A LEARNING APPROACH TO CONTENT-BASED IMAGE CATEGORIZATION AND RETRIEVAL

Washington Mio
Department of Mathematics, Florida State University, Tallahassee, FL, 32306, USA

Yuhua Zhu, Xiuwen Liu
Department of Computer Science, Florida State University, Tallahassee, FL, 32306, USA

Keywords: Content-based image retrieval, image categorization, image indexing, machine learning, spectral components, dimension reduction, discriminant analysis.

Abstract: We develop a machine learning approach to content-based image categorization and retrieval. We represent images by histograms of their spectral components associated with a bank of filters and assume that a training database of labeled images – that contains representative samples from each class – is available. We employ a linear dimension reduction technique, referred to as Optimal Factor Analysis, to identify and split off "optimal" low-dimensional factors of the features to solve a given semantic classification or indexing problem. This content-based categorization technique is used to structure databases of images for retrieval according to the likelihood of each class given a query image.

1 INTRODUCTION

In this paper, we investigate machine learning techniques for dimension reduction and optimal discrimination, and apply them to content-based categorization and retrieval of images. Large image libraries – such as those found in the World Wide Web and in surveillance and medical databases – are generating a pressing demand for intelligent and scalable systems that can be trained to index and retrieve images according to their contents in a fully automated manner. Classical approaches based on "expert" annotations are simply not a viable option in the presence of massive amounts of data.

For the image categorization problem, we shall assume that a training database containing labeled images representing various different classes is available and the goal is to learn optimal low-dimensional features or "signatures" that can be used to assign a new query image to the correct class. In content-based image retrieval, the objective is to find the top ℓ matches in a database to a query image, where the number ℓ is prescribed by the user. In the proposed approach, retrieval and categorization are closed related. We will use a categorization algorithm to organize a large database according to features learned from a training set. Given a query image I, we will use this organization to estimate the probability that I is associated with a given class and we will retrieve images according to these probabilities.

The problem of classifying images in a database into semantic categories arises in many different levels of generality: for example, the problem can be as broad as separating images that depict an indoor or outdoor scene, or it may involve much more specific categorization into classes such as cars, people, and flowers. As the breadth of the semantic categories may vary considerably, the development of general strategies poses significant challenges. This motivated us to approach the problem in two stages. First, we extract "stable" features that are able to capture a large amount of information about the structure and semantic content of an image. Subsequently, we use learning techniques to identify the factors that have the highest discriminating power for a particular classification problem.

The histogram of an image carries very useful information, however, it tends to have only limited discriminating ability because it encodes the statistics of pixel values, but ignores their relative positions in the image. To remedy the situation, we propose to use histograms of various spectral components of an im-

Mio W., Zhu Y. and Liu X. (2007).
A LEARNING APPROACH TO CONTENT-BASED IMAGE CATEGORIZATION AND RETRIEVAL.
In *Proceedings of the Second International Conference on Computer Vision Theory and Applications - IU/MTSV*, pages 36-43

age as they retain a significant amount of information about texture patterns and edges. The statistics of spectral components have been used in the past primarily in the context of texture analysis and synthesis. In (Zhu et al., 1998), it is demonstrated that marginal distributions of spectral components suffice to characterize homogeneous textures; other studies include (Portilla and Simoncelli, 2000) and (Wu et al., 2000). To provide some preliminary evidence of the discriminating power of spectral histogram (SH) features, in Section 3, we report the results of a retrieval experiment on a database of 1,000 images representing 10 different semantic categories. The relevance of an image is determined by the nearest-neighbor criterion applied to a number of SH-features combined into a single vector. Even without a learning component, we already obtain performances comparable to those exhibited by many existing retrieval systems.

Learning techniques will be employed with a twofold purpose: (a) to identify and split off the most relevant factors of the SH-features for the discrimination of various categories of images; (b) to lower the dimension of the representation to reduce complexity and improve computational efficiency. We adopt a learning strategy that will be referred to as Optimal Factor Analysis (OFA) – a preliminary form of OFA was introduced in (Liu and Mio, 2006) as Splitting Factor Analysis. Given a (small) positive integer k, the goal of OFA is to find an "optimal" k-dimensional linear reduction of the original image features for a particular categorization or indexing problem. Image categorization and retrieval will be based on the nearest neighbor classifier applied to the reduced features, as explained in more detail below. We employ OFA in the context of SH-features, but it will be presented in a more general feature learning framework.

