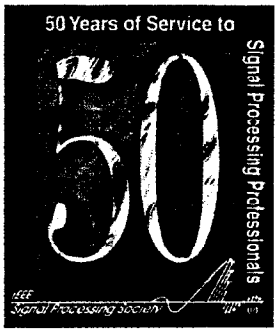
IEEE TRANSACTIONS ON

SIGNAL PROCESSING

A PUBLICATION OF THE IEEE SIGNAL PROCESSING SOCIETY

DECEMBER 1997 VOLUME 45 NUMBER 12 ITPRED (ISSN 1053–587X)



(Contents Continued From Front Cover)

Weak Continuity with Structural Constraints

N. D. Sidiropoulos, J. S. Baras, and C. A. Berenstein

Abstract—**Nonlinear regression and nonlinear regularization are two powerful approaches to segmentation and nonlinear filtering. In this correspondence, we propose a hybrid approach that effectively combines the best of both worlds and can be efficiently implemented via the Viterbi algorithm.**

I. Introduction

Edge detection and its dual problem of segmentation are important in low-level vision [1]. One may choose from a number of possible approaches, including statistical formulations, usually based on Markov models [2]–[10], classic nonlinear filters, e.g., the median, coupled with postfiltering detection [11]–[17], nonlinear regression [18]–[21], nonlinear regularization, e.g., [1], [22]–[25], among others.

Nonlinear regularization admits a Bayesian-Markovian interpretation; nonlinear regression admits a constrained maximum likelihood interpretation (although both were conceived starting from nonstatistical perspectives). Both have unique strengths and some drawbacks. The purpose of this correspondence is to propose and investigate a hybrid nonlinear regression-regularization approach, effectively combining the best of both worlds.

The idea of combining nonlinear regression and regularization is related in spirit to the idea of combining deterministic (rule-based) and statistical prior knowledge about a source one is trying to estimate; cf. the important work of Grenander *et al.* [26], [27].

A. Organization

The rest of this correspondence is organized as follows. In Section II, we review some important background. Our hybrid approach is introduced in Section III; an important result concerning idempotence of the proposed hybrid filter and, therefore, existence of and convergence to root signals is also presented in this section. A specific instance of our hybrid approach is presented in Section IV, which includes two useful design-oriented results and a detailed illustrative simulation experiment, highlighting the features of the proposed hybrid approach and the prior art. Conclusions are drawn in Section V.

II. Background

A. Nonlinear Regularization

For reference purposes, let us define regularization as the following general problem:

Problem 1—**Regularization:** Given $\mathbf{y} = \{y(n)\}_{n=0}^{N-1}$, find $\hat{\mathbf{x}} = \{\hat{x}(n)\}_{n=0}^{N-1}$ to minimize $d(\mathbf{y}, \mathbf{x}) + \mathbf{g}(\mathbf{x})$. Usually, $d(\mathbf{y}, \mathbf{x}) = \sum_{n=0}^{N-1} d_n(y(n), x(n))$.

Note that the term *nonlinear* regularization has to do with whether or not the solution to the above optimization problem is a linear function of its input $\mathbf{y}$; nonlinear regularizing functionals (e.g., quadratic) $\mathbf{g}(\cdot)$ may well lead to a linear solution.

Weak continuity (WC), which was developed by Mumford and Shah [22], [23] and Blake and Zisserman [1] (see also Morel and Solimini [24]), is, in a sense, the next logical step beyond Tikhonov regularization. WC attempts to fit *piecewise*-smooth candidate interpretations to the observable data (thus, the term *weak* continuity).

Since, in practice, we often deal with digital data, i.e., sequences of finite-alphabet variables, in order to avoid unnecessary complication, we present a digital version of discrete-time WC (following Blake and Zisserman [1]).

Problem 2—**Weak Continuity:** Given a (generally real-valued) sequence of finite extent $\mathbf{y} = \{y(n)\}_{n=0}^{N-1} \in \mathbf{R}^N$, find a finite-alphabet sequence $\hat{\mathbf{x}} = \{\hat{x}(n)\}_{n=0}^{N-1} \in \mathcal{A}^N$ (the *reproduction process*; usually, $\mathcal{A}$ is, e.g., $\{0, 1, \ldots, 255\}$) and a sequence of boolean edge markers $\hat{\mathbf{e}} = \{\hat{e}(n)\}_{n=1}^{N-1} \in \{0, 1\}^{N-1}$ (the *edge process*) so that the following cost is minimized.

$$\mathcal{V}_{WC}(\mathbf{y}, \mathbf{x}, \mathbf{e}) = \sum_{n=0}^{N-1} (y(n) - x(n))^2 + \sum_{n=1}^{N-1} \left[\lambda_{WC}^2 (x(n) - x(n-1))^2 (1 - e(n)) + \alpha e(n)\right].$$

Here, α is a nonnegative real.

Manuscript received October 15, 1995; revised May 13, 1997. This work was supported in part by core funds from the NSF ERC program, made available through the Communications and Signal Processing Group of the Institute for Systems Research of the University of Maryland, and industry through the Martin-Marietta Chair in Systems Engineering funds. The associate editor coordinating the review of this paper and approving it for publication was Prof. Peter C. Doerschuk.

N. D. Sidiropoulos was with the Institute for Systems Research and the Department of Electrical Engineering, University of Maryland, College Park, MD 20742 USA. He is now with the Department of Electrical Engineering, University of Virginia, Charlottesville, VA 22903 USA (e-mail: nikos@virginia.edu).

J. S. Baras is with the Institute for Systems Research and the Department of Electrical Engineering, University of Maryland, College Park, MD 20742 USA.

C. A. Berenstein is with the Department of Mathematics and the Institute for Systems Research, University of Maryland, College Park, MD 20742 USA.

Publisher Item Identifier S 1053-587X(97)08555-3.

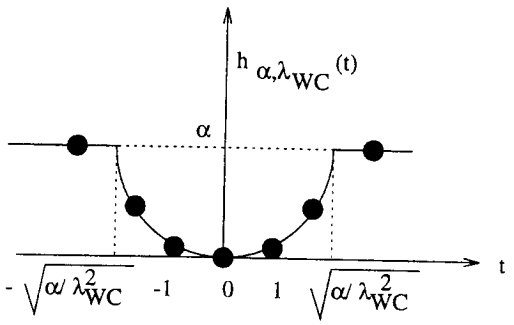


Fig. 1. WC regularizing functional.

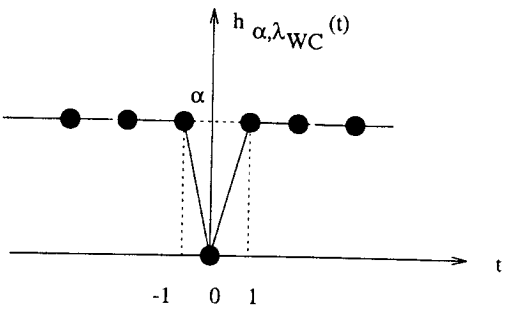


Fig. 2. WC reg. functional: case of $\lambda_{\mathrm{WC}}^2 > \alpha$.

