RF*IPF: A Weighting Scheme for Multimedia Information Retrieval*

James Z. Wang†
School of Information Sciences and Technology
The Pennsylvania State University
University Park, Pennsylvania, USA
jwang@ist.psu.edu

Yanping Du‡
Department of Electrical Engineering
The Pennsylvania State University
University Park, Pennsylvania, USA
ydu@cse.psu.edu

Abstract

*Region-based approach has become a popular research trend in the field of multimedia database retrieval. In this paper, we present the Region Frequency and Inverse Picture Frequency (RF*IPF) weighting, a measure developed to unify region-based multimedia retrieval systems with text-based information retrieval systems. The weighting measure gives the highest weight to regions that occur often in a small number of images in the database. These regions are considered discriminators. With this weighting measure, we can blend image retrieval techniques with TF*IDF-based text retrieval techniques for large-scale Web applications. The RF*IPF weighting has been implemented as a part of our experimental SIMPLIcity image retrieval system and tested on a database of about 200,000 general-purpose images. Experiments have shown that this technique is effective in discriminating images of different semantics. Additionally, the overall similarity approach enables a simple querying interface for multimedia information retrieval systems.*

1 Introduction

Content-based image retrieval is the retrieval of relevant images from an image database based on automatically derived features. The need for efficient content-based image retrieval has increased tremendously in many application areas such as biomedicine, crime prevention, military, commerce, culture, education, entertainment, and Web image classification and searching.

There are many general-purpose image search engines. In the commercial domain, IBM QBIC [3, 14] is one of the earliest developed systems. Recently, additional systems have been developed at IBM T.J. Watson [21], VIRAGE [4], NEC C&C Research Labs [12], Bell Laboratory [13], Interpix (Yahoo), Excalibur, and Scour.net. In academia, MIT Photobook [15, 16] is one of the earliest. Berkeley Blobworld [1], Columbia VisualSEEK and WebSEEK [20], CMU Informedia [22], UIUC MARS [11], UCSB NeTra [10], UCSD, Stanford WBIIS [25], and Stanford SIMPLIcity [24]) are some of the recent systems.

Many earlier content-based image retrieval systems used color histogram and color layout to index the content of images. Region-based approach has recently become a popular research trend. Region-based retrieval systems attempt to overcome the deficiencies of color histogram and color layout search by representing images at the object-level. A region-based retrieval system applies image segmentation to decompose an image into regions, which correspond to objects if the decomposition is ideal. The object-level representation is intended to be close to the perception of the human visual system (HVS).

Region-based retrieval systems include the Netra system [10], the Blobworld system [1], the query system with color region templates [21], and our recently developed SIMPLIcity (Semantics-sensitive Integrated Matching for Picture LIbraries). SIMPLIcity uses semantics type classification and an integrated region matching (IRM) scheme to provide efficient and robust region-based image matching [9].

Region-based image matching is a difficult problem because of inaccurate segmentation [19, 7, 8, 23]. The IRM measure [9] developed by Li et al. at Stanford University is a similarity measure of images based on region representations. It incorporates the properties of all the segmented regions so that information about an image can be fully used. With IRM, region-based image-to-image matching can be

*This work was supported in part by the National Science Foundation under Grant No. IIS-9817511. and an endowment from the PNC Foundation. The authors wish to thank the help of Jia Li and Gio Wiederhold. An on-line demonstration is provided at URL: http://wang.ist.psu.edu

†J. Wang is also with Department of Computer Science and Engineering, The Pennsylvania State University. Research started when J. Wang was with the Departments of Biomedical Informatics and Computer Science at Stanford University.

‡Y. Du is now with Cisco Systems, Inc.

performed. The overall similarity approach reduces the adverse effect of inaccurate segmentation, helps to clarify the semantics of a particular region, and enables a *simple* querying interface for region-based image retrieval systems. Experiments have shown that IRM is comparatively more effective and more robust than many existing retrieval methods.

