# Learning a Discriminative Sparse Tri-value Transform

Zhenhua Qu[1]
[1]*School of Information Science and Technology*
*Sun Yat-Sen University, China*
qzhua3@gmail.com

Guoping Qiu[2]
[2]*School of Computer Science*
*University of Nottingham, United Kingdom*
qiu@cs.nott.ac.uk

Pong C Yuen[*]
[*]*Department of Computer Science*
*Hong Kong Baptist University, Hong Kong*
pcyuen@comp.hkbu.edu.hk

## Abstract

*Simple binary patterns have been successfully used for extracting feature representations for visual object classification. In this paper, we present a method to learn a set of discriminative tri-value patterns for projecting high dimensional raw visual inputs into a low dimensional subspace for tasks such as face detection. Unlike previous methods that use predefined simple transform bases to generate tens of thousands features first and then use machine learning to select the most useful features, our method attempts to learn discriminative transform bases directly. Since it would be extremely hard to develop analytical solutions, we define an objective function that can be solved using simulated annealing. To reduce the search space, we impose sparseness and smoothness constraints on the transform bases. Experimental results demonstrate that our method is effective and provides an alternative approach to effective visual object classification.*

## 1. Introduction

For pattern recognition tasks such as visual object classification, e.g., face detection, one of the important challenges is to extracting discriminative representation features. Normally, a transform is used to transform the raw pixels into feature vectors. To counter the ”curse of dimensionality” and to facilitate fast computation, the transformed features should be of low dimensional. There have been a lot of work in the literature addressing this problem, such as Principle Component Analysis (PCA)[6][11], Linear Discriminat Analysis (LDA)[9] and the Independent Component Analysis(ICA) [3].

The results of Viola and Jones [10] have demonstrated that using simple tri-value transform bases is sufficient to extract effective representation features for visual object classifications. The computational advantage of tri-value patterns has also made this type of transform attractive. Another very interesting transform is random transform (RT) where the bases are generated randomly [2][5]. For these types of transforms, a posterior selection is oftern prefered [10][4].

Our new method’s transform bases are tri-value patterns in the set $\{-1, 0, +1\}$ similar to those used in [10][5][12] where the value 0’s corresponding to those pixels that are not involved in the computation. However, we take a different approach to these previous methods. In random transform, the transform bases are generated randomly with certain predefined distribution; in Viola and Jones, the transform patterns have predefined shapes and they are used to generate tens of thousand of features first and then machine learning is used to select the relevant features. In [12], the method is just for $\{0, 1\}$ binary bases. Instead, in this work, we attempt to derive the transform bases directly and the contribution of this paper can be regarded as introducing an alternative method to developing simple tri-value transform bases for feature extraction and representation for visual object classification.

## 2. Problem Formulation

The problem that we are interested is illustrated in Fig. 1 where an input image is convolved with a set of binary pattern (tri-value or tri-state bases) to produce a feature vector which is then used by a classifier to determine to identity of the input. It is not difficult to see the importance of a proper pattern design which should

978-1-4244-2175-6/08/$25.00 ©2008 IEEE

help producing more discriminative feature vectors.

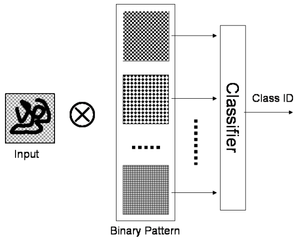


**Figure 1. Project onto tri-value bases.**

Generally speaking, the problem of finding the binary/tri-value bases is an integer programming problem. Since the computation complexity is NP-hard and the dimension of such bases is relatively high (several hundred), an exhaustive search approach is intractable. However if we can provide some heuristic information, such as the sparseness constraints or local structural constraints, the search space can be significantly reduced and a randomized optimization algorithm such as simulated annealing and genetic algorithm can be used to solve such kind of problem. It is obviously very difficult to simultaneously optimize all the bases. Instead, we extract the bases in a sequential manner. That is, we firstly find the most *discriminative one*, and then repeated this process until we get enough bases. Therefore, the problem we are facing is a discrete projection pursuit problem which can be formulated as follow:

**Problem Definition**: Given the sample set $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^N$ and corresponding class labels $\{c_i\}_{i=1}^N$ , we search for projection vectors $\mathbf{W} = \{\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_d\}$ that satisfy

$$\mathbf{W} = \arg\max{}_{\mathbf{W}} F(\mathbf{W}) + \alpha \cdot S(\mathbf{W}) + \beta \cdot C(\mathbf{W}) \quad (1)$$

s.t.$\mathbf{w}_i(k) \in A$, where $A$ is the alphabet, e.g. $A = \{0, 1\}$ for binary basis and $A = \{-1, 0, 1\}$ for the tri-value case. The objective function in (2) is composed of three terms, the discriminate term $F(\mathbf{w})$, sparseness terms $S(\mathbf{W})$ and the clique constraint term $C(\mathbf{W})$. The coefficient $\alpha$ and $\beta$ controls the trade off between them. And considering the sequential way of obtaining these bases, the $i$'th base can be obtained by turning the objective function into

$$\mathbf{w}_i = \arg\max{}_{\mathbf{w}} F(\mathbf{w}) + \alpha \cdot S(\mathbf{w}) + \beta \cdot C(\mathbf{w}) \quad (2)$$

The discriminative term $F(\mathbf{w})$ in (2) measures the separability of the input patterns (we consider two-class inputs here). It should maximize the distance between different classes while minimize the within class distance. Based on this principle, several discriminative functions can be used[8]. Here we mainly concentrated on the *Fisher Discriminant Function*.

$$F(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{S}_b \mathbf{w}}{\mathbf{w}^T \mathbf{S}_w \mathbf{w}} \quad (3)$$

(2) Taking into account the discreteness and sparseness of the tri-state bases, it is more convenient to write this formula in an equivalent form to reduce computation complexity while avoiding numerical instability.

$$F(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{S}_w^{-1} \mathbf{S}_b \mathbf{w}}{\mathbf{w}^T \mathbf{w}} \quad (4)$$

For the $\mathbf{S}_w$ which is too singular to invert, a PCA dimension reduction before the LDA is necessary. Then the objective function becomes

$$F(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{G}^T \hat{\mathbf{S}}_w^{-1} \hat{\mathbf{S}}_b \mathbf{G} \mathbf{w}}{\mathbf{w}^T \mathbf{w}} \quad (5)$$

where the matrix $\mathbf{G}$ is formed by the first $m$ principle components obtained from the training pattern features. $\hat{\mathbf{S}}_w$ and $\hat{\mathbf{S}}_b$ are the within and between class scatter matrices after the PCA dimension reduction.

The sparseness function $S(\mathbf{w})$ measures the sparseness of the bases. When we refer to tri-value bases where $A = \{-1, 0, 1\}$, it is the number of non-zero coefficients in the bases and can simply be written as $\mathbf{w}^T \mathbf{w}$. As will be shown later, if the base being optimized is constrained by a predefined constant sparseness, this term as well as the denominator of Equation (5) will remain constant in the optimization process.

The clique constraint function $C(\mathbf{w})$ controls the piece-wise smoothness of the tri-state bases. This term is introduced with two purposes: First, by applying it we can integrate some prior knowledge into the optimization process, such as the smoothness or geometrical characteristic of the pattern. This could eliminate a lot of bad candidate bases in our search space and leads to a faster convergence speed. Second, when the discriminative power of the pursuited base gradually falls, this constraint can force the non-zero coefficients concentrating on a discriminative local part.

Here we only considered a class of clique constraint functions which can be written in quadratic form.

$$C(\mathbf{w}) = \sum_{\psi \in \{-1,0,+1\}} \overline{\mathbf{w}}_\psi \mathbf{D}_\psi \overline{\mathbf{w}}_\psi$$

where $\overline{\mathbf{w}}_\psi$ is obtain from $\mathbf{w}$ using $\overline{w}_\psi(k) = \begin{cases} 1, w(k) = \psi \\ -1, w(k) \neq \psi \end{cases}$. The distance matrix $\mathbf{D}_\psi$ is assumed to have the form

$$\mathbf{D}_\psi = (d_{i,j}) = \left( e^{-\frac{dist(\Delta x_{ij}, \Delta y_{ij})}{\sigma}} \right) = \left( e^{-\frac{(\Delta x_{ij}^2 + \Delta y_{ij}^2)}{\sigma}} \right)$$

where $\Delta x_{ij}$ and $\Delta y_{ij}$ are the horizontal and vertical distance of two different entries in a base which should be defined with some prior knowledge to reveal the correlation between the two entries. In this work, we are interested in bases in which zeros and non-zero coefficient will cluster separately and this term can be intergrated into the sparseness term as $\alpha\overline{\mathbf{w}}_0(\mathbf{D}_0+\beta/\alpha\mathbf{I})\overline{\mathbf{w}}_0$.