Image retrieval strategies employing a variety of methods have been investigated in (Wang et al., 2001), (Carson et al., 1999), (Rubner et al., 1997), (Smith and Li, 1999), (Yin et al., 2005), (Hoi et al., 2006). Further references can be found in these papers. Some of these proposals employ a relevance feedback mechanism in an attempt to progressively improve the quality of retrieval. Although not discussed in this paper, a feedback component can be incorporated to the proposed strategy by gradually adding to the training set images for which the quality of retrieval was low.

A word about the organization of the paper. In Section 2, we describe the histogram features that will be used to characterize image content. Preliminary retrieval experiments using these features are described in Section 3. Section 4 contains a discussion of Optimal Factor Analysis, and Sections 5 and 6 are devoted to applications of the machine learning methodology to image categorization and retrieval. Section 7 closes the discussion with a summary and a few remarks on refinements of the proposed methods.

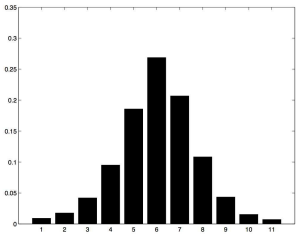
2 SPECTRAL HISTOGRAM FEATURES

Let I be a gray-scale image and F a convolution filter. The spectral component I_F of I associated with F is the image I_F obtained through the convolution of I and F, which is given at pixel location p by

$$I_F(p) = F * I(p) = \sum_q F(q) I(p-q), \qquad (1)$$

where the summation is taken over all pixel locations. For a color image, we apply the filter to its R,G,B channels. For a given set of bins, which will be assumed fixed throughout the paper, we let $h(I,F)$ denote the corresponding histogram of I_F. We refer to $h(I,F)$ as the spectral histogram (SH) feature of the image I associated with the filter F. If the number of bins is b, the SH-feature $h(I,F)$ can be viewed as a vector in $\mathbb{R}^b$. Figure 1 illustrates the process of obtaining SH-features. Frames (a) and (b) show a color image and its red channel response to a Laplacian filter, respectively. The last panel shows the 11-bin histogram of the filtered image.

(a) (b)



(c)

Figure 1: (a) An image; (b) the red-channel response to a Laplacian filter; (c) the associated 11-bin histogram.

If $\mathcal{F} = \{F_1, \ldots, F_r\}$ is a bank of filters, the SH-features associated with the family $\mathcal{F}$ is the collection $h(I,F_i)$, $1 \leqslant i \leqslant r$, combined into the single m-dimensional vector

$$h(I,\mathcal{F}) = (h(I,F_1), \ldots, h(I,F_r)), \qquad (2)$$

where $m = rb$. For a color image, $m = 3rb$. Banks of filters used in this paper typically include Gabor filters of different widths and orientations, gradient filters, and Laplacian of Gaussians.

3 SH-FEATURES FOR RETRIEVAL

To offer some evidence that SH-features are desirable for image retrieval, we perform a preliminary retrieval experiment using the Euclidean distance between histograms. To be able to compare the results with those reported in (Wang et al., 2001), we use the same subset of the Corel data set consisting of 10 semantic categories, each with 100 images. We refer to this data set as Corel-1000. The categories are listed in Table 1 and three samples from each category are shown in Figure 2. As the examples suggest, even within a semantic category, significant variations are observed among the images.

Table 1: Image categories in Corel-1000.

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

We utilize a bank of 5 filters and apply each filter to the R, G, and B channels of the images to obtain a total of 15 histograms per image. Each histogram consists of 11 bins so that the SH-feature vector $h(I,\mathcal{F})$ has dimension 165. For a query image I from the database, we calculate the Euclidean distances between $h(I,\mathcal{F})$ and $h(J,\mathcal{F})$, for every J in the database, and rank the images according to increasing distances. For comparison purposes, as in (Wang et al., 2001), we calculate the weighted precision and the average rank, which are defined next. 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

$$p(I) = \frac{1}{100}\sum_{\ell=1}^{100}\frac{n_\ell}{\ell}. \qquad (3)$$

For a query image I, rank order all 1,000 images in the database, as described above. The average rank $r(I)$ is the mean value of the ranks of all images that belong to the same class as I. Figures 3(a) and 3(b) show the mean values

