Information contents of fracture lines

H. C. G. Leitão[1] **J. Stolfi**[2]

[1] Institute of Computing, Fluminense Federal University
Niterói, RJ, Brazil — hcgl@pgcc.uff.br

[2] Institute of Computing,University of Campinas
13083-970 Campinas, SP, Brazil — stolfi@dcc.unicamp.br

ABSTRACT

Reassembling unknown broken objects from a large collection of fragments is a common problem in archaeology and other fields. Computer tools have recently been developed by the authors and by others, which try to help by locating pairs of fragments with matching outline shapes. These tools have been succesfully tested on small collections of fragments. Here we address the question of whether such tools can be expected to work for practical instances of the problem (10^3 to 10^5 fragments). To that end, we describe here a method to measure the average amount of information contained in the shape of a fracture line of given length. This parameter tells us how many false matches we can expect to find for that fracture among a given set of fragments; and we show that outline comparison should give useful results even for large instances.

Keywords: fractures, pattern recognition, contour matching, fractals

1 Introduction

Reassembling unknown broken objects from a large collection of irregular fragments is a problem that arises in several contexts, such as archaeology (ceramic vessels), failure analysis (debris), paleontology (fossil bones), conservation (mural paintings), and so on. Large instances of the problem—involving tens of thousands of of randomly shaped and featureless fragments—are not uncommon, and their reassembly often requires years of tedious and delicate work. The most difficult part of the problem is finding the pairs of *matching fragments*, those that were adjacent in the original object.

This problem could obvioulsy use some computer help; and indeed programs have been developed, by us and by others [Gama 98, Üçolu97, Burde89] that, given a set fragment outlines like the ones shown in figure 1, can identify a substantial fraction of the matching pairs, as shown in figure 2, at reasonable computing cost.

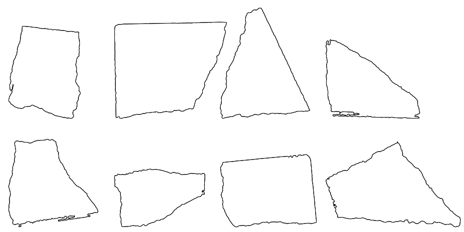


Figure 1: Digitized outlines of ceramic fragments.

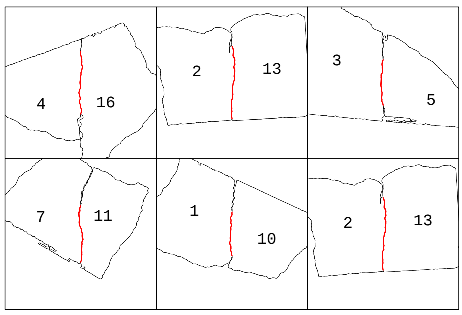


Figure 2: Some matching fragment pairs.

1.1 Scaling up to large problems

While computer matching has proven effective for small instances of the problem (about a hundred fragments), it is not obvious that it will work for realistic instances, with 10^3 to 10^5 fragments.

One may wonder that, among such large collections, there will be far too many "false positives"—pairs that were not adjacent in the original object, but whose outlines have the same shape, just by chance. A program that produced thousands of false matches for each fragment would not be of much help.

For a rough analysis of the question, suppose we have N fragment outlines, with average perimeter L. For a given point p on the boundary of one fragment, there are $O(NL)$ points on other fragments that could be matched to p in the reconstructed object. In order to identify the correct match q, we need to extract $\log_2(NL) + O(1)$ bits of useful information from the shape of the outline in the neighborhood of p— "useful" in the sense that the same bits can be extracted, with high probability, from the outline around q. This observation is encouraging, in that it says that the amount of information required grows very slowly (logarithmically) with the size of the problem.

In fact, experience suggests that the shape of a ceramic fragment contains a lot of information about its matching partner. Anyone who has tried to put back together a broken vase knows that a correct pair of fragments, even relatively small ones, will "fit" together vastly better than an incorrect pair; so that the latter is hardly ever mistaken for the former. The reason is that, for suitable materials, the two sides of a fracture will remain congruent to within a fraction of a millimeter, for most of their length. Given their irregular, random shape, the probability of obtaining such a precise fit among two unrelated fragments is practically nil. See figures 3 and 4.

