Optimal Lower Bound for Generalized Median Problems in Metric Space

Xiaoyi Jiang[1] and Horst Bunke[2]

[1] Department of Electrical Engineering and Computer Science
Technical University of Berlin
Franklinstrasse 28/29, D-10587 Berlin, Germany
jiang@cs.tu-berlin.de
[2] Department of Computer Science, University of Bern
Neubrückstrasse 10, CH-3012 Bern, Switzerland
bunke@iam.unibe.ch

Abstract. The computation of generalized median patterns is typically an NP-complete task. Therefore, research efforts are focused on approximate approaches. One essential aspect in this context is the assessment of the quality of the computed approximate solutions. In this paper we present a lower bound in terms of a linear program for this purpose. It is applicable to any pattern space. The only assumption we make is that the distance function used for the definition of generalized median is a metric. We will prove the optimality of the lower bound, i.e. it will be shown that no better one exists when considering all possible instances of generalized median problems. An experimental verification in the domain of strings and graphs shows the tightness, and thus the usefulness, of the proposed lower bound.

1 Introduction

The concept of average, or mean, is useful in various contexts. In sensor fusion, multisensory measurements of some quantity are averaged to produce the best estimate. Averaging the results of several classifiers is used in multiple classifier systems in order to achieve more reliable classifications. In clustering and machine learning, a typical task is to represent a set of (similar) objects by means of a single prototype. Interesting applications of the average concept have been demonstrated in dealing with shapes [6], binary feature maps [10], 3D rotation [3], geometric features (points, lines, or 3D frames) [15], brain models [4], anatomical structures [17], and facial images [13].

In structural pattern recognition symbolic structures, such as strings, trees, or graphs, are used for pattern representation. One powerful tool in dealing with these data structures is provided by the generalized median. Given a set S of input patterns, the generalized median is a pattern that has the smallest sum of distances to all patterns in S (see Section 2 for a formal definition).

T. Caelli et al. (Eds.): SSPR&SPR 2002, LNCS 2396, pp. 143–151, 2002.
© Springer-Verlag Berlin Heidelberg 2002

The computation of generalized median of symbolic structures is typically an NP-complete task. Therefore, research efforts are focused on approximate algorithms. One essential aspect in this context is the assessment of the quality of the computed approximate solutions. Since the true optimum is unknown, the quality assessment is not trivial in general. In this paper we present an optimal lower bound in terms of a linear program for this purpose. It is applicable to any pattern space. The only assumption we make is that the distance function used for the definition of generalized median is a metric.

The outline of the paper is as follows. In Section 2 we first introduce the generalized median of patterns. Then, we present the LP-based lower bound and discuss its optimality in Sections 3 and 4. The results of an experimental verification in the domains of strings and graphs are reported in Section 5 to show the usefulness of the lower bound. And finally, some discussion conclude the paper.

2 Generalized Median of Patterns

Assume that we are given a set S of patterns in an arbitrary representation space U and a distance function $d(p,q)$ to measure the dissimilarity between any two patterns $p,q \in U$. An important technique for capturing the essential information of the given set of patterns is to find a pattern $\overline{p} \in U$ that minimizes the sum of distances to all patterns from S, i.e.

$$\overline{p} \;=\; \arg\min_{p\in U} \sum_{q\in S} d(p,q).$$

Pattern $\overline{p}$ is called a *generalized median* of S. If the search is constrained to the given set S, the resultant pattern

$$\hat{p} \;=\; \arg\min_{p\in S} \sum_{q\in S} d(p,q)$$

is called a *set median* of S. Note that neither the generalized median nor the set median is necessarily unique.

Independent of the underlying representation space we can always find the set median of N patterns by means of $N(N-1)/2$ distance computations. The computational burden can be reduced if the distance function is a metric [9]. For non-metric distance functions an approximate set median search algorithm has been reported recently [12]. Note that the generalized median is the more general concept and therefore usually a better representation of the given patterns than the set median.

If U is the universe of real numbers and the distance function $d(p,q)$ is the absolute (squared) difference of p and q, then the generalized median simply corresponds to the scalar median (mean) known from statistics. Scalar median represents a powerful technique for image smoothing. Its extension to vector spaces [1,2] provides a valuable image processing tool for multispectral/color images and optical flow.