$$\bar{p_i} = \frac{1}{100}\sum_{I\in C_i} p(I) \text{ and } \bar{r_i} = \frac{1}{100}\sum_{I\in C_i} r(I), \qquad (4)$$

of the weighted precision and average rank within each class C_i, $1 \leqslant i \leqslant 10$. High retrieval performance is reflected in high mean precision and low mean rank. Note that even without a learning component, the results obtained using SH-features and SIMPLIcity are essentially comparable. Both perform considerably better than color histograms with the earth mover's distance (EMD) investigated in (Rubner et al., 1997). In Figure 3, color histograms 1 and 2 refer to EMD applied to histograms with a different number of bins. The results for SIMPLIcity and color histograms have been reported in (Wang et al., 2001).

4 OPTIMAL FACTOR ANALYSIS

We introduce Optimal Factor Analysis (OFA), a linear feature learning technique whose goal is to find a linear mapping that reduces the dimension of the data representation while optimizing the discriminative ability of the K-nearest neighbor (KNN) classifier as measured by its performance on given training data. We assume that a given ensemble of data in Euclidean space $\mathbb{R}^m$ is divided into training and cross-validation sets, each consisting of labeled representatives from P different classes of objects. For an integer c, $1 \leq c \leq P$, we denote by $x_{c,1},\ldots,x_{c,t_c}$ and $y_{c,1},\ldots,y_{c,v_c}$ the training and cross-validation elements that belong to class c.

If $A\colon \mathbb{R}^m \to \mathbb{R}^k$ is a linear transformation, the quantity

$$\rho(y_{c,i};A) = \frac{\min_{c\neq b,j}\|Ay_{c,i} - Ax_{b,j}\|^p}{\min_j \|Ay_{c,i} - Ax_{c,j}\|^p + \varepsilon} \qquad (5)$$

provides a measurement of how well the nearest-neighbor classifier applied to the transformed data identifies the cross-validation element $y_{c,i}$ as belonging to class c. Here, $\varepsilon > 0$ is a small number used to prevent vanishing denominators and $p > 0$ is an exponent that can be adjusted to regularize ρ in different ways. In this paper, we set $p = 2$. A large value of $\rho(y_{c,i};A)$ indicates that, after the transformation A is applied, $y_{c,i}$ lies much closer to a training sample of the class it belongs to than to those of other classes. A value $\rho(y_{c,i};A) \approx 1$ indicates a transition between correct and incorrect decisions by the nearest neighbor classifier. The special case of the function ρ, where $p = 1$, was used in the development of

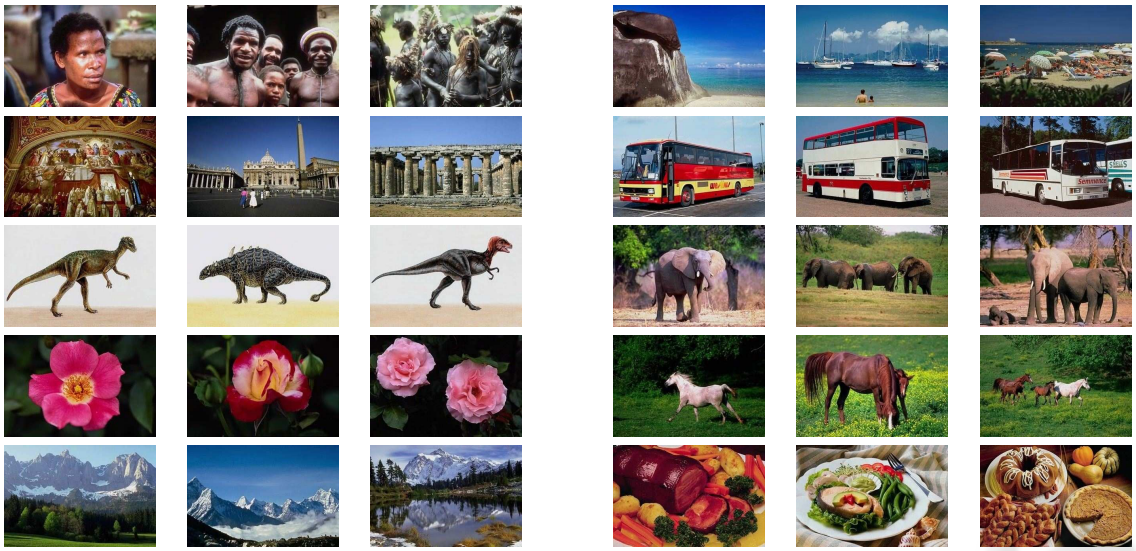


Figure 2: Samples from the dataset Corel-1000: three images from 10 classes, each consisting of 100 images.

