Two-dimensional phase unwrapping with use of statistical models for cost functions in nonlinear optimization

Curtis W. Chen and Howard A. Zebker

Department of Electrical Engineering, Stanford University, Stanford, California 94305-9515

Received April 21, 2000; revised manuscript received August 7, 2000; accepted August 15, 2000

Interferometric radar techniques often necessitate two-dimensional (2-D) phase unwrapping, defined here as the estimation of unambiguous phase data from a 2-D array known only modulo 2π rad. We develop a maximum *a posteriori* probability (MAP) estimation approach for this problem, and we derive an algorithm that approximately maximizes the conditional probability of its phase-unwrapped solution given observable quantities such as wrapped phase, image intensity, and interferogram coherence. Examining topographic and differential interferometry separately, we derive simple, working models for the joint statistics of the estimated and the observed signals. We use generalized, nonlinear cost functions to reflect these probability relationships, and we employ nonlinear network-flow techniques to approximate MAP solutions. We apply our algorithm both to a topographic interferogram exhibiting rough terrain and layover and to a differential interferogram measuring the deformation from a large earthquake. The MAP solutions are complete and are more accurate than those of other tested algorithms. © 2001 Optical Society of America

OCIS codes: 280.6730, 120.3180, 350.5030.

1. INTRODUCTION

Two-dimensional (2-D) phase unwrapping has received a great deal of attention in recent years, owing in large part to the advent of synthetic aperture radar (SAR) interferometry. In this application, multiple coherent radar images of the same area are combined to form interferograms, with the 2-D interferometric phase arrays providing extremely fine measurements of surface topography,[1] deformation,[2] or velocity.[3] Phase, however, can be measured only modulo 2π rad, so physical quantities derived from interferometric phase data are wrapped with respect to some modulus or ambiguity and often must be unwrapped to provide meaningful information.

We propose a new approach for choosing phase-unwrapped solutions given wrapped data. Through the use of nonlinear cost functions, we cast the phase-unwrapping problem as a maximum *a posteriori* probability (MAP) estimation problem, developing approximate models for the statistics and expected properties of interferometric SAR signals. Although faithfully modeling elaborate probability relationships is an enormous theoretical task, we sidestep much of this complexity and focus instead on designing a working algorithm for use in an applied, practical sense. We also describe a technique based on nonlinear network optimization for approximately solving the posed estimation problem, and we demonstrate our algorithm's performance on interferometric SAR data sets that measure both rugged topography and surface deformation from a large earthquake.

Strictly, phase unwrapping is an impossible problem, because an unwrapped phase array necessarily contains information not available in the wrapped array; all phase-unwrapping algorithms therefore rely on at least some assumptions. The most basic and most common of these assumptions is that the true unwrapped phase field varies slowly enough that neighboring phase values are within one half cycle (π rad) of one another throughout much of the interferogram. Where this is the case, unwrapped phase values may be obtained by simply integrating the neighboring-pixel wrapped phase differences—called gradients in the phase-unwrapping literature—from pixel to pixel. The difficulty of phase unwrapping, of course, stems from the fact that in nearly all phase fields of interest, some gradients do exceed one half cycle, and their incorrect integration results in global unwrapping errors. The task of a phase-unwrapping algorithm therefore reduces to locating and accommodating such gradients.

While these gradients can be handled in a variety of ways, many popular algorithms pose phase unwrapping as a constrained-optimization problem. In such a problem, an objective function maps unwrapped solutions to scalar values. A solver routine is then used to find a solution that minimizes the value of the objective function, either exactly or approximately, while meeting predefined problem constraints that ensure the solution's validity. The main differences between many phase-unwrapping algorithms can thus be viewed as differences between the algorithms' objective functions and respective minimization techniques.

An objective function can be any function of the set of all phase values or, equivalently, any function of the set of all phase gradients. In the interest of computational efficiency, however, the objective function is commonly assumed to be separable so that it can be written in the form

$$\text{minimize}\left\{\sum_{i}\sum_{j} g^{(r)}_{i,j}(\Delta\phi^{(r)}_{i,j}, \Delta\psi^{(r)}_{i,j}) + \sum_{i}\sum_{j} g^{(a)}_{i,j}(\Delta\phi^{(a)}_{i,j}, \Delta\psi^{(a)}_{i,j})\right\}, \qquad (1)$$

where $\Delta\phi^{(r)}$ and $\Delta\psi^{(r)}$ are the range components of the unwrapped and the wrapped phase gradients, respectively, and $\Delta\phi^{(a)}$ and $\Delta\psi^{(a)}$ are their azimuthal counterparts in a 2-D range-azimuth coordinate system (for side-looking imaging radars, range and azimuth are the across-track and along-track directions). Wrapped gradients are always assumed to be between $-\pi$ and π. The functions $g(\,\cdot\,)$ are called cost functions by analogy to minimizing the total cost of the solution.

Ghiglia and Romero[4] suggested a phase-unwrapping framework under which the cost functions are restricted to the form

$$g(\Delta\phi, \Delta\psi) = w|\Delta\phi - \Delta\psi|^{p}. \qquad (2)$$

Here the cost functions all have the same shape, as determined by the constant p, and independent weights w determine each cost function's scaling. The resulting objective function defines the weighted minimum L^p-norm problem, or for compactness of notation, simply the L^p problem. When $p = 2$, the problem is a weighted least-squares minimization problem, and many approaches to solving it have been introduced.[5–7] Alternatively, when $p = 1$, the linearity of the objective function permits efficient solution.[8,9] In the limit as p approaches zero (henceforth $p = 0$ or L^0), the objective is to minimize the weighted number of locations where the unwrapped and the wrapped gradients differ; several algorithms for this problem have been proposed as well,[4,10–12] although as described below, none are actually able to solve it exactly. Note that while we reference these algorithms for completeness, we treat L^p cost functions themselves as generic optimization criteria, independent of the methods used to minimize them and separate from any specific algorithm implementation.

Regardless of the objective function used, the optimization problem may be constrained such that congruence is required between the unwrapped and the wrapped phase arrays. Corresponding unwrapped and wrapped phase values may then differ only by integer numbers of cycles, so the assumption of congruence makes the solution space discrete. Nevertheless, the cost functions described above can still be used to compare different allowable solutions. Moreover, for the L^0 and L^1 norms with integer parameters, a noncongruent optimum can be no better than a congruent one.[11,13] For the L^2 norm, however, a noncongruent optimum is generally better than one computed under the assumption of congruence, though the former loses L^2 optimality, even over the set of strictly congruent solutions, if it is forced into congruence in a postoptimization processing step.[14]

Goldstein *et al.*[10] pointed out that for a properly unwrapped, congruent phase array, the sum (integral) of phase gradients around a closed, directed loop of 2 × 2 pixels is always zero. A nonzero result, called a residue, may arise in a wrapped phase array, however, when gradients are incorrectly assumed to be less than one half cycle. Residues thus indicate the presence of inconsistencies with the assumption that all gradients are less than one half cycle. Moreover, in the unwrapped array, discontinuities, or lines of gradients greater than one half cycle, necessarily run between residues of opposite signs.

For the L^1 metric, Costantini[9] recognized that existing, generic minimum-cost-flow (MCF) solver routines could be exploited through the explicit adoption of a network-flow model[13] for the phase-unwrapping problem (see Fig. 1). Each 2 × 2 residue loop integral of the wrapped phase is a node in this network; single units of surplus and demand (negative surplus) are assigned to nodes of positive and negative residues, respectively, and flow is allowed to travel on arcs connecting neighboring nodes. In this paper, arcs are bidirectional and flow magnitudes are integers between $-\infty$ and ∞, with sign denoting the direction of flow. Constraints are defined so that the net flow out of each node (total flow out minus total flow in) must be equal to the surplus of the node. When flow is conserved in this way, the solution is called feasible, and a feasible solution is optimal if it also minimizes the value of the objective function. Since arcs in the network correspond to phase gradients, the direction and the magnitude of flow on an arc physically represent the sign and the difference, in cycles, between the unwrapped and the wrapped phase gradients associated with that arc. In this way, for a given wrapped phase array, any feasible network flow exactly corresponds to a valid unwrapped phase array.

Using ideas from network theory, Chen and Zebker[11] showed that the L^0 phase-unwrapping problem is *NP*-hard and therefore cannot be solved exactly in polynomial time (unless problem classes *P* and *NP* are equivalent).[15] Algorithms using the L^0 objective thus compute only approximate rather than exact solutions. (Note that although in some contexts the term "approximate" implies performance guarantees or complexity bounds, we use it here only colloquially.) Furthermore, because it is a generalization of the L^0 problem, the optimization problem of Eq. (1) is *NP*-hard as well in the absence of simplifying assumptions. Although Carballo[16] outlined conditions under which L^1 methods can be used to solve congruent nonlinear problems, his conditions reduce to the well-known requirement of convexity.[13] The nonconvex L^0 problem remains *NP*-hard.

Nevertheless, an objective function's computational difficulty does not necessarily detract from its appeal. For

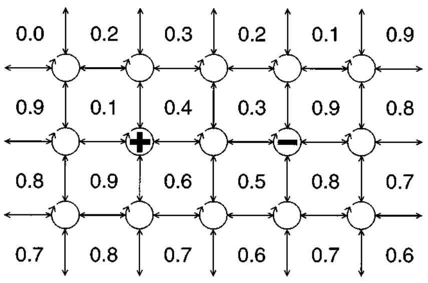


Fig. 1. Example network equivalent of the phase unwrapping problem. The numbers represent a 2-D array of phase samples (normalized to one cycle). Each 2 × 2 clockwise loop integral of wrapped phase gradients is a node in the network, and positive and negative residues result in supply and demand nodes. Neighboring nodes are connected by arcs, or possible flow paths.

example, it has been demonstrated qualitatively and confirmed empirically that the L^0 metric is well suited to data where the true unwrapped phase field contains physical discontinuities.[10,11,14] However, there are no physical reasons that exactly optimal L^p solutions must be correct. L^p norms are simply abstract mathematical quantities, which have empirically led to workable solutions.