In this paper we try to turn the above intuition into a quantitative statement. Specifically, we describe a method for determining the average amount of useful information contained in a piece of fragment outline of given length, given a sample of correctly matched fragment outlines. This method is not used in the matching algorithm proper, which has been described elsewhere [Gama 99]. Its purpose is to enable *a priori* analysis of the effectiveness of shape comparison. The data provided by this method can be used to estimate the number of false matches that one can expect to find among a large collection of segments, the minimum length of common boundary that is needed for reliable matching, and the precision with which the outlines must be digitized, and so forth.

We illustrate the method with an artificial but fairly realistic sample of ceramic fragments. The information contents we observe (about 17.2 bits per centimeter) means that, for a fracture 1.1 cm long, we can expect about one false match every 200 fragments or so.

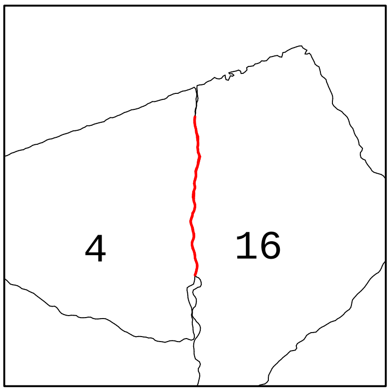


Figure 3: Two matching fragments – actual size.

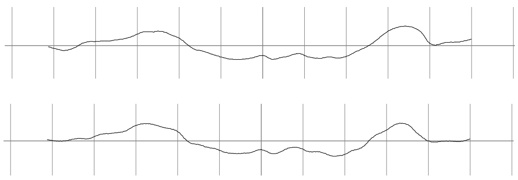
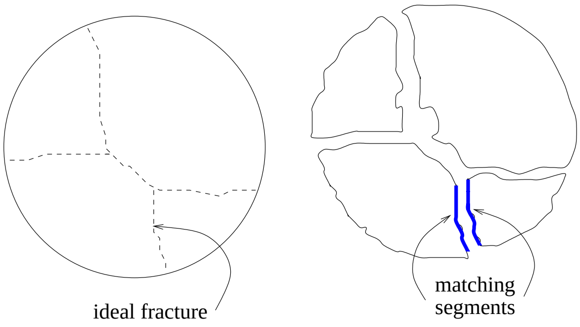


Figure 4: Corresponding parts of two matching fragments, magnified. Grid lines are 1 mm apart. The outlines were digitized at a nominal resolution of 300 dpi (0.085 mm/pixel) and smoothed with a Gaussian filter of characteristic length $\sigma = 0.085$ mm.

2 Fracture model

Our fragment matching algorithms are specialized for objects with a smooth and locally flat surface, such as tiles, plates, tablets, large vases, frescoes, etc. The algorithms' input consists of the digitized *fragment contours* or *outlines*, modeled as a collection of plane curves.

We assume that two fragments which were adjacent in the original object were separated there by an *ideal frature line* of zero thickness. The concrete manifestation of that line is a pair of *matching segments* on the contours of those two fragments. See figure 5. Note that, in general, it is not possible to identify the endpoints of these segments without knowing the matching fragment.



(a) (b)

Figure 5: Original object with ideal fracture lines (a) and the observed fragment contours (b).

Needless to say, two matching segments will never be precisely congruent: there will be some differences, either real (e.g. due to loss of small fragments) or artificial (due to errors in the contour extraction process, such as parallax, shadowing, quantization, etc.). The useful information contents of a piece of contour is determined by the magnitude of all these errors, relative to the size of the characteristic details that can be used to identify the matching piece.

3 Interpreting curves as signals

Before we can apply the tools of information theory to this problem, we must turn each curve into a *signal*—a real function of some real parameter t. The transformation must turn matching contour segments into similar signals, even if the fragments were digitized in random orientations.