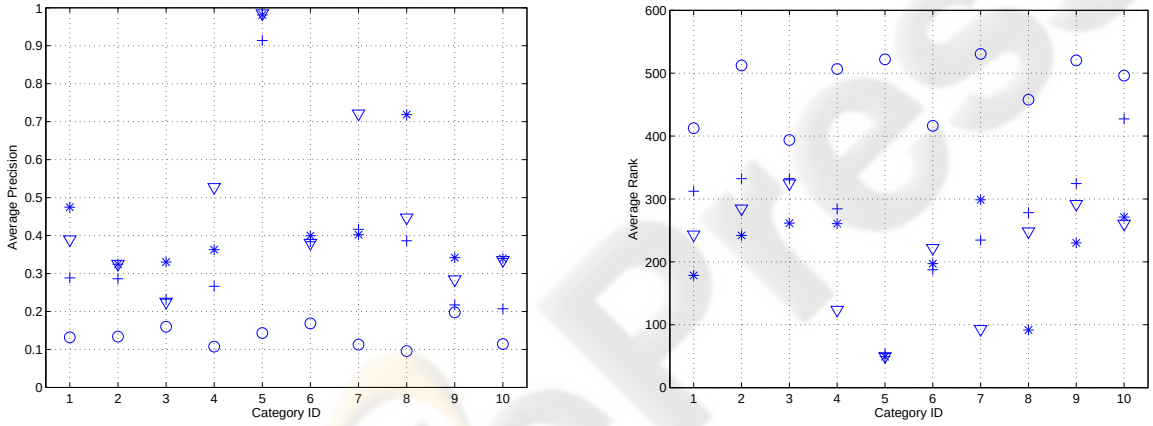


Figure 3: (a) Plots of $\bar{p}_i$ and $\bar{r}_i$, $1 \leqslant i \leqslant 10$. The methods are labeled as follows: ($\triangledown$) spectral histogram; ($*$) SIMPLIcity; ($\circ$) color histogram 1; (+) color histogram 2.

Optimal Component Analysis (Liu et al., 2004). Note that expression (5) can be easily modified to reflect the performance of the KNN classifier.

The idea is to choose a transformation A that maximizes the average value of $\rho(y_{c,i};A)$ over the cross-validation set. To control bias with respect to particular classes, we scale $\rho(y_{c,i};A)$ with a sigmoid of the form

$$\sigma(x) = \frac{1}{1+e^{-\beta x}} \tag{6}$$

before taking the average. We identify linear maps $A\colon \mathbb{R}^m \to \mathbb{R}^k$ with $k \times m$ matrices, in the usual way, and define a performance function $F\colon \mathbb{R}^{k\times m} \to \mathbb{R}$ by

$$F(A) = \frac{1}{P}\sum_{c=1}^{P}\left(\frac{1}{v_c}\sum_{i=1}^{v_c}\sigma\left(\rho(y_{c,i};A)-1\right)\right). \tag{7}$$

Scaling an entire dataset does not change decisions based on the nearest-neighbor classifier. This is reflected in the fact that F is (nearly) scale invariant; that is, $F(A) \approx F(rA)$, for $r > 0$. Equality does not hold exactly if $\varepsilon \neq 0$, but in practice, ε is negligible. Thus, we fix the scale and optimize F over matrices A of unit Frobenius norm. Let

$$\mathbb{S} = \left\{A \in \mathbb{R}^{k\times m} :\ \|A\|^2 = \operatorname{tr}(AA^T) = 1\right\} \tag{8}$$

be the unit sphere in $\mathbb{R}^{k\times m}$. The goal of OFA is to maximize the performance function F over $\mathbb{S}$; that is, to find

$$\hat{A} = \operatorname*{argmax}_{A\in\mathbb{S}} F(A). \tag{9}$$

Due to the existence of multiple local maxima of F, the numerical estimation of $\hat{A}$ is carried out with a stochastic gradient search. We omit the details since

the optimization strategy is similar to that employed in (Liu et al., 2004), but much simpler because the search is performed over a sphere instead of a Grassmann manifold.