In an effort to strengthen the physical foundations of optimization objectives, we introduce here new objectives based on generalized, statistical cost functions. That is, we allow the cost functions $g(\cdot)$ in Eq. (1) to take any form and to vary in shape for different parts of the interferogram. We then choose cost functions that maximize the conditional probability of a solution given the wrapped phase, image intensity, and interferogram coherence. This implies the use of application-specific objective functions that yield greater accuracy than can be obtained from "one-size-fits-all" algorithms. Separately examining both topographic and deformation-mapping applications of SAR interferometry, we derive simple statistical models for each, making assumptions about the physical characteristics of the specific measurements. We outline in Appendix A a procedure based on nonlinear network optimization for approximating solutions to these estimation problems. Using this procedure, we demonstrate the performance of our algorithm on both topographic and differential SAR interferograms.

2. STATISTICAL FRAMEWORK FOR GENERALIZED COST FUNCTIONS

A constrained-optimization approach to phase unwrapping involves balancing two aims: (1) the objective function should provide accurate solutions when minimized, and (2) solutions in terms of this objective function should be computable efficiently. The lack of complete information in a wrapped phase array and the *NP*-hardness of the general unwrapping problem sometimes make this balance a difficult one, but recent advances on the latter aim call now for work on the former. Specifically, the solver routine of the dynamic-cost-cycle-canceling (DCC) algorithm proposed by Chen and Zebker[11] can approximate not only L^0 solutions but solutions for *any* separable objective function (see Appendix A). We therefore derive in this section statistically based, application-specific, generalized cost functions $g(\cdot)$ for computing maximally accurate solutions.

Much work to date has addressed the subject of advantageous weights for L^p cost functions,[14] but L^p cost functions all have the same shape for a given objective function (Fig. 2). Most investigations have thus treated cost-function shape and scaling as two distinct issues. The generalized cost functions of Eq. (1), however, handle both together, as the cost function for each phase difference in the interferogram has its own individual form, arbitrary and independent of all others.

We use this flexibility to design cost functions based on MAP estimation. That is, given a 2-D wrapped phase field Ψ, we develop an objective function such that its minimization results in an estimate $\hat{\Phi}$ of the true unwrapped phase field Φ, where $\hat{\Phi}$ approximately maximizes the conditional probability density function (PDF) $f(\Phi|\Psi)$. Of course, joint PDF's such as this are very difficult to formulate exactly, but realistic solver routines cannot guarantee exact solutions to the *NP*-hard minimization problem in any event. Consequently, our approach is geared toward the design of a working algorithm that—even if inexact—provides significantly better results than other existing algorithms. Following this philosophy, we approximate the joint PDF's with simplifying models whose refinement we leave for later work. We present here the general methodology and practical application of our approach, focusing on physical insight rather than mathematical detail; the success of this approach is demonstrated in Section 3.

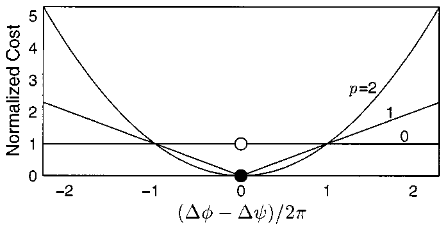


Fig. 2. Normalized cost functions for the L^p family of objective functions. The abscissa is the difference in cycles between the unwrapped and the wrapped gradients.

In our notation, capital letters denote arrays and lowercase letters denote individual entries in those arrays. That is, $\Delta\Phi$ is the set of all unwrapped gradients $\Delta\phi$ in an interferogram, where $\Delta\phi$ is any particular row-wise or column-wise gradient. Uppercase and lowercase letters are not random variables and their specific instances. We denote both probability mass and density functions by $f(\cdot)$. Unless otherwise noted by a subscript, $f(\cdot)$ describes the random variable(s) in its argument. Thus, for example, $f(\Phi|\Psi)$ is equivalent to $f_{\Phi|\Psi}(\Phi|\Psi)$.

We begin by changing the problem variables, without loss of generality, so that our goal is to estimate the set of all unwrapped gradients $\Delta\Phi$ given the set of all wrapped gradients $\Delta\Psi$. We assume that the PDF $f(\Delta\Phi|\Delta\Psi)$ is separable in that individual unwrapped phase gradients are statistically independent given their wrapped counterparts and given the knowledge that the resulting unwrapped phase field Φ is a residue-free (irrotational) surface. This latter condition is enforced by our solver routine (see Appendix A), which ignores invalid, nonfeasible solutions. Of course, our assumption of independence is not strictly correct, but its viability is born out by empirically verified results, and it is in any case required for computational tractability. Thus

$$f(\Delta\Phi|\Delta\Psi) = \prod_k f(\Delta\phi_k|\Delta\psi_k). \tag{3}$$

The product with index k in this expression is taken over all rows and columns for the sets of both row-wise and column-wise (range and azimuth) gradients. Using logarithms, we transform the maximization of this product into a minimization of sums:

$$\text{minimize}\left\{-\sum_k \log[f(\Delta\phi_k|\Delta\psi_k)]\right\}. \tag{4}$$

Comparing Eq. (4) with Eq. (1), we define our cost functions to be the negative logarithms of the unwrapped-gradient PDF's:

$$g_k(\Delta\phi_k,\ \Delta\psi_k) = -\log[f(\Delta\phi_k|\Delta\psi_k)]. \tag{5}$$

To avoid the phase-gradient underestimation effects typical of noncongruent solutions,[14,17,18] we enforce congruence between unwrapped and wrapped phase values. The conditional probability of an unwrapped gradient may then be expressed in terms of its nonconditional PDF as

$$f(\Delta\phi|\Delta\psi) = \begin{cases} \dfrac{f_{\Delta\phi}(\Delta\phi)}{\sum\limits_{m=-\infty}^{\infty} f_{\Delta\phi}(\Delta\psi + m2\pi)} & \text{if } \Delta\phi = \Delta\psi + n2\pi \\ 0 & \text{otherwise} \end{cases}, \tag{6}$$

where n and m are integers. The denominator of the fraction does not depend on $\Delta\phi$ and has no effect on the minimization problem. Consequently, we need to model only the individual unwrapped-gradient distributions $f(\Delta\phi)$ and evaluate them at integer-cycle offsets from the wrapped phase.

As with any MAP estimate, however, our unwrapped solution benefits from the inclusion of additional information. All SAR interferograms are described by intensity and coherence information, so we explicitly rewrite $f(\Delta\phi)$ as the conditional PDF $f(\Delta\phi|I,\ \rho)$, where I is the average of the intensities of the SAR images forming the interferogram and ρ is the magnitude of the interferogram complex correlation coefficient.[19] Although the interferogram phase statistics may be written very compactly in this form, the conditional PDF actually embodies untold complexity. Moreover, different applications have different statistics, so we must treat each separately. We concentrate here on two applications of SAR interferometry: topography and deformation.

A. Topography Measurements

We examine in this subsection the specific statistics and resulting cost functions for topographic SAR interferometry. With unwrapped phase measuring the elevation of the target surface,[1] we decompose the true unwrapped gradients into their topographic and phase-noise parts:

$$\Delta\phi = \Delta\phi_{\text{topo}} + \Delta\phi_{\text{noise}}. \tag{7}$$

The term ϕ_{noise} here represents the combined phase noise from all sources, not the phase of the complex noise. Given the coherence magnitude, the phase noise is unrelated to the intensity and is independent of topographic slope, so the conditional unwrapped-gradient PDF is the convolution of the probability densities corresponding to Eq. (7):

$$f(\Delta\phi|I,\ \rho) = f(\Delta\phi_{\text{topo}}|I,\ \rho) * f(\Delta\phi_{\text{noise}}|\rho). \tag{8}$$

We first consider the unwrapped-gradient noise distribution, the second term on the right-hand side of Eq. (8). Lee *et al.*[20] derived analytical expressions for multilook interferometric phase-noise PDF's; these PDF's can be approximated by normal distributions in areas of high correlation. Since both topography and correlation usually vary slowly, neighboring phase-noise terms can be treated as identically distributed as well as independent. Their difference $\Delta\phi_{\text{noise}}$ is then a zero-mean Gaussian random variable whose variance $\sigma^2_{\Delta\psi}$ is twice that of the individual phase-noise variances σ^2_{ψ}. We calculate σ^2_{ψ} from the phase-noise standard deviation σ_{ψ}, shown in Fig. 3 as a function of ρ and the equivalent number of independent looks N_i. Our model for σ_{ψ} is based on similar plots given by Li and Goldstein,[21] Rodriguez and Martin,[22] and Lee *et al.*[20] We note that when the correlation is low, the normal approximation is less valid—in the worst case, the distribution of $\Delta\phi_{\text{noise}}$ is triangular, resulting from the convolution of two uniform $(-\pi, \pi)$ distributions. Rather than evaluate the hypergeometric functions of the exact PDF expression, however, we maintain the normal approximation even for low coherence.

In calculating $\sigma^2_{\Delta\psi}$ from the observed coherence, we must also be aware of the bias introduced by the common coherence estimator

$$\hat{\rho} = \left| \frac{\sum\limits_{k=1}^{N} s_{1k} s^*_{2k}}{\sqrt{\sum\limits_{k=1}^{N} |s_{1k}|^2 \sum\limits_{k=1}^{N} |s_{2k}|^2}} \right|, \tag{9}$$

where $\hat{\rho}$ is the biased estimate of the true coherence magnitude ρ, s_1 and s_2 are the signals forming the interferogram, * denotes complex conjugation, and N is the number of complex pixels (looks) used for the estimate. Touzi *et al.*[23] found an exact expression for the expected value of $\hat{\rho}$ in terms of ρ, N_i, and another hypergeometric function; we use a piecewise-linear model based on these results to estimate ρ from $\hat{\rho}$.

Independently, Carballo[16] used a similar, but not identical, approach to formulate weights for an L^1 minimization problem. He gave little attention to the statistics of the topographic part of Eq. (8), though, instead assuming that unwrapped and wrapped gradients virtually never differ by more than one cycle. However, it has been shown that because of topographic effects, multiple-cycle differences are in fact very important in topographic phase unwrapping.[11,14]

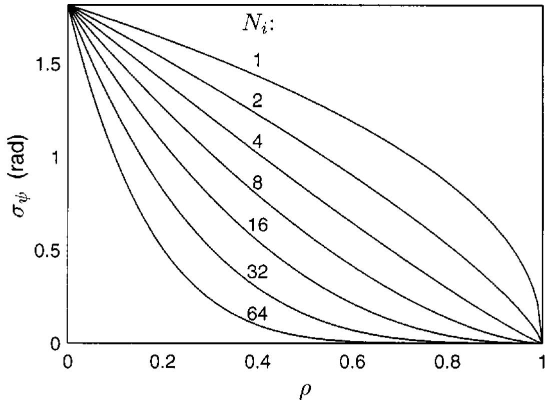


Fig. 3. Model interferometric phase-noise standard deviation σ_ψ as a function of interferogram coherence ρ for different numbers of independent looks N_i.

