PROCEEDINGS OF SPIE
SPIE—The International Society for Optical Engineering

Nonlinear Image Processing and Pattern Analysis XII

Edward R. Dougherty
Jaakko T. Astola
Chairs/Editors

22–23 January 2001
San Jose, USA

Sponsored by
IS&T—The Society for Imaging Science and Technology
SPIE—The International Society for Optical Engineering

SPIE
P
PROCEEDINGS SERIES

Volume 4304

PROCEEDINGS OF SPIE
SPIE—The International Society for Optical Engineering

Nonlinear Image Processing and Pattern Analysis XII

Edward R. Dougherty
Jaakko T. Astola
Chairs/Editors

22–23 January 2001
San Jose, USA

Sponsored by
IS&T—The Society for Imaging Science and Technology
SPIE—The International Society for Optical Engineering

Published by
SPIE—The International Society for Optical Engineering

Volume 4304

SPIE is an international technical society dedicated to advancing engineering and scientific applications of optical, photonic, imaging, electronic, and optoelectronic technologies.

The papers appearing in this book compose the proceedings of the technical conference cited on the cover and title page of this volume. They reflect the authors' opinions and are published as presented, in the interests of timely dissemination. Their inclusion in this publication does not necessarily constitute endorsement by the editors or by SPIE. Papers were selected by the conference program committee to be presented in oral or poster format, and were subject to review by volume editors or program committees.

Please use the following format to cite material from this book:

Author(s), "Title of paper," in *Nonlinear Image Processing and Pattern Analysis XII*, Edward R. Dougherty, Jaakko T. Astola, Editors, Proceedings of SPIE Vol. 4304, page numbers (2001).

ISSN 0277-786X
ISBN 0-8194-3982-7

Published by
SPIE—The International Society for Optical Engineering
P.O. Box 10, Bellingham, Washington 98227-0010 USA
Telephone 1 360/676-3290 (Pacific Time) • Fax 1 360/647-1445
http://www.spie.org/

Printed in the United States of America.

Conference Committee

Conference Chairs

Edward R. Dougherty, Texas A&M University (USA)
Jaakko T. Astola, Tampere University of Technology (Finland)

Program Committee

Sergey V. Ablameyko, Institute of Engineering Cybernetics (Belarus)
Sos S. Agaian, University of Texas at San Anotonio (USA)
Junior Barrera, Universidade de São Paulo (Brazil)
Yidong Chen, National Institutes of Health (USA)
Reiner Creutzburg, University of Applied Sciences (Germany) and Tampere University of Technology (Finland)
Paul D. Gader, University of Missouri/Columbia (USA)
Robert M. Haralick, University of Washington (USA)
Vladimir V. Lukin, National Aerospace University (Ukraine)
Françoise J. Prêteux, Institut National des Télécommunications (France)
Giovanni Ramponi, University of Trieste (Italy)
Akira Taguchi, Musashi Institute of Technology (Japan)
Arthur R. Weeks, University of Central Florida (USA)

Session Chairs

1 Pattern Recognition
Karen O. Egiazarian, Tampere University of Technology (Finland)

2 Applications
Sos S. Agaian, University of Texas at San Antonio (USA)
Peter D. Dolan, MIT Lincoln Laboratory (USA) and Tufts University (USA)

3 Filters
Akira Taguchi, Musashi Institute of Technology (Japan)
John C. Handley, Xerox Corporation (USA)

Contents

FILTERS

Nonlinear Filtering and Pattern Recognition: Are They the Same?

Edward R. Dougherty,[a] Junior Barrera[b]

[a]Department of Electrical Engineering, Texas A&M University
[b]Departamento de Ciencia de Computacao, Universidade de Sao Paulo

ABSTRACT

Statistical design of window-based nonlinear filters for signal and image processing is closely related to pattern recognition. The theory of pattern recognition is concerned with estimating the errors of optimal classifiers and with designing classifiers from sample data whose errors are close to minimal. Of special importance is the Vapnik-Chervonenkis theory, which relates the design cost to the VC dimension of a classification rule. This paper discusses both constraint and design costs for the design of nonlinear filters, and discusses the relation to the theory of pattern recognition. As to the question posed in the title, the paper argues that nonlinear filtering possesses its own integrity because classification rules and constraints depend on signal and image properties, both in theory and the manner in which expert knowledge is applied in design.

Keywords: optimal filters, pattern recognition, nonlinear filters, VC dimension

1. INTRODUCTION

During the last decade there has been a movement towards statistical design of nonlinear filters for signal and image processing. In particular, there has been continuing development of design techniques for translation-invariant window-based digital filters. To the extent that filter design involves finding optimal (close-to-optimal) functions of a finite number of observation random variables, it is connected to computational learning theory and pattern recognition [1, 2]. Owing to algorithm complexity, most early statistical design was confined to binary operators on fairly small windows. Recently there has been increasing attention given to larger windows and gray-scale operators. Various techniques are being proposed to overcome limits to both sample data and design time. The problem is especially acute for gray-scale images. Effort is focused on finding appropriate classes in which to constrain designed filters to mitigate the data requirement. Constraining the class from which to select a filter reduces the error of a designed filter owing to estimation from sample data, at the cost of having to search in a class not containing the optimal filter. For a bounded sample size, the problem is one of constraint error versus filter-estimation error. This problem has a long history in the theory of pattern recognition and has generated an extensive theory regarding the behavior of classifiers for increasing sample size [3].

