

LODENet: A Holistic Approach to Offline Handwritten Chinese and Japanese Text Line Recognition

Huu-Tin Hoang*, Chun-Jen Peng*, Hung Vinh Tran, Hung Le
Cinnamon AI Labs
Minato, Tokyo 105-0001, Japan
Email: {tin, larry, xing, toni}@cinnamon.is

Huy Hoang Nguyen
University of Oulu
Oulu, Finland
Email: huy.nguyen@oulu.fi

Abstract—One of the biggest obstacles in Chinese and Japanese text line recognition is how to present their enormous character sets. The most common solution is to merely choose and represent a small subset of characters using one-hot encoding. However, such an approach is costly to describe huge character sets, and ignores their semantic relationships. Recent studies have attempted to utilize different encoding methods, but they struggle to build a bijection mapping. In this work, we propose a novel encoding method, called LOGographic DEComposition encoding (LODEC), that can efficiently perform a 1-to-1 mapping for all Chinese and Japanese characters. As such, LODEC enables to encode over 21,000 Chinese and Japanese characters by 520 fundamental elements. Moreover, to handle the vast style variety of handwritten texts in the two languages, we propose a novel deep learning (DL) architecture, called LODENet, together with an end-to-end training scheme, that leverages auxiliary ground truths generated by LODEC or other radical-based encoding methods. We systematically performed experiments on both Chinese and Japanese datasets, and found that our approach surpassed the performance of state-of-the-art baselines. Furthermore, empirical evidence shows that our method can gain significantly improvement using synthesized text line images without the need for domain knowledge.

I. INTRODUCTION

Optical character recognition (OCR) is a crucial component in document analysis systems [1]. OCR technology is critical since it offers effective storage and information retrieval solutions, which help to speed up the manuscript digitization process, and make it more efficient for people to access or analyze them [2]. In addition, as commercial documents are very common in daily activities and are kept being accumulated over time, OCR-based systems are in high demand to convert them into digital forms.

Input data in the offline form of OCR methods are raw bitmaps or compressed images that store printed or written texts on a background characterized by their colors, shapes, and textures. Whereas printed texts of a certain font are consistent throughout documents, handwritten

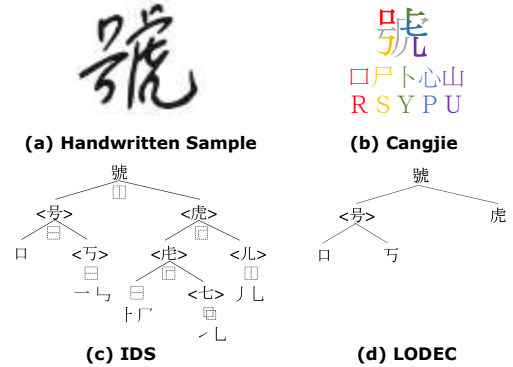


Fig. 1: (a) A handwritten sample of the character 號 and its corresponding encoding methods: (b) Cangjie, (c) Ideographic Description Sequences (IDS) and (d) LODEC.

texts erratically vary during writing, which makes offline handwritten OCR challenging.

In the scope of this study, we specifically focus on offline handwritten Chinese and Japanese text line images. While Chinese and Japanese Kanji are logographic systems, Japanese Kana – Hiragana and Katakana – are syllabic ones. Concerning Kanji, there are several well-known challenges. Unlike Latin phonograms representing limited phonetic elements (e.g., 26 characters in the English alphabet), Chinese and Japanese logographic characters are significantly numerous since they possess both semantic and phonetic traits [3]. As such, the Kangxi dictionary has over 47,000 Kanji characters (21,448 of them have Unicode correspondences), each of which consists of a unique set of radicals and basic components with distinctive semantic meanings [4]. Therefore, the first objective in developing deep learning (DL) based text line OCR for such data is how to encode their semantic relationships embedded in radicals and the huge character set. Then, the next target is to have an appropriate DL architecture, together with a suitable training strategy, to leverage the encoded data.

For the aforementioned demands and challenges, a huge number of studies have emerged to exclusively tackle

*Corresponding author

Kanji-based OCR [5]–[16]. For Kanji encoding, a naive method is to select a small subset of characters and describe them using one-hot encoding. However, one-hot encoding treats each character as an independent category and ignores the fact that common radicals with the same semantic meanings appear repeatedly among different characters. Thus, several logographic encoding methods such as Cangjie [5] and Ideographic Description Sequences (IDS) [6] have been arisen to decompose logograms into radicals and fundamental components. Figures 1 (b) and (c) depict how Cangjie and IDS encode a Kanji, respectively.

