PERFORMANCE OF A COMPACT FEATURE VECTOR IN CONTENT-BASED IMAGE RETRIEVAL

Gita Das
Clayton School of Information Technology, Monash University, Victoria 3800, Australia

Sid Ray
Clayton School of Information Technology, Monash University, Victoria 3800, Australia

Keywords: CBIR, feature representation, sample size, dimensionality.

Abstract: In this paper, we considered image retrieval as a dichotomous classification problem and studied the effect of sample size and dimensionality on the retrieval accuracy.

Finite sample size has always been a problem in Content-Based Image Retrieval (CBIR) system and it is more severe when feature dimension is high. Here, we have discussed feature vectors having different dimensions and their performance with real and synthetic data, with varying sample sizes. We reported experimental results and analysis with two different image databases of size 1000, each with 10 semantic categories.

1 INTRODUCTION

Content-Based Image Retrieval (CBIR) where images are being represented by their visual features has been an active research topic in recent years. The application domain of CBIR encompasses a wide range including medical, defence and security surveillance systems. The selection of features e.g. colour, shape, colour-layout etc. and their proper representation e.g. colour histogram, statistical moments etc. are very important for good system retrieval. The concept of co-occurrence matrix in texture has been known for long (Haralick et al., 1973), however, its use in colour has been reported very recently (Huang, 1998), (Shim and Choi, 2003), (Ojala et al., 2001). A Colour Co-occurrence Matrix (CCM) represents how the spatial correlation of colour changes with distance i.e. pixel positions. In (Shim and Choi, 2003), a Modified Colour Co-occurrence Matrix is used where a CCM of Hue is simplified to represent the number of colour pairs between adjacent pixels (4-neighbourhood). They did not consider the adverse effect of ignoring Saturation and Value components of colour. The diagonal elements in a CCM convey the colour information of the entire image whereas the non-diagonal elements represent the shape information in an indirect way (Das and Ray, 2005), (Shim and Choi, 2003). We described (Das and Ray, 2005) a compact feature representation based on the elements of CCMs in HSV (Hue, Value, Saturation) space. The feature vector consists of all diagonal elements and one representative value for all non-diagonal elements of the CCM. Although the addition of features contributes to better retrieval, it brings up the problem of Curse of Dimensionality (Hughes, 1968), (Duda et al., 2001). Hence, dimension reduction has become a critical issue in feature representation and image indexing of CBIR systems (Wu et al., 2000). In (Das and Ray, 2005), we tried to reduce dimension without compromising the retrieval accuracy. Experimental results reveal that diagonal elements of CCMs are much more in number (about 80%) compared to the non-diagonal elements (about 20%). This is in line with that reported in (Shim and Choi, 2003). As the diagonal elements in the feature vector are the majority, manipulating them in any way may contribute to loss of information content of images significantly. Also, it is worth noting that most of the non-diagonal elements are zero. Thus representing all the non-diagonal elements with a single Sum-Average (Haralick et al., 1973) value (for details, see section 3) attributes to several benefits: i) the Sum-Average of non-diagonal elements would be less sensitive to noise and thus enhance retrieval performance, ii) the dimension is reduced significantly, thus reducing online computation and retrieval time, iii) compared to other methods of dimension reduction e.g. Principal Component Analysis (PCA)(Wu et al., 2000), com-

Das G. and Ray S. (2007).
PERFORMANCE OF A COMPACT FEATURE VECTOR IN CONTENT-BASED IMAGE RETRIEVAL.
In *Proceedings of the Second International Conference on Computer Vision Theory and Applications - IU/MTSV*, pages 241-246

puting Sum-Average is very simple and easy.

Thus, for HSV=[16,3,3] the feature dimension is 148 in original dimension and 25 in reduced dimension. For rest of the paper, we refer the original feature space as 148-D and the compact one as 25-D. With reduced dimension, we obtained improved performance and faster retrieval.

In the past people have tried PCA (Principal Component Analysis) (Sinha and Kangarloo, 2002), (Martinez and Kak, 2001), (Swets and Weng, 1996), a useful statistical technique that finds the most significant features that describe a data set. PCA is suitable for CBIR where we have basically a two-class (Relevant and Non-relevant) classification problem, and the training sample size is usually small (Martinez and Kak, 2001). To demonstrate the goodness of 25-D feature vector we used the first 25 eigen vectors (or principal components) using PCA from original 148-D and the feature vector thus derived will be called PCA25-D.

The rest of the paper is organized as follows. Section 2 gives an overview of our work. Section 3 provides a description of feature vectors, similarity measure and evaluation methodology. Section 4 details our experimental setup and result analysis while section 5 gives the conclusions and future work proposals.