This paper, which is a much shortened version of a paper on the relationship between nonlinear signal processing and pattern recognition [4], briefly discusses their commonality, in particular with regard to constraint and estimation costs for filter design. It also comments upon the way in which nonlinear signal processing, insofar as it concerns filter design, is an independent subject from pattern recognition theory. From a very abstract perspective, one might claim that it is not. Of course, from a sufficiently abstract perspective, pattern recognition is submerged in statistical estimation theory. This would miss the point. As for signal processing, its role is emphasized by the theory of pattern recognition itself. Filter design is not simply a question of data; rather, it depends on appropriate constraints, and these are functions of the signal classes and operators under consideration. Moreover, these constraints may be the result of expert knowledge concerning the signal processes. This is especially so in image processing, where visual understanding can play a key role in constraining filter design.

In signal (image) processing, a filter estimates an ideal random signal F from an observed random signal G. For a window-based filter, a d-point window W is placed at a point z, thereby determining a random vector $\mathbf{X}$ of G-values in W. A function ψ is applied to $\mathbf{X}$ to form an estimator $\psi(\mathbf{X})$ of the value $Y = F(z)$. The *W-operator* Ψ is defined by $\Psi(G)(z) = \psi(\mathbf{X})$. If F and G are jointly stationary, then ψ is independent of z. Filter design is a form of inverse problem. We consider F to be operated on by a random *system transformation* Γ to produce the observed signal $G = \Gamma(F)$. The problem is to find an optimal estimator $\Psi(G)$ for F. For a W-operator, the joint random signal process (F, G) induces a distribution on the $(d + 1)$-vector $(\mathbf{X}, Y)$. This induced distribution determines the optimal filter. Often, we know nothing about the distribution; however, in many cases there are signal properties which imply that the optimal filter belongs to some subclass of filters. If Γ is extensive, meaning that $G \geq F$, then Ψ must be antiextensive, meaning $\Psi(G) \leq G$. Generally, from either theory or experience, we might know that a constraint will provide (or likely provide) a good suboptimal filter.

Because a translation-invariant window-based filter is defined via a function ψ, filter design involves estimation of a random variable Y by a function ψ of a random vector $\mathbf{X} = (X_1, X_2,..., X_d)$. For nonlinear filtering, the commonly used error measure is the mean-absolute error (MAE), $\varepsilon[\psi] = E[|Y - \psi(\mathbf{X})|]$, where the expectation is with respect to the joint distribution of $(\mathbf{X}, Y)$. In the digital setting, each component vector X_k has a range $L = \{0, 1,..., l\}$, and Y has range $M = \{0, 1,..., m\}$. An optimal filter is defined by a function ψ_d whose error, ε_d, is minimum among all functions ψ: $L^d \rightarrow M$. In the special case when $M = \{0, 1\}$, so that filter output is binary, an optimal filter is given by the thresholded conditional expectation

$$\psi(\mathbf{X}) = \begin{cases} 0, \text{if } E[Y \mid \mathbf{X}] \leq 1/2 \\ 1, \text{if } E[Y \mid \mathbf{X}] > 1/2 \end{cases} \tag{1}$$

Filter design involves using a set S_n of sample pairs $(\mathbf{X}^{(1)}, Y^{(1)}), (\mathbf{X}^{(2)}, Y^{(2)}),..., (\mathbf{X}^{(n)}, Y^{(n)})$ to from an estimate ψ_n of ψ_d. The error, ε_n, of ψ_n cannot be less than ε_d owing to the optimality of ψ_d. Letting $\Delta_n = \varepsilon_n - \varepsilon_d$ denote the *design cost*, the error of the designed filter is decomposed as $\varepsilon_n = \varepsilon_d + \Delta_n$. Hence, the expected error of the designed filter is

$$E[\varepsilon_n] = \varepsilon_d + E[\Delta_n] \tag{2}$$

It is common in nonlinear digital signal processing to estimate the conditional probabilities determining the optimal filter and to use these estimates to determine ψ_n. This method yields a consistent estimate of ε_n: $E[\Delta_n] \rightarrow 0$ as $n \rightarrow \infty$. The essential problem for nonlinear filtering is that satisfactory filtering often requires large windows, especially for images, and it is often impossible to get large enough samples to sufficiently reduce $E[\Delta_n]$. To ease the design problem, optimization is constrained to some subclass $\mathcal{C}$ of filters. If ψ_C is an optimal filter in $\mathcal{C}$ with error ε_C and design error $\Delta_{n,C}$, then $\varepsilon_C \geq \varepsilon_d$ and $E[\Delta_{n,C}] \leq E[\Delta_n]$. The error of a designed constrained filter, $\psi_{n,C}$, possesses the decomposition $\varepsilon_{n,C} = \varepsilon_C + \Delta_{n,C}$. The *constraint cost* is $\Delta_C = \varepsilon_C - \varepsilon_d$. Hence, $\varepsilon_{n,C} = \varepsilon_d + \Delta_C + \Delta_{n,C}$, and

$$E[\varepsilon_{n,C}] = \varepsilon_d + \Delta_C + E[\Delta_{n,C}] \tag{3}$$

Constraint is statistically beneficial if and only if $E[\varepsilon_{n,C}] \leq E[\varepsilon_n]$, which is true if and only if

$$\Delta_C \leq E[\Delta_n] - E[\Delta_{n,C}] \tag{4}$$

The savings in design error must exceed the cost of constraint. Since $E[\Delta_n] - E[\Delta_{n,C}] \rightarrow 0$ as $n \rightarrow \infty$, a constraint can only be beneficial for samples that are not too large.