There are several limitations when using these encodings. On one hand, Cangjie and IDS are not bijective functions (1-to-1 mappings) and thus, they sometimes create ambiguous and incorrect transcriptions. On the other hand, IDS is too complicated in terms of encoding length for DL training. Hence, as far as our knowledge goes, DL-based studies in the literature have not adopted IDS into their methods yet.

With respect to DL models for Kanji OCR, many important methods have been developed to capture sequential features of Kanji text lines [7], [13], [16]. Unfortunately, an end-to-end DL-based method that well utilizes both original and radical-based ground truth data is still missing.

For the aforementioned issues, we propose a holistic method that offers an efficient encoding for both Chinese and Japanese and a novel DL architecture, as well as an end-to-end training scheme, for leveraging logographic, syllabic, and radical data simultaneously. Firstly, our method includes a decomposition method, called LOGraphic DEComposition encoding (LODEC), which is a compact version of IDS [6]. Instead of decomposing logograms ultimately to fundamental strokes, which are not always visible in practical writings, as in IDS, we limit the decomposition to a predefined set of radicals such that LODEC can perform a 1-to-1 mapping overall logograms (illustrated in Figure 1 (d)). In addition, syllabic Kana characters are decomposed into unique tuples of voiced consonants and pronunciation changes. Therefore, our LODEC is not only a bijective function over logographic and syllabic characters, but also has a smaller search space compared to the well-known IDS [6].

The second component of our method is an end-to-end OCR training scheme that leverages auxiliary data produced by LODEC to improve the performance of the recognition task. To achieve it, we develop a DL architecture, called LODENet, consisting of a module to extract radical-based features from input data and a conversion network to decode logographic and syllabic characters from the previous radical-based features. Thanks to our conversion network, our architecture collaborates well with any DL architecture for sequential recognition and an arbitrary radical-based encoding method. Thus, our study is the very firstly reported one that enables us to make use of the IDS encoding method. On top of those, we

utilize a weighted sum of two Connectionist Temporal Classification (CTC) losses [17] corresponding to radical and original predictions to ensure an end-to-end training manner. In summary, our contributions are as follows:

- 1) We propose the LODEC encoding method that can fully represent all logograms and syllabic characters of Chinese and Japanese scripts,
- 2) We propose an end-to-end training scheme that can be plugged in any sequential architecture and radical-based encoding method,
- 3) We propose the LODENet architecture equipped with the conversion network that learns to transcribe Japanese and Chinese contents from radical-based features,
- 4) We systematically conducted experiments to assess the effect of our components (with statistical tests), the ability of our method to learn from extra generated text line images, and also show that our method achieves state-of-the-art (SOTA) results on two public offline handwritten Chinese datasets such as CASIA [18] and SCUT-EPT [19], and one private Japanese dataset.

II. RELATED WORKS

A. Logographic Encoding Methods

The Cangjie, invented in 1976, is the first Chinese input method to be applied on the QWERTY keyboard [5]. It is a shape-based encoding using 26 fundamental radicals to represent components and auxiliary shapes in Chinese characters. In its 5th version¹, all characters are encoded by at most 5 codes, making it feasible to generate fixed-length labels for the recognition task [16]. However, with such limited codes, it cannot describe all radicals and auxiliary shapes. For instance, “aa”, “aaa”, and “aaam” are the ambiguous encoding sequences of the sets (昌, 開), (晶, 唱), and (壘, 疊, 壘) respectively.

Chinese Japanese Korean (CJK) unified ideographs, the common character scripts of the 3 countries, are well presented by an encoding standard called Ideographic Description Sequence (IDS) [6]. Each encoding sequence is a composition of Ideographic Description Characters (IDCs) and Description Characters (DCs)², defined as a Unicode block with 12 codes from “𠩺” (U+2FF0) to “𠩺” (U+2FFB), which enables to describe logographic layouts. DCs, the basic elements of the decomposition, are either radicals, basic components decomposed from radicals, or fundamental strokes (See Figure 1(c)). However, IDS does not guarantee a 1-to-1 mapping over all logograms. One counterexample is that both characters “土” (mud) and “士” (man) are encoded as “𠩺十一”. Additionally, IDS encoding sequences massively vary in terms of encoding length (as in Table I), and its fundamental strokes are

¹<https://github.com/Jackchows/Cangjie5>

²<https://github.com/cjkvi/cjkvi-ids>

not always available in Unicode or handwritten texts in practice.

In this study, we aim to have both the bijection characteristic and simplicity. We adopt the idea of IDS and simplify its decomposition process. As such, we define a set of radicals, which ones tend to write as single strokes in reality. Subsequently, we merely perform the IDS decomposition until reaching the predefined radicals.

