Hypothesis Testing in High-Dimensional Space with the Sparse Matrix Transform

Leonardo R. Bachega, Charles A. Bouman
Purdue University
School of Electrical and Computer Engineering
West Lafayette, IN, 47907-2035, USA
{lbachega, bouman}@purdue.edu

James Theiler
Los Alamos National Laboratory
Space and Remote Sensing Group
Los Alamos, NM 87545, USA
jt@lanl.gov

***Abstract*—This paper discusses the use of the Sparse Matrix Transform (SMT) to model the covariance structure of high-dimensional data in the likelihood ratio test used for hypothesis testing. The SMT has been shown to produce more accurate estimates of covariance matrices when the number of training samples n is much less than the number of dimensions p of the data. Several experiments with face recognition and hyperspectral images show that SMT-based hypothesis testing can be superior to other methods in at least two general aspects: First, the SMT-based method is more robust to the size of the training set, remaining accurate even when only a few training samples are available; Second, the total computation required to apply the method is very low, making it attractive for use in low-power devices, or in applications requiring fast computation.**

I. Introduction

Statistical hypothesis testing is widely used in signal processing and machine learning. According to the seminal *Neyman-Pearson* lemma [1], when deciding between two alternative hypotheses, the test with most discrimination power depends on one's knowledge of the ratio between the likelihoods under both hypotheses, and therefore, the knowledge of the data covariance matrices under both hypotheses. In practice the true covariances are not known and we need to rely on estimates from available training sets.

However, when the data dimensionality p is large, the number of training samples, n available to estimate the covariances involved in the likelihood ratio test is small compared to p, making conventional covariance estimates to behave poorly. As argued in [2], this $n \ll p$ scenario is rather common. Nevertheless, even if one had enough samples to obtain accurate covariance matrix estimates, when p is large, the amount of computation required to compute their eigen-decomposition and the memory space required to store them would both be prohibitive, limiting the practical application of such tests.

The Sparse Matrix Transform (SMT) [3], [4] is capable of successfully modeling the covariance structure of high dimensional data in the scenario when $n \ll p$, and requiring low computational cost when applied. In this paper we investigate the SMT deployment to estimate the covariance matrices involved in log-likelihood ratio for hypothesis testing. We look at three different flavors of hypothesis testing: matched filtering, power detection and classification.

Results in detection involving hyperspectral images and face recognition suggest that the accuracy of detectors and classifiers relying on SMT is better than of competing methods when few training samples are available, while the computation associated with its application is significantly lower. In the case when the true covariances are known, a sparse representation of the covariances by the SMT can reduce the computation required for the likelihood ratio test while yielding to similar accuracy to the exact method.

II. The Sparse Matrix Transform (SMT)

The essence of our method is to use SMTs to provide full-rank estimates of the $p \times p$ covariance matrices used in the detection and classification frameworks discussed in Section III.

A. Design of the SMT transform

The SMT design consists of estimating the full set of eigenvectors and associated eigenvalues for a general p-dimensional signal. More specifically, the objective is to estimate the orthonormal matrix E and diagonal matrix Λ such that the signal covariance can be decomposed as $R = E\Lambda E^t$, and to compute this estimate from n independent training vectors, $Y = [\mathbf{y}_1, \cdots, \mathbf{y}_n]$. This is done by assuming the samples are i.i.d. Gaussian random vectors and computing the constrained maximum log-likelihood (ML) estimates of E and Λ. In [3], we show that these constrained ML estimates are given by

$$\hat{E} = \arg\min_{E \in \Omega_K} \left\{ \left| \mathrm{diag}(E^t S E) \right| \right\} \tag{1}$$

$$\hat{\Lambda} = \mathrm{diag}(\hat{E}^t S \hat{E}) , \tag{2}$$

where $S = \frac{1}{n} Y Y^t$ is the sample covariance matrix, and Ω_K is the set of allowed orthonormal transforms.

If $n > p$ and Ω_K is the set of all orthonormal transforms, then the solution to (1) and (2) is the diagonalization of the sample covariance, i.e, $\hat{E}\hat{\Lambda}\hat{E}^t = S$. However, the sample covariance is a poor estimate of the covariance when $n < p$.

