HANDWRITTEN/PRINTED TEXT SEPARATION USING PSEUDO-LINES FOR CONTEXTUAL RE-LABELING

By: Ahmad Montaser Awal
Abdel Belaïd
Vincent Poulain d’Andecy
CONTEXT

- Administrative documents are
  - Noisy
  - Annotated...

- Separation of scripts in administrative documents
  - Annotation extraction
  - Sending each script to a specialized system
  - Noise removal
Chère Madame,

Vous avez récemment fait part de la création d'un observatoire que vous aimeriez développer.

Je la remercie de tous cœur pour ce nouveau projet et je souhaite lui apporter un soutien complet.

Bien cordialement,

de 13/14/08

Je vous remercie d'avoir annulé l'annuaire de 23€ et de m'avoir remis le chèque de remboursement du 31/10/08.

Ci-joint, CEF de 86,50€ pour mon abonnement de Nov. 08.

De vous recevoir les chèques au fur et à mesure de leur arrivée.

Si vous êtes raisonnable, je vous remercie d'avance. Je souhaite que vous me rendiez le service.

PQ. A CEF de 86,50€ d'octobre 08 au 27/10/08.

Sincères salutations,

[Signature]

P.S. A CEF de 86,50€ du 27/10/08 au 27/10/08.
STATE OF THE ART

- Printed/handwritten text separation systems share the main steps
  - Preprocessing
    Removing very small/large connected components
  - Document segmentation
    Segment the document into basic units
  - Classification
    Assign each unit to a text class
  - Contextual re-labeling
    Correct classification errors using neighborhood information
STATE OF THE ART
DOCUMENT SEGMENTATION

- Text line level (Pal et al. 2001)(Kavallieratou et al. 2004)
  - Lines are assumed to be homogeneous (mono-class)
  - Segmentation using the horizontal projection profiles

- Word level
  - Grouping connected components to approximate words
  - Distance based (Zheng et al. 2004) (Shetty et al. 2007)
  - Morphological operations (Peng et al. 2011) (Zagoris et al. 2014)

- Character level (Fan et al. 1998)
  - Non-cursive scripts (Chinese documents)
  - X-Y cut algorithm
STATE OF THE ART CONTEXTUAL RE-LABELING

Step 1: Define the neighborhood of a given word
- 4 Nearest Neighbors (Peng et al. 2013) (Zheng et al. 2007)
- 6 Nearest Neighbors (Shetty et al. 2007)

Step 2: Define criteria to re-label a word based on the labels of its neighborhood
- Majority voting (kandan et al. 2007)
- Probabilistic models
  - Markov Random Field (MRF) (Zheng et al. 2007) (Peng et al. 2013)
  - Conditional Random Field (CRF) (Shetty et al. 2007)
PROPOSED SYSTEM OVERVIEW

SEGMENTATION

- Differently from most of existing works, the document is first segmented into pseudo-lines before being segmented into pseudo-words

- Pseudo-line
  - A set of connected components where:
    - Horizontal distances < $d_H$
    - Vertical distances < $d_V$

- Pseudo-word
  - A set of connected components belonging to the same pseudo-line
  - Horizontal distance < $ws$ (word spacing distance estimated automatically for each pseudo-line)
**Improved Segmentation – Heuristic**

- Avoid vertical connection caused by handwritten annotations

- Use CCs horizontal overlapping

\[
o(c_1, c_2) = \frac{h_1 \cap h_2}{\max(h_1, h_2)}
\]

O = 0%  
O = 30%  
O = 50%  
O = 100%
IMPROVED SEGMENTATION – HEURISTIC

De plus agricole, capitale
Vous en toute d

Mlle A
CARMA
B.P. 28 166
**Pseudo-words classification**

- A pseudo-word is characterized by 137 features.
- A multiclass Support vector machines SVM is used to classify a pseudo-word into:
  - Handwritten text
  - Printed text
  - Noise
CONTEXTUAL RELABELING

- Some classification errors could be corrected using contextual neighborhood
- The label of each pseudo-word is updated based on those of its neighbors
- Local neighborhood
  - K nearest neighbors*
  - Confidence propagation *
  - Conditional Random Fields
- Using pseudo-lines
  - Probabilistic model (CRF)
  - Static model

The separation problem can be modeled by CRF.