If $(x(n)-x(n-1))^2$ is too large, one has the option of declaring an edge in between $x(n)$ and $x(n-1)$ by choosing $e(n)=1$, and thus paying only α, instead of $\lambda_{\mathrm{WC}}^2(x(n)-x(n-1))^2$. One can first minimize with respect to the edge process and then minimize the resulting functional with respect to the reproduction process. Since the first sum in the combined cost does not depend on the edge process, it is easy to see [1, pp. 43, 112–114] that the optimization above is equivalent to minimizing

$$\mathcal{V}'_{\mathrm{WC}}(\mathbf{y},\mathbf{x}) = \sum_{n=0}^{N-1}(y(n)-x(n))^2 + \sum_{n=1}^{N-1} h_{\alpha,\lambda_{\mathrm{WC}}}(x(n)-x(n-1))$$

by appropriate choice of reproduction process $\mathbf{x}$, where $h_{\alpha,\lambda_{\mathrm{WC}}}: \mathbf{Z} \mapsto \mathbf{R}$ is defined as

$$h_{\alpha,\lambda_{\mathrm{WC}}}(t) = \begin{cases} \lambda_{\mathrm{WC}}^2 t^2, & t^2 < \frac{\alpha}{\lambda_{\mathrm{WC}}^2} \\ \alpha, & \text{otherwise} \end{cases}$$

This is depicted in Fig. 1. The associated optimal edge process can be implicitly inferred, once the optimal reproduction process is determined, by level tests on the first-order residuals $\hat{x}(n)-\hat{x}(n-1)$ of the optimal reproduction process.

From the form of $\mathcal{V}'_{\mathrm{WC}}$, one may readily see why WC is the next logical step beyond Tikhonov regularization: WC replaces the quadratic regularizer in Tikhonov regularization with a hard-limited quadratic. In general, classical optimization techniques, like steepest descent, are not applicable to nonconvex problems like WC [28], even if these problems only involve continuous variables; this is due to the existence of local minima [28].

There exist essentially two ways to go about solving WC: dynamic programming (DP) [29] and the so-called *graduated nonconvexity* (GNC) algorithm [1]. The GNC is suitable for optimization over $\hat{\mathbf{x}} \in \mathbf{R}^N$, i.e., the continuous-valued case, and it does not lend itself to discrete-valued problems, i.e., $\hat{\mathbf{x}} \in \mathcal{A}^N$ [28]. There are two DP algorithms for WC: one that works by DP over "time" (in a manner very similar to the Viterbi algorithm [30]–[32]) and *requires* $\mathbf{x}$ to be quantized [33] and another that works by DP over "edges" [28] and works for either continuous or discrete-valued $\mathbf{x}$, i.e., either $\hat{\mathbf{x}} \in \mathbf{R}^N$ or $\hat{\mathbf{x}} \in \mathcal{A}^N$. The latter is much slower than the former for moderate $|\mathcal{A}|$. Here, we consider the discrete-valued problem and opt for the former; throughout, we use DP over "time" to solve WC in $O(|\mathcal{A}|^2 N)$. DP is exact, i.e., it provides a true minimizer; GNC has been proven to do so for a large class of inputs [1] but not for an arbitrary input. The drawback of DP is that it does not easily generalize in higher dimensions (however, cf. [34]). The GNC, in comparison, carries over quite effortlessly in higher dimensions. The GNC is a special case of *mean field annealing* [35].

As mentioned earlier, WC and the GNC can be interpreted from a Bayesian estimation viewpoint; they are closely related to *maximum a posteriori* (MAP) inference for Markov models and associated *annealing*-type algorithms [2]–[10], [35].

A related optimization has been advocated by Leclerc [36] (also cf. [7]) based on the *minimum description length* (MDL) principle of Rissanen [37]. The MDL principle can be related to an instance of the MAP principle with a certain suitable choice of prior. In Leclerc's formulation, one seeks to minimize

$$\mathcal{V}_{\mathcal{MDL}}(\mathbf{y},\mathbf{x}) = \sum_{n=0}^{N-1}\frac{(y(n)-x(n))^2}{\sigma^2} + \sum_{n=1}^{N-1}\lambda_{\mathrm{MDL}}[1-\delta(x(n)-x(n-1))]$$

by appropriate choice of reproduction process $\mathbf{x}$, where δ is the Kronecker delta function, and σ^2 is noise variance. Here, $\lambda_{\mathrm{MDL}} \geq 0$. We should note that this cost is only an approximation of the MDL objective function obtained under certain assumptions. MDL, in general, need not take this form.

Leclerc pointed out that in the case of one-dimensional (1-D) data, one can readily figure out a DP program to minimize $\mathcal{V}_{\mathcal{MDL}}$ and provided a GNC-like algorithm for two-dimensional (2-D) data.

Both WC and MDL seek to minimize a cost of the following general form.

$$\mathcal{V}(\mathbf{y},\mathbf{x}) = \sum_{n=0}^{N-1} d_n(y(n),x(n)) + \sum_{n=1}^{N-1} g_n(x(n),x(n-1)).$$

In case $\mathbf{x} \in \mathcal{A}^N$, $|\mathcal{A}| < \infty$, Leclerc's MDL formulation is a special case of WC. Indeed, if λ_{WC} is sufficiently large (i.e., $\lambda_{\mathrm{WC}}^2 > \alpha$), then t being an integer, $h_{\alpha,\lambda_{\mathrm{WC}}}(t) = \alpha[1-\delta(t)]$. This is depicted in Fig. 2. If, in addition, $\alpha = \lambda_{\mathrm{MDL}}\sigma^2$, then WC reduces to Leclerc's MDL approach.

Both WC, and Leclerc's MDL approach are powerful and meritorious paradigms; however, in the context of edge detection in the presence of impulsive noise, both exhibit a shortcoming; they are susceptible to noise-induced outliers[1] that are locally inconsistent with the data. Consider an input consisting of a single Kronecker delta of height Δ. If $(\frac{\Delta}{\sigma})^2 > 2\lambda_{\mathrm{MDL}}$, then Leclerc's MDL approach will preserve this delta; similarly, if $\Delta^2 > \frac{\alpha}{\lambda_{\mathrm{WC}}^2}$ and $\Delta^2 > 2\alpha$, then WC will also preserve it. Thus, for each given choice of respective optimization parameter(s), one can find a sufficiently large Δ so that both WC and Leclerc's MDL approach will preserve outliers of magnitude $\geq \Delta$.

WC and MDL are susceptible to these outliers because they both stipulate a model that classifies powerful outliers as information-bearing signals. In the context of segmentation, this means that outliers are segmented as separate regions (which can later be merged with other more significant regions). However, in the context of edge detection in the presence of strong impulsive noise interference, this

[1] In the digital world, there is really no such thing as an impulse; a better substitute term would be *outlier*, or *outlying burst*.

behavior is undesirable; these outliers are usually associated with the noise rather than the signal.

Of course, there is no universal agreement on what constitutes an edge and what constitutes an outlier, and we will certainly steer clear of offering a suggestion. Even though defining an edge or an outlier can be a delicate and potentially troublesome task, defining *what distinguishes an edge from an outlier* is arguably easier. The following axiom adopts a simple and intuitive viewpoint.

True edges in the data should be consistent in the sense that they should manifest themselves as jump level changes in between two locally approximately flat regions of sufficient breadth, and this is what distinguishes an edge from an outlier.

This leads to nonlinear regression ideas.

B. Nonlinear Regression

Nonlinear regression exploits prior knowledge about the signal and the noise by picking a solution (estimate) from a *characteristic set*, $\mathcal{C}$ of candidate solutions compatible with given prior knowledge about the signal with the goal of minimizing a noise-induced distortion measure between the solution and the observation; see the following problem.