4.1 An Interpretation of OFA

We interpret the dimension reduction via a linear map $A\colon \mathbb{R}^m \to \mathbb{R}^k$ and the Euclidean metric in the reduced space in terms of the original m-dimensional features. If A is a rank r matrix, take a singular value decomposition

$$A = U\Sigma V^T, \tag{10}$$

where U and V are orthogonal matrices of dimensions k and m, respectively, and Σ is a $k \times m$ matrix whose $r \times r$ northwest quadrant is diagonal with positive eigenvalues and whose remaining entries are all zero. Let H be the r-dimensional subspace of $\mathbb{R}^m$ spanned by the first r columns of V and denote the orthogonal projection of a vector $x \in \mathbb{R}^m$ onto H by x_H. Then,

$$Ax \cdot Ay = y^T(A^TA)x = y^TKx = y_H^TKx_H, \tag{11}$$

for any $x, y \in \mathbb{R}^m$, where $K = A^TA$ is a positive semi-definite symmetric matrix. In particular,

$$\|Ax - Ay\|^2 = (x_H - y_H)^TK(x_H - y_H). \tag{12}$$

This means that the Euclidean distance between feature vectors in the reduced space $\mathbb{R}^k$ can be interpreted as the distance between the projected vectors x_H and y_H in the original feature space with respect to the new metric

$$d(x_H, y_H) = \sqrt{(x_H - y_H)^TK(x_H - y_H)}. \tag{13}$$

Note that the subspace H is spanned by the eigenvectors of K associated with its non-zero eigenvalues, so that (13) does define a metric on H. Thus, OFA can be viewed as a technique to learn from a training set an optimal subspace of the feature space $\mathbb{R}^m$ for dimension reduction and an inner product whose associated metric is optimal for categorization based on the nearest-neighbor classifier.

4.2 A Leave-One-Out Strategy

In applications of OFA to image retrieval, or in situations where the training set is not very large, a leave-one-out strategy is adopted during the optimization process. In other words, for a candidate linear map A, the value $F(A)$ of the performance function is replaced with the average value of F over several passes, as follows. In each pass, the cross-validation set consists of a single element taken from the training set and $F(A)$ is calculated according to (7). Then, the average value over the entire training set is used to quantify performance in the optimization process.

5 IMAGE CATEGORIZATION

We report the results of several image categorization experiments with the Corel-1000 data set described in Section 3. In each experiment, we placed an equal number of images from each class in the training set and used the remaining ones as query images to be indexed by the nearest neighbor classifier applied to a reduced feature learned with OFA. Initially, an image is represented by an SH-feature vector $h(I, \mathcal{F})$ of dimension 165 obtained from the 11-bin histograms associated with 5 filters applied to the R, G, and B channels; OFA was used to reduce the dimension to $k = 9$. Table 2 shows the categorization performance: T denotes the total number of images in the training set, and categorization performance refers to the rate of correct indexing using all $1,000 - T$ images outside the training set as queries.

Table 2: Results of categorization experiments with the Corel-1000 data set. T is the number of training images and the dimension of the reduced feature space is 9.

