Summer Internship

Centre for Development of Advanced Computing

CDAC, Pune

A presentation by

Kimya Gandhi

08625005

IDC, IIT Bombay

Objective of project

- To study font design in Indian scripts
- To study Indian scripts and their diversity
- Get acquainted with softwares and standards
- Learn the process of font design; the research and methodology required.

Introduction to CDAC & GIST

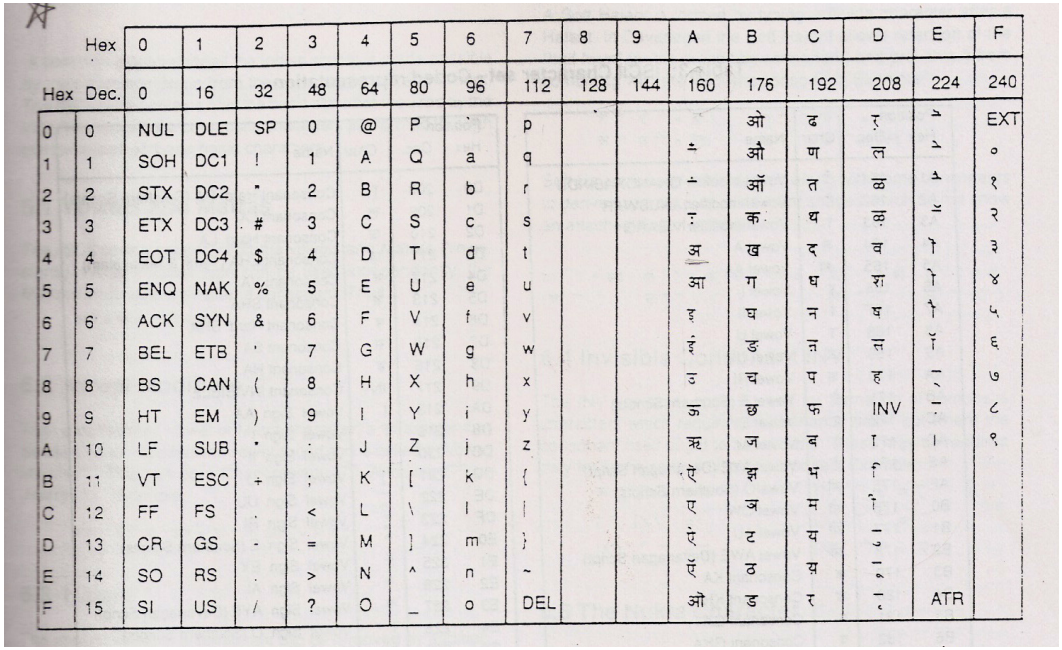
- Centre for Development of Advanced Computing (CDAC) has been a pioneer in developing and proliferating use of Indian languages on the computer.
- Their Graphic and Intelligence based Script Technology (GIST) has been a a bridge between Indian languages and IT
- They are in tune with technologies worldwide
- It is well known for its research and digitization of Indian languages and making it available to the masses.
- The division brags a collection of fonts in the various scripts of India, Devanagari, Gurumukhi, Bengali, Tamil, Telegu, Malyalama, Gujrati, Sinhalese to name a few.

Project Outline

- Project: Font Design
- Script: Devanagari
- Application: Optical Character Recognition (OCR) systems
- Understanding OCR technology and its application
- Study of devanagari script and nature of its characters
- To create a font that would be specifically give accurate results for character recognition.

Digitization and standards

- Need for standards: because one font is used over diverse applications and softwares in different places.
- The need for digitization of Indian scripts: To increase their reach and prevent them from being obsolete
- The use of one script for writing different languages
- UNICODE, ASCII, ISCII
- Opentype font formats



Hex	Dec.	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
		0	16	32	48	64	80	96	112	128	144	160	176	192	208	224	240
0	0	NUL	DLE	SP	0	@	P	`	p			ं	ओ	ढ	ृ	ॅ	EXT
1	1	SOH	DC1	!	1	A	Q	a	q			ः	औ	ण	ॄ	ॆ	०
2	2	STX	DC2	"	2	B	R	b	r			ऄ	ऑ	त	ॅ	े	१
3	3	ETX	DC3	#	3	C	S	c	s			अ	क	थ	ॆ	ै	२
4	4	EOT	DC4	$	4	D	T	d	t			आ	ख	द	े	ॉ	३
5	5	ENQ	NAK	%	5	E	U	e	u			इ	ग	ध	ै	ॊ	४
6	6	ACK	SYN	&	6	F	V	f	v			ई	घ	न	ॉ	ो	५
7	7	BEL	ETB	'	7	G	W	g	w			उ	ङ	ऩ	ॊ	ौ	६
8	8	BS	CAN	(	8	H	X	h	x			ऊ	च	प	ो	्	७
9	9	HT	EM	)	9	I	Y	i	y			ऋ	छ	फ	ौ	॑	८
A	10	LF	SUB	*	:	J	Z	j	z			ऌ	ज	ब	्	॒	९
B	11	VT	ESC	+	;	K	[	k	{			ऍ	झ	भ	॑		
C	12	FF	FS	,	<	L	\	l	|		ऎ	ञ	म	॒		
D	13	CR	GS	-	=	M	]	m	}			ए	ट	य			
E	14	SO	RS	.	>	N	^	n	~			ऐ	ठ	र			
F	15	SI	US	/	?	O	_	o	DEL			ऑ	ड	ऱ		ATR	

8- bit Code Table of the Latin and Indian Script Alphabet

Optical Character Recognition (OCR)

- Understanding Optical Character Recognition (OCR): it is the mechanical or electronic translation of images of handwritten, typewritten or printed text into machine-editable text.

- OCR for Latin script: a largely solved problem

- Applications of OCR: credit cards, postal services, data entry, preserving and digitizing old documents or legal documents

- Reasons for using OCR
  - Reduce data entry errors
  - Consolidate data entry
  - Handle peak loads
  - Human readable
  - Many printing techniques
  - Scanning corrections

Cheque number printed in OCR readable font

Devanagari script

- An overview of the script: written from left to right, no distict cases, *shirorekha* or the top line is a distict feature, based on phonetic structure
- Devanagari in manuscripts: a brief study of the precedents of the present day used devanagari.
- Analysing different writing systems
- Nature of the script and alphabet: consonants, conjuncts, vowels, vowel signs, diacritic marks, punctuations.
- Anatomy of the devanagari letterforms

Consonant Signs with Following ă

Velars	क	ख	ग	घ	ङ
	ka	kha	ga	gha	ṅa
Palatals	च	छ	ज	झ	ञ
	ca	cha	ja	jha	ña
Linguals	ट	ठ	ड	ढ	ण
	ṭa	ṭha	ḍa	ḍha	ṇa
Dentals	त	थ	द	ध	न
	ta	tha	da	dha	na
Labials	प	फ	ब	भ	म
	pa	pha	ba	bha	ma
Semivowels	य	र	ल	व	
	ya	ra	la	va	
Sibilants	श	ष	स		
	śa	ṣa	sa		
Aspirate	ह				
	ha				

Devanagari font design

- Font design in devanagari is a complex process: no complete set.
- The non linear and complex nature of the script
- The difference in physical form of the same letter used in different languages
- Inappropriate placement of vowel signs (*matras*), consonant clusters form complex conjucts and ligatures.

मुद्राक्षरलेखनकला

Non linear nature of script

अल्पित नूपुर

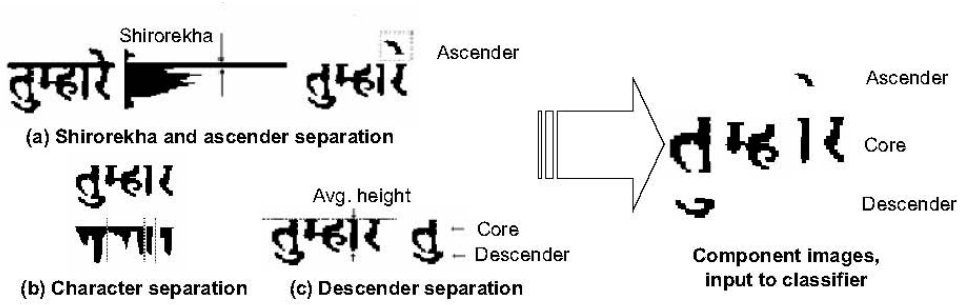
Inappropriate placement and spacing of matras

OCR and Devanagari script

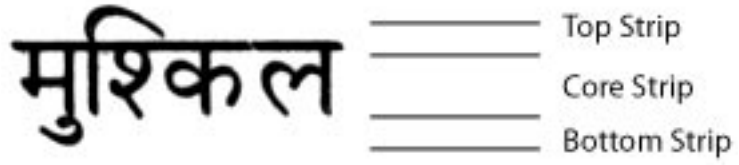
- Chitrankan: OCR software for Indian languages

- Process of character recognition: feature extraction, segmentation

- Problems with devanagari character recognition
  - recognition difficult in smaller point sizes
  - ambiguity in characters with similar structures
  - non linear nature of script
  - shirorekha breaks after a word as also in characters like अ, श, झ
  - ligatures and complex conjuncts



