On Lower Bounds for the Dimensions of Dot-Matrix Characters to Represent Farsi and Arabic Scripts

Behrooz Parhami
Department of Electrical and Computer Engineering
University of California
Santa Barbara, CA 93106-9560, USA

Abstract: *This paper reports the results of a preliminary study to establish the minimal dot matrix size for producing legible Farsi and Arabic scripts. For the generation of legible Farsi and Arabic scripts without much regard to aesthetic quality, a 9×9/2 matrix (9 rows by 9 half-pitch columns) is shown to be minimal. In the area of high-quality output for electronic publishing and the like, where aesthetics is important, results of a Farsi character recognition study are used to establish a 16-row variable-width dot matrix as minimal.*

1. Introduction

Until recently, the computer industry was relatively insensitive to the status of writing as an important art form in Farsi and Arabic (F&A) speaking countries. For many years, the F&A scripts resulting from computer printers and displays resembled the so-called *architectural script* derived from *Kufi* by artists who wanted to embed holy and historical writings on brick or ceramic facades of mosques and palaces hundreds of years ago (Fig. 1). Such a script might have been acceptable for decorative purposes, but it is quite inappropriate for efficient transfer of information. The problems were in part due to properties of F&A scripts, distinguishing them from the scripts of technologically advanced countries for whom the design of most computer display units and printers were optimized. However, the scarcity of systematic studies of F&A scripts in relation to the capabilities and limitations of computer printer and display technologies also played a role.

Dot-matrix character formation, which has long been the preferred method of displaying textual data, dominates today's printer technologies as well. Industrialized countries, aided by a four-decade history of computer applications and a lucky coincidence (in that, with the exception of Japan, their Latin-based scripts can be encoded in relatively small dot matrices), have found acceptable solutions to the font design problem for such printers and modern high-resolution displays. F&A speaking countries have been less fortunate in that they face the font design problem for high quality dot-matrix displays and printers without the benefit of extensive experimentation with older technologies. The need for a systematic study of dot-matrix character representation in these languages is thus apparent.

This paper reports the results of a preliminary study in this direction, leading to two main findings on the lower bound for the dot matrix size: A lower bound pertaining to the generation of legible scripts, without much regard to aesthetic quality, for use in cost-critical systems with a need for F&A user interfaces (e.g., portable monitoring and remote test equipment using low-resolution displays) and another bound for high-quality output for electronic publishing and the like, where both legibility and aesthetic quality are important. We will refer to F&A symbols by 2-digit hex codes (Fig. 2). Dot matrices considered either have the same horizontal and vertical resolutions or use half-pitch columns (Fig. 3).

The rest of this paper is organized as follows. Section 2 deals with the design of several small dot matrices, starting with the absolutely minimal 7×5 (7 rows, 5 columns) dot matrix. Legibility problems are categorized with respect to unique features of F&A scripts in order to allow an incremental derivation of the required minimal matrix. It is then shown that a 7×9/2 matrix still leads to difficulties and that to rectify the observed legibility problems, at least a 9×9/2 or 9×*w* (9 rows, variable width) matrix is needed. Section 3 deals with issues in the design of dot matrices for high-quality output. Here, arguing that legibility is related to ease of automatic recognition, results and digitized text samples from a character recognition study are used to establish a 16-row variable-width dot matrix as minimal.

2. Minimal Acceptable Dot-Matrix Size

Although one encounters some difficulty in implementing the full ISO or ASCII character set on a 7×5 dot matrix, such a matrix has been found adequate in some informatics applications. Unfortunately, a 7×5 dot matrix is totally inadequate for designing legible F&A symbols. Nevertheless, an attempt at such a design is instructive for three reasons:

a. It clearly shows where the most serious legibility and aesthetic problems exist and how the matrix should be expanded to alleviate or minimize these problems.

b. It provides a basis for judging the improvements, in legibility and aesthetic quality, gained from expanding the matrix in horizontal and/or vertical directions.

c. It yields an absolutely minimal, though somewhat cryptic, font for use in applications with limited output variety, or where speed/cost is of paramount importance.