B. OCR with One-hot Encoding

One of the primitive approaches in text line OCR is to selectively choose a subset of logograms and utilize one-hot scheme to encode them. Following that direction, Sahu *et al.* [8] proposed an encoder-decoder architecture that used Convolutional Neural Networks (CNN) as an encoder to extract visual features. Several attention-based encoder-decoder architectures were proposed for Japanese historical document recognition and Chinese text line image recognition [9], [10]. However, those methods could only recognize fixed-length text lines.

Notably, Shi *et al.* [7] proposed a Convolutional Recurrent Neural Network (CRNN) architecture consisting of a CNN-based feature extractor and a sequential component using Recurrent Neural Network (RNN). CRNN could handle text sequences with arbitrary length without character segmentation nor predefined lexicon. Based on this architecture, many follow-up studies improved Chinese text line OCR performance by replacing RNN with Long-Short Term Memory (LSTM) or Bidirectional LSTM [11], [12].

C. OCR with Decomposition Methods

1) *Single Character Decomposition*: Before the prevalence of DL models in offline handwritten logographic text line OCR, Wang *et al.* [13] studied a decomposition method for logogram recognition using radical decomposition and hierarchical radical matching. T.Q. Wang *et al.* [20] proposed using deep residual network to recognize position-dependent radicals. Then, Radical Analysis Network (RAN) series, proposed by Zhang *et al.* [14], applied decomposition methods with an encoder-decoder architecture for single Chinese character recognition. Wang *et al.* [15] applied RAN with densely connected architecture (DenseRAN).

2) *Text Line Decomposition*: Inspired by the Chinese Cangjie input method, Bluche *et al.* [16] proposed the Multi-Dimensional Long Short-Term Memory Recurrent Neural Network (MDLSTM-RNN) for basic *character-level* recognition. MDLSTM-RNN with CTC identified sub-character radicals and subsequently reconstructed characters using a post-processing step with a 3-gram language model. In such a two-step manner, whereas the inference is fast, the final performance was negatively affected.

In this work, we propose an end-to-end training scheme for *offline handwritten Chinese and Japanese text line*

Method	No. of codes	Bijection function	Encoding length Mean & Std.
One-hot	21448	Yes	21448±0
Cangjie	26	No	109±19
IDS	387	No	3618±1722
LODEC (Ours)	520	Yes	1856±691

TABLE I: The amounts of codes and encoding lengths to represent 21,448 Kanji characters having Unicode correspondences in different encoding methods.

recognition that includes a novel LODENet architecture tailored to exploit the strength of both generated radical-based and original logographic ground truth.

III. PROPOSED METHODS

A. Logographic Decomposition Encoding

We propose an encoding method, called LODEC, that is inspired by IDS [6] and based on an observation that in reality, ones tend to write a radical, a combination of glyphs, in a single stroke with a specific cursive pattern as illustrated in Figure 1(a). Consequently, LODEC targets to identify unique shapes of logographic characters rather than fundamental glyphs or partial shapes as in IDS [6] or Cangjie [5], respectively.

Let \mathcal{R} denote the set of 214 Kangxi and 115 CJK radicals corresponding to the Unicode characters from U+2F00 to U+2FD5 and from U+2E80 to U+2EF3, respectively. Unlike IDS [6], our LODEC method performs a hierarchical decomposition until it reaches DCs that belong to \mathcal{R} . Thereby, we resolve a problem of IDS in which not all glyphs are defined in the Unicode standard.

LODEC reduces 21,448 Kanji Unicode blocks to the set of 520 selective radicals and basic components. Although LODEC has more fundamental elements than IDS, our encoding method produces 33% more compact and over 2 times more stable encoded sequences with respect to encoding length compared to the opponent (see Table I).

When applied to Japanese Kana, LODEC decreases the number of characters from 189 to 108. It also isolates the dakuten, indicating voiced consonants, and the han-dakuten, indicating a specific change of pronunciation, as unique components in its outputs (e.g., 𐰇 is a combination of 𐰆 and the dakuten 𐰇).

We eliminate the confusion of similar characters by mapping them to a single Unicode according to the Equivalent Ideograph Dictionary (EID)³. In practice, we encode a sequence of logograms by consequently applying Algorithm 1 to replace each logogram with its LODEC encoding. We use the notation $\{Y'_t\}_{t=1}^{T'} = \text{LODEC}(\{Y_t\}_{t=1}^T)$ to denote the encoding process from a sequence of T logograms to a sequence of T' radicals ($T' \geq T$). We concretely describe the implementation of LODEC in Algorithm 1.

