Addressing the ImageClef 2009 Challenge Using a Patch-based Visual Words Representation

U. Avni[1], J. Goldberger[2] and H. Greenspan[1,3]
[1]Tel-Aviv University, Israel
[2] Bar-Ilan University, Israel
[3] IBM Almaden Research Center, San Jose, CA, USA

Abstract

This paper describes our participation at the ImageClef 2009 medical annotation task. In this task we have used the bag-of-words approach for image representation. We submitted one run, using support-vector-machines trained on the visual word histograms in multiple scales. In this task our result ranked first, with error score of 852.8.

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. This work is an enhancement of the classification system we have submitted to the medical annotation challenge in ImageClef 2008 [12].

Medical Image Annotation Task

In this task we are presented with 12,729 classified x-ray images, labeled according to four label sets, labels are based on labeling standards from the last four years. Label sets from 2005 and 2006 contain 57 and 116 labels, respectively. Label sets from 2007 and 2008 contain 116 and 196 IRMA codes. The goal is to classify about 2000 unseen images according the four label sets. Error evaluation scheme is described in the task website.

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 built a dictionary from a random subset of 400 images from the database. The dictionary building process extracts patches of a fixed size of 9x9 pixels with a grid of 6 pixels spacing, patches are normalized to have 0 mean and 1 variance. We then compute a covariance matrix of a set of roughly 2,000,000 patches, and apply PCA to find its eigenvectors. The 6 vectors with the highest energy are shown in Figure 1.

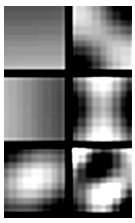

Fig. 1: ***PCA components***

These eigenvectors are later used as a base for the rest of the patches in the database. Patch center coordinates are added to the feature set, in order to include information about the visual words layout. Running k-means algorithm on this set produces 1000 dictionary visual words. A sample dictionary is displayed in Figure 2.

The dictionary building process is repeated in 3 image scales: full resolution, 1/2 scale and 1/8 scale. The resulting dictionary is a concatenation of the 3 dictionaries from the 3 scales. In the image representation step, patches are extracted from each image using a dense grid- around every pixel. An image is represented as a word histogram over the multi-scale dictionary.

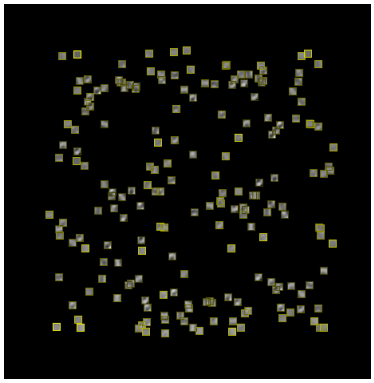

Fig. 2: *Dictionary layout*

Image classification is performed on the word histograms by an SVM classifier with χ^2 kernel. Multi-class classification is implemented using one-vs-one heuristic. In the training step each IRMA code is treated as a separate label, without using the hierarchical nature of the code. The classifier output is returned without using wildcards, except for replacing trailing '0's with '*'. This does not damage the error score if the last digit in the true code is '0', and can reduce the error score if the last digit is non-zero.

Experiments and Results

Kernel parameter γ and SVM tradeoff parameters C were exhaustively searched to minimize cross-validation average error over 5 experiments, where in each experiment 2000 random images served as test data. Parameter tweaking was done using the 2007 code labels. Figure 3 displays the error landscape of the scanned parameters space. The optimal parameters set was used in all four classification tasks.

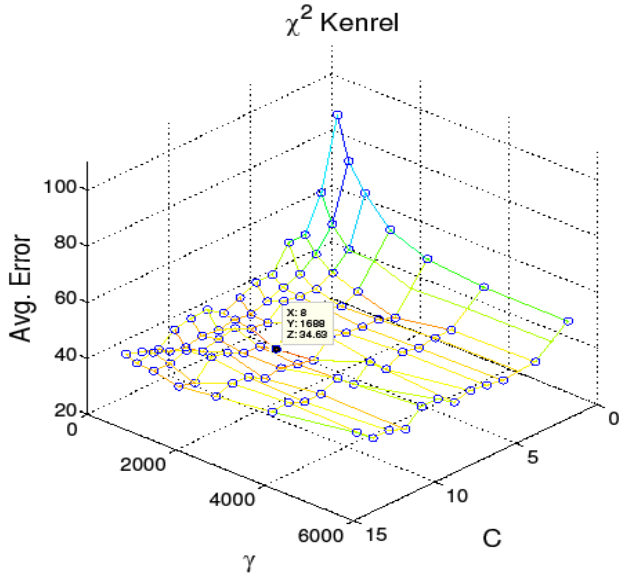

Fig. 3: ***SVM classifier parameters***

The classification on the actual test data of our run is shown in Table 1.

Table 1: Score and ranking of the submitted medical image annotation run

Run	*2005*	*2006*	*2007*	*2008*	*Sum*
Error score	356	263	64.3	169.5	852.8
Rank	1st	2nd	1st	1st	1st

The total running time for the whole system, training and classification, was approximately 90 minutes on a dual quad-core Intel Xeon 2.33 GHz.

Summary

We presented a classification system for large medical databases, based on compact bag-of-features image representation. The system achieves comparatively good results in the ImageClef 2009 medical annotation challenge, while maintaining efficient computation times.

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. U. Avni, J. Goldberger, H. Greenspan., (2008) TAU MIPLAB at ImageClef 2008, in working notes of the 2008 CLEF Work-shop, Aarhus, Denmark