THE 16TH INTERNATIONAL CONFERENCE ON FRONTIERS IN HANDWRITING RECOGNITION

CHARACTER AND TEXT RECOGNITION OF KHMER HISTORICAL PALM LEAF MANUSCRIPTS

Dona Valy, Michel Verleysen, Sophea Chhun, and Jean-Christophe Burie
Overview

- Khmer Palm Leaf Manuscripts
- Task 1: Isolated Character Classification
- Task 2: Word/Text Recognition
- Conclusion
KHMER PALM LEAF MANUSCRIPTS
Introduction

- Palm Leaf Manuscripts or Sleuk Rith in Khmer
  - [Sleuk: leaf] + [Rith: to bind/tie together]
Challenges

- Degradations and defects

![Degradations and defects images]
Challenges

- Ambiguity of certain characters
  - Khmer alphabet (more or less 70 symbols)
  - Similarity between characters
Challenges

- Sequential order of characters composing a word
  - Khmer alphabet (more or less 70 symbols)
  - Irregularity of how characters are combined into words
SleukRith Set

- A collection of annotated data created from 657 pages of digitized Khmer palm leaf manuscripts
- Composed of 3 types of annotated data:
  - Character/Glyph
  - Word
  - Line

Available at [https://github.com/donavaly/SleukRith-Set](https://github.com/donavaly/SleukRith-Set)
Statistics of SleukRith Set

<table>
<thead>
<tr>
<th>Data</th>
<th>Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annotated Characters/Glyphs</td>
<td>301,626</td>
</tr>
<tr>
<td>Annotated Words</td>
<td>73,359</td>
</tr>
<tr>
<td>Text Lines</td>
<td>3,245</td>
</tr>
</tbody>
</table>

Character and word image patches

Available at [https://github.com/donavaly/SleukRith-Set](https://github.com/donavaly/SleukRith-Set)
TASK 1: ISOLATED CHARACTER CLASSIFICATION

-system
\[ c_1 : p_1 \]
\[ c_2 : p_2 \]
\[ \ldots \]
\[ c_n : p_n \]
Isolated Character Dataset

Data normalization

(a). Original image, (b). Gray scaled and resized to 48x48, (c). Normalized

Dataset:
- Train: ~113k
- Test: ~91k
- Number of classes: 111
Network 1.1: CNN
Network 1.2: Column LSTM
Network 1.3: Row-Column LSTM
Network 1.4: CNN-LSTM
Experiments and Results

- **Training configurations:**
  - Batch size: 300
  - Samples are reshuffled after each epoch
  - Stop condition:
    - average loss does not improve after $N = 10$ consecutive tests
    - each test is done for every 50 iterations

- **Results: top-k error rate**

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Error Rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Top 5</td>
</tr>
<tr>
<td>Network 1.1: CNN</td>
<td>0.65</td>
</tr>
<tr>
<td>Network 1.2: Column LSTM</td>
<td>1.05</td>
</tr>
<tr>
<td>Network 1.3: Row-Column LSTM</td>
<td>0.82</td>
</tr>
<tr>
<td><strong>Network 1.4: Conv-LSTM</strong></td>
<td><strong>0.46</strong></td>
</tr>
</tbody>
</table>
TASK2: WORD/TEXT RECOGNITION
Annotated Word Dataset

- **Character-Class Map**
  - $I_w, n_{col}$
  - $c_w$, $c_h$:
    - Cell height and width
  - $n_{row} = I_h/c_h$, $n_{col} = I_w/c_w$

- **Dataset**:
  - **Train**: $\sim16k$
  - **Test**: $\sim8k$
  - **Number of character-classes**: 134
    - Including 1 token class for background or blank space
General Architecture
Network 2.1: 1D-LSTM

- LSTM Layer of Network 2.1
Network 2.2: 2D-LSTM

- LSTM Layer of Network 2.2
Experiments

- Training configurations:
  - Batch size: 30
  - Samples are sorted and batched according to their width
    - (a). Initial sample order
    - (b). Sort by the width of each sample
    - (c). Pad each sample to the maximum width in the batch
    - (d). Shuffle batch order

- Stop condition:
  - average loss does not improve after $N = 30$ consecutive tests
  - each test is done for every 50 iterations
Results

- **Measurement**
  - Top-k error rate: average error rate of all cells in the predicted character-class map

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Error Rate (%)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Top 5</td>
<td>Top 1</td>
</tr>
<tr>
<td>Network 2.1: 1D-LSTM</td>
<td>8.46</td>
<td>32.01</td>
</tr>
<tr>
<td><strong>Network 2.2: 2D-LSTM</strong></td>
<td><strong>2.40</strong></td>
<td><strong>20.49</strong></td>
</tr>
</tbody>
</table>

(a). Original word image  
(b). Ground truth character-class map  
(c). Result predicted by Network 2.1  
(d). Result predicted by Network 2.2
CONCLUSION
Conclusion

- We present different approaches for two tasks on medium size datasets constructed from Khmer palm leaf manuscripts:
  - Isolated character classification
  - Word/text recognition
- The predicted character-class map from Task 2 can be used further to generate the final transcription of the word image
  - CTC and/or encoder-decoder mechanism
Thank you for your attention!