This material is based upon work supported by, or in part by, the U.S. Army Research Laboratory and the U.S. Army Research Office under contract/grant number 56541-CI.

In order to improve the accuracy of the covariance estimate, we will impose the constraint that Ω_K be the set of sparse matrix transforms (SMT) of order K. More specifically, we will assume that the eigen-transformation has the form

$$E = \prod_{k=1}^{K} E_k = E_1 \cdots E_K \ , \tag{3}$$

where each E_k is a planar rotation over some (i_k, j_k) coordinate pair by an angle θ_k, and K is the model order parameter.

Intuitively, each Givens rotation, E_k, plays the same role as the butterflies of a fast Fourier transform (FFT). In fact, the SMT is a generalization of both the FFT and the orthonormal wavelet transform. However, since both the ordering of the coordinate pairs, (i_k, j_k), and the values of the rotation angles, θ_k, are unconstrained, the SMT can model a much wider range of transformations. It is often useful to express the order of the SMT as $K = rp$, where r is the average number of rotations per coordinate, being typically very small: $r < 5$. The optimization of (1) is non-convex, so we use a greedy optimization approach in which we select each rotation, E_k, in sequence to minimize the cost. The greedy optimization can be done fast if a graphical constraint can be imposed to the data [4]. The parameter r can be estimated using cross-validation over the training set[3], [4] or using the minimum length description criterion proposed in [5].

B. Application of the SMT transform

Typically, r is small (< 5), so that the computation to apply the SMT to a vector of data is very low, i.e, $2r + 1$ floating-point operations per coordinate. Therefore, we can apply the SMT decorrelating transform to p-dimensional random vectors in only $(2r + 1)p$ steps.

III. Hypothesis Testing

Let $\mathbf{x}$ be a p-dimensional random vector drawn from a multivariate normal distribution. One seeks to decide between the hypotheses

$$\begin{array}{ll} \mathcal{H}_0 : & \mathbf{x} \sim \mathcal{N}(\mu_A, R_A) \\ \mathcal{H}_1 : & \mathbf{x} \sim \mathcal{N}(\mu_B, R_B) \ , \end{array} \tag{4}$$

where $\mathcal{H}_0$ and $\mathcal{H}_1$ are referred as the *null* and *alternative* hypotheses respectively. The *Neyman-Pearson* lemma [1] states that the log-likelihood ratio test

$$l(\mathrm{x}) = \log \left\{ \frac{p(\mathrm{x}; \mathcal{H}_1)}{p(\mathrm{x}; \mathcal{H}_0)} \right\} \gtrless \eta \tag{5}$$

maximizes the probability of detection $p(\mathcal{H}_1; \mathcal{H}_1)$ for a fixed probability of false alarm $p(\mathcal{H}_1; \mathcal{H}_0)$, which is controled by the threshold η.

Below, we discuss how the log-likelihood ratio test in (5) is used to test alternative hypotheses in the context of three common problems in signal processing, involving detection and classification of random signals.

A. Matched Filter

Let $\mathrm{t} \in \mathbb{R}^p$ be a deterministic signal buried in additive random clutter $\mathbf{w} \sim \mathcal{N}(0, R)$. The random vector $\mathbf{x}$ is measured and one wants to make a decision on whether the signal t is present (i.e, $\mathbf{x} = \mathrm{t} + \mathbf{w}$), or the measurement contains only clutter (i.e, $\mathbf{x} = \mathbf{w}$), by testing the hypotheses

$$\begin{array}{ll} \mathcal{H}_0 : & \mathbf{x} \sim \mathcal{N}(0, R) \\ \mathcal{H}_1 : & \mathbf{x} \sim \mathcal{N}(\mathrm{t}, R) \ . \end{array} \tag{6}$$

In this case, the log-likelihood ratio test in (5) has the form of an inner product: $l(\mathbf{x}) = \mathrm{q}^t \mathbf{x} \gtrless \eta'$, where the vector $\mathrm{q} \triangleq R^{-1}\mathrm{t}$ is called a *matched filter*, and its detection capability is measured directly by the signal-to-clutter statistic [6]:

$$SCR = \frac{(\mathrm{q}^t\mathrm{t})^2}{\mathbb{E}\left\{(\mathrm{q}^t\mathbf{x})^2\right\}} = \frac{(\mathrm{q}^t\mathrm{t})^2}{\mathrm{q}^t\mathbb{E}\left\{\mathbf{x}\mathbf{x}^t\right\}\mathrm{q}} = \frac{(\mathrm{q}^t\mathrm{t})^2}{\mathrm{q}^t R \mathrm{q}} \ . \tag{7}$$