The complete font design is omitted here, but a sample text using the font appears in Fig. 4. Careful inspection of Fig. 4, and of the font table, reveals the following legibility problems (the 2-digit hex codes are column-row pairs in the code table of Fig. 2):

1. Letters with three dots: C8, CA, CD, CE, D7, D9.
2. Letters with complex shapes or fine detail: C3, C6, DA, E1, ED, EE, F8.
3. Tall letters: C3, C4, E2, E3, E4, E5, ED, EE, EF, F0, F1.

As for aesthetic quality, the figures point to the following shortcomings resulting from a combination of limited matrix size and legibility requirements:

4. Elimination of curvatures and slants: B5, B7, B8, CD, ED, EE, F7.
5. Elimination of serrates: B3, D8, D9, DA, E1, FB.
6. Exaggeration of fine features or detail: A4, BB, D3, D4, EF, F8.
7. Disproportionate representation: B4, AC, BB, C5, DB, DC, DD, DE, ED, EE, F8.

To these, one must add the two-width approximation of symbols which assume a multitude of widths, ranging almost continuously from 1 to 6 units, in high-quality printed text.

Fig. 5 shows that some of these difficulties can be overcome by using a 7×9/2 dot matrix. In particular, Problems 1, 4, 5, and to some extent 6, are effectively dealt with. Some designers have eliminated Problem 3 by using a lowered (off-center) connection line between adjacent characters. However, this creates serious difficulties elsewhere, as only two rows are left below the connection line for representing a great deal of detail.

Hence, the question of what dot matrix size is an acceptable minimum reduces to "how many additional rows and/or columns will enable us to deal with Problems 2, 3, and 7?". The answer appears to be 2 additional rows, provided that an off-center connection line is used. Thus we arrive at a 9×9/2 matrix as the absolute minimum for legible and aesthetically not very unpleasant symbols. Fig. 6 shows the resulting script.

The two additional rows have been used principally for a more natural representation of tall letters and better positioning of upper dots as compared to the 7×9/2 design. Notice how this relatively minor change enhances both legibility and aesthetic quality to a great extent (Fig. 7). Note also that proportional, as opposed to two-width, spacing of symbols improves both legibility and aesthetic quality. While we are nowhere close to the aesthetic quality of the samples in Fig. 1, we seem to have achieved the goal of legibility. Where half-pitch row dots cannot be used, a 9×7 dot matrix provides comparable quality.

3. Larger Matrices and Aesthetic Quality

Many existing fonts provide high quality results, at least with larger point sizes (e.g. Fig. 8). Representative samples from printers/displays (Fig. 9) show considerable improvement over earlier systems and technologies, some of which were included in my 1984 book on computer appreciation. However, published explanations of how dot matrix sizes were chosen and how the inevitable quality/efficiency tradeoffs were dealt with are still lacking.

Availability of inexpensive printers, with resolutions approaching or exceeding 100 dots per centimeter, can easily lure designers into using matrices that are too large. Unfortunately, display terminals are still a few steps behind printers in terms of resolution. Also, dot matrices that are larger than needed have other implications in storage and transmission of bit patterns representing the fonts that must be taken into account. Hence, it is natural to inquire about the smallest dot matrix size that can lead to acceptable quality in run of the mill applications. No doubt, artists and publishers will remain interested in higher resolutions. My point is to obtain output quality that is sufficient for ordinary correspondence, electronic mail, office memos, technical reports, and the like.

Whereas in the case of small matrices the limited design space and existing fonts could be used to find near-optimal designs by inspection, it is extremely costly to design and test alternative fonts in large (say 16×16 or larger) matrices. It is therefore imperative to develop some analytic or experimental design guidelines as aids in reducing the search space. Fortunately, data gathered in connection with an automatic Farsi character recognition study provide some useful clues as to how to proceed. Fig. 10 shows ten fragments from digitized printed text samples that were used in the study. The samples include a representative variety of characters and different typefaces, taken from newspaper headlines, all of which were successfully recognized by the automatic recognition system.