A well-known rotation-invariant representation of a curve is the graph of its curvature $\kappa(t)$ as a function of its arc-length t from an arbitrary reference point. However, since the curvature is essentially a second derivative, it tends to magnify the effect of small-scale noise, and its shape looks quite different from the shape of the curve. While these defects may not be significant for the Fourier-based analysis below, it seemed prudent ot use representation as close as possible to the original curve itself.

Therefore, we have chosen to use a *shape function* derived from the curve segment as described below. We assume that the curve segment in question has length L and is given by $n + 1 = 2^k + 1$ equally spaced sample points $c_0, .. c_n$ on the plane. The shape function s is conceptually defined on the interval $[0 \,..\, L]$, and is computed as $n + 1$ real sample values $s_0, s_1, .. s_n$, with $s_0 = s_n = 0$, by this recursive procedure:

1. if $n = 0$, return $s_0 = 0$.
2. let r be the index of the middle sample, $r = n/2$. Recursively convert the sequences $c_0, .. c_r$ and $c_r, .. c_n$ to signals $s_0, .. s_r$ and $s_r, .. s_n$.
3. let α be the angle between the vectors $u = c_r - c_0$ and $v = c_n - c_r$ Add to the samples $s_0, .. s_n$ a sequence $g_0, .. g_n$ with $g_0 = g_n = 0$, $g_r = \alpha L/4$, and other values defined by linear interpolation between these values (i.e. a triangular pulse of height h). Return the sequence $s_0, .. s_n$.

This transformation is fully invertible: given the length L, the point c_0, the line c_0c_n, and the samples s_i, the original points c_i can be reconstructed by running the algorithm in reverse. This representation does not magnify the small-scale noise, and moreover it preserves qualitatively the shape of the curve. See figure 6.

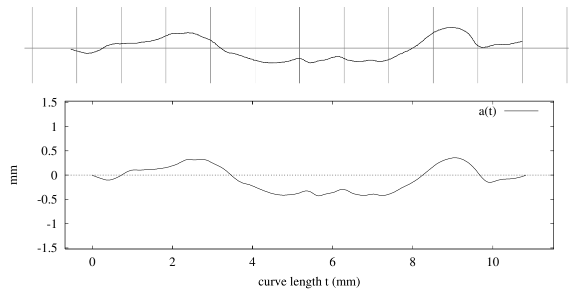


Figure 6: A curve and its shape function.

One drawback of this transformation is that a local disturbance in the curve may change its length, and therefore cause a global shift of the shape function from that point on. Nevertheless, one verifies experimentally that the shape functions of corresponding contour pieces, like the ones shown in figure 4, generally remain in sync over most of their length, as shown in figure 7.

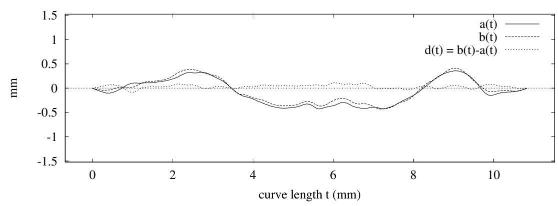


Figure 7: The shape functions of the two corresponding segments shown in figure 4.

4 Information contents of outlines

We can view a digitized fragment contour abstractly as a signal (the fracture line) corrupted by noise (the material losses and data acquisition errors). Specifically, the shapes of the two corresponding segments of a pair can be written as $a(t) = s(t) + n'(t)$ and $b(t) = s(t) + n''(t)$, where s is the shape of the ideal fracture line, and n', n'' are "noise" functions that represent loss of material, acquisition errors, etc.

Since the shape function $s(t)$ of a curve, as defined above, is known only at a finite set of uniformly spaced t values $t_j = j\delta$, for $0 \leq j < n$, we can express it as a *discrete Fourier series*

$$s_j = s(t_j) = \sum_{k=-m}^{m} S_k \exp(\frac{2\pi\mathbf{i}k}{T} t_j) \qquad (1)$$

where $T = n\delta$ is the *period*, $m = \lfloor n/2 \rfloor$ is the *maximum frequency*, and $\mathbf{i} = \sqrt{-1}$ is the imaginary unit. (Recall that the Fourier coefficients S_k of a real-valued series satisfy $S_{-k} = S_k^*$ for all k; moreover S_0 is real, and S_m is real when n is even. Therefore, we get exactly n degrees of freedom in the coefficients S_k.)

