**Indian Institute of Information Technology Allahabad**

**Department of Information Technology**



**Mini Project Course 2021**

**Progress Report**

**on**

**Handwritten and Machine Printed Text**

**Detection**

As Presentation of C1 by

| **Enrollment No.** | **Name** | **GitHub ID.** |
| --- | --- | --- |
| **IIT2019239** | **Mrityunjaya Tripathi** | **Error404m** |
| **IIT2019222** | **Rauank Singh Rathore** | **Error404r** |
| **IIB2019006** | **Amanjeet Kumar** | **Amanjeetk11** |
| **IIT2019202** | **Jyoti Verma** | **Jyo123-verma** |

1. **Abstract:**

in a way to archive proper documents or books, starting from sets of short documents with annotations . It is necessary to separate the machine printed and handwritten text before applying different recognition methodologies to each. Because a document image having both types of texts (machine printed and handwritten text) may give rise to significant issues within a digitisation and recognition pipeline. In this project we will strive to figure out and isolate handwritten text from machine printed text using the Bag of Visual Words paradigm (BoVW).

1. **Introduction:**

The presence of printed and handwritten text in the same document image gives rise to significant issues since each modality requires different treatment to recognize the corresponding characters

In this project work we will see a method based on the Bag of Visual Words paradigm is presented for the separation of the machine printed and handwritten text. It mainly consists of three stages:

first one is The Page Segmentation stage in which it will detects blocks of interest on the document image, second one is the Block Descriptor Extraction stage, which calculates the descriptors of the extracted blocks using the BoVW model and the third one is Block Descriptor Extraction stage which characterizes the blocks by a Support Vector Machine classifier. as handwritten, machine printed or noise.

1. **Literature survey:** Nowadays, We can observe a rapidly growing number of digitization initiatives in libraries and archives, involving a variety of document types. Among several other obstacles, the presence of printed and handwritten text in the same document image gives rise to significant issues since each modality requires different treatment to recognize the corresponding characters.

The elements of handwriting recognition interface typically include:

* a pen or stylus for the user to write with.
* a touch sensitive surface, which may be integrated with, or adjacent to, an output display.
* a software application which interprets the movements of the stylus across the writing surface, translating the resulting strokes into digital text.

The process of handwriting recognition can be broken down into a few general steps:

* preprocessing,
* feature extraction and
* classification

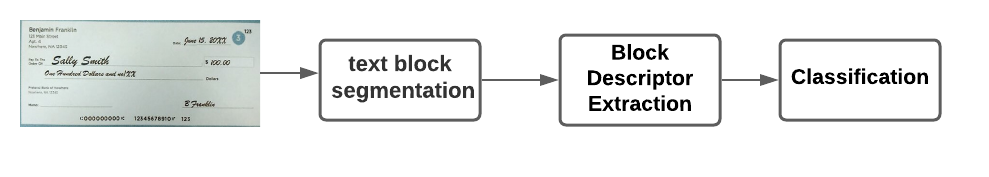
1. **The proposed approach :** The BoVW model is inspired by the Bag of Words (BoW) model employed in information retrieval in which a document is described by a set of words. Accordingly, the BoVW model for document images comprises a set of visual words to describe the image content.

A visual word is expressed by a set of features that correspond to local image

information of the image pixels which is identified by the image keypoints.These features are grouped in a number of clusters.

In particular, each local point belongs to a visual word which corresponds to the closest center of the cluster calculated by a distance function such as Euclidean, Manhattan, etc. Finally, the image is represented by a vector which denotes the corresponding descriptor and it reflects the frequency of each visual word that illustrates the BoVW paradigm.

The proposed model for the separation of machine printed from handwritten text has these main stages :



1. **Text Block Segmentation**: The objective of this stage is to detect blocks of interest in the document image. During this stage a number of challenges have to be addressed which could be grouped in four categories:

