TELKOMNIKA Telecommunication, Computing, Electronics and Control
Vol. 18, No. 4, August 2020, pp. 1892~1896
ISSN: 1693-6930, accredited First Grade by Kemenristekdikti, Decree No: 21/E/KPT/2018
DOI: 10.12928/TELKOMNIKA.v18i4.11176

A principal component analysis-based feature dimensionality reduction scheme for content-based image retrieval system

Oluwole A. Adegbola[1], Ismail A. Adeyemo[2], Folasade A. Semire[3], Segun I. Popoola[4], Aderemi A. Atayero[5]
[1,2,3]Department of Electronic and Electrical Engineering, Ladoke Akintola University of Technology, Nigeria
[4,5]Department of Electrical and Information Engineering, Covenant University, Nigeria

Article Info

Article history:

Received Sep 13, 2018
Revised Mar 23, 2020
Accepted Apr 3, 2020

Keywords:

Content-based image retrieval system
Feature dimensionality reduction
Low-level visual feature
Principal component analysis

ABSTRACT

In content-based image retrieval (CBIR) system, one approach of image representation is to employ combination of low-level visual features cascaded together into a flat vector. While this presents more descriptive information, it however poses serious challenges in terms of high dimensionality and high computational cost of feature extraction algorithms to deployment of CBIR on platforms (devices) with limited computational and storage resources. Hence, in this work a feature dimensionality reduction technique based on principal component analysis (PCA) is implemented. Each image in a database is indexed using 174-dimensional feature vector comprising of 54-dimensional colour moments (CM54), 32-bin HSV-histogram (HIST32), 48-dimensional gabor wavelet (GW48) and 40-dimensional wavelet moments (MW40). The PCA scheme was incorporated into a CBIR system that utilized the entire feature vector space. The k-largest eigenvalues that yielded a not more than 5% degradation in mean precision were retained for dimensionality reduction. Three image databases (DB10, DB20 and DB100) were used for testing. The result obtained showed that with 80% reduction in feature dimensions, tolerable loss of 3.45, 4.39 and 7.40% in mean precision value were achieved on DB10, DB20 and DB100.

Corresponding Author:

Segun I. Popoola,
Department of Electrical and Information Engineering,
Covenant University,
P.M.B. 1023, Ota, Nigeria.
Email: segun.popoola@covenantuniversity.edu.ng

1. INTRODUCTION

One of the challenges of relevance feedback (RF) in image retrieval is the inherent 'curse of dimensionality' occasioned by small sample size with high feature dimension. Therefore, for RF techniques which are based on training classifier using feedback examples, the curse of dimensionality can deteriorate the classifier performance, thereby leading to poor retrieval results. To mitigate this problem, a technique that relies on the properties of the feedback examples for selecting a lower dimensional feature, that will serve as good representative for classification can be employed. In this way, a significant dimensionality reduction can be achieved by removing irrelevant or redundant features, thus leading to a significant decrease in training time and memory complexities, and better classifier performance [1, 2]. Approaches for feature dimensionality reduction have been grouped into two [3]: (a) those that involves linear or nonlinear mapping from the original feature space to a new one of lower dimensionality. Notable among these are linear discriminant analysis [4]

and principal component analysis [1, 5-7]; (b) those that directly reduce the number of the original features by selecting a subset of them that still retains sufficient information for classification. In general, approaches in this category can be grouped into two namely: filter methods and wrapper methods [8].

The filter methods are generally not classifier dependent as they acquire no feedback from the classifiers, but depend on indirect assessments like distance measure to estimate classification performance on the other hand, the wrapper methods are classifiers dependent and are known to yield better classification performance [8, 9]. Many features selection methods for classification have been proposed in the literature, [10] with many experimental results in favour of the wrapper methods [8, 11, 12]. However, in spite of good classification performance, the wrapper methods have limited application due to high computational complexity, especially when applied to support vector machine (SVM) classifiers.