B. Power Detector

Let the p-dimensional random vector $\mathbf{x}$ be drawn from a multivariate normal distribution with the same mean under both hypotheses but different covariances. The general hypotheses in (4) become

$$\begin{array}{ll} \mathcal{H}_0 : & \mathbf{x} \sim \mathcal{N}(0, R_A) \\ \mathcal{H}_1 : & \mathbf{x} \sim \mathcal{N}(0, R_B) \ . \end{array} \tag{8}$$

For instance, the hypothesis test in (8) also corresponds to the problem of anomalous change detection in multispectral imagery modeled by Gaussian distributions [7].

We can compute the generalized eigen-decomposition [8] that diagonalizes both R_A and R_B simultaneously, allowing us to decorrelate the vector $\mathbf{x}$ under both hypotheses using

$$\widetilde{\mathbf{x}} = \widetilde{E}_B^t \Lambda_A^{-1/2} E_A^t \mathbf{x} \ , \tag{9}$$

where E_A and Λ_A are the eigenvectors and eigenvalues [1] given by

$$R_A = E_A \Lambda_A E_A^t \ ,$$

and $\widetilde{\Lambda}_B$ and $\widetilde{E}_B$ are the eigenvalues and eigenvectors of the matrix $\widetilde{R}_B$ given by

$$\widetilde{R}_B \triangleq \Lambda_A^{-1/2} E_A^t R_B E_A \Lambda_A^{-1/2} = \widetilde{E}_B \widetilde{\Lambda}_B \widetilde{E}_B^t \ .$$

The linear transformation of (9) is equivalent to the Fisher linear discriminant (FLD) that is used to maximize the ratio of the between class to within class scatter [8], [9], [10].

In this new space, the hypotheses in (8) are written in terms of $\widetilde{\mathbf{x}}$ and become

$$\begin{array}{ll} \mathcal{H}_0 : & \widetilde{\mathbf{x}} \sim \mathcal{N}(0, I) \\ \mathcal{H}_1 : & \widetilde{\mathbf{x}} \sim \mathcal{N}(0, \widetilde{\Lambda}_B) \ . \end{array} \tag{10}$$

Since $\mathbf{x}$ and $\widetilde{\mathbf{x}}$ are related by an invertible linear transformation, the log-likelihood ratio of (5) can be shown to be

$$\begin{aligned} l(\mathrm{x}) &= \log \left\{ \frac{p(\mathrm{x}; \mathcal{H}_1)}{p(\mathrm{x}; \mathcal{H}_0)} \right\} && (11) \\ &= -\sum_{i=1}^{p} \left(\frac{1}{\tilde{\lambda}_{Bi}} - 1 \right) \tilde{x}_i^2 + \sum_{i=1}^{p} \log \tilde{\lambda}_{Bi} \ , && (12) \end{aligned}$$

[1] All eigenvalues in Λ_A are assumed here to be non-zero

where $\widetilde{\lambda}_{Bi}$ is the ith diagonal element of $\widetilde{\Lambda}_B$ and $\widetilde{x}_i$ is the ith coordinate of the vector $\widetilde{\mathbf{x}}$.

