ACM Multimedia, Augsburg, Germany, September 23-29, 2007.

Feature Selection Using Principal Feature Analysis

Yijuan Lu
University of Texas at San Antonio
One UTSA Circle
San Antonio, TX, 78249
lyijuan@cs.utsa.edu

Ira Cohen
Hewlett-Packard Labs
1501 Page Mill Road
Palo Alto, CA 94304
Ira.cohen@hp.com

Xiang Sean Zhou
Siemens Medical Solutions USA, Inc.
51 Valley Stream Parkway
Malvern, PA, 19355
xiang.zhou@siemens.com

Qi Tian
University of Texas at San Antonio
One UTSA Circle
San Antonio, TX, 78249
qitian@cs.utsa.edu

ABSTRACT

Dimensionality reduction of a feature set is a common preprocessing step used for pattern recognition and classification applications. Principal Component Analysis (PCA) is one of the popular methods used, and can be shown to be optimal using different optimality criteria. However, it has the disadvantage that measurements from all the original features are used in the projection to the lower dimensional space. This paper proposes a novel method for dimensionality reduction of a feature set by choosing a subset of the original features that contains most of the essential information, using the same criteria as PCA. We call this method Principal Feature Analysis (PFA). The proposed method is successfully applied for choosing the principal features in face tracking and content-based image retrieval (CBIR) problems. Automated annotation of digital pictures has been a highly challenging problem for computer scientists since the invention of computers. The capability of annotating pictures by computers can lead to breakthroughs in a wide range of applications including Web image search, online picture-sharing communities, and scientific experiments. In our work, by advancing statistical modeling and optimization techniques, we can train computers about hundreds of semantic concepts using example pictures from each concept. The ALIPR (Automatic Linguistic Indexing of Pictures - Real Time) system of fully automatic and high speed annotation for online pictures has been constructed. Thousands of pictures from an Internet photo-sharing site, unrelated to the source of those pictures used in the training process, have been tested. The experimental results show that a single computer processor can suggest annotation terms in real-time and with good accuracy.

Categories and Subject Descriptors

I.4.7 [**Feature Measurement**]

General Terms

Algorithms, Theory, Performance, Experimentation

Keywords

Feature Extraction, Feature Selection, Principal Component Analysis, Discriminant Analysis

1. INTRODUCTION

In many real world problems, reducing dimension is an essential step before any analysis of the data can be performed. The general criterion for reducing the dimension is the desire to preserve most of the relevant information of the original data according to some optimality criteria. In pattern recognition and general classification problems, methods such as Principal Component Analysis (PCA), Independent Component Analysis (ICA) and Fisher Linear Discriminate Analysis (LDA) have been extensively used. These methods find a mapping from the original feature space to a lower dimensional feature space.

In some applications it might be desired to pick a subset of the original features rather then find a mapping that uses all of the original features. The benefits of finding this subset of features could be in saving cost of computing unnecessary features, saving cost of sensors (in physical measurement systems), and in excluding noisy features while keeping their information using "clean" features (for example tracking points on a face using easy to track points- and inferring the other points based on those few measurements).

Variable selection procedures have been used in different settings. Among them, the regression area has been investigated extensively. In [1], a multi layer perceptron is used for variable selection. In [2], stepwise discriminant analysis for variable selection is used as inputs to a neural network that performs pattern recognition of circuitry faults. Other regression techniques for variable selection are described in [3]. In contrast to the regression methods, which lack unified optimality criteria, the optimality properties of PCA have attracted research on PCA based variable selection methods [4, 5, 6, 7]. As will be shown, these methods have the disadvantage of either being too computationally expensive, or choosing a subset of features with redundant information. This work proposes a computationally efficient method that exploits the structure of the principal components of a feature set to find a subset of the original feature vector. The chosen subset of features is shown empirically to maintain some of the optimal properties of PCA.

The rest of the paper is as follows. The existing PCA based feature selection methods are reviewed in Section 2. The proposed method, Principal Feature Analysis (PFA), is described in Section 3. We apply PFA to face tracking and content-based image retrieval problems in Section 4. Discussion will be given in Section 5.

2. BACKGROUND AND NOTATION

We consider a linear transformation of a random vector $X \in \Re^n$ with zero mean and covariance matrix Σ_x to a lower dimension random vector $Y \in \Re^q, q < n$

$$Y = A_q^T X \tag{1}$$

with $A_q^T A_q = I_q$ where I_q is the $q \times q$ identity matrix.

In PCA, A_q is a $n \times q$ matrix whose columns are the q orthonormal eigenvectors corresponding to the first q largest eigenvalues of the covariance matrix Σ_x. There are ten optimal properties for this choice of the linear transformation [4]. One important property is the maximization of the "spread" of the points in the lower dimensional space which means that the points in the transformed space are kept as far apart as possible, and therefore retaining the variation in the original space. Another important property is the minimization of the mean square error between the predicted data and the original data.

