DIP Project

Clustering based Automatic Document Separation

Team Name: Mehtaab Singh

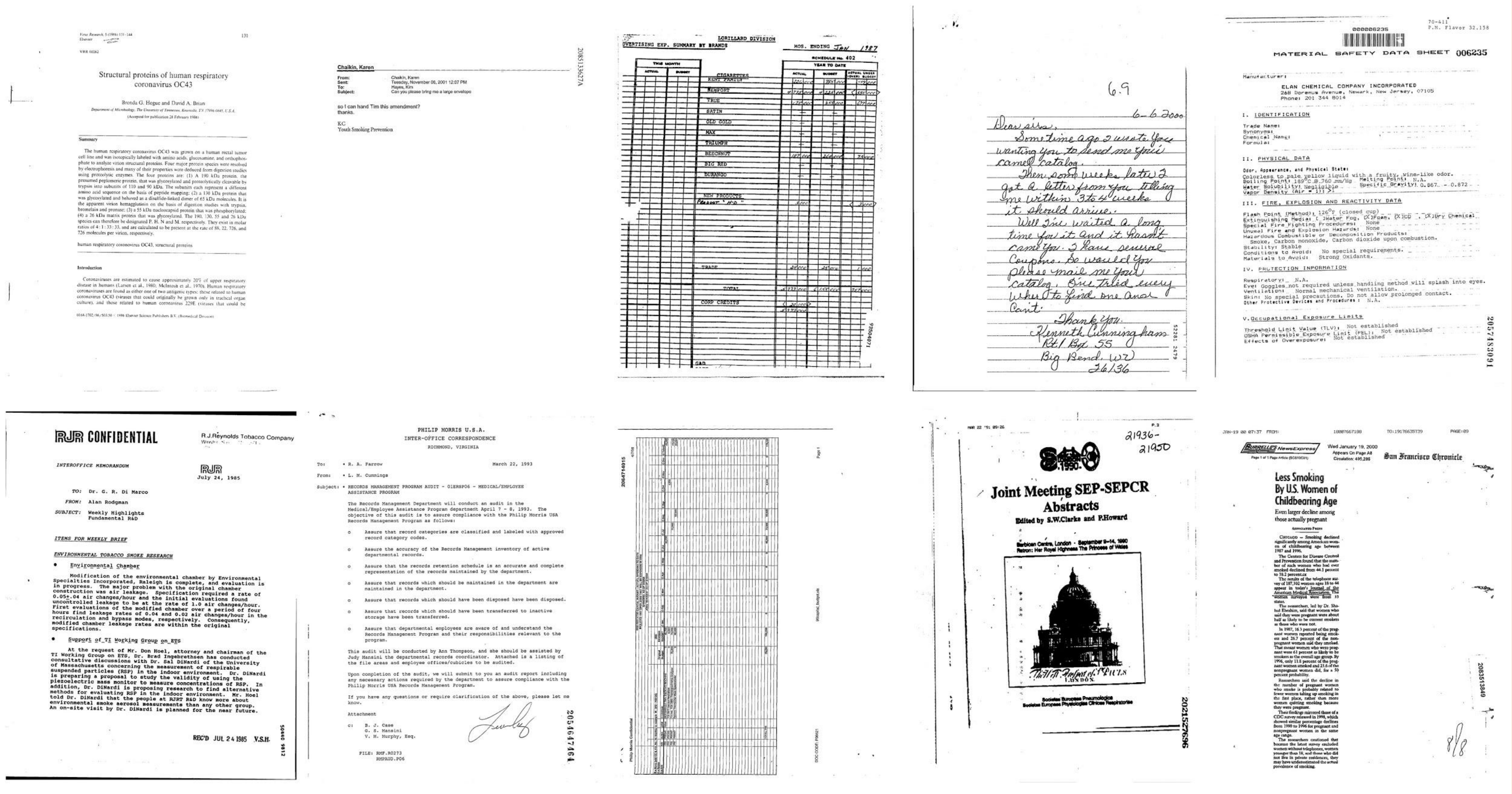
Mentor TA - Abhishek Prusty

Mehtaab Singh - 20171034 - Btech. ECE
Amrit Preet Singh - 2019909001 - Btech. Ext.