A fundamental problem of nonlinear digital signal processing is to find constraints for which Eq. 4 is satisfied. The benefit of a constraint is dependent upon the class of ideal and observed signals under consideration — equivalently, the class of ideal signals and system transformation Γ. $\mathcal{C}$ may be defined theoretically in accordance with knowledge of Γ, such as $\mathcal{C}$ consisting of antiextensive filters when Γ is extensive, or it may be that experience has shown that a certain constraint works well in a given setting. There is often no fine line between these two situations. The case when Γ is extensive is in some sense too easy. Statistically, there is no constraint, because the optimal filter lies in $\mathcal{C}$ and $\Delta_C = 0$.

2. PATTERN CLASSIFICATION

Henceforth, we consider binary filters in the context of pattern recognition. A binary filter ψ is called a *classifier* and serves as a decision function between two classes having *labels* 0 and 1. The input to the classifier is a random vector of real-valued *features* that are not independent of the binary label random variable Y. The mean-absolute error of the classifier reduces to $\varepsilon[\psi] = P(\psi(\mathbf{X}) \neq Y)$, the probability of erroneous classification. An optimal classifier and its error are called a *Bayes classifier* and the *Bayes error*, respectively. For d observed features, we denote the optimal classifier and its error by ψ_d and ε_d, respectively. An optimal classifier is given by Eq. 1.

Design of a classifier ψ_n from a set S_n of sample vector-label pairs requires a *classification rule*. In fact, the classification rule is really a sequence of classification rules depending on n. The Bayes error ε_d is estimated by the error ε_n of ψ_n. There is a *design cost* $\Delta_n = \varepsilon_n - \varepsilon_d$, ε_n and Δ_n being sample-dependent random variables. The expected design cost is $E[\Delta_n]$, the expectation being relative to S_n. The expected error of ψ_n is decomposed according to Eq. 2. A classification rule is said to be *consistent* for the distribution of $(\mathbf{X}, Y)$ if $E[\Delta_n] \to 0$ as $n \to \infty$. It is *universally consistent* if consistent for any distribution.

For constraint class $\mathcal{C}$, sample S_n, and classifier $\psi \in \mathcal{C}$, the *empirical error*, $\hat{\varepsilon}_n[\psi]$, for the sample is the fraction of times that $\psi(\mathbf{X}^i) \neq Y^i$ for $(\mathbf{X}^i, Y^i) \in S_n$. It is the error rate on the sample. Since S_n is a random sample for the distribution $(\mathbf{X}, Y)$, the empirical error estimates $P(\psi(\mathbf{X}) \neq Y)$. Since $\psi_{\mathcal{C}}$ is the classifier in $\mathcal{C}$ minimizing this probability, choosing the classifier with minimal empirical error is a reasonable classification rule for optimization relative to $\mathcal{C}$. We denote this classifier by $\hat{\psi}_{n,\mathcal{C}}$ and its error by $\hat{\varepsilon}_{n,\mathcal{C}}$. For the *empirical error rule*, the design error is $\hat{\Delta}_{n,\mathcal{C}} = \hat{\varepsilon}_{n,\mathcal{C}} - \varepsilon_{\mathcal{C}}$. In the decomposition of Eq. 3, $E[\varepsilon_{n,\mathcal{C}}]$ and $E[\Delta_{n,\mathcal{C}}]$ are replaced by $E[\hat{\varepsilon}_{n,\mathcal{C}}]$ and $E[\hat{\Delta}_{n,\mathcal{C}}]$, respectively. For $\mathcal{C}$ fixed, our concern is the convergence of $E[\hat{\Delta}_{n,\mathcal{C}}]$.

For a measurable set $A \subset \Re^d \times \{0, 1\}$, let $\nu[A] = P((\mathbf{X}, Y) \in A)$. Let $\nu_n[A]$ be the empirical measure of A based on the sample S_n, meaning that $\nu_n[A]$ is the fraction of the sample points that fall in A. Define

$$\hat{\Delta}_n(\mathcal{C}) = \sup_{A \in \mathcal{A}_{\mathcal{C}}} |\nu_n[A] - \nu[A]| \tag{5}$$

where

$$\mathcal{A}_{\mathcal{C}} = \{(\mathcal{K}[\psi] \times \{0\}) \cup (\mathcal{K}[\psi]^c \times \{1\}) : \psi \in \mathcal{C}\} \tag{6}$$

and $\mathcal{K}[\psi] = \{\mathbf{x}: \psi(\mathbf{x}) = 1\}$ is the *kernel* of ψ. It can be shown that

$$\hat{\Delta}_{n,\mathcal{C}} \leq 2\hat{\Delta}_n(\mathcal{C}) \tag{7}$$

and therefore

$$P(\hat{\Delta}_{n,\mathcal{C}} > \tau) \leq P(\hat{\Delta}_n(\mathcal{C}) > \tau/2) \tag{8}$$

Hence, to bound $P(\hat{\Delta}_{n,\mathcal{C}} > \tau)$, it is sufficient to bound $P(\hat{\Delta}_n(\mathcal{C}) > \tau/2)$.