(a) (b) (c)

Figure 1. Visual similarity does not always imply semantic similarity when viewing pathology slides. (a) an outdoor photo (b) an overall visually similar photo (c) a semantically similar photo

(a) (b) (c)

Figure 2. Visual similarity does not always imply semantic similarity when viewing pathology slides. (a) a pathology slide (b) an overall visually similar slide (c) a semantically similar slide

Like many existing image distance measures, the discriminating power of the IRM measure is limited by the way it functions, i.e., measuring the distance between two images based only on the information within the two images themselves. In Figure 1, we show three pictures, a picture of people with a large area of sky in the background, a picture of a house with similar background, and a picture of people with a small area of sky in the background. Given that the picture library contains only outdoor scenes, most people consider the third image as closer in semantics to the first image than the second image to the first image. However, most image retrieval systems cannot make this distinction because they rely on overall visual similarity.

Similarly, the problem arises in the retrieval of biomedical images. For example, Figure 2 shows three pathology slides. Based on global visual features, the third image is closer to the first image. However, pathologists find it more useful if the image retrieval system returns the second image as a closer image.

In this paper, we present the Region Frequency and Inverse Picture Frequency (RF*IPF) weighting, a relatively simple weighting measure developed to further enhance the discriminating efficiency of IRM based on the characteristics of the entire picture library. The definition of RF*IPF is in some way close to the definition of the Term Frequency and Inverse Document Frequency (TF*IDF) weighting [18], a highly effective techniques in document retrieval. The combination of RF*IPF and IRM is more effective than the IRM itself in a variety of image retrieval applications. Additionally, this weighting measure provides a better unification of content-based image retrieval and text-based image retrieval.

The remainder of the paper is organized as follows. In Section 2, the similarity measure based on segmented regions is defined. In Section 3, we describe the experiments we performed and provide results. We conclude in Section 4.

2 The Similarity Measure

In this section, we focus on the novel similarity measure we developed. We briefly describe the segmentation process and related notations in Section 2.1. The feature space analysis process is described in Section 2.2. In Section 2.3, we give details of the RF*IRF weighting. The new image matching scheme is given in Section 2.4. We describe the combination scheme for RF*IPF-based image retrieval and TF*IDF-based image retrieval in Section 2.5.

2.1 Region segmentation

Semantically-precise image segmentation is extremely difficult and is still an open problem in computer vision. We attempt to develop a robust matching metric that can reduce the adverse effect of inaccurate segmentation. The segmentation process in our system is very efficient because it is essentially a wavelet-based fast statistical clustering process on blocks of pixels.

To segment an image, SIMPLIcity partitions the image into blocks with $t \times t$ pixels and extracts a feature vector for each block. The k-means algorithm is used to cluster the feature vectors into several classes with every class corresponding to one region in the segmented image. Six features are used for segmentation. Three of them are the average color components in a $t \times t$ block. The other three

represent energy in high frequency bands of wavelet transforms [2], that is, the square root of the second order moment of wavelet coefficients in high frequency bands. We use the well-known LUV color space, where L encodes luminance, and U and V encode color information (chrominance). The LUV color space has good perception correlation properties. We chose the block size t to be 4 to compromise between the texture detail and the computation time.

Let N denote the total number of images in the image database. For the i-th image, denoted as R_i, in the database, we obtain a set of n_i feature vectors after the region segmentation process. Each of these n_i d-dimensional feature vectors represents the dominant visual features (including color and texturečiteKaru:1996) of a region, the shape of that region, the rough location in the image, and some statistics of the features obtained in that region.

2.2 Feature space analysis

Because the definition of the RF*IPF weighting depends on the entire picture library, we process and analyze the characteristics of the d-dimensional feature space.

Suppose feature vectors in the d-dimensional feature space are $\{x_i : i = 1, ..., L\}$, where L is the total number of regions in the picture library. Then $L = \sum_{i=1}^{N} n_i$.