We therefore consider next the topographic phase term of Eq. (8). If the topography-independent (flat-earth) phase signature is removed, the true unwrapped topographic phase is approximately related to the relative surface elevation z by[1]

$$\phi_{\text{topo}} = \frac{-4\pi B_{\perp}}{\lambda r \sin\theta} z, \tag{10}$$

where $B_{\perp}$ is the component of the interferometer baseline perpendicular to the radar signal, λ is the signal wavelength, and θ is the look angle with respect to nadir. The unwrapped gradient $\Delta\phi_{\text{topo}}$ thus depends nearly linearly on the physical surface slope Δz, although care must be taken to properly interpret measurements in the SAR range-azimuth coordinate system.

Separating the intensity and correlation dependencies of the unwrapped gradient, we rewrite its conditional PDF from Eq. (8) as

$$f(\Delta\phi_{\text{topo}}|I,\, \rho) = \frac{f(\Delta\phi_{\text{topo}}|I) f(\rho|\Delta\phi_{\text{topo}})}{f(\rho|I)}, \tag{11}$$

assuming that ρ is independent of I given $\Delta\phi_{\text{topo}}$. The denominator on the right-hand side does not depend on $\Delta\phi_{\text{topo}}$ and may be dropped. Examining the first PDF of the numerator, we now model the relationship between topography and intensity; that is, we model the dependence of the radar image brightness on surface slope. This relationship is evident on inspection of any SAR image, and efforts have been devoted to both the recovery of topographic information from SAR intensity alone and to the radiometric correction of topographic effects.[24,25]

The coherent nature of SAR images—the very quality that permits interferometry—also causes speckle in the intensity, however, complicating the inference of topographic information from brightness. On the basis of ideas from Lopes *et al.*,[26] we use an adaptive speckle-removal filter that computes the mean intensity $E[I]$ of a variably oriented rectangular window, selecting the orientation that maximizes contrast with the local background. Although speckle statistics and removal have been examined extensively,[27] we assume perfect speckle removal to avoid unduly complicating our model PDF's. We then normalize $E[I]$, dividing by a coarse moving-window average.

We relate the normalized, despeckled intensity to the topography, using

$$E[I] = C\sigma^0 A, \tag{12}$$

where σ^0 is the normalized radar cross section, A is the area of the illuminated surface contributing to the measurement, and C is a scaling factor that may be ignored for normalized intensity data. Topography enters this equation, as both σ^0 and A are functions of the local signal incidence angle θ_i. To obtain expressions for them, we use a facet model to represent the ground surface demarcated by a single range-azimuth pixel (Fig. 4). In the model's right-handed $(x,\, y,\, z)$ coordinate system, earth curvature is neglected, and x and z are aligned with increasing ground range and elevation. The ground-range and azimuth pixel spacings are denoted by Δx and Δy, and their corresponding components of local elevation change are denoted by Δz_r and Δz_a. Through Eq. (10), Δz_r and Δz_a are therefore nearly proportional to the topographic parts of the unwrapped phase gradients. Note that while the slant-range bin spacing Δr is constant, the ground-range spacing depends on the local slope, so

$$\Delta x = \frac{\Delta r}{\sin\theta} + \frac{\Delta z_r}{\tan\theta}. \tag{13}$$

Assuming that the SAR viewing direction is exactly normal to the sensor velocity vector (i.e., zero squint angle), we thus derive the following two equations:

$$\cos\theta_i = \frac{(\Delta z_r/\Delta x)\sin\theta + \cos\theta}{[(\Delta z_r/\Delta x)^2 + (\Delta z_a/\Delta y)^2 + 1]^{1/2}} \tag{14}$$

$$A = [(\Delta y \Delta z_r)^2 + (\Delta x \Delta z_a)^2 + (\Delta x \Delta y)^2]^{1/2}. \tag{15}$$

Related expressions have also been derived elsewhere.[28]

Guided by Eq. (12), we next examine the normalized cross section σ^0, for which a variety of models exist.[29] We adopt one used by Goering *et al.*[25] because of its ease of evaluation given Eq. (14):

$$\sigma^0 = \begin{cases} k_{ds}\cos^2\theta_i + \cos^n 2\theta_i \cos\theta_i & \text{if } \cos 2\theta_i > 0 \\ k_{ds}\cos^2\theta_i & \text{otherwise} \end{cases}. \tag{16}$$

Here the parameter k_{ds} affects the ratio of diffuse to specular backscatter, n determines the sharpness of the specular peak with incidence angle, and a constant scaling factor has been dropped.

Incorporating the scattering model of Eq. (16) and the viewing geometry relations of Eqs. (14) and (15) into Eq. (12), we arrive at a model for the SAR intensity as a function of topography. For the scattering model parameters, we use values of $k_{ds} = 0.02$ and $n = 8$, finding them to give good agreement between the real and the simulated SAR intensity images shown in Fig. 5. (These parameters might differ for other terrain types.) The simulated image is generated from a digital elevation model (DEM) with only range components of slope used in our brightness model, as explained below. Section 3 contains more detail about both the SAR data and the DEM. As evidenced by the agreement between the two images, intensity is a valuable source of information about surface topography and is adequately reproduced with our model.

In Fig. 6 we plot the expected image intensity as a function of the range slope Δz_r for the case of zero slope in azi-

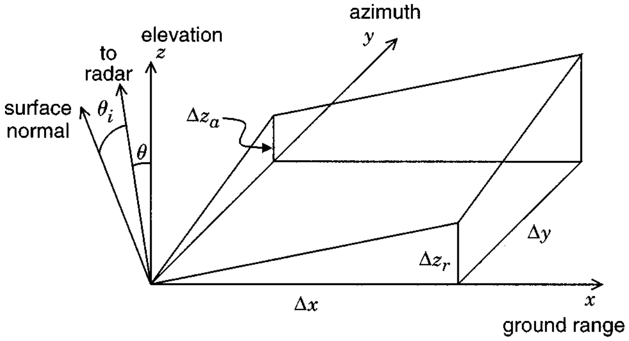


Fig. 4. Facet model used to relate topography to the brightness of a single SAR range-azimuth pixel. Note that Δx depends on both Δr and Δz_r.

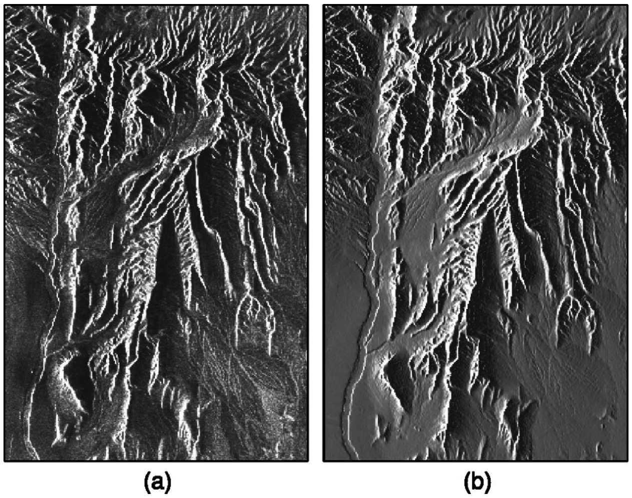


Fig. 5. Comparison of (a) actual, normalized SAR image intensity with (b) simulated intensity from a DEM and scattering model.

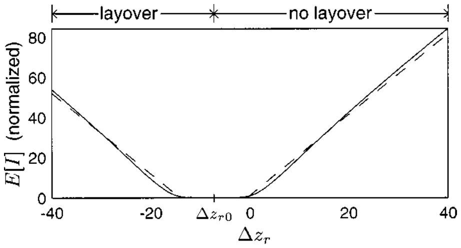


Fig. 6. Model intensity as a function of slant-range elevation change for zero azimuth slope. The solid line represents the expected intensity $E[I]$ from Eq. (12) and the dashed line is a piecewise linear approximation to the solid line.

muth ($\Delta z_a = 0$). The curve is normalized so that a horizontal surface ($\Delta z_r = 0$) has unit expected intensity. It is also symmetric about its zero where the local signal incidence angle is 90°, which occurs when $\Delta z_r = \Delta z_{r0} = -\Delta r \cos \theta$. Values of Δz_r between Δz_{r0} and zero correspond to surfaces sloping away from the radar, and values of Δz_r greater than zero correspond to surfaces sloping toward the radar. The intensity increases without bound as Δz_r becomes large and the local incidence angle approaches zero ($\Delta z_r \rightarrow \infty$, $\Delta z_r/\Delta x \rightarrow \tan \theta$, $\cos \theta_i \rightarrow 1$), because a single range bin then encloses infinite area (i.e., the facet model becomes invalid). A surface for which $\Delta z_r < \Delta z_{r0}$ may be in either layover or shadow. Surfaces in shadow usually have negligible brightness, so we assume that if $\Delta z_r < \Delta z_{r0}$, the surface is in layover. For example, if $\Delta z_r = -\Delta r/\cos \theta$, the surface is vertical, like the face of a cliff. Such a surface is physically sloped toward the radar, but the true elevation decreases as the slant range increases. As Δz_r becomes infinitely negative, θ_i again approaches zero and $E[I]$ again becomes infinite.

Because variations in image intensity are much more dependent on slopes in range than in azimuth, as described by Guindon,[24] we deal with range and azimuth gradients separately, beginning with the former. Qualitatively speaking, foreshortening effects from the SAR imaging geometry tend to make the range components of significant slopes much greater than the azimuth components. When the azimuth components are more significant than the range components, the total slope is often relatively low and presents little difficulty in phase unwrapping. Hence with $\Delta z_a = 0$, we can invert the relation illustrated in Fig. 6 to find an expected slant-range slope from our computed value of $E[I]$. To facilitate this inversion, we further simplify our brightness model by using a piecewise-linear approximation to $E[I]$ (dashed line in Fig. 6).