Let $\mathcal{A}$ be any class of measurable sets in $\Re^d$. If $\{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n\}$ is a point set in $\Re^d$, let

$$N_{\mathcal{A}}(\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n) = |\{\{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n\} \cap A : A \in \mathcal{A}\}| \tag{9}$$

be the cardinality of the class of distinct subsets of $\{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n\}$ created by intersection with sets in $\mathcal{A}$. The n^{th} *shatter coefficient* of $\mathcal{A}$ is defined by

$$\xi(\mathcal{A}, n) = \max_{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n} N_{\mathcal{A}}(\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n) \tag{10}$$

It is possible that $\xi(\mathcal{A}, n) = 2^n$, Then $\mathcal{A}$ is said to *shatter* $\{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n\}$. If $\xi(\mathcal{A}, n) < 2^n$, then $\xi(\mathcal{A}, j) < 2^j$ for $j > n$. The shatter coefficient of a filter class $\mathcal{C}$ is defined by $\Xi(\mathcal{C}, n) = \xi(\mathcal{A}_{\mathcal{C}}, n)$.

The *Vapnik-Chervonenkis* theorem states that, for any filter class $\mathcal{C}$, sample size n, and $\tau > 0$,

$$P\left(\hat{\Delta}_n(\mathcal{C}) > \tau\right) \leq 8\Xi(\mathcal{C}, n)e^{-n\tau^2/8} \tag{11}$$

[5]. From Eq. 8,

$$P(\hat{\Delta}_{n,\mathcal{C}} > \tau) \leq P(\hat{\Delta}_n(\mathcal{C}) > \tau/2) \leq 8\Xi(\mathcal{C}, n)e^{-n\tau^2/32} \tag{12}$$

A corollary of the Vapnik-Chervonenkis theorem is that

$$E[\hat{\Delta}_{n,\mathcal{C}}] \leq 4\sqrt{\frac{\log(8e\Xi(\mathcal{C}, n))}{2n}} \tag{13}$$

The *Vapnik-Chervonenkis* (*VC*) *dimension*, V_C, of a class C is the largest integer n for which $\Xi(C, n) = 2^n$. If $\Xi(C, n) = 2^n$ for all n, then $V_C = \infty$. If its VC dimension is finite, then C is called a *VC class*. For $V_C > 2$, $\Xi(C, n) \leq n^{V_C}$. From Eqs. 12 and 13, for finite $V_C > 2$,

$$P(\hat{\Delta}_{n,C} > \tau) \leq 8n^{V_C} e^{-n\tau^2/32} \tag{14}$$

$$E[\hat{\Delta}_{n,C}] \leq 4\sqrt{\frac{V_C \log n + 4}{2n}} \tag{15}$$

While this bound is encouraging asymptotically, it shows the problem with trying to reduce the constraint cost for small samples. For instance, the VC dimension of a k-neuron two-layer neural network exceeds kd for k even and exceeds $(k-1)d$ for k odd. Increasing the number of neurons to improve the approximation capability of the network rapidly increases the sample size required to have an acceptable design cost.

3. SIGNAL PROCESSING AND PATTERN RECOGNITION

The Vapnik-Chervonenkis and related bounds have been applied to various classification rules in pattern recognition. They can also be applied to filters for signal processing. Consider an increasing binary operator $\psi: L^d \to \{0, 1\}$. The *basis*, $\mathcal{B}[\psi]$, of ψ is the set of minimal elements in $\mathcal{K}[\psi]$, and ψ possesses the supremum representation

$$\psi(\mathbf{x}) = \bigvee\{\varepsilon_{\mathbf{b}}(\mathbf{x}): \mathbf{b} \in \mathcal{B}[\psi]\} \tag{16}$$

where the erosion operator $\varepsilon_{\mathbf{b}}$ is defined by $\varepsilon_{\mathbf{b}}(\mathbf{x}) = 1$ if and only if $\mathbf{x} \geq \mathbf{b}$. It can be shown that the VC dimension of the class of m-erosion classifiers is bounded by d^m. From Eq. 14,

$$P(\hat{\Delta}_{n,C} > \tau) \leq 8n^{d^m} e^{-n\tau^2/32} \tag{17}$$

for all $\tau > 0$.

Constraint can be imposed from theoretical considerations, but it also can be imposed based on expert knowledge. Envelope constraint involves two humanly designed filters, α and β, such that $\alpha \leq \beta$, and a designed filter ψ must lie in the envelope determined by α and β, meaning that $\alpha \leq \psi \leq \beta$. C is the class of all filters ψ such that $\alpha \leq \psi \leq \beta$. Envelope design is a form of hybrid human-machine filter design [6]. If the human designed envelope contains the optimal filter, then $\Delta_C = 0$; if not, then $\Delta_C > 0$ and the constraint is beneficial if and only if Eq. 4 is satisfied. For expert human design, envelope constraint is often beneficial even for relatively large samples. The classification rule used here is $\psi_n = (\phi_n \vee \alpha) \wedge \beta$, where ϕ_n is designed from estimates of the conditional probabilities. For any estimate filter ϕ_n, the two sides of Eq. 4 are given by

