Application of Principal Components Analysis and Gaussian Mixture Models to Printer Identification

Gazi N. Ali, Aravind K. Mikkilineni, Edward J. Delp, and Jan P. Allebach
School of Electrical and Computer Engineering, Purdue University,
West Lafayette, Indiana, USA
Pei-Ju Chiang and George T. Chiu
School of Mechanical Engineering, Purdue University, West Lafayette, Indiana, USA

Abstract

Printer identification based on a printed document has many desirable forensic applications. In the electrophotographic process (EP) quasiperiodic banding artifacts can be used as an effective intrinsic signature. However, in text only document analysis, the absence of large midtone areas makes it difficult to capture suitable signals for banding detection. Frequency domain analysis based on the projection signals of individual characters does not provide enough resolution for proper printer identification. Advanced pattern recognition techniques and knowledge about the print mechanism can help us to device an appropriate method to detect these signatures. We can get reliable intrinsic signatures from multiple projections to build a classifier to identify the printer. Projections from individual characters can be viewed as a high dimensional data set. In order to create a highly effective pattern recognition tool, this high dimensional projection data has to be represented in a low dimensional space. The dimension reduction can be performed by some well known pattern recognition techniques. Then a classifier can be built based on the reduced dimension data set. A popular choice is the Gaussian Mixture Model where each printer can be represented by a Gaussian distribution. The distributions of all the printers help us to determine the mixing coefficient for the projection from an unknown printer. Finally, the decision making algorithm can vote for the correct printer. In this paper we will describe different classification algorithms to identify an unknown printer. We will present the experiments based on several different EP printers in our printer bank. The classification results based on different classifiers will be compared.∗

Introduction

In our previous work, we have described the intrinsic and extrinsic features that can be used for printer identification.[1] Our intrinsic feature extraction method is based on frequency domain analysis of the one dimensional projected signal. If there are a sufficient number of samples in the projected signal, the Fourier transform gives us the correct banding frequency. When we work with a text-only document, our objective is to get the banding frequency from the projected signals of individual letters. In this situation, there are not enough samples per projection to give high frequency domain resolution. Significant overlap between spectra from different printers makes it difficult to use it as an effective classification method. The printer identification or classification task is closely related to various pattern identification and pattern recognition techniques. The intrinsic features are the patterns that are used to recognize an unknown printer. The basic idea is to create a classifier that can utilize the intrinsic signatures from a document to make proper identification. A Gaussian mixture model (GMM) or the tree based classifier is suitable for the classification part; but the initial dimension reduction is performed by principal component analysis.

Principal component analysis (PCA) is often used as a dimension-reducing technique within some other type of analysis.[2] Classical PCA is a linear transform that maps the data into a lower dimensional space by preserving as much data variance as possible. In the case of intrinsic feature extraction, PCA can be used to reduce the dimension of the projected signal. The proper number of components can be chosen to discriminate between different printers. These components are the features that can be used by the classifier (GMM or tree classifier).

The Gaussian mixture model defines the overall data set as a combination of several different Gaussian distributions. The parameters of the model are determined by the training data. Once the parameters are selected, the model is used to predict the printer, based on projections from the unknown printer.

Binary tree structured classifiers are formed by repeated splits of the original data set. Tree classifiers are also suitable for the printer identification problem. Properly

grown and pruned tree leaves should represent different printers.

In Fig. 1, the printer identification process is described briefly with the help of a flowchart. The printed document is scanned and the scanned image is used to get the one dimensional projected signal from individual characters1. PCA provides the lower dimensional feature space. GMM or tree classifier works on the feature space to correctly identify the unknown printer.

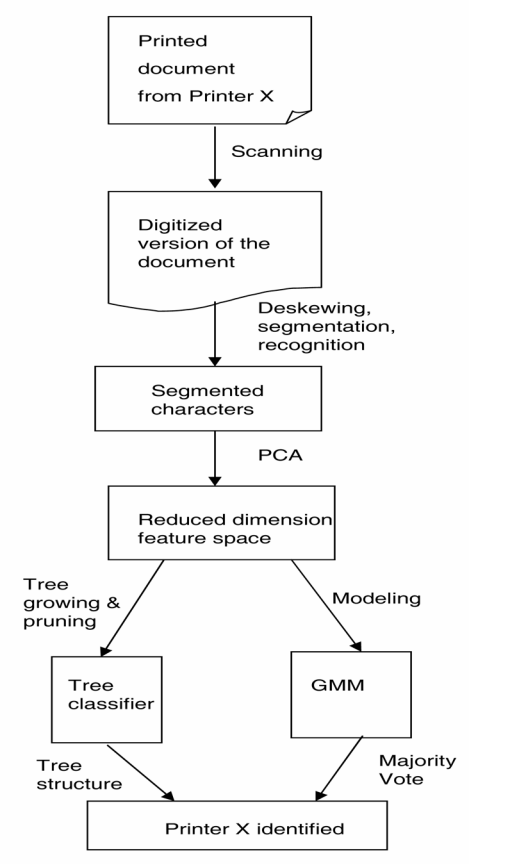

Figure 1. Steps followed in printer characterization.