PCA is a dimensionality reduction technique that transforms the original set of features into a smaller subset that account for as much of the total variation in the data as possible [13]. It is widely used in the area of pattern recognition, computer vision and signal processing [7]. Several optimality properties of PCA have been identified namely: variance of extracted features is maximized; the extracted features are uncorrelated; finds best linear approximation in the mean-square sense and maximizes information contained in the extracted feature [14].

These properties of PCA have attracted research on PCA-based variable selection methods [7, 13-18] and has been applied to relevance feedback in both document and image retrieval systems [1, 5, 6]. In [1], a novel PCA-based feature dimensionality reduction scheme (or approach) was proposed for the RF framework with a view to capturing the subjective class implied in the positive examples. Similarly, the works of Cox, et al, [19] and Vasconcelos & Lippman [20], employed Bayesian learning to integrate user's feedback for updating image probability distribution and subsequently re-rank images in the database.

It was reported that the scheme (or approach) reduced the average retrieval time and significantly reduced storage space utilization. However, the precision measure in top 20 retrieval results in four feedback iterations was 45%. This may be due to the failure of Bayesian classifiers to use the few available image samples gathered over the feedback iterations to estimate the class probability distribution. It was stated by Yin, Bhanu, Chang and Dong [21] that one of the shortcomings of the Bayesian approach is that it requires more feedback iterations to gather more samples, which is not always available in real time retrieval systems, to effectively estimate the probability distribution of the image samples.

In other to address the computational complexity issue, a SVM-based technique, termed filtered and supported sequential forward search, was proposed feature selection [3]. The technique integrates the filter and wrapper parts into one scheme by leveraging on their unique strengths. Results of experimental on both synthetic and real data showed effectiveness of the method regarding classification accuracy. However, given the fact that much smaller data, compared to what obtains in CBIR system, was used to evaluate the system, an average run-time of 16.23 seconds was recorded Such a lengthy run time is not acceptable for CBIR system with RF framework.

2. MATERIALS AND METHODS

2.1. Feature extraction

Feature extraction is one very crucial task in CBIR application, and it is the core of any such system [4]. The extraction of suitable features from the images influences to a great extent the choice of the indexing structure and the query processing unit. In view of this, various methods of feature extraction to extract various types of visual contents from the images have been developed and are being improved upon overtime [22, 23]. Three generic domain image databases (DB10, DB20 and DB100) were employed with each image database indexed using two colour models (CM54 and HIST32) and two texture models (GW54 and WM40). Adegbola, Aborisade, Popoola and Atayero [24] presents detailed description of various image database and feature extraction models.

2.2. Feature selection model

In a generic system, it is extremely difficult to know the particular feature model(s) to be used to uniquely identify certain groups of images. Therefore, a combination of several image feature models is usually employed with the assumption that at least one will have the ability to capture the unique identity of the targeted images. This approach poses several challenges. First, because the image features are cascaded as a flat vector, such arrangement may increase the chances of *diluting* the feature component that uniquely identifies the targeted image group. This may also lead to what is known as *curse of dimensionality* in CBIR system that employs machine learning techniques for relevance feedback. Cost of feature extraction algorithm is another issue which may become prohibitive as the number of feature descriptors increases. In view of this, including too many features is obviously not feasible for application involving human-machine interaction. Since such system is expected to be fast enough for smooth

interaction, the selection of most appropriate features to relduce computational burden becomes imperative and to achieve this, a procedure that uses Principal Component Analysis is employed in this work.

Assume a binary classification problem, given a set of label training data $\{(\boldsymbol{X}_i, y_i)\ i = 1, N \mid y_i \neq 0\}$ where sample $\boldsymbol{X}_i \in \mathbb{R}^d$ and $y_i \in \{-1, 1\}$. Let

$$F = \{f_1, f_2, \ldots, f_d\} \quad (1)$$

be the set of all features under examination, and let

$$S = \{(\boldsymbol{X}_i, y_i) \mid i = 1, 2, \ldots, N\} = \left\{\left[x_i^1, x_i^2, \ldots, x_i^d\right]^T, y_i \mid i = 1, 2, \ldots, N\right\} \quad (2)$$

denote the training set containing *N* training pairs, where x_i^d is the numerical value of feature f_d for the *i*th training sample. The goal of dimensionality reduction is to find a minimal set of features $F_s = \{f_{s1}, f_{s2}, \ldots, f_{sk}\}$ to represent the input vector *X* in a lower dimensional space as

$$X_s = \{x_{s1}, x_{s2}, \ldots, x_{sk}\} \quad (3)$$

where $k < d$, while the classification obtained in the low-dimensional space still yields the desired accuracy.

