GEOMBINATORIC ASPECTS OF PROCESSING LARGE IMAGES AND LARGE SPATIAL DATABASES

by Jan Beck[1], Vladik Kreinovich[1,2], and Brian Penn[2]

[1]Department of Computer Science and
[2]NASA Pan-American Center for
Earth and Environmental Studies (PACES)
University of Texas at El Paso
El Paso, TX 79968, USA
emails janb@utep.edu,
bpenn@geo.utep.edu, vladik@cs.utep.edu

Abstract. *Computer processing can drastically improve the quality of an image and the reliability and accuracy of a spatial database. A large image (database) does not easily fit into the computer memory, so we process it by downloading pieces of the image. Each downloading takes a lot of time, so, to speed up the entire processing, we must use as few pieces as possible.*

Many algorithms for processing images and spatial databases consist of comparing the value at a certain spatial location with values at nearby locations. For such algorithms, we must select (possibly overlapping) sub-images in such a way that for each point, its neighborhood (of given radius) belongs to a single sub-image. We reformulate the corresponding optimization problem in geometric terms, and use this reformulation to provide some information about the solution. Namely, for images, the optimal sub-images should be bounded by straight lines or circular arcs; for non-homogeneous spatial databases, we deduce an explicit expression for the curvature of the boundaries in terms of the data density in different points.

A practical problem: brief description. Computer processing can extract useful information from an image or from a spatial database; it can also drastically improve the quality of an image and the reliability and accuracy of a spatial database; see, e.g., [Güting 1994] and [Zaniolo et al. 1997].

Many algorithms for processing images and spatial databases consist of comparing the value at a certain spatial location with values at nearby locations. Let us give two examples:

- The first example concerns *satellite photos*. When we process satellite photos of the earth's surface, it is very important to detect *edges*: in geophysical analysis, edges can describe rifts; in geographic analysis, edges can describe roads or rivers; in the agricultural analysis, edges can indicate the boundary between different crops and/or different fields, etc. Usual algorithms for checking whether a given point belongs to the edge take into consideration the image intensity at this point and at the neighboring points (see, e.g., [Penn 1991], [Penn et al. 1993], [Penn et al. 1994], and references therein).
- Another example concerns *gravity measurements*. At our university, we keep a database of gravity measurements in the region. These measurements are very useful in geophysics; see, e.g., [Birt et al. 1997], [Braile et al. 1997], [Fliedner et al. 1996], [Tesha et al. 1997]. The database was compiled from measurement done by different groups, which used equipment of different quality. Some of the measurement results may be erroneous. To get good quality data, we must therefore eliminate such erroneous records. One possible way towards eliminating such records is to take into consideration the fact that there is a physical upper bound B on the gradient of the gravity force. Therefore, if the difference between the values of gravity measured at two nearby points exceeds $B \cdot d$ (where d is the distance between them), this means that one of the measurements was erroneous.

Some images and spatial databases are so large that they do not easily fit into a computer memory. For example, a typical Landsat satellite photo consists of $\approx$ 6,000$\times$6,000 pixels, which does not fit into some computer's operating memory. Therefore, we must process it by downloading and processing sub-images. For algorithms which compare the intensity at each point with the intensity at the neighboring points, we must select these sub-images in such a way that for each point, its neighborhood (of given small radius $r > 0$) belongs to a single sub-image.

Therefore, these sub-images should overlap: indeed, for a borderline point of a sub-image (of a point whose distance to the borderline is smaller than r), its neighborhood does not belong to this sub-image, so this point must belong to a different sub-image as well.

Each downloading takes a lot of time, so, to speed up the entire processing, we must use as few pieces as possible.

This requirement can be further ramified. We not only want to fit the data into a computer memory, we also want to leave space for processing this data; the more space we leave for processing, the more sophisticated algorithms we will be able to use to process this data. So, when we select a number of pieces, the next optimization problem is to find a sub-division into this many pieces in which the largest amount of information contained in each piece is the smallest possible.

In this paper, we reformulate the problem of optimal subdivision into images in purely geombinatoric terms.