Problem 3—**Nonlinear Regression:**

$$\text{minimize: } \sum_{n=0}^{N-1} d_n(y(n), x(n))$$
$$\text{subject to: } \mathbf{x} = \{x(n)\}_{n=0}^{N-1} \in \mathcal{C}.$$

Nonlinear regression may be interpreted as a generalized projection or as constrained maximum likelihood, provided that the noise sequence can be assumed to be independent $d_n(y(n), x(n)) = d_n(y(n) - x(n))$ and equal to minus the logarithm of the noise marginal at time n evaluated at $y(n) - x(n)$.

Observe that if $d_n(\cdot, \cdot)$ is a distance for all n (and even under milder conditions [20], [21]), then the *root set* (or *domain of invariance*, which is the class of signals that are invariant under the regression) of nonlinear regression is precisely the characteristic set of the regression. This kind of precise control over the root set is certainly appealing, as is the closest nonlinear filtering analog[2] to controling a linear filter's passband. Observe that this type of control is not, in general, available in nonlinear regularization approaches, like WC, whose input–output analysis is very difficult [1], [24]. One may work out results that exclude certain signals from the root set of WC, and we will do this in the sequel. A full characterization of root signal structure for WC appears to be very difficult, and this difficulty carries over, in part, to our proposed hybrid regression-regularization approach.

Specific instances of nonlinear regression can be found in [18]–[21]. These include the following problem.

Problem 4—**VORCA Filtering** *[20]:*

$$\text{minimize: } \sum_{n=0}^{N-1} d_n(y(n), x(n))$$
$$\text{subject to } \mathbf{x} = \{x(n)\}_{n=0}^{N-1} \in P_M^N$$

where P_M^N is the set of all sequences of N elements of $\mathcal{A}$ that are piecewise constant of plateau (run) length $\geq M$. This regression explicitly formalizes the axiom that edges should be consistent in the sense of exhibiting sufficient breadth in both directions. This regression can be efficiently implemented via the Viterbi algorithm in time $O((|\mathcal{A}|^2 + |\mathcal{A}|(M-1))N)$ [20].

[2]Although the concept of a nonlinear filter's root signal set is far less powerful than the concept of passband for linear filters because the principle of superposition does not hold.

Locally monotonic regression [18], [19] is another example. This regression is the optimal counterpart of iterated median filtering. It involves the concept of *local monotonicity*, which we need to define. Local monotonicity is a property of sequences that appears in the study of the set of root signals of the median filter [11], [12], [14]–[17]; it constraints the roughness of a signal by limiting the rate at which the signal undergoes changes of trend (increasing to decreasing or vice versa).

Let $\mathbf{x}$ be a real-valued sequence (string) of length N, and let γ be any integer less than or equal to N. A *segment* of $\mathbf{x}$ of length γ is any substring of γ consecutive components of $\mathbf{x}$. Let $\mathbf{x}_i^{i+\gamma-1} = \{x(i), \ldots, x(i+\gamma-1)\}$, $i \geq 0, i+\gamma \leq N$ be any such segment. $\mathbf{x}_i^{i+\gamma-1}$ is monotonic if either $x(i) \leq x(i+1) \leq \cdots \leq x(i+\gamma-1)$ or $x(i) \geq x(i+1) \geq \cdots \geq x(i+\gamma-1)$.

Definition 1: A real-valued sequence $\mathbf{x}$ of length N is *locally monotonic* of degree $\alpha \leq N$ (or *lomo-*α or, simply, *lomo*, in case α is understood) if each and every one of its segments of length α is monotonic.

Throughout the following, we assume that $3 \leq \alpha \leq N$. A sequence $\mathbf{x}$ is said to exhibit an increasing (resp. decreasing) transition at coordinate i if $x(i) < x(i+1)$ (resp. $x(i) > x(i+1)$). If $\mathbf{x}$ is locally monotonic of degree α, then $\mathbf{x}$ has a constant segment (run of identical symbols) of length at least $\alpha - 1$ in between an increasing and a decreasing transition; the reverse is also true [11], [18]. If $3 \leq \alpha \leq \beta \leq N$, then a sequence of length N that is lomo-β is lomo-α as well; thus, the *lomotonicity* of a sequence is defined as the highest degree of local monotonicity that it possesses [18].

In the 1-D finite-data case, iterations of median filtering are known to converge, regardless of the original input (modulo some pathological cases) to a locally monotonic signal of lomo-degree related to the size of the median window and resembling the original input. However, this resemblance cannot be quantified, and, in general, the result of iterated median filtering is not the best (e.g., in the l_1, or l_2 sense) locally monotonic approximation of the original input signal. This gave rise to the idea of locally monotonic regression, which was proposed by Restrepo and Bovik [18]. They developed an elegant mathematical framework in which they studied locally monotonic regressions in $\mathbf{R}^N$. The problem was that their regression algorithms entailed a computational complexity that was exponential in N (the size of the sample). Motivated by this observation, and the fact that median filtering of digital signals always results in digital signals, Sidiropoulos proposed the following problem.

Problem 5—**Digital Locally Monotonic Regression:** [21]

$$\text{minimize: } \sum_{n=0}^{N-1} d_n(y(n), x(n))$$
$$\text{subject to: } \mathbf{x} = \{x(n)\}_{n=0}^{N-1} \in \Lambda(\alpha, N, \mathcal{A})$$

where $\Lambda(\alpha, N, \mathcal{A})$ is the set of all sequences of N elements of $\mathcal{A}$, which are locally monotonic of lomo-degree α [21].

This latter problem can be efficiently solved via the Viterbi algorithm in time $O(|\mathcal{A}|^2 \alpha N)$, i.e., *linear* in N [21].

Both approaches are robust in the sense of suppressing impulsive noise while preserving salient edge signals. However, neither take into account *edge strength*, i.e., the magnitude of jump level changes. This often results in undesirable ripple in the solution, and it happens exactly because pure nonlinear regression does not *explicitly* account for roughness/complexity, i.e., unlike WC, it does not incorporate a roughness/complexity penalty into the cost function: As long as a solution remains within the characteristic set of the regression, it may follow relatively insignificant input features.

III. Weak Continuity with Structural Constraints

We have seen that nonlinear regression, by virtue of its reliance on hard structural constraints, is robust in the presence of outliers, yet it may trace relatively insignificant edge features. On the other hand, nonlinear regularization (and WC in particular) ranks the importance of edge features by means of their significance in terms of the incurred approximation error [1], yet it does not exhibit the same degree of robustness in the presence of outliers. It appears quite natural, then, to endow WC with improved robustness by proposing the following hybrid optimization.

Problem 6—**Weak Continuity with Structural Constraints (WCSC):**

$$\text{min: } \mathcal{V}(\mathbf{y},\mathbf{x}) = \sum_{n=0}^{N-1} d_n(y(n),x(n)) + \sum_{n=1}^{N-1} g_n(x(n),x(n-1))$$

$$\text{subject to: } \mathbf{x} \in \mathcal{C}$$

where $\mathcal{C}$ is the set of all sequences of N elements of $\mathcal{A}$ satisfying some local hard structural constraint. Here again, $d(\mathbf{x},\mathbf{y}) = \sum_{n=0}^{N-1} d_n(y(n),x(n))$ is a fidelity measure, and $g(\mathbf{x}) = \sum_{n=1}^{N-1} g_n(x(n),x(n-1))$ is a roughness-complexity measure.

When $\mathcal{C} = P_M^N$, *runlength-constrained weak continuity* (RC-WC) results; similarly, if $\mathcal{C} = \Lambda(\alpha, N, \mathcal{A})$, then *locally monotonic weak continuity (LM-WC)* results. VORCA is a special case of RC-WC, and so is WC, MDL. Digital locally monotonic regression is a special case of LM-WC, and so is WC, MDL.