2 OVERVIEW OF OUR WORK

In this paper, we studied mainly two issues:

1. *Behaviour of feature vectors with real data and synthetic data*: In this paper we have discussed three feature vectors with an emphasis on 25-D and 148-D and their behaviour with real data and synthetic data. In real domain, a number of parameters are involved and it is difficult to isolate each one's contribution to the ultimate retrieval accuracy. Whereas, with synthetic data we can have more control over data distribution. We explained that in 25-D compact feature vector the correlation among the feature components are much less compared to 148-D and the assumption of feature independence is maintained.
2. *Behaviour of feature vectors at varying relevant class sizes*: In reality, an image database comprises a number of semantic categories, each category having a different number of samples. Precision (a performance evaluation parameter which will be discussed in Section 3.3) for a query image belonging to a category having more number of samples may be higher compared to the one belonging to a category having less number of samples. So, given a feature vector describing the images in a database, it is important to know the relation of precision to the sample size (i.e. the number of samples) of relevant category. In (Huijsmans and Sebe, 2005), Huijsmans and Sebe presented some results keeping sample size of Relevant class constant while varying that of Non-relevant class. They reported results on accuracy based on one query category only. In our study, we varied the sample size for each semantic category at a time, measured precision for the category and then averaged results of all categories to obtain precision for the whole data set. This way we get a detailed and more representative picture of system performance.

3 METHODOLOGY

For rest of the paper, we used the following nomenclature:

N: Number of images in the database

C: Number of semantic categories in the database

Q, I: Query image and Database image respectively

M: Number of components in the feature vector i.e. feature dimension

L: Quantization levels in H,S,V matrices

N_r: Scope i.e. the number of top retrieved images returned to the user

3.1 Feature Representation and Indexing

Let P be the $L \times L$ co-occurrence matrix whose element p_{xy} indicates the number of times a pixel with colour level x occurs, at a distance d, relative to pixels with colour level y. The Sum-Average as described in (Haralick et al., 1973) has been modified in (Das and Ray, 2005) and is as follows:

$$Sum\_ndiag = \sum_{x=1}^{L-1} \sum_{y=x+1}^{L} (x+y) p_{xy} \tag{1}$$

where *Sum_ndiag* are Sum-Average of non-diagonal elements of P. We chose HSV colour model as it is known to be perceptually uniform. We tried to make the spatial correlation more sensitive to Hue and less sensitive to Value and Saturation. We experimented with different levels of quantization and found HSV=[16,3,3] to be a good choice. This finding is in line with (Ojala et al., 2001). We chose co-occurrence distance d=3 and used pixel pairs in both vertical and horizontal directions. Thus we obtained symmetric matrices and needed only upper diagonal elements to consider. For a 16x16 matrix, the number of diagonal elements is 16 and the the number of non-diagonal elements is 120. For a 3x3 matrix, this number is 3 for both diagonal and non-diagonal elements. In our method, we represented all non-diagonal elements by a single value. Thus, for HSV=[16,3,3] the feature dimension is 148 in original space and 25 in reduced space.

As different feature components have different range (or values), we normalized them so that they lie within [0,1] and each component contributes equally in the similarity measure.

In PCA, the first principal component gives the direction along which the variance of data is maximum, the second principal component is the direction of maximum variance of data which is orthogonal to the first principal component, and so on. We constructed PCAD25-D using the first 25 components (they contribute to almost 100% variation in data). This also makes performance comparison with 25-D feature vector on the same platform.

To find the similarity between I and Q, we used Minkowski distance measure, a commonly used metric in CBIR,

$$D(I,Q) = \sum_{i=1}^{M} |f_{iI} - f_{iQ}| \tag{2}$$

where, f_i is the i^{th} normalized feature component. This metric is computationally simple and produces fairly good results.

3.2 Behaviour of Feature Vectors with Real Data and Synthetic Data

In the similarity measure we assume that the features are independent of each other. This, in reality, can be a pretty strong assumption. So, even if each individual feature has discriminative power, together they may not work as expected because of inter-dependence.

We experimented with synthetic data to have more control over data distribution. We used the mean and standard deviation of each category from real data set to randomly generate 100 points for each category. Here, Gaussian distribution was used. To keep things simple, we assumed the features to be uncorrelated. The covariance of two statistically independent variables is always zero. However, the reverse is not always true. For the special case of Gaussian distribution, zero covariance does imply independence. Thus we expect to have better result with synthetic data as compared to real data. Let X be the dataset consisting of N vectors, each being M-dimensional.

$$X = [\mathbf{x}(1), \mathbf{x}(2), ...\mathbf{x}(k), ..\mathbf{x}(N)]^T, \tag{3}$$

where,

$$\mathbf{x}(k) = [x_1(k), ..x_M(k)]$$

The covariance matrix obtained from the dataset X gives a measure of how strongly its components are related. The diagonal elements of the covariance matrix indicates the variance of feature components whereas the non-diagonal elements represent the covariance between the components. Let us denote $\mathbf{R}$ to be the correlation matrix. Given any pair of components, x_i and x_j, we denote their correlation as

$$r_{ij} = \frac{cov(x_i, x_j)}{s_i s_j} \tag{4}$$

where s_i and s_j are the standard deviations of x_i and x_j respectively.