The goal of the feature clustering algorithm is to partition the features into k groups with centroids $\hat{x}_1, \hat{x}_2, ..., \hat{x}_k$ such that

$$D(k) = \sum_{i=1}^{L} \min_{1 \le j \le k} (x_i - \hat{x}_j)^2 \tag{1}$$

is minimized. That is, the average distance between a feature vector and the group with the nearest centroid to it is minimized. Two necessary conditions for the k groups are:

1. Each feature vector is partitioned into the cluster with the nearest centroid to it.
2. The centroid of a cluster is the vector minimizing the average distance from it to any feature vector in the cluster. In the special case of the Euclidean distance, the centroid should be the mean vector of all the feature vectors in the cluster.

These requirements of our feature grouping process are the same requirements as those of the k-means algorithm [5].

If the Euclidean distance is used, the k-means algorithm results in hyper-planes as cluster boundaries. That is, for the feature space $\mathbb{R}^d$, the cluster boundaries are hyper-planes in the $d-1$ dimensional space $\mathbb{R}^{d-1}$.

Both the initialization process and the stopping criterion are critical in the process. We initialize the algorithm adaptively by choosing the number of clusters k by gradually increasing k and stop when a criterion is met. We start with $k = 2$. The k-means algorithm terminates when no more feature vectors are changing classes. It can be proved that the k-means algorithm is guaranteed to terminate, based on the fact that both steps of k-means (i.e., assigning vectors to nearest centroids and computing cluster centroids) reduce the average class variance. In practice, running to completion may require a large number of iterations. The cost for each iteration is $O(kn)$, for the data size n. Our stopping criterion is to stop after the average class variance is smaller than a threshold or after the reduction of the class variance is smaller than a threshold.

2.3 The RF*IPF weighting

The RF*IPF weighting consists of two parameters: the Region Frequency (RF) and the Inverse Picture Frequency (IPF).

For each region feature vector x_i of the image R_j, we find the closest group centroid from the list of k centroids computed in the feature analysis step. That is, we find c_0 such that

$$\| x_i - \hat{x}_{c_0} \| = \min_{1 \le c \le k} \| x_i - \hat{x}_c \| \ . \tag{2}$$

Let's denote N_{c_0} as the number of pictures in the database with at least one region feature closest to the centroid $\hat{x}_{c_0}$ of the image group c_0. Then we define

$$IPF_i = log\left(\frac{N}{N_{c_0}}\right) + 1 \tag{3}$$

where IPF_i is the Inverse Picture Frequency of the feature x_i.

Now let's denote M_j as the total number of pixels in the image R_j. For images in a size-normalized picture library, M_j are constants for all j. Denote $P_{i,j}$ as the area percentage of the region i in the image R_j. Then, we define

$$RF_{i,j} = log(P_{i,j} M_j) + 1 \tag{4}$$

as the Region Frequency of the i-th region in picture j. Then RF measures how frequently a region feature occurs in a picture.

We can now assign a weight for each region feature in each picture. The RF*IPF weight for the i-th region in the j-th image R_j is defined as

$$W_{i,j} = RF_{i,j} * IPF_i \ . \tag{5}$$

Clearly, the definition is close to that of the TF*IDF (Term Frequency times Inverse Document Frequency) weighting in text retrieval.

2.4 Image matching

After computing the RF*IPF weights for all the L regions in all the N images in the image database, we store these weights for the image matching process.

To define the similarity measure between two sets of regions, we assume that the image R_1 and image R_2 are represented by region sets $R_1 = \{r_1, r_2, ..., r_m\}$ and $R_2 = \{r'_1, r'_2, ..., r'_n\}$, where r_i or r'_i is the descriptor of region i. Denote the distance between region r_i and r'_j as $d(r_i, r'_j)$, which is written as $d_{i,j}$ in short. To compute the similarity measure between region sets R_1 and R_2, $d(R_1, R_2)$, we first compute all pair-wise region-to-region distances in the two images. Our matching scheme aims at building correspondence between regions that is consistent with our perception. To increase robustness against segmentation errors, we allow a region to be matched to several regions in another image. A matching between r_i and r'_j is assigned with a significance credit $s_{i,j}$, $s_{i,j} \geq 0$. The significance credit indicates the importance of the matching for determining similarity between images. The matrix $S = \{s_{i,j}\}$, $1 \leq i \leq n$, $1 \leq j \leq m$, is referred to as the significance matrix.

