Application of SPCA Algorithm in Image Dimensionality Reduction

Xian.Wei Wu
Information Engineering College
Ningbo Dahongying College
Ningbo, People's Republic of China
wxw786@qq.com

Wen.Yang Yu, Yu.Bin Yang
Information Engineering College
Ningbo Dahongying College
Ningbo, People's Republic of China
seayuweya@163.com

Abstract—In the page, We discuss several dimensionality reduction methods for image feature, and then focus on the one: SPCA(Simple Primary Component Analysis), which is simple fast and exceeding algorithm of data-oriented PCA algorithm. In order to better understand the SPCA algorithm, Some well-designed experiments of image compression and image retrieval are taken to compare these algorithms. By experiment 1, we get the result: PCA matrix algorithm is best in performance but worst in speed, and GHA is better in speed ,but worst in performance, and the results show that SPCA is out-standing not only in performance, but also in speed. By experiment 2, we get the desired result: using the image feature after SPCA almost get the same performance of original image feature, but much better than original image feature in speed. The conclusion is: SPCA algorithm can be applied in many field, especially in image compression and image retrieval.

Keywords-Dimensionality Reduction;SPCA;PCA;GHA;

Image Feature

I. INTRODUCTION

With the problem more complex, high-dimensional data processing is becoming increasingly important. On the one hand, high-dimensional data make it difficult to understand the relationship between the data, on the other hand, makes high-dimensional data storage, transmission, treatment becomes more difficult. High dimensional data has become a bottleneck of problem solving. Therefore, dimensionality reduction techniques[1] is a key to achieve an effective dimensionality reduction purposes. PCA (Principal Component Analysis) method is a very effective dimensionality reduction method. The PCA methods include PCA matrix, SPCA, GHA and other algorithms, In this paper we do experiments to compare the performance of these algorithms in high-dimensional image feature dimension reduction compression which highlights the SPCA algorithm is simple, fast, meanwhile high performance. This paper is organized as follows: First, give an overview of the PCA then, introduce a variety of specific algorithms, and do experiment with a variety of algorithms comparison Finally, make out the experimental results.

II. PCA(PRINCIPAL COMPONENT ANALYSIS) [2]

PCA is a non-parametric method of ,pattern recognition. By using a small number of features describe the samples in order to reducing the dimension of feature space. It is the goal to represent high dimensional data in a low dimensional subspace with the minimum mean square error and finally reach the purpose of reducing dimensions.

. Specifically, assume that there is an n-dimensional vector X, want to down to become k-dimensional data ($k \leq n$) If we simply cut off X, brought the sum of the mean square error variance equal to lay down the various components. This requires the existence of a reversible linear transformation T, such that the TX truncated sense of mean square error at the best, it is clear that some of the transformed component has a lower variance.

For the dimensions $m \times n$ input matrix X, it can be expressed as:

$$X = t_1 p_1^T + t_2 p_2^T + \cdots + t_k p_k^T + E \quad (1)$$

. Where $k \leq n$, t_i called the principal component scores, p_i called the principal component loadings, E for the remainder of the X. The formula is equivalent to:

$$X = TP^T + E \tag{2}$$

T expressed approximately as

$$T = XP \tag{3}$$

Can be shown when the i-th eigenvalue p_i is taken as the covariance matrix of the matrix X corresponding to the descending order of the feature vector time, T is the maximum variance. For each pair t_i, p_i, is rearrange in accordance with the value corresponding to the characteristic feature vectors in descending order, the first pair t_1, p_1, get the first primary element vector and principal component factor of the largest amount of information, and the rest, and so on.

For primary componet variable $T = (t_1, t_2, ..., t_k)^T$, t_k is a linear function of X; t_k make the maximum variance t_k fully reflect the changes in X; while each of t_k, tries not to correlate that is extracted principal component variables contain no duplicate information.

From the viewpoint of statistical pattern recognition, the main component analysis is a dimension reduction process in fact, it ignores the linear component has a smaller variance, the larger the error term has retained, thus reducing the number of valid data is represented by the dimension. However, it is a linear algorithm, the only extract data related to the linear feature.

Depending on the method of calculation of feature vectors, PCA methods are generally divided into two categories: one is the matrix method, and the other is the data-oriented methods. Matrix method is through the matrix calculations to accurately solve the eigenvalues and eigenvectors, resulting principal component vector. While the data-oriented methods use numerical computation rather than matrix operations to obtain approximate principal component vectors, the biggest advantage is that when the dimension size increases, the computational efficiency can solve the bottleneck problem of matrix operations encountered.

