CONTENT-BASED IMAGE CATEGORIZATION AND RETRIEVAL USING NEURAL NETWORKS

Yuhua Zhu, Xiuwen Liu

Department of Computer Science
Florida State University
Tallahassee, FL 32306

Washington Mio

Department of Mathematics
Florida State University
Tallahassee, FL 32306

ABSTRACT

We propose a neural network based method for organizing images for content-based image retrieval. We use spectral histogram features, the histograms of filtered images to capture the spatial relationship among pixels as well as global appearance of images. We then find the optimal combination of spectral histogram features using optimal factor analysis to reduce the dimension of features and maximize the discrimination. The reduced features are then used as input to a multiple layer perceptron, which is trained to categorize images based on content using back propagation. For a query image, images are retrieved from different classes based on the categorization probability for the query image. Experimental results on a subset of Corel dataset demonstrate the effectiveness of the proposed method and comparisons show that the proposed method gives significant improvement over other methods.

1. INTRODUCTION

Along with the availability of massive image and video datasets, organizing and finding relevant images has become an interesting and challenging problem with numerous applications. Briefly, the central functionality of any content-based image retrieval (CBIR) system is to give top matches to a query image from large datasets. It is clear that an effective implementation of a CBIR system requires categorization, indexing and retrieval of images, which relies on effective and efficient features to characterize image content. While there are numerous implementations of CBIR systems, their performance is often not satisfactory. The large variation within a class requires discriminative features and effective techniques to maximize retrieval performance.

In this paper, we use a family of spectral histogram features to characterize image content; spectral histograms capture the spatial relationship among pixels through filtering and global appearance of images through histogram, which has shown to be particularly effective for image retrieval [1]. To maximize the effectiveness of the spectral histogram features and reduce the complexity for training and retrieval, we use optimal factor analysis (OFA) to find the optimal combination of features for categorization. The learned features are then used as input to a neural network (a multiple-layer perceptron to be specific), which is trained to categorize images using back propagation. Using the trained neural network, we then categorize and organize images based on classes. During retrieval, for a query image, we first estimate the probability for each class and then retrieve images based on the estimated probability distribution. A distinctive advantage of the proposed method compared to other methods is that by using learning and categorization we avoid a complete ordering of images, which leads to significant performance improvement and computational efficiency.

The rest of the paper is organized as follows. In Section 2 we give a brief description of spectral histogram features along with optimal factor analysis. Section 3 describes our approach to retrieval and experimental results and comparisons are given in Section 4. We conclude the paper with a discussion on future research directions.

2. OPTIMAL FEATURES FOR RETRIEVAL

Here we describe the features that are used for retrieval. One of the distinctive advantages of the proposed method is that we use a learning technique to learn optimal features for discrimination, which improves the discrimination while at the same time reducing the computation complexity.

2.1. Spectral histogram features

Histogram of images is used often in CBIR systems. As statistics of pixel values, histogram captures useful characteristics of images but ignores the relative spatial positions in the images. In other words, histogram does not capture spatial patterns that are important to characterize image content. To overcome this limitation, we use histograms of various spectral components of an image as the signature, which are used in texture analysis and synthesis [2]. The statistics of spectral components retain a great amount of information of images' marginal distributions and texture patterns, hence providing

great discriminating power [3, 4].

More formally, let I be a gray scale image and F a filter, image I_F be the spectral component of I associated with the filter F, which is obtained by convolving I and F. Then we compute the histogram of I_F, which will be denoted as $h(I, F)$. Note that by using different filters, one can characterize different spatial patterns in images, and these histograms collectively are called spectral histograms. For color images, the filters are applied to different color bands, resulting in three histograms for each filter. In our experiment, we utilize a bank of 5 filters, each of them consists of 11 bins in the corresponding histogram. The filters are applied to the three bands of RGB image, resulting in a SH-feature vector of 165.

To demonstrate the effectiveness of the SH features, we have compared the retrieval results using SH features to results reported in [5]. The comparison shows that SH features are comparable to the best results in [5]. For details, see [1].