In dealing with strings, the popular Levenshtein edit distance is usually used. Under this distance function the set median string problem is solvable in polynomial time. However, the computation of generalized median strings turns out to be NP-complete [5,16]. Several approximate approaches have been reported in the literature; see [8] for a discussion. In [7] the concept of generalized median graphs is defined based on graph edit distance. Also here we are faced with an NP-complete computation problem.

An approximate computation method gives us a solution $\tilde{p}$ such that

$$\mathrm{SOD}(\tilde{p}) \;=\; \sum_{q\in S} d(\tilde{p},q) \;\geq\; \sum_{q\in S} d(\overline{p},q) \;=\; \mathrm{SOD}(\overline{p})$$

where SOD stands for sum of distances and $\overline{p}$ represents the (unknown) true generalized median. The quality of $\tilde{p}$ can be measured by the difference $\mathrm{SOD}(\tilde{p}) - \mathrm{SOD}(\overline{p})$. Since $\overline{p}$ and $\mathrm{SOD}(\overline{p})$ are unknown in general, we resort to a lower bound $\varGamma \leq \mathrm{SOD}(\overline{p})$ and measure the quality of $\tilde{p}$ by $\mathrm{SOD}(\tilde{p}) - \varGamma$. Note that the relationship

$$0 \;\leq\; \varGamma \;\leq\; SOD(\overline{p}) \;\leq\; SOD(\tilde{p})$$

holds. Obviously, $\varGamma = 0$ is a trivial, and also useless, lower bound. We thus require $\varGamma$ to be as close to $\mathrm{SOD}(\overline{p})$ as possible. In the next two sections we present such a lower bound and prove its optimality (in a sense to be defined later). The tightness of the proposed lower bound will be experimentally verified in Section 5 in the domain of strings and graphs.

It is worth pointing out that a lower bound is not necessarily needed to compare the relative performance of different approximate methods. But it is very useful to indicate the closeness of approximate solutions to the true optimum. Such an absolute performance comparison is actually the ultimate goal of performance evaluation.

3 LP-Based Lower Bound

We assume that the distance function $d(p,q)$ be a metric. Let the set S of input patterns be $\{q_1, q_2, \ldots, q_n\}$. The generalized median $\overline{p}$ is characterized by:

$$\begin{aligned}
&\text{minimize } \mathrm{SOD}(\overline{p}) \;=\; d(\overline{p},q_1) + d(\overline{p},q_2) + \cdots + d(\overline{p},q_n) \text{ subject to} \\
&\forall i,j \in \{1,2,\ldots,n\}, i \neq j, \;\begin{cases} d(\overline{p},q_i) + d(\overline{p},q_j) \;\geq\; d(q_i,q_j) \\ d(\overline{p},q_i) + d(q_i,q_j) \;\geq\; d(\overline{p},q_j) \\ d(\overline{p},q_j) + d(q_i,q_j) \;\geq\; d(\overline{p},q_i) \end{cases} \\
&\forall i \in \{1,2,\ldots,n\},\; d(\overline{p},q_i) \geq 0
\end{aligned}$$

Note that the constraints except the last set of inequalities are derived from the triangular inequality of the metric $d(p,q)$. By defining n variables $x_i,\ i = 1,2,\ldots,n$, we replace $d(\overline{p},q_i)$ by x_i and obtain the linear program LP:

minimize $x_1 + x_2 + \cdots + x_n$ subject to

$$\forall i,j \in \{1,2,\ldots,n\}, i \neq j, \ \begin{cases} x_i + x_j \ \geq \ d(q_i,q_j) \\ x_i + d(q_i,q_j) \ \geq \ x_j \\ x_j + d(q_i,q_j) \ \geq \ x_i \end{cases}$$

$$\forall i \in \{1,2,\ldots,n\}, \ x_i \geq 0$$

If we denote the solution of LP by $\mathit{\Gamma}$, then we have:

Theorem 1. *The true generalized median $\overline{p}$ satisfies $\mathit{\Gamma} \leq SOD(\overline{p})$. That is, $\mathit{\Gamma}$ is a lower bound for $SOD(\overline{p})$.*

