STROKE IDENTIFICATION FROM TEXT

Project Supervisor : Prof. S. Sanyal

Group Members
Kaustubh Shamshery (IIT2014147)
Prateek Agarwal (IIT2014068)
Pratik Mangalore (IIT2014112)
Sulyab T V (IIT2014125)

INTRODUCTION AND MOTIVATION

Recognition of printed characters and handwriting is a subfield of prime importance in computer vision and artificial intelligence domains. Digitalization of information for safe storage and efficient manipulation has become a necessity in this era of information explosion.

Character recognition has been implemented very efficiently using many techniques in the past few decades. However, in this project, we explore a new method - identifying the strokes required to generate scripts. Here, strokes mean the basic curves which add up to produce whole letters. Identification of strokes is important as it gives and takes from how the human brain recognizes objects. The human brain uses information of orientation and spatial relations to identify objects. Using this method to identify characters might, in turn, provide more insight into the intricate ways of brain functioning.

PROBLEM SPECIFICATION

Given a dataset consisting of grayscale images of characters in a script, the strokes that completely generate the script are to be identified using unsupervised learning approach.

The probability that a particular pixel of a particular character is black, can be expressed using the probability distribution of strokes as

$$P(x_i|c_j) = \sum_{k \in S} P(s_k|c_j) P(x_i|s_k) \quad (1)$$

where x_i is the state of the i^{th} pixel (black/white), c_j is the j^{th} letter in the script, s_k is the k^{th} stroke that has been identified, and S is the set of all strokes.

THE APPROACH - PLSA

The expression (1) contains latent (hidden) variables in its RHS, and hence cannot be determined directly. Hence, we adopted the PLSA (Probabilistic latent semantic analysis) model, which is a statistical technique for automated document indexing. In PLSA, the probability of each variable is modelled as expressions involving conditionally independent multinomial distributions.

PLSA MODEL

The PLSA model consists of three elements -

1. A set of documents, $D = \{d_1, d_2, \ldots, d_N\}$
2. A set of classes/concepts/topics, $Z = \{z_1, z_2, \ldots, z_K\}$
3. A set of words, $W = \{w_1, w_2, \ldots, w_M\}$

The joint probability distribution is then defined as (2)

$$P(w,d) = P(d)\sum_z P(z|d)P(w|z) = \sum_z P(z)P(d|z)P(w|z)$$

The above expression can be obtained directly from (1). Thus, our problem can be fit into PLSA model and solved by taking D as the set of sample images, Z as the set of strokes and W as the set of pixels.

DOCUMENT-TERM MATRIX

Document-term matrices describe the frequency of words that occur in a set of documents. Documents are represented in rows and words are represented in columns.

For our problem, the Document-Term Matrix would look as follows -

	Pixel 1	Pixel 2	...	Pixel M
Image 1				
Image 2				
...				
Image N				

As the images are in grayscale mode with 1-byte depth, each of the cells would contain an integer in the range 0-255.

LEARNING THE PARAMETERS

The parameters involving latent variable z are learned by maximizing the log-likelihood of data. The likelihood function to be maximized here is

$$L = \prod_{i=1}^{N} \prod_{j=1}^{M} P(d_i, w_j)^{n(d_i, w_j)} \tag{3}$$

where $n(d_i,w_j)$ is the number of occurrences of j^{th} word in i^{th} document. The log-likelihood is obtained as

$$\log L = \sum_{i=1}^{N} n(d_i) \left[\log P(d_i) + \sum_{j=1}^{M} \frac{n(d_i, w_j)}{n(d_i)} \log \sum_{k=1}^{K} P(w_j|z_k) P(z_k|d_i) \right] \tag{4}$$

where $n(d_i)$ is the number of words in d_i

LEARNING THE PARAMETERS - EM ALGORITHM

The log-likelihood function of PLSA given in (4) is maximized using Expectation-Maximization (EM) algorithm to learn the parameters.

The EM algorithm is as follows. It has a time complexity of thismuch.