T	Categorization Performance
600	84.5%
400	84.3%
200	73.9%

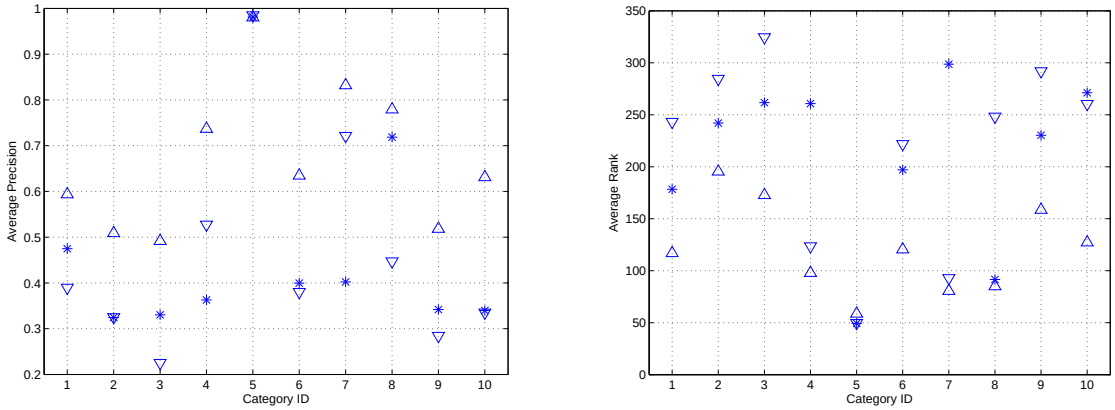
6 IMAGE RETRIEVAL

We now use the classifier learned for content-based image categorization to retrieve images according to their contents. We begin with the remark that the classifier was optimized to categorize query images correctly according to the nearest neighbor criterion, but not necessarily to rank matches to a query image correctly according to distances in feature space. Thus, in contrast with the retrieval strategy based solely on distances adopted, e.g., in (Wang et al., 2001) and (Hoi et al., 2006), we propose to exploit the strengths of the image categorization classifier in a more essential way.

Let $A\colon \mathbb{R}^m \to \mathbb{R}^k$ be the optimal linear dimension-reduction map learned with OFA. If I is an image and $h(I, \mathcal{F}) \in \mathbb{R}^m$ is the associated SH-feature vector, we let x denote its projection to $\mathbb{R}^k$; that is,

$$x = Ah(I, \mathcal{F}). \tag{14}$$

If there are P classes of images, for each i, $1 \leqslant i \leqslant P$, let x_i be the reduced feature vector of the training image in class i, which is closest to x. To each i, we assign a probability $p(i|I)$ that I belongs to class i, as



(a) (b)

Figure 4: (a) Average precision within each class; (b) average rank. The methods are labeled as follows: ($\triangledown$) spectral histogram; ($*$) SIMPLIcity; ($\triangle$) OFA-400.

follows:

$$p(i|I) = \frac{e^{-\|x-x_i\|^2}}{\sum_{j=1}^{P} e^{-\|x-x_j\|^2}}. \qquad (15)$$

Given a query image I and a positive integer ℓ, the goal is to retrieve a ranked list of ℓ images from the database. We assume that all images in the database have been indexed according to content using the classifier learned with OFA. Rank the classes according to the probabilities $p(i|I)$. We retrieve images as follows: select as many images as possible from the most likely class; once that class is exhausted, we proceed similarly with the second most likely class and iterate the procedure until ℓ images are obtained. Within each class, the images are retrieved and ranked according to their Euclidean distances to I as measured in the reduced feature space.

6.1 Experimental Results

We report the results of retrieval experiments with the Corel-1000 dataset. To quantify performance in an objective manner, we only use query images that are part of the database. Since each class contains 100 images, the maximum possible number of matches to a query image is 100, where a match is an image that belongs to the same class. We first compare retrieval results using OFA learning with those obtained with SIMPLIcity and spectral histograms, as described in Section 3. We calculated the mean values $\overline{p_i}$ and $\overline{r_I}$ of the weighted precision and rank as defined in (4). The plots shown in Figure 4 show a significant improvement in retrieval performance with a learning component. OFA was used with 400 training images.

We further quantify retrieval performance, as follows. For an image I and a positive integer ℓ, let m_ℓ be the number of matching images among the top ℓ returns. Define

$$p_\ell(I) = \frac{m_\ell(I)}{\ell} \quad \text{and} \quad r_\ell(I) = \frac{m_\ell(I)}{100}, \qquad (16)$$

which are the precision and recall rates for ℓ returns for image I. The average precision and average recall for the top ℓ returns are defined as

$$p_\ell = \frac{\sum_I p_\ell(I)}{1000} \quad \text{and} \quad r_\ell = \frac{\sum_I r_\ell(I)}{1000}, \qquad (17)$$

respectively. Here, the sum is taken over all 1,000 images in the database. Note that, for a perfect retrieval system, $p_\ell = 1$, for $1 \leqslant \ell \leqslant 100$, and gradually decays to $p_{1000} = 0.1$; similarly $r_\ell = 1$, for $\ell \geq 100$ decaying to $r_1 = 0.01$.

Table 3 shows several values of the average precision and average recall based on a 9-dimensional classifier learned with T training images. The full average-precision-recall plots are shown in Figure 5. Figure 6 shows the top 10 returns for a few images in the database for a classifier trained with $T = 400$ images. In each group, the first image is the query image, which is also the top return.