Now, suppose we want to choose a subset of the original variables/features of the random vector *X*. This can be viewed as a linear transformation of *X* using a transformation matrix

$$A_q = \begin{bmatrix} I_q \\ [\mathbf{0}]_{(n-q)\times q} \end{bmatrix} \tag{2}$$

or any matrix that is permutation of the rows of A_q. Without loss of generality, lets consider the transformation matrix A_q as given above and rewrite the corresponding covariance matrix of *X* as

$$\Sigma = \begin{bmatrix} \{\Sigma_{11}\}_{q\times q} & \{\Sigma_{12}\}_{q\times(n-q)} \\ \{\Sigma_{21}\}_{(n-q)\times q} & \{\Sigma_{22}\}_{(n-q)\times(n-q)} \end{bmatrix} \tag{3}$$

In [4] it is shown that it is not possible to satisfy all of the optimality properties of PCA using the same subset. Finding the subset which maximizes $|\Sigma_Y| = |\Sigma_{11}|$ is equivalent to maximization of the "spread" of the points in the lower dimensional space, thus retaining the variation of the original data.

Minimizing the mean square prediction error is equivalent to minimizing the trace of

$$\Sigma_{22|1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}. \tag{4}$$

This can be seen since the retained variability of a subset can be measured using

$$\textit{Retained Variability} = \left(1 - \frac{trace(\Sigma_{22|1})}{\sum_{i=1}^{n}\sigma_i^2}\right)\cdot 100\% \sum_{i=1}^{n}\sigma_i^2 \tag{5}$$

where σ_i is the standard deviation of the i'th feature.

Thus, to find an optimal subset of q features, one of the two quantities above is computed for all possible combinations of q features. This method is very appealing since it satisfies well-defined properties. Its drawback is in the complexity of finding the subset. It is not computationally feasible to find this subset for a large feature vector.

Another method, proposed in [5], uses the principal components as the basis for the feature selection. A high absolute value of the *i'th* coefficient of one of the principal components (PC) implies that the x_i element of *X* is very dominant in that axes/PC. By choosing the variables corresponding to the highest coefficients of each of the first q PC's, the same projection as that computed by PCA is approximated. This method is a very intuitive and computationally feasible method. However, because it considers each PC independently, variables with similar information content might be chosen.

The method proposed in [6] and [7] chooses a subset of size q by computing its PC projection to a smaller dimensional space, and minimizing a measure using the Procrustes analysis [8]. This method helps reduce the redundancy of information, but is computationally expensive because many combinations of subsets are explored.

In our method, we exploit the information that can be inferred by the PC coefficients to obtain the optimal subset of features. But unlike the method in [5], we use all of the PC's together to gain a better insight on the structure of our original features. So we can choose variables without redundancy of information.

3. PRINCIPAL FEATURE ANALYSIS

Let *X* be a zero mean *n*-dimensional random feature vector. Let Σ be the covariance matrix of *X*. Let *A* be a matrix whose columns are the orthonormal eigenvectors of the matrix Σ:

$$\Sigma = A\Lambda A^T \tag{6}$$

where $\Lambda = \begin{bmatrix} \lambda_1 & & & \\ & . & 0 & \\ & 0 & . & \\ & & & \lambda_n \end{bmatrix}$, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$

$\lambda_1, \lambda_2, \ldots \lambda_n$ are the eigenvalues of Σ and $A^T A = I_n$. Let A_q be the first q columns of *A* and let $V_1, V_2, \ldots V_n \in \Re^q$ be the rows of the matrix A_q.

Each vector V_i represents the projection of the *i'*th feature (variable) of the vector *X* to the lower dimensional space, that is, the q elements of V_i correspond to the weights of the *i'th* feature on each axis of the subspace. The key observation is that features that are highly correlated or have high mutual information will have similar absolute value weight vectors V_i (changing the sign has no statistical significance [5]). On the two extreme sides, two independent variables have maximally separated weight vectors; while two fully correlated variables have identical weight vectors (up to a change of sign). To find the best subset, we use the structure of the rows V_i to first find the subsets of features that are highly correlated and follow to choose one feature from each subset. The chosen features represent each group optimally in terms of high spread in the lower dimension, reconstruction and insensitivity to noise. The algorithm can be summarized in the following five steps:

Step 1 Compute the sample covariance matrix, or use the true covariance matrix if it is available. In some cases it is preferred to use the correlation matrix instead of the covariance matrix [5]. The correlation matrix is defined as the $n \times n$ matrix whose *i,j'th* entry is

$$\rho_{ij} = \frac{E[x_i x_j]}{E[x_i^2]E[x_j^2]} \tag{7}$$

This representation is preferred in cases where the features have very different variances from each other, and using the regular covariance form will cause the PCA to put very heavy weights on the features with the highest variances. See [5] for more details.