It should be noted that the incorporation of hard structural constraints is not the only way to handle outliers in the context of nonlinear regularization, e.g., cf. [9].

It is not difficult to see that RC-WC and LM-WC can be solved using exactly the same resources and computational structures as VORCA and digital locally monotonic regression, respectively. The extension to weak continuity (i.e., the incorporation of the first-order roughness-complexity measure $g(\mathbf{x}) = \sum_{n=1}^{N-1} g_n(x(n),x(n-1))$ into the cost functional) essentially comes for free; we skip the details and refer to [20] and [21]. One basically has to set up a suitable Viterbi trellis and specify the cost of one-step state transitions. The resulting complexity of RC-WC, LM-WC is $O((|\mathcal{A}|^2 + |\mathcal{A}|(M-1))N)$, $O(|\mathcal{A}|^2\alpha N)$, respectively. Observe that these algorithms work for *any* choice of fidelity and roughness-complexity measures of the above general form. It should be noted that one could consider roughness-complexity measures of order higher than one. Yet, this entails a significant increase in computational complexity of the resulting Trellis-type implementation. For this reason, we chose to work with first-order roughness-complexity measures.

A. Existence of and Convergence to the WCSC Root Set

Observe that the WCSC problem above always has a solution, albeit not necessarily a unique one.[3] We have the following important characterization theorem.

Theorem 1: If $d(\cdot,\cdot)$ is a distance metric[4] and we resolve ties by selecting a solution of least roughness-complexity,[5] then WCSC is an idempotent operation, i.e., it converges to an element of its root set in just one application. This is true for all characteristic sets $\mathcal{C}$ and, therefore, also for pure WC.

Proof: Consider an arbitrary input $\mathbf{y}$, and let $\hat{\mathbf{x}}$ be a corresponding WCSC solution computed in accordance with the tie-breaking strategy in the statement of the theorem. In addition, let $\tilde{\mathbf{x}}$ be a solution to the WCSC problem *for input* $\hat{\mathbf{x}}$. Suppose that $\tilde{\mathbf{x}}$, $\hat{\mathbf{x}}$ are distinct. Clearly, both $\hat{\mathbf{x}}$ and $\tilde{\mathbf{x}}$ are necessarily in $\mathcal{C}$. Therefore, from optimality of $\hat{\mathbf{x}}$ for input $\mathbf{y}$ over $\mathcal{C}$, it follows that

$$d(\mathbf{y},\hat{\mathbf{x}}) + g(\hat{\mathbf{x}}) \le d(\mathbf{y},\tilde{\mathbf{x}}) + g(\tilde{\mathbf{x}}). \tag{1}$$

On the other hand, from optimality of $\tilde{\mathbf{x}}$ for input $\hat{\mathbf{x}}$ over $\mathcal{C}$, it follows that

$$d(\hat{\mathbf{x}},\tilde{\mathbf{x}}) + g(\tilde{\mathbf{x}}) \le d(\hat{\mathbf{x}},\hat{\mathbf{x}}) + g(\hat{\mathbf{x}})$$

or, since $d(\cdot,\cdot)$ is a distance metric

$$d(\hat{\mathbf{x}},\tilde{\mathbf{x}}) + g(\tilde{\mathbf{x}}) \le g(\hat{\mathbf{x}}). \tag{2}$$

Add $d(\mathbf{y},\hat{\mathbf{x}})$ to both sides of this inequality to obtain

$$d(\mathbf{y},\hat{\mathbf{x}}) + d(\hat{\mathbf{x}},\tilde{\mathbf{x}}) + g(\tilde{\mathbf{x}}) \le d(\mathbf{y},\hat{\mathbf{x}}) + g(\hat{\mathbf{x}}).$$

By the triangle inequality, we have that $d(\mathbf{y},\tilde{\mathbf{x}}) \le d(\mathbf{y},\hat{\mathbf{x}}) + d(\hat{\mathbf{x}},\tilde{\mathbf{x}})$; it then follows that

$$d(\mathbf{y},\tilde{\mathbf{x}}) + g(\tilde{\mathbf{x}}) \le d(\mathbf{y},\hat{\mathbf{x}}) + g(\hat{\mathbf{x}}). \tag{3}$$

From (1) and (3), it follows that

$$d(\mathbf{y},\tilde{\mathbf{x}}) + g(\tilde{\mathbf{x}}) = d(\mathbf{y},\hat{\mathbf{x}}) + g(\hat{\mathbf{x}})$$

i.e., there exists a tie between $\tilde{\mathbf{x}}$ and $\hat{\mathbf{x}}$ for input $\mathbf{y}$. Given the tie-breaking strategy in the statement of the theorem, it follows that

$$g(\hat{\mathbf{x}}) \le g(\tilde{\mathbf{x}}) \tag{4}$$

since $\hat{\mathbf{x}}$ is *a* least roughness-complexity solution for input $\mathbf{y}$ over $\mathcal{C}$. However, (2), combined with the fact that $d(\cdot,\cdot)$ is a distance metric, and the assumption that $\tilde{\mathbf{x}}$, $\hat{\mathbf{x}}$ are distinct, implies that

$$g(\tilde{\mathbf{x}}) < g(\hat{\mathbf{x}}). \tag{5}$$

Inequalities (4) and (5) constitute a contradiction; it follows that the hypothesis that $\tilde{\mathbf{x}}$, $\hat{\mathbf{x}}$ are distinct is false. This deduction works for arbitrary $\mathbf{y}$; the proof is therefore complete. □

This is a very useful result because it demonstrates that provided distortion is a distance metric, the root set of WCSC is well defined, and, in fact, one application of WCSC is sufficient for convergence to a root signal, regardless of choice of roughness-complexity measure $g(\cdot)$ and characteristic set $\mathcal{C}$. This is a highly desirable property, both from a theoretical and from a practical viewpoint [38].

What is the root set of WCSC? It is obvious that (provided distortion is a distance) the root set of WCSC is a subset of its characteristic set $\mathcal{C}$. Actually, it is possible to show that the root set is usually a *proper* subset of $\mathcal{C}$. We will provide some results in this direction in the following section, although, in general, a complete root signal analysis of WCSC appears to be very hard. Still, knowing that the root set is a subset of $\mathcal{C}$ is better than what we can currently say about pure WC.

B. Design

Given the general WCSC formulation above, one needs to choose d_n, g_n, and the characteristic set $\mathcal{C}$ for a particular problem in hand. From a Bayesian perspective, the formulation above is tantamount to MAP estimation of a signal $\mathbf{x}$ in additive noise, provided that $d_n(y(n),x(n)) = d_n(y(n)-x(n))$, the noise sequence can be assumed to be independent (and independent of the signal) with

[3] Nonuniqueness usually complicates the analysis of optimization problems, e.g., cf. [20] and the proof of the next theorem.

[4] This is usually the case in practice.

[5] This is easy to implement in a trellis computation by keeping track of the roughness-complexity measure accrued so far by partial solutions using an auxiliary state variable by virtue of the fact that roughness-complexity is a sum of state transition costs.

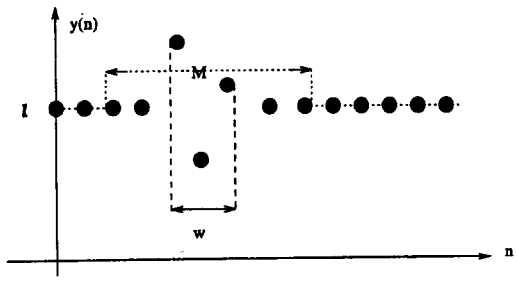


Fig. 3. Isolated outlying burst of width w.