III. DIMENSIONALITY REDUCTION METHODS

A. PCA matrix algorithm [3]

① to obtain a set of raw data t_k, where t_k is the i-th point, the point itself is n-dimensional, i = 1,2 ..., m

② Center for each point: first calculate of the average $\overline{X} = \frac{\sum_{i=1}^{m} X_i}{m}$, then adjust each point $X_i - \overline{X}$

③ calculate the covariance matrix C, where $C_{ij} = (X_i - \overline{X})(X_j - \overline{X})$

④ to obtain the eigenvalues λ and eigenvectors α of the covariance matrix by the C to give the corresponding eigenvalues of the eigenvectors descending according to the arrangement, as long as the choice is generally in front of the k ($k \leq n$) can be approximated representation of the original feature vector data .

GHA and SPCA algorithms both are PCA methods

B. GHA (Generalized Hebb algorithm) [4]

The algorithm is the use of unsupervised neural network The input is n neurons x_i, the output is L neurons y_j, synaptic weight value, w_{ji} where n is the number of dimensions, L is the number of principal components selected, i = 1,2 ..., n, j = 1,2, ..., L.

Specific algorithm is as follows:

① At time t = 1, the synaptic weights of the network w_{ji} is initialized, it takes a small random number, assigned to a learning factor η small positive number.

② For t = 1, j = 1,2, ..., L and i = 1,2 ..., n, is calculated

$$y_j(t) = \sum_{i=1}^{n} w_{ji}(t) x_i(t) \tag{4}$$

$$\Delta w_{ji}(t) = \eta \left[y_j(t) x_i(t) - y_j(t) \sum_{k=1}^{j} w_{ki}(t) y_k(t) \right] \tag{5}$$

Where $x_i(t)$ is the i-th component, of input vector $X(t)$, L is the main component of the desired number.

③t increased 1 (t = t + 1) then go to ②, and continue until it reaches stable.

For larger t, j synapses value converges to the covariance matrix of the input vector of the j-th eigenvalue eigenvector corresponding i-th component.

C. SPCA（Simple Principal Component Analysis）[5]

SPCA is a data-oriented method and like Hebbian learning [6], the algorithm does not explicitly calculate nor diagonalize the covariance matrix. Also SPCA does not require the tuning of learning parameters and convergence is obtained with very few iterations.

The SPCA algorithm[5] is given as:

①collect of an dimensional data set,V={v1,v2,…,vm}

②X={ x1,x2,…,xm },which are obtained by subtracting which is the average value of form the center of gravity of set of the vectors, is used as input vectors.

③The column vector is defined as connection weights between the inputs and output. The first weight is used to approximate the first eigenvector. The output function has a value given by:

$$y_1 = \left(e_1\right)^T x_i \tag{6}$$

④using the following equation (7)and (8),repetitive calculation of arbitrary vector suitably given as an initial value is carried out. As a result, the vector can approach the same direction as.

$$e_1^{k+1} = \frac{\sum_{i=0}^{m} \Phi\left(y_1, x_i\right)}{\left\| \sum_{i=0}^{m} \Phi\left(y_1, x_i\right) \right\|} \tag{7}$$

$$y_1 = \left(e_1^k\right)^T x_i \tag{8}$$

Where Φ is a threshold function given as

$$\Phi\left(y_1, x_i\right) = \begin{cases} x_i & y_1 \geq 0 \\ 0 & y_1 < 0 \end{cases} \tag{9}$$

⑤using following equation, we remove the first principal component vector from the dataset in order to find the next principal component.

$$x'_i = x_i - \left(e_1^T x_i\right) e_1 \tag{10}$$

We obtain principal components because we substitute with and with in equation (7),(8)and perform repetitive calculation once again.

If the same operation is being done enough for the dataset, we obtain the principal component which is stronger in contribution rate by turns.

IV. EXPERIMENTS

Experiment 1:

This experiment is using the Lena standard image of 256×256. Firstly, the image is divided into blocks (n1×n2), then each d1×d2 (d1=256/n1, d2=256/n2) pixels, then extract the I pixel point of each of the mapped as new (n1×n2) of the I point dimension (second block is itself is n1×n2 dimension, the first component of it from the original image of the first block of the I pixel, its second components from the original image in the I pixel, and so on), resulting in a set of original data (including i=1,2... , d1 × d2). With these data, we use the three dimension reduction algorithm, get the main element vector, is reconstructed and the original image using principal component vectors, and comparison with the original image distortion.