Let A_k, B_k, S_k, N'_k, and N''_k be the Fourier coefficients of a, b, s, n', and n'', respectively. We can generally assume that the coefficients S_k, N'_k, and N''_k are independent random variables with zero-mean, symmetric Gaussian distributions over the complex plane. We can assume also that N'_k and N''_k have the same variance $\hat{N}_k$. Then the information given by each coefficient A_k about the corresponding coefficient B_k [Lathi68] is

$$\begin{aligned} I_k &= \log\left[\frac{\hat{A}_k\hat{B}_k}{\hat{S}_k\hat{N}_k + \hat{A}_k\hat{N}_k}\right] \\ &= \log\left[\frac{(\hat{S}_k + \hat{N}_k)^2}{(2\hat{S}_k + \hat{N}_k)\hat{N}_k}\right] \qquad (2) \end{aligned}$$

(All logarithms here are on base 2.) The total information about b carried by a is then simply $I_{\text{tot}} = \sum_{k=0}^{m} I_k$. Note that the summation includes only the terms with positive k, since the Fourier coefficients with negative k are determined by the constraint $S_{-k} = S_k^*$. Moreover, if the signals are shape functions as defined in section 3, we must leave out the term I_0, since the condition $a_0 = 0$ implies that coefficient A_0 can be computed from other coefficients.

Determining $\hat{S}_k$ and $\hat{N}_k$. Unfortunately, we have no direct information about the variance of the original signal $\hat{S}_k$ (the shape function of the ideal fracture line) or of the noise $\hat{N}_k$ (the difference between the fracture lines and the observed contours). However, we can estimate these parameters by comparing sections of fragment contours that are known to correspond to the same fracture line in the original object — such as the highlighted part in figure 3.

Let's then denote by $a(t)$ and $b(t)$, for $t \in [0..T]$, the shape functions of two corresponding pieces of contours, as in figure 7, selected so that the midpoints $a(T/2), b(T/2)$ of the two graphs correspond to the same point of the ideal fracture line. Let $m(t) = [a(t) + b(t)]/2$ be the average of the two signals, and $d(t) = a(t) - b(t)$ their difference. See figure 8. Then the Fourier coefficients M_k and D_k of signals m and d have variances

$$\begin{aligned} \hat{M}_k &= \text{var}\left[\frac{S_k + N'_k + S_k + N''_k}{2}\right] \\ &= \hat{S}_k + \frac{1}{2}\hat{N}_k \\ \hat{D}_k &= \text{var}\left[(S_k + N'_k) - (S_k + N''_k)\right] \\ &= 2\hat{N}_k \end{aligned}$$

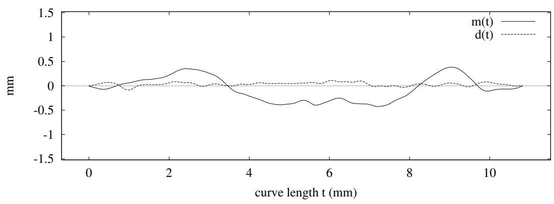


Figure 8: The average $m(t) = [a(t)+b(t)]/2$ and difference $d(t) = a(t)-b(t)$ of the shape functions of figure 7.