If there is no layover, $\Delta z_r > \Delta z_{r0}$ and the observed intensity corresponds uniquely to some expected range slope Δz_{rI}. This slope, the elevation change from one pixel to the next, can then be related to the unwrapped gradient $\Delta \phi$ through Eq. (10). We therefore expect a peak in $f(\Delta \phi_{\text{topo}}^{(r)}|I)$ at $\Delta \phi_I$, where the unwrapped gradient corresponds to Δz_{rI}.

Significant phase-unwrapping errors are often caused by layover, however, and in its presence, the observed pixel intensity is not directly related to the desired slope. This is because multiple parts of the imaged surface fall into the same range bin, as illustrated in Fig. 7. A large phase discontinuity arises from the elevation jump between range bins r_0 and r_1, but the brightness of range bin r_1 does not indicate the magnitude of this discontinuity. Furthermore, range bin r_2 appears bright because it contains part of the front face of the mountain, yet its unwrapped phase is usually assumed to represent the elevation of the mountain's back face.

To model the effects of topography on intensity where there is layover, we note that the elevation change from range bin r_0 to r_1 in Fig. 7 is related to the full height of the illustrated mountain. We note also that the height of the mountain is equal to the integral of the slope along one of its faces. Therefore, because slope is related to brightness through Eq. (12), we can estimate the severity a layover discontinuity by examining the intensities of all range bins containing part of the face in layover (range

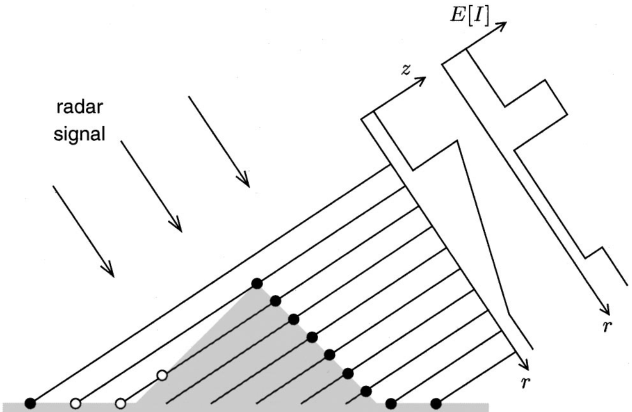


Fig. 7. Profile of a mountain in layover. The range bins r_0–r_9 represent contours of constant range from the radar. The elevation z and mean intensity $E[I]$ are plotted as they map into slant range for this profile. Because of layover, multiple parts of a surface may map into the same range bin; the solid and open circles represent intersections of the ground surface with the range contours. Unwrapped phase values are assumed to represent elevations at the solid circles, but echoes from the locations of the open circles complicate the topography–intensity relationship when there is layover.

bins r_1 and r_2 in Fig. 7). We therefore compute Δz_{lay}, the expected maximum elevation change of an unwrapped gradient straddling a layover discontinuity, for each pixel as follows. First, we compare the local intensity to some threshold value. If the intensity is below the threshold, layover is unlikely; otherwise, we convert the intensity to a layover-face range slope, using the relation of Fig. 6 over the layover regime ($\Delta z_r < \Delta z_{r0}$). We then integrate the similarly calculated, intensity-derived range slopes for the next several pixels in increasing range to obtain Δz_{lay}. The true elevation change may be less than Δz_{lay}, though, since some of the bright pixels contributing to the integrated value may actually correspond to slopes that are not in layover ($\Delta z_r > \Delta z_{r0}$). Layover thus forces us to incorporate more features into our model for $f(\Delta \phi_{\text{topo}}^{(r)}|I)$. In addition to the peak at $\Delta \phi_I$, we expect the PDF to have a wide, platformlike section with an upper cutoff at $\Delta \phi_{\text{lay}}$, the unwrapped gradient corresponding to Δz_{lay}. This section reflects the probability that $\Delta \phi_{\text{topo}}^{(r)}$ indeed represents a layover discontinuity.

Layover causes further complications, however. A pixel whose unwrapped phase corresponds to a mild, negatively sloped back face, like range bin r_2 in Fig. 7, may appear bright because it contains part of the face in layover. The true unwrapped gradient would then be unrelated to the computed values for Δz_{rI} and Δz_{lay}. Since a gradient straddling a layover discontinuity cannot easily be distinguished from the gradients in the bright area beyond the layover discontinuity, we must account for the probability of the latter in $f(\Delta \phi_{\text{topo}}^{(r)}|I)$. We therefore include in the PDF a peak at a slightly negative value near zero, denoted $\Delta \phi_{\text{back}}$. The resulting PDF is shown in Fig. 8. It is asymmetric, because as range increases, an upward step in elevation due to layover is usually much more likely than a downward one.

Having a qualitative model for $f(\Delta \phi_{\text{topo}}^{(r)}|I)$, we now return to Eq. (11) and consider $f(\rho|\Delta \phi_{\text{topo}}^{(r)})$, the conditional PDF of the coherence given the unwrapped gradient. Although we have already described the relationship between the coherence and the phase noise (Fig. 3), ρ is also related to the topographic part of the unwrapped gradient through spatial or baseline decorrelation. Zebker and Villasenor[19] obtained an expression for this decorrelation factor, denoted ρ_s:

$$\rho_s = \max\left\{0, 1 - \frac{2|B_\perp|R_r}{\lambda r|\tan \theta_i|}\right\}. \tag{17}$$

Here r is the slant range, R_r is the slant-range resolution, and the azimuth slope is assumed to be zero. Thus as the range slope increases and the local incidence angle decreases, the correlation decreases as well. In the presence of layover, though, several different parts of the surface map into the same pixel, and the multiplicity of local incidence angles then makes Eq. (17) inapplicable. However, correlation measures for areas in layover are generally very low because of actual decorrelation as well as the random complex superposition of the different signals contributing to the interferometric phase. This property has been exploited by phase-unwrapping schemes that use correlation information alone for weighting L^p cost functions.[6,14]

Because there are usually other contributions to decorrelation as well, ρ_s is an upper bound on the expected statistical correlation ρ. For fixed ρ and variable $\Delta \phi_{\text{topo}}^{(r)}$, then, the function $f(\rho|\Delta \phi_{\text{topo}}^{(r)})$ has a cutoff at $\Delta \phi_\rho$, the unwrapped gradient corresponding to the range slope for which $\rho = \rho_s$. That is, the observed correlation places an upper limit on the expected slope; if the correlation is high, the maximum expected slope is small. Without more specific knowledge of other decorrelation sources, we assume that for fixed ρ, $f(\rho|\Delta \phi_{\text{topo}}^{(r)})$ is constant when $\Delta \phi_{\text{topo}}^{(r)} \leqslant \Delta \phi_\rho$ and is zero when $\Delta \phi_{\text{topo}}^{(r)} > \Delta \phi_\rho$ (see Fig. 8).

Guided by Eq. (11), we now combine the PDF's that relate topography to intensity and correlation, obtaining a model for $f(\Delta \phi_{\text{topo}}^{(r)}|I, \rho)$, the conditional PDF of an unwrapped range gradient's topographic component. As this PDF is proportional to the product of $f(\Delta \phi_{\text{topo}}^{(r)}|I)$ and $f(\rho|\Delta \phi_{\text{topo}}^{(r)})$, its shape is similar to the former, but its upper cutoff $\Delta \phi_{\max}$ is determined by the lesser of the layover cutoff $\Delta \phi_{\text{lay}}$ and the correlation cutoff $\Delta \phi_\rho$ (see Fig. 8):

$$\Delta \phi_{\max} = \min\{\Delta \phi_{\text{lay}}, \Delta \phi_p\}. \tag{18}$$

Finally, with the topographic PDF $f(\Delta \phi_{\text{topo}}^{(r)}|I, \rho)$, we can proceed to form our MAP cost functions through Eqs. (5) and (8). Because the sharp peaks of $f(\Delta \phi_{\text{topo}}^{(r)}|I, \rho)$ are likely much narrower than the Gaussian noise PDF with which they are convolved, the resulting full PDF $f(\Delta \phi^{(r)}|I, \rho)$ is characterized mainly by the width of the noise PDF and the critical unwrapped-gradient values $\Delta \phi_{\text{back}}$, $\Delta \phi_I$, and $\Delta \phi_{\max}$ illustrated in Fig. 8. We assume that after convolution with the noise PDF, the narrow peaks at $\Delta \phi_I$ and $\Delta \phi_{\text{back}}$ can be modeled by a single, wide hump, as shown in Fig. 9. That is, the convolution removes much of the apparent structure from $f(\Delta \phi_{\text{topo}}^{(r)}|I, \rho)$, resulting in a simpler form with fewer parameters. The negative logarithm of $f(\Delta \phi^{(r)}|I, \rho)$ is the continuous-valued cost function $g^{(r)}(\Delta \phi^{(r)})$, which we can

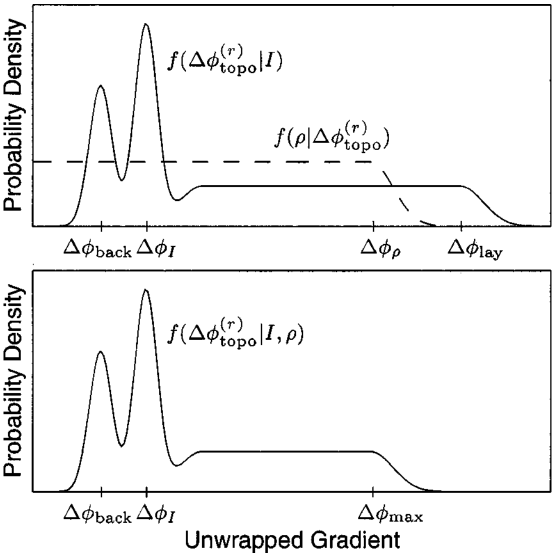


Fig. 8. Model PDF's for the topographic component of an unwrapped range gradient, conditional on observed intensity and correlation values. The PDF in the bottom panel is proportional to the product of the curves in the top panel. Note that the dashed curve in the top panel is $f(\rho|\Delta \phi_{\text{topo}}^{(r)})$ for fixed ρ and variable $\Delta \phi_{\text{topo}}^{(r)}$, not vice versa.