³<http://www.unicode.org/review/pri344/EquivalentUnifiedIdeograph-draft2.txt>

Algorithm 1 LODEC encoding generation

c : logogram, \mathcal{R} : Radical set, EID : Equivalent Ideograph Dictionary, IDS : IDS dictionary, IDC : IDC set, S : encoding set

```

procedure ENCODE( $c$ )
   $L \leftarrow$  empty list
  if  $c \in \mathcal{R}$  then                                 $\triangleright$  stop-at-radical
     $L.append(c)$ 
  end if
  if  $c \in EID \wedge EID[c] \in \mathcal{R}$  then                 $\triangleright$  stop-at-radical
     $L.append(EID[c])$ 
  end if
  if  $L \neq \emptyset$  then                                 $\triangleright$  reach leaf
    return  $L$ 
  end if
  for  $c$  in  $IDS[c]$  do                                 $\triangleright$  decompose logogram
     $L \leftarrow L + \text{Encode}(c)$ 
  end for
   $L \leftarrow L - IDC$                                  $\triangleright$  remove IDCs
  while  $L \in S$  do
     $L \leftarrow L + \text{duplicate\_code}$                  $\triangleright$  ensure bijection
  end while
   $S \leftarrow S \cup L$ 
  return  $L$ 
end procedure

```

B. End-to-end Training Scheme

Our training scheme leverages both characters and radicals from data in an end-to-end manner. The center of the scheme is our LODENet that consists of major components such as a feature extractor, a sequential decoder, and a conversion network. Specifically, LODENet firstly extracts fine-grained features of input image and predicts radicals at an intermediate layer. The prediction, together with radical ground truth, forms the first loss. Subsequently, our model learns to convert the intermediate output representing radicals to original characters. Thus, we obtain the second loss based on ground truth character labels (see Figure 2). In the next sections, we present the components of LODENet and its losses in detail.

1) Visual Feature Extractor and Sequential Decoder:

The input of our architecture is a 1-channel image with a fixed height h and a flexible width w . We utilize a CNN to extract fine-grained features of both radicals and characters from text line images. This component of LODENet is a customized version of the previously developed network [7]. We aim to ultimately shrink the shape of feature maps to $h' \times w' \times c$, where $h' = 4$ in the case of $h = 64$, $w' < w$ and c is the number of filters, using convolutional blocks of a 3×3 convolutional layer with a stride of 1, a rectified linear unit (ReLU) [21], batch normalization [22], and 4 max-pooling layers.

Notably, we keep the height of the final feature map as $h' > 1$ to preserve different radical features at the same horizontal position. Then, we reshape the tensor to $w' \times h' \times c$ representing the input for the sequential decoder. Table II presents the detailed description of this partial architecture in the case that h is 64.

Layer name	Output shape	Layer structure
Input	$64 \times w \times 1$	
Conv1_1	$64 \times w \times 64$	Conv 3×3 , 64, $S = 1$
Conv1_2	$64 \times w \times 64$	Conv 3×3 , 64, $S = 1$
MaxPool_1	$32 \times w/2 \times 64$	MaxPooling 2×2 , $S = (2, 2)$
Conv2_1	$32 \times w/2 \times 128$	Conv 3×3 , 128, $S = 1$
Conv2_2	$32 \times w/2 \times 128$	Conv 3×3 , 128, $S = 1$
Conv2_3	$32 \times w/2 \times 128$	Conv 3×3 , 128, $S = 1$
MaxPool_2	$16 \times w/2 \times 128$	MaxPooling 2×2 , $S = (2, 2)$
Conv3_1	$16 \times w/4 \times 256$	Conv 3×3 , 256, $S = 1$
Conv3_2	$16 \times w/4 \times 256$	Conv 3×3 , 256, $S = 1$
Conv3_3	$16 \times w/4 \times 256$	Conv 3×3 , 256, $S = 1$
MaxPool_3	$8 \times w/2 \times 256$	MaxPooling 2×2 , $S = (2, 2)$
Conv4_1	$8 \times w/8 \times 512$	Conv 3×3 , 512, $S = 1$
Conv4_2	$8 \times w/8 \times 512$	Conv 3×3 , 512, $S = 1$
Conv4_3	$8 \times w/8 \times 512$	Conv 3×3 , 512, $S = 1$
MaxPool_4	$4 \times w/8 \times 512$	MaxPooling 2×2 , $S = (2, 1)$
Reshape	$w/8 \times 2048$	
BiGRU	$w/8 \times K$	Bidirectional GRU

TABLE II: The feature extractor and the sequential encoder architecture in LODENet with fixed input height of 64. S , w and K denote the stride of an operator, the input width, and the size of radical set, respectively.