Thus, given a sample of matching segment pairs, we can compute the variances $\hat{M}_k$ and $\hat{D}_k$, and then estimate $\hat{S}_k, \hat{N}_k$ by the formulas

$$\hat{S}_k = \hat{M}_k - \frac{1}{4}\hat{D}_k \qquad \hat{N}_k = \frac{1}{2}\hat{D}_k \qquad (3)$$

Therefore, by formula (2), the amount of information provided by the frequency-k component of curve a about the same component of its partner b is

$$\begin{aligned} I_k &= \log\left[\frac{(\hat{A}_k)^2}{\left(2(\hat{M}_k - \frac{1}{4}\hat{D}_k) + \frac{1}{2}\hat{D}_k\right)\left(\frac{1}{2}\hat{D}_k\right)}\right] \\ &= \log\left[\frac{(\hat{A}_k)^2}{\hat{M}_k\hat{D}_k}\right] \qquad (4) \end{aligned}$$

When estimating the variances $\hat{M}_k$ and $\hat{D}_k$, we must note that they are used as arguments to the logarithm function, which is highly nonlinear in this case. Therefore, instead of computing the variances by the usual formula, it is safer to expand formula (4) into

$$I_k = 2\log\hat{A}_k - \log\hat{M}_k - \log\hat{D}_k \qquad (5)$$

then estimate the term $\log\hat{A}_k$ by averaging $\log(|A_k|^2)$ for several segments, and similarly for $\log\hat{M}_k$ and $\log\hat{D}_k$.

Consistency check. As a consistency check, let's consider what would happen if $a(t)$ and $b(t)$ were the shape functions of two unrelated contour segments with the same length. In this case, we have $a = s' + n'$ and $b = s'' + n''$, where s' and s'' are independent signals. The variances of the coefficients M_k and D_k would be

$$\begin{aligned} \hat{M}_k &= \frac{1}{2}(\hat{S}_k + \hat{N}_k) = \frac{1}{2}\hat{A}_k \\ \hat{D}_k &= 2(\hat{S}_k + \hat{N}_k) = 2\hat{A}_k \end{aligned}$$

Formula 5 would then evaluate to

$$I_k = 2\log\hat{A}_k - \log(\frac{1}{2}\hat{A}_k) - \log(2\hat{A}_k) = 0$$

as it should.

5 Experimental results

To test this theory, we shattered five unglazed ceramic tiles into about a hundred fragments, ranging from 10 to 50 mm diameter. We digitized those fragments with a 300 dpi flatbed scanner, and extracted their outlines with simple thresholding and contour-following algorithms. To remove the quantization noise, we smoothed the outlines with a geometric Gaussian filter [Gama 99], with characteristic width $\sigma = 1\,\text{pixel}\ (= 0.085\,\text{mm})$, and resampled each set with uniform stepsize $0.25\,\text{pixel}\ (= 0.022\,\text{mm})$. Some of those contours are shown in figures 1 and 3.

From these contours, we selected 40 pairs of fragments that were adjacent in the original tiles, and extracted by hand the approximately matching parts of their outlines, 128 pixels (10.8 mm) long. We converted these trimmed curve segments to shape functions $a(t)$ and $b(t)$, as explained in section 3, and computed the mean and difference signals $m(t)$ and $d(t)$ for each pair. Figures 4 and and 8 show one of these pairs.

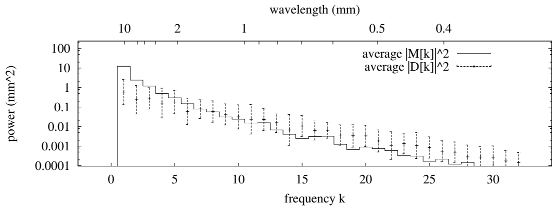


Figure 9: Average power spectra of the mean ($\hat{M}_k$) and difference ($\hat{D}_k$) signals for 40 matching pairs.

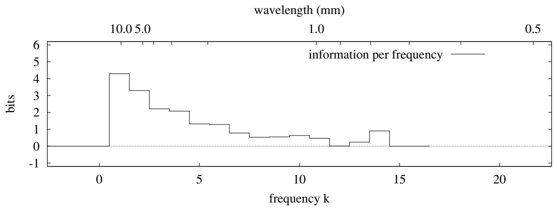


Figure 10: Useful information contents I_k per frequency k, computed from a sample of 40 corresponding segments.

