User Manual

Lipi Indic Character Recognizers 4.0

Contents

1 Introduction

The Lipi Indic Character Recognizers package *lipi-reco-indic-char 4.0* contains separate recognizers for isolated handwritten *consonants* & *matras* written in different Indic scripts. The scripts supported are Devnagari, Tamil, Telugu and Bangla. They are built using the Lipi Core Toolkit, and are meant for users interested in integrating such recognition capabilities into an application.

This document describes the steps for installing the recognizer and integrating it into an application.

The recognizer package comes with the following shape recognizers:

- **DEVNAGARI:** For Devnagari consonants and matras
- **TAMIL:** For Tamil consonants and matras
- **TELUGU:** For Telugu consonants and matras
- **BANGLA:** For Bangla consonants

The recognizers use the Nearest Neighbor(nn)/Active-dtw(activedtw)/Neural Network(neuralnet) shape recognition methods from the Core Toolkit.

Note that these are separate recognizers for the corresponding input types. It is up to the application to ensure that the recognizer is provided with the right input. The recognizers return one or more candidates from the characters they were trained for, along with confidence values.

Users interested in understanding how to create such recognizers using the toolkit and using the different recognition methods available in Lipi Core Toolkit may refer to the Core Toolkit User Manual.

2 Prerequisites

This section describes the prerequisites for installing and using the package.

2-1 Supported platforms and environment

lipi-reco-indic-char 4.0 has been tested on the following platforms:

- Windows 7, 32 and 64 bit
- Ubuntu 10.10, 32 and 64 bit

2-2 Disk space requirements

lipi-reco-indic-char 4.0 has packages for Windows and Linux.

Package	Package size	Disk space required
lipi-reco-indic-char4.0.0-x86.exe	32 MB	40 MB
lipi-reco-indic-char4.0.0-x64.exe	32 MB	40 MB
lipi-reco-indic-char4.0.0-linux.tar.gz	32 MB	32 MB

2-3 Lipi Toolkit

Lipi Toolkit 4.0 needs to be installed before installing *lipi-reco-indic-char 4.0* package. The environment variable LIPI_ROOT needs to be set to point to the Lipi Toolkit installation directory.

For instructions, please refer section 2 of Getting Started.

3 Installation and Setup

This section describes the steps for installing the recognizer package on the target machine and setting up the environment.

3-1 Installing the recognizer package

lipi-reco-indic-char 4.0 is available in the form of packages for Windows (32 and 64 bit) and Linux. Select and download the package for your development platform.

Windows

Once downloaded, you can install the package on Windows platform by double clicking the exe (lipi-reco-indic-char4.0.0-x86.exe OR lipi-reco-indic-char4.0.0-x64.exe) and as a result, Indic recognizers get installed as folders under $LIPI_ROOT/projects. The config file lipiengine.cfg also gets updated with project and logical name of these recognizers automatically as a part of this installation.

Linux

On the Linux platform, tar command can be used to untar-uncompress the compressed package:

```
$ tar -xzvf lipi-reco-indic-char4.0.0-linux.tar.gz
```

Once uncompressed, Indic directories have to be copied to the projects location under $LIPI_ROOT.

```
$ cp -r bangla/ tamil/ devnagari/ telugu/ $LIPI_ROOT/projects
```

Unlike in Windows, lipiengine.cfg does not get updated automatically with the project and logical names of the recognizers. These must be added explicitly to lipiengine.cfg as shown below:

Append the below entries to $LIPI_ROOT/projects/lipiengine.cfg:

```
SHAPEREC_BANGLA=bangla(default)
SHAPEREC_TAMIL=tamil(default)
SHAPEREC_DEVNAGARI=devnagari(default)
SHAPEREC_TELUGU=telugu(default)
```

4 Recognizers in the package

Most Indic scripts are descendents of the ancient Brahmi script and are defined as ``syllabic alphabets'' in that the unit of encoding is a syllable. In general, these syllabic units are the smallest units of isolated writing in the Indic scripts, and hence the nearest thing to isolated characters. These units are written from left to right.

These syllabic units correspond to

• Isolated vowels (e.g., 'u', represented as ఉ in Telugu, உ in Tamil)

• Isolated consonants (with inherent neutral vowel a) (e.g, 'ka', represented as క). The vowel may be muted by a special diacritic (called *halantha* in Devnagari).

• CV combinations where a consonant (C) has been modified by a vowel (V) and is indicated by a vowel diacritic (e.g. 'k$\bar{a}$', represented as కా). These vowel diacritics are called *matras* in Devnagari.