The experiments were carried out with three different blocks of the original image (8 × 8, 16 × 16, 32 × 32), and each block, with different number of principal components are compared experimentally recorded all the time efficiency of the algorithm and the peak signal to noise ratio (PSNR). At the same time, also SPCA and GHA iteration is divided two cases, specific data, please refer to the following table:

Note: the PCA algorithm over data refers to the matrix of the PCA algorithm, the number refers to the number of iterations, the learning factor used by GHA 0.0001. By using the time (s) and PSNR (Peak Signal to Noise Ratio) (DB) to measure the difference of these algorithms. The PSNR (DB) index is used to measure the similarity of image reconstruction and the original image. For the 8 bit binary image, PSNR definition[7] is as follows:

$$PSNR = 10 \times \lg_{10} \frac{255^2}{\frac{1}{MN} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} \left[I(i,j) - I'(i,j) \right]^2} \tag{11}$$

TABLE1 PERFORMANCE COMPARISON OF PCA, SPCA AND GHA

Algo.	B	Nu Of PC	Com ratio	Situation 1			Situation 2		
				Times	Time (s)	PSNR	Times	Time (s)	PSNR
PCA	8 × 8	4	1/16		0.1	20.5			
SPCA				10	0.1	19.9	100	1.0	20.0
GHA				10	0.3	17.0	100	3.1	18.6
PCA		8	1/8		0.1	22.6			
SPCA				10	0.2	22.1	100	1.9	22.1
GHA				10	0.6	15.4	100	6.0	18.7
PCA		16	1/4		0.1	25.5			
SPCA				10	0.4	25.2	100	3.78	25.2
GHA				10	1.2	12.7	100	11.9	18.5
PCA	16 × 16	16	1/16		8.2	27.0			
SPCA				10	0.4	26.5	100	3.7	26.5
GHA				10	1.2	11.6	100	11.9	15.8
PCA		32	1/8		8.2	30.3			
SPCA				10	0.9	30.0	100	7.4	30.0
GHA				10	2.4	8.7	100	24.1	13.7
PCA		64	1/4		8.7	36.1			
SPCA				10	1.7	35.4	100	14.7	35.4
GHA				10	4.7	6.1	100	47.1	11.5
PCA	32 × 32	64	1/16		710.2	177.9			
SPCA				10	1.8	103.4	100	15.3	103.0
GHA				10	5.9	6.7	100	60.8	7.2
PCA		128	1/8		712.4	177.6			
SPCA				10	3.5	21.67	100	30.4	21.7
GHA				10	25.1	4.1	100	257.5	5.0
PCA		256	1/4		715.8	177.4			
SPCA				10	7.0	13.8	100	60.0	13.5
GHA				10	49.7	0.9	100	506.7	2.6

The (I, J) represents the pixel coordinates, and the representation of the original image and the reconstruction of the original image of each pixel gray value respectively, M, N representing the number of columns.

Through the experiments we can see that when the block is 16×16, the compression ratio is 1/16 case, SPCA case of PSNR 26.5db, PCA PSNR 27db, both with similar performance, but the algorithm of time is not the same, PCA algorithm with a time of 8.2 seconds, while the SPCA is only 0.4 seconds. Under the same conditions of the GHA algorithm, PSNR 11.6db, a time of 1.2 seconds, the performance and the time is also not as SPCA algorithm. In the higher dimensional case, the speed of the SPCA algorithm is hundreds of times the speed of PCA algorithm. Other conditions of the experiment also shows that the SPCA algorithm is not only superior performance, almost no difference with PCA matrix) algorithm (visual effect is shown in Fig .1), more important is, when the original data dimension reach a larger scale, PCA algorithm of matrix is difficult to adapt to, even if the calculation, its speed also far less than the speed of SPCA algorithm. For example, in Table 1, block of 32×32, PCA computing time for more than 700 seconds, while the SPCA only a few seconds. For the GHA algorithm, alt-

hough theoretically, iteration through an infinite number of times of GHA, and decreased gradually learning factor, can the theory of infinite into the main element of the value, but from the above experimental results, in the same iteration, GHA algorithm of time more than SPCA algorithm, and the performance is far as SPCA. In conclusion, SPCA algorithm is the reduction of a dimension reduction algorithm of both performance and speed dimension algorithm, especially suitable for large dimension scale data, and real-time situations.

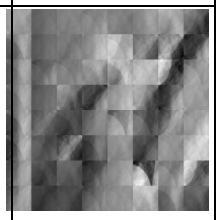
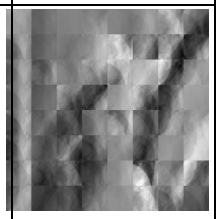
algorithm	8×8 block, 4 primary components	16×16 block, 16 primary components	32×32 block, 64 primary components
SPCA			
	PSNR=19.9	PSNR=26.5	PSNR=103.4
PCA			
	PSNR=20.5	PSNR=27.0	PSNR=172.9

Figure 1 Intuitionistic performance comparison of SPCA and PCA under the same compression ratio(1/16)

Experiment 2:

In this experiment, there is total 1361 color endoscopic images, including 2 classes: one is 169 images having cancer cell, the other is remain images having no cancel cell. Analyzing these images, we find obvious color differences of the two class images. Therefore, we take color feature[8] in our retrieval experiments. We choose 10 images randomly from the 169 images as query-example images for the following experiment:

From the literature [5] was informed that SPCA algorithm can obtain rapid and effective dimensionality reduction under the prerequisite of almost no loss of performance. In the experiment, using SPCA algorithm we compress 256-dimensional color histograms feature and color correlograms feature into 16 dimensions separately, so the dimension of the final integrated feature[9] become 16+16=32.