Segmentation driven OCR



Segmentation for recognition; Chitrankan

Character design

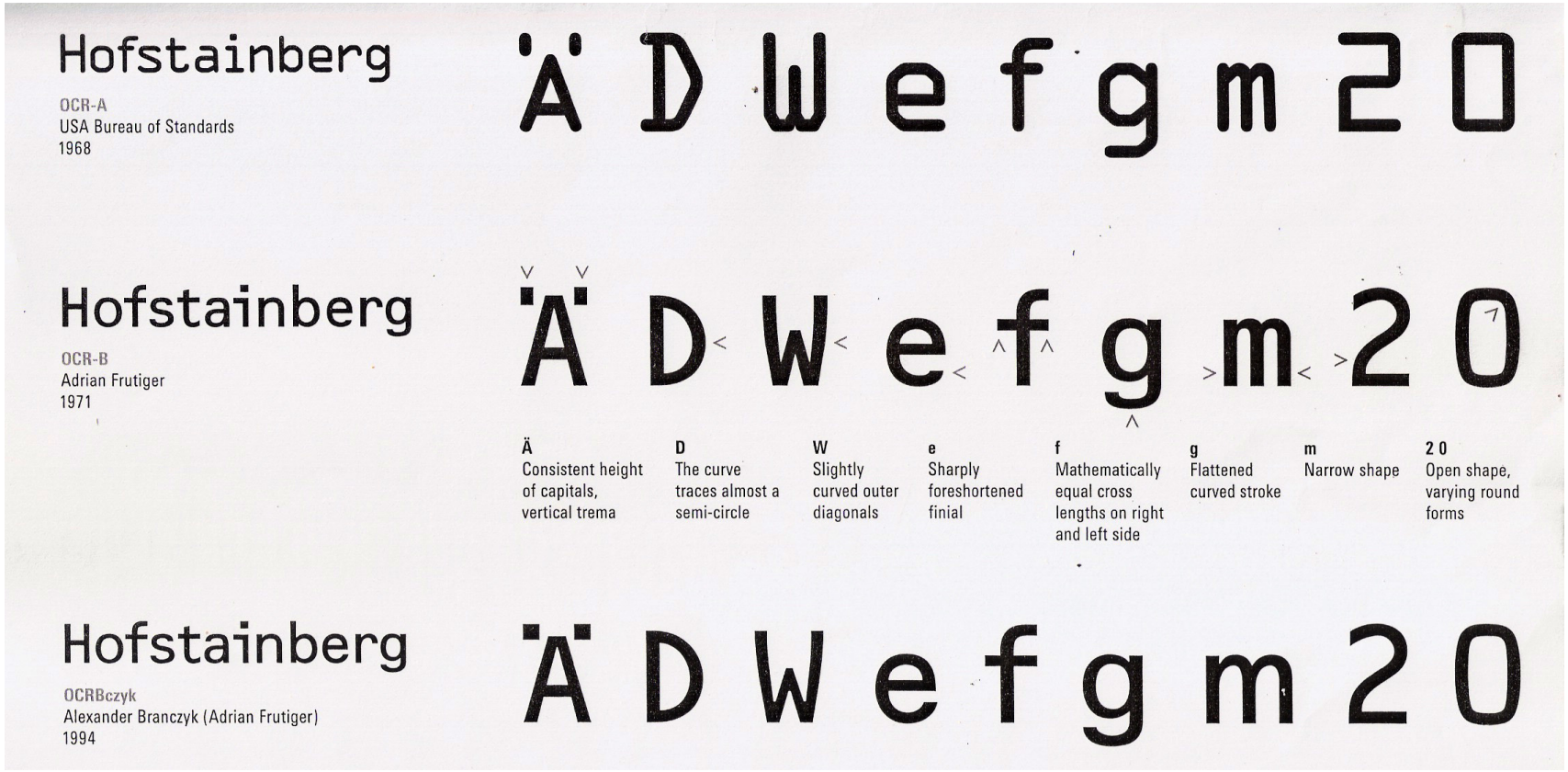
- The study of problems faced by recognition of devanagari script led to setting parameters for font design
    - characters that could be scanned with accuracy and speed
    - simple and open counters
    - readable in small point sizes
    - machine and human readable

- The stages involved in the design of letterforms included following stages
    - case studies; OCR specific fonts
    - visual references; understanding proportions
    - study of existing fonts
    - explorations for construction of letterforms
    - character set design

Case studies

- OCR-A, OCR-B

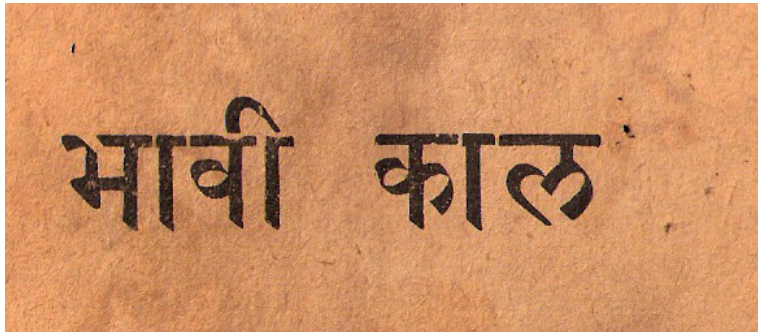
- Understanding methodology and logic that went behind designing the font

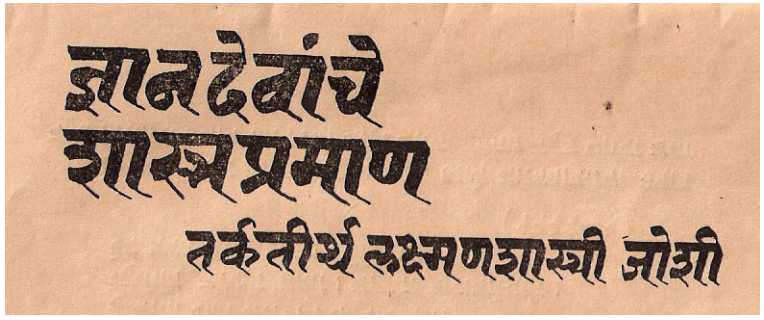


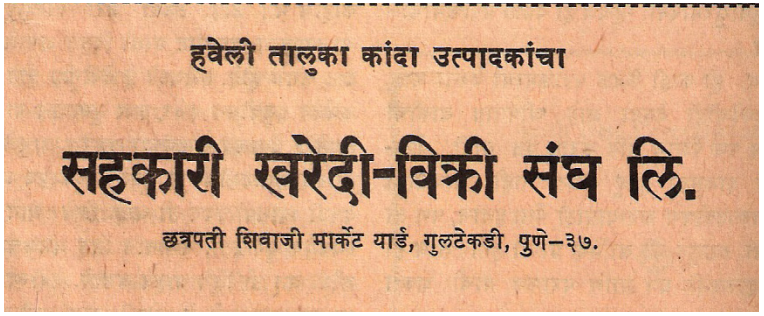
Above: Comparison between OCR-A and OCR-B

