

14th INTERNATIONAL CONFERENCE ON FRONTIERS IN HANDWRITING RECOGNITION

1-4 September

ICFHR 2014

Crete Island - Greece

A NOVEL TRANSCRIPT MAPPING TECHNIQUE FOR HANDWRITTEN DOCUMENT IMAGES

Nikolaos Stamatopoulos, Georgios Louloudis and Basilis Gatos



DEMOKRITOS
NATIONAL CENTER FOR SCIENTIFIC RESEARCH

tranScriptorium

Outline

- Introduction
- Proposed Methodology
 - Local Approach
 - Global Approach
 - Combination
- Experimental Results
- Conclusions

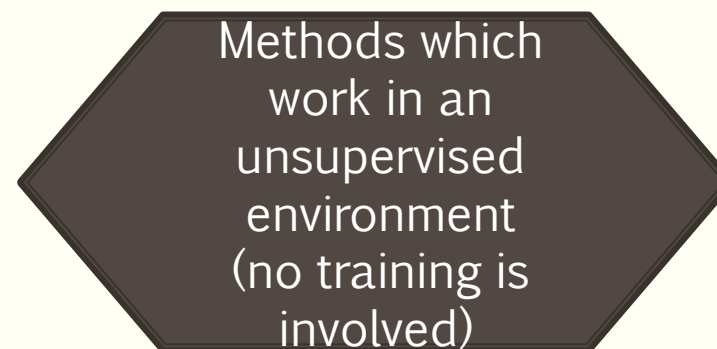
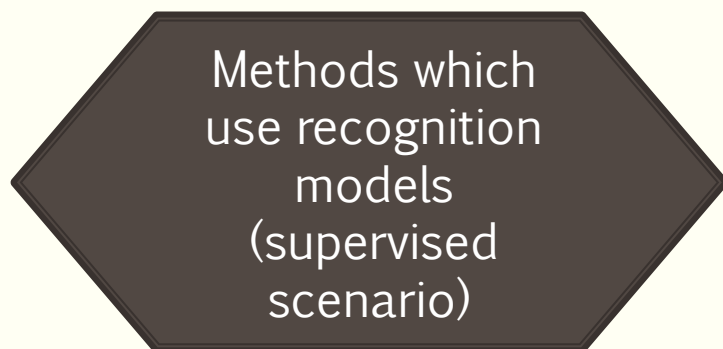
Transcript Mapping

- Align the correct text information to a segmentation result produced automatically.
- A minimum user involvement for the correction of segmentation errors is necessary.
- Fast generation of benchmarking/training datasets.