Reformulating the problem in geombinatoric terms. In accordance with our description of the sub-division problem, we are looking for a division of the original domain D into several sub-domains $D_1, \ldots, D_n$ such that for each point p from the domain D, there is at least one sub-domain D_i which contains the entire neighborhood $N_r(p) = \{q \in D \mid d(p,q) \leq r\}$ (where $d(p,q)$ is the Euclidean distance between the points p and q).

For each sub-domain D_i, we can denote, by $\widetilde{D}_i$, the set of all points $p \in D$ for which $N_r(p) \subseteq D_i$. So, all we need is to divide D into n sub-domains $\widetilde{D}_i$, and then take, as D_i, the set of all points from D which are $\leq r$-close to one of the points from D_i, i.e.,

$$D_i = \{p \mid d(p,q) \leq r \text{ for some } q \in \widetilde{D}_i\}.$$

We want to select a subdivision D_i in such a way that the largest of the amounts of information $I = \max_i I(D_i)$ corresponding to n sub-domains is the smallest possible (so that the remaining size of the memory should be the largest possible).

Let us deduce two things from here. First, the amount of information $I(D_i)$ corresponding to each sub-domain D_i must be bounded by the computer memory size I_0. If for one of the sub-domains D_i, the amount of information $I(D_i)$ is larger than for the others, then we can take some part of the corresponding set $\widetilde{D}_i$ and distribute it between sets $\widetilde{D}_j$ with smaller $I(D_j)$; as a result, the largest amount of information I would decrease. Therefore, in the optimal subdivision, we should have all domains with the exact same amount of information $I(D_1) = \ldots = I(D_n)$.

Second, if two sub-domains $\widetilde{D}_i$ and $\widetilde{D}_j$ have a common interior point, then the points from this intersection are served twice; so we can eliminate this part from one of the sets $\widetilde{D}_i$ and thus, decrease the size of at least one of the sets D_i; after that, we can reshuffle the sub-domains and decrease I. Thus, we can assume that for the optimal subdivision, different sub-domains $\widetilde{D}_i$ and $\widetilde{D}_j$ can only intersect in border points.

Let us first consider the problem of sub-dividing the satellite image. The satellite image is easier to analyze because, in contrast to, e.g., gravity databases, it is informationally homogeneous in the sense that the pixels containing information are uniformly distributed among the image, so for each sub-domain D_i of the domain D covered by the image, the amount of information about this sub-domain is simply proportional to its area $A(D_i)$. So, the requirement that $I(D_1) = \ldots = I(D_n) = I$ can be reformulated as $A(D_1) = \ldots = A(D_n) = A$, and the optimization means $A \to \min$.

If there were no neighborhoods involved (i.e., if we had $r = 0$), then we would simply take non-intersecting sub-domains D_i, so the problem would be: to divide the domain D into several sub-domains D_i of equal area. Since $r > 0$, we have to take the difference $D_i - \widetilde{D}_i$ into consideration. Namely, the area $A(D_i)$ is equal to the area $A(\widetilde{D}_i)$ plus the area of the neighborhood difference $D_i - \widetilde{D}_i$. The neighborhood size r is usually small. For small r, and for a small boundary area of length ΔL, the size of the corresponding piece of the neighborhood is equal to $\Delta L \cdot r + o(r)$; therefore, the total area of the neighborhood is equal to

$$A(D_i - \widetilde{D}_i) = L(\partial \widetilde{D}_i) \cdot r + o(r), \tag{1}$$

where $L(\partial \widetilde{D}_i)$ is the total length of the borderline $\partial \widetilde{D}_i$ separating this sub-domain $\widetilde{D}_i$ from other sub-domains $\widetilde{D}_j$ (of course, the borderline which coincides with the border of the large domain D should not be counted). From (1), we conclude that $A(\widetilde{D}_i) = A(D_i) - L(\partial\widetilde{D}_i)\cdot r + o(r) = A_0 - L(\partial\widetilde{D}_i)\cdot r + o(r)$. Since the sub-domains $\widetilde{D}_i$ provide a non-intersecting covering of the original domain D, we conclude that