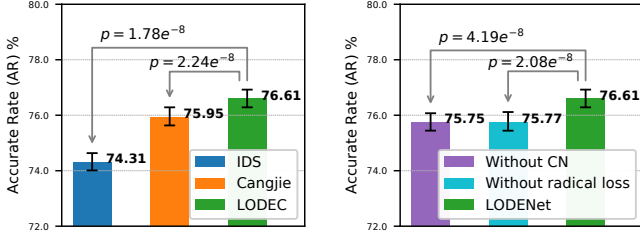
Subsequently, we target to predict the first output, which is related to radicals instead of logographic characters as in [7]. We utilize a bidirectional Gated Recurrent Unit (BGRU) [23] with a hidden size of 512 to exploit the sequential relations of the extracted features. The output of the BGRU is a tensor $R \in \mathbb{R}^{w' \times K}$ representing radical predictions, where K is the radical set size $|\mathcal{R}|$.

2) *Conversion Network*: After obtaining logits of radical predictions R , we aim to convert them to original character-level logits C . As such, we employ a conversion network (CN) that is composed of convolutional and sequential layers. The convolutional layers are inspired by the Inception module [24], where we simultaneously use 4 1D convolutions with different kernel sizes of 1×1 , 1×3 , 1×5 , and 1×7 . By intuition, the 1×1 kernel is suitable for non-decomposable characters (e.g. symbols, Latin, and numbers). The other larger kernels are designed to process groups of radicals that form composable ones. We set the number of filters to 256 for each 1D convolution. The resulting feature maps of all the convolutional layers are concatenated to a single feature map with the size of $1 \times w' \times 1024$, which can be interpreted as a sequence of w' 1024-dimensional vectors $\{X_t\}_{t=1}^{w'} = \text{Inception}(R)$. To capture sequential relationships between radicals, we utilize another Bidirectional GRU to the final feature map as $C_t = \text{BGRU}(X_t)$ resulting in the character-level logits $C \in \mathbb{R}^{w' \times N}$, where N is the number of characters in the vocabulary. The character-level logits are then used to compute CTC loss [17] with the true labels of original logographic characters:

$$\text{CTC}_{\text{logo}} = \text{CTC} \left(\{C_t\}_{t=1}^{w'}, \{Y_t\}_{t=1}^T \right),$$

where $\{Y_t\}_{t=1}^T$ is the ground truth character sequence.

The existence of the intermediate radical representations indicates that the radical and original character



(a) Ablation study on the binding of LODENet and each encoding method (b) Ablation study on the conjoining of LODENet and each encoding method

Fig. 3: Ablation studies on different components of our method. The charts shown ARs and 95% confidence intervals of the training settings. The green bars are our combination of LODENet and LODEC. The arrows indicate one-sided Wilcoxon signed-rank tests between our method and other settings.

generated or crawled from Wikipedia pages, which is detailedly presented in Section IV-A3. As such, we in turn trained our model on data from the CASIA with or without each extra set generated above. The results of those trained models were also compared to the baseline [28]. We present the detailed results in Section IV-C.

Finally, we compared our entire method to a variety of baselines. For the Chinese datasets, our baselines were CNN + MDLSTM + CTC [11], CNN + MDirLSTM + CTC [29], CNN + LSTM + CTC [7] [30] [31] [19], and CMAM [26]. For the Japanese dataset, we chose the state-of-the-art [26] to compete with. Detailed results are presented in Section IV-D.

3) *Implementation Details*: We implemented a traditional CRNN baseline model using the same feature extractor and sequential decoder as LODENet for a fair comparison. By default for LODENet, we used our proposed LODEC to encode Chinese and Japanese to radical encodings. Input images were converted to gray-scale (1 channel) and proportionally re-scaled to a fixed height h . All images in a batch were padded to the same width according to the longest one.

For efficient training, we derived a training procedure based on Curriculum Learning [32]. To be specific, we sorted the text line images ascendingly by the number of characters. Each model was trained with the data sorted in that order for the first 5 epochs. From the 6-th epoch, we shuffled the training data and the model was fed with a random batch of data. We used Adam optimizer [33] with a base learning rate of $1e-4$, and a weight decay of $1e-5$. The batch size was set at 8 across all experiments.

To compute confidence intervals, we applied the bootstrap method that resampled 1,000 times with replacement. To prepare data for Wilcoxon tests, we split the test set of SCUT-EPT to 40 folds, each of which had 250 samples, and computed the AR for each fold.