Transcript Mapping

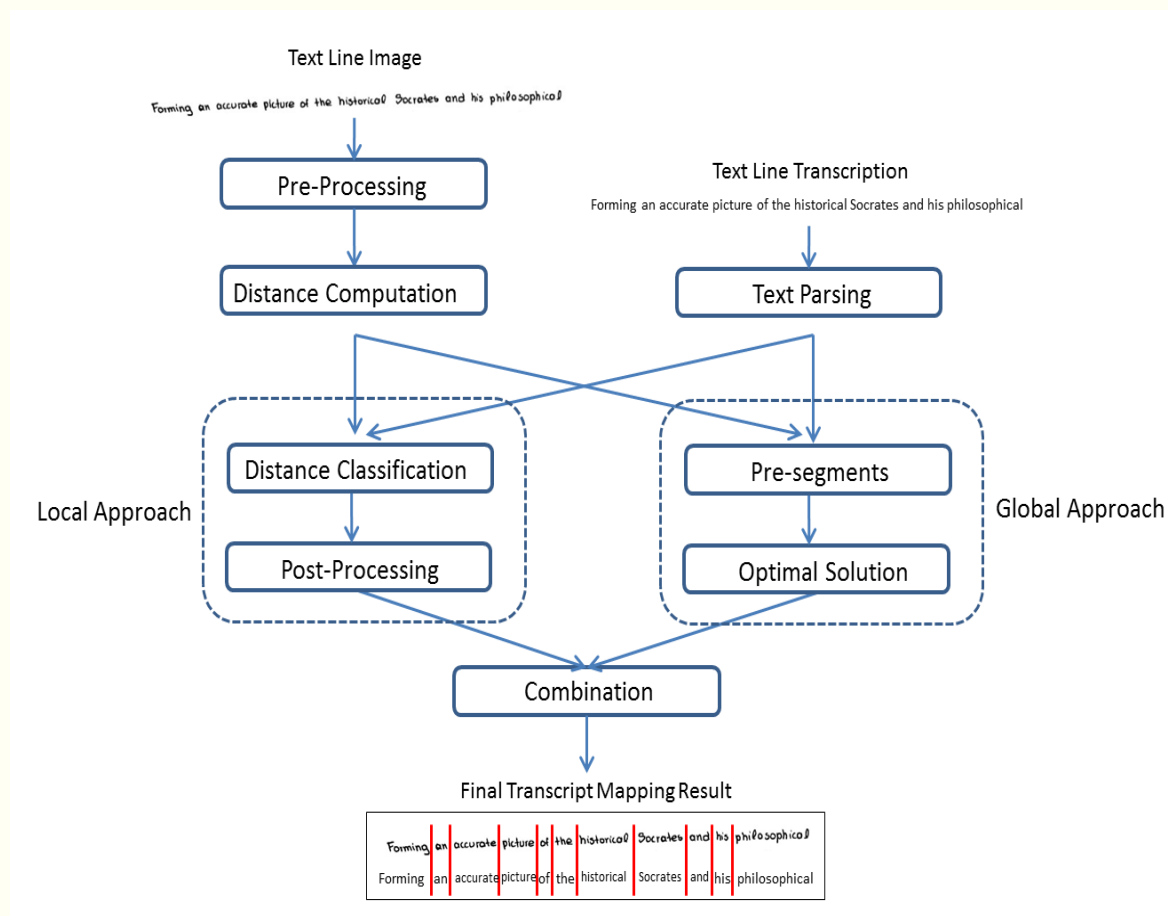
- Transcript mapping techniques can be classified into two main categories according to the algorithm which is used for the alignment.



- Supervised methods lead to high performance but have the disadvantage of needing a training phase which makes necessary the existence of annotated data beforehand.

Proposed Methodology

- Guided by the number of words as well as the characters per word of a text line.
- Combines the results of a local and a global approach using a scoring algorithm.
- Local Approach: A modification of our previous method [1].
- Global Approach: The optimal segmentation result among several segmentation hypotheses is produced by minimizing a suitable cost function.



Proposed Methodology

Text Parsing

- Transcription contains useful information which can be used in order to correctly segment a document image into words.

Image

Socrates was a classical Greek philosopher.

Transcription

Socrates was a Classical Greek philosopher.

Number of words

NW=6

Number of characters

Socrates

$NC_1 = 8$

was

$NC_2 = 3$

a

$NC_3 = 1$

Classical

$NC_4 = 9$

Greek

$NC_5 = 5$

philosopher.

$NC_6 = 12$

Proposed Methodology

Pre-processing

ανθρώπους διότι δεν γνωρίζουν τι κάνουν με το να τον

Original Text Line

ανθρώπους διότι δεν γνωρίζουν τι κάνουν με το να τον

Skew Correction

ανθρώπους διότι δεν γνωρίζουν τι κάνουν με το να τον

Slant Correction

Proposed Methodology

Distance Computation

- Calculate the distance of adjacent overlapped components (OC) in the text line image.
 - An OC is defined as a set of connected components whose projection profiles overlap in the vertical direction.