2.2. Optimal factor analysis

To maximize the discrimination of SH features and reduce dimension and thus computation for retrieval and categorization, we employ optimal factor analysis to find the optimal linear combination of SH features by optimizing the discriminative ability of K-nearest-neighbor classifier. More specifically, a given ensemble of data is divided into training and cross-validation sets, each consisting of labeled representatives from P different classes. The training and cross-validation elements that belong to class c are denoted by $x_{c,1}, \ldots, x_{c,t_c}$ and $y_{c,1}, \ldots, y_{c,v_c}$ respectively, where $1 \leq c \leq P$. Let $A\colon \mathbb{R}^m \to \mathbb{R}^k$ be a linear transformation, quantity $\rho(y_{c,i}; A)$ gives the measurement of how well the nearest neighbor classifier identifies the cross-validation element $y_{c,i}$ as belonging to class c.

$$\rho(y_{c,i}; A) = \frac{\min_{c \neq b,j} \|Ay_{c,i} - Ax_{b,j}\|^p}{\min_j \|Ay_{c,i} - Ax_{c,j}\|^p + \epsilon} \qquad (1)$$

The larger $\rho(y_{c,i}; A)$ is, the closer $y_{c,i}$ lies to a training sample of the class it belongs to than other classes. Then the goal is to choose a transformation A that maximize the average value of $\rho(y_{c,i}; A)$ over the cross validation set. This leads to an optimization problem on a sphere in a high dimensional space, where we use stochastic gradient optimization to find optimal or close to optimal solutions; see [1] for details.

Note that A defines a linear operator on the original feature space, which gives a new metric on the reduced space. In this regard, OFA can be viewed as a technique to learn from a training set an optimal subspace of the original feature space whose associated metric is optimal for categorization based on the nearest neighbor classifier.

3. NEURAL-NETWORK BASED CATEGORIZATION AND RETRIEVAL

To organize images, we use a multiple-layer perceptron (the most widely used neural network family) for categorization. Specifically, we use a three-layer perceptron with the number of hidden units being a parameter (we use 40 hidden units for all the experiments). The input to the neural network is the learned features using OFA; we use a dimension of 9 for the experiments. The neural network is trained using standard back propagation algorithm with a momentum term.

After it is trained, we use the neural network learned for image categorization to retrieve images according to probabilities with which they are associated with each class. We do retrieval in this way by noticing the fact that any classifier optimized to categorize query images does not necessarily rank matches to a query image correctly and thus our method exploits the categorization of the trained neural-network in a more essential way.

More specifically, let I be a query image. We first compute the SH features and then use the learned A to reduce the dimension. The reduced features are used as input to the learned neural-network we obtain a P-dimensional output vector: out, where P is number of classes. For each i, $1 \leqslant i \leqslant P$, we assign a probability that I belongs to class i, as follows:

$$p(i|I) = \frac{e^{-\|1-out_i\|^2/2\sigma^2\tau}}{\sum_{j=1}^{P} e^{-\|1-out_j\|^2/2\sigma^2\tau}}, \qquad (2)$$

where

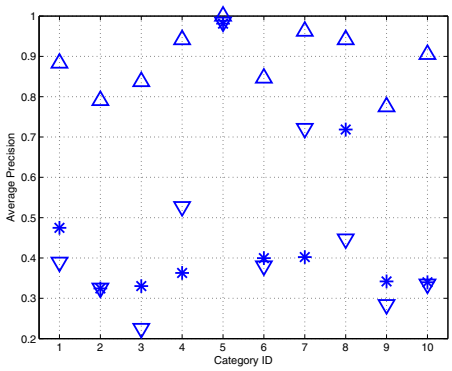
$$\sigma^2 = \frac{1}{P-1} \sum_{j=1}^{P} \|1 - out_j\|^2 \qquad (3)$$

and τ is a "temperature" parameter analogous to that employed in simulated annealing.

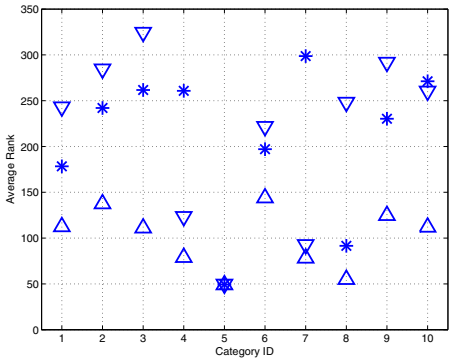
Given a query image I and a positive integer ℓ, the goal of CBIR system is to retrieve top ℓ images listed by rank from the database. Before retrieval starts, all images in the database are indexed according to content using the learned neural-network. The images are retrieved in several passes. In each pass, we retrieve the maximum number possible of images from all classes so that the number of images retrieved from each class is proportional to probabilities $p(i, I)$. In each pass, images in class i_1 are ranked higher than those in class i_2, if $p(i_1, I) > p(i_2, I)$. Once a class is exhausted, we only retrieve from the remaining classes in the subsequent passes, until ℓ images are obtained.