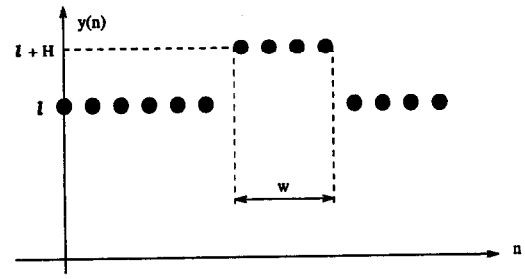


Fig. 4. Isolated constant segment of saliency $\mu = w \cdot H$.

marginal at time n given by $e^{-d_n(\cdot)}$, and the signal prior is $e^{-g(\mathbf{x})} = e^{-\sum_{n=1}^{N-1} g_n(x(n),x(n-1))}$ over $\mathcal{C}$, and zero elsewhere. Therefore, at least in principle, d_n, g_n, and $\mathcal{C}$ can be estimated from training data.

The choice of d_n is relatively easier; e.g., $d_n(y(n) - x(n))$ proportional to $|y(n) - x(n)|$ means one expects to be dealing with Laplacian (long-tailed) noise. The choice of g_n and $\mathcal{C}$ is far more critical as it constitutes the *signal model*, which is usually much harder to infer from limited training data. It is for this reason that for the purposes of segmentation and nonlinear filtering, we choose to restrict $\mathcal{C}$ to be P_M^N or $\Lambda(\alpha, N, \mathcal{A})$, which have been proven to be useful characteristic sets from a pure regression viewpoint, and g_n to λ^2 times a WC-type hardlimited notch function. This suggests a useful class of signal models that is not apparent from a Bayesian perspective and reduces the choice of signal model down to selecting two parameters.[6]

With these choices, what remains to be investigated is the interplay between M or α and λ^2. We know that at least for some specific choices, e.g., $M = 1$, leading to WC, MDL, or $\lambda^2 = 0$, leading to VORCA, or digital locally monotonic regression, we may expect good nonlinear filtering results. The point is, can we make even better choices? To see this, let us consider a concrete instance of RC-WC.

IV. A Specific Instance of RC-WC

Let $d_n(y(n), x(n)) = |y(n) - x(n)|$, $g_n(x(n), x(n-1)) = \lambda^2[1 - \delta(x(n) - x(n-1))]$ for all n, and let $\mathcal{C} = P_M^N$, i.e., consider

$$\text{minimize: } \mathcal{V}(\mathbf{y}, \mathbf{x}) = \sum_{n=0}^{N-1} |y(n) - x(n)| + \lambda^2 \sum_{n=1}^{N-1} [1 - \delta(x(n) - x(n-1))]$$

$$\text{subject to: } \mathbf{x} \in P_M^N.$$

We will need the following definitions.

Definition 2: An *isolated outlying burst of width* $w < M$ is a deviation from a plateau of the type depicted in Fig. 3.

Definition 3: An *isolated profile of saliency (sum of absolute deviations, which here is simply the width-strength product)* $\mu = w \cdot H$ is an equidistant deviation from a plateau of the type depicted in Fig. 4.

In the strict sense, isolated means that the entire input consists of the given feature; in practice, it means that the given feature is away from possible interactions with other input features.[7] The following two claims provide guidelines on how to choose M, λ^2. These claims apply to *this particular instance of RC-WC.*

[6]Note that other classes of signal models have been investigated in the context of MRF's, e.g., cf. [4] and references therein.

[7]This type of analysis of isolated features is typical of WC, and it is necessitated by analytical difficulties in dealing with potential interactions, e.g., cf. [1, pp. 58, 67, 100, 143, and 215].

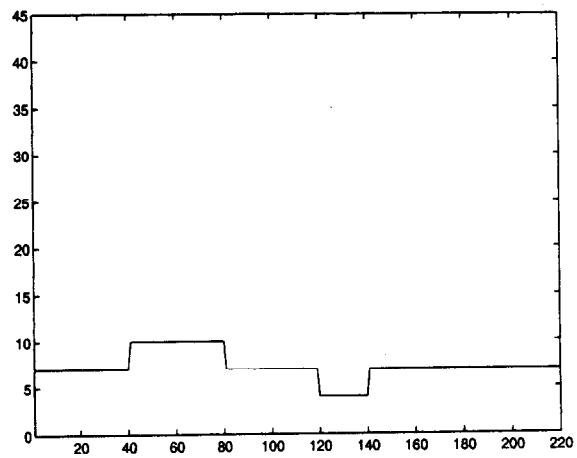


Fig. 5. Noise-free test data.

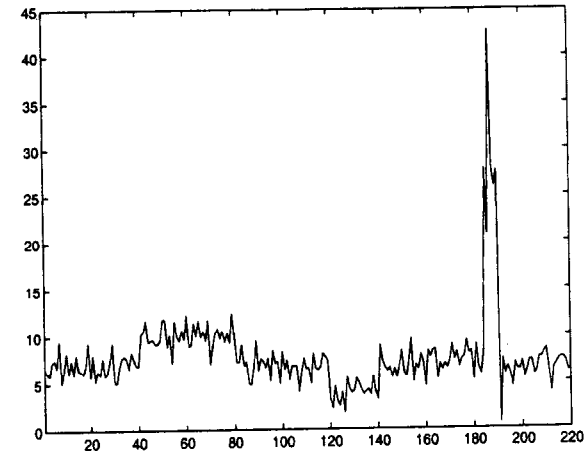


Fig. 6. Noisy input data.

Claim 1: Assume that M is odd. RC-WC eliminates all isolated outlying bursts of width $w \leq \frac{M-1}{2}$, regardless of λ^2, and the same is true for $\lambda^2 = 0$, i.e., plain VORCA filtering with respect to the above choice of $d_n(\cdot, \cdot)$.

Proof: With reference to Fig. 3, since $w < M$, the next best candidate (modulo a shift that is irrelevant here) after just drawing a straight line at the plateau level would be one consisting of just three constant segments, the middle of which is of width M, as shown with a dotted line in Fig. 3. This is because any two-segment solution would incur a cost that can be made as large as one wishes (this is where the assumption that one deals with *isolated* features comes into play). Now, the level of this middle segment should be chosen optimally to minimize the sum of absolute errors. This amounts to constant regression over M symbols under a least absolute error

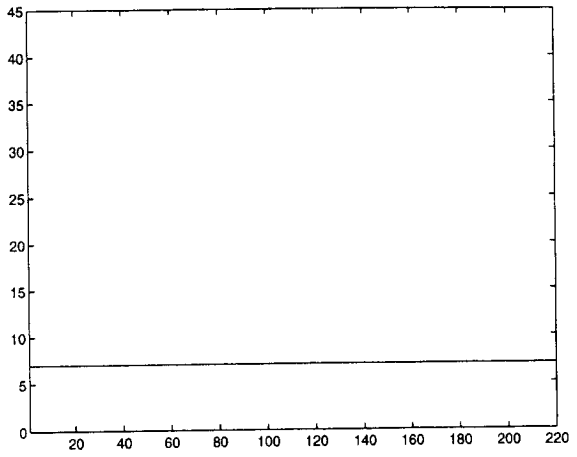


Fig. 7. Output of WC, $\lambda^2 = 55$.