• Clusters of 2 or more consonants modified by vowels (CCV, CCCV, and so forth) (e.g. 'kro', represented as క్రో). Consonant conjuncts such as CC, CCC etc are called *samyuktasharas* in Devnagari).

Theoretically, the number of syllables runs into the hundreds, though a much smaller subset is used in practice. From recognition point of view, considering each syllable as a pattern class (symbol) is practically impossible. The approach used here is to identify a smaller subset of basic units sufficient to cover the entire set of syllables. These basic units need not always be the constituent graphemes of the syllables because of the structural complexities. For example, certain vowel 'maatra' can be fused inseparably with the underlying consonant, as in the syllable 'చి' in Telugu. In such a case, recognition based on considering the consonant and vowel as basic units would have to deal with segmentation problems. In practice, the basic units are determined by various factors like ease of segmentation of characters into these units and not just by linguistic criteria.

Indic scripts are generally non-cursive in writing style and hence pen-up usually separates the basic graphemes though not always. So, the basic graphemes of the script i.e. independent vowels, consonants, vowel diacritics and consonant modifiers are included in the symbol sets of these recognizers. In addition, the symbol set also contains some symbols which do not have linguistic interpretation but have stable pattern across writers and help reduce the total number of symbols to be collected. Note that *CV combinations* are *not* included in the symbol set, in order to be recognized the consonant and the vowel diacritic must be interpreted individually. Similarly, only a few consonant conjuncts that have distinctive shapes are included as symbols in this version of the recognizers. A conjunct may still be interpreted by writing it as two (or more) isolated consonants and using the vowel muting diacritic.

A more detailed description of the structure of Indic scripts may be found at http://www.hpl.hp.com/india/documents/papers/HPL-2008-45.pdf.

The tables below list the symbol sets for the different recognizers along with their symbol ids, and UNICODE values in parentheses.

Note: For some symbols that are part of the recognizers there is no corresponding UNICODE value. These symbols have been shown with UNICODE value as (0).

4-1 Devnagari

Devnagari is descendent of the ancient Brahmi script and is the script of many languages including Hindi, Marathi, Gujarati, Nepali and Sanskrit. Devnagari is written in non-cursive or discrete form from left to right. Each consonant carries an inherent vowel a. This inherent vowel can be muted by the diacritic *halanta*, or can be modified by other vowels diacritics. Two or more consonants can be combined together with vowels to form ligatures, or conjuncts or samyuktasharas.

The symbol set represented in the recognizer includes the *shirorekha* or headline (symbol 0), isolated vowel symbols (1-11), consonants and selected conjuncts (12-46), vowel and other modifiers (47-64), for a total of 65 symbols. The following table lists out the consonants and matras symbols present in the recognizer along with their class ids and UNICODE symbols in "class id (UNICODE)" format.

— 0 (0)	ओ 10 (0913)	झ 20 (091D)	ध 30 (0927)	व 40 (0935)	ु 50 (0941)	ः 60 (0903)
अ 1 (0905)	औ 11 (0914)	ञ 21 (091E)	न 31 (0928)	श 41 (0936)	ू 51 (0942)	़ 61(094D)
आ 2 (0906)	क 12 (0915)	ट 22 (091F)	प 32 (092A)	ष 42 (0937)	ृ 52 (0943)	ऽ 62 (093C)
इ 3 (0907)	ख 13 (0916)	ठ 23 (0920)	फ 33 (092B)	स 43 (0938)	े 53 (0947)	८ 63 (0)
ई 4 (0908)	ग 14 (0917)	ड 24 (0921)	ब 34 (092C)	ह 44 (0939)	ै 54 (0948)	∧ 64 (0)
उ 5 (0909)	घ 15 (0918)	ढ 25 (0922)	भ 35 (092D)	क्ष 45 (0915 094D 0936)	ो 55 (094B)	
ऊ 6 (090A)	ङ 16 (0919)	ण 26 (0923)	म 36 (092E)	त्र 46 (0924 094D 0930)	ौ 56 (094C)	
ऋ 7 (090B)	च 17 (091A)	त 27 (0924)	य 37 (092F)	ा 47 (093E)	ं 57 (0902)	
ए 8 (090F)	छ 18 (091B)	थ 28 (0925)	र 38 (0930)	ि 48 (093F)	ॅ 58 (0945)	

ऐ 9 (0910)	ज 19 (091C)	द 29 (0926)	ल 39 (0932)	ी 49 (0940)	ॅ 59 (0901)	

Table 1: Symbol set of Devnagari Character Recognizer

4-2 Telugu

Telugu script has 18 vowels and 36 consonants of which thirteen vowels and thirty five consonants are in common usage. Of all the Indic scripts, the Telugu script has the largest number of vowels and consonants. Theoretically, the number of syllables is $O(10^4)$ though a much smaller subset is used in practice.

