ICDAR 2015 competition on text line detection in historical documents
(2015)

Murdock, Michael Reid, Shawn Hamilton, Blaine Reese, Jackson

December 3, 2018

Abstract

Describing the ANDAR-TL competition (ANncestry Document Analysis and Recognition Text Lines), the used documents are labeld with the coordinate location of the first character of the left-most word on each line and the objective is to use the training data to train a system to detect the locations of text-line origin points in the images.

1 Introduction

Investigating and comparing general methods that can reliably and robustly identify the origin point for a text? line in the presence of various noise conditions, interfering annotations, and the artifacts common to historical documents

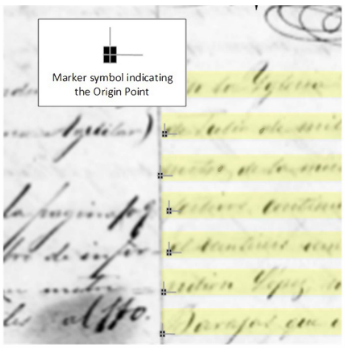


Figure 1: The start of a text-line, called an origin point is shown in the inset

2 Dataset

Accompanying each image in the trammg set is a metadata file (MDAT), which consists of an origin point of each text-line in the image. An origin point for a text-line is the (x, y) coordinate located at the intersection of the baseline of the first character of the first word in the line and the left-most edge of that character.

3 Scoring methodology

An image is processed by the participant's text line detector to produce a set of Estimate Points (EP), which are individually scored against ground truth Origin Points.

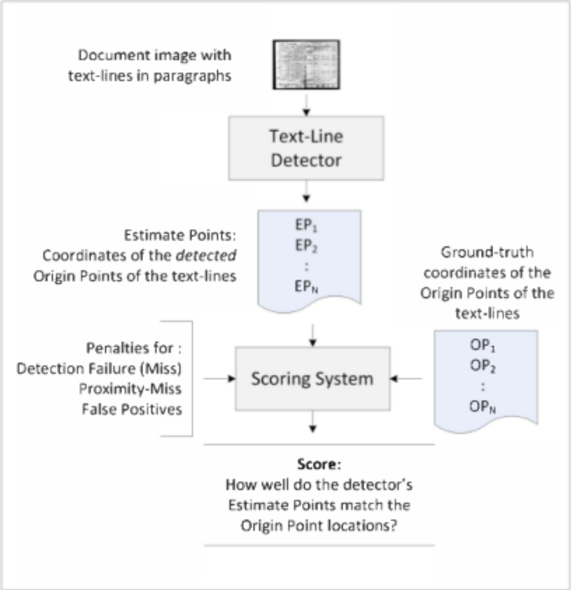

Document image with text-lines in paragraphs
Text-Line Detector
Estimate Points: Coordinates of the *detected* Origin Points of the text-lines
EP$_1$
EP$_2$
:
EP$_N$
Ground-truth coordinates of the Origin Points of the text-lines
OP$_1$
OP$_2$
:
OP$_N$
Penalties for : Detection Failure (Miss) Proximity-Miss False Positives
Scoring System
Score:
How well do the detector's Estimate Points match the Origin Point locations?

Figure 2: The start of a text-line, called an origin point is shown in the inset

The precise location of the origin point can be quite subjective and even individual operators are quite inconsistent in the way they determined the location of the origin points. Therefore, in an attempt to eliminate the challenges related to scoring and comparing systems in the presence of noisy ground truth, a prediction is scored as a hit if it is within a square region around the origin point. This first tolerance region is referred to as R_1 and an EP within R_1, is assigned a penalty of zero.
An EP within R_2 is assigned a penalty proportional to its distance from the origin point, with a max value clamped at the penalty of a miss, which was set to twenty. And a False-Positive is an estimate that is inside R_1 or R_2, but it's an extra EP. Since

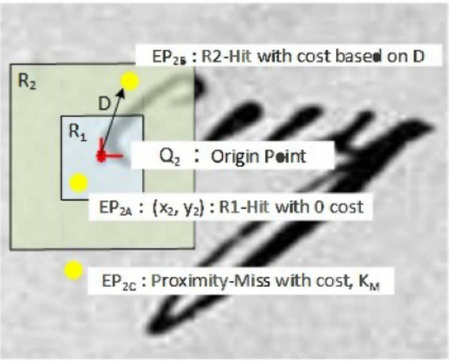

EP$_{2B}$: R2-Hit with cost based on D
R$_2$
D
R$_1$
Q$_2$: Origin Point
EP$_{2A}$: (x$_2$, y$_2$) : R1-Hit with 0 cost
EP$_{2C}$: Proximity-Miss with cost, K$_M$

Figure 3: An example of tolerance region around an OP Q2

4 Overview of the systems

4.1 A2iA System I

based on a convolutional neural network that takes normalized pixel values as raw inputs and directly outputs the positions of the origin points, along with a confidence score that is used to decide whether or not the origin point is really here (the confidence score is similar to a posterior probability).
In the network architecture: LSTM layer was used along with convolution, a "sliding window" approach is used, and a non linear function (sigmoid) is applied on the outputs of the network that

correspond to positions.