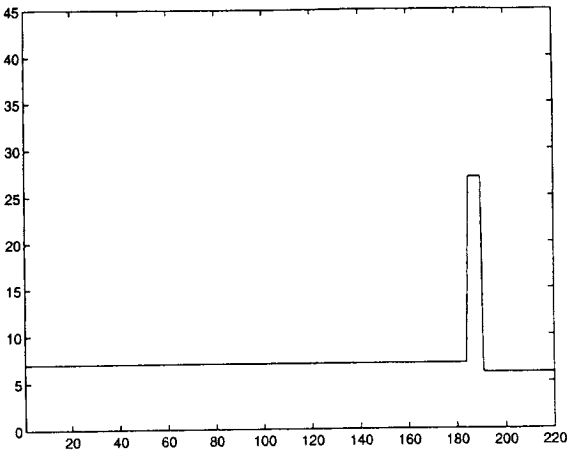


Fig. 8. Output of WC, $\lambda^2 = 50$.

criterion, and it is well known [18] that the answer is provided by the median of these M symbols. However, since only $w \leq \frac{M-1}{2}$ of these M symbols are potentially different from the plateau level (l in Fig. 3), it follows that the absolute majority of these M symbols is equal to the plateau level, and, therefore, the median produces this level at its output: The best solution amounts to simply drawing a straight line at the plateau level. □

Claim 2: RC-WC suppresses all isolated profiles of saliency (width-strength product) $\mu = w \cdot H < 2\lambda^2$, i.e., mends the weak edges at the endpoints of such profiles, and the same holds for $M = 1$, i.e., plain WC with respect to the above choice of $d_n(\cdot,\cdot), g_n(\cdot,\cdot)$.

Proof: In reference to Fig. 4, the next best candidate after just drawing a straight line at the plateau level would be (if allowed by the runlength constraint) one consisting of just three constant segments exactly following the input in Fig. 4 (again, this is where the assumption that one deals with isolated features comes into play). Such a candidate would incur a cost of at least $2\lambda^2$, whereas the straight line solution carries a cost of $\mu = w \cdot H < 2\lambda^2$. □

The overall conclusion is that this particular instance of RC-WC suppresses features of either i) width $w \leq \frac{M-1}{2}$ (M: odd), *regardless of strength*, or ii) saliency $\mu = w \cdot H < 2\lambda^2$. This allows us to essentially separately fine tune two important aspects of filter behavior. Given an estimate of maximum outlying burst duration, we pick M to eliminate outlying bursts. Given that we desire to suppress insignificant profiles producing spurious weak edges, where significance is quantified by profile saliency, we pick λ^2. In a nutshell, M controls outlier rejection, whereas λ^2 controls residual ripple.

A. An Illustrative Simulation Experiment

Fig. 6 depicts a noisy input sequence. This input has been generated by adding noise on synthetic piecewise-constant data, which is depicted in Fig. 5. The noise is white Gaussian; a simulated error burst has also been added to test outlier rejection capability. The outlying burst in Fig. 6 has length 6 and saliency (here, sum of absolute burst errors) 120. The noiseless signal in Fig. 5 consists of two rectangular pulses. The first has length 40 and saliency 120; the second has length 20 and saliency 60.

In reality, one rarely has a precise noise model available, and practitioners will opt for, e.g., tried-and-true l_2 or l_1 distance metrics, depending on whether the noise appears to be closer to Gaussian (short-tailed) or Laplacian (long-tailed), respectively. If the noise appears to be mixed (as is the case here due to the simulated outlying burst), this choice is not obvious. We chose the l_1 metric because it provides for improved outlier rejection, although this

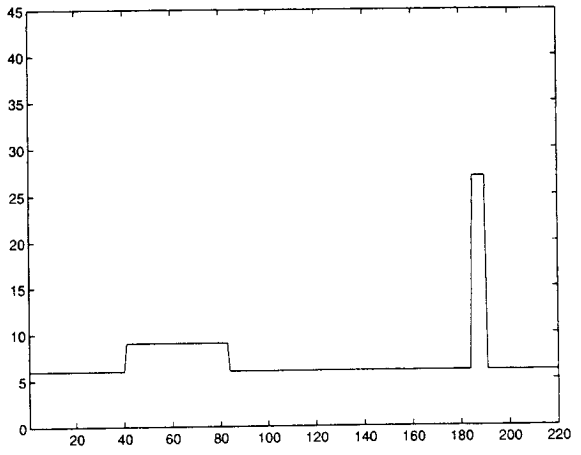


Fig. 9. Output of WC, $\lambda^2 = 45$.

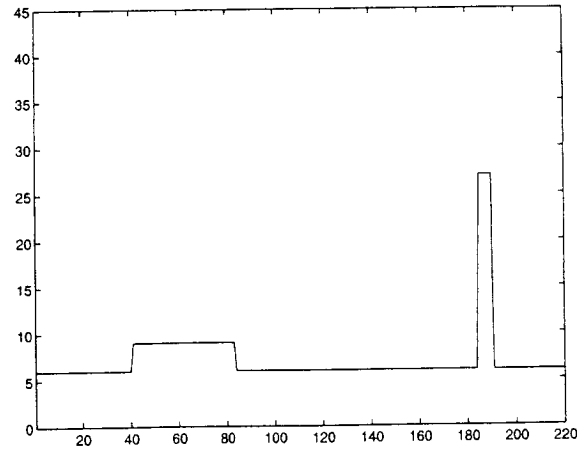


Fig. 10. Output of WC, $\lambda^2 = 20$.

choice does not appear to be critical.[8] We selected $\mathcal{C} = P_M^N$ since this is a natural parameterized constraint set for piecewise constant signals. Finally, we chose g_n to be a hardlimited MDL-type notch $g_n(x(n), x(n-1)) = \lambda^2[1 - \delta(x(n) - x(n-1))]$ for all n.

With these choices, we may use the claims above to help us pick appropriate values for the two optimization parameters. Accordingly, we selected the values $\lambda^2 = 15$ and $M = 15$. This way, we may

[8] Qualitatively comparable results have been obtained using the l_2 metric.

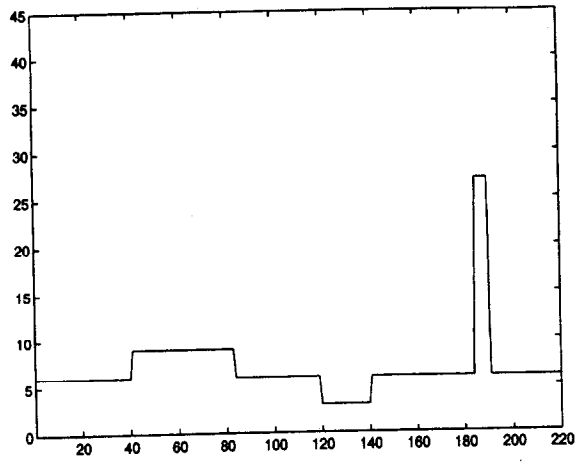


Fig. 11. Output of WC, $\lambda^2 = 15$.

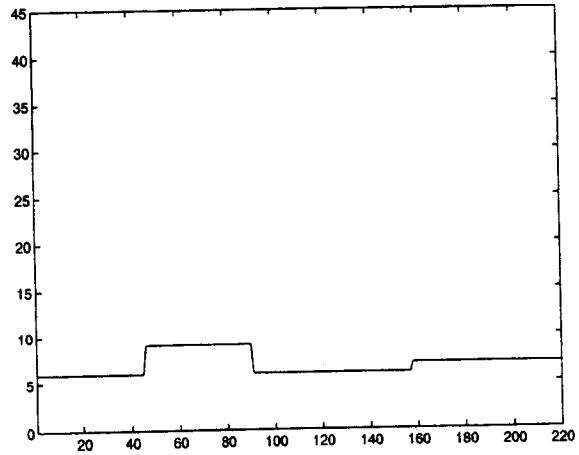


Fig. 12. Output of VORCA, $M = 45$.

guarantee the suppression of any isolated outlying burst of length up to 7 (which is just above what is required to suppress the simulated burst) and any isolated constant segment of saliency less than 30 (which is conservatively below the saliency of the weakest signal feature).