Initialization

Initialize the parameters p(z|d) and p(w|z) randomly

repeat until convergence :

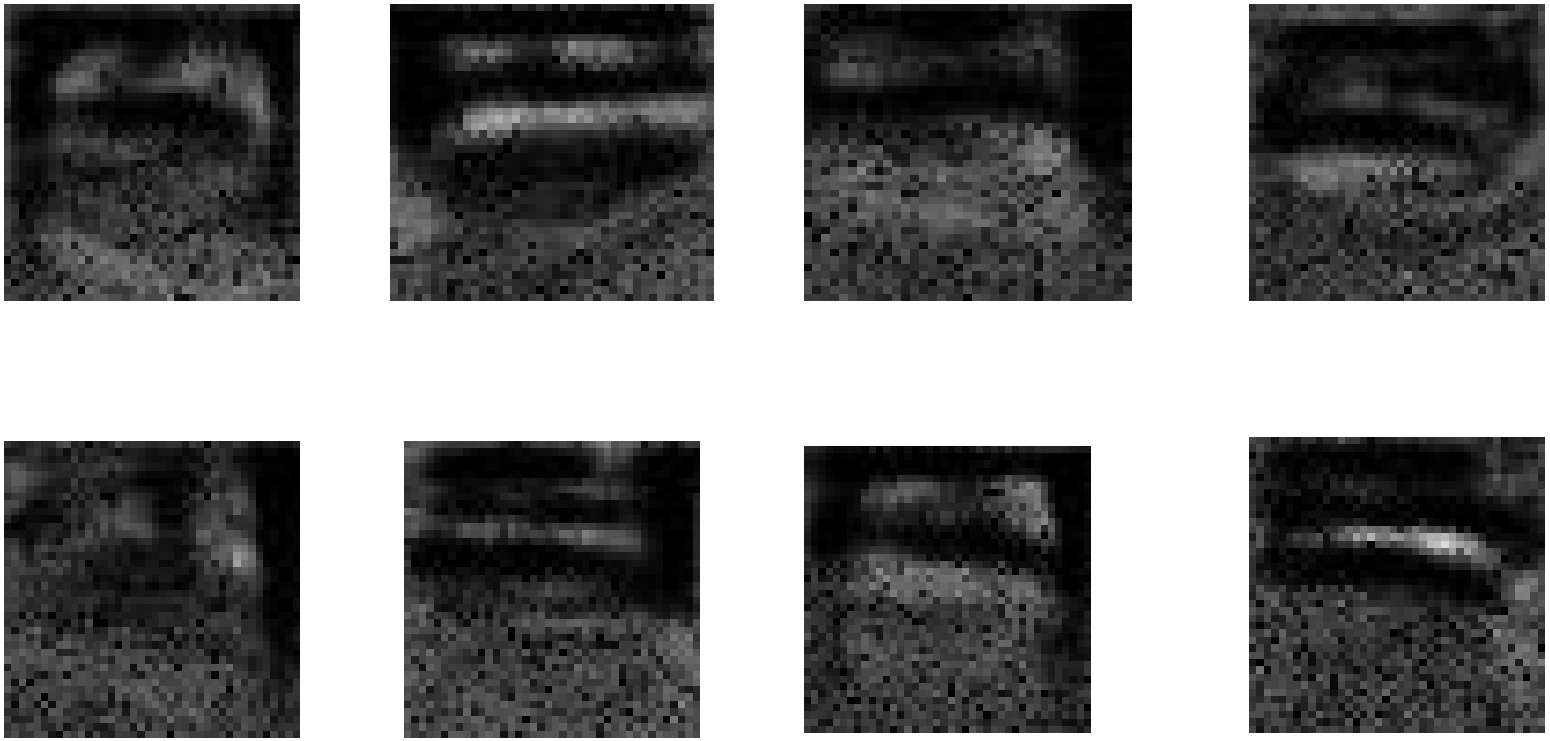
E-step : $$P\left(z_k|d_i,\, w_j\right) = \frac{P\left(w_j|z_k\right)P\left(z_k|d_i\right)}{\sum_{l=1}^{K} P\left(w_j|z_l\right)P\left(z_l|d_i\right)}$$

M-step : $$P\left(w_j|z_k\right) = \frac{\sum_{i=1}^{N} n(d_i, w_j)P\left(z_k|d_i,\, w_j\right)}{\sum_{m=1}^{M}\sum_{i=1}^{N} n(d_i, w_m)P\left(z_k|d_i,\, w_m\right)}$$

$$P\left(z_k|d_i\right) = \frac{\sum_{i=1}^{M} n(d_i, w_j)P\left(z_k|d_i,\, w_j\right)}{n(d_i)}$$