7 CONCLUSION

We represented images using histograms of their spectral components for content-based image categorization and retrieval. A learning technique was developed to reduce the dimension of the representation and optimize the discriminative ability of the nearest-neighbor classifier. Several experiments were carried out and the results indicate a very significant improvement in retrieval performance over a number

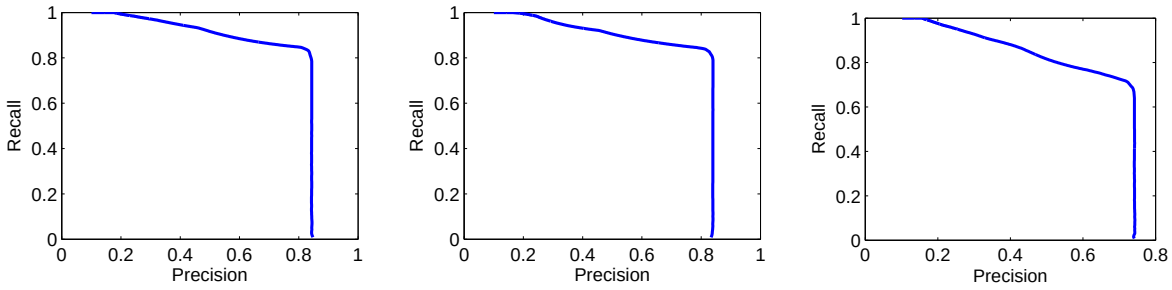


Figure 5: Corel-1000: plots of the average-precision × average-recall for 600, 400, and 200 training images.

Table 3: Retrieval results with T training images. Average retrieval precision (p_ℓ) and recall (r_ℓ) for the top ℓ matches.

$T = 600$

ℓ	10	20	40	70	100	200	500
p_ℓ	0.842	0.0842	0.843	0.843	0.832	0.466	0.198
r_ℓ	0.084	0.168	0.337	0.590	0.832	0.930	0.992

$T = 400$

ℓ	10	20	40	70	100	200	500
p_ℓ	0.840	0.0840	0.840	0.841	0.828	0.460	0.199
r_ℓ	0.084	0.168	0.336	0.589	0.828	0.919	0.996

of existing retrieval systems. Refinements to obtain sparse representations and incorporate kernel techniques to cope with nonlinearity in data geometry, retrieval strategies for real-time execution, as well as a user feedback component will be investigated in future work.

ACKNOWLEDGEMENTS

This work was supported in part by NSF grants CCF-0514743 and IIS-0307998, and ARO grant W911NF-04-01-0268.

REFERENCES

Carson, C., Thomas, M., Belongie, S., Hellerstein, J., and Malik, J. (1999). Blobworld: a system for region-based image indexing and retrieval. In *Proc. Visual Information Systems*, pages 509–516.

Hoi, S., Liu, W., Lyu, M., and Ma, W.-Y. (2006). Learning distance metrics with contextual constraints for image retrieval. In *Proc. CVPR 2006*.

Liu, X. and Mio, W. (2006). Splitting factor analysis and multi-class boosting. In *Proc. ICIP 2006*.

Liu, X., Srivastava, A., and Gallivan, K. (2004). Optimal linear representations of images for object recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 26:662–666.

Portilla, J. and Simoncelli, E. (2000). A parametric texture model based on joint statistics of complex wavelet coeficients. *International Journal of Computer Vision*, 40:49–70.

Rubner, Y., Guibas, L., and Tomasi, C. (1997). The earth mover's distance, multi-dimensional scaling, and color-based image databases. In *Proc. DARPA Image Understanding Workshop*, pages 661–668.