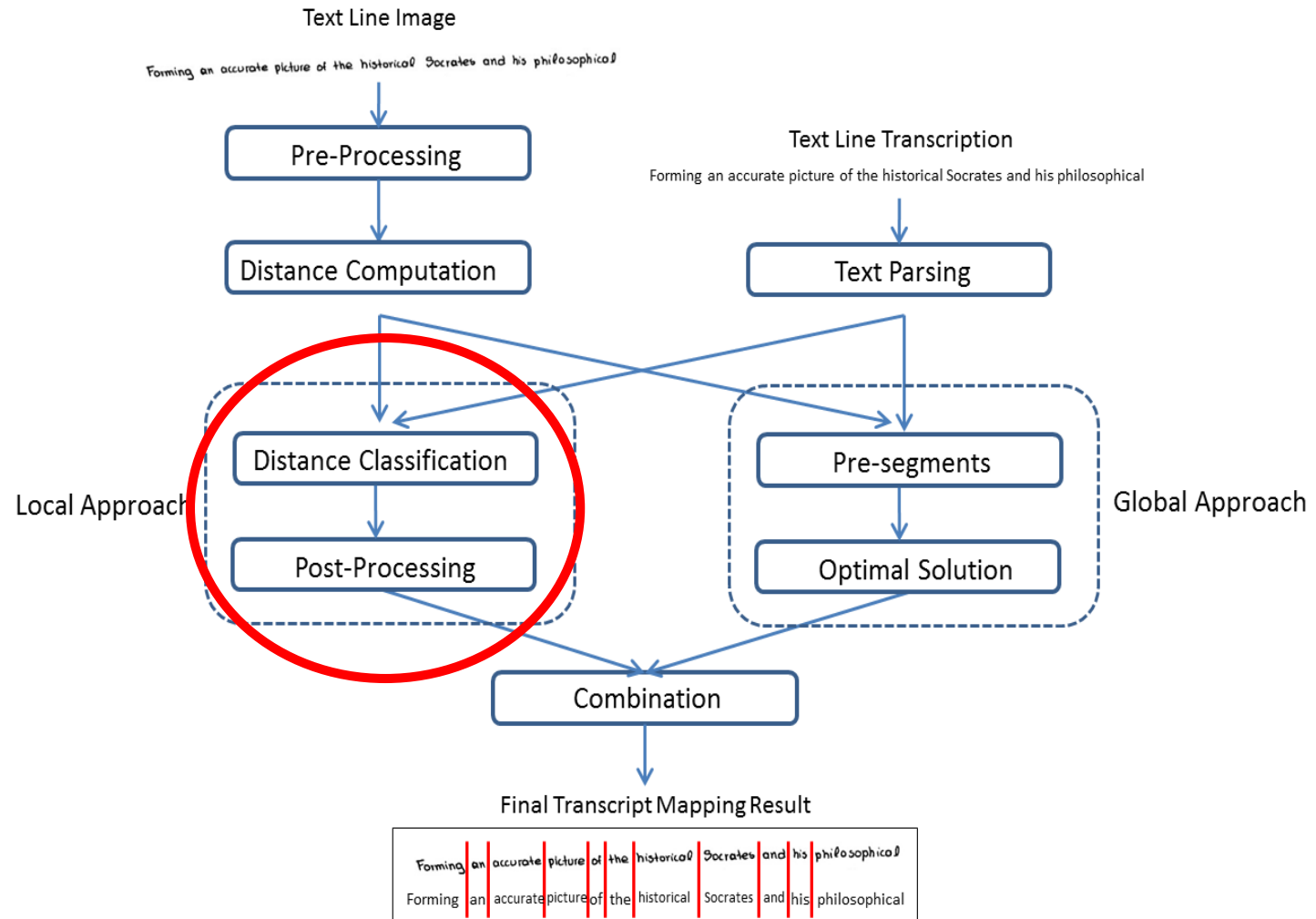


Connected Components: 10

Overlapped Components: 3

- Distance: The minimum Euclidean distance among the Euclidean distances of all pairs of points of the two adjacent overlapped components.