2.3. Principal component analysis

PCA is a statistical priocedure for high dimensionality reduction of feature space. It uses orthogonal transformation to decorrelate a set of correlated feature space to enhance variance by emphazising the directions of principal variation of dataset [25]. Consider a set of d-dimensional vectors $\{\boldsymbol{x} = [x_1, \ldots, x_d]^T\}$ with distribution centred at the origin, $E(x) = 0$. The covariance is obtained using (4)

$$r_{ij} = E\{(x_i - \boldsymbol{x_i})(x_j - \boldsymbol{x_j})\} = E\{x_i x_j\}, \quad (4)$$

where E is the expectation operator. The parameters r_{ij} can be arranged to form the $d \times d$ covariance matrix

$$R_x = E\{(x - \boldsymbol{x})(x - \boldsymbol{x})^T\} = E\{xx^T\} \quad (5)$$

Assuming $det(R_x) \neq 0$, then by applying eigenvector decomposition, R_x can be decomposed into the product of three matrices:

$$R_x = W\Lambda W^{-1} \quad (6)$$

where, $\Lambda = \mathrm{diag}\{\lambda_1, \ldots, \lambda_\mathrm{d}\}$ is the Eigenvalue matrix. $W = [\mathrm{w}_1, \ldots, \mathrm{w}_\mathrm{d}]^\mathrm{T}$ forms a set of orthonormal basis vectors called Eigenvectors.

For dimensionality reduction, only the set of orthonormal bases vectors resulting from the *k*-largest Eigenvalues are retained. This will result into significant feature dimensionality reduction. Normally, the *k*-largest Eigenvalues that constitutes 95% of the total Eigenvalues are retained for dimensionality reduction. However, this work employed precision/recall graph to determine the dimension of feature to be retained. This is a more objective choice, since the resulting lower dimensional feature vectors are used for distance (similarity) measurement in image retrieval system with relevance feedback. Consequently, the number of feature dimension retained is based on a 5% maximum loss constraint imposed on the precision/recall graph.

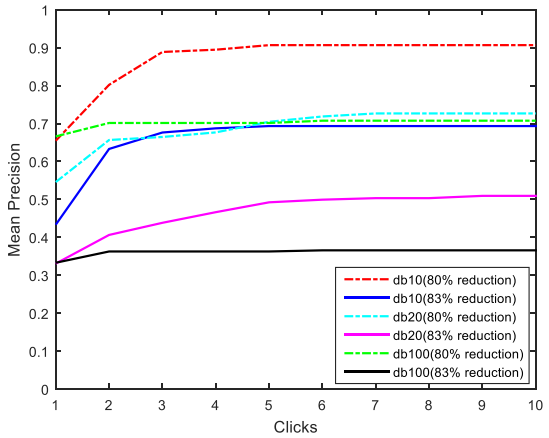
3. RESULTS AND ANALYSIS

Combination of visual descriptors results to increase in the dimension of the resulting feature vector. Normally, the resulting feature model, which is the concatenation of individual feature vectors, could have very high dimensions and thus increase the latency of RF scheme even on a medium-size image database. Hence, in order to mitigate the curse of dimensionality problem associated with machine learning based RF scheme, reducing the dimensions of feature vectors may be necessary. In this study, principal component analysis (PCA) is integrated to the developed OC-SVM RF for the purpose of feature vector dimensionality reduction.