Repo URL-https://github.com/TheIndianCoder/A-Clustering-Based-Algorithm-for-Automatic-Document-Separation.git

Problem Statement

Given a set of document images, find their relative similarity.



Practical applications:-

- First step in Document Image Processing (DIP) tasks:
  - document retrieval,
  - information extraction and text recognition,
- Web search
- Paperless office
- Enhanced Database indexing & retrieval

Pipeline of project

- Week 1: Problem definition, related work exploration.
- Week 2,3: Dataset exploration, Algorithmic scrutiny.
- Week 3,4: Textual Similarity & Average Word Height
- Week 4,5: Average Character Width & Average Word spacing
- Week 5,6: Average Line spacing and Classifier selection.
- Week 7: Classifier implementation and Training.
- Week 8: General Code improvements and testing.

Algorithm:-

- Step 1:

Find the textual similarity between two document pages.

- How is this achieved?

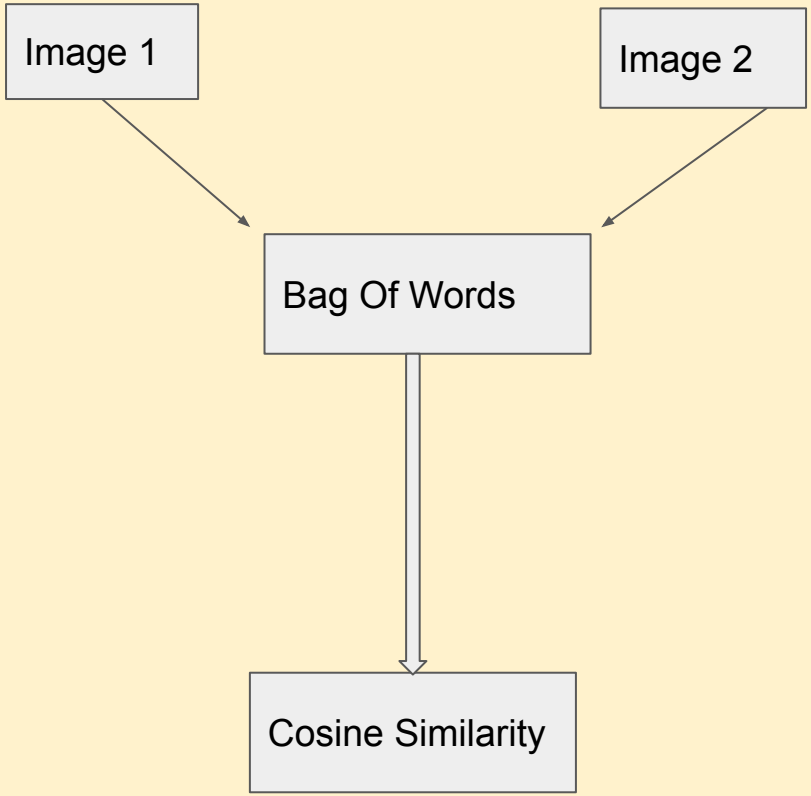
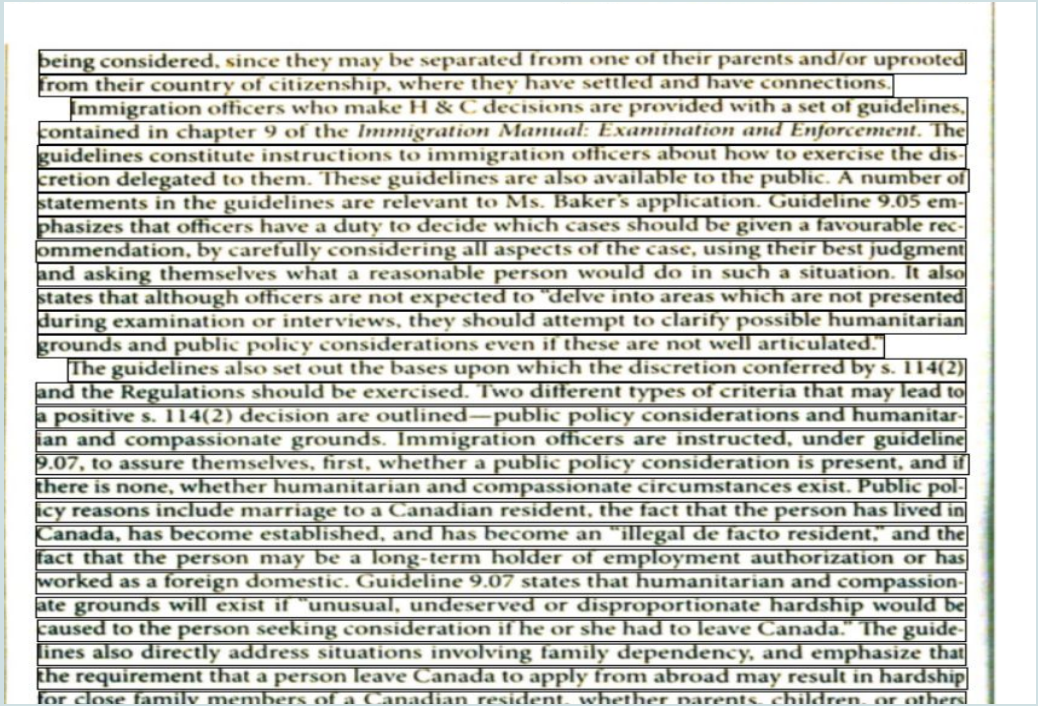