Proposed Methodology – Local Approach



Proposed Methodology – Local Approach

Distance Classification

- Classify the distances as inter-word distances or intra-word distances.
 - Use a local threshold for every text line.
 - Select as threshold the largest distance which produces equal or larger number of words from the actual number of words NW .

to sarrire from ambiguity
 d_1 d_2 d_3

$NW=4$

$d_1 > d_2 > d_3$

to sarrire from ambiguity

Threshold = d_1

to sarrire from ambiguity

Threshold = d_2

to sarrire from ambiguity

Threshold = d_3



Proposed Methodology – Local Approach

Post-processing

- Split or merge a detected word when its width deviates from a statistical estimation based on the number of the characters of the word.

Average character width: $AW = \frac{\sum_{j=1}^{ND} W_j}{\sum_{l=1}^{NW} NC_l}$

—————> Width of detected word (pixels)
—————> Number of characters (transcription)

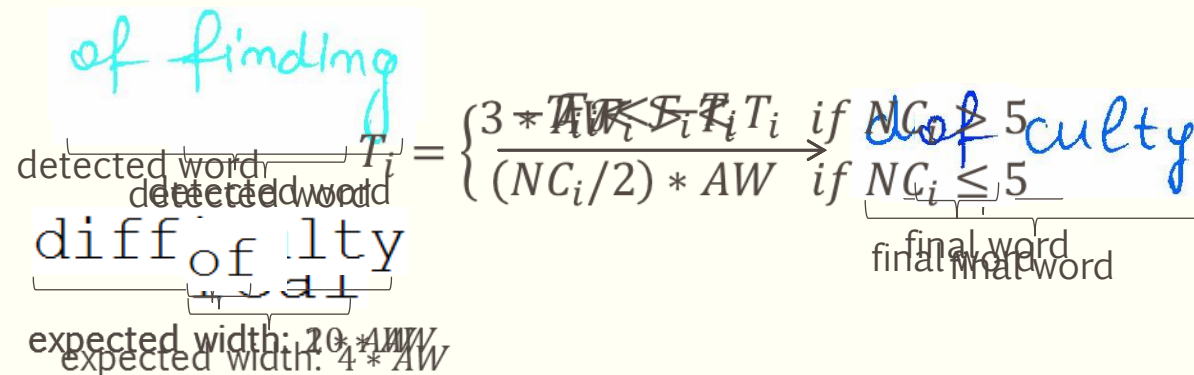
Cost Function of word i : $\mathcal{F}_i = \underbrace{(NC_i * AW)}_{\text{expected width}} - \underbrace{W_i}_{\text{width of detected word}}$

Proposed Methodology – Local Approach

Post-processing

- Split or merge a detected word when its width deviates from a statistical estimation based on the number of the characters of the word.

- (1) $-T_i < \mathcal{F}_i < T_i$ The word i has been detected correctly.
- (2) $\mathcal{F}_i > T_i$ The word i has to be merged with the following detected word.
- (3) $\mathcal{F}_i < -T_i$ The word i has to be split.



Proposed Methodology – Local Approach

The difficulty of finding the real

(a)

The difficulty of finding the real

$$F_1 = -1.72 \quad F_2 = 196.25$$

$$T_1 = 51.63 \quad T_2 = 103.26$$

The difficulty of finding the real

(b)

The difficulty of finding the real

$$F_1 = -1.72 \quad F_2 = 1.25 \quad F_3 = -276.15$$

$$T_1 = 51.63 \quad T_2 = 103.26 \quad T_3 = 34.42$$

The difficulty of finding the real

(c)