According to (Nicolas et al. 2007), the probability of a pseudo-word $w$ is given by:

$$P(X_w | Y_L, Y_C) = \lambda_L f_L + \lambda_C f_C$$

**Contextual features**

- Local classification probabilities of left/right neighbors
- Structural features extracted from the pseudo-word and each neighbor
  - Height ratio
  - Position ratios
  - Density ratio
RE-LABELING USING PSEUDO-LINES

- Ideally, a pseudo-line represents a text line of the document.
- More than 90% of pseudo-lines contain one type of text (printed or handwritten).
- Pseudo-lines define, implicitly, a global horizontal neighborhood relation between the pseudo-words.
RE-LABELING USING PSEUDO-LINES

- The **dominant class** $C_D$ in a pseudo-line is the class with the highest cardinality.
- In case of equality of cardinalities, the dominant class is the one with highest average confidence of its pseudo-words.
- The label of a pseudo-word is updated:
  - Using a CRF model
  - If it verifies the following condition:

$$ (f_i < cf) \lor (|h_i - h_D| < d) $$

Classification Confidence

Certainty factor

Regularity factor
RE-LABELING USING PSEUDO-LINES

EXAMPLES

No Change

Handwritten  Printed  Noise

ICHFR 2014

A.M.AWAL, A. Belaïd and V.P.d'Andecy
EXPERIMENTATION

- **Evaluation**
  - **Pixel level**
    \[ \text{pixRate} = \frac{\text{pixels correctly recognised}}{\text{total number of pixels}} \]
  - **Pseudo-word level**
    \[ \text{pwRate} = \frac{\text{pseudo-words correctly recognized}}{\text{total number of pseudo-words}} \]

- **Documents**
  - **Training DB**
    - 107 documents (32706 pseudo-words)
      - H: 5888; P: 18078; N: 8740
  - **Test DB**
    - 202 documents (82142 pseudo-words)
      - H: 11970; P: 43705; N: 25190
  - All documents are labeled at the pixel level
## RESULTS (1/2)

<table>
<thead>
<tr>
<th>System</th>
<th>H%</th>
<th>P%</th>
<th>N%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Previously proposed system*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proposed system without contextual re-labeling</td>
<td>97.7</td>
<td>96.5</td>
<td>94.3</td>
</tr>
<tr>
<td>k-NN</td>
<td>95.5</td>
<td>97.5</td>
<td>92.3</td>
</tr>
<tr>
<td>Confidence propagation</td>
<td>97.8</td>
<td>96.6</td>
<td>94.0</td>
</tr>
<tr>
<td>New relabeling methods</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CRF</td>
<td>98.5</td>
<td>97.1</td>
<td>94.2</td>
</tr>
<tr>
<td>Pseudo-lines (CRF): Probabilistic</td>
<td>98.9</td>
<td>97.5</td>
<td>93.5</td>
</tr>
<tr>
<td>Pseudo-lines: Deterministic</td>
<td>98.3</td>
<td>99.2</td>
<td>87.9</td>
</tr>
<tr>
<td>Improved segmentation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pseudo-lines: Deterministic</td>
<td>99.1</td>
<td>99.2</td>
<td>90.1</td>
</tr>
</tbody>
</table>

## RESULTS (2/2)

<table>
<thead>
<tr>
<th>System</th>
<th>Docs</th>
<th>pwRate</th>
<th>pixRate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>H%</td>
<td>P%</td>
</tr>
<tr>
<td>[kandan et al. 2007]</td>
<td>150</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>[Zheng et al. 2004]</td>
<td>94</td>
<td>93.0</td>
<td>98.0</td>
</tr>
<tr>
<td>[Peng et al. 2013]</td>
<td>82</td>
<td>93.8</td>
<td>95.7</td>
</tr>
<tr>
<td>[Shetty et al. 2007]</td>
<td>27</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>[Hamrouni et al. 2014]</td>
<td>32</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Proposed system</td>
<td>202</td>
<td>97.3</td>
<td>99.5</td>
</tr>
</tbody>
</table>
CONCLUSION AND PERSPECTIVES

- Distance based segmentation is not always enough to obtain ‘good’ pseudo-words
  - Heuristics could improve and solve some segmentation problems
- A better performance using pseudo-line based contextual relabeling
- A very good performance compared to the state of the art systems
- In future work:
  - Feature selection
  - Ambiguity layer
Thank you