$$A(D) = \sum A(\widetilde{D}_i) = \sum A(D_i) - \sum L(\partial\widetilde{D}_i)\cdot r + o(r) =$$

$$n\cdot A - \sum L(\partial\widetilde{D}_i)\cdot r + o(r).$$

Thus,

$$A = \frac{A(D)}{n} + \sum L(\partial\widetilde{D}_i)\cdot r + o(r).$$

For small r, therefore, $A \to \min$ iff the sum $\sum L(\partial\widetilde{D}_i)$ is the smallest possible. The requirement that $A(D_i) = \text{const}$ can be approximately reformulated as $A(\widetilde{D}_i) = \text{const}$. So, in geometric terms, the problem can be formulated as follows:

Problem 1. *Given a domain D and an integer n, divide D into sub-domains $\widetilde{D}_1,\ldots,\widetilde{D}_n$ of equal area $A(D)/n$ for which the total length of the subdividing lines is the smallest possible.*

Result. Let us first describe the general result:

Proposition 1. *For the optimal subdivision corresponding to Problem 1, the borderlines between sub-domains $\widetilde{D}_i$ are either straight lines or circular arcs.*

Proof. Let us take any two close points on the border between $\widetilde{D}_i$ and $\widetilde{D}_j$. If we could further decrease the length of the borderline between these points without changing the area of the domain restricted by this borderline, then we would be able to decrease the total length of the borderline curves. Thus, for an optimal subdivision, the borderline must have the smallest possible length among all the curves surrounding a domain of given area. Thus, this borderline must be a solution to the isoperimetric problem, and hence, either a straight line (if the area is 0), or a circular arc (if the area is non-zero). The proposition is proven.

Example. For a rectangular domain of size $H \times V$, it is natural to divide it into small rectangles. If we divide a horizontal size of length H into n_H equal parts, and the vertical size of length V into n_V equal parts, then we get a total of $n = n_H \times n_V$ sub-domains of size $(H/n_H) \times (V/n_V)$ with the total borderline length of $(n_H - 1) \cdot V + (n_V - 1) \cdot H$. Optimization is tough because n_H and n_V take discrete values. We can get an idea of how n_H and n_V depend on H and V if we – approximately – treat n_H and n_V as continuous variables. The Lagrange multiplier method leads to an explicit solution to the resulting continuous optimization problem $(n_H-1)\cdot V+(n_V-1)\cdot H \to \min$ under the condition that $n_H \cdot n_V = n$ (for given n): namely, we get

$$(n_H - 1) \cdot V + (n_V - 1) \cdot H + \lambda \cdot (n_H \cdot n_V - n) \to \min,$$

hence $V + \lambda \cdot n_V = 0$ and $H + \lambda \cdot n_H = 0$, i.e., $H/n_H = V/n_V$, and the domain D is sub-divided into squares.

Case of spatial databases. In spatial databases, there may be more records in some areas and less records in others. Here, for each spatial point p, we have a *density* $\rho(p)$ of measurement points around p. This means that a small domain ΔD is area ΔA around p has $\rho(p) \cdot \Delta A$ bits of information, and the total amount of information in a sub-domain D_i is now equal to the integral $I(D_i) = \int \rho(p)\,\mathrm{d}p$.

The arguments similar to the ones given in the above case lead to the conclusion that all sub-domains should have the same information, and that the total integral $\oint \rho(p)$ along the borderlines should be the smallest possible. The corresponding analogue of the isoperimetric problem can be explicitly solved by using the calculus of variations (see, e.g., [Hermann 1977], [Rassias et al. 1985]): For a piece of a borderline between the two points which is described by a curve $y(x)$ (so that $y_1 = y_2 = 0$ at the ends), we have $\int \rho(x, y(x)) \cdot \sqrt{1 + (y')^2}\,\mathrm{d}x \to \min$ under the condition that $\int_x \int_{y=0}^{y(x)} \rho(x, y)\,\mathrm{d}x\mathrm{d}y = \text{const}$, we can use the Lagrange multiplier method to reduce it to unconditional optimization problem, and then apply variational techniques to conclude that

$$\frac{d}{dx}\left(\rho \cdot \frac{y'}{\sqrt{1+(y')^2}}\right) + \lambda \cdot \rho = 0,$$

i.e., in terms of the curvature $k(s)$ of the borderline curve, that

$$k(s) = C - \frac{1}{\rho} \cdot \frac{d\rho(s)}{ds},$$

where C is a constant ($= -\lambda$), and $\rho(s)$ is a value of data density along the curve. Knowing $k(s)$, we can easily compute the shape of the corresponding borderline curve.