In this work we considered two different distance functions: the cluster constraint and the symmetric constraint. The cluster constraint defined the distance function $dist(\Delta x_{ij}, \Delta y_{ij})$ with

$$\Delta x_{ij} = \operatorname{mod}(|x_i - x_j|, \left\lceil \frac{x_{\max}}{2} \right\rceil)$$

where $x_{max}$ is the width of the face patch and $\Delta y_{ij}$ can be defined similarly. It encourages non-zero coefficients to gather together and the connectivity is considered in a circular way that to avoid non-zeros coefficients to be "pushed" to the four corners by the zeros. The symmetric constraint take into consideration the symmetric characteristic of human face, if one portion of the pattern is considered to be non-discriminative (correspondent to zeros coefficients) then the symmetric counter part of it is very likely to be unimportant too. So the distant function $dist(\Delta x_{ij}, \Delta y_{ij})$ is defined with

$$\Delta x_{ij} = \left| x_i - \left\lceil \frac{x_{\max}}{2} \right\rceil \right| - \left| x_j - \left\lceil \frac{x_{\max}}{2} \right\rceil \right|$$

And $\Delta y_{ij}$ can be defined similarly.

## 3. Optimizing the Objective Function

We optimize the above objective function with the simulated annealing (SA) algorithm[7]. In practice, we optimize $N = 100$ bases of different sparseness in parallel and choose the one with the largest objective function as the result. Each of these bases is specified a constant sparseness defined by $\mathbf{S}(\mathbf{w}_i) = 240 - 2i, i = 0, \ldots, N-1$ and made zero mean by forcing an equal number of $+1$ and $-1$. Consequently, only swapping is needed in the permutation operation of SA algorithm.

The algorithm works by swapping the position of two randomly selected entries of a base $\mathbf{w}_i$. If the base $\hat{\mathbf{w}}_i$ resulted by this alteration increases the objective function, then accept the change. Otherwise, accept the change with a probability of $e^{\frac{J(\hat{\mathbf{w}}_i)-J(\mathbf{w}_i)}{T}}$. The temperature $T$ is set to a sufficient large value $T_{max}$in the initialization to make the algorithm have enough time to reach the optimum.

Initialization: $T = T_{\max}$,$\mathbf{W} = \phi$

*i*) Generate a set of random bases $\{\mathbf{w}_i | i = 0..N-1\}$ with different sparseness.

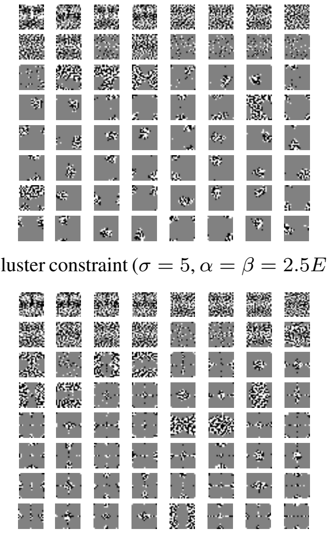
*ii*) Optimize each $\mathbf{w}$ with SA algorithm. Randomly swapping two entry of $\mathbf{w}_i$ to obtain a new base $\hat{\mathbf{w}}_i$. If $J(\hat{\mathbf{w}}_i) - J(\mathbf{w}_i) > 0$ or $e^{\frac{J(\hat{\mathbf{w}}_i)-J(\mathbf{w}_i)}{T}} > r$ , $\mathbf{w}_i = \hat{\mathbf{w}}_i$ . The random number $r$ obeys a uniform distribution between $[0, 1]$. Repeat this process for 3000 times and then decrease as $T = kT$, $k = 0.98$.

*iii*) Select the base with the maximum objective function value among $\{\mathbf{w}_i | i = 0..N-1\}$. Subtract $\mathbf{w}$ from $\mathbf{X}$ and $\mathbf{W} = \mathbf{W} \cup \{\mathbf{w}\}$.

## 4. Experimental Results

An face/non-face classification task is conducted to evaluate the effectiveness of our method. By using the MIT database [1], we have 2429 faces and 4,548 non-faces for training and 472 faces and 23,573 non-faces for testing. The $19 \times 19$ image patches are preprocessed with histogram equalization and mean subtraction to remove the influence of different illuminant.

We compared the classification performance of our selected tri-value bases to the results obtained with PCA, LDA, and RT. In the experiment, the image patches are firstly projected into a lower dimension feature space with aforementioned methods and then fed to a statistical classifier for training and testing. Three different feature dimension $(10, 30, 64)$ and two types classifiers: the Naive Bayesian classifier and support vector machine are investigated in this experiment.