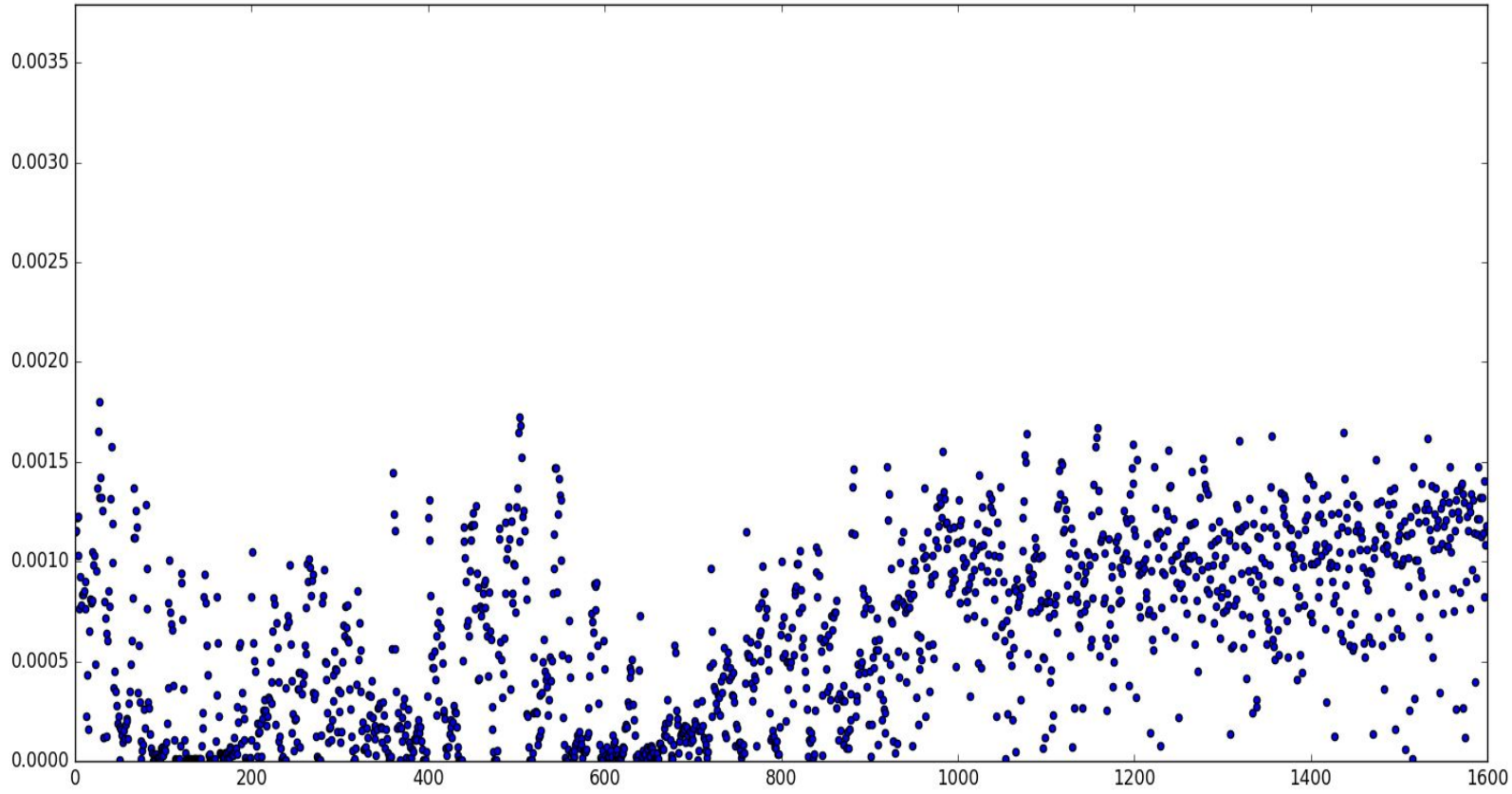
PROGRESS

EM Algorithm was run on 2517 samples with the assumption that the script has a total of 8 strokes.



Progress

Probability Plots of a Stroke



FUTURE GOALS

- Decompose the ‘strokes’ obtained into simpler, basic strokes without changing the unsupervised nature of the algorithm
- Optimize the algorithm and reduce its running time
- Introduce modifications to make sure EM does not get stuck at local maxima
- Devise a method to extract the minimum number of strokes required to completely and accurately generate a given script, while taking care of overfitting issues
- Refining the stroke Image.

REFERENCES

[1] T. Hofmann, “Probabilistic latent semantic analysis,” in Proceedings of the fifteenth conference on uncertainty in artificial intelligence, 1999, pp. 289-296.

[2] T. Hofmann, “Unsupervised learning by probabilistic latent semantic analysis,” Machine learning, vol. 42, iss. 1-2, pp. 177-196, 2001.

[3] DataEra.org, “Notes on EM and pLSA,” http://dataera.org/2014/04/notes-on-em-and-plsa/

THANK YOU