Visual references: Letterpress

Letterpress printed ads and book covers







Visual references: Writing systems

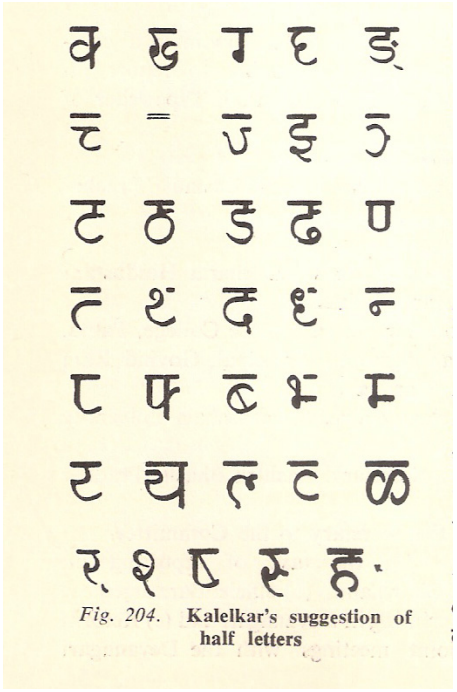


Fig. 204. **Kalelkar's suggestion of half letters**

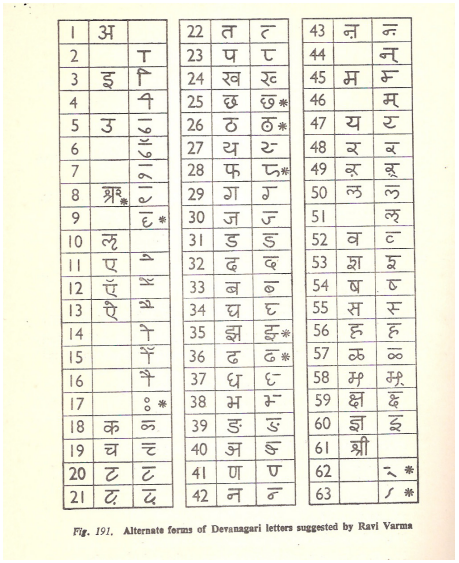


Fig. 191. **Alternate forms of Devanagari letters suggested by Ravi Varma**

Fig. 176. **The Sagariya Lipi : Calligraphy.**

Fig. 177. **The Sagariya Lipi : Type script. A—63 types**

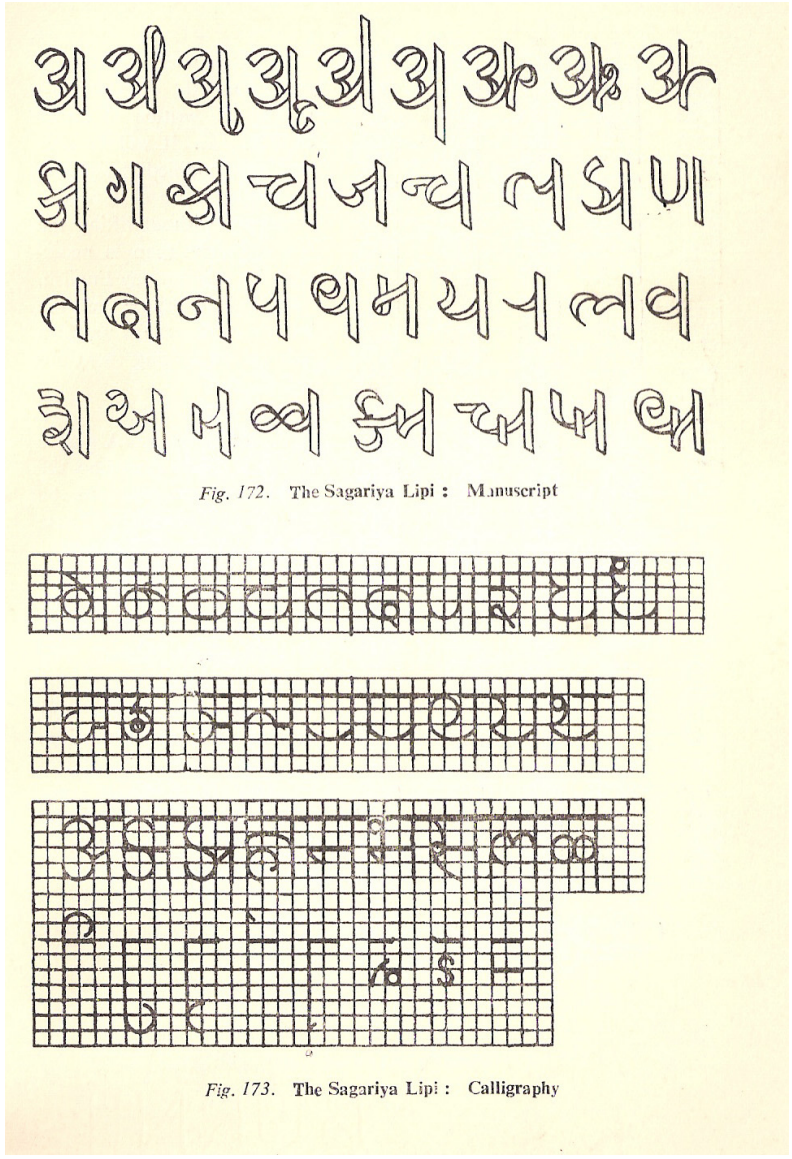


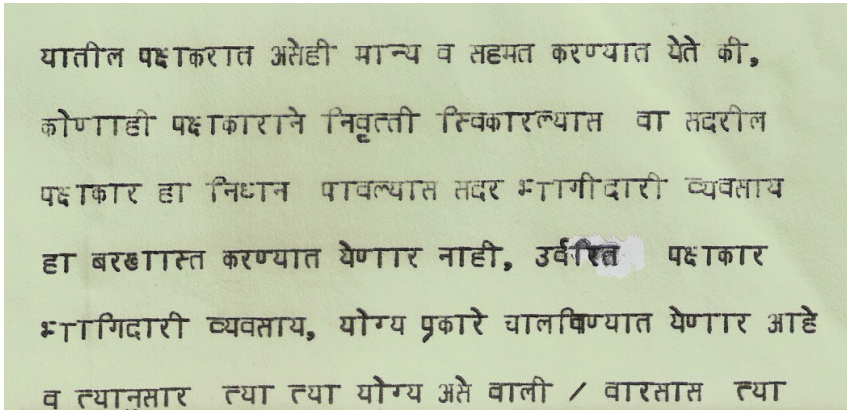
Fig. 172. **The Sagariya Lipi : Manuscript**

Fig. 173. **The Sagariya Lipi : Calligraphy**

Linear writing systems

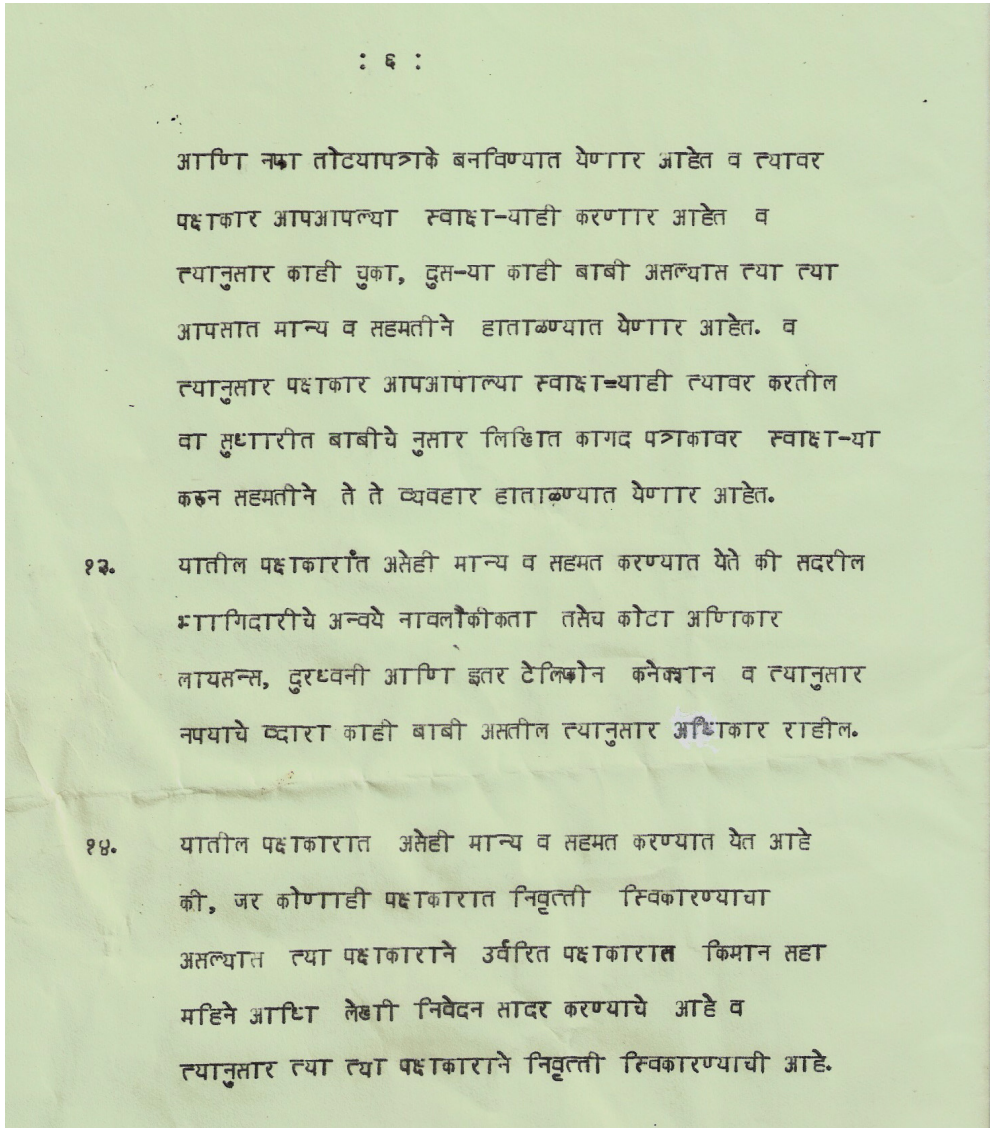
Other writing systems and proportions.

Visual references: Typewritten text



यातील पक्षकारात असेही मान्य व सहमत करण्यात येते की,
कोणाही पक्षकाराने निवृत्ती स्विकारल्यास वा सदरील
पक्षकार हा निधन पावल्यास सदर भागीदारी व्यवसाय
हा बरखास्त करण्यात येणार नाही, उर्वरित पक्षकार
भागिदारी व्यवसाय, योग्य प्रकारे चालविण्यात येणार आहे
व त्यानुसार त्या त्या योग्य असे वाली / वारसास त्या

Above; Devanagari text typewriiten page section & right; the entire page



: ६ :

आणि नवा तोटयापत्राके बनविण्यात येणार आहेत व त्यावर
पक्षकार आपआपल्या स्वाक्षा-याही करणार आहेत व
त्यानुसार काही चुका, दुरुस्या काही बाबी असल्यास त्या त्या
आपसात मान्य व सहमतीने हाताळण्यात येणार आहेत. व
त्यानुसार पक्षकार आपआपाल्या स्वाक्षा=याही त्यावर करतील
वा सुधारीत बाबीचे नुसार लिखित कागद पत्राकावर स्वाक्षा-या
करुन सहमतीने ते ते व्यवहार हाताळण्यात येणार आहेत.