Through the curve of Fig .2 ,we find that after feature compressed no significant difference in retrieval performance, but feature dimensions from 512 drop to 32, thus effectively reduce the amount of calculation, and ultimately improve the speed of retrieval, the average of retrieval time reduce from 7.61S to 1.78S. This retrieval advantage is especially made in the large-scale database.

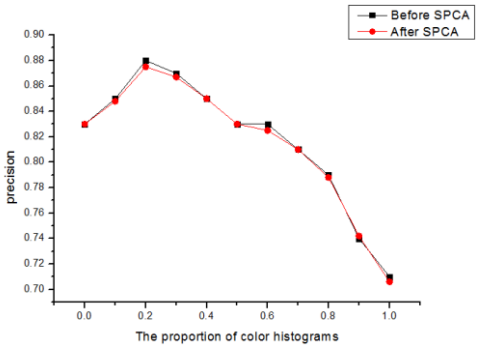


Figure 2 Performance curves of integrated features under the two methods share different circumstances before and after the SPCA Compression

Figure 3 Examples of retrieval results of integrated features after using SPCA

Fig .3 is the examples of retrieval results of integrated features after dimensionality reduction, in line with similar distance of images from small to large. The first is the query-sample. Integrated feature made from 20% color histograms feature and 80% color correlograms[10] feature. the 3th,9th,16th, 17th image are unrelated images and the rest are related.

Through this experiment, we can make the conclusions that using SPCA on high-dimensional feature can largely enhance the retrieval speed.

REFERENCES

[1] Li Zhuo,Bo Cheng and Jing Zhang, "A comparative study of dimensionality reduction methods for large-scale image retrieval" Neurocomputing, vol. 141, Oct. 2014, pp. 202-10, doi:10.1016/j.neucom.2014.03.014.

[2] Chunyu Chen; Keyu Xie, "Face Recognition Based on Two-dimensional Principal Component Analysis and Kernel Principal Component Analysis" Information Technology Journal, vol. 11, 2012, pp. 1781-5, doi:10.3923/itj.2012.1781.1785.

[3] Good, RP; Kost, D; Cherry, GA, "Introducing a Unified PCA Algorithm for Model Size Reduction" IEEE TRANSACTIONS ON SEMICONDUCTOR MANUFACTURING, vol. 23, MAY 2010, pp. 201-209, doi:10.1109/TSM.2010.2041263.

[4] Liao Ke; Pu Yifei; Zhou Jiliu, "A public adaptive watermark algorithm for color images based on principal component analysis of generalized Hebb" Proc. 2004 International Conference on Intelligent Mechatronics and Automation (IEEE Cat. No.04EX952), 2004, pp. 890-5

[5] Matthew Partridge and Rafael A. Calvo, "Fast Dimensionality Reduction and Simple PCA" Intelligent Data Analysis, vol. 2, Aug. 1998 , pp. 203-214

[6] Huyck, CR ; Mitchell, IG, "Post and pre-compensatory Hebbian learning for categorisation" COGNITIVE NEURODYNAMICS, vol. 8,AUG 2014 , pp.299-311, doi:10.1007/s11571-014-9282-4.

[7] Wang, SH ; Lin, T, "United coding method for compound image compression " MULTIMEDIA TOOLS AND APPLICATIONS , vol. 71,AUG 2014 , pp.1263-1282

[8] Li Guang-Li; Zhang Hong-bin, "Research on tumor image retrieval system based on relevancy discriminant by color feature " Computer Engineering and Design , vol. 33,Nov. 2012 , pp.4272-7

[9] Wang, XY ; Zhang, BB; Yang, HY, "Content-based image retrieval by integrating color and texture features " MULTIMEDIA TOOLS AND APPLICATIONS , vol. 68,FEB 2014 , pp.545-569, doi:10.1007/s11042-012-1055-7.

[10] Lei, JS, "Image Annotation Using Sub-block Energy of Color Correlograms" Proc. ARTIFICIAL INTELLIGENCE AND COMPUTATIONAL INTELLIGENCE, 2009, vol. 5855 pp. 555-562