By construction, a correlation is always a number between -1 and 1. Correlation inherits the symmetry property of covariance. To understand the feature dependence better and to explain our results we have introduced the following parameter α:

$$\alpha = \frac{\sum_{i=1}^{M} \sum_{j=i+1}^{M} |r_{ij}|}{M(M-1)/2} \tag{5}$$

In eqn (5), a high value of α indicates that the feature correlation is high.

To find the statistical significance of correlation coefficient we used t-test given by the following formula (Spiegel, 1998):

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} \tag{6}$$

where r is the correlation coefficient between two variables and N is the number of samples. The probability of the t-test indicates whether the observed correlation coefficient occurred by chance, if the true correlation is zero. To state in another way, t-test is a measure to find whether the correlation between two variables is significantly different from zero.

In our case, we have multiple feature components and in eqn (6), we have replaced r by α. This allows us to test the statistical significance of the average correlation value. Note that this is only an approximation of the t-test that is applicable to a pair-wise correlation coefficient test.

3.3 Impact of Sample Size on Accuracy

To study the effect of sample size, we varied the relevant class size (*R*) while keeping the non-relevant class size (*NR*) constant. We used a random subset of the original class size in order to avoid any bias in choosing images. We used precision and recall (Das and Ray, 2005), two widely used evaluation parameters in CBIR field, as a measure of system performance. We calculated precision of a category by averaging precision of all the images in the category used as query image. The final precision for any sample size is obtained by averaging results from all semantic categories in the database. This way we can have the most appropriate representation of the system performance.

4 EXPERIMENTAL STUDY

We experimented with two databases having the same number of semantic categories. All images are of 256×256 pixels size. An image in the retrieved list is considered to be relevant if that image comes from the same category as the query image, otherwise, non-relevant. While changing the *R* value we used a random subset of the original set and averaged the precision over 3 random subsets. This way we can minimize bias in precision, if any, due to selection of images. For all experiments we used a scope value of 20, a value that is not too high for user's point of view and not too low to have any relevant image retrieved.

4.1 Image Database and Ground Truth

1. DB1: This consists of 1000 images from 10 semantic categories (Flower, Leaf, Face, Fish, Dam, Car, Aeroplane, Leopard, Ship and Wristwatch). Each category contains 100 images. We chose 500 images randomly for training and the rest 500 images for testing. For *R*=50, each category contains 50 samples (or images). For *R*=40, 40 samples are taken randomly from 50 samples for each category at a time whereas all other categories are kept intact (i.e *NR*=450). Hence, for *R*=40 sample size, the total number of images in the database is 490.
2. DB2: This consists of 1000 images from 10 semantic categories (Africa, Beach, Dinosaurs, Elephants, Roses, Horses, Mountains, Food and Historical buildings). Each category contains 100 samples. This WANG database is a subset of Corel database and is freely available for research at the Pensnsylvania State University website http://wang.ist.psu.edu/. We picked up 500 images randomly for training and the rest 500 for testing. We changed *R* = 50, 40, 30, 20, 10 while *NR* = 450.

Table 1: Values for Correlation Matrix elements.

			Avg of diag	Avg of non-diag	α
DB1	**Real Data**	148-D	1.000	0.229	0.229
		25-D	1.000	0.158	0.158
	Synthetic Data	148-D	1.000	0.077	0.077
		25-D	1.000	0.081	0.081
DB2	**Real Data**	148-D	1.000	0.221	0.221
		25-D	1.000	0.157	0.159
	Synthetic Data	148-D	1.000	0.100	0.100
		25-D	1.000	0.100	0.100

4.2 Results Analysis

4.2.1 Behaviour of Feature Vectors with Real and Synthetic Data

Table 1 shows the co-variance values calculated for both real and synthetic data, for original dataset containing all 100 images per category, thus total number of images in the dataset is 1000. For real data, the average of non-diagonal elements is more with 148-D compared to 25-D. This is true irrespective of datasets. This is because 25-D has been constructed form 148-D by combining features in an intelligent way. The t-test for both 25-D and 148-D with real data showed statistical significance with 99% confidence. For synthetic data, average of non-diagonal elements is more or less the same for both 148-D and 25-D. This is true for both datasets. The t-test shows a confidence level of 98%, though ideally the α value for synthetic data should be very close to zero. The reason behind this is the way we constructed the synthetic dataset. The way we generated synthetic data it is only assured that there is no feature correlation within a semantic class, however, feature correlation is possible over entire dataset.