One can easily argue that legibility of a script is related to the ease of automatic recognition. The fact all of the samples used in the aforementioned study came from newspaper headlines is significant here, as these fonts have evolved over the years to near optimal designs with respect to legibility and aesthetics. The samples in Fig. 10 have the following parameters, where the numbers of upper/lower rows are relative to the connecting line:

Sample (Fig. 10)	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)	Rel. Avg.
Pen width (pw)	12	8	4	4	6	6	7	5	5	8	1.0 pw
Upper rows	22	21	12	17	24	24	20	19	15	21	3.2 pw
Lower rows	16	15	8	12	17	17	16	14	10	16	2.3 pw
Total matrix height	50	44	24	33	47	47	43	38	30	45	6.5 pw

One conclusion of the character recognition study was that to have near-perfect automatic recognition of printed Farsi texts, the digitization hardware must be able to take 4 samples across the *pen width* (which is the width of the widest strokes in the font; e.g., the width of the line connecting the symbol C9 to its tail C1). Incidentally, one reason we used newspaper headlines was that our digitization hardware consisted of a low-resolution fax machine, providing 7×12 samples in a square centimeter. Attempts to recognize texts with 3 samples across the pen width (e.g., secondary headlines) led to some problems in determining boundaries between letters, figuring out the number of dots, and distinguishing between similar symbols. With 2 samples, the problems became insurmountable.

Unlike our automatic system, human readers can use contextual information to remove any ambiguity and to correctly recognize even illegible text. However, in many computer-oriented applications, there is very little or no context to rely on. For example in displaying database records, items such as names, birth dates, and various ID & account codes must be independently readable. Furthermore, as any teacher who has graded homeworks and exams can testify, excessive reliance on context to decipher written information is highly strenuous.

Thus, I believe that contextual information must be regarded as a safety margin to be used for increasing the speed and accuracy of reading, rather than as a primary recognition aid. These observations, along with the fact the our pattern recognition system used features such as curvatures, dots, holes, concavities, and mass distribution/density, which are pretty much what human readers use for character recognition, lead me to believe that effortless reading by humans and preservation of symbol shapes (aesthetic quality) cannot be achieved with fewer than 3 (or perhaps even 4) dots across the pen width.

The samples shown in Fig. 10 exhibit a ratio ranging from 4 to 8 between the matrix height and pen width. Taking the smallest value of 4 for this ratio along with a pen width of 4 dots, leads to a 16-row dot matrix, perhaps with 7 upper and 5 lower rows (the upper-to-lower ratio ranges from 1.25 to 1.5 for our samples). The resulting symbols would look somewhat heavy and also quite dense when multiple dots or other details exist. A pen width of 3 dots leads to somewhat slimmer strokes and allows us to expand the number of lower rows to 6. Either way, a dot-matrix height of 16 appears to be the minimum acceptable.

4. Conclusion

I started thinking about the problem of minimal dot-matrix representations for F&A symbols in late 1985. My work in this area has remained in a preliminary stage in view of the shift of focus in my research program after leaving Iran. Since I have not seen this problem addressed in the intervening years, I felt that even these preliminary results might be worth reporting in order to point out the need for detailed studies in this area.

On the relatively old 21-inch monitor (870×1152 dots) that I am using to compose this paper, I can see 60 lines of text with no problem or strain. According to the above observations, the same display can adequately show no more than about 40 F/A text lines with similar quality. Whereas desktop monitors have already advanced to the point that an entire page of F/A text, including very small typefaces, can be easily displayed, there is a vast segment of the computing field (laptops and personal portable systems) where lower resolutions and smaller screen sizes will be the norm in the foreseeable future. Such applications can benefit from judicious trimming of the dot matrix used to form the symbols.

Acknowledgment

Contributions of Mr. Anoosh Hosseini in providing the font designs and samples of Farsi output for several software and hardware systems are gratefully acknowledged.