(a) cluster constraint ($\sigma = 5, \alpha = \beta = 2.5E{-}3$)

(b) symmetry constraint ($\sigma = 2, \alpha = \beta = 2.5E - 3$)

**Figure 2. The 64 discriminative binary patterns learned with two different constraints, where -1's are back pixel, 0's are gray pixels and +1's are white pixels.**

Fig. 2 visualized the tri-value patterns learned from the face non-face database. It is seen that the first few bases reveal patterns resembles a human face. As more bases are extracted, they become more and more sparse. These sparse bases concentrated on local areas. From the parameter setting of our experiments, the intermediate stage between them is relatively short. Some interesting patterns with a specific shape which show some relationship to the discriminative part of human face are also observed in the transition process.

Fig. 3 shows comparison between the ROC curves of tri-value projection, PCA, LDA and RT. We can see that the tri-value bases successfully acquired the discriminative power of LDA projection and the performance is very close to LDA bases. As a result, the first a few bases (about 10 dim) is much more discriminative than RT. With the increase of base number, the RT which shares a similar perfect reconstruction characteristic with PCA started to get better in this application.

## 5. Concluding Remarks

In this paper, we have introduced a learning method to derive discriminative binary patterns for creating effective representation features for visual object classification. Compared with previous methods that used predefine the binary patterns, our method learns the binary pattern directly. We have introduced an objective function and two constraints to pursuit a set of tri-value bases that is both discriminative and sparse. Experiment results show that the tri-value bases have a similar performance to LDA while reducing the computation complexity.We will go on improving the performance and sparseness of the selected bases using more sophisticate prior knowledge of the patterns and to find more effective methods for our optimization problem.

## Acknowledgements

This work is partially supported by Faculty Research Grant of Hong Kong Baptist University.

## References

[1] *The MIT CBCL face and non-face database*. http://www.ai.mit.edu/projects/cbcl/software-datasets/index.html.

[2] D. Achlioptas. Database-friendly random projections. In *In Proc. of the 20th Symp. on Principles of Database Systems*, 2001.

[3] M. Bartlett, J. Movellan, and T. Sejnowski. Face recognition by independence component analysis,. *IEEE Trans on Neural Networks*, 13(6):1450 –1464, 2002.

[4] Fleuret. Fast binary feature selection with conditional mutual information. *The Journal of Machine Learning Research*, 5:1531–1555, 2004.

[5] N. Goel, G. Bebis, and A. Nefian. Face recognition experiments with random projection. In *In Proc. of SPIE Defense and Security Symposium*, volume 5779, page 426, 2005.

[6] Joliffe. *Principal component analysis*. Springer-Verlag, New York, 1986.

[7] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing, science. *Science*, 220(4598):671–680, 1983.

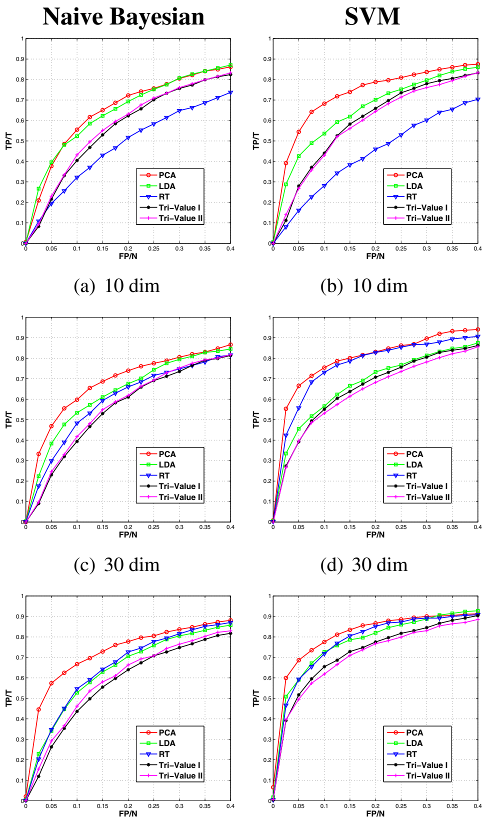
[8] H. Li, T. Jiang, and K. Zhang. Efficient and robust feature extraction by maximum margin criterion. *NISP*, 16, 2004.

[9] D. K. P.N. Belhumeur, J.P. Hespanha. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE TPAMI*, 19:711–720, 1997.

[10] P.Viola and M. Jones. Robust real time object detection. *IJCV*, 57(2):137–154, 2004.

[11] M. Turk and A. Pentland. Face recognition using eigenfaces. In *In Proc. of CVPR1991*, pages 586–591, 1991.

[12] S. Yan, X. Tang, and T. Yuan. Learning semantic patterns with discriminant localized binary projections. In *In Proc. of CVPR 2006*, volume 1, 2006.



(a) 10 dim

(b) 10 dim

(c) 30 dim

(d) 30 dim

(e) 64 dim

(f) 64 dim

**Figure 3. ROC of features extracted with different methods.**