

# Classification-Driven Feature Space Reduction for Semantic-based Image Retrieval

Y. Liu, N. A. Lazar\*, W. E. Rothfus\*\*, M. Buzoianu\* and T. Kanade

The Robotics Institute and \*Statistics Department, Carnegie Mellon University,  
Pittsburgh 15213, USA,

\*\* University of Pittsburgh Medical Center, Pittsburgh, PA [yanxi.tk@cs.cmu.edu](mailto:yanxi.tk@cs.cmu.edu),  
[nlazar@stat.cmu.edu](mailto:nlazar@stat.cmu.edu), [rothfuswe@radserv.arad.upmc.edu](mailto:rothfuswe@radserv.arad.upmc.edu)  
<http://www.cs.cmu.edu/~yanxi/www/home.html>

**Abstract.** This paper summarize our work of the past few years on volumetric pathological neuroimage retrieval under the framework of classification-driven feature selection. In particular, we report our concentrated effort on image feature space reduction for the purposes of (1) reduced computational cost during image retrieval; (2) effective scalability to large image datasets; and (3) improved discriminating power.

## 1 Motivation

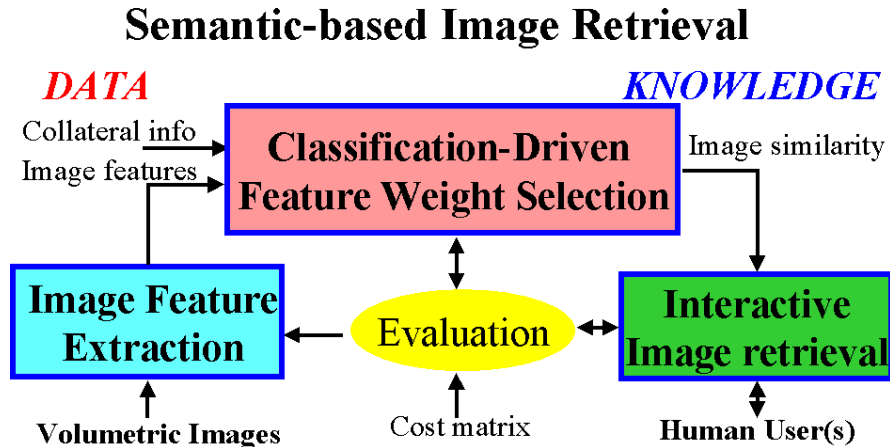
With the current trend towards “paper-less” hospitals, the commercially available Picture Archiving and Communication System (PACS) installed in many hospitals around the world collects several terabytes of on-line image data in individual hospitals monthly, weekly, even daily [2]. However, the utilization of such data for research and education is hampered by the lack of intelligent, effective retrieval capabilities [7]. Using text alone for medical database retrieval has traditionally been the dominating approach for database access. However, text-based methods are limited by predefined vocabularies, which can be subjective, incomplete, coarse, and/or ambiguous [5].

Medical images form an essential and inseparable component of diagnosis, intervention and patient follow-ups. It is therefore natural to use medical images as a front-end index to retrieve medically relevant cases from digital patient databases. Common practice in the image retrieval and pattern recognition community is to map each image into a set of numerical or symbolic attributes called *image indexing features*. Thus each image corresponds to a point in a multidimensional image feature space. Existing “content-based” image retrieval (CBIR) systems [1, 4], depend on general visual properties such as color and texture to classify diverse, two-dimensional (2D) images. These general visual cues, however, often fail to be effective discriminators for image sets taken within a single domain, where images have subtle, domain-specific differences. Furthermore, these global statistical color and texture measures do not necessarily reflect or have proven correspondence to the meaning of an image, i.e. the image semantics, nor are they suitable for handling three-dimensional (3D) volumetric images.

## 2 Our Approach

Our research goal is to establish a systematic framework to extract relevant image features that best reflect image semantics, to automatically construct a semantically discriminating feature subset from a large potential feature set, and to quantitatively evaluate retrieval performance. Figure 1 gives an overview of this approach. The three major components in this scheme are

1. **Feature extraction** maps each volumetric image into a multi-dimensional image feature space;
2. **Feature weighting and image similarity construction** determines the relative scale of the feature space and the best metric for image comparison;
3. **Adaptive image retrieval** captures user intention by choosing the most suitable image similarity.



**Fig. 1.** Overview of a semantic-based classification-driven image retrieval framework.

As a realistic test-case for our methodology, we choose a neuroimage database composed of volumetric CT image sets of hemorrhage (blood), bland infarct (stroke) and normal brains. Justifications for this endeavor are two-fold: first, a database composed of volumetric images and collateral information in a particular medical domain provides objective, semantically well-defined training sets and quantifiable results; second, due to the limitation of the popular color and texture image features used by many existing CBIR systems, finding novel image features and most discriminating feature subsets for medical image characterization becomes crucial.