References

[BECK87] Becker, J.D., "Arabic Word Processing", *Commun. ACM*, Vol. 30, No. 7, pp. 600-610, 1987.

[DILL88] Dillon, A., C. McKnight, and J. Richardson, "Reading from Paper versus Reading from Screen", *The Computer J.*, Vol. 31, No. 5, pp. 457-464, 1988.

[GRIF88] Griffee, A.W. and C.A. Casey, "An Introduction to Typographic Fonts and Digital Font Resources", *IBM Systems J.*, Vol. 27, No. 2, pp. 206-218, 1988.

[HOSS93] Hosseini, A., "Development of a Persian Graphical User Interface", *Proc. 1st Annual Conf. Technical Advances in Developing Countries*, Columbia Univ., July 1993.

[MORR89] Morris, R.A., "Rendering Digital Type: A Historical and Economic View of Technology", *The Computer J.*, Vol. 32, No. 6, pp. 524-532, Dec. 1989.

[PARH77] Parhami, B. and F. Mavaddat, "Computers and the Farsi Language: A Survey of Problem Areas", *Information Processing 77* (Proc. IFIP Congress), North-Holland, 1977, pp. 673-676.

[PARH78] Parhami, B., "On the Use of Farsi and Arabic Languages in Computer-Based Information Systems", *Proc. Symp. Linguistic Implic. Computer-Based Info. Syst.*, New Delhi, Nov. 1978.

[PARH81] Parhami, B., "Language-Dependent Considerations for Computer Applications in Farsi & Arabic Speaking Countries", *System Approach for Development*, North-Holland, 1981, pp. 507-513.

[PARH81a] Parhami, B. and M. Taraghi, "Automatic Recognition of Printed Farsi Texts", *Pattern Recognition*, Vol. 14, Nos. 1-6, pp. 395-403, 1081.

[PARH84] Parhami, B., *Computer Appreciation* (in Farsi *Aashna'ee baa Computer*), Tehran, 1984 (p. 191).

[PARH84a] Parhami, B., "Standard Farsi Information Interchange Code and Keyboard Layout: A Unified Proposal", *J. Inst. Electrical & Telecommun. Engineers*, Vol. 30, No. 6, pp. 179-183, 1984.

[SANA87] Sanati, M., M. Dadashzadeh, and M.B. Dadfar, "Iranian Standard Code for Information Interchange (ISCII)", *Computer Standards & Interfaces*, Vol. 6, pp. 427-432, 1987.

[TAYL90] Tayli, M. and A.I. Al-Salamah, "Building Bilingual Microcomputer Systems", *Commun. ACM*, Vol. 33, No. 5, pp. 495-504, May 1990.

(a)

(b)

طرح های بلندو کوتاه مدت وزارت آموزش وپرورش
برای تامین معلم در مناطق دور افتاده کشور اعلام شد

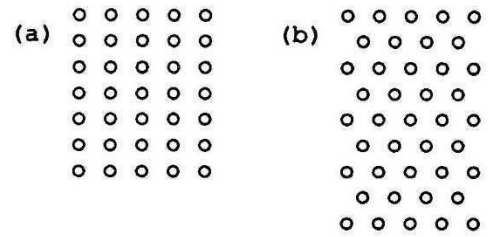
(c)

(d)

Fig. 1. Rather than mimicking the beautiful *Nasta'liq* (a) or the elegant *Naskh* in its traditional or modern printed form (b, c), the outputs of some dot-matrix computer printers and displays have resembled the architectural adaptation of *Kufi* (d), a script that facilitated decorative writing with bricks and ceramic tiles centuries ago.

