2017 International Conference on Advances in ICT for Emerging Regions (ICTer): 311 - 316

An Approach for Digitizing Form Based Images

Pasindu. A. De Alwis[#1], Pumudu A. Fernando[*2]

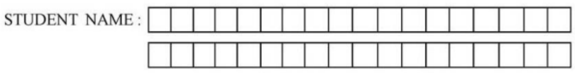
[#]*Informatics Institute of Technology*
No. 57, Ramakrishna Road,
Colombo 06, Sri Lanka.
[1]pasindu094@gmail.com
[2]pumudufernando@gmail.com

Abstract— **Numerous associations still rely on paper-escalated work processes. Due to the fact that the printed and handwritten documents are all around acknowledged and perceived for any authoritative report. The major issues having handled the paper documents are the inability to monitor the lost data, storage and money and time wasted on re-keying data. It is possible to address these problems through a solution that can digitize the data in these paper documents. The most common approach is to identify handwritten of a single person through a template matching approach. In the proposed approach, the template of the document is identified and handwritten areas are extracted through an image processing component and the identification of the handwritten characters are addressed through training the system using a convolutional neural network. The accuracy level of 90% achieved with recognition of form template and 84.67% accuracy level achieved with handwritten character recognition.**

Keywords— **Handwritten Character Recognition, Image Classification, Machine Learning, Pattern Recognition, Template Based Forms**

I. Introduction

In spite of the hype around a paperless future, the reality remains that many organizations still depend on paper-intensive workflows. Printed documents are globally accepted and recognized for any legal document and handwritten signatures are accepted to a greater level than any digital signature [1]. From these documents, most documents are structured documents. When the process of data entry is automated significant cost savings can be realized. Most of these forms have the same format with a separate box for each character as illustrated in fig. 1.



STUDENT NAME :

fig. 1 - Template Based Document Sample

The major problems with handling paper-based document are time spent re-keying data [2], Storage, inability to monitor progress and lost paperwork. In this survey, it was found that the average cost of capturing data from these documents was $2.84 and 25% of the respondents have reported more than $5 per a form.

These problems can be addressed with a solution that can digitize the data in the paperwork. It will reduce the cost of handling paper forms and had keying data. It will save time on both organizational and customers' side, as the data will be ready for further processes.

The proposed approach will automatically identify the template of structured documents and handwritten data despite not all forms having the same structure. This approach will increase the flexibility of handwritten recognition engines and provide greater value to organizations which in turn will increase the productivity and performance of business processes.

II. Background

A numeric system with 10 numerals and the Latin alphabet with 26 signs are used in the English language. Styles of written letterforms may vary between people and typically vary between regions. There are two major forms of handwritten numerals.

- In "In-line" form, all numerals are written with the same height
- In "old style", numerals have different heights.

As an example, there are many variations in the numeral 1. This can resemble the shape of lowercase l. Most people write this numeral as a straight-line top to bottom. Some write this with a serif at the top. This stroke is extended nearly to the baseline by people in some parts of Europe. The shape of this can appear like numeral 7. There is a version of this numeral that has serif at both top and bottom. This can be misread as the uppercase I. Explained variations of numeral 1 and possible variations of numeral 4 can be seen in fig. 2.

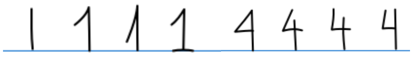


fig. 2 - variations of numerals

There are many alphanumeric characters that happen to have difficulties due to similarities of the appearance [3].

A. *Handwritten Character Recognition*

Handwritten character recognition (HCR) is a process of identifying, segmenting and recognizing characters from an input image. HCR can be divided into two categories as online and offline handwritten character recognition. Online recognition process takes place while the character is under creation. During this process, the character is recognized through the pressure, writing speed, the number of pen strokes and the shape of writing. When compared to online recognition process, offline recognition process has a lower accuracy level. The number of pixels per inch in the image used in online recognition process as raw data while real-time contextual information used in offline recognition [4]. This can be completed with different approaches.

Template matching is a common approach that output is processed by matching the input character or word image against the stored template. This approach will fail if the patterns are distorted, in the change of viewpoint or high variations among the patterns.

Syntactic Techniques are based on the recursive information of complex pattern regarding simpler patterns, by the shape of the object. The syntactic approach requires a huge amount of possibilities to be investigated, a considerable amount of data to be trained and immense computational effort.

978-1-5386-2444-9/17/$31.00 © 2017 IEEE