In the following sections of the paper, dimension reduction technique by PCA, classification by GMM, and tree growing and pruning algorithm by binary tree classifiers are explained. Experimental results related to these algorithms are also provided.

Principal Component Analysis (PCA)

The theory behind principal component analysis is described in detail by Jolliffe.[2] Fukunaga,[3] and Webb.[4] In this section the fundamentals of the PCA are described.

An n dimensional vector $\mathbf{X}$ can be represented by the summation of nlinearly independent vectors.

$$\mathbf{X} = \sum_{i=1}^{n} y_i \phi_i = \Phi \mathbf{Y}, \tag{1}$$

where y_i is the i-th principal component and the ϕ_i's are the basis vectors obtained from the eigenvectors of the covariance matrix of $\mathbf{X}$. Using only $m < n$ of ϕ_i's the vector $\mathbf{X}$ can be approximated as

$$\hat{\mathbf{X}}(m) = \sum_{i=1}^{m} y_i \phi_i + \sum_{i=m+1}^{n} b_i \phi_i, \tag{2}$$

where b_i's are constants. The coefficients b_i and the vectors ϕ_i are to be determined so that $\mathbf{X}$ can be best approximated. If the first my_i's are calculated, the resulting error is

$$\Delta\mathbf{X}(m) = \mathbf{X} - \hat{\mathbf{X}}(m) = \sum_{i=m+1}^{n} (y_i - b_i)\phi_i. \tag{3}$$

The set of m eigenvectors of the covariance matrix of $\mathbf{X}$, which corresponds to the m largest eigenvalues, minimizes the error over all choices of m orthonormal basis vectors. The expansion of a random vector in the eigenvectors of covariance matrix is also called the discrete version of the Karhunen-Loéve expansion.[3]

Method of Canonical Variates

In the printer identification problem, we have additional information about the class label from the training data. We can get training samples from different known printers. The class label will represent different printer models. This additional class label information provides the optimal linear discrimination between classes with a linear projection of the data. This is known as the method of canonical variates.[5] Here the basis vectors are obtained from the between class and within class covariance matrices. Due to singularity in the covariance matrix, this method has to be implemented with the help of simultaneous digonalization.[6] This version of PCA is described in detail by Webb.[4] We have to diagonalize two symmetric matrices S_W and S_B simultaneously:

1. The within class covariance matrix S_W is whitened. The same transformation is applied to the between class covariance matrix S_B. Let us denote by S'_B the transformed S_B.
2. The orthonormal transformation is applied to diagonalize S'_B.

The complete mathematical formulation can be found in Webb[4] and Fukunaga.[6]

Experimental Results

To perform the experiments, we have used the printers in our printer bank.[1] The experimental procedure is depicted in Fig. 2.

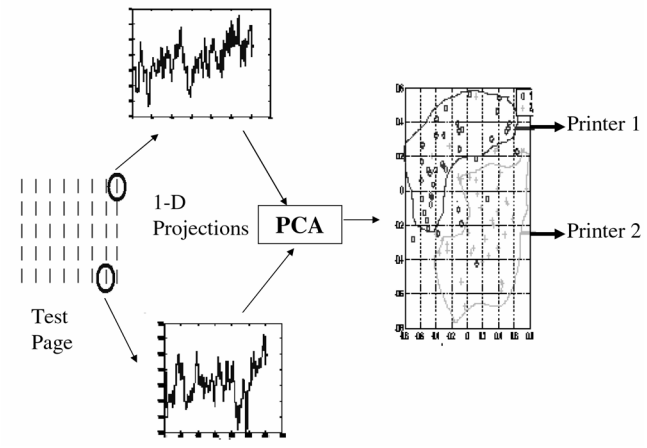

Figure 2. Principal component analysis using 1D projected signal.

The test page has the letter 'I' in 10pt., 12pt. and 14pt. size in Arial font. Each test page has 40-100 letters. From each letter, a one dimensional projected signal is extracted. The projected signals are mean subtracted and normalized. This step is done to remove variability due to long term trends, such as cartridge depletion and printer wear, and other factors which are not stable intrinsic features. The projected signals from different printers are concatenated into a large data matrix. The Canonical Variates method is applied to this data matrix to get the principal components.