The distance between the two region sets is the summation of all the weighted matching strength, i.e.,

$$d_{IRM}(R_1, R_2) = \sum_{i,j} s_{i,j} d_{i,j} \ .$$

This distance is the integrated region matching (IRM) distance defined in [9].

We now combine the IRM distance with the RF*IPF weighting in the process of choosing the significance matrix S. A natural issue to raise is what constraints should be put on $s_{i,j}$ so that the admissible matching yields good similarity measure. In other words, what properties do we expect an admissible matching to possess? The first property we want to enforce is the fulfillment of significance. We computed the significance W_{i,R_1} of r_i in image R_1 and r'_j in image R_2 is W_{j,R_2}, we require that

$$\begin{aligned} \sum_{j=1}^{n} s_{i,j} &= p_i = \frac{W_{i,R_1}}{\sum_{l=1}^{m} W_{l,R_1}}, \ i = 1, ..., m \\ \sum_{i=1}^{m} s_{i,j} &= q_j = \frac{W_{j,R_2}}{\sum_{l=1}^{n} W_{l,R_2}}, \ j = 1, ..., n \ . \end{aligned}$$

The fulfillment of these significance constraints ensures that all the regions play a role for measuring similarity. The algorithm is given in [9].

2.5 Combining RF*IPF and TF*IDF in multimedia retrieval

We now consider a database of multimedia documents, such as Web pages. Each document is composed of both images and text. The combination of the RF*IPF weighting, the TF*IDF weighting, and the IRM metric provides a general framework for matching multimedia documents. The distance between an image R_1 and another image R_2 can be computed by

$$d(R_1, R_2) = d_{IRM}(R_1, R_2) d_T(R_1, R_2)$$

where D_T is the text distance between the two images using a conventional TF*IDF-based text retrieval method.

3 Experiments

The RF*IPF weighting has been implemented and compared with the first version of our experimental SIMPLIcity image retrieval system. We tested the system on a general-purpose image database (from COREL) including about $200,000$ pictures, which are stored in JPEG format with size 384×256 or 256×384. To conduct a fair comparison, we use only picture features in the retrieval process.

3.1 Speed

On a Pentium III 800MHz PC using the Linux operating system, it requires approximately 60 hours to compute the feature vectors for the $200,000$ color images of size 384×256 in our general-purpose image database. On average, one second is needed to segment an image and to compute the features of all regions. Fast indexing has provided us with the capability of handling outside queries and sketch queries in real-time.

The matching speed is fast. When the query image is in the database, it takes about 1.5 seconds of CPU time on average to sort all the images in the 200,000-image database using our similarity measure. If the query is not in the database, one extra second of CPU time is spent to process the query.

3.2 Accuracy on image categorization

We conducted extensive evaluation of the system. One experiment was based on a subset of the COREL database, formed by 10 image categories, each containing 100 pictures. These categories are africa, beach, buildings, buses, dinosaurs, elephants, flowers, horses, mountains, and food. Within this database, it is known whether any two images are of the same category. In particular, a retrieved image is considered a match if and only if it is in the same category as the query. This assumption is reasonable since the

10 categories were chosen so that each depicts a distinct semantic topic. Every image in the sub-database was tested as a query, and the retrieval ranks of all the rest images were recorded. We computed the precision within the first 100 retrieved images for each query. The recall within the first 100 retrieved images was not computed because it is proportional to the precision in this special case. The total number of semantically related images for each query is fixed to be 100.