The following table lists out the symbols present in the recognizer along with their class ids and UNICODE symbols in "class id (UNICODE)" format.

అ 0 (0C05)	ఐ 10 (0C10)	ౘ 20 (0C1E)	ణ 30 (0C23)	మ 40 (0C2E)	ౕ 50 (0)	ౠ 60 (0C4A)
ఆ 1 (0C06)	ఒ 11 (0C12)	చ 21 (0C1A)	త 31 (0C24)	య 41 (0C2F)	ఱ 51 (0C31)	ౡ 61(0C4B)
ఇ 2 (0C07)	ఓ 12 (0C13)	ఛ 22 (0C1B)	థ 32 (0C25)	ర 42 (0C30)	ా 52 (0C3E)	ౢ 62 (0C4D)
ఈ 3 (0C08)	ఔ 13 (0C14)	జ 23 (0C1C)	ద 33 (0C26)	ల 43 (0C32)	ి 53 (0C3F)	ౣ 63 (0)
ఉ 4 (0C09)	ం 14 (0C02)	ఝ 24 (0C1D)	ధ 34 (0C27)	వ 44 (0C35)	ీ 54 (0C40)	౤ 64 (0)
ఊ 5 (0C0A)	ః 15 (0C03)	ఞ 25 (0C19)	న 35 (0C28)	శ 45 (0C36)	ు 55 (0C41)	
ఋ 6 (0C0B)	క 16 (0C15)	ట 26 (0C1F)	ప 36 (0C2A)	ష 46 (0C37)	ూ 56 (0C42)	
ఌ 7 (0C60)	ఖ 17 (0C16)	ఠ 27 (0C20)	ఫ 37 (0C2B)	స 47 (0C38)	ృ 57 (0C46)	
ఎ 8 (0C0E)	గ 18 (0C17)	డ 28 (0C21)	బ 38 (0C2C)	హ 48 (0C39)	ె 58 (0C47)	
ఏ 9 (0C0F)	ఘ 19 (0C18)	ఢ 29 (0C22)	భ 39 (0C2D)	ళ 49 (0C33)	ై 59 (0C48)	

Table 2: Symbol set of Telugu Character Recognizer

4-3 Tamil

Tamil, like most of the other Indic scripts, is defined as a ``syllabic alphabet'' in that the unit of encoding is a syllable. In general, these syllabic units are the smallest units of isolated writing in the Indic scripts, and hence the nearest thing to isolated characters.

These syllabic units correspond to

- isolated vowels (e.g., 'u', represented as உ)
- isolated consonants (with inherent neutral vowel a) (e.g, 'ka', represented as க)
- CV combinations where a consonant has been modified by a vowel and is indicated by a vowel diacritic (e.g. 'ku', represented as □)
- Clusters of 2 or more consonants modified by vowels (CCV, CCCV, and so forth) (e.g. 'kshū', represented as □◌ூ)

The set of 24 symbols represented in this collection includes in addition to independent V and C graphemes CV combinations where the vowel diacritics attach above or below the base C grapheme or are otherwise difficult to segment, and those vowel diacritics that occur as distinct characters to the left or right of the base C. Tamil has very few consonant symbols compared to other Indic scripts because multiple consonant sounds (e.g. "k", "kh", "g", "gh") are represented using a single symbol (க) and resolved by the speaker using the context of the word.

Like all Indic scripts, there is no tradition of writing Tamil in boxes; however informal observation of several native Tamil writers revealed that they could write "characters" as defined here, in boxes consistently with no or minimal training.

The following table lists out the symbols present in the recognizer along with their class ids and UNICODE symbols in "class id (UNICODE)" format.

