Project Proposal

Matching handwritten document images using CNN

Submitted by:

Anjana K	Shivani Naik	S Maneesha	Vaishali Jain
MT2016011	MT2016088	MT2016119	MT2016149

Title:

Matching handwritten document images to determine similarity score using CNN

Abstract:

Our aim is to calculate similarity between documents written by different individuals. We will assign a similarity score for a pair of documents based on various factors. This score would be decided irrespective of factors like word form variation, order in which words appear in different documents, format of the document and paraphrasing. We will be using word spotting which is a known technique to retrieve text from images. The project would use CNN (Convolutional Neural Network) for feature extraction and prediction. Applications of this include detecting plagiarism and automatic scoring of answer sheets given the answer key.

End Goal:

Given a set of digitized handwritten documents, we would predict how similar the two handwritten documents are.

Dataset:

IIIT-HWS Data Set is a synthetic handwritten dataset of 1 million word images which are very similar to natural handwriting. It is formed from 750 handwritten fonts. We are planning to use a subset of this data set.

Proposed Plan Of Execution:

The project can be divided into the following phases:

- Document segmentation which is dividing the document into its constituent words.
- Training CNN to extract relevant features for each word class.
- Computing similarity between a pair of documents using the word distribution of each document, which will be predicted using the trained CNN.

Main Challenges:

- Segmenting the document into words is a challenging task because of variable placement of page elements, skewed lines and irregular spacing between different words. We need to detect all the words from the document with relatively high efficiency.
- Writing styles of individual and page constraints may result in word overflow which poses a problem during segmentation.
- Presence of stopwords (eg. the, is, or) in the document contribute to the noise as they don’t have significance in calculating similarity and have to be removed accordingly.
- Paraphrasing is a common technique to hide the fact that a document has been copied. This involves changing the order of words such that meaning remains the same but it doesn’t come out as an exact copy of the sentence in other document.

Learning objective:

- Understanding the concepts of Convolutional Neural Networks and their application to various problems.
- Using Image Processing techniques such as segmentation, for dividing the image into constituent components.