4. EXPERIMENTAL RESULTS

To evaluate the effectiveness of the proposed method and perform fair comparisons, we use a subset of the Corel data set



(a)



(b)

Fig. 1: (a) Average precision within each class; (b) average rank. The methods are labeled as follows: ($\triangledown$) spectral histogram; ($*$) SIMPLIcity; ($\triangle$) neural-network-OFA-600.

consisting of 10 labeled categories, each with 100 images as in [5].[1] The categories are listed in Table 1.

Table 1: Image categories in Corel-1000.

1	African People & Villages
2	Beach
3	Buildings
4	Buses
5	Dinosaurs
6	Elephants
7	Flowers
8	Horses
9	Mountains & Glaciers
10	Food

The reported results of retrieval experiments are all on Corel-1000 data set. The temperature used in probability calculation was set to $\tau = 0.002$. All the training images are part of the whole library. Since each class contains 100 images, there should be at most 100 possible matches to a query image. In order to compare the results with others', we calculate the weighted precision and the average rank, which are

[1] The images were made available by Wang [5] and we acknowledge this.

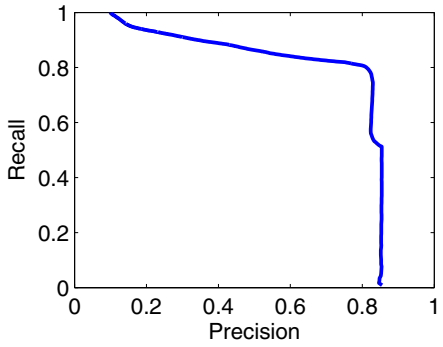


Fig. 2: Corel-1000: plots of the average-precision $\times$ average-recall for using neural network retrieval with 600 training images.

defined as follows: The retrieval precision for the top ℓ returns, is n_ℓ/ℓ, where n_ℓ is the number of correct matches. The weighted precision for a query image I is

$$p(I) = \frac{1}{100}\sum_{\ell=1}^{100}\frac{n_\ell}{\ell}. \tag{4}$$

Rank is referred to as the position in retrieval sequence. Average rank $r(I)$ is the mean value of ranks of all images belonging to the same class as I. The mean values

$$\bar{p}_i = \frac{1}{100}\sum_{I\in C_i} p(I) \text{ and } \bar{r}_i = \frac{1}{100}\sum_{I\in C_i} r(I), \tag{5}$$

of the weighted precision and average rank within each class C_i , $1 \leqslant i \leqslant 10$ will be used as comparison criterion. We compare retrieval results using OFA learning neural-network based retrieval with those obtained with SIMPLIcity[5] and spectral histogram without OFA learning. OFA was used with 600 training images. The comparison is shown in Figue 1. It is clear that retrieval performance is improved significantly with OFA learning.

Retrieval performance is further quantified as follows. For an image I and a positive integer ℓ, m_ℓ is the number of relevant images among the top ℓ retrieval.

$$p_\ell(I) = m_\ell(I)/\ell \quad \text{and} \quad r_\ell(I) = m_\ell(I)/100\,, \tag{6}$$

$p_\ell(I)$ and $r_\ell(I)$ are precision and recall rates for top ℓ retrieval for image I respectively. The average precision and average recall for the top ℓ retrieval are calculated using

$$p_\ell = \frac{1}{1000}\sum_I p_\ell(I) \quad \text{and} \quad r_\ell = \frac{1}{1000}\sum_I r_\ell(I)\,, \tag{7}$$

Several values of the average precision and average recall are shown in Table 2 for neural-network based OFA learning retrieval. Figue 2 shows the full average-precision-recall plots. Figue 3 shows the top 10 matched images for 3 of the 10 classes. In each group, the first image is the query image, which is also the top match.

ℓ	10	20	40	70	100	200	500
p_ℓ	0.866	0.864	0.864	0.832	0.814	0.446	0.191
r_ℓ	0.087	0.173	0.346	0.582	0.814	0.892	0.957

Table 2: Retrieval results with 600 training images. Average retrieval precision (p_ℓ) and recall (r_ℓ) for the top ℓ matches.

