Discriminant Non-negative Matrix Factorization and Projected Gradients for Frontal Face Verification.

Irene Kotsia, Stefanos Zafeiriou, and Ioannis Pitas

Aristotle University of Thessaloniki, Department of Informatics, Box 451, 54124, Greece
{ekotsia,dralbert,pitas}@aiia.csd.auth.gr
http://www.aiia.csd.auth.gr

Abstract. A novel *Discriminant Non-negative Matrix Factorization* (DNMF) method that uses projected gradients, is presented in this paper. The proposed algorithm guarantees the algorithm's convergence to a stationary point, contrary to the methods introduced so far, that only ensure the non-increasing behavior of the algorithm's cost function. The proposed algorithm employs some extra modifications that make the method more suitable for classification tasks. The usefulness of the proposed technique to the frontal face verification problem is also demonstrated.

Key words: Non-negative Matrix Factorization, projected gradients, frontal face verification.

1 Introduction

Over the past few years, the *Non-negative Matrix Factorization* (NMF) algorithm and its alternatives have been widely used, especially in facial image characterization and representation problems [3]. NMF aims at representing a facial image as a linear combination of basis images. Like *Principal Component Analysis* (PCA), NMF does not allow negative elements in either the basis images or the representation coefficients used in the linear combination of the basis images, thus representing the facial image only by additions of weighted basis images. The nonnegativity constraints introduced correspond better to the intuitive notion of combining facial parts to create a complete facial image.

In order to enhance the sparsity of NMF, many methods have been proposed for its further extension to supervised alternatives by incorporating discriminant constraints in the decomposition, the so-called DNMF or Fisher-NMF (FNMF) methods [3]. The intuitive motivation behind DNMF methods is to extract bases that correspond to discriminant facial regions and contain more discriminative information about them. A procedure similar to the one followed in the NMF decomposition [6] regarding the calculation of the update rules for the weights and the basis images was also used in the DNMF decomposition [3].

In this paper, a novel DNMF method is proposed that employs discriminant constraints on the classification features and not on the representation coefficients. Projected gradient methods are used for the optimization procedure to ensure that the limit point found will be a stationary point (similar methods have been applied to NMF [5]). Frontal face verification experiments were conducted and it has been demonstrated that the proposed method outperforms the other discriminant non-negative methods.

2 Discriminant Non-Negative Matrix Factorization Algorithms

2.1 Non-Negative Matrix Factorization

An image scanned row-wise is used to form a vector $\mathbf{x} = [x_1 \dots x_F]^T$ for the NMF algorithm. The basic idea behind NMF is to approximate the image $\mathbf{x}$ by a linear combination of the basis images in $\mathbf{Z} \in \Re_+^{F \times M}$, whose coefficients are the elements of $\mathbf{h} \in \Re_+^M$ such that $\mathbf{x} \approx \mathbf{Zh}$. Using the conventional least squares formulation, the approximation error $\mathbf{x} \approx \mathbf{Zh}$ is measured in terms of $L(\mathbf{x}||\mathbf{Zh}) \triangleq ||\mathbf{x} - \mathbf{Zh}||^2 = \sum_i (x_i - [\mathbf{Zh}]_i)^2$. Another way to measure the error of the approximation is using the Kullback-Leibler (KL) divergence, $KL(\mathbf{x}||\mathbf{Zh}) \triangleq \sum_i (x_i \ln \frac{x_i}{[\mathbf{Zh}]_i} + [\mathbf{Zh}]_i - x_i)$ [6] which is the most common error measure for all DNMF methods [3]. A limitation of KL-divergence is that it requires both $\mathbf{x}_i$ and $[\mathbf{Zh}]_i$ to be strictly positive (i.e., neither negative nor zero values are allowed).