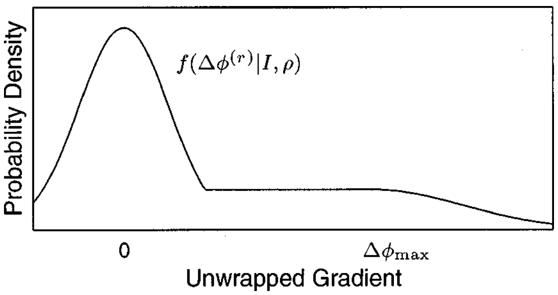


Fig. 9. Model conditional PDF for an unwrapped range gradient, with both topographic and noise components included.

evaluate at integer-cycle offsets from $\Delta\psi^{(r)}$ to obtain discrete gradient costs (i.e., flow costs) for our topographic phase-unwrapping problem.

When $\Delta\phi_{\max}$ is small compared with the phase-noise standard deviation $\sigma_{\Delta\phi}$, layover is unlikely. Our model PDF $f(\Delta\phi^{(r)}|I, \rho)$ is then Gaussian with mean $\Delta\phi_I$, the phase value suggested by the single-pixel intensity. The cost function $g^{(r)}(\Delta\phi^{(r)})$ is therefore a parabola centered on $\Delta\phi_I$ whose width is $\sigma_{\Delta\phi}$ (see Fig. 10). Since variances sum,

$$\sigma^2_{\Delta\phi} = \sigma^2_{\Delta\psi} + \sigma^2_{\text{meas}}. \tag{19}$$

The first term on the right-hand side represents the phase-noise variance calculated from the measured correlation (Fig. 3), and the second term is a constant representing uncertainty in our estimates of $E[I]$ and ρ.

When the upper cutoff $\Delta\phi_{\max}$ is large compared with $\sigma_{\Delta\phi}$, the cost function must include a shelflike region that accounts for the probability of layover. This region extends out to $\Delta\phi_{\max}$ for positive gradient values, and its cost is based on the conditional probability of layover. In our implementation, we assume a constant layover cost $g^{(r)}_{\text{lay}}$ whose value we derive empirically. The central parabola of our layover cost function is centered on $\Delta\phi = 0$, our assumed mean of $\Delta\phi_I$ and $\Delta\phi_{\text{back}}$. The variance $\sigma^2_{\Delta\phi}$ includes an extra term σ^2_{lay}, which represents uncertainty in the location of the true peak due to layover:

$$\sigma^2_{\Delta\phi} = \sigma^2_{\Delta\psi} + \sigma^2_{\text{meas}} + \sigma^2_{\text{lay}}. \tag{20}$$

Beyond $\Delta\phi_{\max}$, the cost function increases quadratically at a rate related to $\sigma_{\Delta\phi}$.

Thus the cost function for a particular range gradient is illustrated in Fig. 10 and may be quantified as follows. Let $\Delta\phi^{(r)}_{\text{crit}}$ equal $[g^{(r)}_{\text{lay}}\sigma^2_{\Delta\phi}]^{1/2}$ with $\sigma^2_{\Delta\phi}$ calculated from Eq. (20). If $\Delta\phi_{\max} \geqslant \Delta\phi^{(r)}_{\text{crit}}$,

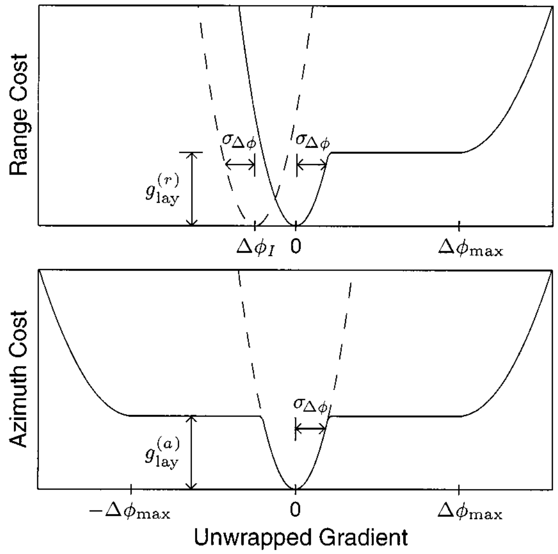
$$g^{(r)}(\Delta\phi) = \begin{cases} \dfrac{\Delta\phi^2}{\sigma^2_{\Delta\phi}} & \text{if } \Delta\phi \leqslant \Delta\phi^{(r)}_{\text{crit}} \\ g^{(r)}_{\text{lay}} & \text{if } \Delta\phi^{(r)}_{\text{crit}} < \Delta\phi \leqslant \Delta\phi_{\max}, \\ \dfrac{(\Delta\phi - \Delta\phi_{\max})^2}{c\,\sigma^2_{\Delta\phi}} + g^{(r)}_{\text{lay}} & \text{if } \Delta\phi > \Delta\phi_{\max} \end{cases} \tag{21}$$

where c is a constant. Conversely, if $\Delta\phi_{\max} < \Delta\phi^{(r)}_{\text{crit}}$, we define $\sigma^2_{\Delta\phi}$ according to Eq. (19) and assume that

$$g^{(r)}(\Delta\phi) = \frac{(\Delta\phi - \Delta\phi_I)^2}{\sigma^2_{\Delta\phi}}. \tag{22}$$

We have thus far considered cost functions only for unwrapped gradients in range, yet gradients in azimuth are equally important. Because the physical mechanisms behind the two are similar, their cost functions have similar shapes. The parameter $\sigma^2_{\Delta\phi}$ is the same for both range and azimuth gradients because the same noise processes affect both. The parameter $\Delta\phi_{\max}$ is also the same since layover-induced discontinuities often have both range and azimuth components. The layover-discontinuity shelf cost for azimuth, $g^{(a)}_{\text{lay}}$, may be greater than $g^{(r)}_{\text{lay}}$, though, as discontinuities may be more likely across range gradients.

The most important difference between range and azimuth cost functions, however, is the two-sided nature of the azimuth cost functions. This symmetry follows from the azimuthal symmetry of the SAR imaging geometry. Accounting for it, we quantify the azimuth cost functions as follows. Let $\Delta\phi^{(a)}_{\text{crit}}$ equal $[g^{(a)}_{\text{lay}}\sigma^2_{\Delta\phi}]^{1/2}$ with $\sigma^2_{\Delta\phi}$ calculated from Eq. (20). If $\Delta\phi_{\max} \geqslant \Delta\phi^{(a)}_{\text{crit}}$,



Parameter	Physical Source
$\sigma_{\Delta\phi}$	ρ
$\Delta\phi_{\max}$	integrated $E[I]$, ρ
$\Delta\phi_I$	single-pixel $E[I]$
g_{lay}	conditional layover probability

Fig. 10. Example cost functions for unwrapped topographic range (top) and azimuth (bottom) gradients in the presence (solid curves) and absence (dashed curves) of layover. Note that the abscissa is the unwrapped gradient $\Delta\phi$ itself, not $\Delta\phi - \Delta\psi$. The model parameters are based on the physical observables as shown and differ throughout the interferogram.

$$g^{(a)}(\Delta\phi) = \begin{cases} \dfrac{\Delta\phi^2}{\sigma_{\Delta\phi}^2} & \text{if } |\Delta\phi| \leq \Delta\phi_{\text{crit}}^{(a)} \\ g_{\text{lay}}^{(a)} & \text{if } \Delta\phi_{\text{crit}}^{(a)} < |\Delta\phi| \leq \Delta\phi_{\max}. \\ \dfrac{(|\Delta\phi| - \Delta\phi_{\max})^2}{c\,\sigma_{\Delta\phi}^2} + g_{\text{lay}}^{(a)} & \text{if } |\Delta\phi| > \Delta\phi_{\max} \end{cases} \tag{23}$$

When $\Delta\phi_{\max} < \Delta\phi_{\text{crit}}^{(a)}$, layover-induced discontinuities are unlikely, and

$$g^{(a)}(\Delta\phi) = \frac{\Delta\phi^2}{\sigma_{\Delta\phi}^2}, \tag{24}$$

where $\sigma_{\Delta\phi}^2$ is calculated from Eq. (19).

Although this cost model does suggest that some azimuth gradients will indeed be very large, it does not necessarily invalidate our assumption above of small azimuth gradients. This assumption holds as long as local range gradients are more significant than their azimuth counterparts. Moreover, as explained above, gradients that straddle layover discontinuities do not affect the intensity and correlation in proportion to their magnitude.

As shown in Fig. 10 (dashed curves), the layover-absent cost functions for both range and azimuth consist simply of parabolas with widths determined by the local correlation. These cost functions are thus similar in shape to weighted least-squares (L^2) cost functions, although here we assume congruence between the unwrapped and wrapped solutions. On the other hand, in the presence of layover, the cost functions are somewhat similar to L^0 cost functions. Our MAP cost functions therefore receive some validation from both the success of existing least-squares algorithms on noisy interferograms lacking layover and the success of L^0 algorithms where layover is present.

Note, however, that our cost functions are not always centered on the wrapped gradients $\Delta\psi$ as L^p cost functions are. Having modeled the statistics of a particular unwrapped gradient, we might expect the likelihood of an extra cycle of phase to be much higher in one direction than the other. For example, consider a wrapped gradient $\Delta\psi = 0.4$ cycle. If we believe the topographic slope to be zero ($\Delta\phi_{\text{topo}} = 0$), it is much more probable that the true unwrapped gradient $\Delta\phi$ is wrapped from -0.6 than from 1.4 cycle. Our generalized cost functions thus nicely treat unwrapped phase gradients as random draws from fixed probability distributions. Moreover, although above we use cost functions centered at either $\Delta\phi = \Delta\phi_I$ or $\Delta\phi = 0$, more elaborate models are possible for the cost-function minima. They might, for example, be determined by averaging wrapped gradients over local neighborhoods.

While our statistical cost functions establish topographic phase unwrapping as a nonlinear optimization problem, we note again that application-specific assumptions were used in their development. Thus, while aiming to increase accuracy for one particular phase-unwrapping problem, we sacrifice general applicability to others. We must therefore develop additional cost functions specific to other phase-unwrapping problems as well.