$$\begin{aligned} E[\Delta_n] - E[\Delta_{n,C}] = & \sum_{\mathbf{x} \notin \mathcal{K}[\psi_d] \cup \mathcal{K}[\beta]} \delta_{\mathbf{x}} P(\mathbf{x} \in \mathcal{K}[\phi_n]) + \sum_{\mathbf{x} \in \mathcal{K}[\psi_d] \cap \mathcal{K}[\alpha]} \delta_{\mathbf{x}} P(\mathbf{x} \notin \mathcal{K}[\phi_n]) \\ & + \sum_{\mathbf{x} \in \mathcal{K}[\psi_d] - \mathcal{K}[\beta]} \delta_{\mathbf{x}} P(\mathbf{x} \notin \mathcal{K}[\phi_n]) + \sum_{\mathbf{x} \in \mathcal{K}[\alpha] - \mathcal{K}[\psi_d]} \delta_{\mathbf{x}} P(\mathbf{x} \in \mathcal{K}[\phi_n]) \end{aligned} \tag{18}$$

$$\Delta_C = \sum_{\mathbf{x} \notin \mathcal{K}[\psi_d] - \mathcal{K}[\beta]} \delta_{\mathbf{x}} + \sum_{\mathbf{x} \notin \mathcal{K}[\alpha] - \mathcal{K}[\psi_d]} \delta_{\mathbf{x}} \tag{19}$$

Since ϕ_n is a consistent estimator, the kernel relations show that $E[\Delta_n] - E[\Delta_{n,C}] \to 0$ as $n \to \infty$.

A important issue in pattern recognition is feature selection. The idea is to try to find features that are nonredundant but still provide small classification error. Using less features reduces the VC dimension of the problem, but it can also mean increasing the Bayes error. If we restrict ourselves to a subset of the possible features, this can be viewed as constraining the classifier to a the class of functions that do not have the left-out variables as arguments. Thus, feature selection is really an issue of constraint cost.

More generally, one can consider transforming input variables. Design cost can be reduced by transforming **X** to a random vector of lower dimension. This involves a transformation $\mathbf{Z} = \zeta(\mathbf{X})$. The designed filter depends on a classification rule being applied to the transformed sample. If the dimension of **Z** is $a < d$, then two Bayes classifiers can be considered, ψ_d and ψ_a, for (**X**, Y) and (**Z**, Y), respectively. There is also the filter $\psi_a\zeta$ relative to **X**. A filter, ϕ_n, is designed relative to the sample for (**Z**, Y), and the resulting designed filter relative to (**X**, Y) is $\psi_n = \phi_n\zeta$. This filter is likely to be suboptimal relative to ψ_d on (**X**, Y). If so, then estimation is not ncessarily consistent relative to (**X**, Y), even if it is consistent relative to (**Z**, Y). In pattern recognition, dimension reduction via an orthogonal transformation φ represents this kind of reduction. φ is applied to **X** to provide representation of **X** in terms of orthogonal component vectors. Some components of $\varphi(\mathbf{X})$ are deleted to form **Z**. Letting ω denote the transformation that deletes the components, $\zeta = \omega\varphi$.

In signal processing, multiresolution design provides dimension reduction by designing a filter at a low resolution with the intention being that there is sufficient error reduction to offset the loss of information [7]. Let $\mathcal{D}_0$ denote the space of observation vectors for W, and Let $P = \{\mathcal{D}_{01}, \mathcal{D}_{02}, \ldots, \mathcal{D}_{0n}\}$ be a partition of $\mathcal{D}_0$. An equivalence relation on $\mathcal{D}_0$ is defined by P: $\mathbf{x} \sim \mathbf{x}'$ if and only if **x** and **x'** lie in the same subset of P. Let $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n$ be distinct binary vectors, let $\mathcal{D}_1 = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n\}$, and assign the correspondence $\mathcal{D}_{0k} \leftrightarrow \mathbf{z}_k$. $\mathcal{D}_1$ forms a new, lower resolution space. There is a *resolution mapping* ρ: $\mathcal{D}_0 \rightarrow \mathcal{D}_1$ defined by $\rho(\mathbf{x}) = \mathbf{z}_k$ if $\mathbf{x} \in \mathcal{D}_{0k}$, and the partition is defined via the inverse sets of ρ, namely, $\rho^{-1}(\mathbf{z}_k) = \mathcal{D}_{0k}$. Each $\mathbf{z} \in \mathcal{D}_1$ has a corresponding equivalence class $\mathcal{C}[\mathbf{z}] = \rho^{-1}(\mathbf{z})$. $\mathcal{D}_0$ is a discrete probability space with probability mass $P(\mathbf{x})$. Optimal filtering is relative to the product space $\{0, 1\} \times \mathcal{D}_0$ with probability mass $P(y, \mathbf{x})$. Probability masses are induced on $\mathcal{D}_1$ by $P(\mathbf{z}) = P(\rho^{-1}(\mathbf{z}))$ and on $\{0, 1\} \times \mathcal{D}_1$ by $P(y, \mathbf{z}) = P(\{y\} \times \rho^{-1}(\mathbf{z}))$. An operator ϕ on $\mathcal{D}_1$ induces an operator ψ_ϕ on $\mathcal{D}_0$ by $\psi_\phi(\mathbf{x}) = \phi(\rho(\mathbf{x}))$. ψ_ϕ is *resolution constrained* in accordance with the partition P: if $\mathbf{x} \sim \mathbf{x}'$, then $\psi_\phi(\mathbf{x}) = \psi_\phi(\mathbf{x}')$. Conversely, if ψ is any operator on $\mathcal{D}_0$ satisfying the resolution constraint, then it induces an operator ϕ_ψ on $\mathcal{D}_1$ by $\phi_\psi(\mathbf{z}) = \psi(\mathbf{x})$, where **x** is any vector in $\mathcal{C}[\mathbf{z}]$. Let $\mathcal{O}_0$ and $\mathcal{O}_1$ denote the classes of binary operators on $\mathcal{D}_0$ and $\mathcal{D}_1$, respectively. The mapping $\psi \rightarrow \phi_\psi$ defines a surjection $\mathcal{O}_0 \rightarrow \mathcal{O}_1$ and the mapping $\phi \rightarrow \psi_\phi$ defines an injection $\mathcal{O}_1 \rightarrow \mathcal{O}_0$. If $\mathcal{R}_0$ is the subset of $\mathcal{O}_0$ composed of operators satisfying the resolution constraint and $\psi \in \mathcal{R}_0$, then $\psi \rightarrow \phi_\psi$ defines a bijection $\mathcal{R}_0 \rightarrow \mathcal{O}_1$ whose inverse is given by $\phi \rightarrow \psi_\phi$. This bijective relation allows identification of operators on $\mathcal{D}_1$ with resolution constrained operators on $\mathcal{D}_0$. This identification is MAE preserving: the MAE as an estimator of Y of an operator ψ on $\mathcal{D}_0$ satisfying the resolution constraint is equal to the MAE of ϕ_ψ as an estimator of Y: $\varepsilon[\phi_\psi] = \varepsilon[\psi]$. Thus, the optimal filter on $\mathcal{D}_1$ induces the optimal resolution-constrained filter on $\mathcal{D}_0$, and optimization can take place on $\mathcal{D}_1$. Error expressions have been derived for multiresolution design, and the entire matter can be extended to several iterations of resolution reduction. The partitioning of $\mathcal{D}_0$ need not occur according to a partitioning of W.