Proof: In the initial characterization the quantities $d(\overline{p}, q_i)$ are dependent of each other. The linear program LP results from replacing $d(\overline{p}, q_i)$ by x_i and is defined in contrast by n totally independent variables x_i. Consequently, LP poses less conditions than the initial characterization and its solution $\mathit{\Gamma}$ thus must be smaller than or equal to SOD($\overline{p}$). QED

The linear program LP has $\frac{3n^2-n}{2}$ inequality constraints and we may apply the popular simplex algorithm [14] to find out the solution. Note that, despite its exponential worst-case computational complexity, the simplex algorithm turns out to be very efficient in practice and is used to routinely solve large-scale linear programming problems.

4 Optimality Issue

For a fixed n value, any set S of n patterns specifies $N = \frac{n(n-1)}{2}$ distances $d(p,q), \ p,q \in S$, and can be considered as a point in the N-dimensional real space $\Re^N$. Due to the triangular inequality required by a metric, all possible sets of n patterns only occupy a subspace $\Re^N_*$ of $\Re^N$. Abstractly, any lower bound is therefore a function $f: \ \Re^N_* \rightarrow \Re$. The lower bound $\mathit{\Gamma}$ derived in the last section is such a function.

Does a lower bound exist that is tighter than $\mathit{\Gamma}$? This optimality question is interesting from both a theoretical and a practical point of view. The answer and the implied optimality of the LP-based lower bound $\mathit{\Gamma}$ is given by the following result.

Theorem 2. *There exists no lower bound that is tighter than $\mathit{\Gamma}$.*

Proof: Given a point $b \in \Re^N_*$, we denote the solution of the corresponding linear program LP by $(x_1, x_2, \ldots, x_n)$. We construct a problem instance of $n+1$ abstract patterns $q_1, q_2, \ldots, q_n, q_{n+1}$. The $\frac{n(n-1)}{2}$ distances $d(q_i, q_j), \ 1 \leq i,j \leq n$, are taken from the coordinates of b. The remaining distances are defined by $d(q_{n+1}, q_i) = x_i, 1 \leq i \leq n$. The distance function d is clearly a metric. Now we compute the generalized median $\overline{p}$ of $\{q_1, q_2, \ldots, q_n\}$. Since $\mathit{\Gamma} = x_1 + x_2 + \cdots + x_n$ is a lower bound, we have SOD($\overline{p}$) $\geq \mathit{\Gamma}$. On the other hand, the pattern q_{n+1} satisfies:

$$\begin{aligned} \text{SOD}(q_{n+1}) &= d(q_{n+1}, q_1) + d(q_{n+1}, q_2) + \cdots + d(q_{n+1}, q_n) \\ &= x_1 + x_2 + \cdots + x_n \\ &= \mathit{\Gamma} \end{aligned}$$

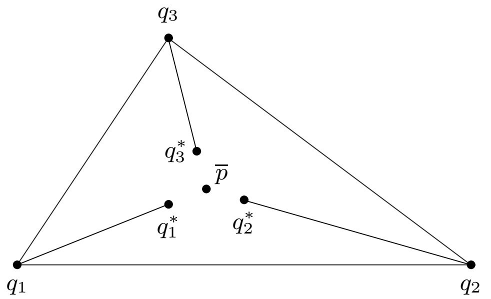

Fig. 1. The lower bound Γ cannot be reached by the generalized median $\overline{p}$

Consequently, q_{n+1} is a generalized median of $\{q_1, q_2, \ldots, q_n\}$. This means that, for each point in $\Re_*^N$, we can always construct a problem instance where the lower bound Γ is actually reached by the generalized median. Accordingly, no lower bound can exist that is more tight than the LP-based lower bound Γ. QED

