Image Indexing Using Weighted Color Co-occurrence Matrix and Feature Selection

Dong Liang[1, 2], Edmund Y. Lam[2]
[1](Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, 200030, China)
[2](Department of Electrical and Electronic Engineering, The University of Hong Kong, Pokfulam Road, Hong Kong)

***Abstract*- In this paper, image indexing based on Weighted Color Co-occurrence Matrix (WCCM) feature and Isolation Parameter-based feature selection is introduced. In this method, Isolation Parameter (IP) is used to indicate the visual perception complexity and conduct feature selection for each query image. When indexing images from database in the reduced feature space, the similarities of diagonal elements and non-diagonal elements of CCM feature are weighted separately with different values based on the Isolation Parameters of query image and images from database. The experiments show that the proposed method provides better results than Modified Color Co-occurrence Matrix (MCCM) based method and Sub-range Cumulative Histogram (SCH) based method.**

I. INTRODUCTION

Color Co-occurrence Matrix (CCM) [1, 2] is a kind of commonly used color feature representation in image retrieval. In simplified CCM, the homogeneous color region of the image contributes to the diagonal elements of CCM and non-homogeneous region (color edge) to the non-diagonal elements. When indexing an image, the similarity between the query image and images in the database are computed regarding diagonal elements and non-diagonal elements as a whole. But indexing image with CCM feature will ignore the shape information since the number of diagonal elements is far greater than that of non-diagonal elements. Modified Color Co-occurrence Matrix (MCCM) [3] was proposed to overcome this problem, where the similarities of diagonal elements and non-diagonal elements of CCM are measured respectively with equal weights. However, weighting equally the similarities of homogeneous regions and non-homogeneous regions is not a good choice. For example, if a query image and an image in the database all consist of few homogeneous regions, the similarity of homogeneous regions should play a more important role. When they all consist of many small regions, the similarity of non-homogeneous regions should play a more important role.

On the other hand, given the dimension of low-level features is usually high (the dimension of CCM is 256), feature selection is a very important step in image retrieval and largely affects retrieval accuracy.

In this paper, image indexing using Weighted Color Co-occurrence Matrix (WCCM) with feature selection is proposed. A subset of CCM features is selected based on Isolation Parameter of each query image, and then the similarities of diagonal elements and non-diagonal elements of CCM are assigned with different weights based on the visual complexity of the query and images in the database.

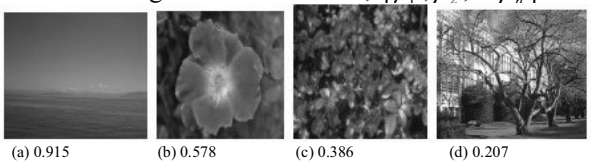
II. FEATURE SELECTION BASED ON ISOLATION PARAMETR

In image retrieval, feature selection can not only reduce the cost of retrieval by reducing the number of features, but in some cases it can also provide better performance [4-5]. For CCM, it is usually sparse and thus sensitive to noise. Therefore, a feature selection method is proposed in this paper. The idea is to consider the contribution of each bin of CCM feature of the query image in image retrieval. The easiest way is to calculate the fitness value of each bin using Bayesian decision function liked and then select a set of features that performs the best. In this method, the database is classified into two categories based on the Isolation Parameters of images.

Isolation Parameter $p_k \in (0,1)$ is introduced in [6] to demonstrate the visual complexity of image content. It is small when image consists of many small regions, and big when image consists of a few homogeneous regions. Fig.1 shows the Isolation Parameters of different images, from left to right, the image becomes more intricate, and the Isolation Parameter becomes smaller. From this figure, we can see that Isolation Parameter is in correspondence with the complexity of human visual perception. In this paper, threshold $p_T = 0.5$ is defined based on the statistics from the database. If $p_k \geq p_T$, we consider that image mainly consists of few of homogeneous regions, and if $p_k < p_T$, the image mainly consists of many small regions.

The Isolation Parameter-based feature selection method is described below:

- For each query image *Q*, the *n* dimensional low-level feature $F^Q = \{f_1^Q, f_2^Q, \cdots f_n^Q\}$ is extracted and Isolation Parameter p_k^Q is computed.
- Count the frequency of each bin of the feature vector in query image *Q*. $\{C_1^Q, C_2^Q, \cdots C_n^Q\}$.
- Define a subset *S* of image dataset *I*, in which the Isolation Parameter p_k^S of each image belongs to the same region with p_k^Q. For example, $p_k^Q, p_k^S \geq p_T$ or $p_k^Q, p_k^S < p_T$. Then count the frequency of each bin of feature vector in the subset, $\{\alpha_1^S, \alpha_2^S, \cdots \alpha_n^S\}$.
- Count the frequency of each bin of feature vector in all the images in the database *I*, $\{\beta_1^I, \beta_2^I, \cdots \beta_n^I\}$.