C. Classification

Let $\mathbf{y}_0$ and $\mathbf{y}_1, \cdots, \mathbf{y}_\mathcal{K}$ all be p-dimensional random vectors, and assume that each vector is formed by $\mathbf{y}_k = \mathbf{x}_k + \mathbf{w}_k$, where $\mathbf{x}_k \sim \mathcal{N}(0, R_x)$ is an unknown p-dimensional signal, and $\mathbf{w}_k \sim \mathcal{N}(0, R_w)$ is additive p-dimensional noise. Our objective is to classify the vector $\mathbf{y}_0$ as a member of the class $k \in \{1, \cdots, \mathcal{K}\}$ if the pair of vectors $\mathbf{y}_0$ and $\mathbf{y}_k$ constitute a match, i.e, they both originated from the same signal: $\mathbf{x}_0 = \mathbf{x}_k$. Therefore, under the hypothesis of a match, the difference $\Delta\mathbf{y}_k = \mathbf{y}_k - \mathbf{y}_0 \sim \mathcal{N}(0, 2R_w)$. Alternatively, under the hypothesis that $\mathbf{y}_0$ and $\mathbf{y}_k$ are *not* a match we have that $\Delta\mathbf{y}_k \sim \mathcal{N}(0, 2(R_x + R_w))$. In summary, the probability density of the random vector $\Delta\mathbf{y}_k$ is given by

$$\begin{array}{lll} \mathcal{H}_0: & \Delta\mathbf{y}_k \sim \mathcal{N}(0, 2(R_x + R_w)) & \text{if } \mathbf{x}_0 \neq \mathbf{x}_k \\ \mathcal{H}_1: & \Delta\mathbf{y}_k \sim \mathcal{N}(0, 2R_w) & \text{if } \mathbf{x}_0 = \mathbf{x}_k \ . \end{array} \tag{13}$$

The maximum likelihood selection of $\hat{k}$ is given by

$$\hat{k} = \arg\max_k \left\{ \log \left[\frac{p(\Delta\mathrm{y}_k; \mathcal{H}_1)}{p(\Delta\mathrm{y}_k; \mathcal{H}_0)} \right] \right\} . \tag{14}$$

Following the same lines of Section III-B, we can compute the generalized eigen-decomposition of both R_x and R_w, thus allowing the computation of $\Delta\widetilde{\mathbf{y}}_k$ from $\Delta\mathbf{y}_k$, which is decorrelated under both hypotheses. As a result, the hypotheses in (13) are equivalent to

$$\begin{array}{lll} \mathcal{H}_0: & \Delta\widetilde{\mathbf{y}}_k \sim \mathcal{N}(0, 2(\widetilde{\Lambda}_x + I)) & \text{if } \mathbf{x}_0 \neq \mathbf{x}_k \\ \mathcal{H}_1: & \Delta\widetilde{\mathbf{y}}_k \sim \mathcal{N}(0, 2I) & \text{if } \mathbf{x}_0 = \mathbf{x}_k \ . \end{array} \tag{15}$$

The selection of $\hat{k}$ in (14) can be written in terms of the coordinates of $\Delta\widetilde{\mathbf{y}}_k$ and the diagonal elements of $\widetilde{\Lambda}_x$, resulting in the expression

$$\begin{aligned} \hat{k} &= \arg\max_k \left\{ \log \left[\frac{p(\Delta\mathrm{y}_k; \mathcal{H}_1)}{p(\Delta\mathrm{y}_k; \mathcal{H}_0)} \right] \right\} \\ &= \arg\min_k \left\{ \sum_{i=1}^{p} \left(\frac{\widetilde{\lambda}_{xi}}{1 + \widetilde{\lambda}_{xi}} \right) \Delta\widetilde{y}_{ki}^2 \right\} , \end{aligned} \tag{16}$$

where $\widetilde{\lambda}_{xi}$ is the ith diagonal element of $\widetilde{\Lambda}_x$ and $\Delta\widetilde{y}_{ki}$ is the ith coordinate of the vector $\Delta\widetilde{\mathbf{y}}_k$.

D. Hypothesis Testing using SMT

In Section III-B, the generalized eigendecomposition of the covariance matrices R_A and $\widetilde{R}_B$ is a key step for the computation of the log-likelihood test (12). We use the SMT to perform the generalized eigendecomposition of $R_A = E_A \Lambda_A E_A^t$ and $\widetilde{R}_B = \widetilde{E}_B \widetilde{\Lambda}_B \widetilde{E}_B^t$, with r_1 and r_2 rotations per coordinate respectively. We apply the SMT for the computation of the following steps:

1) Compute $\mathbf{x}' = \Lambda_A^{-1/2} E_A^t \mathbf{x}$, requiring $(2r_1+1)p$ floating-point operations. At the end, we may choose to clip a fraction of the p dimensions and keep only αp of them, with $\alpha \in [0,1]$.
2) Compute $\widetilde{\mathbf{x}} = \widetilde{E}_B^t \mathbf{x}'$, requiring $(2r_2+1)\alpha p$ operations.
3) Compute the sum in (12) , equiring a total of $2\alpha p$ floating-point operations.