In order to apply the NMF algorithm, the matrix $\mathbf{X} \in \Re_+^{F \times T} = [x_{ij}]$ should be constructed, where x_{ij} is the i-th element of the j-th image vector. In other words, the j-th column of $\mathbf{X}$ is the facial image $\mathbf{x}_j$. NMF aims at finding two matrices $\mathbf{Z} \in \Re_+^{F \times M} = [z_{i,k}]$ and $\mathbf{H} \in \Re_+^{M \times T} = [h_{k,j}]$ such that:

$$\mathbf{X} \approx \mathbf{ZH}. \tag{1}$$

After the NMF decomposition, the facial image $\mathbf{x}_j$ can be written as $\mathbf{x}_j \approx \mathbf{Zh}_j$, where $\mathbf{h}_j$ is the j-th column of $\mathbf{H}$. Thus, the columns of the matrix $\mathbf{Z}$ can be considered as basis images and the vector $\mathbf{h}_j$ as the corresponding weight vector. The vector $\mathbf{h}_i$ can be also considered as the projection of $\mathbf{x}_j$ in a lower dimensional space.

The defined cost for the decomposition (1) is the sum of all KL divergences for all images in the database:

$$D(\mathbf{X}||\mathbf{ZH}) = \sum_j KL(\mathbf{x}_j||\mathbf{Zh}_j) = \sum_{i,j} \left(x_{i,j} \ln(\frac{x_{i,j}}{\sum_k z_{i,k} h_{k,j}}) + \sum_k z_{i,k} h_{k,j} - x_{i,j} \right). \tag{2}$$

The NMF factorization is the outcome of the following optimization problem:

$$\min_{\mathbf{Z},\mathbf{H}} D(\mathbf{X}||\mathbf{ZH}) \text{ subject to} \tag{3}$$
$$z_{i,k} \geq 0,\ h_{k,j} \geq 0,\ \sum_i z_{i,j} = 1,\ \forall j.$$

2.2 Discriminant Non-Negative Matrix Factorization

In order to formulate the DNMF algorithm, let the matrix $\mathbf{X}$ that contains all the facial images be organized as follows. The j-th column of the database $\mathbf{X}$ is the ρ-th image of the r-th image class. Thus, $j = \sum_{i=1}^{r-1} N_i + \rho$, where N_i is the cardinality of the image class i. The r-th image class could consist of one person's facial images, for face recognition and verification problems. The vector $\mathbf{h}_j$ that corresponds to the j-th column of the matrix $\mathbf{H}$, is the coefficient vector for the ρ-th facial image of the r-th class and will be denoted as $\boldsymbol{\eta}_\rho^{(r)} = [\eta_{\rho,1}^{(r)} \dots \eta_{\rho,M}^{(r)}]^T$. The mean vector of the vectors $\boldsymbol{\eta}_\rho^{(r)}$ for the class r is denoted as $\boldsymbol{\mu}^{(r)} = [\mu_1^{(r)} \dots \mu_M^{(r)}]^T$ and the mean of all classes as $\boldsymbol{\mu} = [\mu_1 \dots \mu_M]^T$. Then, the within-class scatter matrix for the coefficient vectors $\mathbf{h}_j$ is defined as:

$$\mathbf{S}_w = \sum_{r=1}^{K} \sum_{\rho=1}^{N_r} (\boldsymbol{\eta}_\rho^{(r)} - \boldsymbol{\mu}^{(r)})(\boldsymbol{\eta}_\rho^{(r)} - \boldsymbol{\mu}^{(r)})^T \tag{4}$$

whereas the between-class scatter matrix is defined as:

$$\mathbf{S}_b = \sum_{r=1}^{K} N_r (\boldsymbol{\mu}^{(r)} - \boldsymbol{\mu})(\boldsymbol{\mu}^{(r)} - \boldsymbol{\mu})^T. \tag{5}$$

The matrix $\mathbf{S}_w$ defines the scatter of the sample vector coefficients around their class mean. The dispersion of samples that belong to the same class around their corresponding mean should be as small as possible. A convenient measure for the dispersion of the samples is the trace of $\mathbf{S}_w$. The matrix $\mathbf{S}_b$ denotes the between-class scatter matrix and defines the scatter of the mean vectors of all classes around the global mean $\boldsymbol{\mu}$. Each class must be as far as possible from the other classes. Therefore, the trace of $\mathbf{S}_b$ should be as large as possible.

To formulate the DNMF method [3], discriminant constraints have been incorporated in the NMF decomposition inspired by the minimization of the Fisher's criterion [3]. The DNMF cost function is given by:

$$D_d(\mathbf{X}||\mathbf{ZH}) = D(\mathbf{X}||\mathbf{ZH}) + \gamma \mathrm{tr}[\mathbf{S}_w] - \delta \mathrm{tr}[\mathbf{S}_b] \tag{6}$$

where γ and δ are non-negative constants. The update rules that guarantee a non-increasing behavior of (6) for the weights $h_{k,j}$ and the bases $z_{i,k}$, under the constraints of (2), can be found in [3]. Unfortunately, the update rules only guarantee a non-increasing behavior for (6) and do not ensure that the limit point will be stationary.

