

CHAPTER 2

DATASET

2.1 KANNADA SCRIPT

Kannada is one of the major four major Dravidian languages of South India, spoken predominantly in Karnataka. The Kannada alphabet evolved primarily in the Kadamba and Chalukya scripts. Kannada has become a classical language status from November 1st, 2008. Kannada is written using its own unique script unlike a few other languages that depend on other languages for their script. Kannada script is written horizontally from left to right and the concept of lower and upper case is absent. Kannada is a non-cursive script. This means characters are isolated within a word. Kannada is one of the scheduled languages of India and official and administrative language of Karnataka. Also, all the government forms are available in both the state language and English. So, people can interact with computers in the native language through the medium of handwriting. Thus, arises the need for developing character recognition system for Kannada language, especially for the offline system despite intense research the result comparable to human is still not been able to achieve due to data scarcity.

Kannada language consists of 49 phonemic letters (Varnamaale). This is divided into three groups: they are 13 Vowels, 34 consonants and other two special characters (called Anuswara and Visarga) as shown in Figure 2.1. The script has its own numerals. It has vowel modifiers and consonant modifiers for each vowel and consonants. This combination is called Syllable (Akshara). Vowels are independent of themselves, but the consonants are dependent on the vowels. The consonants are divided into two sets they are Vargeeya and Avargeya Vyanjanas. The vowel modifiers for the consonant 'ka' is shown in Figure 2.2. Similar all consonants can have vowel modifiers.

ಅ	ಆ	ಇ	ಈ	ಉ	ಊ	ಋ	
ಎ	ಏ	ಐ	ಒ	ಓ	ಔ	ಅಂ	ಆಃ
ಕ	ಖ	ಗ	ಘ	ಙ			
ಚ	ಛ	ಜ	ಝ	ಞ			
ಟ	ಠ	ಡ	ಢ	ಣ			
ತ	ಥ	ದ	ಧ	ನ			
ಪ	ಫ	ಬ	ಭ	ಮ			
ಯ	ರ	ಲ	ವ	ಶ	ಷ	ಸ	ಹ ಳ

Figure 2.1: Kannada Aksharas

2.2 DATABASE COMPLEXITY REDUCTION

Each consonant combines with each of the vowels to form a compound character (CV combination) called Akshara. Optionally, an Akshara can have one or more consonants preceding a CV combination forming a canonical structure of ((C)C) CV. Thus, the number of theoretically possible combinations of Kannada characters is huge and is listed in Table 2.1.

Char Type	V	C	CV	CCV	CCCV	TOTAL
Possible Combinations	13	36	468	16848	606528	623893
Example	ಅ	ಕ	ಕಾ	ಕ್ರೀ	ಕ್ಷೀ	

**Table2.1: Number of possible combinations of
Kannada characters (V: Vowels, C: Consonants)**

From Table 2.1, it is clear that considering each combination as a separate class for recognition increases the computational cost and may reduce the recognition accuracy. Also, it is not practically feasible to collect all the combinations during data collection. In this section, a method is described to get a comprehensive set of symbols that can be employed in the recognition of any Kannada Akshara.

All vowel modifiers change the shape of the consonant in the CV combination. Based on their relative position in a CV combination, vowel modifiers can be classified into different types:

- a) Some vowel modifiers have significant overlap in the writing direction. CVs shown in Figure 2.3 are treated as distinct classes.

ಕ ಕಾ ಕಿ ಕು ಕೂ ಕೆ ಕೊ ಕೌ

Figure 2.2: Vowel Modifiers Which Have Significant Overlap in The Writing Direction

b) For some vowel modifiers, a special pattern is written at the bottom right of the modified consonant. From the point of view of recognition, it would suffice to recognize the modifier separately and then append it to the corresponding base character.



Figure 2.3: The Consonant with The Vowel Modifiers Below

c) Three vowel modifiers are written separately towards the right of the consonant. These vowel modifiers are segmented and recognized as separate classes, which reduces the number of classes.