The steps above amount to a total of $[2(r_1+\alpha r_2)+3\alpha+1]p$, i.e, $O(p)$ floating-point operations, were $\alpha \in [0,1]$. These same steps are used to compute the generalized eigen-decomposition of R_x and R_w in Section III-C, and the log-likelihood ratio used in (16).

IV. Experimental Results

A. Face Recognition

The SMT classification developed in Section III-C is applied to the task of *face recognition*. We evaluate the SMT-based face recognition with the FERET test protocol and dataset [11], and compare it against the LDA face recognition method [10], a conceptually similar method but that relies on dimensionality reduction to handle the high-dimensional face data. We also compare with a regularized version of LDA using shrinkage covariance estimation. The FERET protocol splits the data into two *disjoint* sets: the *training* set, with face images of 221 individuals/ three different frontal images per individual, and the *gallery* set, with face images of 160 individuals/ four different frontal images per individual. After training the classifier with images of the *training* set, we simulate the recognition process by randomly picking one image from the *gallery* set and searching it against the whole gallery. The system returns all candidates sorted by the likelihood of being a match. If the searched individual appears among the top ρ likely matches in a fraction f of all the searches, we say the rank-ρ recognition rate is f.

Fig. 1 compares the recognition rates of several classifiers, each using a different method for covariance estimation. The SMT is used both as a standalone method for the covariance estimation, referred as SMT, as well as the shrinkage target, referred as S-SMT. Both SMT-based methods are compared with shrinkage toward identity (S-I) and the LDA face recognition method [10]. The Shrinkage/SMT (S-SMT) performs best among all compared methods. The SMT and Shrinkage/Identity (S-I) methods exhibit almost identical accuracies. Finally, all regularized methods compared are more accurate than the LDA.

As discussed in the Section III-D, the computational cost associated with the application of the SMT is $O(p)$, compared to $O(p^2)$ required to apply the S-SMT and the S-I methods. Therefore, the SMT can be deployed in an environment with limited computational resources delivering competitive accuracy to the one of the computationally expensive shrinkage estimation.

B. Hyperspectral Image Processing

We use hyperspectral data to measure the performance of the matched filter and the power detector described in Sections III-A and III-B respectively.

Fig. 2 shows the area under the ROC curve for the power detector presented in Section III-B using several methods. The true covariances R_A and R_B are known. In such scenario, the accuracy of the SMT-based method approaches the one of

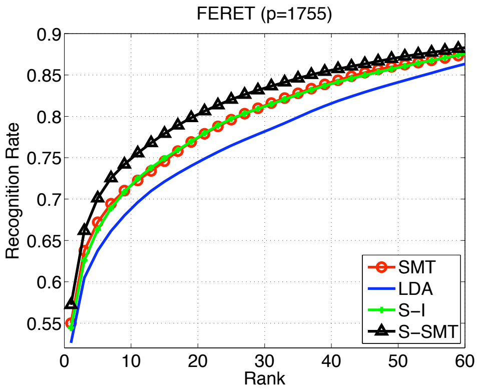

Fig. 1. Face recognition rates for ranks 1-60 using different classifiers, SMT, LDA, Shrinkage/Identity (S-I), and Shrinkage/SMT (S-SMT), trained with 221 individuals / 3 images per individual.

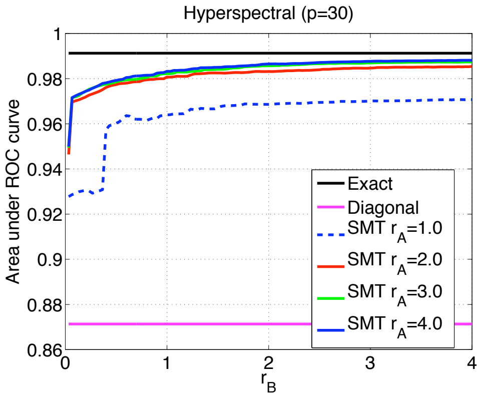

Fig. 2. Area under the ROC curve for the SMT as the number of Givens rotations varies. Only a few SMT's Givens rotations are necessary to get most of the detection accuracy given by the exact generalized eigen-decomposition of the true covariance matrices.

the exact generalized eigen-decomposition with only a small number of Givens rotations per coordinate.