At this point two remarks are in order. For most problems in practice it is likely that the lower bound Γ cannot be reached by the generalized median. The first reason is a fundamental one and is illustrated in Figure 1, where we consider points in the plane. The distance function is defined to be the Euclidean distance of two points. Let $\overline{p}$ be the true generalized median of q_1, q_2, and q_3. Then, $x_i = |q_i\overline{p}|, \ i = 1, 2, 3$, satisfy the constraints of the linear program LP. Now we select a point q_i^* on the line segment $q_i\overline{p}$ such that $|q_i^*\overline{p}| = \epsilon$ (an infinitely small number). Due to the small amount of ϵ, $x_i^* = |q_i q_i^*|$ satisfy the constraints of LP as well. But in this case we have $x_1^* + x_2^* + x_3^* < x_1 + x_2 + x_3 = \mathrm{SOD}(\overline{p})$. As a consequence, the solution of LP, i.e. the lower bound Γ, is constrained by:

$$\Gamma \ \leq \ x_1^* + x_2^* + x_3^* \ < \ \mathrm{SOD}(\overline{p})$$

and therefore not reached by the generalized median $\overline{p}$. Fundamentally, this example illustrates the decoupled nature of the quantities x_i in LP in contrast to $d(\overline{p}, q_i)$ in the original problem of generalized median computation. By doing this, however, the solution x_i of LP may not be physically realizable through a single pattern $\overline{p}$.

The special property of a concrete problem may also imply that the lower bound Γ is not reached by the generalized median. We consider again points in the plane, but now with integer coordinates only. The distance function remains the Euclidean distance. An example is shown in Figure 2 with four points q_1, q_2, q_4, and q_4. The lower bound Γ turns out to be $2\sqrt{34}$ corresponding to $x_1 = x_2 = x_3 = x_4 = \frac{\sqrt{34}}{2}$. This lower bound is satisfied by $p(\frac{5}{2}, \frac{3}{2})$, which is unfortunately not in the particular space under consideration. Any point with integer coordinates will result in a SOD value larger than Γ.

It is important to point out that Theorem 2 only implies that we cannot specify a better lower bound than the solution of LP, when considering *all* possible instances of generalized median problems. An improved lower bound may

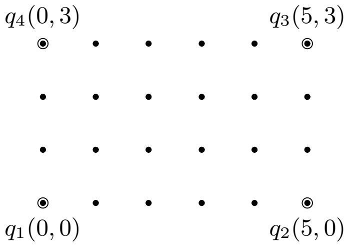

Fig. 2. The point $p(\frac{5}{2}, \frac{3}{2})$ reaching the lower bound is not in the problem space

still be computed for a *particular* problem instance. For the problem in Figure 2, for example, the constraint $x_1 + x_3 \geq d(q_1, q_3) = \sqrt{34}$ can be replaced by $x_1 + x_3 \geq \sqrt{34} + \Delta$ for some $\Delta > 0$. The reason is that no point with integer coordinates lies on the line segment q_1q_3 and the corresponding constraint can thus be made tighter. The constraint $x_2 + x_4 \geq d(q_2, q_4) = \sqrt{34}$ can be modified in a similar manner. As a final result, the modified constraints may lead to a tighter lower bound.

5 Experimental Verification

A lower bound is only useful if it is close to SOD($\overline{p}$) where $\overline{p}$ represents the (unknown) true generalized median pattern. In this section we report the results of an experimental verification in the domain of strings and graphs to show the tightness, and thus the usefulness, of the proposed lower bound. We used the MATLAB package to solve the linear program LP.

5.1 Median Strings

The median concept can be used in OCR to combine multiple classification results for achieving a more reliable final classification [11]. In doing so we may obtain multiple classification results either by applying different classifiers to a single scan of a source text or by applying a single classifier to multiple scans of the text.

To verify the usefulness of the LP-based lower bound in this context we conducted a simulation by artificially distorting the following text which consists of 448 symbols (including spaces):

> There are reports that many executives make their decisions by flipping a coin or by throwing darts, etc. It is also rumored that some college professors prepare their grades on such a basis. Sometimes it is important to make a completely ’unbiased’ decision; this ability is occasionally useful in computer algorithms, for example in situations where a fixed decision made each time would cause the algorithm to run more slowly. Donald E. Knuth