The difficulty of finding the real

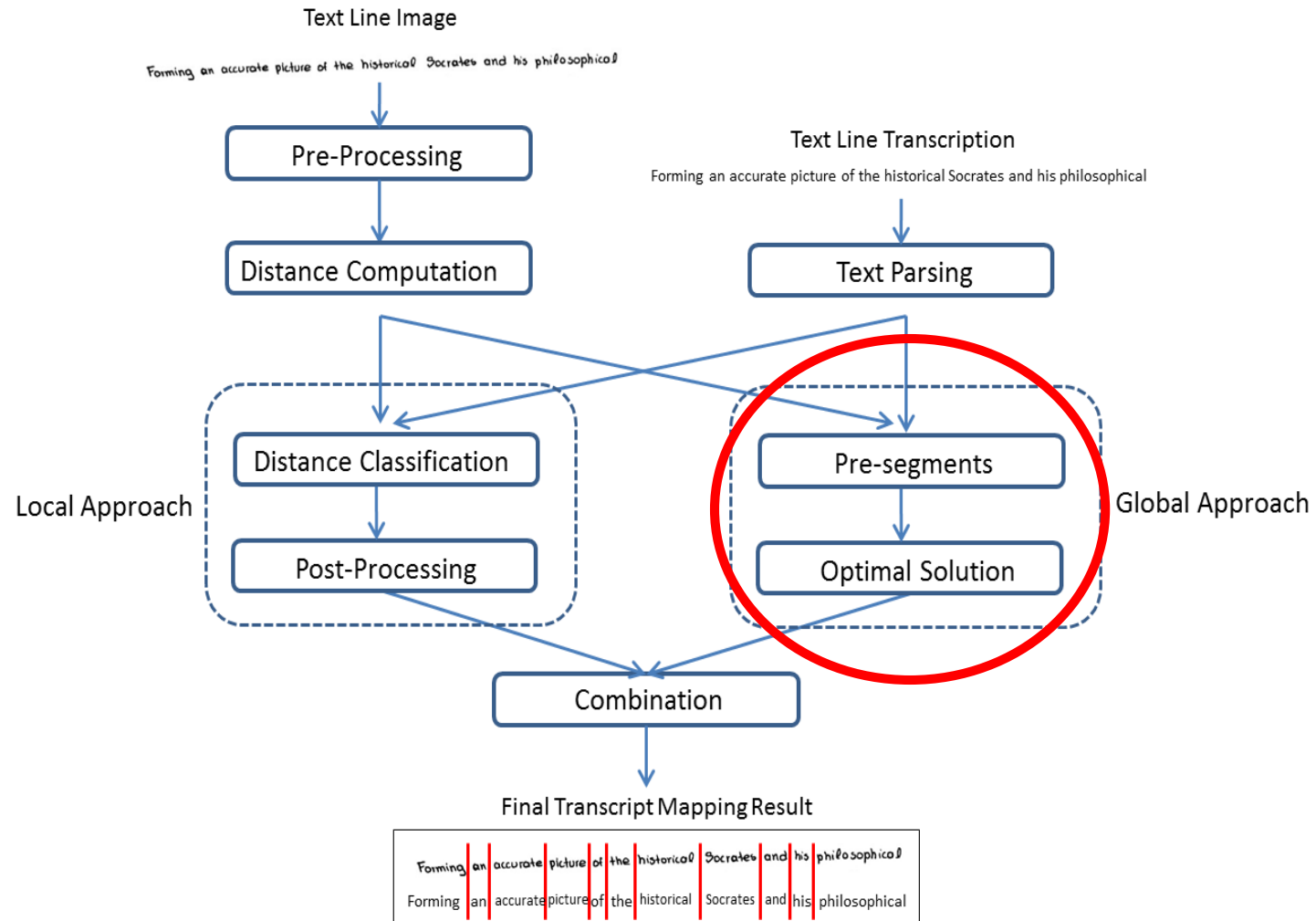
$$F_1 = -1.72 \quad F_2 = 1.25 \quad F_3 = 14.14 \quad F_4 = -2.02 \quad F_5 = -5.27 \quad F_6 = -5.70$$

$$T_1 = 51.63 \quad T_2 = 103.26 \quad T_3 = 34.42 \quad T_4 = 103.26 \quad T_5 = 51.63 \quad T_6 = 68.84$$

The difficulty of finding the real

(d)

Proposed Methodology – Global Approach



Proposed Methodology – Global Approach

Pre-segments

- Classify the distances as inter-word distances or intra-word distances.
 - Use a local threshold for every text line.
 - Select as threshold the largest distance which produces $NW + n$ words.
 - n is a parameter related to the desired over-segmentation flavor of the result.