Image 1
Image 2
Bag Of Words
Cosine Similarity

OCR

1) Line Finding
2) Baseline Fitting
3) Fixed Pitch Detection
4) Proportional Word Finding
5) Word Recognition
6) Chopping joined characters and associating broken characters

Baseline fitting



being considered, since they may be separated from one of their parents and/or uprooted from their country of citizenship, where they have settled and have connections.

Immigration officers who make H & C decisions are provided with a set of guidelines, contained in chapter 9 of the *Immigration Manual: Examination and Enforcement*. The guidelines constitute instructions to immigration officers about how to exercise the discretion delegated to them. These guidelines are also available to the public. A number of statements in the guidelines are relevant to Ms. Baker's application. Guideline 9.05 emphasizes that officers have a duty to decide which cases should be given a favourable recommendation, by carefully considering all aspects of the case, using their best judgment and asking themselves what a reasonable person would do in such a situation. It also states that although officers are not expected to "delve into areas which are not presented during examination or interviews, they should attempt to clarify possible humanitarian grounds and public policy considerations even if these are not well articulated."

The guidelines also set out the bases upon which the discretion conferred by s. 114(2) and the Regulations should be exercised. Two different types of criteria that may lead to a positive s. 114(2) decision are outlined—public policy considerations and humanitarian and compassionate grounds. Immigration officers are instructed, under guideline 9.07, to assure themselves, first, whether a public policy consideration is present, and if there is none, whether humanitarian and compassionate circumstances exist. Public policy reasons include marriage to a Canadian resident, the fact that the person has lived in Canada, has become established, and has become an "illegal de facto resident," and the fact that the person may be a long-term holder of employment authorization or has worked as a foreign domestic. Guideline 9.07 states that humanitarian and compassionate grounds will exist if "unusual, undeserved or disproportionate hardship would be caused to the person seeking consideration if he or she had to leave Canada." The guidelines also directly address situations involving family dependency, and emphasize that the requirement that a person leave Canada to apply from abroad may result in hardship for close family members of a Canadian resident, whether parents, children, or others

Algorithm Page_Seg_Line_Det ()

STEP 1: Initialize DLC as 0.

STEP 2: Determine horizontal profile of D_I.

a. Loop if a peak appears then there is a long run of black pixels.