4. CONCLUSION

The general probabilistic theory of pattern recognition applies to nonlinear signal processing. But it does not subsume the need to utilize signal theory, operator theory, or expert knowledge in the design of filters. In fact, it shows quite the opposite. The need for constraints, classification rules, and prior knowledge in filter design is clarified by pattern recognition theory. All of these involve signal and image processing proper. Moreover, while a great deal of attention has been paid in pattern recognition theory to asymptotic results that apply to very large samples, it is clear that much practical application depends on samples that are relatively small in comparison to the number of potential features. While there are general formulas for constraint and design costs in some cases, these tend to be rather abstract and need to be improved by use of appropriate signal models. Moreover, very little is known about the VC dimensions of important filter classes for image and signal processing.

5. REFERENCES

1. J. Barrera, E. R. Dougherty, and N. S. Tomita, "Automatic Programming of Binary Morphological Machines by Design of Statistically Optimal Operators in the Context of Computational Learning Theory," *Electronic Imaging*, 6 (1), 54-67, 1997.
2. E. R. Dougherty, and J. Barrera, "Logical Image Operators," in *Nonlinear Filters for Image Processing*, eds. E. R. Dougherty and J. T. Astola.

3. L. Devroye, L. Gyorfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, Springer-Verlag, New York, 1996.
4. E. R. Dougherty, and J. Barrera, "Pattern Recognition Theory in Nonlinear Signal Processing," in preparation.
5. V. Vapnik, and A. Chervonenkis, "On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities," *Theory of Probability and its Applications*, **16**, 264-280, 1971.
6. J. Barrera, E. R. Dougherty, and M. Brun, "Hybrid Human-Machine Binary Morphological Operator Design: An Independent Constraint Approach," *Signal Processing*, **80** (8), 1469-1487, 2000.
7. E. R. Dougherty, J. Barrera, G. Mozelle, S. Kim, and M. Brun, "Multiresolution Analysis for Optimal Binary Filters," to appear in *Mathematical Imaging and Vision*, **14** (1), 2001.

Color-invariant shape moments for object recognition

Qiang Zhou and Mehmet Celenk*

School of Electrical Engineering and Computer Science
Stocker Center, Ohio University, Athens, OH 45701

ABSTRACT

Geometric moments have been widely used in many shape recognition and object classification tasks. These monomials are usually computed from binary or gray-level images for the object shape recognition invariant to rotation, translation, and scaling. In this paper, we attempt to calculate the shape related moments from color images, and study their noise immunity and color invariance property for the application areas of face recognition and content-based image retrieval. To this end, we describe a computationally efficient method for converting a vector-valued color image into a gray scale for robust moment computation. Geometric moments are calculated from the resultant scalar representation of a color image data, and proven to be robust shape descriptors for the face and flower images. The generated shape invariants appear to have better noise immunity than the Hu moments and exhibit characteristics invariant to hue changes in the object colors. As compared to the Zernike polynomials, the proposed feature set has higher discriminatory power although the Zernike polynomials present superior noise rejection capability. Robust performance, computational efficiency, high noise immunity, and hue invariance property of the new approach are particularly useful for fast image retrieval tasks requiring high query accuracy.