A criterion of 5% maximum degradation in mean precision value was used to determine the dimension of feature vector to keep. The effect of feature vector dimensionality reduction is shown in Figure 1. The maximum mean precision values obtained on DB10, DB20 and DB100 were 0.9067, 0.7266 and 0.7275 respectively, for 80% reduction in feature vector dimension. While a reduction of feature

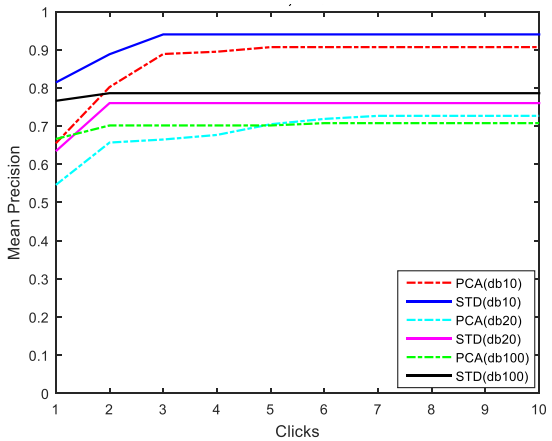
dimension by 83% for DB10, DB20 and DB100 resulted into mean precision values of 0.6933, 0.5093 and 0.3657 respectively.

Figure 2 shows the comparison between the OC-SVM RF that used the whole 174-dimensional feature (STD) and the OC-SVM RF with PCA that used 35-dimensional features (PCA). The maximum mean precision values of 0.9400, 0.7600 and 0.7860 were achieved on the DB10, DB20 and DB100 respectively for the STD. The maximum mean precision achieved with PCA on the DB10, DB20 and DB100 were 0.9067, 0.7266 and 0.7275 respectively. Thus an 80% reduction in feature dimension, yielded tolerable degradation of 3.54%, 4.39% and 7.4% in maximum mean precision performance on DB10, DB20 and DB100 respectively.



Database	DB10		DB20		DB100	
% reduction	**80%**	83%	**80%**	83%	**80%**	83%
Mean Precision	**0.9067**	0.6933	**0.7266**	0.5093	**0.7275**	0.3657

Figure 1. Mean precision result of the OC-SVM RF with PCA of different dimensionality reduction



Database	DB10		DB20		DB100	
% reduction	**STD**	PCA	**STD**	PCA	**STD**	PCA
Mean Precision	**0.9400**	0.9067	**0.7600**	0.7266	**0.7860**	0.7275

Figure 2. Mean precision result of the OC-SVM relevance feedback with PCA utilizing 80% dimensionality reduction

4. CONCLUSION

In CBIR system designed for generic image databases, it is general practice to represent images using combination of several different image features with a view to capturing extra information that may improve retrieval accuracy. This usually results in high dimensionality of visual feature vectors for CBIR system with classifier-based relevance feedback scheme. In this paper, the issue of curse of dimensionality is addressed using a PCA-based feature selection approach. The feature selection model was incorporated

into an exisisting OC-SVM RF retrieval system. The findings revealed that by allowing a 5% loss tolerance in mean precison, it was possible to achieve 80% reduction in feature vector dimensionality, while attempt to increase the percentage reduction of feature vector dimension resulted into poor retrieval results.

References

[1] Z. Su, S. Li, and H. Zhang, "Extraction of feature subspaces for content-based retrieval using relevance feedback," *Proceedings of the ninth ACM international conference on Multimedia*, pp. 98-106, 2001.

[2] A. Marakakis, N. Galatsanos, A. Likas, and A. Stafylopatis, "Relevance feedback for content-based image retrieval using support vector machines and feature selection," *Artificial Neural Networks – ICANN 2009*, Springer, vol. 5768, pp. 942-951, 2009.

[3] Y. Liu and Y. F. Zheng, "FS_SFS: A novel feature selection method for support vector machines," *Pattern Recognition*, vol. 39, no. 7, pp. 1333-1345, 2006.

[4] K. P. Chung, "Intelligent content-based image retrieval framework based on semi-automated learning and historic profiles," Thesis, Murdoch University, 2007.

[5] Y. Lu, I. Cohen, X. S. Zhou, and Q. Tian, "Feature selection using principal feature analysis," *Proceedings of the 15th ACM international conference on Multimedia*, pp. 301-304, 2007.

[6] R. Tavoli and F. Mahmoudi, "PCA-based relevance feedback in document image retrieval," *IJCSI International Journal of Computer Science Issues*, vol. 9, no. 2, pp. 497-502, July 2012.