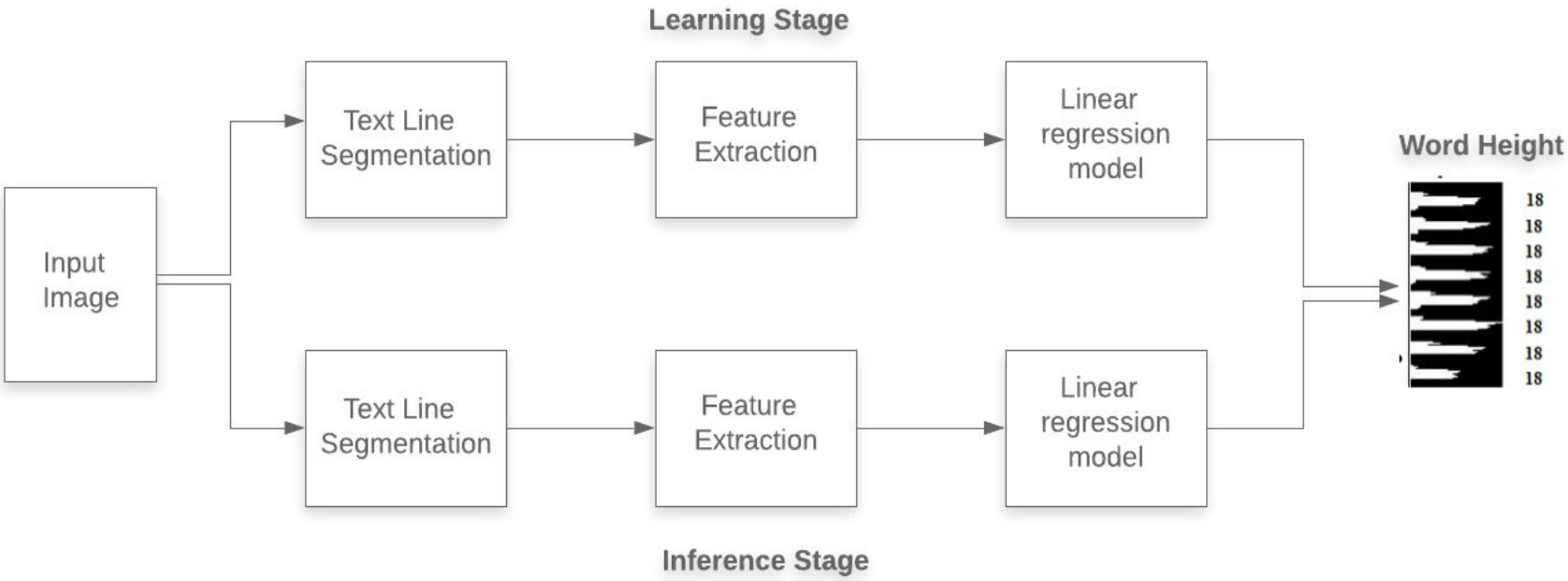
b. DLC = DLC + 1.

c. Else there is a long run of white pixels.

d. Continue STEP 2 until no peak remains uncovered.

Shalini Puri et. al. 2015

Word height detection



- Line Spacing Estimates
  - Difference of coordinates of the lines in baseline fitting step.

- Character Width Estimates
  - Chop words along the detected line on a fixed pitch

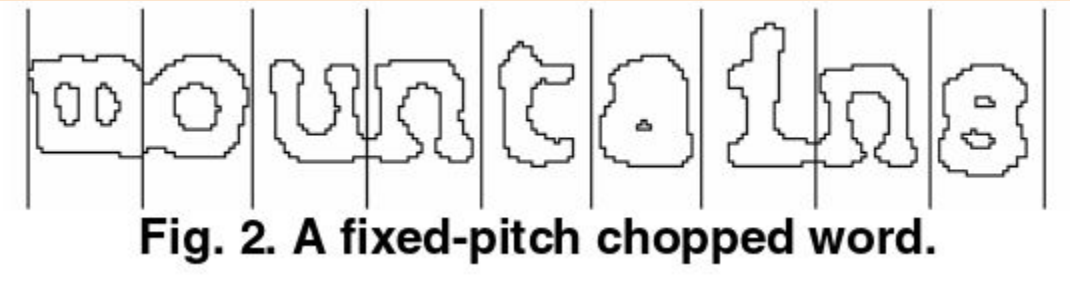


Fig. 2. A fixed-pitch chopped word.

Ray Smith, Google Inc.

- Word spacing detection
  - Difference of pixel coordinates of consecutive words.

Features Analysis

- Textual features:
  - bag of words approach
- Layout features:
  - Character Width.
  - Word height,
  - Line spacing,
  - Word Spacing.
- Why not take Color?
- Headers, footers, page numbers, text to white ratio, Intersection of structural rectangles etc.