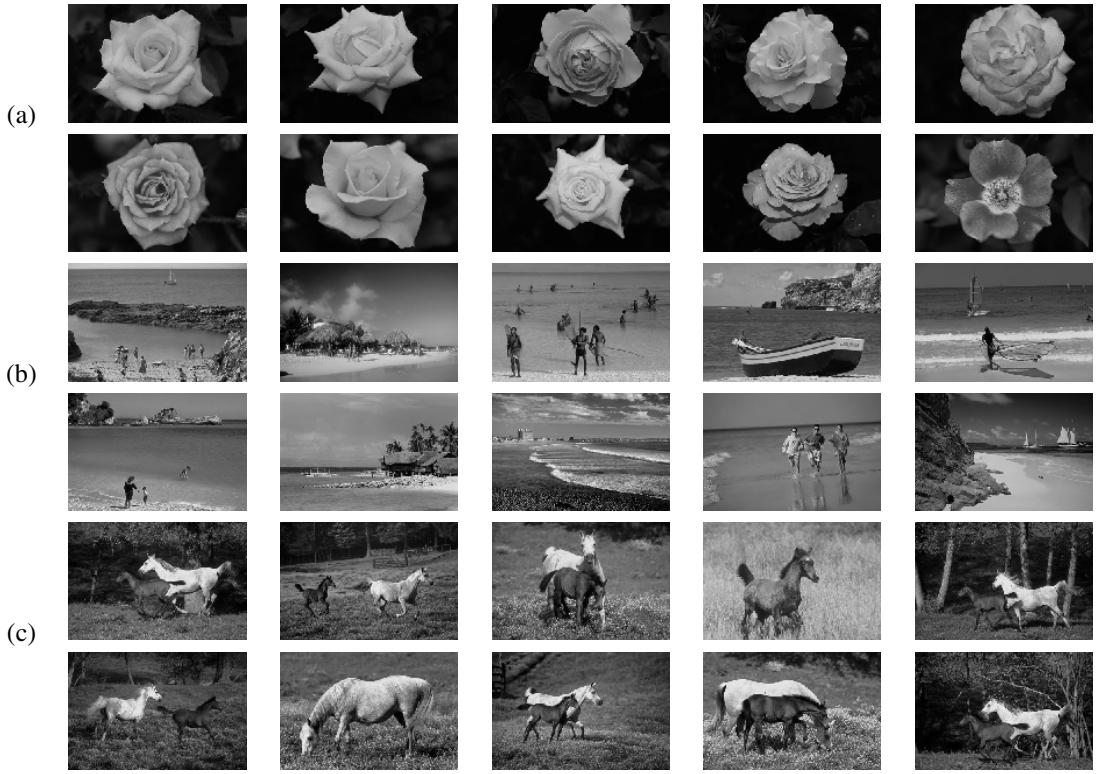


Fig. 3: Examples of top ten returns. In each group, the first image is the query, and also the top return.

5. CONCLUSION

We use spectral histogram of images as features for content-based image retrieval and OFA learning technique is employed to reduce feature dimension and optimize discriminativeness of features. A neural network is applied as a classifier for image categorization and probability estimation in retrieval process. Our experiments and the results indicate significant performance improvement compared to pior work. In the future we will refine the neural-network and OFA learning to cope with nonlinearity structures in data, and investigate in integrating user feedback and real-time animation with our method.

6. REFERENCES

[1] W. Mio., Y. Zhu, and X. Liu, "A learning approach to content-based image categorization and retrieval," in *Proc. 2nd International Conference on Computer Vision Theory and Applications 2007*, 2007.

[2] S.C. Zhu, Y. Wu, and D. Mumford, "Filters, random fields and maximum entropy (FRAME)," *International Journal of Computer Vision*, vol. 27, pp. 1–20, 1998.

[3] J. Portilla and E.P. Simoncelli, "A parametric texture model based on joint statistics of complex wavelet coeficients," *International Journal of Computer Vision*, vol. 40, pp. 49–70, 2000.

[4] Y.N. Wu, S.C. Zhu, and X. Liu, "Equivalence of Julesz ensembles and FRAME models," *International Journal of Computer Vision*, vol. 38, pp. 247–265, 2000.

[5] J. Wang, J. Li, and G. Wiederhold, "SIMPLIcity: Semantics-sensitive integrated matching for picture libraries," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 23, no. 9, pp. 947–963, 2001.