Figure 2.4: The Consonant with Vowel Modifiers Written Separately to The Right

d) There are some special cases of these modifiers, where they are written from below the consonant. The corresponding CV combinations are again treated as distinct classes as shown in fig 2.6

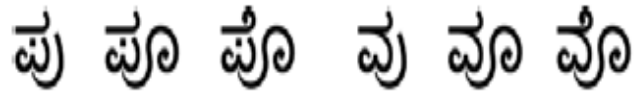


Figure 2.5: Special Consonants for Which Few Vowel Modifiers Start Below the Consonant and Written Towards Right

Hence while recognizing a character in a word, first the base character is recognized and then the vowel modifiers and Vatt Aksharas are recognized and appended to base character.

2.3 DATASET USED FOR THE PROJECT

The average number of samples collected for each class is 25.

[illegible]

Figure 2.6 Kannada Main Aksharas

ອ ປ ຕ ວ ນ ຈ ພ ງ ຊ ຋ ຍ ັ ື ູ
 ະ ົ ຼ ຽ ຾ ຿ ບ ຢ ຣ ລ ອ ຟ ຜ ຝ
 ພ ຸ ມ

Figure 2.7 Kannada Vatt Aksharas

Next, we scan a Kannada document and create PNG image, and then we run the image through a segmentation algorithm to extract the characters. The code separates the individual characters as Main and Vatt aksharas from the words and saves all of them into separate images. The results for one such word are shown below:



Figure 2.8 Segmented Characters

All the extracted characters from the PNG image of the documents are now categorized into separate classes of Main aksharas and Vattaksharas manually as shown in the figures below. The mainaksharas are labelled from 0 to 339 and the vattaksharas are labelled from 0 to 31.

The classes of Main Aksharas and Vatt Aksharas are based on the label shown in the figure 2.9.

Kannada	UID	Kannada	UID
೦೦	0	೦೦	0
೦೧	1	೦೧	1
೦೨	2	೦೨	2
೦೩	3	೦೩	3
೦೪	4	೦೪	4
೦೫	5	೦೫	5
೦೬	6	೦೬	6
೦೭	7	೦೭	7
೦೮	8	೦೮	8
೦೯	9	೦೯	9
೧೦	10	೧೦	10
೧೧	11	೧೧	11
೧೨	12	೧೨	12
೧೩	13	೧೩	13
೧೪	14	೧೪	14
೧೫	15	೧೫	15
೧೬	16	೧೬	16
೧೭	17	೧೭	17
೧೮	18	೧೮	18
೧೯	19	೧೯	19
೨೦	20	೨೦	20
೨೧	21	೨೧	21

Fig. 2.9 Main Aksharas and Vatt Aksharas Classes After Labelling

Now the dataset is ready in order to train the CNNs.

2.4 UNICODE

Unicode is defined as “a computing industry standard for the consistent encoding, representation, and handling of text expressed in most of the world's writing systems.” The Unicode concept further replaced the 8-bit encoding system such as ASCII. Unicode uses 16 bits (specifically UTF-16 uses 16 bits), which is way more than enough to represent characters in all of the world’s living languages, as well as historic scripts such as Brahmi. UTF-16 assigns each of its characters with a unique 16-bit identification number. The codes for Kannada characters are in the range of 0x0C82 to 0x0CF2. This range is reserved exclusively for Kannada characters.

In our application we define our own UID for each Kannada character for both the Mainaksharas and the vattaksharas, then for printing we translate it to Unicode using our own lookup table. A Kannada character may be a single Unicode character or may actually be a combination of two Unicode characters. The figure shows the Unicode format for Kannada literature and the other figure shows the snippet of our own Unicode lookup table along with the UID.