Open problems. We have shown that the problem of processing large images leads to a geombinatoric problem: of dividing a given domain D into a given number n of sub-domains of equal area with the smallest possible total length of borderlines. It is therefore useful to determine such optimal subdivisions for basic domains such as rectangles.

Acknowledgments. This work was supported in part by NASA under cooperative agreement NCC5-209, by NSF grant No. DUE-9750858, by United Space Alliance, grant No. NAS 9-20000 (PWO C0C67713A6), by the Future Aerospace Science and Technology Program (FAST) Center for Structural Integrity of Aerospace Systems, effort sponsored by the Air Force Office of Scientific Research, Air Force Materiel Command, USAF, under grant number F49620-95-1-0518, and by the National Security Agency under Grant No. MDA 904-98-1-0561. The authors are thankful to Ann Q. Gates and G. Randy Keller for useful discussions.

References

C. S. Birt, P. K. H. Maguire, M. A. Khan, H. Thybo, G. R. Keller, and J. Patel, "The influence of pre-existing structures on the evolution of the southern Kenya Rift Valley: Evidence from seismic and gravity studies", *Tectonophysics*, 1997, Vol. 278, pp. 211–242.

L. W. Braile, W. J. Hinze, and G. R. Keller, "New Madrid seismicity, gravity anomalies, and interpreted ancient rift structures", *Seismology Research Letters*, 1997, Vol. 67, pp. 599–610.

M. M. Fliedner, S. D. Ruppert, P. E. Malin, S. K. Park, G. R. Jiracek, R. A. Phinney, J. B. Saleeby, B. P. Wernicke, R. W. Clayton, G. R. Keller, K. C. Miller, C. H. Jones, J. H. Luetgert, W. D. Mooney, H. L. Oliver, S. L. Klemperer, and G. A. Thompson, "Three-dimensional crustal structure of the southern Sierra Nevada from seismic fan profiles and gravity modeling", *Geology*, 1996, Vol. 24, pp. 367–370.

R. H. Güting, "An introduction to spatial database systems", *VLDB Journal*, 1994, Vol. 3, pp. 357–399.

R. Hermann, *Differential geometry and the calculus of variations*, Math. Sci. Press, Brookline, MA, 1977.

B. S. Penn, "Identifying Edges and Linear Features in Remotely Sensed data using Neural Networks", *Proceedings of Geotech/Chatauqua*, Lakewood, Colorado, September 21–24, 1991, pp. 181-189.

B. S. Penn, A. Gordon, and R. F. Wendlandt, "Finding Edges in Satellite Images," *Proceedings of the 9th Thematic Conference – Geologic Remote Sensing*, February 8–11, 1993, Pasadena, CA, USA.

B. S. Penn, A. Gordon, and R. F. Wendlandt, "Using Neural Networks to Locate Edges and Linear Features in Satellite Images," *Computers & Geosciences*, 1994, Vol. 19, No. 10, pp. 1545–1565.

G. M. Rassias and T. M. Rassias (eds.), *Differential geometry, calculus of variations, and their applications*, M. Dekker, New York, 1985.

A. L. Tesha, A. A. Nyblade, G. R. Keller, and D. I. Doser, "Rift localization in suture-thickened crust: Evidence from Bouguer gravity anomalies in northeastern Tanzania, East Africa", *Tectonophysics*, 1997, Vol. 278, pp. 315–328.

C. Zaniolo, S. Ceri, C. Faloutsos, R. T. Snodgrass, V. S. Subrahmanian, and R. Zicari, *Advanced Database Systems*, Morgan Kaufmann, Menlo Park, CA, 1997.