The PCA using five different printer models is shown in Fig. 3. Each projection has 168 samples. The high dimensional data is represented only by the first two principal components. The classes (different printers) are well separated. A sixth printer is added as a 'test' printer. The sixth printer is an HP Laserjet 4050 and the projections from this printer (❑) overlap with those of the other Laserjet 4050 (o). The projections from the Laserjet 1000 (×) and Laserjet 1200 (✷) overlap because of the similarities in their banding characteristics.[1] It should be noted that the Samsung ML-1450 (+) and the Okipage 14e (◇) show well separated classes.

Gaussian Mixture Model for Classification

The dimension of the projected signal is reduced by PCA. The next step is to classify the printers using the features. The Gaussian mixture model (GMM) is a generative model. The posterior probability of a data point can be determined using Bayes' theorem. A model with m components is given by

$$p(\mathbf{x}) = \sum_{j=1}^{m} P(j)p(\mathbf{x}|j). \tag{4}$$

The parameter $P(j)$ is called the mixing coefficients. The component density function $p(x|j)$ is Gaussian with spherical covariance matrix and data dimension d.

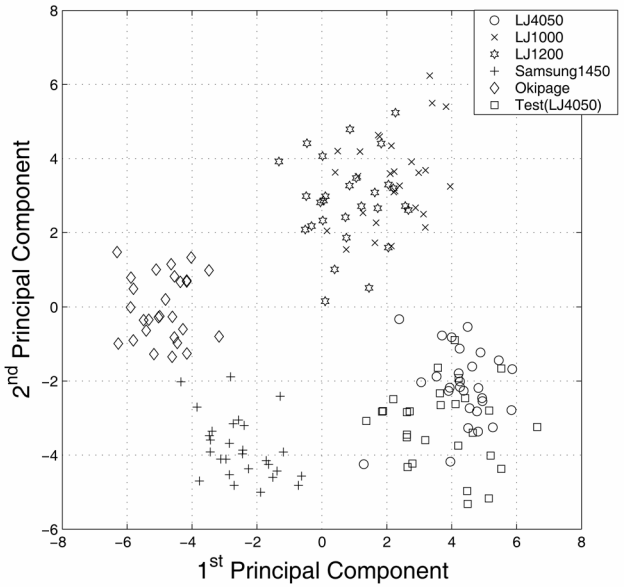

Figure 3. Representation of the projected signal by the first two principal components.

$$p(\mathbf{x}|j) = \frac{1}{(2\pi\sigma_j^2)^{\frac{d}{2}}} \exp\left\{-\frac{\|\mathbf{x}-\boldsymbol{\mu}_j\|^2}{2\sigma_j^2}\right\}. \tag{5}$$

Suppose that the data set contains projections from different printers A_1, ..., A_k. The classification can be done by using the posterior probabilities of class membership $P(A_k|x)$.[7] The density model for class A_t is built by training the model only on data from that class. This gives an estimate of $p(x|A_t)$. Then by using the Bayes' theorem,

$$P(A_t|\mathbf{x}) = \frac{p(\mathbf{x}|A_t)P(A_t)}{p(\mathbf{x})}. \tag{6}$$

The prior probabilities $P(A_t)$ are determined by the fraction of samples class A_t in the training set.

Parameter Estimation

The parameters of a Gaussian mixture are determined from a data set by maximizing the data likelihood. The most widely used method is the expectation-maximization (EM) algorithm.[8,9] The EM algorithm works by using the log likelihood expression of the complete data set. The maximization is repeatedly performed over the modified likelihood. Basically, the EM algorithm iteratively modifies the GMM parameters to decrease the negative log likelihood of the data set.[7] The source or class of each data point x^i is known during the training process. The maximization steps are explained in detail by Render.[10] The resulting equations are

$$P^{k+1}(j) = \frac{1}{N}\sum_{i=1}^{N} P^k(j|\mathbf{x}^i), \tag{7}$$

where $P^{(k+1)}(j)$ is the mixing coefficient at the $(k + 1)$-th iteration. N is the total number of samples, and $P^k(j|x^i)$ is the posterior probability from class i at the k-th iteration. The mean and the variance are updated by using the following relations

$$\boldsymbol{\mu}_j^{k+1} = \frac{\sum_{i=1}^{N} P^k(j|\mathbf{x}^i)\mathbf{x}^i}{\sum_{i=1}^{N} P^k(j|\mathbf{x}^i)}. \tag{8}$$

$$(\sigma_j^{k+1})^2 = \frac{1}{d}\frac{\sum_{i=1}^{N} P^k(j|\mathbf{x}^i)\|\mathbf{x}^i - \boldsymbol{\mu}_j^{k+1}\|^2}{\sum_{i=1}^{N} P^k(j|\mathbf{x}^i)}. \tag{9}$$

The initialization of the model is performed by the k-means algorithm. First, a rough clustering is done; and the number of clusters is determined by the number of printer models in the training set. Then each data point is assumed to belong to the closest cluster center. Initial prior probability, mean, and variances are calculated from these clusters. After that, the iterative part of the algorithm starts by using Eqs.(7)-(9).