	A	B	C	D	E	F		A	B	C	D	E	F
0	≡	٠	×	ـ	ـ	ل	8	(	٨	ب	ـ	غ	ه
1	!	١	ـ	غ	ف	ل	9	)	٩	ت	ث	ف	ی
2	"	٢	ں	خ	ط	م	A	*	:	ث	ص	ف	ب
3	#	٣	آ	د	ظ	ـ	B	+	؛	ج	ّ	ق	ّ
4	℟	۴	ا	ذ	ظ	ن	C	،	<	ـ	ـ	ق	|
5	٪	۵	ء	ر	ظ	ـ	D	-	=	ج	ُ	ک	٠
6	÷	۶	ث	ز	ع	و	E	.	>	ـ	ٌ	گ	—
7	'	٧	ب	ژ	غ	ه	F	/	؟	ح	ـ	لا	DEL

Fig. 2. A Farsi extension of the ISO international standard 8-bit code: Hexadecimal column and row numbers correspond to the high-order and low-order 4-bits of the 8-bit code, respectively (so, A3 denotes the symbol "#"); Only the 6 columns A through F, containing Farsi symbols, are shown here.



(a) (b)

Fig. 3. Dot matrix definitions: (a) 7×5 dot matrix, (b) 9×9/2 dot matrix using 9 half-pitch columns.

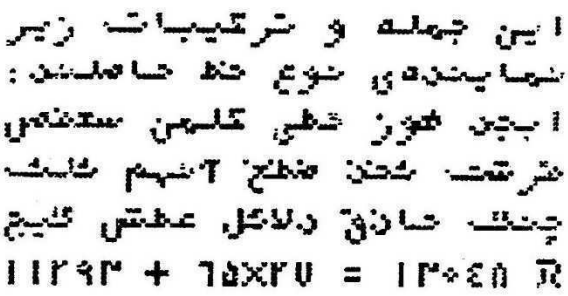


Fig. 4. Legibility of 7×5 dot-matrix Farsi and Arabic symbols is inadequate, even when contextual and semantic information are taken into account.

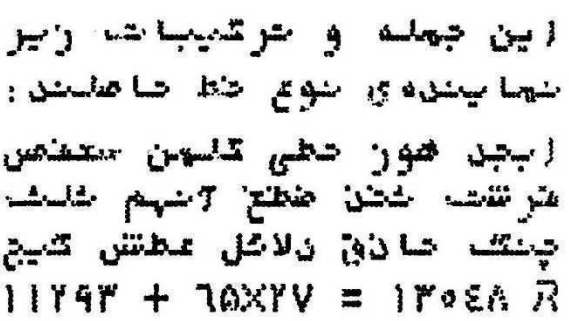


Fig. 5. Legibility of 7×9/2 dot-matrix Farsi and Arabic symbols, using half-pitch row dots, shows considerable improvement over Fig. 4.

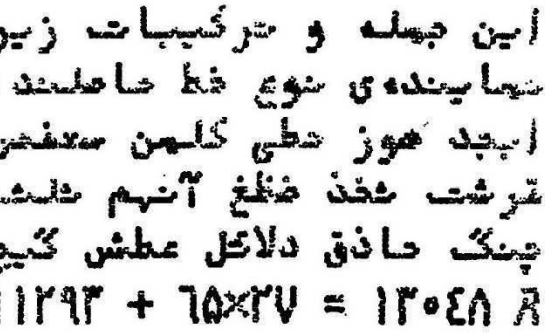


Fig. 6. Legibility of 9×9/2 dot-matrix Farsi and Arabic symbols further improves (see also Fig. 7).

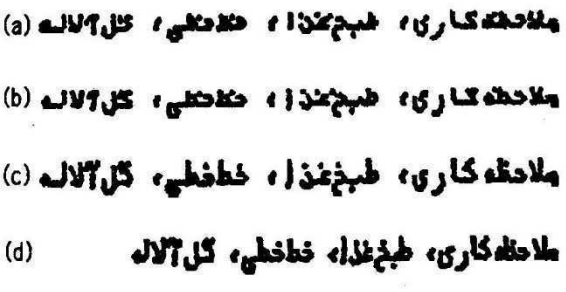


Fig. 7. Sample text line featuring *tall* letters with (a) 7×5, (b) 7×9/2, (c) 9×9/2, and (d) 9×*w* (proportionally spaced) dot-matrix symbols.