Step 2 Compute the Principal components and eigenvalues of the Covariance/Correlation matrix as defined in equation (6).

Step 3 Choose the subspace dimension q and construct the matrix A_q from A. This can be chosen by deciding how much of the variability of the data is desired to be retained. The retained variability can be computed using:

$$Variability\ Retained = \frac{\sum_{i=1}^{q} \lambda_i}{\sum_{i=1}^{n} \lambda_i} \cdot 100\% \qquad (8)$$

Step 4 Cluster the vectors $|V_1|, |V_2|, ..., |V_n| \in \Re^q$ to $p \geq q$ clusters using K-Means algorithm. The distance measure used for the K-Means algorithm is the Euclidean distance. The reason to choose p greater then q in some cases is if the same retained variability as the PCA is desired, a slightly higher number of features is needed (Usually 1-5 more are enough).

Step 5 For each cluster, find the corresponding vector V_i which is closest to the mean of the cluster. Choose the corresponding feature, x_i , as a principal feature. This step will yield the choice of p features. The reason for choosing the vector nearest to the mean is twofold. This feature can be thought of as the central feature of that cluster- the one most dominant in it, and which holds the least redundant information of features in other clusters. Thus it satisfies both of the properties we wanted to achieve- large "spread" in the lower dimensional space, and good representation of the original data.

For clarity it should be noted that the clustering is the representation of the features in the lower dimensional space, and not of the projection of the measurements to that space (as in [6]).

The complexity of the algorithm is of the order of performing PCA, because the K-Means algorithm is applied on just n q-dimensional vectors.

4. EXPERIMENTAL RESULTS

4.1 Facial Motion Feature Points Selection

The first experiment is selection of the important points that should be tracked on a human face in order to account for the non-rigid motion. This is a classic example of the need for feature selection because it can be very expensive and difficult to track many points on the face reliably.

We capture video sequences of the subjects with markers on their face. Tracking of these markers is done automatically using template matching, with manual error correction. Thus we have reliable tracking results for the entire video sequence (60 sec at 30 frames per second) of human facial motion performing several normal actions - smiling, frowning, acting surprised, and talking. The images in Figure 1 demonstrate some facial motions that appear in the video sequence. There are a total of 40 facial points that are being tracked.

For the principal feature analysis, the points are split into two groups, upper face (eyes and above) and lower face. Each point is represented by its horizontal and vertical direction, and therefore the actual number of features we have is 80. We apply PFA on the data, retaining 90% of the variability in the data. Figure 2 shows the results of the analysis. The chosen features are marked by arrows, which display the principal directions chosen for those feature points. It can be seen that the chosen features correspond to physical based models of the face, i.e., horizontal motion is chosen for the middle lip point, with more features chosen around the lips than other lower facial regions. Vertical motion features are chosen for the upper part of the face (with the inner eyebrows chosen to represent the horizontal motion which appeared in the original sequence). This implies that much of the lower-face region's motion can be inferred using mainly lip motion tracking (an easier task from the practical point of view). In the upper part of the face, points on the eyebrows are chosen, and in the vertical direction mostly, which is in agreement with the physical models. It can also be seen that less points are needed for tracking in the upper part of the face (7 principal motion points) then the lower part of the face (9 principal motion points) since there are fewer degrees of freedom in the motion of the upper part of the face.

Figure 1. Examples of Facial Expressions

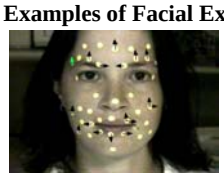

Figure 2. Principal Motion Features chosen for the facial points. The arrows show the motion direction chosen (Horizontal or Vertical)

The example shows that the PFA can model a difficult physical phenomenon such as the face motion to reduce complexity of existing algorithms by saving the necessity of measuring all of the features. This is in contrast to PCA, which will need the measurements of all original motion vectors to do the same modeling.

4.2 Feature Selection in Content-based Image Retrieval

The second experiment is designed to test the output of the principal feature analysis (PFA) algorithm for the purpose of content-based image retrieval. In a typical content-based image database retrieval application, the user has an image that he or she is interested in and wants to find similar images from the entire database. A two-step approach to search the image database is adopted. First, for each image in the database, a feature vector characterizing some image properties is computed and stored in a feature database. Second, given a query image, its feature vector is computed, compared to the feature vectors in the feature database, and images most similar to the query images are returned to the user. The features and the similarity measure used to compare two feature vectors should be efficient enough to

match similar images as well as being able to discriminate dissimilar ones.

Image similarity is mainly based on low-level features, such as color, texture and shape, although this might not always match what a user perceives as similar. We use 9 color and 20 wavelet moments (WM) as feature representation.