We carried out similar evaluation tests for color histogram match. We used LUV color space and a matching metric similar to the EMD described in [17] to extract color histogram features and match in the categorized image database. Two different color bin sizes, with an average of 13.1 and 42.6 filled color bins per image, were evaluated. We call the one with less filled color bins the Color Histogram 1 system and the other the Color Histogram 2 system. Figure 3 shows the performance as compared with the RF*IPF-based SIMPLIcity system. Clearly, both of the two color histogram-based matching systems perform much worse than the RF*IPF-based system in almost all image categories. The performance of the Color Histogram 2 system is better than that of the Color Histogram 1 system due to more detailed color separation obtained with more filled bins. However, the Color Histogram 2 system is so slow that it is impossible to obtain matches on larger databases. SIMPLIcity runs at about twice the speed of the faster Color Histogram 1 system and gives much better searching accuracy than the slower Color Histogram 2 system. The overall performance of the RF*IPF-based system is close to that of the original system which uses area percentages of the segmented regions as significant constraints. However, the two systems return different results for individual queries because they are designed to emphasize different semantics of the images. RF*IPF is better suited to certain applications such as biomedical image databases.

4 Conclusions and Future Work

The Region Frequency and Inverse Picture Frequency (RF*IPF) weighting is a measure designed to combine region-based multimedia retrieval systems with text-based information retrieval systems. The weighting measure has been implemented as part of the IRM metric in the experimental SIMPLIcity image retrieval system. Tested on a database of about 200,000 general-purpose images, the technique has demonstrated high efficiency and robustness. We will further evaluate the method on special image databases (e.g., biomedical), and very large multimedia document databases (e.g., WWW, video).

References

[1] C. Carson, M. Thomas, S. Belongie, J. M. Hellerstein, J. Malik, "Blobworld: a system for region-based image indexing and retrieval," *Proc. Int. Conf. on Visual Information Systems*, D. P. Huijsmans, A. W.M. Smeulders (eds.), Springer, Amsterdam, The Netherlands, June 2-4, 1999.

[2] I. Daubechies, *Ten Lectures on Wavelets*, Capital City Press, 1992.

[3] C. Faloutsos, R. Barber, M. Flickner, J. Hafner, W. Niblack, D. Petkovic, W. Equitz, "Efficient and effective querying by image content," *Journal of Intelligent Information Systems: Integrating Artificial Intelligence and Database Technologies*, vol. 3, no. 3-4, pp. 231-62, July 1994.

[4] A. Gupta, R. Jain, "Visual information retrieval," *Communications of the ACM*, vol. 40, no. 5, pp. 70-79, May 1997.

[5] J. A. Hartigan, M. A. Wong, "Algorithm AS136: a k-means clustering algorithm," *Applied Statistics*, vol. 28, pp. 100-108, 1979.

[6] K. Karu, A.K. Jain, R.M. Bolle, "Is there any texture in the image?," *Pattern Recognition*, vol. 29, pp. 1437-1446, 1996.

[7] J. Li, R. M. Gray, "Context-based multiscale classification of document images using wavelet coefficient distributions," *IEEE Trans. on Image Processing*, vol. 9, no. 9, pp. 1604-1616, September 2000.

[8] J. Li, R. M. Gray, R. A. Olshen, "Multiresolution image classification by hierarchical modeling with two dimensional hidden Markov models," *IEEE Trans. on Information Theory*, vol. 46, no. 5, pp. 1826-1841, August 2000.

[9] J. Li, J. Z. Wang, G. Wiederhold, "IRM: Integrated region matching for image retrieval," *Proc. ACM Multimedia Conference*, 147-156, Los Angeles, ACM, October, 2000.

[10] W. Y. Ma, B. Manjunath, "NaTra: A toolbox for navigating large image databases," *Proc. IEEE Int. Conf. Image Processing*, pp. 568-71, 1997.

[11] S. Mehrotra, Y. Rui, M. Ortega-Binderberger, T.S. Huang, "Supporting content-based queries over images in MARS," *Proc. IEEE Int. Conf. Multimedia Computing and Systems*, pp. 632-3, Ottawa, Ont., Canada 3-6 June 1997.

