Texture,Template matching for Language Detection in Document Images

Deepak Jannarapu, Nallagatla Manikanta, Dr.Vinod Pankajakshan, IIT Roorkee

Abstract

The problem of recognising the language of a document image has large number of applications in many fields. We analysed and improved two approaches for language detection- texture and template matching. Local binary patterns are used for texture features. For template matching character templates of three languages English, Hindi, Telugu are handpicked from their respective scripts. Comparing both approaches, texture features, achieved over 95% accuracy on the classification of document images from 3 languages.

Introduction

The world we live in is becoming increasingly multilingual and, at the same time, increasingly automated. Hence the need of automatic language identification is increasing. Most used 3 methods namely texture analysis, template matching and statistical analysis are found to be accurate and feasible. The details about these methods are explained in the further sections.

Methods

After image processing step shown in figure 1, respective features are extracted from the image.

Texture approach:

Using local binary patterns texture features are extracted. To calculate the LBP value for a pixel in the gray scale image, we compare the central pixel value with the neighbouring pixel values. The whole process is shown in the figure 2

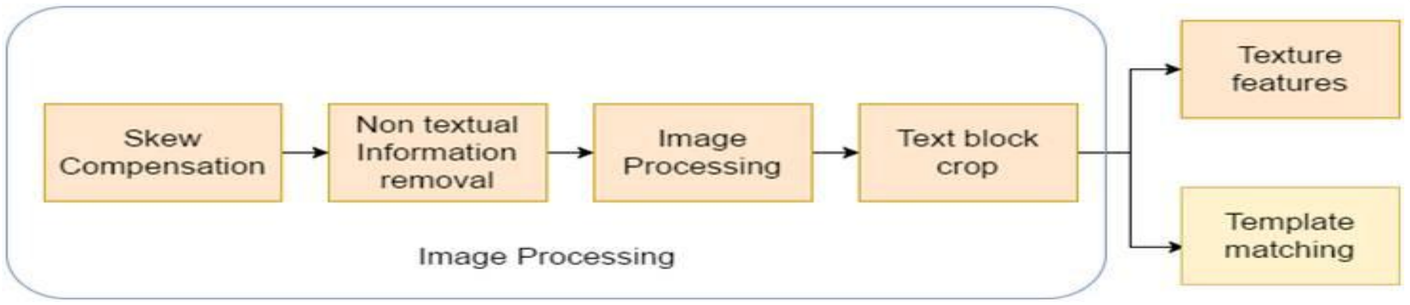

***fig 1**: Method*

***fig 2**: Local Binary Pattern example*

Template matching:

Character templates of three languages English, Hindi, Telugu are handpicked from their respective scripts. Then after performing image processing as above, the instances of these templates are searched in test document images. Sample of templates for each language are shown in introduction.

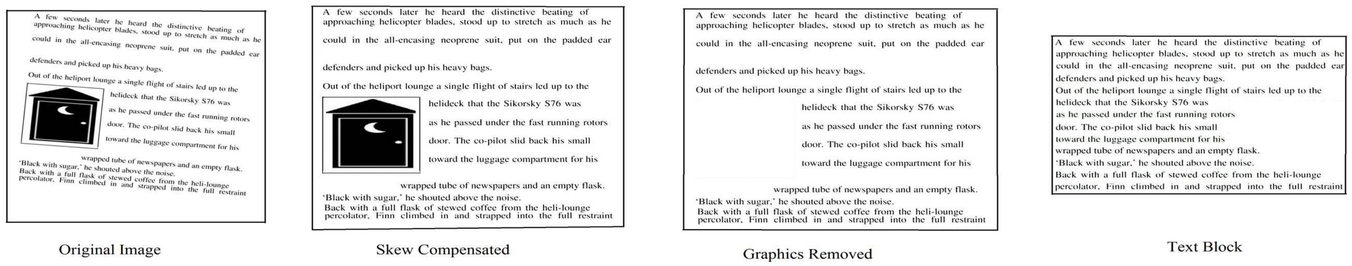

***fig 3**: Image Processing sample result*

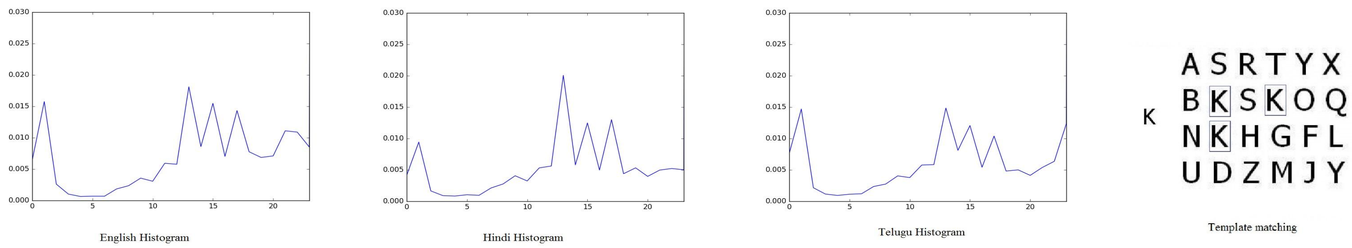

***fig 4**: Texture, template matching sample results*

Results

80 document examples were chosen for each of: English, Hindi and Telugu. Images contain graphics, and would resemble reasonably closely the output from a document segmentation system. Figure 3 shows step by step processing of a sample image used in the experiments. The images were divided into 60 training and 20 test images per language. For non text removal, connected component of more than 2.25 times average area of a character is removed. For template matching, 52 for English, 425 for Hindi, 756 for Telugu template characters are hand picked. The texture features approach achieved over 95% accuracy on the classification of 60 test document images from 3 languages are very promising over 73% accuracy with template matching features.

Uniqueness

Texture features approach is rotation invariant as local binary patterns are rotation invariant. As we are using texture for language classification texture approach is independent on number of lines work, size of image. This approach is also working for scanned documents also which has various applications like indexing, digitalizing a hard copy.

Conclusion and Future Study

We analysed and improved two methods of language detection in scanned documents for three different languages. Our system can also overcome noise which is in the form of moderate skew, numerals, foreign characters, illustrations, and blurred or fragmented characters. But main drawback of texture approach is that it is sensitive for font size in training images. The next step in our research is to link this algorithm with appropriate language character identification. Number of languages to be identified can also be increased. system need to overcome the sensitivity of font size.

References

[1] Busch, Andrew, Wageeh W. Boles, and Sridha Sridharan. "Texture for script identification." *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[2] NGADI, Mohammed, et al. "The performance of LBP and NSVC combination applied on face classification.“,2016