Extensions:

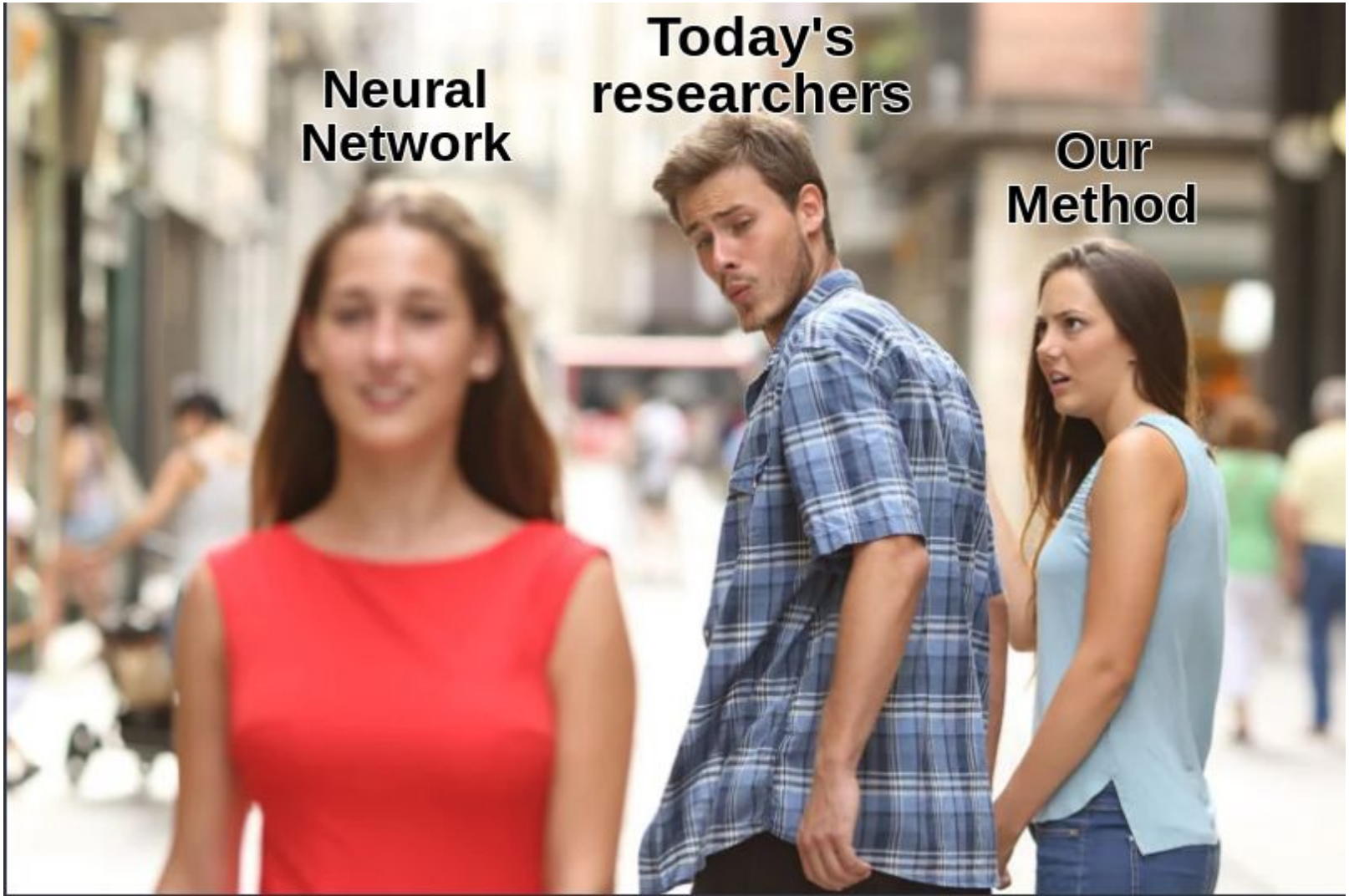
- Neural network representation.
- Enhanced OCR.
- Weighted features.

Limitations

1) English language model.
2) Hand written documents.
3) Equal priority to features

Division of Work

- Mehtaab Singh:
  - Line Finding
  - Baseline Fitting
  - Bag of Words
- Amrit Preet Singh
  - Character Width.
  - Word height,
  - Line spacing,
  - Word Spacing.
- Literature review, algorithmic srutiny, code integration done collaboratively.


Neural Network
Today's researchers
Our Method

Queries are welcome!