(a) 0.915 (b) 0.578 (c) 0.386 (d) 0.207

Figure 1. Isolation parameters of four images.

1-4244-0549-1/06/$20.00 ©2006 IEEE.

- Compute fitness value of each bin of feature vector, $V^Q = \{V_1^Q, V_2^Q, \cdots V_n^Q\}$,

$$V_i^Q = C_i^Q \times \log\left(\beta_i^I \left(1-\alpha_i^S\right) / \alpha_i^S \left(1-\beta_i^I\right)\right) \quad (1)$$

- Sort the fitness value sequence with descending order $SV^Q = sort\left(V^Q\right)$. Then we can define threshold T to indicate how much the selected feature contribute to the whole feature. Suppose h features are selected when

$$\sum_{j=1}^{h} SV_j^Q \Big/ \sum_{t=1}^{n} SV_t^Q \geq T \quad (2)$$

Record the positions of h features in V^Q, then for images in database I, the same positions of features of are selected to form a new feature space. h dimensional features of V^Q are used as the query features to index the images.

III. Weighted CCM Feature

Let M^Q and M^I be co-occurrence matrices of the query image Q and image in the database I. In matching stage, For CCM feature, the similarity between the query image Q and image from the database I is given by:

$$S^{CCM}(Q,I) = \sum_{(i,j)} \min\left\{M^Q(i,j), M^I(i,j)\right\} \Big/ \sum_{(i,j)} M^Q(i,j) \quad (3)$$

We can see from (3) that if CCM is used as a whole feature vector to measure similarity without modification, important shape information will be overwhelmed by color information.

The similarity of modified CCM feature[3] is given by:

$$S^{MCCM}(Q,I) = 0.5S_1(Q,I) + 0.5S_2(Q,I) \quad (4)$$

where $S_1(Q,I)$ and $S_2(Q,I)$ are the similarity of diagonal and non-diagonal elements respectively. We can see that for MCCM feature, the similarities of diagonal elements and non-diagonal elements are measured separately but are assigned the same weights.

Here we propose weighted CCM feature, where different weights are assigned for the similarity of diagonal feature and non-diagonal feature based on Isolation Parameter of images.

$$S^{WCCM} = w_1 S_1(Q,I) + w_2 S_2(Q,I) \quad (5)$$

There are three instances given different Isolation Parameters:

- If $p_k^Q \geq p_T$ and $p_k^I \geq p_T$, we strengthen the similarity of homogeneous region:

$$w_1 = 2 - \left|p_k^I - p_k^Q\right|, w_2 = 1 \quad (6)$$

We can see that the closer between p_k^Q and p_k^I, the bigger of w_1, and then the bigger of $w_1 S_1(Q,I)$, thus $S^{WCCM}(Q,I) > S^{MCCM}(Q,I)$, which means that image in the database I becomes more relevant to the query Q on WCCM feature than on MCCM feature.

- If $p_k^Q < p_T$ and $p_k^I < p_T$, we strengthen the similarity of non-homogeneous region:

$$w_1 = 1,\ w_2 = 2 - \left|p_k^I - p_k^Q\right| \quad (7)$$

In the same way, $S^{WCCM}(Q,I) > S^{MCCM}(Q,I)$, which also means that image in the database I becomes more relevant to the query Q on WCCM feature than on MCCM feature.

- If $p_k^Q \geq p_T$ and $p_k^I < p_T$ or $p_k^Q < p_T$ and $p_k^I \geq p_T$, we weaken the similarities of both homogeneous region and non-homogeneous region:

$$w_1 = w_2 = 1 / \left(1 + \left|p_k^I - p_k^D\right|\right) \quad (8)$$

w_1 and w_2 are less than 1, $S^{WCCM}(Q,I) < S^{MCCM}(Q,I)$, image in the database I becomes less relevant to the query Q on WCCM feature than on MCCM feature when they have different content complexity.

From the analysis above, we can see that the weighting is like a non-linear mapping. After weighting, the images that have similar color and visual complexity become more relevant to the query image and those with different color and visual complexity become less relevant to the query image.

