A Classification Framework for Content-Based Image Retrieval

Selim Aksoy
Insightful Corporation
1700 Westlake Ave. N., Suite 500
Seattle, WA, 98109-3044
saksoy@insightful.com

Robert M. Haralick
Graduate Center
City University of New York
New York, NY
haralick@gc.cuny.edu

Abstract

A challenging problem in image retrieval is the combination of multiple features and similarity models. We pose the retrieval problem in a two-level classification framework with two classes: the relevance class and the irrelevance class of the query. The first level maps high-dimensional feature spaces to two-dimensional probability spaces. The second level uses combinations of simple linear classifiers trained in these multiple probability spaces to compensate for errors in modeling probabilities in feature spaces. Similarity is computed using joint posterior probability ratios instead of the common way of computing distances in feature spaces and taking their weighted combinations. Experiments on two groundtruthed databases show that the proposed classification framework performs significantly better than the common geometric framework of distances and allows a well-defined and effective way of combining multiple features and similarity measures.

1. Motivation and problem definition

In the rapidly growing content-based image retrieval (CBIR) literature, there has been an enormous amount of work on developing features but similarity measures have not received significant attention. Low-level features like color and texture cannot always capture high-level notion of similarity and distance-based similarity measures often retrieve images that are quite irrelevant to the query image.

An important observation is that different features and different similarity measures perform differently for different types of images. Therefore, developing a framework to combine features and similarity measures looks promising to improve the overall performance. Related approaches in the CBIR literature include appending multiple feature vectors [8], linear or Boolean combinations of distances based on individual feature vectors [2], hierarchical classifiers using different features in each level for pre-defined image classes (e.g. city vs. landscape) [11], using relevance feedback to update weights in a weighted linear combination of multiple features and distance values [9], neural networks [6], and boosting [10]. However, operating in high-dimensional feature spaces often requires complex non-linear tools like neural networks or suffers from the limitations of low-level features and geometric distances between them.

We pose the retrieval problem in a classification framework. The goal is to minimize the classification error in a setting of two classes: the relevance class and the irrelevance class. Given a pair of images, one being the query image and the other one being an image in the database, the pair should be assigned to the relevance class if two images are similar and to the irrelevance class if they are not. Pattern recognition literature provides many choices for a classifier. Since the Bayes classifier gives the theoretical minimum classification error [3], it is the ideal choice for the classifier. Since it uses posterior probabilities to make the decision, the posterior probabilities are the ideal features for classification. The discriminant function to classify the image pair (ξ_i, ξ_j) into the relevance class $\mathcal{A}$ or the irrelevance class $\mathcal{B}$ can be represented in the posterior ratio form

$$\Delta(\xi_i, \xi_j) = \frac{P(\mathcal{A}|(\xi_i, \xi_j))}{P(\mathcal{B}|(\xi_i, \xi_j))} = \frac{P((\xi_i, \xi_j)|\mathcal{A})P(\mathcal{A})}{P((\xi_i, \xi_j)|\mathcal{B})P(\mathcal{B})} \quad (1)$$

which gives the decision rule

$$\text{assign } (\xi_i, \xi_j) \text{ to } \begin{cases} \text{class } \mathcal{A} & \text{if } \Delta(\xi_i, \xi_j) > 1 \\ \text{class } \mathcal{B} & \text{if } \Delta(\xi_i, \xi_j) \leq 1. \end{cases} \quad (2)$$

Then, images can be retrieved by ranking them according to their corresponding posterior ratios instead of geometric distances in the feature space.

This paper focuses on methods to compute the posterior probabilities in Eq. (1). We propose a two-level modeling of probability. In the first level, class-conditional probabilities for the feature vectors are computed using simple parametric models. This can also be interpreted as a mapping from high-dimensional feature spaces to two-dimensional probability spaces. Then, classifiers are trained in these two-dimensional spaces instead of the high-dimensional feature spaces. Since these probabilities are only estimates of the true probabilities, the classifiers trained in the probability spaces implicitly perform a second level modeling of probabilities to compensate for errors in modeling probabilities in the feature spaces. Furthermore, the Bayesian formulation provides a natural way to combine models estimated for multiple feature spaces and multiple classifiers trained on these models. We show that the probabilistic setting performs significantly better than the geometric setting where distances and their weighted linear combinations are used.