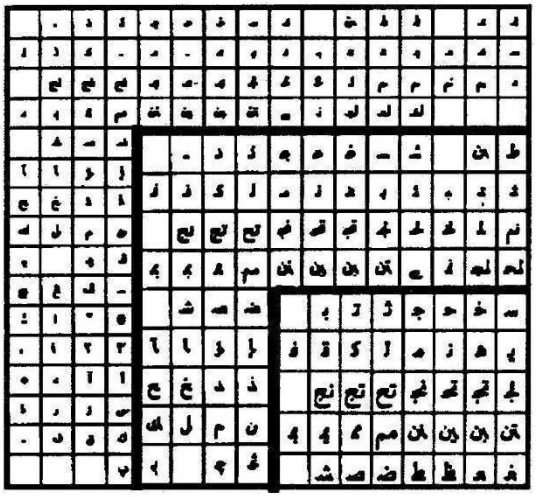


Fig. 8. The Tehran font with different point sizes: 12-point on the upper left, 18-point in the center, and 36-point on the lower right.

Fig. 9. Samples of printer and display outputs exhibiting the improved quality of Farsi output.

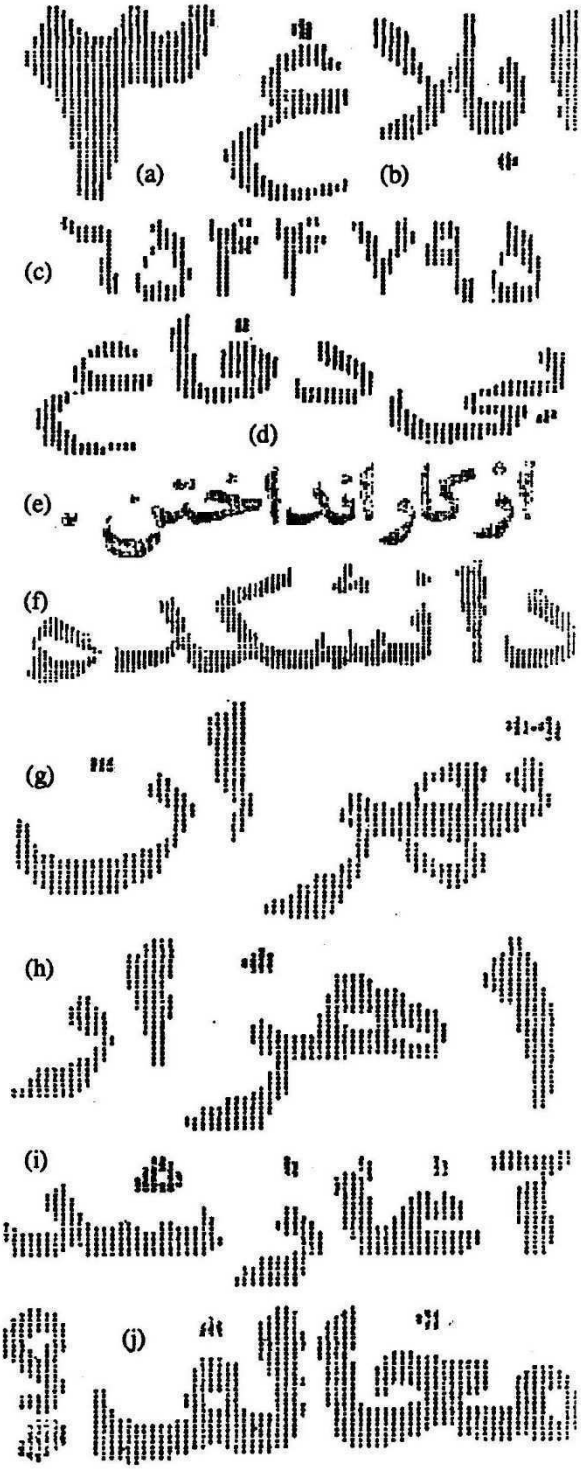


Fig. 10. Fragments of 10 digitized newspaper headlines, from a 1981 Farsi character recognition study, exhibiting various font styles and sizes.