Experimental Results

The feature space is obtained by PCA as described in the previous section. The model is based on only the first two principal components (d = 2). Model initialization is performed by seven iterations of the k-means algorithm. The initial parameters are used by the iterative EM algorithm to get the means and variances of the different classes. The EM algorithm is terminated by either the number of iterations or by a threshold depending on the amount of change in the parameters between successive iterations. The classification is done by majority vote. For each projection from the unknown printer, the posterior probabilities of all the classes are determined. The unknown projection belongs to the class with highest probability. This operation is performed with all the projections from the unknown printer. The class with highest number of votes represents the model of the unknown printer.

In Table 1, the classification result for five different printer models is presented. In the printer bank, we have two printers for each model1. One printer is used for training; and the other printer is used for testing. The initial training data set is created by forty projections from each different printer model. Then the sixth printer is added as an unknown printer to check the performance of the classifier. Forty projections from each printer are used for the experiment during training and testing. For example, when Laserjet 4050 is tested, the classifier predicts that all the 40 projections are from the Laserjet 4050 class, i.e. the classification is 100% correct. The Sam-sung ML-1450 and Okipage 14e are also identified correctly. Due to the close banding characteristics between the Laserjet 1200 and Laserjet 1000, the correct classification rate is decreased. This result confirms the outcome from the banding analysis[1] and PCA.

Table 1. Classification of Five Printers Using GMM*

Class / Test	LJ4050	LJ1200	LJ1000	Oki	SS	CCR (%)
LJ4050	**40**	0	0	0	0	100
LJ1200	0	**25**	15	0	0	62.5
LJ1000	0	35	**5**	0	0	12.5
Oki	0	0	0	**40**	0	100
SS1450	0	0	0	0	**40**	100

* LJ=Laserjet, Oki=Okipage 14e, SS= Samsung ML-1450, and CCR= Correct Classification Rate

Classification and Regression Tree (CART)

A classification tree is a multistage decision process that uses subsets of features at different levels of the tree. The construction of the tree involves three steps:
1. Splitting rule selection for each internal node.
2. Terminal node determination.
3. Class label assignment to terminal nodes.

The approach we have used in our experiments is based on CART.[11] The Gini criterion is used as the splitting rule. The terminal nodes are determined when there is pure class membership. In the printer identification problem, the class labels of the training printers are known; and the unknown printer can be given a temporary label. After the classification process, the class label of the unknown printer can be determined.

Experimental Results

The CART algorithm also works in the reduced dimension feature space generated by PCA. At each node the impurity function is defined by the Gini criterion. The algorithm tries to minimize the node impurity by selecting the optimum splitting. When the node impurity is zero; a terminal node is reached. When the complete tree is grown in this manner, it is overfitted. So the tree is pruned by the cross-validation method.

The experimental setup is similar to that for the GMM classification. Five different printer models are used for training; and a sixth printer is added as a test printer. Figure 4 shows the first two splitting on the data set. The first splitting is done at the value of -0.8473 of the second principal component. This split separates the Laserjet 1000 and Laserjet 1200 from the other printers. Similarly a second split at -1.15 of the first principal component separates the Laserjet 4050 from the Okipage 14e and Samsung ML1450.

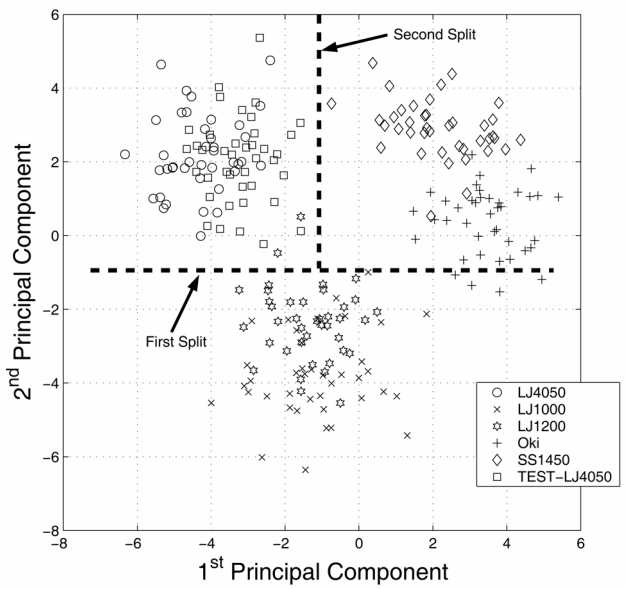

Figure 4. Splitting of the projected signal in the principal component domain by Gini criterion.