१३. यातील पक्षकारांत असेही मान्य व सहमत करण्यात येते की सदरील
भागिदारीचे अन्वये नावलौकीकता तसेच कोटा अधिकार
लायसन्स, दुरध्वनी आणि इतर टेलिफोन कनेक्शन व त्यानुसार
नपयाचे व्दारा काही बाबी असतील त्यानुसार अधिकार राहील.

१४. यातील पक्षकारात असेही मान्य व सहमत करण्यात येत आहे
की, जर कोणाही पक्षकारात निवृत्ती स्विकारण्याचा
असल्यास त्या पक्षकाराने उर्वरित पक्षकारास किमान सहा
महिने आधि लेखी निवेदन सादर करण्याचे आहे व
त्यानुसार त्या त्या पक्षकाराने निवृत्ती स्विकारण्याची आहे.

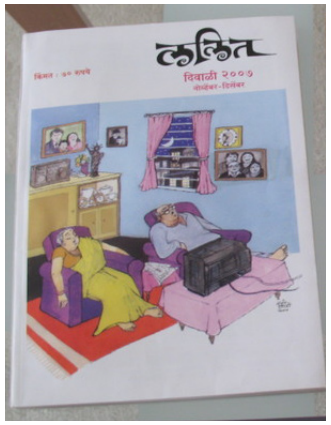
Observations:

Quality of printed page; no grey mass

Spreading and blotting of ink in complex characters or ligatures

Matras and their placement, spacing

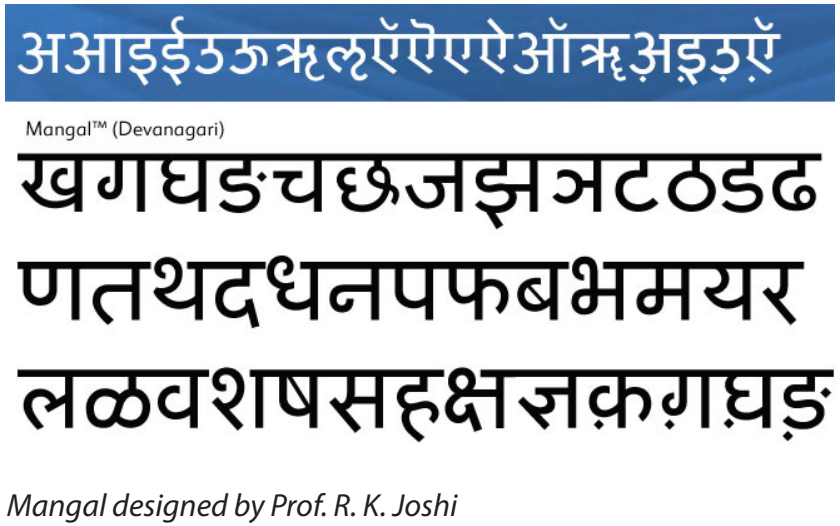
Visual references: Book covers, learning aids



Visual references: Existing fonts

यह सीडॅक की योगेश मुद्रलिपि है।
निम्नलिखित
बीजाङ्कुरित
कॉङ्क्रर्र
काकचेष्टा बकोध्यानम्
ष्ठ्यौः

DV-OT Yogesh designed by GIST team, CDAC



Mangal designed by Prof. R. K. Joshi

Character design

- The process of designing characters included:
  - character listing; *varnamala*
  - categorizing characters according to various parameters
  - openness, number of strokes, joineries with shirorekha
  - geometric construction of form

- The study of OCR, devanagari and visual references led to setting parameters for design of the font which are:
  - geometric construction
  - uniform weight and width of characters
  - one pixel gap between characters
  - simple joineries
  - open counters