Please note that for real data, α for PCA25-D is zero for both data sets. This is because in the principal components are orthogonal and hence, they have zero correlation.

4.2.2 Behaviour of Feature Vectors at Varying Relevant Class Sizes

In previous section we have shown performance of the feature vectors for *R* = 50. However, as *R* changes, precision will change. In this section we will demonstrate the performance of the feature vectors for varying *R*. For all three vectors, precision with synthetic data is better than with real data. This is true for all values of *R* and for both datasets. However, for 25-D

and PCA25-D the curves for real and synthetic follow the variation very closely.

From Figures 3(a) and (b), it is evident that with real data 25-D performs better than PCA25-D and 148-D for all values of relevant class size. The inferior performance of PCA25-D can be attributed to the lack of consideration of the within-class and between-class variation of data in PCA. As *R* increases from 10 to 20, precision with 25-D increases by 15.5% whereas precision with 148-D increases by only 10%. When *R* changes from 40 to 50, the improvement in precision is 5.23% with 25-D but 4.37% with 148-D. This shows the strength of our feature vector against the variation in sample size and especially, at small sample size. In context to CBIR, using a higher scope means we looking for a larger neighbourhood. This means precision will fall and recall will increase. In Figures 4(a) and (b), for DB2, we shown the performance of 25-D and 148-D from a different evaluation angle consisting of both precision and recall. In Figure 4(a), for *R*=50, at recall = 100%, precision is 18.7% for 25-D whereas it is 12.87% for 148-D. This means a scope of 267 and 390 respectively. For *R*=10, at recall 100%, 25-D shows a precision of 9.78% and 148-D shows 5.2%. This means a scope of 102 for 25-D and 192 for 148-D. These values clearly indicate that for lower value of sample size, 25-D performs even better compared to 148-D.

5 CONCLUSIONS AND FUTURE DIRECTIONS

1. The online computation with 25-D is much less compared to 148-D. This reduction in number of computations will be very significant in today's real situation where image database size is already very big and is increasing day by day.
2. Irrespective of data sets and feature dimension, synthetic data always performs better than real data. This is expected as in synthetic data we did not consider any feature correlation.
3. For 25-D and PCA25-D, with real data set, the variation of precision with *R* values follows that of synthetic data pretty closely, unlike with 148-D. This means feature re-weighting method where we assume the features are independent of each other is more suitable for 25-D compared to 148-D. We introduced a new parameter, α, to explain the feature correlation.
4. Irrespective of data set being real or synthetic, for all feature vectors precision is more sensitive for smaller *R* values compared to higher *R* values.
5. For both DB1 and DB2, with real data, for varying relevant class size 25-D performs the best.

   We find that small sample issue is one of the major bottlenecks in CBIR research. In the future, we plan to investigate the small sample issue in more details. Also, the experiments will be extended to larger data sets.

REFERENCES

Das, G. and Ray, S. (2005). A compact feature representation and image indexing in content-based image retrieval. In *Proceedings of Image and Vision Computing New Zealand (IVCNZ 2005)*, pages 387–391, Dunedin, New Zealand.

Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern Classification, 2nd ed.*

Haralick, R. M., Shanmugam, K., and Dinstein, I. (1973). Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-3, No.6:610–621.

Huang, J. (1998). Color-spatial image indexing and applications. *PhD Dissertation, Cornell University*.

Hughes, G. F. (January 1968). On the mean accuracy of statistical pattern recognizers. *IEEE Transactions on Information Theory*, IT-14(1):55–63.

Huijsmans, D. P. and Sebe, N. (February 2005). How to complete performance graphs in content-based image retrieval: Add generality and normalize scope. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 27(2).

Martinez, A. M. and Kak, A. C. (February 2001). PCA versus LDA. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2).

Ojala, T., Rautiainen, M., Matinmikko, E., and Aittola, M. (2001). Semantic image retrieval with hsv correlograms. In *Proc. 12th Scandinavian Conference on Image Analysis*, pages 621–627, Bergen, Norway.

Shim, S. and Choi, T. (2003). Image indexing by modified color co-occurrence matrix. In *International Conference on Image Processing*.

Sinha, U. and Kangarloo, H. (2002). Principal component analysis for content-based image retrieval. *RadioGraphics*, (22 (5)):1271–1289.

Spiegel, M. R. (1998). Schaum's outline series theory and problems of statistics. *McGraw-Hill, 2nd edition, 1998*.

Swets, D. and Weng, J. (1996). Using discriminant eigenfeatures for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(8):831–836.

Wu, P., Manjunath, B., and Shin, H. (2000). Dimensionality reduction for image retrieval. In *Proceeding IEEE International Conference on Image Processing (ICIP 2000)*, pages 726–729, Vol. 3, Vancouver, Canada.