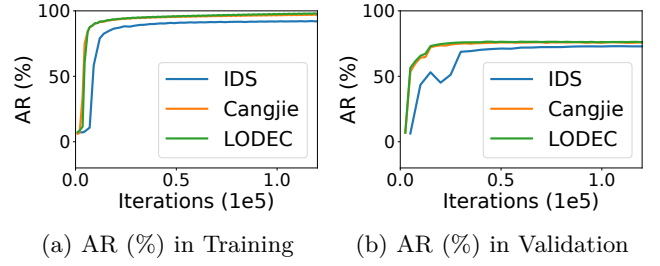


Fig. 4: Learning curves of LODENet with different encoding methods on the SCUT-EPT dataset.

Model	Character (%)		Radical (%)	
	AR \uparrow	CR \uparrow	AR \uparrow	CR \uparrow
CNN + MDLSTM + CTC [11]	73.26	78.30		
CNN + MDirLSTM + CTC [29]	73.65	78.53		
CNN + LSTM + CTC [7] [30] [31] [19]	75.97	80.26		
CMAM [26]	74.45	82.14		
LODENet (Ours)	76.61	82.91	77.81	85.40
LODENet + Random texts (Ours)	77.36	84.64	79.27	87.47
LODENet + Wikipedia texts (Ours)	77.61	83.74	79.25	86.33

TABLE IV: Testing results on SCUT-EPT. The "LODENet" indicates our model only training on standard SCUT-EPT train set (40K text lines) without any augmentation. "LODENet + Random texts" and "LODENet + Wikipedia texts" are the results of our models when training with additional synthesized data.

To generate auxiliary text line images from Wikipedia contents, we randomly selected 12 writers from 1,020 indexes of writers of the CASIA-HWDB 1.0-1.2 dataset, which consists of isolated characters, gathered character images, and finally concatenated ones with the same write index. Following this synthesis scheme, we created additional 120,000 training samples to expand the CASIA-HWDB 2.0-2.2. For the second source, we randomly crawled data from more than 6,000 pages of Simplified Chinese Wikipedia. Then, we randomly split each paragraph into 21,000 sentences and sorted them by length. Subsequently, we chose 10,000 sentences with around 15 characters per text. Finally, we used these selected to generate another extra set of 120,000 training images in the same manner as for the first source.

B. Ablation Studies

1) *The Binding of LODEC and LODENet*: Figure 4 shows that Cangjie [5] or LODEC required less iterations to converge and gained better ARs compared to the original IDS [6]. Moreover, the results of the one-sided Wilcoxon signed-rank test in Figure 3a indicate that the combination of LODEC and LODENet significantly outperformed the combinations with Cangjie or IDS ($p < 1e-7$ in all the cases) with respect to AR. The IDS' drawbacks in terms of long encoding length and high variance substantially affected its performance.

2) *The End-to-end Training Scheme*: We observed significant performance drops when removing either the conversion network or the radical branch. The statistical tests in Figure 3b empirically proved the importance the end-to-end training scheme as p -values in both the comparisons are less than $1e-7$.

C. Learning from Synthesized Data

Results on both the handwritten Chinese datasets in Tables IV and V show that our LODENet gained substantial improvements when trained with auxiliary set of synthesized data. It is worth noting that extra data from either of the data sources helped our model to increase both its AR and CR by at least 5% on CASIA data.

Experimental results reveal that meaningful content from Wikipedia pages did not result in a significant improvement compared to random contents. As such, the AR confidence intervals of the two models with random texts and Wikipedia contents are [91.89%, 92.41%] and [91.71%, 92.24%], respectively. Moreover, the two-sided Wilcoxon test comparing them results in $p = 0.179$.

Model	Character (%)		Radical (%)	
	AR \uparrow	CR \uparrow	AR \uparrow	CR \uparrow
HIT-1 [25]	83.58	86.15		
HIT-2 [25]	86.73	88.76		
Wang et al. [34]	88.79	90.67		
CNN + SMDLSTM + CTC [28]	86.64	87.43		
LODENet (Ours)	86.82	87.83	88.51	91.09
LODENet + Random texts (Ours)	92.00	92.89	92.79	95.25
LODENet + Wikipedia texts (Ours)	92.16	93.04	92.77	94.65

TABLE V: Testing results on CASIA. The "LODENet" indicates model only training on standard CASIA-HWDB 2.0-2.2 dataset (52,2K text lines) without any augmentation, while CNN + SMDLSTM + CTC [28] used 1M training lines. HIT-1, HIT-2, and [34] were trained on both isolated characters CASIA-HWDB 1.0-1.2 and unconstrained text lines from CASIA-HWDB 2.0-2.2.