Keywords: Shape recognition, geometric moments, color covariance, noise immunity, hue invariance

1. INTRODUCTION

Geometric moments provide rich information about the image, and are widely used in many shape recognition and object classification tasks. Their information content stems from the fact that moments provide an equivalent representation of an image in the sense that an image can be reconstructed from its moments of all orders.[1] Thus each moment coefficient conveys a certain amount of information about an image content. The seven moments of Hu[2] are invariant to the object rotation, translation, and scaling.[7] A number of other moment-based features that are invariant to more general transformations have also been proposed.[3-5,8] An analysis by Liao and Pawlak[6] shows that the approximation error for a digital image moment calculation increases with the coarseness of the sampling grid as well as with the order of the moments. The geometric moments can also be viewed as projections of the gray level scale on the basis functions formed by the monomials. Since these monomials are not orthogonal, the resulting geometric moment features are not optimal in the information redundancy sense. Alternatively, moments can also be derived from the orthogonal set of Zernike polynomials over the unit circle. The magnitude of the Zernike moments is invariant to rotation,[1] and translation and scaling invariance is treated by Khotanzad and Hong.[9] Computational aspects of the Zernike moments are examined by Mukundan and Ramakrshnan[10] and the accuracy issue is studied by Liao and Pawlak.[11] Comparative studies of the performance of the Zernike moments against the moments of Hu for character recognition have demonstrated that the former behave better in noisy environments.[12] Zernike moments are used to alleviate geometry and illumination invariance.[13] Pseudo-Zernike moments have been also proposed and used.[14,15]

Shape descriptors of these moments are usually computed from binary or gray-level images and then converted into a set of features so that a class of objects can be recognized invariant to their rotation, translation, and scaling. In this paper, we attempt to calculate the shape related moments directly from color images, and study their noise immunity and color invariance property for the application areas of face recognition and content-based image retrieval. To this end, we describe a computationally efficient method in section 2 and a computational procedure in section 3. The main objective here is to convert a vector-valued color image into a

* Correspondence: Email: celenk@bobcat.ent.ohiou.edu; Telephone: 740-593-1581; Fax: 740-593-0007

scalar-function f for robust moment computation. In turn, a set of geometric moments computed from the resultant feature (f) generates a set of object shape descriptors, which are robust to object and background noise as discussed in section 4. In section 5, for a selected set of images involving human faces and flowers, we demonstrate experimentally that the extracted geometric moments are highly immune to the sensory noise and also invariant to changes in hue of the object colors considered in the experiments. At the end of the paper, we present conclusion and further research topics.

2. DESCRIPTION OF THE OVERALL APPROACH

The proposed algorithm considers a color image in the (R, G, B) tristimulus representation. First it computes the color covariance matrix of the whole image as the second-order statistical approximation of the image color distribution. The given picture is then divided into 8x8 non-overlapping blocks of local areas in the spatial plane. For each 8x8 block, we calculate the 3x1 mean color vector and 3x3 color covariance matrix. Difference between the color covariance matrices of two neighboring blocks is computed so that the resultant difference matrix becomes independent of hue variations in the input scene. Magnitude of the mean color vector of each block is scaled by multiplying it with the trace of the corresponding difference covariance matrix and the trace of the color covariance matrix of the whole image. Each block is then represented by its scaled mean color vector magnitude. This helps reducing the size of the excessive color data and capturing the hue invariant color features of the object in question. The first seven geometric moments of Hu and the first three Zernike moments are computed using the scaled mean magnitudes of the 8x8 blocks. In this step, the object shape information is implicitly incorporated with the object's invariant spectral attributes.

3. MOMENT GENERATION ALGORITHM

The moment generation algorithm developed in this research can be summarized in the following steps of a computational procedure:

Pre-Processing: Calculate the mean intensity of R, G, and B channels of the four corner blocks in an image as an estimate for the background gray-level. If the estimated mean value is sufficiently high, then replace the image with its negative; i.e., for each color channel we replace every individual pixel value with 255 minus the original value.

Step 1: Compute the 3x3 covariance matrix C for the whole image:

$$C = \frac{1}{N \cdot M}\begin{bmatrix} \sum(R_i - \overline{\mu}_R)^2 & \sum(R_i - \overline{\mu}_R)(G_i - \overline{\mu}_G) & \sum(R_i - \overline{\mu}_R)(B_i - \overline{\mu}_B) \\ \sum(G_i - \overline{\mu}_G)(R_i - \overline{\mu}_R) & \sum(G_i - \overline{\mu}_G)^2 & \sum(G_i - \overline{\mu}_G)(B_i - \overline{\mu}_B) \\ \sum(B_i - \overline{\mu}_B)(R_i - \overline{\mu}_R) & \sum(B_i - \overline{\mu}_B)(G_i - \overline{\mu}_G) & \sum(B_i - \overline{\mu}_B)^2 \end{bmatrix} \tag{3.1}$$

Here, N and M are the height and width of the image, and μ_R, μ_G, μ_B are the mean values of three color channels R, G, and B defined by $\overline{\mu}_R = \frac{1}{N \cdot M}\sum R_i$, $\overline{\mu}_G = \frac{1}{N \cdot M}\sum G_i$, and

$\overline{\mu}_G = \frac{1}{N \cdot M}\sum B_i$.

Step 2: Divide the image into blocks of 8x8 sub-images, and for each block compute the corresponding covariance matrix C_b.

Step 3: For each 8x8 block, define a 3x3 color-covariance difference matrix D as $D = C - C_b$.