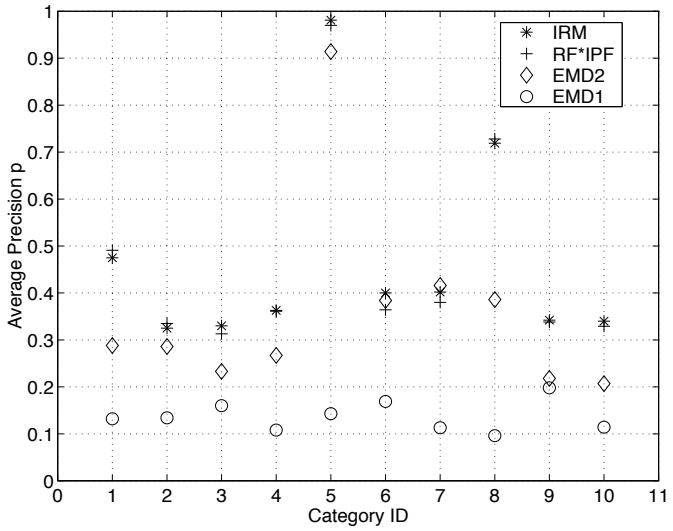

Figure 3. Comparing with color histogram methods on average precision p. Color Histogram 1 gives an average of 13.1 filled color bins per image, while Color Histogram 2 gives an average of 42.6 filled color bins per image. SIMPLIcity partitions an image into an average of only 4.3 regions.

[12] S. Mukherjea, K. Hirata, Y. Hara, "AMORE: a World Wide Web image retrieval engine," *World Wide Web*, vol. 2, no. 3, pp. 115-32, Baltzer, 1999.

[13] A. Natsev, R. Rastogi, K. Shim, "WALRUS: A similarity retrieval algorithm for image databases," *Proc. SIGMOD*, Philadelphia, PA, 1999.

[14] W. Niblack, R. Barber, W. Equitz, M. Flickner, E. Glasman, D. Petkovic, P. Yanker, C. Faloutsos, G. Taubin, "The QBIC project: querying images by content using color, texture, and shape," *Proc. SPIE*, vol. 1908, pp. 173-87, San Jose, February, 1993.

[15] A. Pentland, R. W. Picard, S. Sclaroff, "Photobook: tools for content-based manipulation of image databases," *Proc. SPIE*, vol. 2185, pp. 34-47, San Jose, February 7-8, 1994.

[16] R. W. Picard, T. Kabir, "Finding similar patterns in large image databases," *Proc. IEEE ICASSP*, Minneapolis, vol. V, pp. 161-64, 1993.

[17] Y. Rubner, L. J. Guibas, C. Tomasi, "The earth mover's distance, Shimulti-dimensional scaling, and color-based image retrieval," *Proc. ARPA Image Understanding Workshop*, pp. 661-668, New Orleans, LA, May 1997.

[18] G. Salton, M. J. McGill, *Introduction to Modern Information Retrieval*, McGraw-Hill, NY, 1983.

[19] J. Shi, J. Malik, "Normalized cuts and image segmentation," *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 731-7, San Juan, Puerto Rico, June, 1997.

[20] J. R. Smith, S.-F. Chang, "An image and video search engine for the World-Wide Web," *Proc. SPIE*, vol. 3022, pp. 84-95, 1997.

[21] J. R. Smith, C. S. Li, "Image classification and querying using composite region templates," *Journal of CVIU*, vol. 75, no. 1-2, pp. 165-174, Academic Press, 1999.

[22] S. Stevens, M. Christel, H. Wactlar, "Informedia: improving access to digital video," *Interactions*, vol. 1, no. 4, pp. 67-71, 1994.

[23] J. Z. Wang, J. Li, R. M. Gray, G. Wiederhold, "Unsupervised multiresolution segmentation for images with low depth of field," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 23, no. 1, 85-91, 2001. A short version in ICIAP'99.

[24] J. Z. Wang, J. Li, G. Wiederhold, "SIMPLIcity: Semantics-sensitive Integrated Matching for Picture LIbraries," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 23, 2001.

[25] J. Z. Wang, G. Wiederhold, O. Firschein, X. W. Sha, "Content-based image indexing and searching using Daubechies' wavelets," *International Journal of Digital Libraries*, vol. 1, no. 4, pp. 311-328, 1998.