Proceedings of the 16 th International Conference on Pattern Recognition (ICPR'02)
1051-4651/02 $17.00 © 2002 **IEEE**

The rest of the paper is organized as follows. Models used for features are summarized in Section 2. Operating in the feature space vs. the probability space is discussed in Section 3. Methods for model combination are described in Section 4. Experiments are discussed in Section 5, and conclusions are given in Section 6.

2. Feature extraction and modeling

Each image in our system is represented by multiple texture and color feature vectors like line-angle-ratio statistics, co-occurrence variances, Gabor features, moments features, Tamura features, and color histograms (see [1] for a detailed description). Our main goal is to develop similarity models so only global low-level features are considered in this paper. However, all of the algorithms proposed here can be directly applicable to features computed from regions.

As mentioned in Section 1, the choice for the Bayes classifier comes from the fact that it minimizes the classification error given that we know the true class-conditional distributions and the prior probabilities. However, these distributions are not exactly known in practice but can be estimated from training data. Using the assumption that similarity between images can be based on the closeness of their feature values, we estimate the class-conditional probabilities using feature difference vectors. In the rest of the paper, we assume that a feature difference vector $\mathbf{d} \in \mathbb{R}^{(q\times 1)}$ has multivariate Gaussian distributions under the relevance and irrelevance classes, denoted by $p(\mathbf{d}|\mathcal{A})$ and $p(\mathbf{d}|\mathcal{B})$ respectively (q is different for different feature spaces). Other possible probability models include independently fitted distributions for each feature and Gaussian mixtures [1].

3. Feature space vs. probability space

Feature vectors usually exist in very high-dimensional spaces (e.g. a 60-dimensional space for Gabor features). The curse of dimensionality and sample size are very important factors in classifier design. Duin [4] argued that reliable classifiers in very small sample size problems can be built by using kernel functions to map the high-dimensional feature space to a low-dimensional kernel space. Our proposed probabilistic setting can also be interpreted as a mapping from the high-dimensional feature space to a two-dimensional probability space. Classification can be done either in the feature space using the feature difference vector $\mathbf{d}$, or in the probability space using the class-conditional probabilities $p(\mathbf{d}|\mathcal{A})$ and $p(\mathbf{d}|\mathcal{B})$ as new features.

The class-conditional probabilities computed using parametric density models in the high-dimensional feature space are only estimates of the true probabilities (because of imperfect density modeling, quantization, dimensionality, etc.). However, classifiers trained in the two-dimensional space of class-conditional probabilities impose a second level modeling of probability, i.e. "probability of probability", to compensate for errors in modeling probabilities in the feature space. We used Gaussian linear, Gaussian quadratic, logistic linear, scaled nearest mean, nearest neighbor, Parzen, binary decision tree and feed-forward neural network classifiers [3, 5]. In a system with I feature representations (feature vectors), J probability models and K classifiers, there are $I \times J \times K$ possible configurations for classification in the probability space and $I \times K$ possible configurations for classification in the feature space. Levels of classification in the feature and probability spaces are summarized in Figures 1 and 2 respectively.

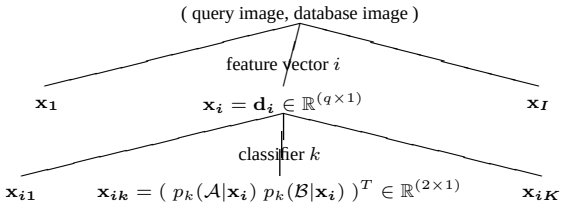


Figure 1: Levels of classification in the feature space for a system with I feature representations and K classifiers. $\mathbf{x}$ represents measurements and $\mathbf{d}$ represents feature difference vectors. For each $\mathbf{x_i}, 1 \le i \le I$, in the feature vector level, K classifiers output the posterior probabilities in $\mathbf{x_{ik}}, 1 \le k \le K$.