3 Projected Gradient Methods for Discriminant Non-Negative Matrix Factorization

Let $\mathbf{E} = \mathbf{X} - \mathbf{ZH}$ be the error signal of the decomposition. The modified optimization problem should minimize:

$$D_p(\mathbf{X}||\mathbf{ZH}) = ||\mathbf{E}||_F^2 + \gamma \mathrm{tr}[\tilde{\mathbf{S}}_w] - \delta \mathrm{tr}[\tilde{\mathbf{S}}_b], \tag{7}$$

under non-negativity constraints, where $||.||_F$ is the Frobenius norm. The within-class scatter matrix $\tilde{\mathbf{S}}_w$ and the between-scatter scatter matrix $\tilde{\mathbf{S}}_b$ are defined using the vectors $\tilde{\mathbf{x}}_j = \mathbf{Z}^T\mathbf{x}_j$ and the definitions of the scatter matrices in (4) and (5).

The minimization of (7) subject to nonnegative constraints yields the new discriminant nonnegative decomposition. The new optimization problem is the minimization of (7) subject to non-negative constraints for the weights matrix $\mathbf{H}$ and the bases matrix $\mathbf{Z}$. This optimization problem will be solved using projected gradients in order to guarantee that the limit point will be stationary. In order to find the limit point, two functions are defined:

$$f_{\mathbf{Z}}(\mathbf{H}) = D_p(\mathbf{X}||\mathbf{ZH}) \text{ and } f_{\mathbf{H}}(\mathbf{Z}) = D_p(\mathbf{X}||\mathbf{ZH}) \tag{8}$$

by keeping $\mathbf{Z}$ and $\mathbf{H}$ fixed, respectively.

The projected gradient method used in this paper, successively optimizes two subproblems [5]:

$$\min_{\mathbf{Z}} f_{\mathbf{H}}(\mathbf{Z}) \text{ subject to, } z_{i,k} \geq 0, \tag{9}$$

and

$$\min_{\mathbf{H}} f_{\mathbf{Z}}(\mathbf{H}) \text{ subject to, } h_{k,j} \geq 0. \tag{10}$$

The method requires the calculation of the first and the second order gradients of the two functions in (8):

$$\begin{aligned} \nabla f_{\mathbf{Z}}(\mathbf{H}) &= \mathbf{Z}^T(\mathbf{ZH} - \mathbf{X}) \\ \nabla^2 f_{\mathbf{Z}}(\mathbf{H}) &= \mathbf{Z}^T\mathbf{Z} \\ \nabla f_{\mathbf{H}}(\mathbf{Z}) &= (\mathbf{ZH} - \mathbf{X})\mathbf{H}^T + \gamma\nabla\mathrm{tr}[\tilde{\mathbf{S}}_w] - \delta\nabla\mathrm{tr}[\tilde{\mathbf{S}}_b] \\ \nabla^2 f_{\mathbf{H}}(\mathbf{Z}) &= \mathbf{HH}^T + \gamma\nabla^2\mathrm{tr}[\tilde{\mathbf{S}}_w] - \delta\nabla^2\mathrm{tr}[\tilde{\mathbf{S}}_b]. \end{aligned} \tag{11}$$

The projected gradient DNMF method is an iterative method that is comprised of two main phases. These two phases are iteratively repeated until the ending condition is met or the number of iterations exceeds a given number. In the first phase, an iterative procedure is followed for the optimization of (9), while in the second phase, a similar procedure is followed for the optimization of (10). In the beginning, the bases matrix $\mathbf{Z}^{(1)}$ and the weight matrix $\mathbf{H}^{(1)}$ are initialized either randomly or by using structured initialization [7], in such a way that their entries are nonnegative. The regularization parameters γ and δ that are used to balance the trade-off between accuracy of the approximation and discriminant decomposition of the computed solution and their selection is typically problem dependent.