## 3 Classification-Driven Feature Space Reduction

In this work, feature space reduction is defined as a mapping from a potential feature space  $F_1 = \{f_1, f_2, \dots, f_n\}$  to another feature space  $F_2 = \{w_1 f_1, w_2 f_2, \dots, w_n f_n\}$  where  $0 \leq w_i \leq 1$ . Since some of the  $w_i$ s may equal 0,  $|F_2| \leq |F_1|$ . Two types of

features are expected to be removed from  $F_1$ : irrelevant features and redundant features. As a result, for any image semantic class  $c_i$  the posterior probabilities  $P(c_i|F_1)$  and  $P(c_i|F_2)$  are equivalent. The net result of the feature space reduction includes: (1) reduced computational cost during image retrieval; (2) effective scalability to large image datasets; and (3) improved discriminating power.

There are many existing feature space reduction approaches. Not all of them are appropriate for our final goal of image retrieval. We are looking for methods that can truly reduce computation cost at retrieval time by discarding unnecessary features. Principal component analysis, for instance, does not meet this standard. We have employed multiple methods for reducing the feature space with improved discriminating power and reduced retrieval cost. All following experiments are carried on the same dataset of 48 subjects, of which 26 are normal, 14 had suffered a hemorrhage and 8 had suffered a stroke. The feature set is 46 dimensions and is composed of statistical features describing human brain asymmetry.

### 3.1 Memory-based Learning

In [3] we have reported in detail our work on using a memory-based learning approach to find the most discriminating feature subset. Based on this a non-deterministic search technique, a 5-10 fold reduction in the size of the feature space is accomplished. Average precision rate for image retrieval across different pathologies is near 80%.

### 3.2 Classification Trees

At each step, search for the best binary split among the features, and in this way, a tree is grown. Here, "best" is defined in terms of the maximum reduction in deviance over all allowed splits. If, for example, a split of some node  $s$  is being considered, into two nodes  $t$  and  $u$ , the reduction in deviance, based on a multinomial probability model for the number of cases in each class at each node [6] is  $D_s - D_t - D_u = 2 \sum_k [n_{tk} \log n_{tk} + n_{uk} \log n_{uk} - n_{sk} \log n_{sk} + n_s \log n_s - n_t \log n_t - n_u \log n_u]$ , where  $n_{tk}$  is the number of observations in class  $k$  at node  $t$ , and so forth.

We divided the dataset 50 times randomly into 2/3-1/3 training-testing sets. A tree was grown on each of the 50 training sets. We found that many of the features appeared consistently; one feature for example was used in 35 of the 50 trees, 31 of those times as the first variable to be split on. The next most frequent feature was used in 22 of the 50 trees. This method has achieved a 12 to 24 fold reduction in the feature space, the best we have seen so far. The best classification rates on testing data have reached 100%.

### 3.3 Discriminant Analysis

A combination of forward selection and linear discriminant analysis is another possibility for carrying out classification-driven reduction of the feature space. We performed discriminant analysis using several different criteria for selecting features. Among these were the augmented variance ratio, which compares within-group and between-group variances and penalizes groups whose centers

are too close together, and Wilks' lambda, which measures dissimilarity among groups.

Best results gave a 5 to 9 fold reduction in the size of the feature space. Classification rates averaged around 80% and for some combinations of training and test data, reached 100%.

## 4 Summary

Under our classification-driven semantic-based image retrieval framework, image classification is used as a tool for feature space reduction. This is accomplished by finding that feature subset which is most effective for image semantics classification. Our effort in looking for the best automated methods for feature space reduction differs from most image retrieval practice where the image indexing features are determined by human system designers. Through extensive experiments using multiple feature selection schemes, we are able to find a subspace as low as 2-4 dimensions in a 46 dimensional feature space and achieve high discriminating capability.

## References

1. D. Faloutsos, R. Barber, M. Flickner, J. Hafner, W. Niblack, D. Petkovic, and W. Equitz. Efficient and effective querying by image content. *Journal of Intelligent Information Systems*, 1994.
2. H.K. Huang and R.K. Taira. Infrastructure design of a picture archiving and communication system. *American Journal of Roentgenology*, 158:743-749, 1992.
3. Y. Liu, F. Dellaert, W.E. Rothfus, A. Moore, J. Schneider, and T. Kanade. Classification-driven pathological neuroimage retrieval using statistical asymmetry measures. In *International Conference on Medical Imaging Computing and Computer Assisted Intervention (MICCAI 2001)*. Springer, October 2001.
4. A. Pentland, R.W. Picard, and S. Sclaroff. Photobook: Content-based manipulation of image databases. *IJCV*, 18(3):233-254, June 1996.
5. H.D. Tagare, C.C. Jaffe, and J. Duncan. Medical image databases: a content-based retrieval approach. *J Am Med Inform Assoc*, 4(3):184-198, May 1997.
6. W.N. Venables and B.D. Ripley. *Modern Applied Statistics with Splus, 2nd edition*. Springer-Verlag, London, Paris, Tokyo, 1997.
7. S.T. Wong and H.K. Huang. Design methods and architectural issues of integrated medical image data base systems. *Comput Med Imaging Graph*, 20(4):285-299, July 1996.