Fig. IV-B shows the detection capability of the matched filter presented in Section III-A measured by the $SCRR = SCR/SCR_0$ statistic, where SCR_0 is the value of the ratio in (7) for the true covariance R. Therefore, normally we expect $SCRR < 1$. When $SCRR = 1$, the detection accuracy is equivalent of the one in the situation that the true covariance R of the clutter is known. We varied the number of training samples n used to estimate $\hat{R}$. The results are averages over 10 trials, each using a different signal t and n different training samples. Notice that the SMT-based detectors perform substantially better than the ones using shrinkage and sample covariance estimates when the training set is small.

V. Conclusions

We presented a framework for hypothesis testing in high-dimensional space using the SMT to model the covariance structure of the high-dimensional data. Results show that the SMT methods for detection and classification can have advantages over other methods in the following important aspects. First, the log likelihood ratio test remains robust when few training samples are available to train the covariance matrices involved. Second, it operates directly in high-dimensional data at a low computational cost. Finally, the SMT can be used to improve the accuracy of shrinkage estimation when it is computationally feasible.

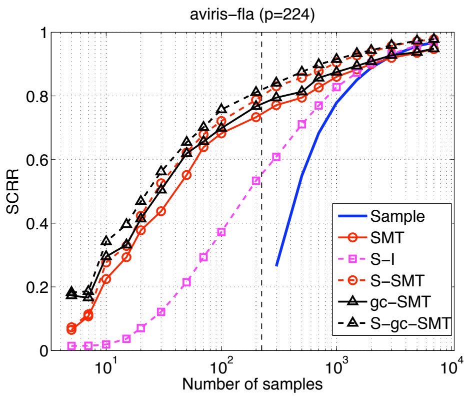

Fig. 3. SCRR for hyper-spectral image AVIRIS-FLA using several different estimators: Sample covariance, SMT, Shrinkage /Identity (S-I), Shrinkage/SMT (S-SMT), graphical-SMT (gc-SMT), and Shrinkage/graphical-SMT (S-gc-SMT). Average of 10 trials (each with different signal t and different set of n samples).

References

[1] S. M. Kay, *Fundamentals of Statistical Signal Processing, Vol.2: Detection Theory*, Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1998.

[2] D. L. Donoho, "High-dimensional data analysis: The curses and blessings of dimensionality," in *Math Challenges of the 21st Century*, Los Angeles, August 8 2000, American Mathematical Society.

[3] G. Cao and C. A. Bouman, "Covariance estimation for high dimensional data vectors using the sparse matrix transform," in *Advances in Neural Information Processing Systems*. 2008, MIT Press.

[4] L. R. Bachega, G. Cao, and C. A. Bouman, "Fast signal analysis and decomposition on graphs using the sparse matrix transform," in *Proceedings of the ICASSP*, Dallas, TX, 2010.

[5] J. Theiler, G. Cao, L. R. Bachega, and C. A. Bouman, "Sparse matrix transform for hyperspectral image processing," *Journal of Selected Topics in Signal Processing*, To appear in special issue on Advances in Remote Sensing Image Processing.

[6] G. Cao, C. A. Bouman, and J. Theiler, "Weak signal detection in hyperspectral imagery using sparse matrix transformation (SMT) covariance estimation," in *First Workshop on Hyperspectral Image and Signal Processing*, 2009.

[7] J. Theiler and S. Perkins, "Proposed framework for anomalous change detection," *ICML Workshop on Machine Learning Algorithms for Surveillance and Event Detection*, pp. 7–14, 2006.

[8] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification (2nd Edition)*, Wiley-Interscience, 2 edition, November 2000.

[9] 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, pp. 711–720, 1997.

[10] W. Zhao, A. Krishnaswamy, R. Chellappa, D. L. Swets, and J. Weng, "Discriminant analysis of principal components for face recognition," 1998.

[11] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 10, pp. 1090–1104, 2000.