3.1 Solving the Subproblem (9)

Consider the subproblem of optimizing with respect to $\mathbf{Z}$, while keeping the matrix $\mathbf{H}$ constant. The optimization is an iterative procedure that is repeated until $\mathbf{Z}^{(t)}$ becomes a stationary point of (9). In every iteration, a proper step size a_t is required to update the matrix $\mathbf{Z}^{(t)}$. When a proper update is found, the stationarity condition is checked and, if met, the procedure stops.

Update the matrix Z For a number of iterations $t = 1, 2, \ldots$ the following updates are performed [5]:

$$\mathbf{Z}^{(t+1)} = P\left[\mathbf{Z}^{(t)} - a_t \nabla f_{\mathbf{H}}(\mathbf{Z}^{(t)})\right] \tag{12}$$

where $a_t = \beta^{g_t}$ and g_t is the first non-negative integer such that:

$$f_{\mathbf{H}}(\mathbf{Z}^{(t+1)}) - f_{\mathbf{H}}(\mathbf{Z}^{(t)}) \leq \sigma \left\langle \nabla f_{\mathbf{H}}(\mathbf{Z}^{(t)}), \mathbf{Z}^{(t+1)} - \mathbf{Z}^{(t)} \right\rangle. \tag{13}$$

The projection rule $P[.] = \max[., 0]$ refers to the elements of the matrix and guarantees that the update will not contain any negative entries. The operator $\langle ., . \rangle$ is the inner product between matrices defined as:

$$\langle \mathbf{A}, \mathbf{B} \rangle = \sum_i \sum_j a_{i,j} b_{i,j} \tag{14}$$

where $[\mathbf{A}]_{i,j} = a_{i,j}$ and $[\mathbf{B}]_{i,j} = b_{i,j}$. The condition (13) ensures the sufficient decrease of the $f_{\mathbf{H}}(\mathbf{Z})$ function values per iteration. Since the function $f_{\mathbf{H}}$ is quadratic in terms of $\mathbf{Z}$, the inequality (13) can be reformulated as:

$$(1-\sigma)\left\langle \nabla f_{\mathbf{H}}(\mathbf{Z}^{(t)}), \mathbf{Z}^{(t+1)} - \mathbf{Z}^{(t)} \right\rangle + \frac{1}{2}\left\langle \mathbf{Z}^{(t+1)} - \mathbf{Z}^{(t)}, \nabla^2 f_{\mathbf{H}}(\mathbf{Z}^{(t+1)}) \right\rangle \leq 0 \tag{15}$$

which is the actual condition checked.

The search of a proper value for a_t is the most time consuming procedure, thus, as few iteration steps as possible are desired. Several procedures have been proposed for the selection and update of the a_t values [8]. The Algorithm 4 in [5] has been used in our experiments and β, σ are chosen to be equal to 0.1 and 0.01 ($0 < \beta < 1$, $0 < \sigma < 1$), respectively. The choice of σ has been thoroughly studied in [5, 8]. During experiments it was observed that a smaller value of β reduces more aggressively the step size, but it may also result in a step size that is too small. The search for a_t is repeated until the point $\mathbf{Z}^{(t)}$ becomes a stationary point.

Check of Stationarity In this step it is checked whether or not in the limit point the first order derivatives are close to zero (stationarity condition). A commonly used condition to check the stationarity of a point is the following [8]:

$$||\nabla^P f_{\mathbf{H}}(\mathbf{Z}^{(t)})||_F \leq \epsilon_{\mathbf{Z}} ||\nabla f_{\mathbf{H}}(\mathbf{Z}^{(1)})||_F \tag{16}$$

where $\nabla^P f_{\mathbf{H}}(\mathbf{Z})$ is the projected gradient for the constraint optimization problem defined as:

$$[\nabla^P f_{\mathbf{H}}(\mathbf{Z})]_{i,k} = \begin{cases} [\nabla f_{\mathbf{H}}(\mathbf{Z})]_{i,k} & \text{if } z_{i,k} > 0 \\ \min(0, [\nabla f_{\mathbf{H}}(\mathbf{Z})]_{i,k}) & z_{i,k} = 0. \end{cases} \tag{17}$$

and $0 < \epsilon_{\mathbf{Z}} < 1$ is the predefined stopping tolerance. A very low $\epsilon_{\mathbf{Z}}$ (i.e., $\epsilon_{\mathbf{Z}} \approx 0$) leads to a termination after a large number of iterations. On the other hand, a tolerance close to one will result in a premature iteration termination.