For $M = 1$, we obtain plain WC, and the results for $\lambda^2 = 55, 50, 45, 20, 15$ are depicted in Figs. 7–11, respectively. Observe that even though the saliency of the outlying burst is the same as that of the first signal pulse, WC first segments the burst (which lacks sufficient consistency) rather than the pulse. Actually, as illustrated by these figures, WC cannot properly segment the signal in this example without also segmenting the burst, i.e., it cannot differentiate between a consistent pulse and a relatively inconsistent outlying burst. This is because WC ranks features by saliency, and saliency is not an unambiguous indicator of consistency; what distinguishes the signal in Fig. 5 from the burst is consistency but not saliency. In addition, observe that (even though we used l_1 instead of l_2 distance), WC exhibits the so-called *uniform localization property* in scale-space. As λ^2 is reduced, new edges may be introduced, but previously detected edges remain stable (lines in scale-space are vertical) [1]. This is a desirable property [1].

For $\lambda^2 = 0$, we obtain plain VORCA, and the results for $M = 45, 40, 30, 25, 15$ are depicted in Figs. 12–16, respectively. Observe that as expected, VORCA first segments out the stronger signal pulse while virtually eliminating the outlying burst. It then proceeds to segment the second (weaker) signal pulse, whereas at the same time

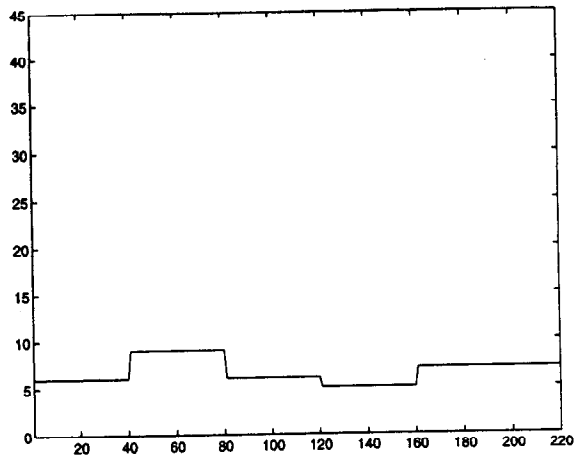


Fig. 13. Output of VORCA, $M = 40$.

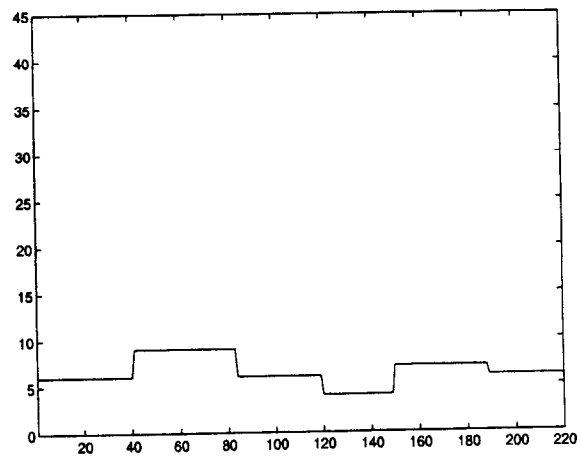


Fig. 14. Output of VORCA, $M = 30$.

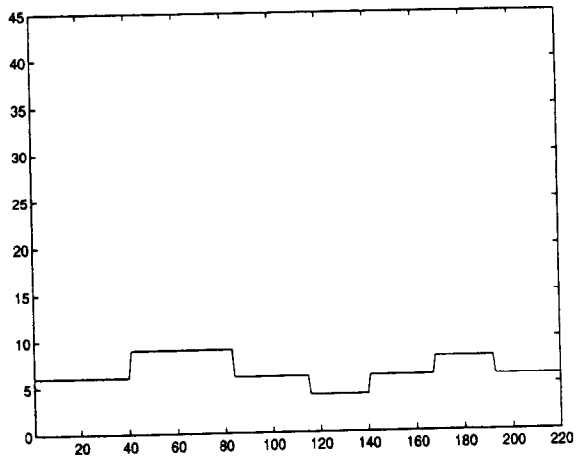


Fig. 15. Output of VORCA, $M = 25$.

producing ripple artifacts due to the burst and the Gaussian noise. These artifacts become progressively significant as M is reduced. Notice that even at $M = 45$, the effect of the burst is never *completely* eliminated due to the end-transient effect. In addition, observe that VORCA does not enjoy the uniform localization property of WC, although edges appear to be stable over wide ranges of values of M.

For $\lambda^2 = 15$, and $M = 15$, we have hybrid RC-WC, and the result is depicted in Fig. 17. RC-WC effectively combines the power

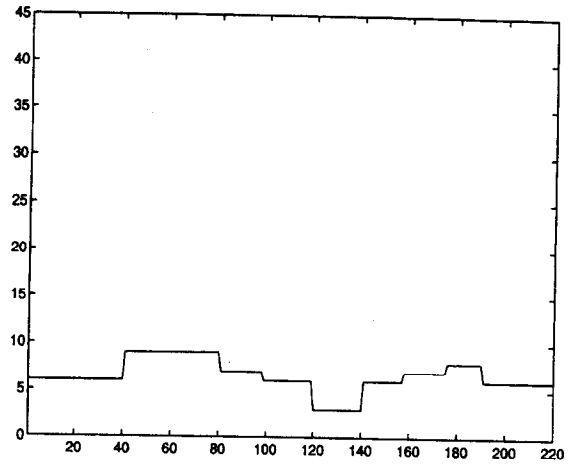


Fig. 16. Output of VORCA, $M = 15$.

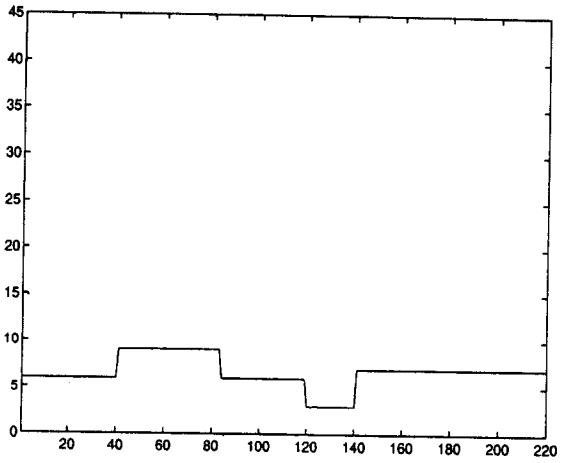


Fig. 17. RC-WC, $M = 15$, $\lambda^2 = 15$, combines the power of both methods.

of both methods. The notable exception relative to WC is the loss of the uniform localization property (in RC-WC, "scale" depends on both λ^2 and M; varying M does not necessarily lead to a stable scale space).

The overall run time is about 2 s for $|\mathcal{A}| = 50$ levels on a SUN SPARC 10 using simple C-code. Much better benchmarks may be expected for smaller alphabets and/or by implementing the algorithm in dedicated Viterbi hardware (cf. [32] and references therein).