Kannada ^{[1][2]}																	
Official Unicode Consortium code chart  (PDF)																	
	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F	
U+0C8x	□	◌̣	◌̤	◌̥	◌̦	ಅ	ಆ	ಇ	ಈ	ಉ	ಊ	ಋ	ೠ		ಎ	ಏ	
U+0C9x	ಐ	◌̧	ಒ	ಓ	ಔ	ಕ	ಖ	ಗ	ಘ	ಙ	ಚ	ಛ	ಜ	ಝ	ಞ	ಟ	
U+0CAx	ಠ	ಡ	ಢ	ಣ	ತ	ಥ	ದ	ಧ	ನ	◌̨	ಪ	ಫ	ಬ	ಭ	ಮ	ಯ	
U+0CBx	ರ	ಱ	ಲ	ಳ	◌̣	ವ	ಶ	ಷ	ಸ	ಹ	◌̣	◌̣	◌̣	಼	ಽ	ಾ	ಿ
U+0CCx	ೀ	ು	ೂ	ೈ	ೃ	◌̣	ೇ	ೋ	ೈ	◌̣	ೊ	ೋ	ೌ	ಋ	ೠ		
U+0CDx	◌̣	◌̣	◌̣	◌̣	◌̣	ೀ	ೈ	◌̣	◌̣	◌̣	◌̣	◌̣	◌̣	◌̣	ಱ	◌̣	
U+0CEx	ಋ	ೠ	ೡ	ೢ	ೣ	೤	೥	೦	೧	೨	೩	೪	೫	೬	೭	೮	
U+0CFx	೯	೺	೻	೼	೽	೾	೿	೺	೻	೼	೽	೾	೿	೺	೻	೼	

Fig. 2.10 Unicodes for Kannada character

Kannada	UID	Hex 1	Hex 2	Dec 1	Dec 2
ಂ	0	0C82	0	3202	0
ಁ	1	0CD5	0	3285	0
ಕ	2	0CB0	0CCD	3248	3277
ಛ	3	0C85	0	3205	0
ಞ	4	0C86	0	3206	0
ತ	5	0C87	0	3207	0
ಥ	6	0C88	0	3208	0
ದ	7	0C89	0	3209	0
ಡ	8	0C8A	0	3210	0
ನ	9	0C8B	0	3211	0
ಢ	10	0C8E	0	3214	0
ಣ	11	0C8F	0	3215	0
ಠ	12	0C90	0	3216	0
ಡ	13	0C92	0	3218	0
ಢ	14	0C93	0	3219	0
ತ	15	0C94	0	3220	0
ಥ	16	0C95	0	3221	0
ಕ	17	0C95	0CBE	3221	3262
ಛ	18	0C95	0CBF	3221	3263
ತ	19	0C95	0CC1	3221	3265
ಥ	20	0C95	0CC2	3221	3266
ಕ	21	0C95	0CC6	3221	3270

Figure 2.11 Lookup table for Main Aksharas

Kannada	UID	Hex 1	Hex 2	Dec 1	Dec 2
ಂ	0	0CC3	0	3267	0
ಁ	1	0CD6	0	3286	0
ಂ	2	0CCD	0C95	3277	3221
ಁ	3	0CCD	0C96	3277	3222
ಂ	4	0CCD	0C97	3277	3223
ಁ	5	0CCD	0C9A	3277	3226
ಂ	6	0CCD	0C9B	3277	3227
ಁ	7	0CCD	0C9C	3277	3228
ಂ	8	0CCD	0C9E	3277	3230
ಁ	9	0CCD	0C9F	3277	3231
ಂ	10	0CCD	0CA0	3277	3232
ಁ	11	0CCD	0CA1	3277	3233
ಂ	12	0CCD	0CA3	3277	3235
ಁ	13	0CCD	0CA4	3277	3236
ಂ	14	0CCD	0CA5	3277	3237
ಁ	15	0CCD	0CA6	3277	3238
ಂ	16	0CCD	0CA7	3277	3239
ಁ	17	0CCD	0CA8	3277	3240
ಂ	18	0CCD	0CAA	3277	3242
ಁ	19	0CCD	0CAB	3277	3243
ಂ	20	0CCD	0CAC	3277	3244

Figure 2.12 Lookup Table for Vatt Aksharas