Gesprächsführung und ihre philosophischen Inhalte sind

NW=6

Gesprächsführung und ihre philosophischen Inhalte sind

NW + 2 = 8

Proposed Methodology – Global Approach

Optimal Solution

- Produce several segmentation results by consecutively merging all neighboring pre-segments in order to have the desired number of words NW .
- Select the optimal segmentation result which minimizes the cost function.

$$C_k = \sum_{i=1}^{NW} \sum_{j=1}^{NW} \left| \frac{W_i}{W_j} - \frac{NC_i}{NC_j} \right|$$

W_i : the width of the i -th detected word (pixels)

NC_i : the number of characters of the i -th word (transcription)

The ratio of the widths of any detected words pair must be approximately equal to the ratio of the number of characters of the corresponding words pair only when the word segmentation result is correct.

Proposed Methodology – Global Approach

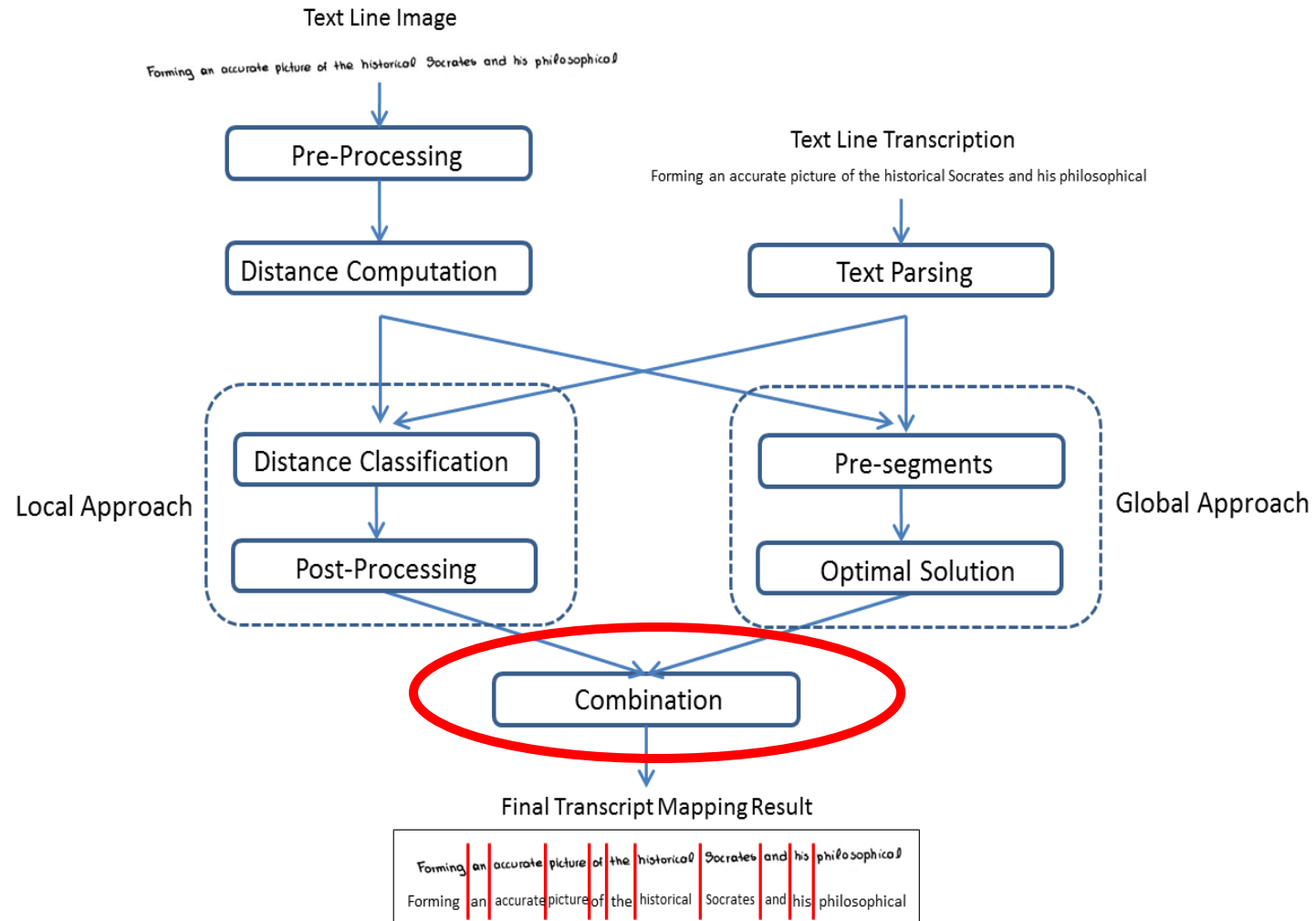
Optimal Solution

$NC_1=16$ $NC_2=3$ $NC_3=4$ $NC_4=15$ $NC_5=7$ $NC_6=4$
Gesprächsführung und ihre philosophischen Inhalte sind
Gesprächsführung und ihre philosophischen Inhalte sind NW=6

Gesprächsführung | und | ihre | philosophischen | Inhalte | sind NW + 2 = 8

Gesprächsführung | und | ihre | philosophischen | Inhalte | sind
← W_1 → ← W_2 → ← W_3 → ← W_4 → ← W_5 → ← W_6 →

Proposed Methodology – Combination



Proposed Methodology – Combination


- The final selection is made based on a scoring algorithm applied on both results.
- The segmentation result with the lowest score is considered as final.
- The scoring algorithm is based on the ranking of both text and image with respect to the word width (number of characters per word in the case of text and width per word in the case of image).