B. Deformation Measurements

Although differential SAR interferometry is closely related to the topographic application described above, it entails different phase statistics and therefore requires a new set of cost functions. In differential interferometry, phase is used to measure surface deformation that may be due to phenomena such as earthquakes, volcanism, or glacial flow. Differential interferograms often contain fewer abrupt changes in unwrapped phase than do topographic interferograms, but very large phase discontinuities are still possible because of surface fracture and shear. Moreover, although low fringe rates may sometimes make differential interferograms relatively easy to unwrap away from these discontinuities, the areas near discontinuities are often of most interest.

We begin our derivation of deformation cost functions by again separating the unwrapped-gradient PDF's into their signal and noise parts:

$$f(\Delta\phi|I,\,\rho) = f(\Delta\phi_{\text{defo}}|I,\,\rho) * f(\Delta\phi_{\text{noise}}|\rho). \tag{25}$$

We assume that the phase noise has the same conditional statistics as in the topography case above. Consequently, we focus on the actual deformation signature, given by[2]

$$\phi_{\text{defo}} = \frac{-4\pi d_r}{\lambda}, \tag{26}$$

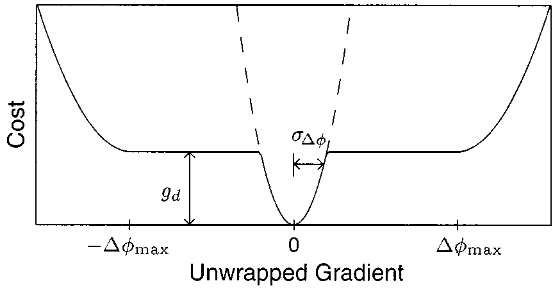
where d_r is the surface displacement parallel to the radar line of sight (other components of displacement are not measurable with this technique).

The unwrapped deformation gradient does not depend on intensity in any obvious way. Although there is sometimes correspondence between topographic features and areas of large deformation changes, this is not always the case. Moreover, while deformation-induced decorrelation lessens an interferogram's magnitude, it has no effect on the individual SAR intensities. Consequently, without further insight into the expected deformation pattern, we assume that the deformation signal is independent of the intensity, so $f(\Delta\phi_{\text{defo}}|I,\,\rho) = f(\Delta\phi_{\text{defo}}|\rho)$.

On the other hand, the relationship between deformation and correlation is much more explicit. Large displacement changes imply distortions of the surface, so they are likely accompanied by significant decorrelation through such mechanisms as temporal scattering changes and local misregistration. We therefore account for the probability of a discontinuity where the correlation is low by using a simple discontinuity model similar to our topographic layover model.

When the phase-noise PDF's and the unwrapped-gradient PDF's are combined as in Eq. (25), the deformation cost functions are thus shaped as illustrated in Fig. 11, where now range and azimuth cost functions have identical shapes. These generic deformation cost functions are quantified analogously to our topographic azimuth cost functions of Eqs. (23) and (24), although different physical quantities are used for our deformation model parameters. Specifically, if the local correlation exceeds some threshold $\rho_{\min}$, the cost function $g(\,\cdot\,)$ there consists only of a parabola centered at zero, whose width is calculated from Eq. (19). Where the correlation falls below $\rho_{\min}$, the cost function also includes shelflike regions representing the probability of a discontinuity in the unwrapped phase. The heights of these shelves, denoted by g_d, are related to the local conditional probability of a phase discontinuity. The maximum probable magnitude for this discontinuity determines $\Delta\phi_{\max}$, the upper shelf cutoff. Without more specific information, we use empirically or experimentally derived constants for $\rho_{\min}$, g_d, and $\Delta\phi_{\max}$. For the results below, these parameters were initially estimated through visual inspection of the interferogram and then refined with further applications of the algorithm.

Thus our deformation cost functions share much with minimum L^p-norm weighting schemes based on thresholded correlation values.[6,14] Our cost functions differ, however, in that not only are they scaled, but their shapes change in relation to our changing confidence in the unwrapped gradient. Furthermore, the upper cutoffs of the discontinuity regions can prevent overestimation of discontinuities, avoiding some global unwrapping errors. Congruence and the asymmetry of our cost functions about the wrapped gradient $\Delta\psi$ should also yield advantages in accuracy, as described above.



Parameter	Physical Source
$\sigma_{\Delta\phi}$	ρ
$\Delta\phi_{\max}$	ρ, assumed maximum discontinuity
g_d	conditional discontinuity probability

Fig. 11. Example cost functions for unwrapped differential phase gradients when a discontinuity is expected (solid curve) and not expected (dashed curve).

For both the topography and the deformation cases, our models include a number of parameters, providing great flexibility and customizability for specific applications. Together with the solver routine described in Appendix A, we call our algorithm SNAPHU, an acronym for "statistical-cost network-flow algorithm for phase unwrapping." We examine its performance in the following section.

3. RESULTS

Our derivation of statistical cost functions followed our goal of developing a useful, working algorithm. We now test this algorithm on actual rather than simulated interferometric SAR data, examining how well it retrieves unwrapped phase fields in both the topography and the deformation cases.

Our test topographic interferogram is formed from two images, acquired 105 days apart, from the European Space Agency's European Remote Sensing-1 satellite. The average intensity from the individual SAR images is shown in Fig. 5(a), and the interferogram is shown in Fig. 12(a) with magnitude displayed in gray scale and phase displayed in color. Radar illumination is from the left with range increasing toward the right. This 1250 × 830 pixel interferogram, formed from five looks in azimuth and a single look in range, depicts a desert region north of Death Valley, California. Shown in Fig. 12(b) is the biased interferogram coherence estimate $\hat{\rho}$, as calculated from Eq. (9) with twenty looks in azimuth and four looks in range. As a topographic reference, we use the

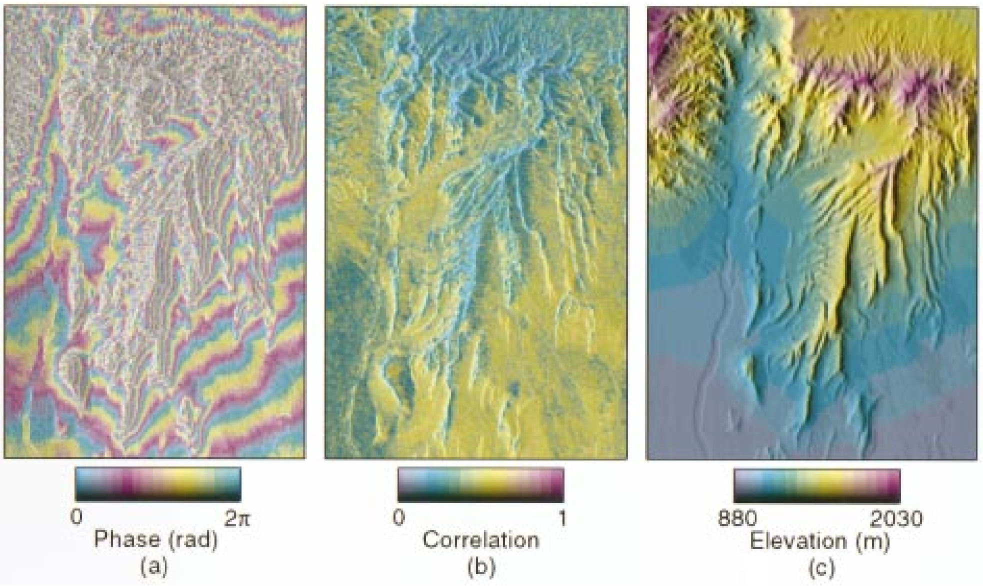


Fig. 12. Topographic test data: (a) interferogram with wrapped phase in color and magnitude in gray scale, (b) biased coherence magnitude, (c) reference DEM with elevation in color and shaded relief in gray scale.

30-m posting U.S. Geological Survey DEM shown in Fig. 12(c), where the elevation is represented in color and a shaded-relief image generated from this DEM is shown in gray scale. The DEM accuracy of 7.5-m rms is sufficient for the interferogram ambiguity height of approximately 80 m.

The variety of topographic features in this interferogram allows us to analyze several problems often apparent in topographic phase unwrapping. Running in azimuth in the middle of the image are long discontinuities resulting from layover, while the top of the image contains areas of rough topography. The bottom is relatively smooth, although it is not without areas of low correlation.

Results from our algorithm are shown in Fig. 13 alongside results from several other algorithms. Here, the color represents relative unwrapped phase error, calculated by subtracting the DEM-derived unambiguous phase from the unwrapped solutions. The interferogram magnitude is again shown in gray scale. Since these algorithms all produce congruent solutions, unwrapping errors can be easily identified as patches differing from their surroundings by integer numbers of cycles. Other differences of less than one cycle may be due to atmospheric artifacts,[30] inaccuracies in the DEM, noise in the interferogram, or artifacts from transforming and registering the DEM to radar coordinates.

SNAPHU results are shown in Fig. 13(a). As is evident from the general homogeneity of color, our algorithm performs well, with errors confined predominantly to layover regions. Overall, 94% of the pixels in the solution are within π rad of the reference phase, and the rms error is 1.94 rad (0.31 cycles). Since few 2π jumps are apparent in Fig. 13(a), most of these errors can be ascribed not to the phase-unwrapping process but to the sources described above.

Chen and Zebker[11] applied other algorithms to the same data set, and those results are reproduced here in Figs. 13(b)–13(d). Figure 13(b) shows results from the Goldstein *et al.*[10] residue-cut algorithm, which performs well where it does unwrap but fails to produce a solution for approximately half the interferogram (shown in black). Figure 13(c) shows results from an L^1 MCF algorithm[9] with edge-detection weights.[11] The MCF solution is good, but the rough area near the top of the interferogram remains incorrectly unwrapped. Figure 13(d) shows results from a transform-based least-squares algorithm with correlation weights and congruence enforced after optimization. Despite the similarities between our statistical cost functions and those of the correlation-weighted L^2 norm, the poor performance of the latter underscores the need for more physical cost functions where the topography is rough. It also highlights the disadvantages of enforcing congruence after rather than during optimization. The SNAPHU solution is clearly the most accurate of those compared here.

While the DEM provides reliable ground truth for the topography data, reference data are more difficult to come by in differential interferometry. However, the subtleties and nuances of real data are also more difficult to simulate faithfully in the latter case, so we again use real data for testing our algorithm. We evaluate the unwrapped solution's quality subjectively, examining its geophysical plausibility.