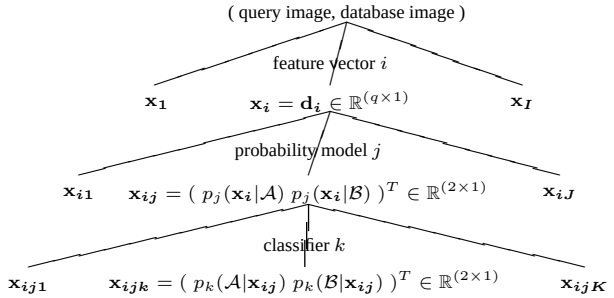


Figure 2: Levels of classification in the probability space for a system with I feature representations, J probability models and K classifiers. $\mathbf{x}$ represents measurements and $\mathbf{d}$ represents feature difference vectors. For each $\mathbf{x_i}, 1 \le i \le I$, J probability models do a mapping to the class-conditional probabilities in $\mathbf{x_{ij}}, 1 \le j \le J$. Then, for each $\mathbf{x_{ij}}$, K classifiers output the posterior probabilities in $\mathbf{x_{ijk}}, 1 \le k \le K$.

4. Feature and similarity combination

Although most of the classifiers may have similar error rates, sets of image pairs misclassified by different classifiers do not necessarily overlap. Classification performance can be further improved by not relying on a single decision but rather by combining the decisions made by the individual classifiers.

The Bayesian framework proposed in this paper provides a natural way to combine multiple measurements on images. Assume that n classifiers with measurement vectors $\mathbf{x_1}, \ldots, \mathbf{x_n}$ are available in our two-class setting. The

Bayesian classifier makes the decision as

$$\text{assign } (\xi_i, \xi_j) \text{ to } \arg\max_{c \in \{\mathcal{A},\mathcal{B}\}} p(c|\mathbf{x_1}, \ldots, \mathbf{x_n}). \quad (3)$$

Computing the joint posterior probability $p(c|\mathbf{x_1}, \ldots, \mathbf{x_n})$ becomes difficult in a practical situation with limited training data. Using equal priors and conditional independence assumptions, further approximations [7, 1] transform the decision rule in Eq. (3) into the following forms:

- Product rule:
  assign (ξ_i, ξ_j) to $\arg\max_{c \in \{\mathcal{A},\mathcal{B}\}} \prod_{i=1}^{n} p(c|\mathbf{x_i})$
- Sum rule:
  assign (ξ_i, ξ_j) to $\arg\max_{c \in \{\mathcal{A},\mathcal{B}\}} \sum_{i=1}^{n} p(c|\mathbf{x_i})$
- Max rule:
  assign (ξ_i, ξ_j) to $\arg\max_{c \in \{\mathcal{A},\mathcal{B}\}} \max_{i=1}^{n} p(c|\mathbf{x_i})$
- Min rule:
  assign (ξ_i, ξ_j) to $\arg\max_{c \in \{\mathcal{A},\mathcal{B}\}} \min_{i=1}^{n} p(c|\mathbf{x_i})$
- Median rule:
  assign (ξ_i, ξ_j) to $\arg\max_{c \in \{\mathcal{A},\mathcal{B}\}} \text{median}_{i=1}^{n} p(c|\mathbf{x_i})$
- Majority vote rule:
  assign (ξ_i, ξ_j) to $\arg\max_{c \in \{\mathcal{A},\mathcal{B}\}} \#\{i \,|\, p(c|\mathbf{x_i}) > 0.5, i = 1, \ldots, n\}$

where $p(c|\mathbf{x_i})$ is the posterior probability given by the classifier i under class c. Since each possible combination of feature vectors, probability models and classifiers gives a set of posterior probabilities (the final level in Figure 2), the classifier combination methods listed above can be directly used to compute posterior ratios to arrive at a final decision about the similarity between images.

5. Experiments

The classification framework proposed in this paper was evaluated using two groundtruthed databases. The first database contains 736 images (texture patches) from the MIT Media Laboratory's VisTex Database with a groundtruth of 46 categories with 16 images in each category. The second database comes from the COREL Photo Stock Library with a total of 1,575 images divided into 18 categories including animals, nature scenes, residential places, cars, etc. Approximately one-third of all images were used for training and the remaining two-thirds were used for testing. (Databases used and experiments presented in this section are described in detail in [1].)