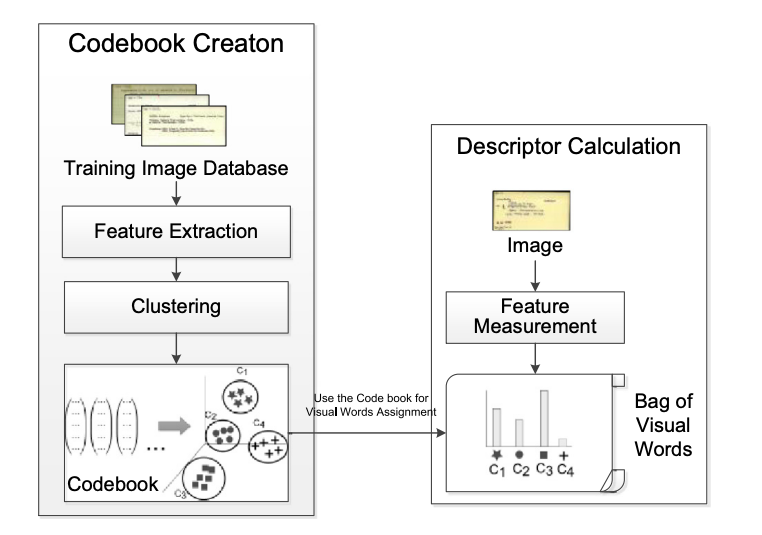
* Binarization problems which stem from different writing instruments
* Overlapping of handwritten and machine printed text
* Overlapping of text with noise
* Combinations of the above categories in which overlapping text (handwritten/machined printed) coexist with noise in different strokes.

1. **Block Descriptor Extraction:** In this stage, the descriptor is calculated based on the BoVW model . Then, a weight is applied based on the statistics of the datasets.

The local features are calculated on the greyscale version of the original document image. Finally, those key points whose corresponding position in the binary image does not match the foreground pixel are rejected.

1. **Classification:** Basically this is the final stage in which the system will classify the block whether it contains handwritten or machine printed text or neither of the both that means simply a noise. In this final stage, a classifier decides if the visual word vector of the block contains handwritten or machine printed text or neither of the above (noise). The proposed approach is based on the Support Vector Machines .

The SVMs are based on statistical learning theory and have been applied to a large number of different classification problems. They are chosen based on their power and their ability that do not require large training sets.



1. **Activity Schedule**

| **Date** | **Work Done** |
| --- | --- |
| 24/08/2021 | Read survey papers about Handwritten and Machine Printed Text Detection |
| 30/09/2021 | Studied about few models and selected one. |
| 10/09/2021 | Understood the research paper discussed earlier. |

1. **Work done so far**

* Selected a specific approach to solve the given problem of Handwritten and Machine Printed Text Detection.
* Selected a dataset for the given problem.
* Understood the proposed model.

1. **Future work**

* Implementing the proposed solution.
* Train on the dataset
* Test the models with different datasets.
* Host on live server (like Heroku, AWS, azuse, Google cloud etc…)

1. **Conclusion:**

In this project, We did using a method based on the Bag of Visual Words paradigm for the separation of the machine printed and handwritten text. It is a generic approach which can deal with document images which originate from datasets that are situated into different machine-printed/handwritten separation context.A text identification system project, able to discriminate between machine-printed and handwritten text-lines. The proposed solution can handle document pages, identifying text areas and splitting each area into text-lines.

1. **References:**

* <https://en.wikipedia.org/wiki/Handwriting_recognition>
* [Notes on the History of Pen-based Computing (YouTube)](https://www.youtube.com/watch?v=4xnqKdWMa_8)
* D. C. Ciresan, U. Meier, J. Schmidhuber. Multi-column Deep Neural Networks for Image Classification. IEEE Conf. on Computer Vision and Pattern Recognition CVPR 2012.
* Y.Zheng, H.Li, D.Doermann, “Text identification in Noisy Document Images Using Markov Random Field”, Proc of 7th ICDAR, 2003, pp.599-603.
* K.C.Fan, L.S.Wang and Y.T.Tu, “Classification of machine-printed and handwritten texts using character block layout variance”, Pattern Recognition, 31(9), 1998, pp.1275-1284.
* V.Pal and B.B.Chaudhuri, “Machine-printed and handwritten text lines identification”, Pattern Recognition Letters, 22, 2001, pp.431-441.
* U. Marti and H. Bunke, “A full English sentence database for off-line handwriting recognition”. Proc. 5th Int. Conference on Document Analysis and Recognition, ICDAR'99, 1999, pp. 705 – 708.
* E.Kavallieratou, N.Fakotakis, and G.Kokkinakis, “Un Off-line Unconstrained Handwriting Recognition System”, International Journal of Document Analysis and Recognition, no 4, 2002, pp. 226-242.