The 2380 × 2548 pixel interferogram shown in Fig. 14(a) depicts the deformation signature resulting from the M7.1 Hector Mine, California earthquake of October 16, 1999. As elsewhere in the paper, the interferogram magnitude is indicated by gray level and the phase by color. Here, rather than topographic contour intervals, each color cycle represents 2.8 cm of relative surface displacement. The data were acquired by ERS-2 during orbits 23027 and 23528, with a perpendicular baseline of 23 m. Topographic phase variations have been removed as much as possible through the use of a topographic ERS tandem interferogram (ERS-1 orbit 24664 and ERS-2 orbit 4991). Note that the fringe rates are in the same direction on either side of the earthquake fault, so across the fault the phase changes by tens of cycles over a narrow spatial region. The peak deformation is greater than

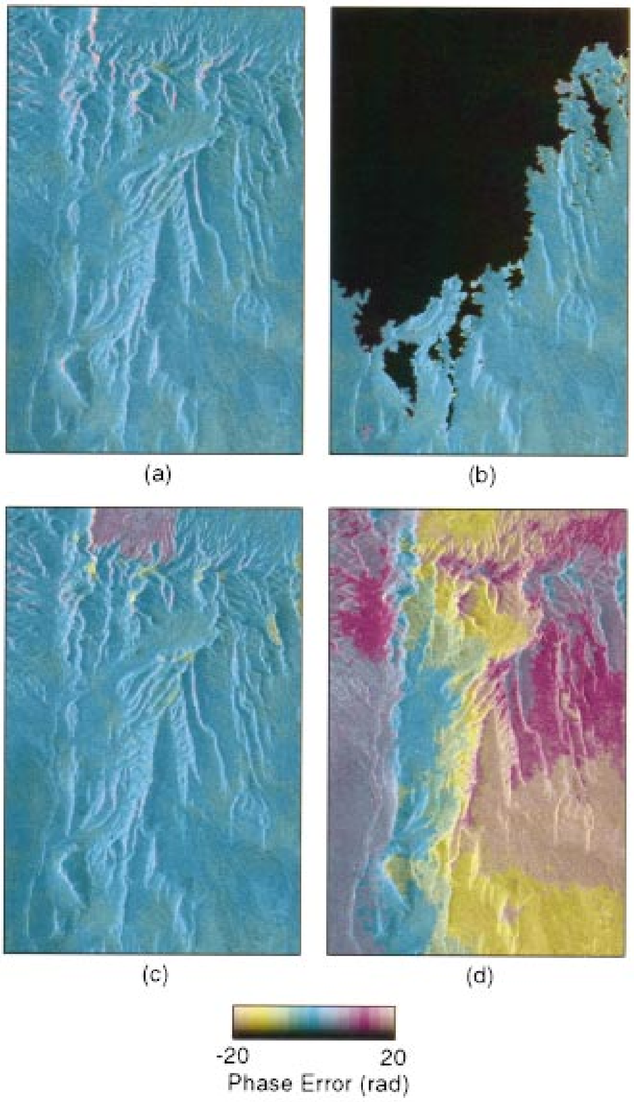


Fig. 13. Algorithm results on the topographic test interferogram of Fig. 12: (a) SNAPHU, (b) reside-cut, (c) MCF, (d) least-squares. The gray scale depicts the interferogram magnitude, and the color represents relative unwrapped phase error with reference to the DEM. Unwrapping errors are manifest as jumps of 2π rad.

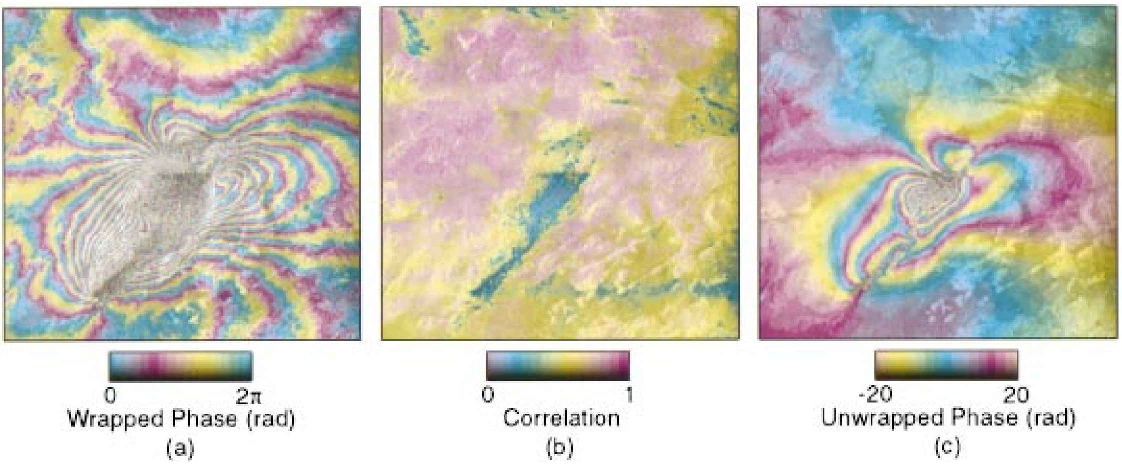


Fig. 14. Results of our algorithm on a differential interferogram: (a) interferogram with wrapped phase in color and magnitude in gray scale, (b) biased coherence magnitude, (c) unwrapped solution from our algorithm, rewrapped modulo 40 rad (6.37 cycles) for display.

20 cycles, or 5.6 m. While some areas along the fault generally appear darker in the interferogram magnitude, these dark areas are likely due to correlation effects. The individual SAR intensity images (not shown) do not suggest any easily discernible relationship between deformation and surface brightness at fine spatial scales. The fault, however, is plainly visible in the interferogram coherence image of Fig. 14(b); as the fringe rates increase toward the fault, the interferogram loses coherence. The coherence estimate shown here is formed from twenty looks in azimuth and four looks in range, while ten looks in azimuth and two looks in range are incorporated into the interferogram. It should be noted that speckle is present in the intensity and magnitude images of Figs. 5 and 12–14, though it may not be visible at the resolutions used for reproduction.

The solution from our algorithm is shown in Fig. 14(c) with the unwrapped phase shown rewrapped modulo 40 rad (6.37 cycles) for display purposes. Away from the fault where fringe rates are low, the interferogram is easy to unwrap, and virtually any algorithm would be expected to perform well. Nearer the fault, the fringe rates increase and coherence decreases, but the solution still does not present any obvious errors. That is, the algorithm performs as well as or better than the authors unwrapping the interferogram by eye. The relatively smooth unwrapped solution is consistent with the intuition that shear or fracture in the surface should be limited to relatively few places, mostly along the fault. By allowing large phase gradients and discontinuities at these locations, the algorithm also appears to capture the high-spatial-frequency information of the interferogram; it thus avoids the distortions characteristic of algorithms prone to excessive smoothing or slope underestimation.[17,18] Where coherence is lost completely in the interferogram, our algorithm produces a smooth, reasonable guess at the unwrapped phase field based on the surrounding valid areas. The overall accuracy of these deformation measurements thus seems to be limited more by the quality of the interferogram itself rather than by our phase-unwrapping algorithm.

The execution times of our algorithm on a 550-MHz Intel Pentium III processor were approximately 100 and 1500 s for the topography and deformation cases (recall that the latter interferogram is larger), and their associated memory requirements were 96 and 490 MB. Execution times for other data sets varied greatly, depending on factors such as the number and density of residues in the interferogram, the values used for model parameters, and the quality of the initialization.

4. CONCLUSIONS

Phase unwrapping, arguably as much an art as a science, is an attempt to solve a problem that we know to be unsolvable. It is more than just an exercise, however, as the difficulty of the problem does not diminish the need for reliable solutions. Therefore, in order to obtain such solutions, we propose a MAP methodology for the phase-unwrapping problem. We introduce simple statistical models and employ nonlinear cost functions to describe the probabilities of particular unwrapped solutions given the wrapped phase, image intensity, and interferogram coherence. Recognizing that these cost functions constitute a nonlinear optimization problem, we also develop network-flow techniques for finding the most physically probable unwrapped solution.

Our tests suggest great promise with this approach. On a topographic SAR interferogram characterized by layover, rough terrain, and low coherence, our algorithm produces a solution in excellent agreement with our reference DEM and significantly more accurate than those of other algorithms applied to the same data set. For our differential test interferogram, no errors are apparent on visual inspection of the solution, although quantitative reference data are unavailable; the long phase discontinuity and high fringe rates associated with the earthquake fault appear to be unwrapped in a geophysically consistent manner, however. Thus, while there is indeed room for refinement in our models both theoretically and empirically, our algorithm's accuracy and reasonable efficiency make it well worth considering for specific unwrapping applications.

The MAP framework implies, however, that no single algorithm will be best for all applications without modification. Since different physical quantities are involved for different applications, each application should have its own statistical models. Thus, while we might desire the simplicity of a single, grand algorithm that is universally applicable, realistic algorithms must generally make specific assumptions about their inputs to achieve the finest overall accuracy.

Assumptions involved in setting up the problem cannot be made independently of the method of solution, however, as an objective that is unsolvable is no more useful than an objective that gives unmeaningful results when solved. In our development, this duality guides the balance between theoretical rigor and computational manageability. This balance, however, might need to be readjusted for the specifics of other applications. In this light, L^p objectives can be seen as easily solvable, coarse approximations to more specific MAP optimization goals. Hence existing, efficient L^p approaches can benefit from weights based on statistical models. Conversely, the approximations used in this paper may be refined to enhance analytical precision at the expense of algorithm complexity. Ultimately, many phase-unwrapping approaches are possible, and it is up to the user to decide which is best for his or her particular needs.

