, L. & Perona, P. (2005). A Bayesian Hierarchical Model for Learning Natural Scene Categories. *Proc. of IEEE Computer Vision and Pattern Recognition*: 524-531.

5. Nowak E.et al. (2006). Sampling strategies for bag-of-features image classification. In *ECCV* 06, 406-503.

6. Jiang Y-G, Ngo C-W & Yang J. (2007): Towards optimal bag-of-features for object categorization and semantic video retrieval. *CIVR* 2007: 494-501

7. Lowe. D. (1999) Object recognition from local scale-invariant features. In *Proc. ICCV*, pages 1150–1157.

8. Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press.

9. Vapnik, V. (1995). The Nature of Statistical Learning Theory. Springer Verlag, New York.

10. Tommasi, T., Orabona, F. & Caputo, B. (2007) CLEF2007 Image Annotation Task: an SVM–based Cue Integration Approach. In Working Notes of the 2007 CLEF Work-shop, Budapest, Hungary.

11. Deselaers, T. et al. (2006). Sparse patch– histograms for object classification in cluttered images. In *DAGM* 2006, *Lecture Notes in Computer Science*, Berlin, Germany, 4174,202–211.

12. C. Buckley, E. M. Voorhees (2004) Retrieval evaluation with incomplete information. SIGIR 2004: 25-32

13. T. Lin, C.-J. Lin, and R. C. Weng. (2007) A Note on Platt's Probabilistic Outputs for Support Vector Machines. *Machine Learning*, 68(3), 267-276.