The "sliding window" approach consists in dividing each document image into frames of size 362x362 pixels, with a step of 180 pixels between consecutive frames. Each 362x362 frame can output a fixed number of OPs (which was arbitrarily chosen to be 20). Each one of the 20 outputs per frame is a 3-uple: (confidence score, horizontal position, vertical position). For each output, the confidence score is compared to a threshold (0.5) to decide whether or not there is a new detected OP, and its position is directly given by the network. No additional post-processing was used.

4.2 A2iA System 2

Based on a text line detector that takes a processed image as input. The image pre-processing algorithm is composed of four main steps:
(1) Global and local contrast enhancement (2) Skew correction (3) A dedicated process to delete left and right margins with vertical lines detection, and (4) An adaptive binary threshold (sliding window algorithm) is coupled with a filtering to delete noisy structures.

4.3 A2iA System 3

A combination of the first and second systems. The OPs predicted by the ConvNN/LSTM system are matched to the OPs predicted by the image processing approach system, using the Munkres matching algorithm that minimizes the sum of the Euclidian distance between OPs. For the match where the OPs are less than 150 pixels distant, this system keeps the location predicted by the RectFiltering system.

Side note: Maximum Bipartite Matching A matching in a Bipartite Graph is a set of the edges chosen in such a way that no two edges share an endpoint. A maximum matching is a matching of maximum size (maximum number of edges). In a maximum matching, if any edge is added to it, it is no longer a matching. There can be more than one maximum matchings for a given Bipartite Graph.

4.4 Institute of Automation, CASIA System I

Proposes a novel approach to segment text lines and identify origin points in historical handwritten documents. Firstly, the input image is binarized by an adaptive thresholding method. Then, the non-textual regions are removed by a simple yet efficient shape filter. The binary image text skew is corrected using the Hough transform. Once deskewed, text lines are segmented via horizontal projection. Finally, a multiple vertical projection strategy is proposed to identify the text line origin points.

4.5 Institute of Automation, CASIA System 2

This method employs basic operations of image processing: radon transform, image rotation, edge detection, and some calculations on density values, without using any machine learning techniques. Empirical results show that the proposed method can achieve good performance under the estimation protocol.

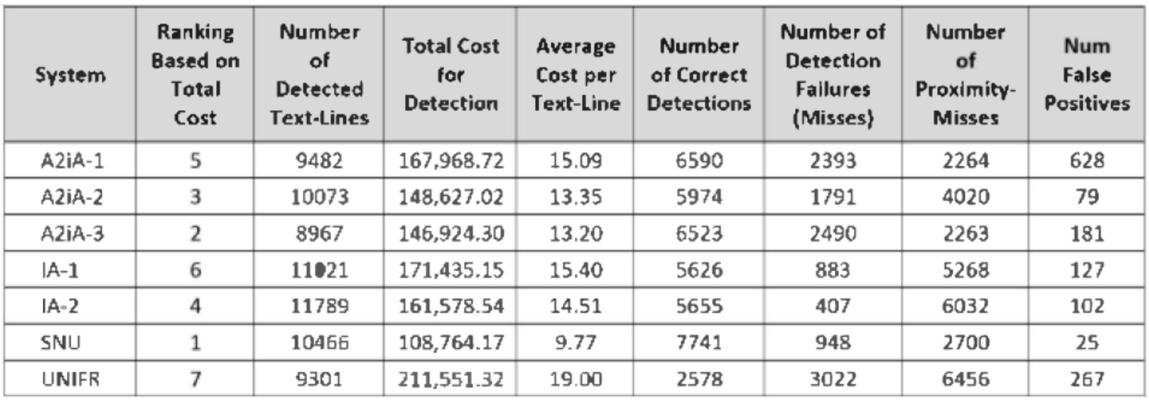
4.6 Seoul National University

This algorithm is composed of two parts: 1) historical document binarization and 2) text line detection.
It normalizes the image using background estimation to compensate the background variation and combine global thresholding result and local thresholding result to achieve high performance. This is futher improved upon by processing the black background region and adding post-processing step. As a result, the algrithm can effectively reject page boundary and artifacts which disturb a correct analysis.

4.7 University of Fribourg

This approach is based on an unsupervised feature learning method. Features are learned automatically from unlabeled pixels. These features are used to train an SVM to extract the text lines. In addition to the simplicity and context independence, this approach requires little prior knowledge. The algorithm learns feature vectors for image patches with a 3-level hierarchy of convolutional autoencoders (CAE).

5 Results



System	Ranking Based on Total Cost	Number of Detected Text-Lines	Total Cost for Detection	Average Cost per Text-Line	Number of Correct Detections	Number of Detection Failures (Misses)	Number of Proximity-Misses	Num False Positives
A2iA-1	5	9482	167,968.72	15.09	6590	2393	2264	628
A2iA-2	3	10073	148,627.02	13.35	5974	1791	4020	79
A2iA-3	2	8967	146,924.30	13.20	6523	2490	2263	181
IA-1	6	11021	171,435.15	15.40	5626	883	5268	127
IA-2	4	11789	161,578.54	14.51	5655	407	6032	102
SNU	1	10466	108,764.17	9.77	7741	948	2700	25
UNIFR	7	9301	211,551.32	19.00	2578	3022	6456	267

Figure 4: An example of tolerance region around an OP Q2

References

[1] Murdock, Michael Reid, Shawn Hamilton, Blaine Reese, Jackson, *ICDAR 2015 competition on text line detection in historical documents*