V. Conclusion

We proposed WCSC, which is a hybrid nonlinear regression-regularization approach for segmentation and nonlinear filtering. The proposed approach draws on earlier work in WC and nonlinear regression, effectively combines the best of both worlds (with the notable exception of the uniform localization property of WC), and can be efficiently implemented via the Viterbi algorithm.

Two types of WCSC have been discussed: RC-WC and LM-WC. Due to space limitations, the emphasis here was on RC-WC. Depending on the kind of roughness-complexity regularizing functional used, LM-WC can be very different from RC-WC. In particular, the characteristic set of RC-WC is a proper subset of that of LM-WC. The latter, e.g., includes ramp signals and all monotonic signals. For relatively mild roughness-complexity penalties, LM-WC may follow ramp edges, whereas RC-WC will convert these to step edges. LM-WC is computationally more complex than RC-WC.

WCSC does not incorporate an *explicit* blur model. It may restore blurred and noisy edges but in a somewhat *ad hoc* manner. If the data is blurred and the blur is, e.g., asymmetric, restoration may fail to properly localize edges. The incorporation of an explicit blur model into the present paradigm may be worthwhile in cases where the present approach fails.

Acknowledgment

The authors would like to thank the anonymous reviewers, whose comments helped improve this manuscript.

References

[1] A. Blake and A. Zisserman, *Visual Reconstruction*. Cambridge, MA: MIT Press, 1987.

[2] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-6, pp. 721–741, Nov. 1984.

[3] D. Geman, S. Geman, C. Graffigne, and P. Dong, "Boundary detection by constrained optimization," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 12, pp. 609–627, July 1990.

[4] D. Geman and C. Yang, "Nonlinear image recovery with half-quadratic regularization," *IEEE Trans. Image Processing*, vol. 4, pp. 932–946, July 1995.

[5] D. Geman and G. Reynolds, "Constrained restoration and the recovery of discontinuities," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 14, pp. 367–384, Mar. 1992.

[6] D. Geiger and F. Girosi, "Parallel and deterministic algorithms from MRF's: Surface reconstruction," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 13, pp. 401–412, May 1991.

[7] P. M. Djuric, "A MAP solution to the off-line segmentation of signals," in *Proc. IEEE ICASSP*, Adelaide, Australia, 1994, vol. IV, pp. 505–508.

[8] M. A. T. Figueiredo and J. M. N. Leitao, "Adaptive discontinuity location in image restoration," in *Proc. IEEE ICIP*, Austin, TX, 1994, vol. I, pp. 665–669.

[9] S. Geman, D. E. McClure, and D. Geman, "A nonlinear filter for film restoration and other problems in image processing," *Comput. Vision, Graphics Image Process.: Graphical Models Image Process.*, vol. 54, no. 4, pp. 281–289, July 1992.

[10] C. Bouman and K. Sauer, "A generalized Gaussian image model for edge-preserving MAP estimation," *IEEE Trans. Image Processing*, vol. 2, pp. 296–310, July 1993.

[11] S. G. Tyan, "Median filtering: Deterministic properties," in *Two-Dimensional Digital Signal Processing II: Transforms and Median Filters*, T. S. Huang, Ed. Berlin: Germany: Springer-Verlag, 1981, pp. 197–217.

[12] N. C. Gallagher, Jr. and G. W. Wise, "A theoretical analysis of the properties of median filters," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 29, pp. 1136–1141, Dec. 1981.

[13] B. I. Justusson, "Median filtering: Statistical properties," in *Two-Dimensional Digital Signal Processing II: Transforms and Median Filters*, T. S. Huang, Ed. Berlin, Germany: Springer-Verlag, 1981, pp. 161–196.

[14] T. A. Nodes and N. C. Gallagher, Jr., "Median filters: Some modifications and their properties," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 30, pp. 739–746, 1982.

[15] N. C. Gallagher, Jr., "Median filters: A tutorial," in *Proc. IEEE Int. Symp. Circuits Syst., ISCAS*, 1988, pp. 1737–1744.

[16] A. C. Bovik, T. S. Huang, and D. C. Munson, "A generalization of median filtering using linear combinations of order statistics," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-31, pp. 1342–1349, 1983.

[17] ——, "The effect of median filtering on edge estimation and detection," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-9, pp. 181–194, Mar. 1987.

[18] A. Restrepo and A. C. Bovik, "Locally monotonic regression," *IEEE Trans. Signal Processing*, vol. 41, pp. 2796–2810, Sept. 1993.

[19] ——, "Statistical optimality of locally monotonic regression," *IEEE Trans. Signal Processing*, vol. 42, pp. 1548–1550, June 1994.

[20] N. D. Sidiropoulos, "The Viterbi optimal runlength-constrained approximation nonlinear filter," *IEEE Trans. Signal Processing*, vol. 44, pp. 586–598, Mar. 1996.

[21] ——, "Fast digital locally monotonic regression," *IEEE Trans. Signal Processing*, vol. 45, pp. 389–395, Feb. 1997.

[22] D. Mumford and J. Shah, "Boundary detection by minimizing functionals," in *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, San Francisco, CA, 1985.
[23] ——, "Optimal approximations by piecewise smooth functions and associated variational problems," *Commun. Pure Applied Math.*, vol. 42, pp. 577–685, 1989.
[24] J.-M. Morel and S. Solimini, *Variational Methods in Image Segmentation*. Boston, MA: Birkhauser, 1994.
[25] D. Terzopoulos, "Regularization of inverse visual problems involving discontinuities," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-8, pp. 413–424, July 1986.
[26] U. Grenander and M. Miller, "Representation of knowledge in complex systems," *J. R. Stat. Soc. B*, vol. 56, no. 4, pp. 549–603, 1994.
[27] M. Miller, B. Roysam, K. Smith, and J. O'Sullivan, "Representing and computing regular languages on massively parallel networks," *IEEE Trans. Neural Networks*, vol. 2, pp. 56–72, Jan. 1991.
[28] A. Blake, "Comparison of the efficiency of deterministic and stochastic algorithms for visual reconstruction," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 11, pp. 2–12, Jan. 1989.
[29] R. Bellman, *Dynamic Programming*. Princeton, NJ: Princeton Univ. Press, 1957.
[30] A. J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Trans. Inform. Theory*, vol. IT-13, pp. 260–269, Apr. 1967.
[31] J. K. Omura, "On the Viterbi decoding algorithm," *IEEE Trans. Inform. Theory*, vol. IT-15, pp. 177–179, Jan. 1969.
[32] H.-L. Lou, "Implementing the Viterbi algorithm," *IEEE Signal Processing Mag.*, vol. 12, pp. 42–52, May 1995.
[33] A. V. Papoulias, "Curve segmentations using weak continuity constraints," M.Sc. thesis, Univ. Edinburgh, Edinburgh, U.K., 1985.
[34] A. A. Amini, T. E. Weymouth, and R. C. Jain, "Using dynamic programming for solving variational problems in vision," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 12, pp. 855–867, Sept. 1990.
[35] G. L. Bibro, W. E. Snyder, S. J. Garnier, and J. W. Gault, "Mean field annealing: A formalism for constructing GNC-like algorithms," *IEEE Trans. Neural Networks*, vol. 3, pp. 131–138, Jan. 1992.
[36] Y. Leclerc, "Constructing simple stable descriptions for image partitioning," *Int. J. Comput. Vision*, vol. 3, no. 1, pp. 73–102, 1989.
[37] J. Rissanen, *Stochastic Complexity in Statistical Inquiry*. Singapore: World Scientific, 1989.
[38] H. J. A. M. Heijmans, *Morphological Image Operators*. Boston, MA: Academic, 1994.