Proposed Methodology – Combination


- Text Ranking: The words are sorted in descending order with respect to their number of characters.

Text	Socrates	was	a	Classical	Greek	philosopher.	Credited	as	one	of	the
Characters	8	3	1	9	5	12	8	2	3	2	3
Ranking	3	6	11	2	5	1	3	9	6	9	6

- Image Ranking: The detected words are sorted in descending order with respect to their widths.

Image											
Width	208	92	32	225	134	268	204	54	80	44	77
Ranking	3	6	11	2	5	1	4	9	7	10	8

- Image Ranking Adjustment:

Image											
Width	208	92	32	225	134	268	204	54	80	44	77
Ranking	3	6	11	2	5	1	3	9	6	9	6

Proposed Methodology – Combination

- Score Calculation: Comparison of Text and Image Ranking

- Score is **zero** at the cases where the ranking values are the same as well as when two adjacent ranking values change position.

T[i]	3	6	11	2	5	1	3	9	6	9	6
Im[i]	3	6	11	2	5	1	3	9	6	9	6
Column_Score	0	0	0	0	0	0	0	0	0	0	0
S	0	0	0	0	0	0	0	0	0	0	0

Final Score = 0

T[i]	5	7	1	4	11	7	1	5	1	7	7
Im[i]	5	7	1	1	11	7	4	5	1	7	7
Column_Score	0	0	0	0	0	0	0	0	0	0	0
S	0	0	0	0	0	0	0	0	0	0	0

Final Score = 0

- Score is equal to the absolute difference of the ranking values

T[i]	7	1	3	7	2	5	5	3	7	7	1
Im[i]	7	1	3	7	7	5	3	5	2	7	1
Column_Score	0	0	0	0	5	0	0	0	5	0	0
S	0	0	0	0	5	5	5	5	10	10	10

Final Score = 10

Experimental Results

- The proposed methodology was evaluated on the test set of the ICDAR2009 Handwriting Segmentation Contest [1].
- The set consists of 200 handwritten document images written in several languages (English, French, German and Greek) that contain **29717** words.
- All handwritten document images come from several writers and they do not include any non-text elements (lines, drawings, etc.)