IV. Experimental Results

The image database used consists of 2103 images with 60 semantic categories. In this paper, the HSV color space is used and color (hue) is quantized to 16 colors because 16 bins are sufficient for proper color invariant object retrieval empirically [7]. When indexing image, we randomly select images from each category as the queries. Retrieval accuracy [8] and average-retrieval-rank [3] are used as the performance criteria.

The baselines for comparison are WCCM without feature selection, MCCM and Sub-range Cumulative Histogram (SCH) feature [9] that is superior to the cumulative histogram and Color Moments. Fig.2 shows average-retrieval-rank of WCCM with and without feature selection, MCCM and SCH. For WCCM with feature selection, T takes 99%, 98%, 97%, 96%, 95%, 90%, 85%, 80%, 75%, 70%, 65%, and 60%. From fig.4, for the criteria of average-retrieval-rank, WCCM with feature selection performs better than other features when $85\% \leq T \leq 99\%$.

Fig.3 shows retrieval accuracy using different features when 11 images are shown to the user. T takes the same value as in fig.2. Retrieval accuracy of WCCM with feature selection is almost identical but little worse than those without feature selection when $95\% \leq T \leq 99\%$ and better than other features. When 35 images are shown to the user, we can see from fig.4 that retrieval accuracy of WCCM with feature selection is better than those without feature selection when $96\% \leq T \leq 99\%$. This trend is consistent with that presented by the criteria of Average-retrieval-rank.

Fig.5 gives the average number of features selected and fig.6 gives the average retrieval time when T takes different values. We can see that when $T = 99\%$, the average number of features selected is about 85.8, whereas the number of features is 256 without feature selection. When the average

retrieval time without feature selection is assigned 1, the average retrieval time is about 0.5 when $T = 99\%$. Two figures indicate that feature selection reduces much redundancy for image indexing.

V. CONCLUSION

In this paper, image indexing based on WCCM with feature selection is introduced. In the proposed method, Isolation Parameter is used to indicate the visual perception complexity of image and direct feature selection for query image. When indexing images from database, the similarity of diagonal elements and non-diagonal elements is assigned different weights based on the Isolation Parameters of query image and images from database. The experiments show the superiority of proposed feature in comparison to MCCM and SCH features.

However, there are problems in color quantization. In future work, CIEL*u*v* color space will be considered to find representative colors whose number equals to the required number of bins.

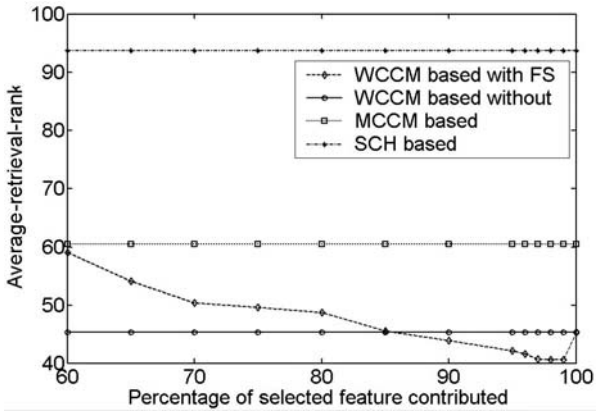


Figure 2. Average-retrieval-rank of four features.

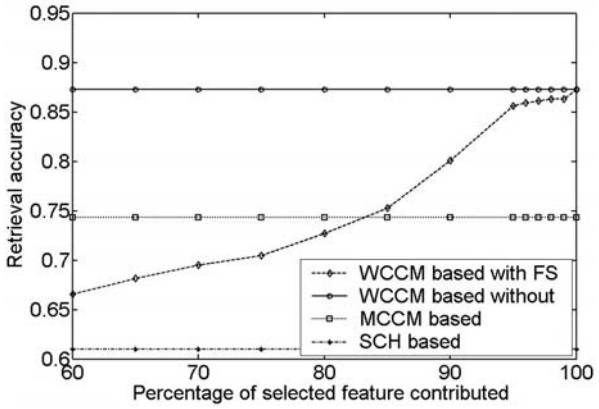


Figure 3. Retrieval accuracy of four features. (No. of images shown=11).