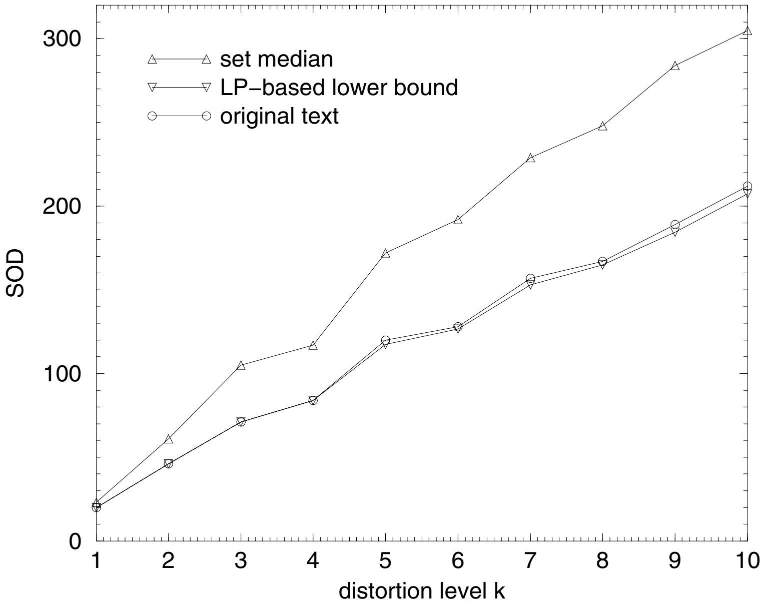

Fig. 3. Verification of lower bound for strings

Totally, ten distortion levels are used, producing $k\%$ ($k = 1, 2 \cdots, 10$) letters in the text to be changed. For each k, five distorted samples of the text are generated. We use the Levensthein edit distance and set the insertion, deletion, and substitution cost each to be one.

Figure 3 summarizes the results of this test series. As a comparison basis, SOD of the original text is also given. The SOD of the (unknown) true generalized median string $\overline{p}$ must be between this curve and the lower bound curve. Clearly, the LP-based lower bound is a very good estimate of SOD($\overline{p}$). In addition the results confirm that the generalized median string is a more precise abstraction of a given set of strings than the set median. It has a significantly smaller SOD value, which corresponds to the representation error.

5.2 Median Graphs

The concept of generalized median graphs was introduced in [7]. We study the LP-based lower bound in this domain by means of random graphs generated by distorting a given initial graph. The initial graph g_0 contains k nodes and $2k$ edges. The node and edge labels are taken from $\{A, B, C, D, E\}$ and $\{F\}$, respectively. Both the graph structure and the labeling of g_0 are generated randomly. The distortion process first randomly changes the labels of 50% of the nodes in g_0. Then, up to two nodes are inserted or deleted in g_0. In case of an insertion the new node is randomly connected to one of the nodes in g_0. If a node in g_0 is deleted, all its incident edges are deleted as well. This way a col-

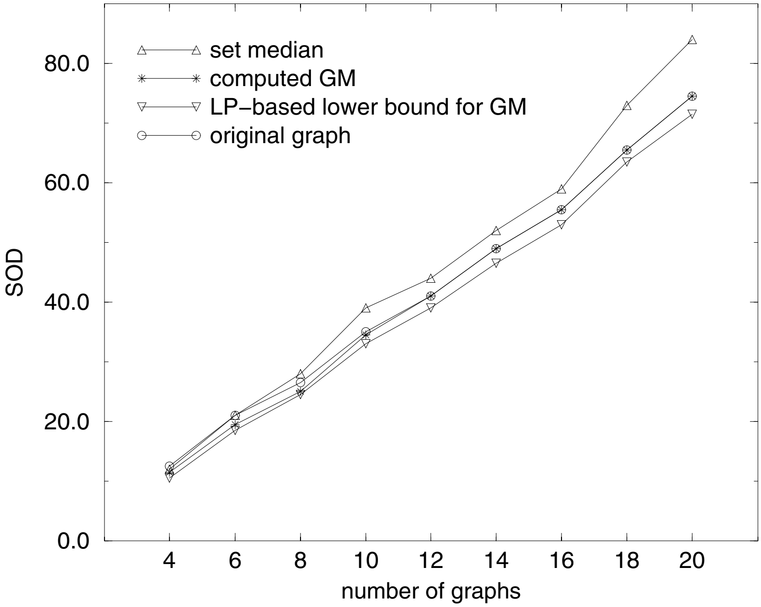

Fig. 4. Verification of lower bound for graphs

lection of 20 distorted graphs are generated for g_0 associated with a particular k value. Based on this procedure, we conducted a series of experiments by using $n \in \{4, 6, \ldots, 20\}$ out of the 20 graphs to test the lower bound. The distance function of two graphs is defined in terms of graph edit operations; see [7] for details.