5.1. Classification performance

We did experiments to evaluate performances of using:

- classifiers trained in high-dimensional feature spaces vs. ones trained in two-dimensional probability spaces
- combinations of classifiers trained on multiple probability spaces corresponding to multiple feature vectors and probability models (multivariate Gaussians).

Simple classifiers, like logistic linear or Gaussian quadratic classifiers, trained in the probability space performed much better than the non-linear classifiers, like Parzen, decision tree and neural network classifiers, in the feature space. This is a very useful result because it allows us to do effective classification by training only simple linear classifiers in the probability space. (These results also agree with those of Duin [4].) Combining outputs of a particular classifier trained on multiple probability spaces corresponding to different feature vectors performed better than the cases without combination or when outputs of different classifiers trained on a particular probability space were used (there is a higher chance of violating conditional independence assumptions in the latter case). The most successful combination rule was the product rule with logistic linear or Gaussian quadratic classifiers.

5.2. Retrieval performance

We did experiments to evaluate performances of using:

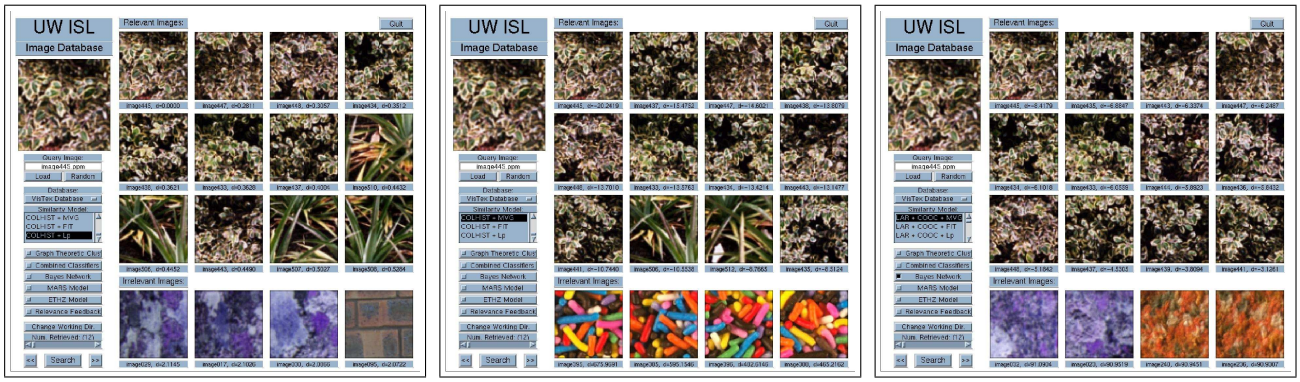
- a single feature vector with Minkowsky L_p metrics vs. posterior ratios from multivariate Gaussians as similarity measures,
- combinations of classifiers trained on multiple probability spaces vs. weighted linear combinations of multiple feature vectors and Euclidean distance values (MARS model from [9]).

Posterior ratios performed significantly better than Minkowsky metrics as similarity measures for individual feature vectors. Classifier combination models that performed the best in classification experiments consistently gave better results than other models in retrieval experiments. They also performed significantly better than linear weighted combinations of distances. The best performing classifier combination was the product rule with logistic linear classifiers. Example queries are given in Figures 3 and 4. Details and precision-recall curves are given in [1].

Effectiveness of simple linear classifiers in improving retrieval results shows the power of the probabilistic framework which simplifies the problem and allows the estimation of less complex models in multiple levels.

6. Conclusions

Numerous feature extraction methods and similarity measures have been proposed in the literature but there is no generally applicable and effective framework to combine multiple features and similarity measures. We posed the retrieval problem in a two-class classification framework where the goal was to minimize the classification error between the relevance and irrelevance classes. We used multivariate Gaussians to model feature vectors to compute posterior probabilities for Bayesian classifiers. However, these posterior probabilities also had uncertainty due to factors like imperfect density modeling, quantization, high dimensionality, etc. To compensate for errors in modeling probabilities in the feature space, we proposed a two-level modeling as the "probability of probability". This setting could be interpreted as a mapping from high-dimensional feature