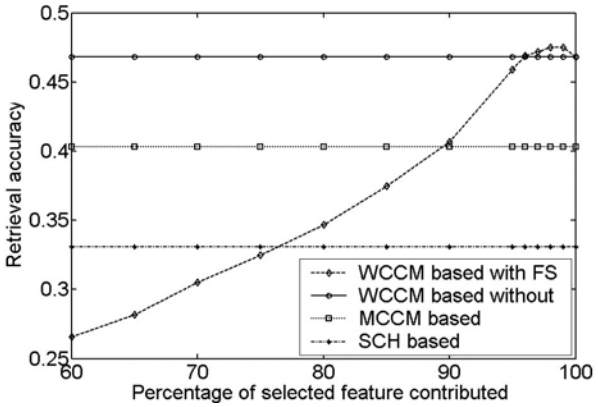


Figure 4. Retrieval accuracy of four features. (No. of images shown=35).

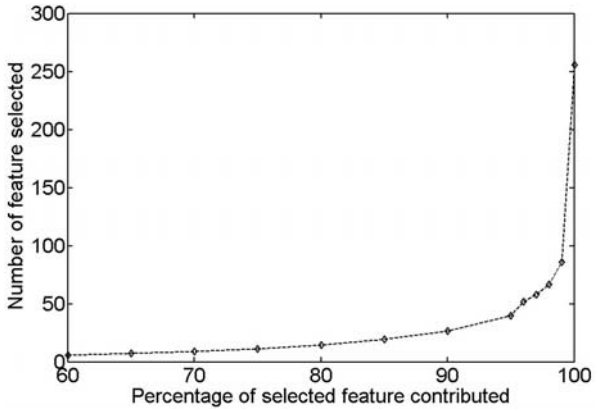


Figure 5. The average number of features selected.

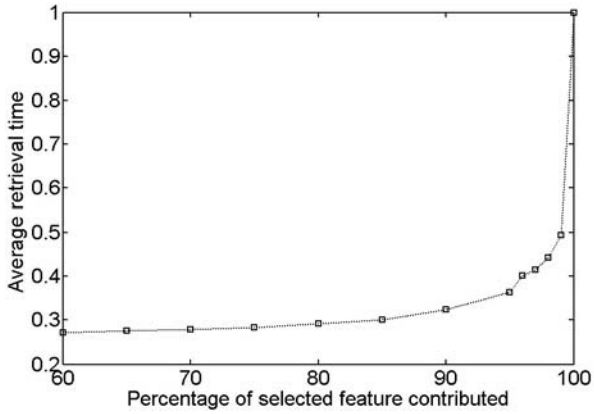


Figure 6. The average retrieval time of features selected.

ACKNOWLEDGMENT

The work described in this paper was partially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project Number HKU 7143/05E).

REFERENCES

[1] V. Kovalev, S. Volmer, "Color Co-occurrence Descriptor for Querying-by-Example", *Multimedia Modeling*, pp.32-38. 1998.

[2] G P Qiu, "Color Image Indexing using BTC", *IEEE Transactions on Image Processing*, 12(1), pp.93-101, 2003.

[3] Seong-O Shim, Tae-Sun Choi, "Image Indexing by Modified Color Co-occurrence Matrix", *IEEE International Conference on Image Processing*, vol. 3, pp.493-496, 2003.

[4] Jennifer G. Dy, Carla E. Brodley, Avi Kak, Lynn S. Broderick, Alex M.Aisen, "Unsupervised Feature Selection Applied to Content-based Retrieval of Lung Images", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(3), pp.373-378, 2003.

[5] Micheline Najjar, Christophe Ambroise, Jean-Pierre Cocquerez, "Feature Selection for Semi Supervised Learning Applied to Image Retrieval", *IEEE International Conference on Image Processing*, vol. 2, pp.559-562, 2003.

[6] Y. Luo, Y.J. Zhang, Y.Y.Gao, "Meaningful Regions Extraction based on Image Analysis", *Chinese Journal of Computer*, 23(12), pp.1313-1319, 2000.

[7] Gevers, T. Smeulders, A.W.M, "PicToSeek: Combining Color and Shape Invariant Features for Image Retrieval" *IEEE Transactions on Image Processing*, 9(1), pp.102-119, 2000.

[8] Xiaofei He, Oliver King, Wei-Ying Ma, Mingjing Li, Hong-Jiang Zhang, "Learning a Semantic Space from User's Relevance Feedback for Image Retrieval", *IEEE Trans. On Circuits and Systems for Video Technology*, 13(1), pp.39-48, 2003.

[9] Y. J. Zhang, Z. W. Liu, Y. He, "Color-based Image Retrieval using Sub-range Cumulative Histogram", *High Technology Letters*, 4(2), pp.71-75, 1998.