Character segregation

number of strokes

अ आ इ ई उ ऊ ए ऐ ओ औ अं अः

ण ध ट थ त्र घ छ ष क्ष ख भ र त य प स द फ ग ह ज क ल झ श

च व ब न म ड त्त श्र ज्ञ रु रू ढ ऋ

one stroke

र ट द ठ ढ ड इ उ

two stroke

ण ग ह ज भ म य प त न व ल च थ छ ए ऊ ई ळ श ज्ञ

three stroke

अ स क ख फ झ त्र ष क्ष ब ऋ

Character segregation

points of contact with *shirorekha*

one point

र ट द ठ ढ ड इ उ ह ज त न व ल ई ऊ ळ च ब क त्र ज्ञ ऋ

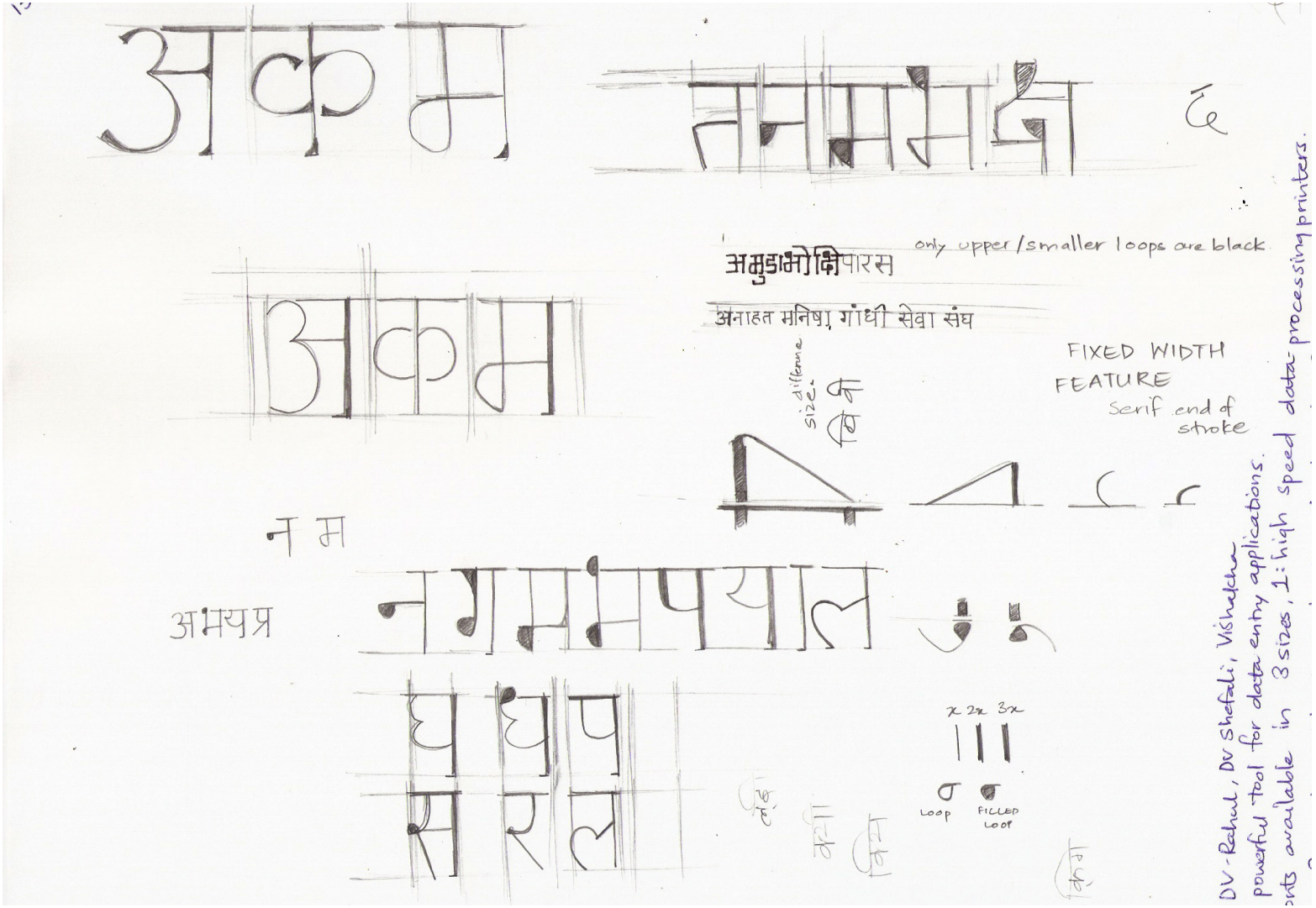
two point

ग म य प छ ए स ख फ झ ष

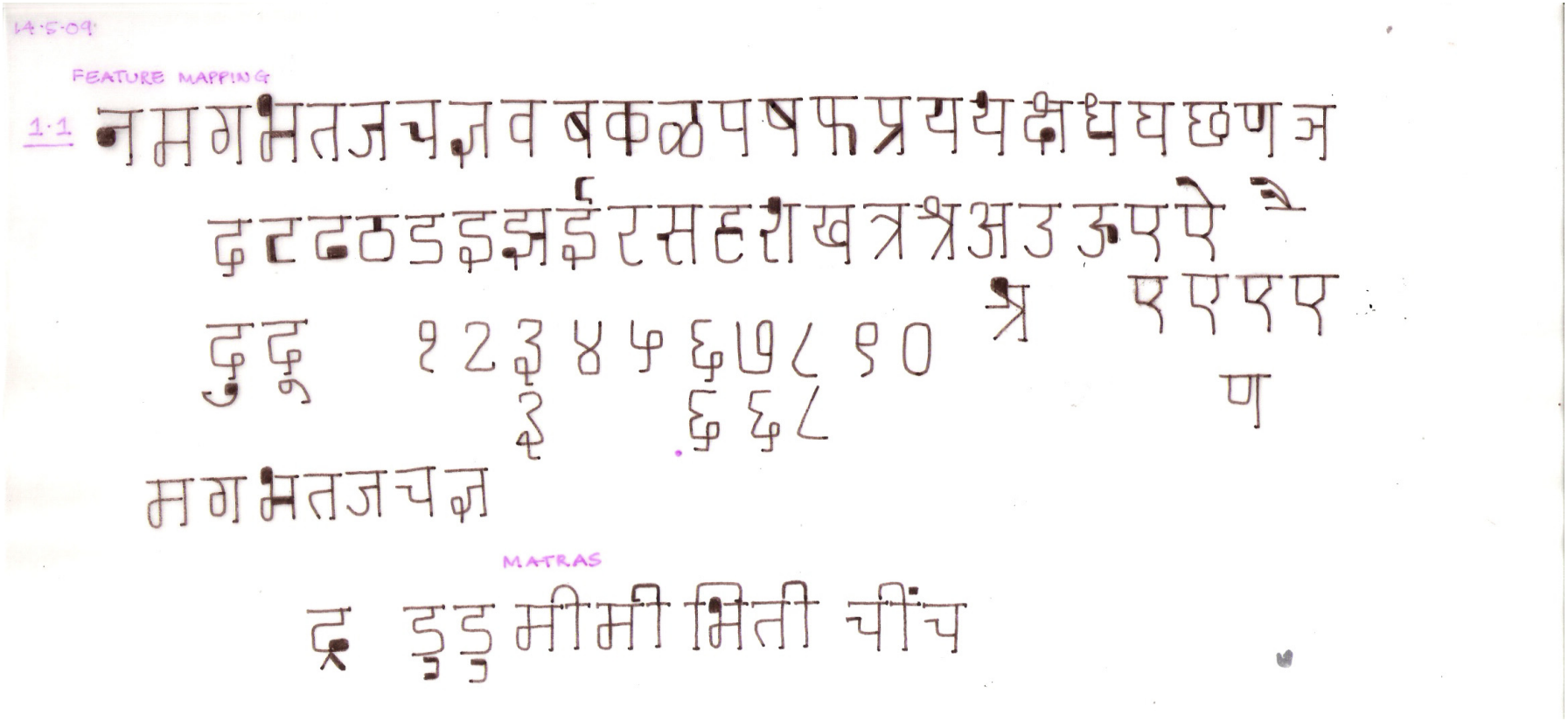
three point

अ क्ष ण भ थ श

Character design; form exploration



Character design; geometric construction

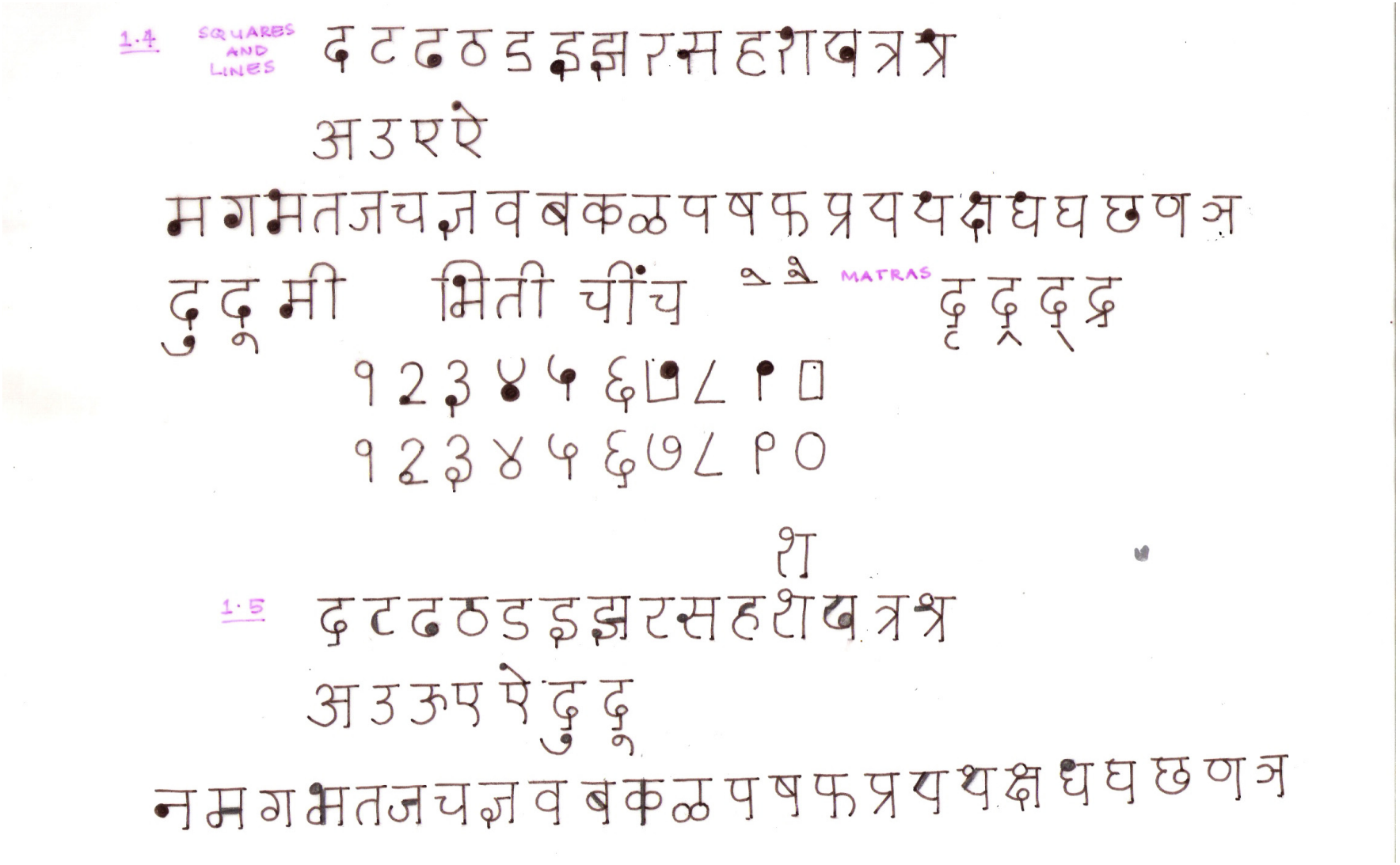


Characters based on lines and slight curves.

Form inspired from the construction of OCR-A

- Letterforms lack the devanagari character

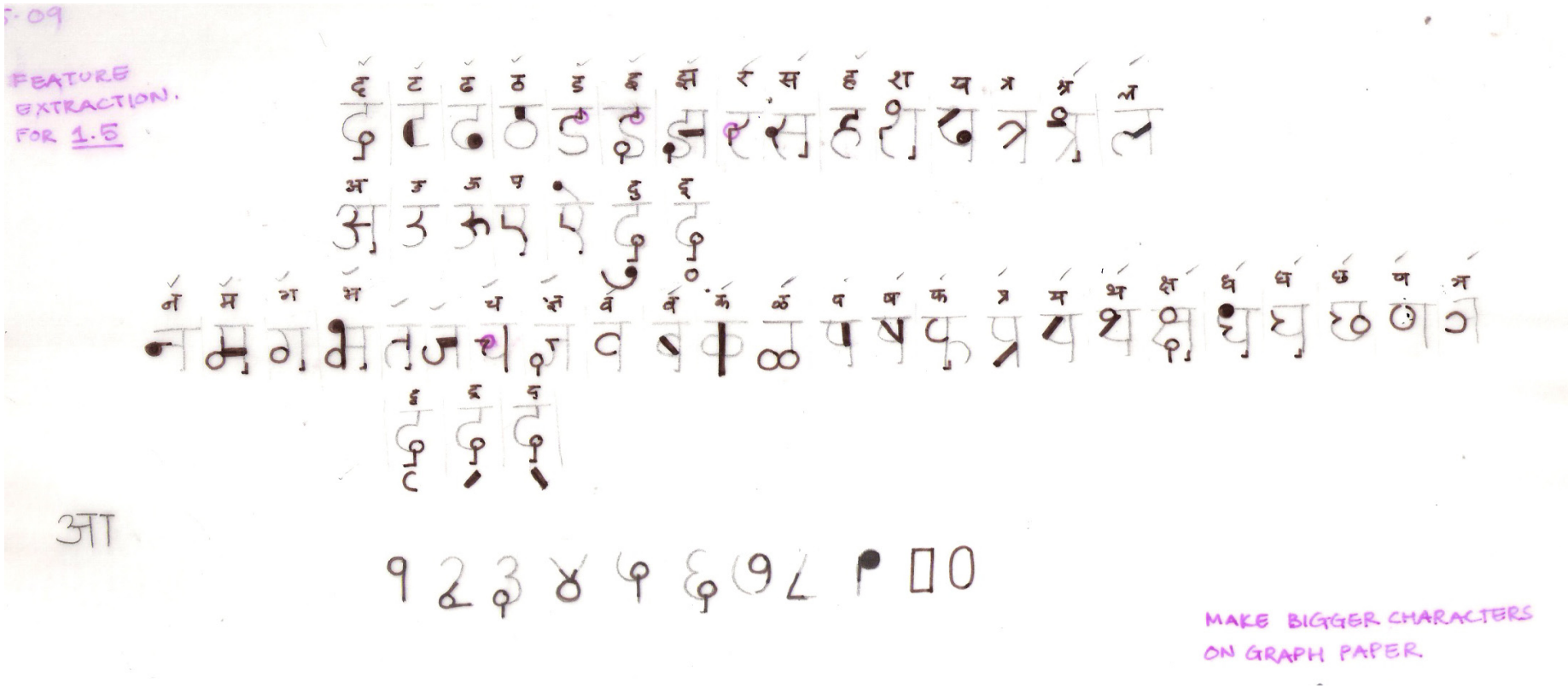
Character design; geometric construction



The circle makes characters more open, readable and restores its natural form

Feature extraction

Feature extraction; highlighting features that distinguish one character from the other.