3.2 Solving the Subproblem (10)

A similar procedure should be followed in order to find a stationary point for the subproblem (10) while keeping fixed the matrix $\mathbf{Z}$ and optimizing in respect of $\mathbf{H}$. A value for a_t is iteratively sought and the weight matrix is updated according to:

$$\mathbf{H}^{(t+1)} = P\left[\mathbf{H}^{(t)} - a_t \nabla f_{\mathbf{Z}}(\mathbf{H}^{(t)})\right] \tag{18}$$

until the function $f_{\mathbf{Z}}(\mathbf{H})$ value is sufficient decreased and the following inequality holds $\langle a, b\rangle$:

$$(1-\sigma)\left\langle \nabla f_{\mathbf{Z}}(\mathbf{H}^{(t)}), \mathbf{H}^{(t+1)} - \mathbf{H}^{(t)} \right\rangle + \frac{1}{2}\left\langle \mathbf{H}^{(t+1)} - \mathbf{H}^{(t)}, \nabla^2 f_{\mathbf{Z}}(\mathbf{H}^{(t+1)}) \right\rangle \leq 0. \tag{19}$$

This procedure is repeated until the limit point $\mathbf{H}^{(t)}$ is stationary. The stationarity is checked using a similar criterion to (16), i.e.:

$$||\nabla^P f_{\mathbf{Z}}(\mathbf{H}^{(t)})||_F \leq \epsilon_{\mathbf{H}} ||\nabla f_{\mathbf{Z}}(\mathbf{H}^{(1)})||_F \tag{20}$$

where $\epsilon_{\mathbf{H}}$ is the predefined stopping tolerance for this subproblem.

3.3 Convergence Rule

The procedure followed for the minimization of the two subproblems, in Sections 3.1 and 3.2, is iteratively followed until the global convergence rule is met:

$$||\nabla f(\mathbf{H}^{(t)})||_F + ||\nabla f(\mathbf{Z}^{(t)})||_F \leq \epsilon \left(||\nabla f(\mathbf{H}^{(1)})||_F + ||\nabla f(\mathbf{Z}^{(1)})||_F\right) \tag{21}$$

which checks the stationarity of the solution pair $\mathbf{H}^{(t)}, \mathbf{Z}^{(t)}$.

4 Experimental Results

The proposed DNMF method will be denoted as Projected Gradient DNMF (PGDNMF) from now onwards. The experiments were conducted in the XM2VTS database using the protocol described in [12]. The images were aligned semi-automatically according to the eyes position of each facial image using the eye coordinates. The facial images were down-scaled to a resolution of 64×64 pixels. Histogram equalization was used for the normalization of the facial image luminance.

The XM2VTS database contains 295 subjects, 4 recording sessions and two shots (repetitions) per recording session. It provides two experimental setups namely, Configuration I and Configuration II [12]. Each configuration is divided into three different sets: the training set, the evaluation set and the test set. The training set is used to create client and impostor models for each person. The evaluation set is used to learn the verification decision thresholds. In case of multimodal systems, the evaluation set is also used to train the fusion manager

[12]. For both configurations the training set has 200 clients, 25 evaluation impostors and 70 test impostors. The two configurations differ in the distribution of client training and client evaluation data. For additional details concerning the XM2VTS database an interested reader can refer to [12].

The experimental procedure followed was the one also used in [3]. For comparison reasons the same methodology using the Configuration I of the XM2VTS database was used. The performance of the algorithms is quoted by the Equal Error Rate (EER) which is the scalar figure of merit that is often used to judge the performance of a verification algorithm. An interested reader may refer to [12, 3] for more details concerning the XM2VTS protocol and the experimental procedure followed. In Figure 1, the verification results are shown for the various tested approaches, NMF [6], LNMF [11], DNMF [3], Class Specific DNMF [3], PCA [9], PCA plus LDA [10] and the proposed PGDNMF. EER is ploted versus the dimensionality of the new lower dimension space. As can be seen, the proposed PGDNMF algorithm outperforms (giving a best $EER \approx 2.0\%$) all the other part-based approaches and PCA. The best performance of LDA has been 1.7% which very close to the best performance of PGDNMF.