D. Comparison with the Baselines

As demonstrated in Table IV, among many strong baselines from the literature, our LODENet (with LODEC) achieved an AR of 76.61% and a CR of 82.91%, claiming SOTA on SCUT-EPT. It is noteworthy that LODENet outperformed the best prior model (CNN+LSTM+CTC) [19] by an AR of 0.64% and a CR of 2.65% without using any augmentation and only applying greedy decoding. In addition, we picked several samples to show LODENet's robustness in challenging cases in Figure 5. The collaboration of radical and character predictions helped our model work well with erased texts with supplements, missing strokes, and connected characters. Although the radical output predictions are not exactly correct, LODENet showed the capability to transform the radical feature composition to correct characters with the conversion network.


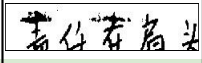
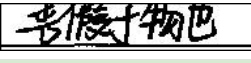
	
Ground truth	与其城市文化有有很大相关性
CRNN	与其城市菊 (8)
LODENet (Ours)	与其城市文化有有很大相关性 (1)
Radical output	与一廿一八土戊丁一少イヒイ木月ソ一大人↑青
	
Ground truth	责任在肩头
CRNN	责任"寿高头 (4)
LODENet (Ours)	责任"在肩头 (1)
Radical output	𠂇貝イ一土"土一尸月头
	
Ground truth	暑假于物也
CRNN	苦物也 (3)
LODENet (Ours)	暑假于物也 (0)
Radical output	一𠂇口イ十𠂇𠂇日又于牛𠂇ノ也

Fig. 5: Picked samples from the test set of SCUT-EPT and their outputs by LODENet and CRNN [7]. Character and radical-based outputs of LODENet are shown in purple and white, respectively. The values in parentheses are corresponding Levenshtein distances.

On the CASIA dataset, without any extra data, our model achieved comparable results to the baseline [28], which added 1 million generated samples. Furthermore, when we added 120,000 auxiliary text lines, which is over 8 times less than the amount in [28], our method surpassed the baseline more than 5% with respect to both AR and CR.

Model	Private Japanese data	
	Validation \downarrow	Test \downarrow
CMAM [26]	17.55	12.99
LODENet w/o conversion network (Ours)	12.36	6.25
LODENet (Ours)	11.70	5.47

TABLE VI: CERs on the Japanese dataset.

LODENet demonstrates significant improvements in this Japanese dataset. Concretely, we outperformed CMAM by huge margins, thereby achieving SOTA results on this dataset. Compared to our implemented CRNN with the same feature extractor and sequential decoder, LODENet again outperformed by CERs of 0.66% and 0.78% on the validation and the test set, respectively.

V. CONCLUSION

We presented a holistic method for offline handwritten Chinese and Japanese text line recognition. Firstly, we proposed LODEC as an alternative for logographic encoding methods when it can perform 1-to-1 mapping overall logographic and syllabic characters in a compact fashion. Then, we proposed an end-to-end training scheme that can well utilize character and radical labels simultaneously. To achieve that, we have the LODENet architecture with CN and a pair of losses for its two prediction branches. Our training scheme is general and allows for plugging any deep CNN architecture and radical encoding method. As far as our knowledge, our method is the first one that can cope with the complex encoding method IDS [6].

We performed systematic experiments and empirical evidence via statistical tests showed that the combination of our components brings significant improvements. Additionally, our method gained improvement leaps when we added synthesized text lines that are merged from single character images with random contents, which do not need domain knowledge.

The idea of our decomposition method can be extended to other applications such as natural language processing and speech-to-text systems for logographic language. It can also be used for a few-shot OCR by identifying the elemental radicals in unseen characters, which will be investigated in our future works.