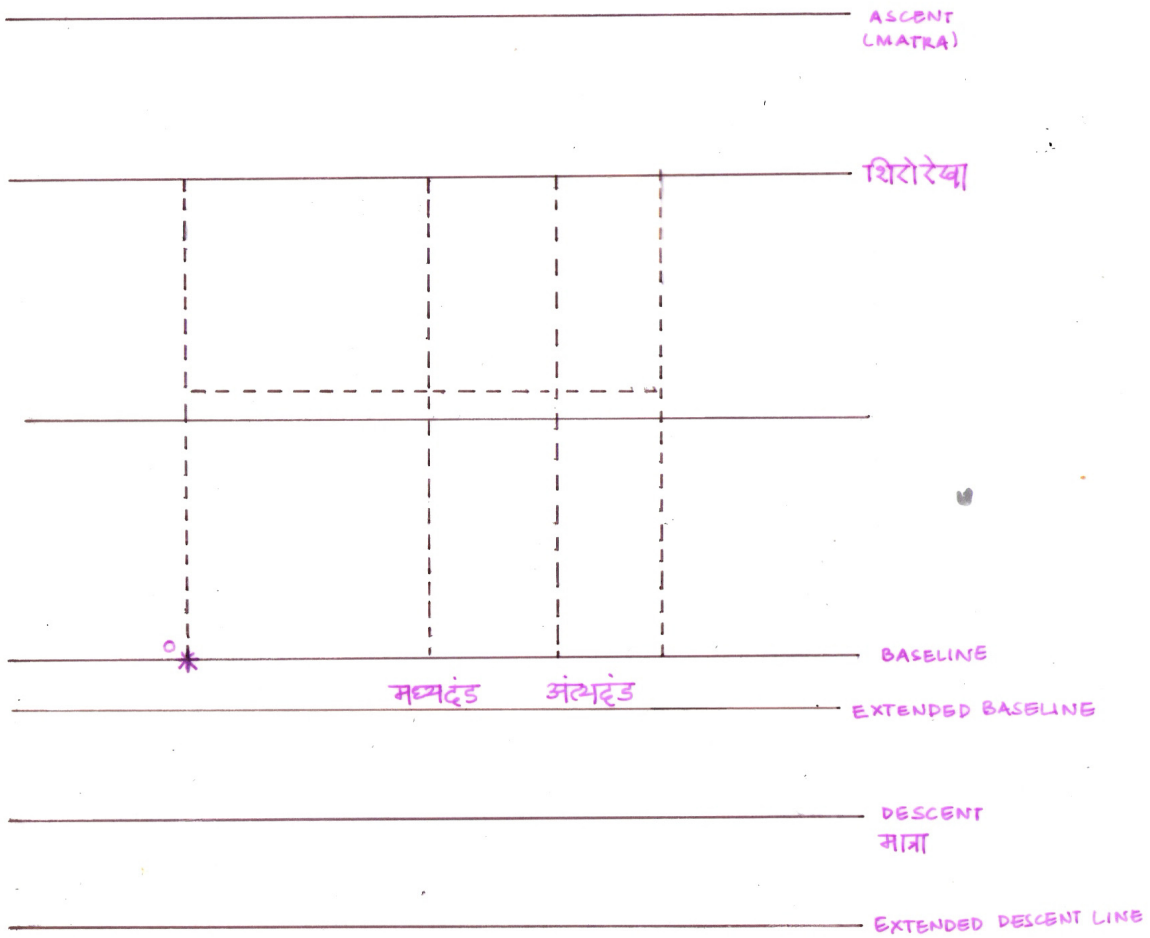


This type of a circle and line construction was finalised and letterforms were constructed on graph sheets with proper measurements and a grid.

The Grid; circle and line

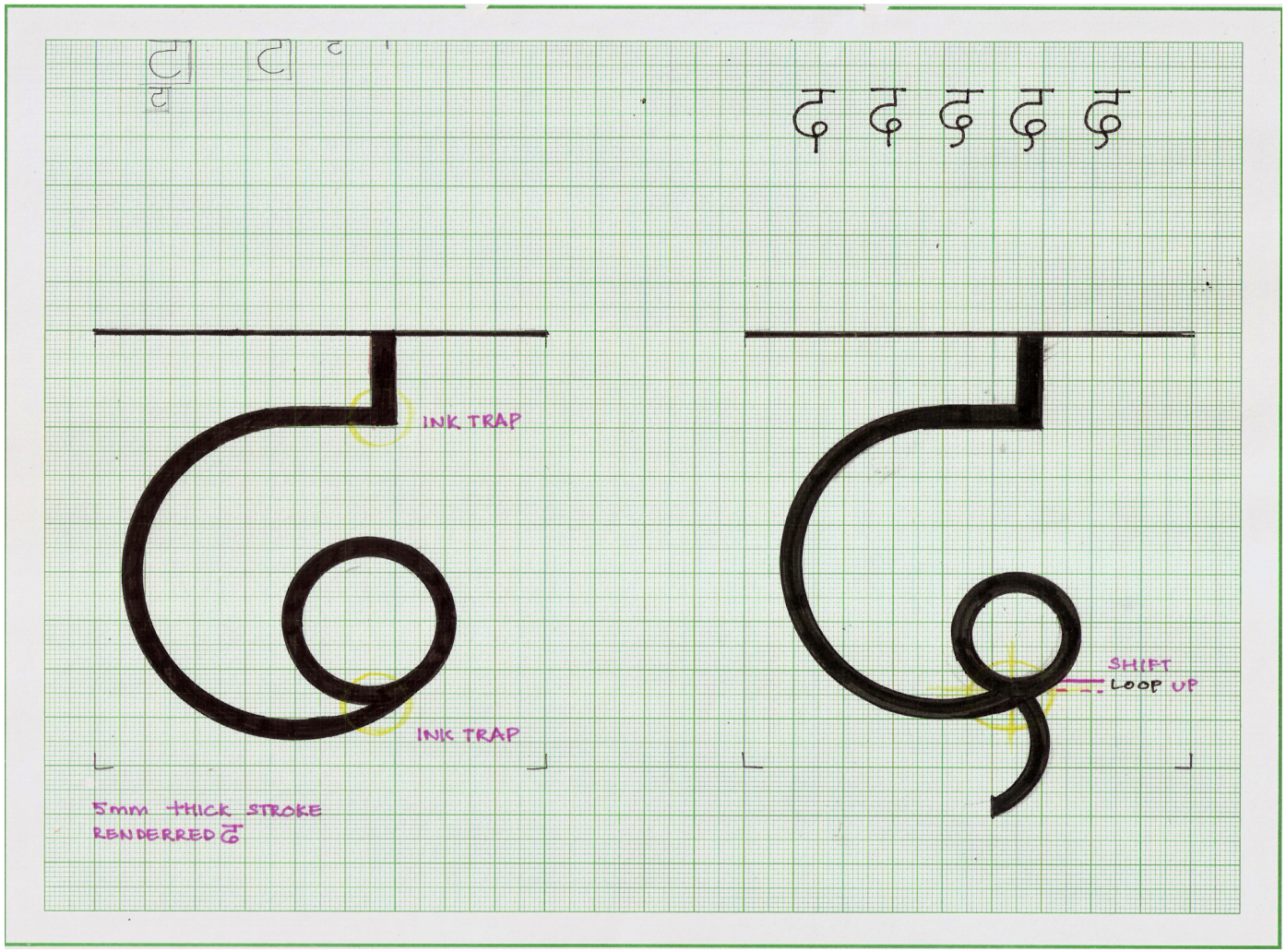
The grid defines framework of the character set