Figures 9 and 10 and table 1 show the estimated variances $\hat{A}_k$, $\hat{M}_k$, and $\hat{D}_k$, and the useful information contents I_k for each component frequency k, as computed by formula (5). Note that the variance $\hat{A}_k$ was estimated by averaging the logarithms of $|A_k|^2$ for all the samples, as discussed in section 4, and then computing the anti-log. The same holds for $\hat{M}_k$ and $\hat{D}_k$.

k	$\hat{A}_k$	$\hat{M}_k$	$\hat{D}_k$	I_k	I_k/L
	mm^2	mm^2	mm^2	bits	$\frac{\text{bits}}{\text{mm}}$
1	11.93	12.29	0.590	4.29	0.396
2	2.371	2.401	0.238	3.30	0.304
3	1.262	1.196	0.287	2.21	0.204
4	0.579	0.493	0.161	2.08	0.192
5	0.366	0.298	0.181	1.32	0.121
6	0.147	0.147	0.060	1.29	0.119
7	0.105	0.079	0.080	0.78	0.072
8	0.069	0.057	0.058	0.53	0.049
9	0.044	0.031	0.042	0.55	0.051
10	0.034	0.023	0.031	0.63	0.058
11	0.022	0.015	0.023	0.47	0.043
12	0.019	0.015	0.023	0.01	0.001
13	0.011	0.006	0.015	0.24	0.022
14	0.007	0.004	0.006	0.90	0.083
15	0.004	0.002	0.010	0.00	0.000
⋮	⋮	⋮	⋮	⋮	⋮
total				18.59	1.716

Table 1: Results for a set of 40 pairs of matching contour segments: power spectra of the contour ($\hat{A}_k$), mean ($\hat{M}_k$), and difference ($\hat{D}_k$) signals, estimated information contents (I_k) and density (I_k/L), per frequency k.

Table 2 shows the information contents condensed by logarithmically-spaced frequency bands (scales of detail), and accumulated up to each scale.

k			wavelength			$I_{\rm bd}$	$I_{\rm bd}/L$
			mm			bits	$\frac{\rm bits}{\rm mm}$
1	..	1	10.84	..	5.42	4.29	0.396
2	..	3	5.42	..	2.71	5.51	0.508
4	..	7	2.71	..	1.35	5.46	0.504
8	..	15	1.35	..	0.68	3.33	0.307
16	..	256	0.68	..	0.00	0.00	0.000
total						18.59	1.715

Table 2: Information contents (I_k) and information density (I_k/L), accumulated by scale of detail (frequency band).

As a control experiment, we repeated the process with 40 pairs of non-matching contour segments. Figures 11 and 12 show the average power spectra and useful information contents I_k (rather, the lack thereof) for that sample.

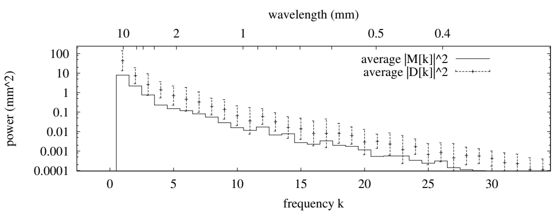


Figure 11: Power spectra of the mean and difference signals for a set of 40 non-corresponding contour segments.

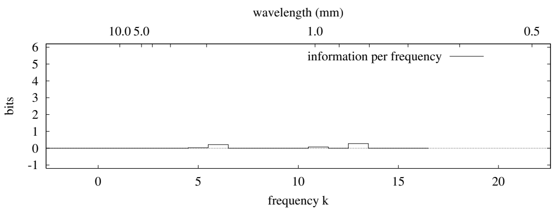


Figure 12: Useful information contents I_k per frequency k computed for a set of 40 non-corresponding contour segments.

6 Conclusions and future work

Based on these results, we conclude that, in our sample curves, the shape of an outline segment 10.8 mm long contains at least 18.6 bits of useful information about the shape of the matching segment. This information lies almost entirely in components 1..15 (wavelengths from 10.8 to 0.72 mm).