Sokrates war ein für das abendläändische Denken grundlegender griechischer Philosoph, der in Athen lebte und wirkte. Seine herausragende Bedeutung zeigt sich nun darin, dass alle griechischen Denker vor ihm als Vorsokratiker bezeichnet werden. Sokrates entwickelte die philosophische Methode eines sokratischen Dialogs, die er Mänetis nannte. Diese Kunst der Gesprächsführung und ihre philosophischen Inhalte sind nur indirekt überliefert worden, da Sokrates selber nichts Schriftliches hinterlassen hat. Mehrere seiner Schüler, der berühmteste unter ihnen Platon, haben sokratische Dialoge verfasst und unterschiedliche Züge seiner Lehre betont. Die unbewusste Haltung des Sokrates in dem gegen ihn wegen angeblich verderblichen Einflusses auf die Jugend und wegen Missachtung der Griechischen Götter geführten Prozess hat zu seinem Nachruhm wesentlich beigetragen. Das Todesurteil nahm er als glückliches Fehlurteil gelassen hin; bis zur Hinrichtung durch den Schierlingsbecher beschäftigten ihn und die zu Besuch im Gefängnis weilenden Freunde und Schüler philosophische Fragen.

Démocrite d'Abdère était un philosophe grec souvent classé parmi les Présocratiques du point de vue philosophique, bien qu'il soit un peu plus jeune que Socrate, et qu'il soit mort quelques trenté années après Socrate. Il est considéré comme un philosophe matérialiste en raison de sa conviction en un Univers constitué d'atomes et de vide, théorie atomiste. Pour Démocrite la nature est composée dans son ensemble de deux principes: les atomes et le vide. L'existence des atomes peut être déduite de ce principe: Rien ne vient du néant, et rien, après avoir été détruit, n'y retourne. Il y a ainsi toujours du plein, i.e. de l'être, et le non-être est le vide. Les atomes sont des corpuscules solides et indivisibles, séparés par des interstices vides, et dont la faible part qu'ils échappent à nos sens. Découverts comme lisses ou ruges, crochus, recourbés ou ronds, ils ne peuvent être affectés ou modifiés à cause de leur dureté.

Experimental Results

- The performance evaluation of segmentation is based on counting the number of the matches between the words detected and the ground truth.
- The performance was recorded in terms of detection rate (DR), recognition accuracy (RA) and F-Measure (FM).

$$DR = \frac{o2o}{N}$$

$$RA = \frac{o2o}{M}$$

$$FM = \frac{2 * DR * RA}{DR + RA}$$

$o2o$ – one-to-one matches , N count of ground-truth words, M count of result words

Experimental Results

Method	M	o2o	DR (%)	RA (%)	FM (%)
ICDAR2009 Winner	29962	28279	95.16	94.38	94.77
Previous Method [1]	29673	28845	97.06	97.21	97.13
Local Approach	29717	29370	98.83	98.83	98.83
Global Approach	29717	29499	99.26	99.26	99.26
Combination	29717	29563	99.48	99.48	99.48

[1] N. Stamatopoulos, G. Louloudis and B. Gatos, "Efficient Transcript Mapping to Ease the Creation of Document Image Segmentation Ground Truth with Text-Image Alignment", 12th International Conference on Frontiers in Handwriting Recognition, pp. 226-231, Kolkata, India, 2010.

Conclusions

- The proposed method fails to correctly detect only 154 words out of 29717.
- Only a very small number of segmentation results needs correction in order to produce the final word segmentation ground truth.
- Representative errors:



- Possible transcription errors affect the proposed method.

Questions