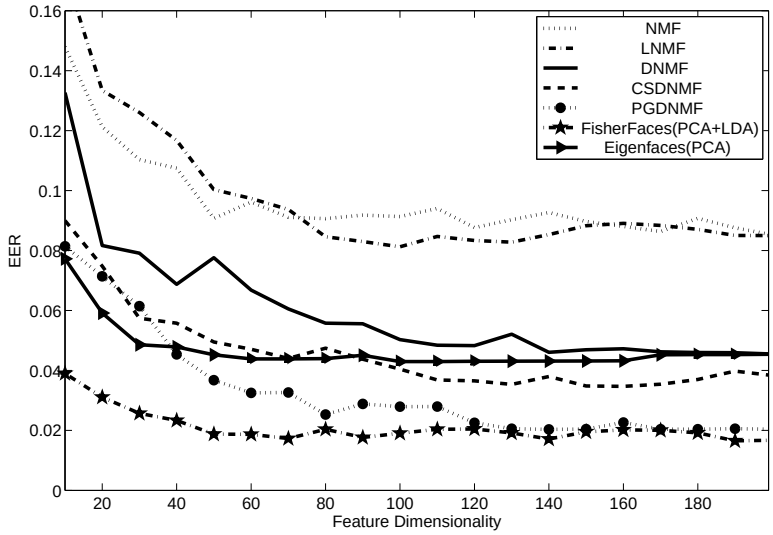

Fig. 1. EER for Configuration I of XM2VTS versus dimensionality.

5 Conclusions

A novel DNMF method has been proposed based on projected gradients. The incorporated discriminant constraints focus on the actual features used for classification and not on the weight vectors of the decomposition. Moreover, we have applied projected gradients in order to assure that the limit point is stationary. The proposed technique has been applied in supervised facial feature extraction for face verification, where it was shown that it outperforms several others subspace methods.

Acknowledgments. This work has been partially supported by the COST 2101 "Biometrics for Identity Documents and Smart Cards", www.cost2101.org.

References

1. D.D. Lee and H.S. Seung : Learning the parts of objects by non-negative matrix factorization. Nature 401, 788–791 (1999)
2. I. Buciu and I. Pitas : Application of non-negative and local non negative matrix factorization to facial expression recognition. In: ICPR 2004, pp. 288–291. Cambridge, United Kingdom (2004)
3. S. Zafeiriou, A. Tefas, I. Buciu and I. Pitas: Exploiting Discriminant Information in Nonnegative Matrix Factorization With Application to Frontal Face Verification. IEEE Transactions on Neural Networks, vol. 17, num. 3, pp. 683–695 (2006)
4. M. Kirby and L. Sirovich : Application of the Karhunen-Loeve Procedure for the Characterization of Human Faces. IEEE Transactions Pattern Analysis and Machine Intelligence, vol. 12, num. 1, 103–108 (1990)
5. C.-J. Lin: Projected gradient methods for non-negative matrix factorization. Technical report, Department of Computer Science, National Taiwan University (2005)
6. D.D. Lee and H.S. Seung : Algorithms for Non-negative Matrix Factorization. In: NIPS 2000, pp. 556–562.
7. I. Buciu, N. Nikolaidis and I. Pitas : On the initialization of the DNMF algorithm. In: Proc. of 2006 IEEE International Symposium on Circuits and Systems, Kos, Greece (2006).
8. C.-J. Lin and J.J. More: Newton's method for large-scale bound constrained problems. SIAM Journal on Optimization, vol. 9, pp. 1100–1127 (1999)
9. M. Turk and A. P. Pentland: Eigenfaces for Recognition. Journal of Cognitive Neuroscience, vol. 3, pp. 71–86 (1991)
10. P. N. Belhumeur, J. P. Hespanha and D. J. Kriegman : Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, num. 7, pp. 711–720 (1997)
11. S.Z. Li, X.W. Hou and H.J. Zhang: Learning Spatially Localized, Parts-Based Representation. In: CVPR 2001, Kauai, HI, USA (2001).
12. K. Messer, J. Matas, J.V. Kittler, J. Luettin and G. Maitre : XM2VTSDB: The Extended M2VTS Database. In: AVBPA'99, pp. 72–77, Washington, DC, USA (1999).