Chapter 5

Issues in Classifier's Performance

5.1 Difficulties, Limitations and Scope

Collecting enough data in Thailand is the most difficult part. It took almost three months to collect from the help of Nepalese natives.

Major challenging problems associated with Nepali on-line handwriting are focussed and the proposed methodology involved in the classifier is taken into account based on the performance with peoples' natural writing styles. In chapter 3, section 3.2, a comprehensive difficulties are explained. The following paragraphs provide the way how the existing problems are fixed. Many of the techniques used in developing a prototype classifier are focussing below along with their problem fixing strategies.

- *Size normalization, cusp elimination and overlapping coordinates deletion*:

  As the size (big or small) of the character is variable from time to time writing, the technique like size normalization is used to transform the character of any size into a standard window (designed in eqn. 4.1) with preserving the strokes' information. This is useful in determining the statistical relation in between/among the strokes. Use of spatial relation in designing a classifier is explained in chapter 4. Cusp elimination improves the classification rate (Fig. 3.11) and co-occurrence coordinates deletion improves the speed of the classifier.

- *Directional property*

  The feature used in the classifier to classify the character contains the directional property of writing. Tangent angle at every position along the trajectory of pen movement provides the way how the symbol/stroke is written and the order as well. This feature produces the correct skeleton of the character's shape (Fig. 3.2). Feature vector sequence stores all information about the symbol. As it preserves the complete skeleton along with the order, there is no problem with the tilted handwriting. No skew transformation is necessary.

- *DTW*

  Alignment of two non-linear sequences is the major problem in few decades before. To some extent, with the introduction of DTW algorithm solved the problem by measuring the distance between them at the cost of time complexity. However, the degree of dimension in feature plays a crucial role in recognition speed. The higher the degree of dimension, the larger is the recognition speed.

- *A Template-Based Approach*

  The approach itself defines the way how it works. A classifier with large number of templates with variability in writing styles leads to receive the higher classification rate. It is claimed that some of the misclassified characters would be classified correctly in case the number of templates are increased. Therefore, it can be concluded that the recognition rate and speed of the system depends on number and size of the templates and test strokes, which are variably determined from a series of experiment.

- *Spatial Information*

  One of the biggest difficulties in natural Nepali handwritten character recognition is the similarity between/among the classes of character. In addition, peoples' writing styles aiding extra contribution to become one character similar to another. (छ, ध), (य, प), (ढ, द), (भ, म), (ध, घ), (थ, य) are some of the confusions pairs, which are often classified due to the consideration of spatial information about the strokes. More specifically, spatial information about the strokes within a character has a significant effect on confusion pairs: (भ, म), (थ, य) and (ध, घ). in correct classification. There are two distinguishing features, one is the location of the 'shirorekha' and next is the small curve in the text part. One of two distinguishing features is enough to classify correctly in these pairs. In many of the cases, small curve in the text is not appeared. Therefore, the location of the 'shirorekha' is the main classifying feature. It is assumed that users certainly left one of the two distinguishing features. The spatial information is carried out based on this assumption. However, it has a record that a very few characters do not show such distinguishing features.

During the test, every character is plotted along with the result (code) and hence it is easy to investigate how the test character is misclassified/rejected. A character is tested stroke by stroke basis. Therefore, it has a chance of misclassification/rejection if one of the strokes does not preserve the right information of the specific symbol. This is how the prototype classifier is related to stroke's features (information). In one hand, the developed prototype classifier recovers some of the challenging problems appearing in the natural handwriting (Nepali), however on the other hand, it has remarkable limitations. Some of the limitations are,

- *Tremor Handwriting*

  The system may not apply for handwriting of children and for tremor handwriting.

- *Fast Handwriting*

  The system is limited to very fast handwriting too. In case of fast handwriting, the plot from the features (the collection of coordinates along the pen trajectory) will not be exactly fit with the off-line graphical representation of the stroke (i.e. the shape can be changed), while it can be possible in case of slow handwriting.

- *Feature similarity*

  Most of the times, characters are misclassified due to the similarity in structure from one class of character to another. This is due to the nature of the script and writing styles too.

- *Two stroke Numeral*

  Two or more than two strokes numerals are often misclassified. The system is trained mostly with (98%) uni-stroke numerals. However, two or more than two strokes numerals can be classified correctly by using the system where such numerals are trained.

Globally, it is believed that there are some of the issues in handwriting technology from which it is always lagging behind. Some of them are,

- For transcription, the keyboard is faster than handwriting for most of the languages like Nepali, English, Urdu and so on. Therefore, handwriting technology is not appropriate here.
- Existing on-line handwriting equipments are not as comfortable and natural as pen and paper, what it should be.

The methodology utilized in building a prototype classifier is based on a template-based approach. Dynamic Programming is used to classify the characters into their classes based on their strokes' information (feature). The scope of the thesis is limited to classify Nepali discrete alphanumeric characters. This methodology can be useful not only in adding other symbols in Nepali but also in extending to other scripts such as, Roman, Chinese, Japanese and other Indian Scripts etc.