Figure 5 shows the complete grown and pruned tree. Each terminal node represents an individual printer. The Laserjet 1000 and Laserjet 1200 have the same parent node because of the similarities in the banding characteristics. The training and the test Laserjet 4050 have the same parent. So, the unknown Laserjet 4050 is identified properly by the tree classifier. The Samsung ML-1450 and the Okipage 14e are represented by separate terminal nodes.

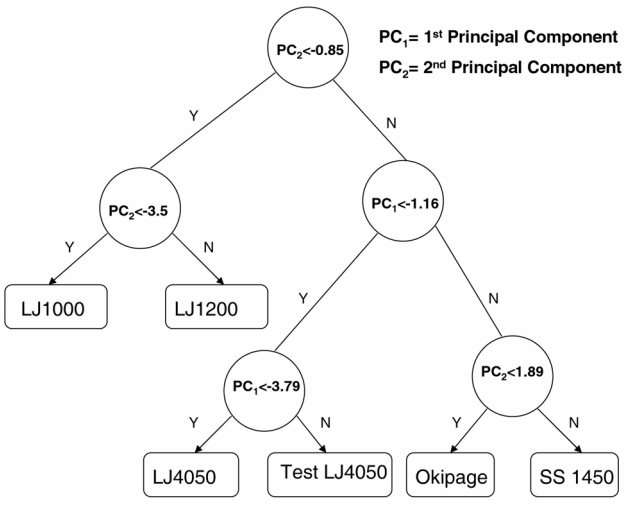

Figure 5. Binary tree structure for classifying an unknown printer. Five printer models are used in training process.

Conclusion

Printer identification from the text-only document requires sophisticated techniques based on feature extraction and pattern recognition. In our work, we presented a method for reducing the dimension of the data set. This reduced dimension data set functions as a feature space for the classifier. We developed two different classifiers based on the Gaussian mixture model and binary tree. If there is distinction in the feature space, both classifiers can identify the unknown printer properly.

References

* This research is supported by a grant from National Science Foundation, under award number 0219893.

1. G. N. Ali, P. C. Chiang, A. K. Mikkilineni, J. .P. Allebach, G. T. Chiu and E. J. Delp, "Intrinsic and Extrinsic Signatures for Information Hiding and Secure Printing with Electrophotographic Devices," in Proceedings of the IS&T's NIP19: International Conference on Digital Printing Technologies, 2003 ,pp. 511-515.
2. I. T. Jolliffe, "Principal Component Analysis," Springer-Verlag, Second Edition, 2002, pp. 199-231.
3. K. Fukunaga, "Statistical Pattern Recognition," Morgan Kaufmann, Second Edition, 1990, pp. 400-407.
4. K. Webb, "Statistical Pattern Recognition," John Wiley & Sons, Second edition, 2002, pp. 319-334.
5. I. T. Nabney, "Netlab: Algorithms for Pattern Recognition," Springer, pp. 273-275.
6. K. Fukunaga, "Statistical Pattern Recognition," Morgan Kaufmann, Second Edition, 1990, pp. 24-33.
7. I. T. Nabney, "Netlab: Algorithms for Pattern Recognition," Springer, pp. 79-113.
8. A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," J. Roy. Stat. Society B, vol 39, 1977, pp. 1-38.
9. C. F. J. Wu, "On the convergence properties of the EM algorithm," The Annals of Statistics, vol 11, no. 1, 1983, pp. 95-103.
10. R. A. Render and H. F. Walker, "Mixture Densities, maximum likelihood, and EM algorithm," SIAM review, vol. 26, no. 2, April 1984, pp. 195-239.
11. L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, "Classification and regression trees," Wadsworth & Brooks, 1984, pp. 18-36.

Biography

Gazi Naser Ali received the B.Sc. degree in electrical engineering from Bangladesh University of Engineering and Technology, in 1998. He is currently pursuing his Ph.D. degree in the School of Electrical and Computer Engineering, Purdue University, West Lafayette. He is a student member of IEEE and IS&T.