The average perimeter L of the fragments in our sample was about 2000 pixels (170 mm). Since the outlines were digitized with a sampling step of 0.25 pixel, each contour had about 2^{11} potential segments that could be matched to a given segment. Thus the 18.6 bits contained in a given 10.8 mm of contour could identify the matching pair, with fixed reliability, among $2^{18.6}/2^{11} = 2^{7.6} \approx 190$ fragments. Or, said another way: among a set of N fragments similar to our sample, we can expect that on the order of $N/190$ fragment outlines will be found to match a given piece of outline 10.8 mm long, to the accuracy of our measurements.

These numbers are of course specific to the unglazed ceramic fragmens used in our test. Unglazed ceramic is particularly appropriate for shape-based reconstruction, because of its highly irregular fracture lines. Still, we conjecture that our basic result — there is enough information in the fragment outlines to solve the problem — holds also for many other materials. For instance, glass fragments scanned at the same resolution will have relatively smooth outlines—meaning lower $\hat{S}_k$'s and hence lower information contents. On the other hand, their outlines are much sharper, and material losses are smaller, so they can be measured with greater accuracy—leading to smaller $\hat{N}_k$. This intuition ought to be checked experimentally.

Of course most practical instances of the fragment reconstruction problem involve three-dimensional objects with curved surfaces, such as vessels and statuary. For such instances, one would probably acquire the fragment outlines through stereo vision techniques, and encode them in some invariant representation (e.g. local curvature in the plane tangent to the object's surface) such that adjacent fragments will have matching outline segments. In that case one could use the techniques of this paper to measure the information contents of the encoded outlines. We believe that the result will be roughly the same as for flat fragments of the same material digitized to the same accuracy.

We expect that the main source of "noise" in real-world instances of the problem will be the erosion of fragment edges, not only from natural causes but mainly from rough handling of the fragments. (Archologists routinely use sieving to separate ceramic fragments from soil, a process which may destroy most of the edge details at sub-millimeter scale.) One could reduce the severity of that problem by tracing the outline of each fragment at a fixed depth relative to the object's surface, rather than at the surface itself. Alternatively, one could use the mean inclination of the fracture surface relative to the object's surface as an additional component of the "signal."

Taking this idea to its natural limit, one should consider fractures as surfaces rather than curves, and use surface-matching techniques (as proposed by Berequet and Sharir [Bareq96] and Levoy [Levoy99]) to find the adjacent fragments. This approach will surely supersede contour-based methods, once ways are found to reduce its formidable computational cost. In any case, it seems likely that the Fourier-based techniques of this paper can be extended to two-dimensional signals, and used to measure the information contents of fracture surfaces (as opposed to one-dimensional contours).

Acknowledgments

We would like to thank the referees for useful comments and suggestions. This work was supported in part by grants from CAPES, CNPQ (process 301016/92-5(NV) and FAPESP.

REFERENCES

[Bareq96] Gill Barequet and Micha Sharir. Partial surface matching by using directed footprints. In *Proc. 12th Annual Symp. on Computational Geometry*, pages 409–, 1996.

[Burde89] Grigore C. Burdea and Haim J. Wolfson. Solving jigsaw puzzles by a robot. *IEEE Transactions on Robotics and Automation*, 5(6):752–764, 1989.

[Gama 98] Helena C. da Gama Leitão and Jorge Stolfi. Automatic reassembly of irregular fragments. Technical report IC-98-06, Institute of Computing(IC), University of Campinas, 1998.

[Gama 99] Helena C. da Gama Leitão. *Reconstrução Automática de Objetos Fragmentados.* PhD thesis, 1999. (In Portuguese.).

[Lathi68] B. P. Lathi. *Communications systems.* John Wiley & Sons, 1968.

[Levoy99] Mark Levoy. Scanning the fragments of the Forma Urbis Romae. Electronic document available at `http: //www.graphics .stanford.edu/`, file `projects/mich /forma-urbis/forma-urbis.html`, May 1999.

[Üçolu97] Göktürk Üçoluk and I. Hakki Toroslu. Reconstruction of 3-d surface object from its pieces. In *The Ninth Canadian Conference on Computational Geometry*, volume 1, Queen's University, Kingston, Ontario, Canada, August 1997.