The height to width ratio is close to one

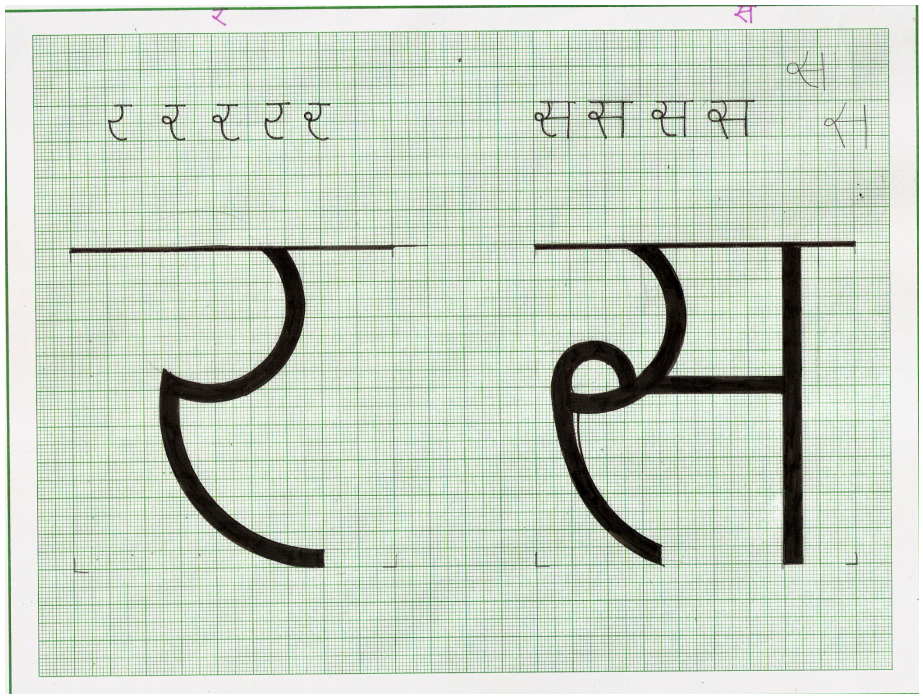


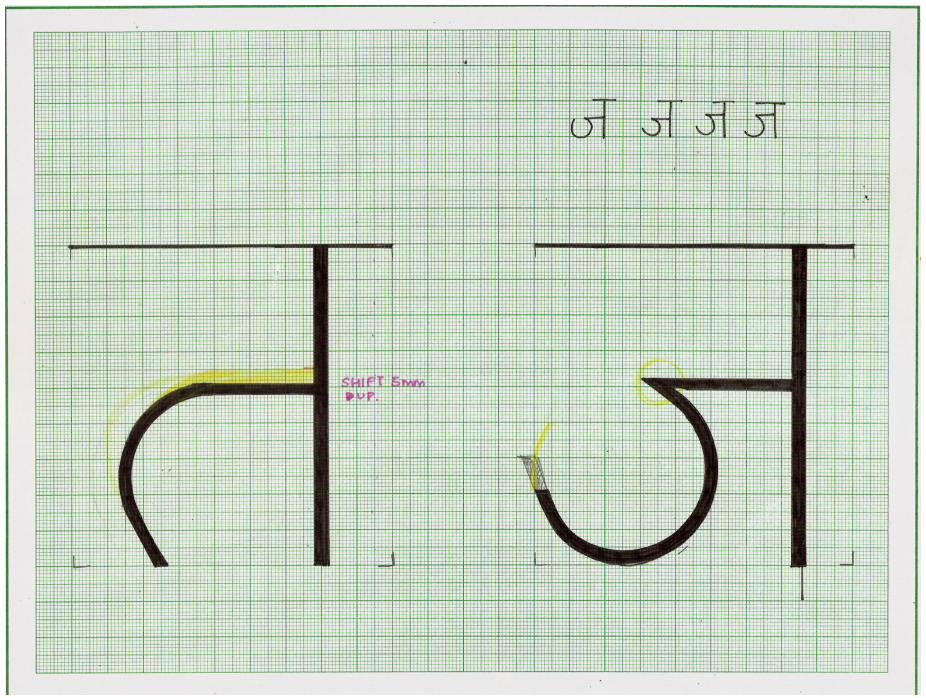
Character design; graph paper

Characters based on the grid were made on graph paper



Character design; graph paper

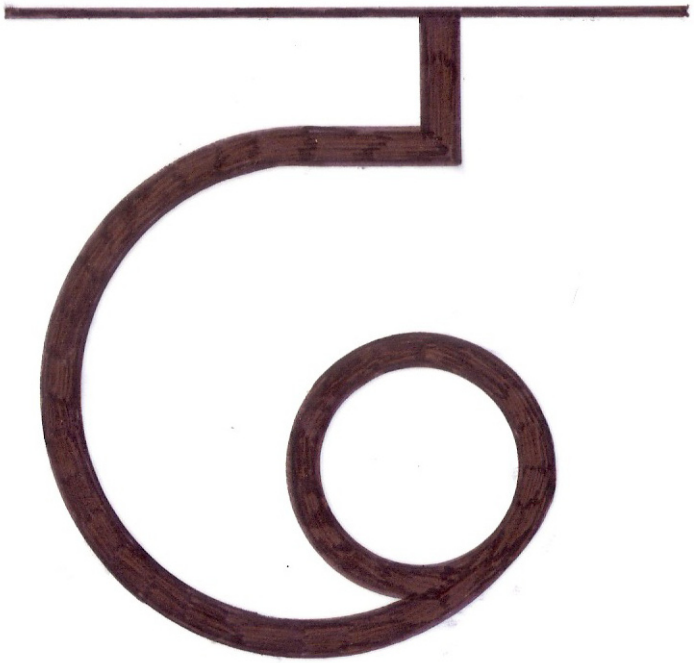




All implicit consonants have been drawn on graph sheets

Character design; tracing sheets

All characters were traced on to sheets to be scanned for digitization



Font design and generation

- Software used for initial test typeface: Fontographer

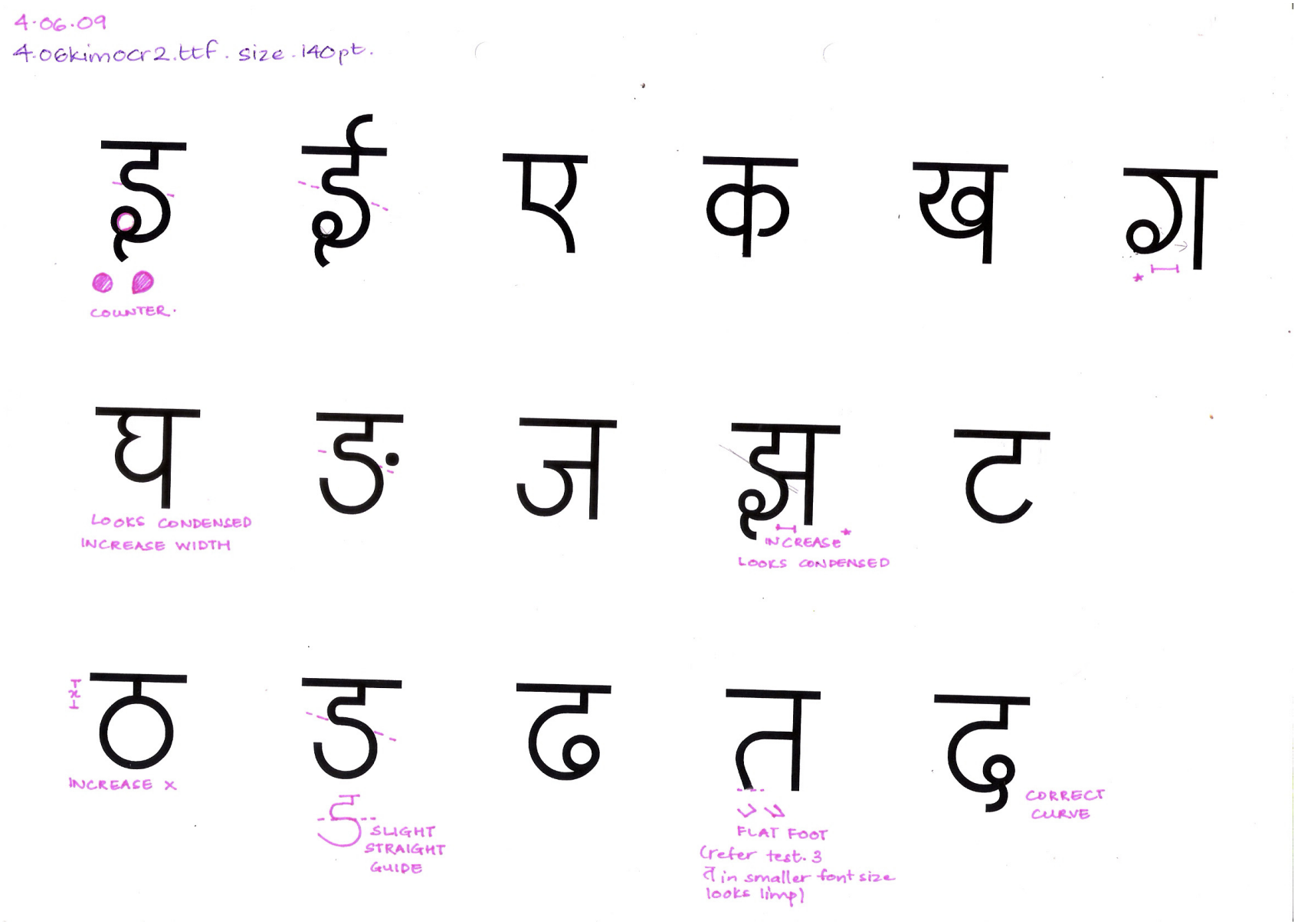
- A new set of alogarithm and codes will be written by the OCR team and the font will be tested simultaneously

- The same grid has been followed and corrections made to the character set include:
  - weight of stroke is increased
  - The distance between shirorekha and hanging characters increased

- Test prints have been reviewed and appropriate corrections have been made

- The first version of the truetype OCR font for devanagari has been sent for testing. The results of which are awaited.