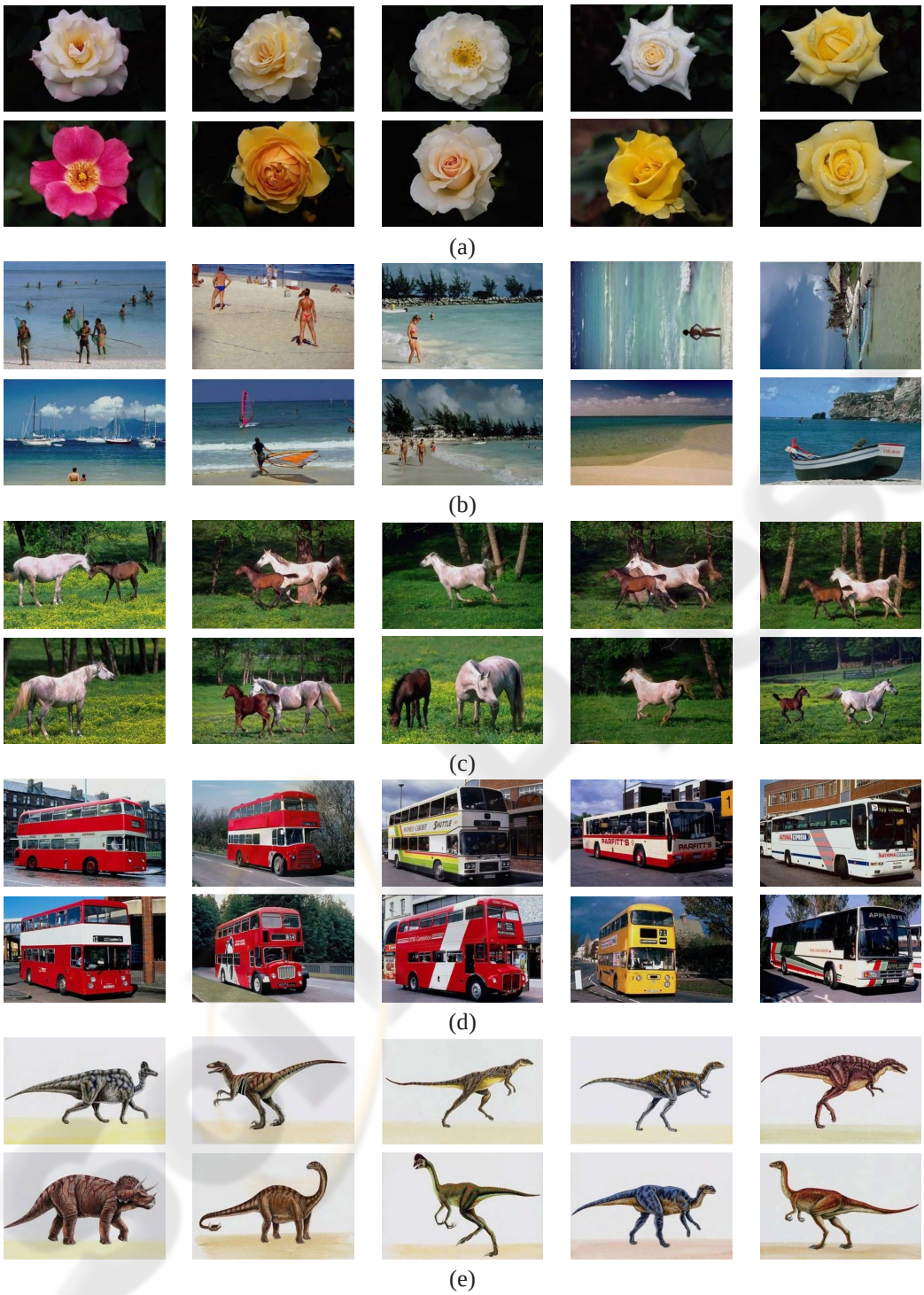
Smith, J. and Li, C. (1999). Image classification and querying using composite region templates. *Computer Vision and Image Understanding*, 75(9):165–174.

Wang, J., Li, J., and Wiederhold, G. (2001). SIMPLIcity: Semantics-sensitive integrated matching for picture libraries. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23(9):947–963.

Wu, Y., Zhu, S., and Liu, X. (2000). Equivalence of Julesz ensembles and FRAME models. *International Journal of Computer Vision*, 38:247–265.

Yin, P.-Y., Bhanu, B., Chang, K.-C., and Dong, A. (2005). Integrating relevance feedback techniques for image retrieval using reinforcement learning. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27(10):1536–1551.

Zhu, S., Wu, Y., and Mumford, D. (1998). Filters, random fields and maximum entropy (FRAME). *International Journal of Computer Vision*, 27:1–20.



(a) (b) (c) (d) (e)

Figure 6: Examples of top ten returns. In each group, the first image is the query, which is also the top return.