Step 4: Generate a feature value (f) for each 8x8 block based on the trace of the difference matrix D computed in step 3:

If ($\overline{\mu}_R^2 + \overline{\mu}_G^2 + \overline{\mu}_B^2$)>T, then

$$f = \sqrt{\mu_R^2 + \mu_G^2 + \mu_B^2} \cdot Trace(D) \qquad (3.2)$$

else

$$f = \sqrt{(\mu_R - \overline{\mu}_R)^2 + (\mu_G - \overline{\mu}_G)^2 + (\mu_B - \overline{\mu}_B)^2} \cdot Trace(D) \qquad (3.3)$$

where μ_R, μ_G, and μ_B are the red, green, and blue mean values of the block.

Step 5: Normalize f linearly by mapping the minimum value to 0 and the maximum value to 255. Using the normalized image function f, compute the (Hu and Zernike) moments for the whole image.

4. NOISE INVARIANT MOMENT GENERATION

In this section, we elaborate on the noise immunity and robustness of the moments generated by the algorithm described in the previous section. For this, we consider 0-mean Gaussian noise with variance σ^2 and denote it as $e \sim N(0, \sigma^2)$. Let us assume that e be a signal independent additive noise affecting the color signals R_i, G_i, B_i of the i-th pixel in the block in question in the same manner.

Considering the 0-mean Gaussian noise model, the feature value f of equations (3.2) and (3.3) can be subject to noise corruption only through the term Trace(D). Notice that the first term of the product for f in both of these equations is related to the magnitude of the mean color vector of a block or the whole image. Since the mean is equal to its true value plus the mean of the Gaussian noise, it does not have any contribution to the noise model of f. As a result, the rest of this section will be devoted to the discussion of the noise effect on the covariance matrix. A detailed mathematical analysis is carried out first and a new feature is derived subsequently. It is shown that the derived feature has the "best" noise resistance that can be generated from covariance matrix based method of section 3.

4.1. Noise Analysis of the Moment Generation Algorithm

The signal-independent noise model can be represented in an analytical form for the red channel as

$$YR_i = R_i + e_i^R \qquad (4.1)$$

Here, YR_i is the observed R-value at the i-th pixel in a block of consideration, R_i is the true value of the red signal, and e_i^R is the noise. The noise is assumed to be zero-mean white noise $e_i^R \sim N(0, \sigma^2)$. The same definition applies to the green (G) and blue (B) channels. For the sake of simplicity, the analysis will be conducted only for the red signal and the same argument will hold for the other two signals.

In accordance with the linear noise model of equation (4.1), the covariance matrix of an image block becomes

$$Con = \frac{1}{n}\begin{bmatrix} \sum(YR_i - \mu_{YR})^2 & \sum(YR_i - \mu_{YR})(YG_i - \mu_{YG}) & \sum(YR_i - \mu_{YR})(YB_i - \mu_{YB}) \\ \sum(YG_i - \mu_{YG})(YR_i - \mu_{YR}) & \sum(YG_i - \mu_{YG})^2 & \sum(YG_i - \mu_{YG})(YB_i - \mu_{YB}) \\ \sum(YB_i - \mu_{YB})(YR_i - \mu_{YR}) & \sum(YB_i - \mu_{YB})(YG_i - \mu_{YG}) & \sum(YB_i - \mu_{YB})^2 \end{bmatrix}$$

(4.2)

Here, μ_{YR} is the mean observed value of the R channel in the block. The same definition applies to the channels G and B.

Lemma 1: *Every element of the covariance matrix* **Con** *can be decomposed into a linear combination of three terms; a noise free term, a term including the Gaussian noise, and a term involving χ^2 noise.*

Proof: Consider $\frac{\sum(YR_i - \mu_{YR})^2}{n}$.

$$\frac{\sum(YR_i - \mu_{YR})^2}{n} = \frac{\sum YR_i^2 - 2\sum YR_i\mu_{YR} + \sum \mu_{YR}^2}{n} = \frac{\sum YR_i^2 - 2n\mu_{YR}^2 + n\mu_{YR}^2}{n} = \frac{\sum YR_i^2 - n\mu_{YR}^2}{n} \quad (4.3)$$

$$\mu_{YR} = \frac{\sum YR_i}{n} = \frac{\sum R_i + \sum e_i^R}{n} = \mu_R + \xi \quad (4.4)$$

where $\xi_R = \frac{\sum e_i^R}{n} \sim N\left(0, \frac{\sigma^2}{n}\right)$

Following equation (4.3), we have

$$\frac{\sum YR_i^2 - n\mu_{YR}^2}{n} = \frac{\sum(R_i + e_i^R)^2 - n(\mu_R + \xi_R)^2}{n}$$

$$= \frac{\left(\sum R_i^2 - n\mu_R^2\right)}{n} + 2\frac{\sum(R_i - \mu_R)e_i^R}{n} + \left(\frac{\sum(e_i^R)^2}{n} - \xi_R^2\right)$$

Define $A = \frac{\left(\sum R_i^2 - n\mu_R^2\right)}{n}$ (4.5)

$$B = \frac{\sum(R_i - \mu_R)e_i^R}{n} = \frac{\sum \Delta R_i e_i^R}{n} \quad (4.6)$$

and

$$C = \left(\frac{\sum(e_i^R)^2}{n} - \xi_R^2\right) \quad (4.7)$$

It is clear that A is the noise free term, B is the term including Gaussian noise, and C is the term involving χ^2 noise. The same results hold for the other elements of the covariance matrix **con**.