APPENDIX A: APPROXIMATE NONLINEAR NETWORK OPTIMIZATION

In this appendix we develop techniques for approximately solving the minimization problem of Eq. (1) with arbitrarily shaped, generalized cost functions $g(\,\cdot\,)$. First we describe how the DCC algorithm introduced by Chen and Zebker[11] may be adapted to solve the generalized problem, and then we describe a more efficient hybrid algorithm. Readers are encouraged to examine the references for background on network theory and its application to phase unwrapping.[9,11,13,31]

As described in Section 1, any wrapped phase field implies a gridlike network (Fig. 1), and any unwrapped solution for it corresponds to a unique feasible flow. The iterative DCC algorithm, similar to Flynn's algorithm,[8] maintains the feasibility of its intermediate solutions by augmenting flow on closed cycles or loops, thereby preserving flow conservation. The algorithm improves the current solution by selecting cycles that have net negative residual or incremental flow costs. Suppose arc a on the network has x_0 units of flow on it and has the cost function $g(x)$. The addition of δ units of flow to a results in a total cost change of $g(x_0 + \delta) - g(x_0)$. For nonlinear cost functions, these residual costs depend on both x_0 and δ, so they must be recalculated after each flow augmentation. The DCC algorithm does exactly this, allowing it to find approximate L^0 solutions. Exact solutions cannot be found, as expected given the *NP*-hardness of the problem.

Though intended for the L^0 objective function, the DCC algorithm may in fact be used with arbitrary cost functions as long as residual costs are calculated appropriately. That is, for an arbitrary cost function $g(x)$, the residual cost c_a for arc a is simply the incremental cost

$$c_a(x_0,\ \delta) = g(x_0 + \delta) - g(x_0), \qquad \text{(A1)}$$

where x_0 is the existing flow on the arc and δ is the flow increment. Since the cost functions are evaluated independently and then summed, as in Eq. (1), the functions $g(\,\cdot\,)$ may take independent, arbitrary shapes—including those defined by L^p norms. If all cost functions are convex, the solution will be exact;[13] otherwise, it will be approximate.

We now apply ideas from efficient linear (MCF or L^1) approaches to the congruent, nonlinear, generalized-cost case. The network simplex algorithm[31] iteratively improves intermediate spanning-tree solutions in which each node is labeled by a potential, the residual cost of sending δ units of flow up the tree to the tree root. A cycle is formed with the addition of any nontree arc to the tree. Thus, if the tree is nonoptimal, a pivot is performed: A nontree arc is added to the tree, flow is possibly augmented on the resulting cycle, and one arc on the cycle is dropped, resulting in a new, better spanning tree. The algorithm pivots from one spanning tree to another, improving node potentials and canceling negative-cost cycles as it goes.

We use this spanning-tree approach in our hybrid algorithm to find negative cycles. As with the DCC algorithm, residual costs are calculated in terms of the current arc flows and the flow increment δ. Because of the nonlinearity of generalized costs, each node is assigned two potentials: One is the residual cost of sending δ units of flow along the tree to the root, and the other is the residual cost of the same flow in the opposite direction. Additionally, each nontree arc a is also associated with an apex node, the node closest to the root on the cycle formed by adding a to the current spanning tree. Consider nodes p and q, and the nontree arc a going from p to q. Let m be the apex node for a, and let $\pi_{\text{in}}^{(\,\cdot\,)}$ and $\pi_{\text{out}}^{(\,\cdot\,)}$ denote inward and outward potentials. The addition of arc a results in a negative cycle if

$$(\pi_{\text{out}}^{(p)} - \pi_{\text{out}}^{(m)}) + (\pi_{\text{in}}^{(q)} - \pi_{\text{in}}^{(m)}) + c_a < 0. \qquad \text{(A2)}$$

Because pivots that make one set of potentials better may make the other set worse, our implementation performs only pivots that augment flow on negative cycles or that improve outward, not inward, potentials $(\pi_{\text{out}}^{(p)} + c_a - \pi_{\text{out}}^{(q)} < 0)$. After each pivot, the potentials and the tree structure are updated as in the network simplex algorithm.[31] Affected apex pointers must be reset as well.

Further speed improvements can be made by adopting a tree growth strategy similar to the one used by Pallottino.[32] Applying this idea to our algorithm, we begin with an arbitrary root node and grow a spanning tree T, maintaining its optimality over each major algorithm iteration. During each major iteration, a set number of new nodes is first added to T in the order prescribed by Dijkstra's algorithm.[13] Then, with use of pivots, T is reoptimized in terms of the nodes it spans. After T spans all nodes, the algorithm terminates.

ACKNOWLEDGMENTS

We thank Margrit Gelautz and E. Weber Hoen for their help and insights during discussions of this research. The anonymous reviewers also provided valuable feedback. We are grateful for data provided by the European Space Agency under several European Remote Sensing announcements of opportunity. This work was supported under a contract from the University of Alaska Geophysical Institute.

Curtis Chen's e-mail address is curtis@nova.stanford.edu.

REFERENCES

1. H. Zebker and R. Goldstein, "Topographic mapping from interferometric SAR observations," J. Geophys. Res. **91**, 4993–4999 (1986).
2. A. K. Gabriel, R. M. Goldstein, and H. A. Zebker, "Mapping small elevation changes over large areas: differential radar interferometry," J. Geophys. Res. **94**, 9183–9191 (1989).
3. R. M. Goldstein and H. A. Zebker, "Interferometric radar measurements of ocean surface currents," Nature (London) **328**, 707–709 (1987).
4. D. C. Ghiglia and L. A. Romero, "Minimum L^p-norm two-dimensional phase unwrapping," J. Opt. Soc. Am. A **13**, 1999–2013 (1996).
5. D. C. Ghiglia and L. A. Romero, "Robust two-dimensional weighted and unweighted phase unwrapping that uses fast transforms and iterative methods," J. Opt. Soc. Am. A **11**, 107–117 (1994).
6. M. D. Pritt, "Phase unwrapping by means of multigrid techniques for interferometric SAR," IEEE Trans. Geosci. Remote Sens. **34**, 728–738 (1996).
7. G. Fornaro, G. Franceschetti, R. Lanari, and E. Sansosti, "Robust phase unwrapping techniques: a comparison," J. Opt. Soc. Am. A **13**, 2355–2366 (1996).
8. T. J. Flynn, "Two-dimensional phase unwrapping with minimum weighted discontinuity," J. Opt. Soc. Am. A **14**, 2692–2701 (1997).
9. M. Costantini, "A novel phase unwrapping method based on network programming," IEEE Trans. Geosci. Remote Sens. **36**, 813–821 (1998).
10. R. M. Goldstein, H. A. Zebker, and C. L. Werner, "Satellite radar interferometry: two-dimensional phase unwrapping," Radio Sci. **23**, 713–720 (1988).
11. C. W. Chen and H. A. Zebker, "Network approaches to two-dimensional phase unwrapping: intractability and two new algorithms," J. Opt. Soc. Am. A **17**, 401–414 (2000).
12. J. R. Buckland, J. M. Huntley, and S. R. E. Turner, "Unwrapping noisy phase maps by use of a minimum-cost-matching algorithm," Appl. Opt. **34**, 5100–5108 (1995).
13. R. K. Ahuja, T. L. Magnanti, and J. B. Orlin, *Network Flows: Theory, Algorithms, and Applications* (Prentice-Hall, Englewood Cliffs, N.J., 1993).
14. D. C. Ghiglia and M. D. Pritt, *Two-Dimensional Phase Unwrapping: Theory, Algorithms, and Software* (Wiley, New York, 1998).
15. M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness* (Freeman, San Francisco, Calif., 1979).
16. G. Carballo, "Statistically-based multiresolution network flow phase unwrapping for SAR interferometry," Ph.D. dissertation (Royal Institute of Technology, Stockholm, Sweden, 2000).
17. R. Bamler, N. Adam, G. W. Davidson, and D. Just, "Noise-induced slope distortion in 2-D phase unwrapping by linear estimators with application to SAR interferometry," IEEE Trans. Geosci. Remote Sens. **36**, 913–921 (1998).
18. H. A. Zebker and Y. Lu, "Phase unwrapping algorithms for radar interferometry: residue-cut, least-squares, and synthesis algorithms," J. Opt. Soc. Am. A **15**, 586–598 (1998).
19. H. A. Zebker and J. Villasenor, "Decorrelation in interferometric radar echoes," IEEE Trans. Geosci. Remote Sens. **30**, 950–959 (1992).
20. J. S. Lee, K. W. Hoppel, S. A. Mango, and A. R. Miller, "Intensity and phase statistics of multilook polarimetric and interferometric SAR imagery," IEEE Trans. Geosci. Remote Sens. **32**, 1017–1028 (1994).
21. F. K. Li and R. M. Goldstein, "Studies of multibaseline spaceborne interferometric synthetic aperture radars," IEEE Trans. Geosci. Remote Sens. **28**, 88–97 (1990).
22. E. Rodriguez and J. M. Martin, "Theory and design of interferometric synthetic aperture radars," IEE Proc. F, Commun. Radar Signal Process. **139**, 147–159 (1992).
23. R. Touzi, A. Lopes, J. Bruniquel, and P. W. Vachon, "Coherence estimation for SAR imagery," IEEE Trans. Geosci. Remote Sens. **37**, 135–149 (1999).
24. B. Guindon, "Development of a shape-from-shading technique for the extraction of topographic models from individual spaceborne SAR images," IEEE Trans. Geosci. Remote Sens. **28**, 654–661 (1990).
25. D. J. Goering, H. Chen, L. D. Hinzman, and D. L. Kane, "Removal of terrain effects from SAR satellite imagery of arctic tundra," IEEE Trans. Geosci. Remote Sens. **33**, 185–194 (1995).
26. A. Lopes, E. Nezry, R. Touzi, and H. Laur, "Structure detection and statistical adaptive speckle filtering in SAR images," Int. J. Remote Sens. **14**, 1735–1758 (1993).
27. C. J. Oliver and S. Quegan, *Understanding Synthetic Aperture Radar Images* (Artech House, Boston, Mass., 1998).
28. L. M. H. Ulander, "Radiometric slope correction of synthetic aperture radar images," IEEE Trans. Geosci. Remote Sens. **34**, 1115–1122 (1996).
29. F. T. Ulaby, R. K. Moore, and A. K. Fung, *Microwave Remote Sensing, Active and Passive* (Addison-Wesley, London, 1981).
30. H. A. Zebker, P. A. Rosen, and S. Hensley, "Atmospheric effects in interferometric synthetic aperture radar surface deformation and topography maps," J. Geophys. Res. **102**, 7547–7563 (1997).
31. J. L. Kennington and R. V. Helgason, *Algorithms for Network Programming* (Wiley, New York, 1980).
32. S. Pallottino, "Shortest-path methods: complexity, interrelations and new propositions," Networks **14**, 257–267 (1984).