Font tests


4·06·09
4·06kimocr2.ttf . size .140pt.
COUNTER.
LOOKS CONDENSED
INCREASE WIDTH
INCREASE
LOOKS CONDENSED
INCREASE x
SLIGHT
STRAIGHT
GUIDE
FLAT FOOT
(refer test. 3
त in smaller font size
looks limp)
CORRECT
CURVE

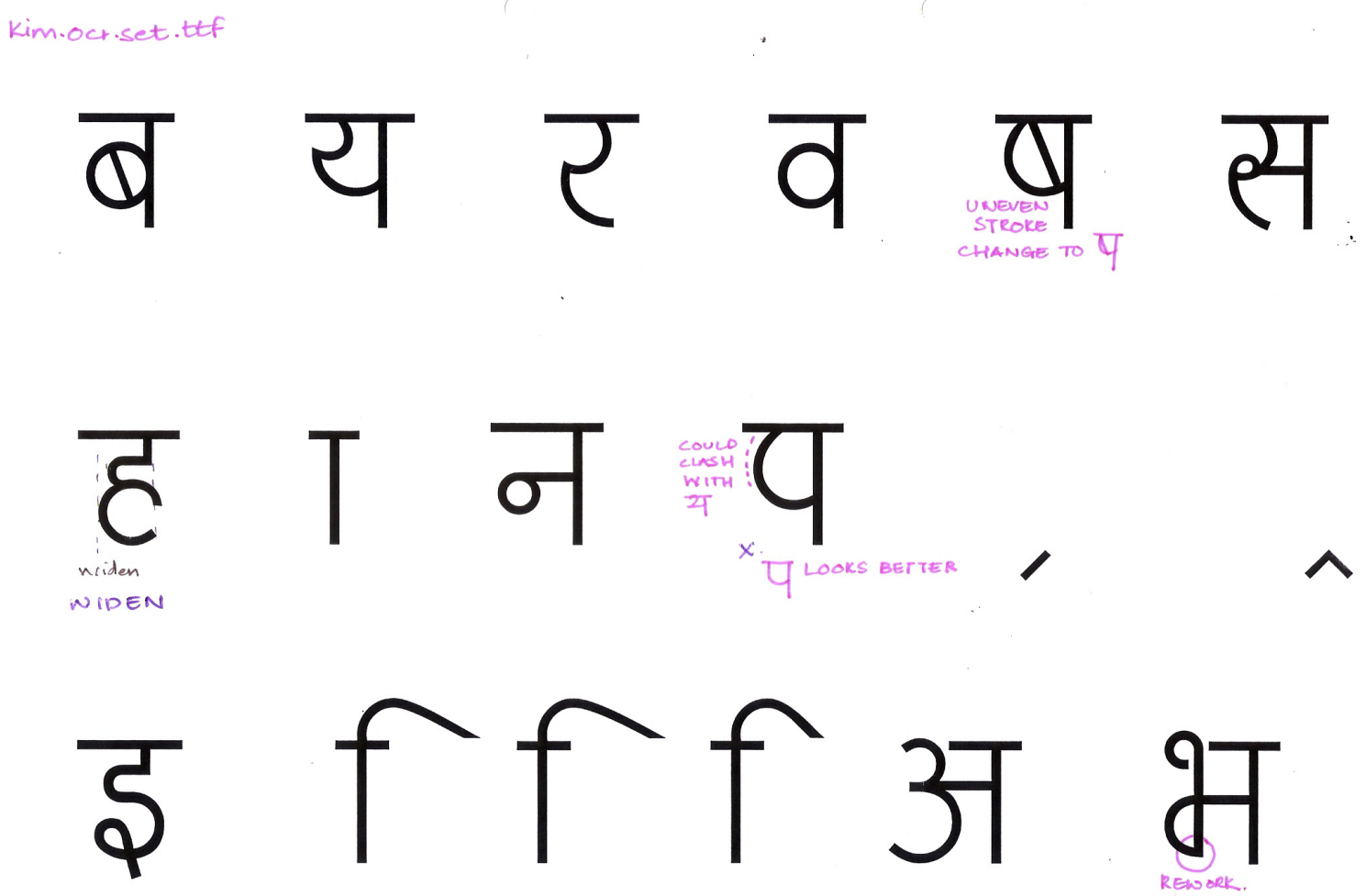
Font tests


4.06.09
4.06kimocr2.tff - 140pt.
प ब य र व ष
LOOP BIGGER
स ह न न
TRY VARIATION
WITHOUT LOOP
COUNTER VARIATION
LIKE IN ह्
MORE OPEN
IN SMALLER
POINT SIZE
-MAKE CURVES FINE
- ABRUPT JOINERIES - CORRECT

Font tests


5.06kimocr.ttf
EXTRA WEIGHT ON LEFT.
- TRY MOVING/ENDING * STROKE TO RIGHT
TRY OPENING LOOP
SLIGHT BIGGER LOOP
MAKE AGAIN

Font tests


Kim.ocr.set.ttf
UNEVEN STROKE CHANGE TO
COULD CLASH WITH
X
LOOKS BETTER
widen
WIDEN
REWORK.

Font tests

Font tests in various point sizes

ईईएकखगघङजझञटठडढतदृननपबयरवषसह

ईईएकखगघङजझञटठडढतदृननपबयरवषसह

ईईएकखगघङजझञटठडढतदृननपबयरवषसह

ईईएकखगघङजझञटठडढतदृननपबयरवषसह

ईईएकखगघङजझञटठडढतदृननपबयरवषसह

ईईएकखगघङजझञटठडढतदृननपबयरवषसह

ईईएकखगघङजझञटठडढतदृननपबयरवषसह

ईईएकखगघङजझञटठडढतदृननपबयरवषसह

ईईएकखगघङजझञटठडढतदृननपबयरवषसह

ईईएकखगघङजझञटठडढतदृननपबयरवषसह

ईईएकखगघङजझञटठडढतदृननपबयरवषसह

ईईएकखगघङजझञटठडढतदृननपबयरवषसह

ईईएकखगघङजझञटठडढतदृननपबयरवषसह

ईईएकखगघङजझञटठडढतदृननपबयरवषसह

ईईएकखगघङजझञटठडढतदृननपबयरवषसह

ईईएकखगघङजझञटठडढतदृननपबयरवषसह

ईईएकखगघङजझञटठडढतदृननपबयरवषसह

अकखगघङंजझटठडढ

तदधनपबयवषसहाइभ

मलऐणपछचउऊ

Font tests

अकखगघङजटठडढतदनपबयवषसहाइभछचउऊतदनपबयवइभमतऐणपयवषसहाइभमत
मतऐणपछचउअकखगघङजझटठडढतदनपबणपछचउऊतदनपबयवषसहाइभपछचउअकखगघङ
मतऐणपछचउटठडढतदनपबयवषसहाइभणपछचउऊतदनपबयवषसहाइभपयवषसहापयवषसहा
मतऐणपछचउऊतदनपबयवषसहाइभणपछचउऊतदनपबयवषसहाइभपछचउअकखगघङ
मतऐणपयवषसहाइभमतऐणपयवषसहाइभमतऐणपयवषसहाइभ

कमल नमन कर
अमर पठण कर
सहज हवन
टालमण भला यहरझ जटगयघ ङहरझऐ खपय हरज गघङहङट ड़यातरिझिपज

कमल नमन कर
अमर पठण कर
सहज हवन
टालमण भला यहरझ जटगयघ ङहरझऐ खपय हरज गघङहङट ड़यातरिझिपज

कमल नमन कर
अमर पठण कर
सहज हवन
टालमण भला यहरझ जटगयघ ङहरझऐ खपय हरज गघङहङट ड़यातरिझिपज

कमल नमन कर
अमर पठण कर
सहज हवन
टालमण भला यहरझ जटगयघ ङहरझऐ खपय हरज गघङहङट ड़यातरिझिपज

कमल नमन कर
अमर पठण कर

सहज हवन
टालमण भला यहरझ जटगयघ ङहरझऐ खपय हरज गघङहङट ड़यातरिझिपज

कमल नमन कर
अमर पठण कर
सहज हवन
टालमण भला यहरझ जटगयघ ङहरझऐ खपय हरज गघङहङट ड़यातरिझिपज

कमल नमन कर
अमर पठण कर
सहज हवन

Future prospects

- Develop the entire character set, including conjuncts, explicit consonants, numerals and punctuations

- Review results from OCR team and revert back with corrections till desired quality is achieved

- Generate a devanagari font that gives optimum accuracy results for character recognition when scanned

- Proposal to continue project as Project 2