அ 0 (0B85)	ஒ 10 (0B93)	ப 20 (0BAA)	ஸ 30 (0BB8)	 40 (0)
ஆ 1 (0B86)	ஃ 11 (0B83)	ம 21 (0BAE)	ஷ 31 (0BB7)	ெ 41 (0BC6)
இ 2 (0B87)	க 12 (0B95)	ய 22 (0BAF)	ஜ 32 (0B9C)	ே 42 (0BC7)
ஈ 3 (0B88)	ங 13 (0B99)	ர 23 (0BB0)	ஹ 33 (0BB9)	ை 43 (0BC8)
உ 4 (0B89)	ச 14 (0B9A)	ல 24 (0BB2)	 34 (0)	
ஊ 5 (0B8A)	ஞ 15 (0B9E)	வ 25 (0BB5)	 35 (0)	

எ 6 (0B8E)	ட 16 (0B9F)	ழ 26 (0BB4)	ா 36 (0BBE)	
ஏ 7 (0B8F)	ண 17 (0BA3)	ள 27 (0BB3)	ி 37 (0BBF)	
ஐ 8 (0B90)	த 18 (0BA4)	ற 28 (0BB1)	ீ 38 (0BC0)	
ஒ 9 (0B92)	ந 19 (0BA8)	ன 29 (0BA9)	ு 39 (0)	

Table 3: Symbol set of Tamil Character Recognizer

4-4 Bangla

Bangla originated from the ancient Indian script, Brahmi through various transformations and it is a mixture of syllabic and alphabetic scripts. This script runs from left to right.

In Bangla, the total number of characters is nearly 300, which consists of a large number of compound characters formed by combination of two or more basic characters. However, there are only 50 basic characters consisting of 11vowels and 39 consonants.

The following table lists out the consonants and vowels symbols present in the recognizer along with their class ids and UNICODE symbols in “class ID (UNICODE)” format.

অ 0 (0985)	ঔ 10 (0994)	ঞ 20 (099E)	ন 30 (09A8)	ষ 40 (09B7)
আ 1 (0986)	ক 11 (0995)	ট 21 (099F)	প 31 (09AA)	স 41 (09B8)
ই 2 (0987)	খ 12 (0996)	ঠ 22 (09A0)	ফ 32 (09AB)	হ 42 (09B9)
ঈ 3 (0988)	গ 13 (0997)	ড 23 (09A1)	ব 33 (09AC)	ড় 43 (09DC)
উ 4 (0989)	ঘ 14 (0998)	ঢ 24 (09A2)	ভ 34 (09AD)	ঢ় 44 (09DD)
ঊ 5 (098A)	ঙ 15 (0999)	ণ 25 (09A3)	ম 35 (09AE)	য় 45 (09DF)

ঋ 6 (098B)	চ 16 (099A)	ত 26 (09A4)	য 36 (09AF)	ং 46 (0982)
এ 7 (098F)	ছ 17 (099B)	থ 27 (09A5)	র 37 (09B0)	ঃ 47 (0983)
ঐ 8 (0990)	জ 18 (099C)	দ 28 (09A6)	ল 38 (09B2)	ঁ 48 (0981)
ও 9 (0993)	ঝ 19 (099D)	ধ 29 (09A7)	শ 39 (09B6)	ৎ 49 (09CE)

Table 4: Symbol set of Bangla Character Recognizer

5 Integrating the Recognizers into Applications

Essentially, once installed the recognizers are available for integration into applications via their logical names. The logical names for all recognizers can be found in lipiengine.cfg under the projects directory of $LIPI_ROOT. The following table lists out the recognizers with their logical names:

Recognizer	Logical names
Bangla	SHAPEREC_BANGLA
Tamil	SHAPEREC_TAMIL
Telugu	SHAPEREC_TELUGU
Devnagari	SHAPEREC_DEVNAGARI

The recognizers can be invoked from an application using these logical names. Digital ink from a digitizer or mouse or from a file may be passed to the recognizer, and the most plausible character (shape) IDs and corresponding confidence values are obtained in return. The section14 of Core Toolkit User Manual describes the process of integrating recognizers into application code.

The accuracy of these Indic character recognizers has been benchmarked using training and test splits of standard datasets (see http://lipitk.sourceforge.net/resources.htm) and has been estimated as below:

Recognizer	Accuracy (%)
Telugu	94.95
Devnagari	89.80
Tamil	90.55
Bangla	92.51

6 References

[1] Lipi Core Toolkit User Manual is available in the following link: **Core toolkit download page**

[2] Muralikrishna Sridhar, Dinesh Mandalapu, Mehul Patel, "Active-DTW : A Generative Classifier that combines Elastic Matching with Active Shape Modeling for Online Handwritten Character Recognition," International Workshop on Frontiers in Handwriting Recognition, 2006