The test dataset from Corel consists of 17695 images. 400 of them are "airplanes", and 100 of them are "American eagles". Table 1 shows the comparisons using the original feature set, PCA, PFA and a feature set of same cardinality as PFA, but spanning a smaller number of clusters. The retrieval results are the average numbers of hits in top 10, 20, 40 and 80 returns using each of 100 airplanes and each of 100 eagles as the query image, respectively. PFA performs best for both airplanes and eagles among all methods. This result demonstrates that using too many features, which are 'noisy' in terms of the classification problem at hand, results in worse classification. In this case PCA does not help because it still uses all of the original features.

Table1. Performance Comparison using original feature set, PCA, PFA, and a 7 features subset spanning fewer clusters than PFA for Corel dataset.

Airplanes	**#Hit in top 10**	**Top 20**	**Top 40**	**Top 80**
10 WM's	3.32	5.75	9.94	17.1
7 PC's	3.37	5.71	9.90	16.8
7 PF's (1,5,6,7,8,9,10)	3.58	6.24	11.2	19.3
(1,2,3,4,8,9,10)	3.20	5.34	9.25	15.6
Eagles	**#Hit in top 10**	**Top 20**	**Top 40**	**Top 80**
10 WM's	1.98	2.82	4.43	6.58
7 PC's	1.96	2.76	4.13	6.27
7 PF's (1,5,6,7,8,9,10)	2.14	2.96	4.67	7.21
(1,2,3,4,8,9,10)	1.78	2.50	3.71	5.75

Table 2. Performance Comparison using the full features, PCA, PFA and randomly selected features.

40 random queries in Corel	**#Hit in top 10**	**Top 20**
22 PC's (7 color, 15 WM)	4.80	8.20
22 PF's (7 color, 15 WM)	4.90	8.15
22 random selected (I) (7 color, 15 WM)	3.88	6.65
22 random selected (II) (7 color, 15 WM)	4.18	6.85
22 random selected (III) (7 color, 15 WM)	4.40	7.18

Table 2 shows results on the 20 wavelet moments plus 9 color moments for the Corel database. 7 out 9 color features and 15 out of 20 wavelet features are selected by PFA. 40 randomly selected images are picked as the query image. The average number of hits in top 10, 20 returned images are calculated for all the query images. The performance of the original feature set and the principal feature set are tested against 3 randomly selected (but the same number of) features. It clearly shows that principal features yield comparable results as that of original set and PCA, and significantly higher results than any of random picks.

Because the number of features in these examples is relatively small, we were able to analyze the ranking of the selected subsets using the optimality criteria suggested in [4] and described in the Section 2. The selected subset was ranked on average in the top 5% of all possible subset selection. For example, for the 15 Wavelet features selected out of 20, the subset was ranked 20 out of 15504 possible combinations.

5. DISCUSSION

In this paper we propose a new method for feature selection named principal feature analysis. The method exploits the structure of the principal components of a set of features to choose the principal features, which retain most of the information, both in the sense of maximum variability of the features in the lower dimensional space and in the sense of minimizing the reconstruction error. The proposed method is applied to two applications, face tracking and content-based image retrieval. The results demonstrate that PFA does have comparable performance to PCA. However, for PCA, all of the original features are needed. This point is the main advantage of PFA over PCA: fewer sensors require or fewer features to compute and the selected features have their original physical meaning. When compared to the optimal features selected in [4], the results show that the PFA features are averagely ranked in the top 5% of all possible combinations.

6. REFERENCES

[1] Lisboa, P.J.G., Mehri-Dehnavi, R. Sensitivity Methods for Variable Selection Using the MLP. *International Workshop on Neural Networks for Identification, Control, Robotics and Signal/Image*, 1996, 330-338.

[2] Lin, T.-S., Meador, J. Statistical Feature Extraction and Selection for IC Test Pattern Analysis. *Circuits and systems*, vol 1., 1992, 391-394.

[3] Hocking, R.R. Development in Linear Regression Methodology: 1959-1982. *Technometrics*, vol. 25, 1983, 219-249.

[4] McCabe, G.P. Principal Variables. *Technometrics*, vol. 26, 1984, 127-134.

[5] Jolliffe, I.T. *Principal Component Analysis*. Springer-Verlag, New-York, 1986.

[6] Krzanowski, W.J. Selection of Variables to Preserve Multivariate Data Structure, Using Principal Component Analysis. *Applied Statistics- Journal of the Royal Statistical Society Series C*, vol.36, 1987, 22-33.

[7] Krzanowski, W.J. A Stopping Rule for structure- Preserving Variable Selection. *Statistics and Computing*, March vol. 6, 1996, 51-56.

[8] Gower, J.C., Statistical Methods of Comparing Different Multivariate Analyses of the Same Data, *Mathematics in the Archaeological and Historical Sciences*), University Press, Edinburgh, 1971, 138-149.