The results of this test series for $k = 6$ are summarized in Figure 4. As an upper bound for SOD($\overline{g}$) of the (unknown) true generalized median graph $\overline{g}$, we give the SOD of the original graph g_0 and an approximate solution found by the method from [7]. Clearly, SOD($\overline{g}$) must be between the minimum of these two curves and the lower bound curve. Also here the LP-based lower bound demonstrates a high predication accuracy.

6 Conclusions

The computation of generalized median patterns is typically an NP-complete task. Therefore, research efforts are focused on approximate approaches. One essential aspect in this context is the assessment of the quality of the computed approximate solutions. In this paper we have presented an optimal lower bound in terms of a linear program for this purpose. It is applicable to any metric pattern space. An experimental verification in the domain of strings and graphs has shown the tightness, and thus the usefulness, of the proposed lower bound.

Acknowledgments

The authors would like to thank J. Crisik for valuable discussions on the topic of this paper.

References

1. J. Astola, P. Haavisto, and Y. Neuvo, Vector median filters, Proceedings of the IEEE, 78(4): 678–689, 1990. 144
2. F. Bartolini, V. Cappellini, C. Colombo, and A. Mecocci, Enhancement of local optical flow techniques, Proc. of 4th Int. Workshop on Time Varying Image Processing and Moving Object Recognition, Florence, Italy, 1993. 144
3. C. Gramkow, On averaging rotations, Int. Journal on Computer Vision, 42(1/2): 7–16, 2001. 143
4. A. Guimond, J. Meunier, and J.-P. Thirion, Average brain models: A convergence study, Computer Vision and Image Understanding, 77(2): 192–210, 2000. 143
5. C. de la Higuera and F. Casacuberta, Topology of strings: Median string is NP-complete, Theoretical Computer Science, 230(1-2): 39–48, 2000. 145
6. X. Jiang, L. Schiffmann, and H. Bunke, Computation of median shapes, Proc. of 4th. Asian Conf. on Computer Vision, 300–305, Taipei, 2000. 143
7. X. Jiang, A. Münger, and H. Bunke, On median graphs: Properties, algorithms, and applications, IEEE Trans. on PAMI, 23(10): 1144–1151, 2001. 145, 149, 150
8. X. Jiang, H. Bunke, and J. Csirik, Median strings: A review, 2002. (submitted for publication) 145
9. A. Juan and E. Vidal, Fast median search in metric spaces, in A. Amin and D. Dori (eds.), *Advances in Pattern Recognition*, Springer-Verlag, 905–912, 1998. 144
10. T. Lewis, R. Owens, and A. Baddeley, Averaging feature maps, Pattern Recognition, 32(9): 1615–1630, 1999. 143
11. D. Lopresti and J. Zhou, Using consensus sequence voting to correct OCR errors, Computer Vision and Image Understanding, 67(1): 39-47, 1997. 148
12. L. Mico and J. Oncina, An approximate median search algorithm in non-metric spaces, Pattern Recognition Letters, 22(10): 1145–1151, 2001. 144
13. A. J. O'Toole, T. Price, T. Vetter, J. C. Barlett, and V. Blanz, 3D shape and 2D surface textures of human faces: The role of "averages" in attractiveness and age, Image and Vision Computing, 18(1): 9–19, 1999. 143
14. C. H. Papadimitriou and K. Steiglitz, *Combinatorial Optimization: Algorithms and Complexity*, Prentice-Hall, Inc., 1982. 146
15. X. Pennec and N. Ayache, Uniform distribution, distance and expectation problems for geometric features processing, Journal of Mathematical Imaging and Vision, 9(1): 49–67, 1998. 143
16. J. S. Sim and K. Park, The consensus string problem for a metric is NP-complete, Journal of Discrete Algorithms, 2(1), 2001. 145
17. K. Subramanyan and D. Dean, A procedure to average 3D anatomical structures, Medical Image Analysis, 4(4): 317–334, 2000. 143