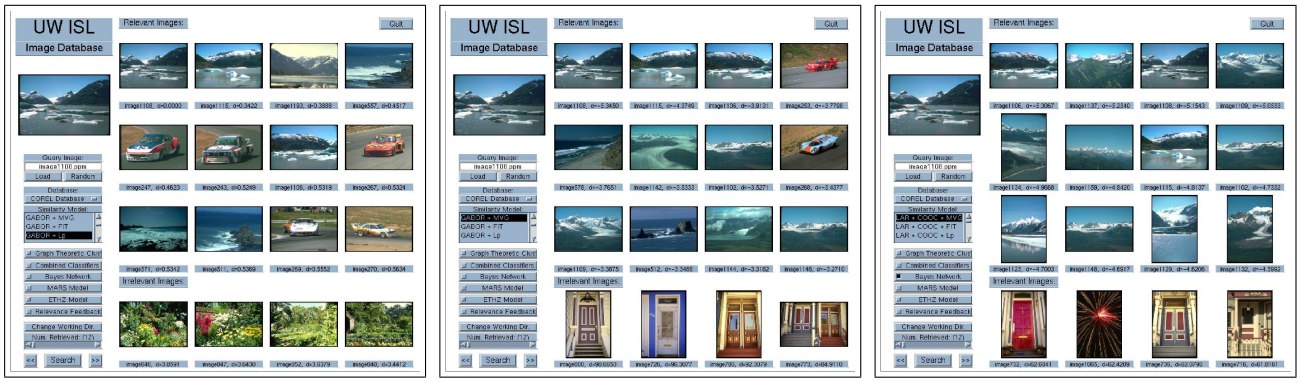


(a) Color histograms and L_p metric (8/12)

(b) Color histograms and multivariate Gaussian (10/12)

(c) Combined classifiers (12/12)

Figure 3: An example query of leaves from the VisTex Database using different similarity measures. The first three rows in the user interface show the best 12 matches and the last row shows the worst 4 matches. The numbers in parentheses in sub-captions show the number of correct matches for each case.



(a) Gabor features and L_p metric (4/12)

(b) Gabor features and multivariate Gaussian (8/12)

(c) Combined classifiers (12/12)

Figure 4: An example query of glaciers and mountains from the COREL Database using different similarity measures.

spaces to two-dimensional probability spaces. We trained simple linear classifiers in multiple two-dimensional probability spaces corresponding to multiple features, and used classifier combination rules to compute joint posterior probabilities for the relevance and irrelevance classes. Posterior ratios were used as similarity measures instead of computing distances in a geometric setting. Experiments showed that the probabilistic framework allows an effective way of combining feature vectors and similarity measures.

References

[1] S. Aksoy. *A Probabilistic Similarity Framework for Content-Based Image Retrieval*. PhD thesis, University of Washington, Seattle, WA, June 2001.

[2] A. P. Berman and L. G. Shapiro. A flexible image database system for content-based retrieval. *Computer Vision and Image Understanding*, 75(1/2):175–195, 1999.

[3] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. John Wiley & Sons, Inc., 2000.

[4] R. P. W. Duin. Classifiers in almost empty spaces. In *Proc. of ICPR*, volume 2, pages 1–7, 2000.

[5] R. P. W. Duin and D. M. J. Tax. Experiments with classifier combining rules. In *Proc. of First Intl. Workshop on Multiple Classifier Systems*, pages 16–29, 2000.

[6] N. Haering and N. de Vitoria Lobo. Features and classification methods to locate deciduous trees in images. *Computer Vision and Image Understanding* 75(1/2):133–149, 1999.

[7] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas. On combining classifiers. *IEEE Trans. on PAMI*, 20(3):226–239, 1998.

[8] C. S. Li and V. Castelli. Deriving texture set for content based retrieval of satellite image database. In *Proc. of ICIP*, pages 576–579, 1997.

[9] Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra. Relevance feedback: A power tool for interactive content-based image retrieval. *IEEE Trans. on Circuits and Systems for Video Technology*, 8(5):644–655, 1998.

[10] K. Tieu and P. Viola. Boosting image retrieval. In *Proc. of CVPR*, volume 1, pages 228–235, 2000.

[11] A. Vailaya, M. Figueiredo, A. Jain, and H. J. Zhang. Content-based hierarchical classification of vacation images. In *Proc. of IEEE Intl. Conference on Multimedia Computing and Systems*, 1999.