[7] M. Suganthy and P. Ramamoorthy, "Principal component analysis based feature extraction, morphological edge detection and localization for fast iris recognition," *Journal of Computer science*, vol. 8, no. 9, pp. 1428-1433, 2012.

[8] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial intelligence*, vol. 97, no. 1-2, pp. 273-324, December 1997.

[9] F. Alonso Atienza, J. L. Rojo Álvarez, A. Rosado Muñoz, J. J. Vinagre, A. García Alberola, and G. Camps Valls, "Feature selection using support vector machines and bootstrap methods for ventricular fibrillation detection," *Expert Systems with Applications*, vol. 39, no. 2, pp. 1956-1967, 2012.

[10] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine learning*, vol. 46, pp. 389-422, 2002.

[11] B.H. Juang and S. Katagiri, "Discriminative learning for minimum error classification (pattern recognition)," *IEEE Transactions on signal processing*, vol. 40, no. 12, pp. 3043-3054, 1992.

[12] H. Watanabe, T. Yamaguchi, and S. Katagiri, "Discriminative metric design for robust pattern recognition," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2655-2662, 1997.

[13] I. Jolliffe, "Principal component analysis," *International encyclopedia of statistical science*, *Springer*, pp. 1094-1096, 2011

[14] A. Tsymbal, S. Puuronen, M. Pechenizkiy, M. Baumgarten, and D. W. Patterson, "Eigenvector-Based Feature Extraction for Classification," *FLAIRS-02 Proceedings*, pp. 354-358, 2002.

[15] W. J. Krzanowski, "Selection of variables to preserve multivariate data structure, using principal components," *Applied Statistics*, vol. 36, no. 1, pp. 22-33, 1987.

[16] W. J. Krzanowski, "A stopping rule for structure-preserving variable selection," *Statistics and Computing*, vol. 6, pp. 51-56, 1996.

[17] M. Mansor, S. Yaacob, M. Hariharan, S. Basah, S. A. Jamil, M. M. Khidir, et al., "Fuzzy k-NN and k-NN Algorithm for fast Infant Cues Detection," *World Congress on Medical Physics and Biomedical Engineering May 26-31, 2012, Beijing, China*, 2013, pp. 1260-1263.

[18] F. Song, Z. Guo, and D. Mei, "Feature selection using principal component analysis," *2010 International Conference on System Science, Engineering Design and Manufacturing Informatization*, pp. 27-30, 2010.

[19] I. J. Cox, M. L. Miller, T. P. Minka, T. V. Papathomas, and P. N. Yianilos, "The Bayesian image retrieval system, PicHunter: theory, implementation, and psychophysical experiments," *IEEE transactions on image processing*, vol. 9, no. 1, pp. 20-37, 2000.

[20] N. Vasconcelos and A. Lippman, "Learning from user feedback in image retrieval systems," *Advances in neural information processing systems*, pp. 977-986, 2000.

[21] P. Y. Yin, B. Bhanu, K. C. Chang, and A. Dong, "Integrating relevance feedback techniques for image retrieval using reinforcement learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 27, no. 10, pp. 1536-1551, 2005.

[22] Y. Rui, T. S. Huang, and S.-F. Chang, "Image retrieval: Current techniques, promising directions, and open issues," *Journal of visual communication and image representation*, vol. 10, no. 1, pp. 39-62, 1999.

[23] F. Long, H. Zhang, and D. D. Feng, "Fundamentals of content-based image retrieval," Multimedia Information Retrieval and Management, *Springer*, pp. 1-26, 2003.

[24] O. A. Adegbola, D. O. Aborisade, S. I. Popoola, and A. A. Atayero, "Performance Evaluation of Visual Descriptors for Image Indexing in Content Based Image Retrieval Systems," *ICCSA 2018: Computational Science and Its Applications – ICCSA 2018*, pp. 539-549, 2018.

[25] K. I. Diamantaras and S. Y. Kung, Principal component neural networks: theory and applications: *John Wiley & Sons*, Inc., 1996.