REFERENCES

- [1] J. Schurmann, N. Bartneck, T. Bayer, J. Franke, E. Mandler, and M. Oberlander, "Document analysis from pixels to contents," *Proceedings of the IEEE*, vol. 80, no. 7, pp. 1101–1119, 1992.
- [2] X. Ding, D. Wen, L. Peng, and C. Liu, "Document digitization technology and its application for digital library in china," in *First International Workshop on Document Image Analysis for Libraries, 2004. Proceedings.* IEEE, 2004, pp. 46–53.
- [3] D. Ghosh, T. Dube, and A. Shivaprasad, "Script recognition—a review," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 32, no. 12, pp. 2142–2161, 2010.
- [4] Kangxi, *Chinese and English dictionary: containing all the words in the Chinese imperial dictionary; arranged according to the radicals, Volume 1.* Printed at Parapattan. Retrieved 2011-05-15., 1842.
- [5] B. Chu, *Handbook of the Fifth Generation of the Cangjie Input Method*, 2009.
- [6] The Unicode Consortium, "The Unicode Standard," Unicode Consortium, Mountain View, CA, Tech. Rep. Version 6.0.0, 2011.
- [7] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *IEEE transactions on pattern analysis and machine intelligence*, 2016.
- [8] D. K. Sahu and M. Sukhwani, "Sequence to sequence learning for optical character recognition," *arXiv preprint arXiv:1511.04176*, 2015.
- [9] A. Le Duc, D. Mochihashi, K. Masuda, and H. Mima, "An attention-based encoder-decoder for recognizing japanese historical documents," 12 2018.
- [10] F. Sheng, C. Zhai, Z. Chen, and B. Xu, "End-to-end chinese image text recognition with attention model," in *International Conference on Neural Information Processing.* Springer, 2017.
- [11] R. Messina and J. Louradour, "Segmentation-free handwritten chinese text recognition with lstm-rnn," in *2015 13th International Conference on Document Analysis and Recognition (ICDAR).* IEEE, 2015.
- [12] C. Zhai, Z. Chen, J. Li, and B. Xu, "Chinese image text recognition with blstm-ctc: a segmentation-free method," in *Chinese Conference on Pattern Recognition.* Springer, 2016.
- [13] Q.-F. Wang, F. Yin, and C.-L. Liu, "Handwritten chinese text recognition by integrating multiple contexts," *IEEE transactions on pattern analysis and machine intelligence*, 2011.
- [14] J. Zhang, Y. Zhu, J. Du, and L. Dai, "Radical analysis network for zero-shot learning in printed chinese character recognition," in *2018 IEEE International Conference on Multimedia and Expo (ICME).* IEEE, 2018.
- [15] W. Wang, J. Zhang, J. Du, Z.-R. Wang, and Y. Zhu, "Denseran for offline handwritten chinese character recognition," in *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR).* IEEE, 2018.
- [16] T. Bluche and R. Messina, "Faster segmentation-free handwritten chinese text recognition with character decompositions," in *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR).* IEEE, 2016.
- [17] F. J. G. A. Graves, S. Fernández and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006.
- [18] C.-L. Liu, F. Yin, D.-H. Wang, and Q.-F. Wang, "Casia online and offline chinese handwriting databases," in *2011 International Conference on Document Analysis and Recognition.* IEEE, 2011.
- [19] Y. Zhu, Z. Xie, L. Jin, X. Chen, Y. Huang, and M. Zhang, "Scut-ept: New dataset and benchmark for offline chinese text recognition in examination paper," *IEEE Access*, 2018.
- [20] T.-Q. Wang, F. Yin, and C.-L. Liu, "Radical-based chinese character recognition via multi-labeled learning of deep residual networks," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 1. IEEE, 2017, pp. 579–584.
- [21] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. icml*, 2013.
- [22] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.
- [23] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.
- [24] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015.
- [25] F. Yin, Q.-F. Wang, X.-Y. Zhang, and C.-L. Liu, "Icdar 2013 chinese handwriting recognition competition," in *2013 12th International Conference on Document Analysis and Recognition.* IEEE, 2013.
- [26] D. Nguyen, N. Tran, and H. Le, "Improving long handwritten text line recognition with convolutional multi-way associative memory," *arXiv preprint arXiv:1911.01577*, 2019.
- [27] F. Wilcoxon, "Individual comparisons by ranking methods," in *Breakthroughs in statistics.* Springer, 1992, pp. 196–202.
- [28] Y.-C. Wu, F. Yin, Z. Chen, and C.-L. Liu, "Handwritten chinese text recognition using separable multi-dimensional recurrent neural network," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR).* IEEE, 2017.
- [29] Z. Sun, L. Jin, Z. Xie, Z. Feng, and S. Zhang, "Convolutional multi-directional recurrent network for offline handwritten text recognition," in *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR).* IEEE, 2016.
- [30] Z. Xie, Z. Sun, L. Jin, H. Ni, and T. Lyons, "Learning spatial-semantic context with fully convolutional recurrent network for online handwritten chinese text recognition," *IEEE transactions on pattern analysis and machine intelligence*, 2017.
- [31] Z. Xie, Z. Sun, L. Jin, Z. Feng, and S. Zhang, "Fully convolutional recurrent network for handwritten chinese text recognition," in *2016 23rd International Conference on Pattern Recognition (ICPR).* IEEE, 2016.
- [32] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proceedings of the 26th annual international conference on machine learning.* ACM, 2009.
- [33] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [34] S. Wang, L. Chen, L. Xu, W. Fan, J. Sun, and S. Naoi, "Deep knowledge training and heterogeneous cnn for handwritten chinese text recognition," in *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR).* IEEE, 2016, pp. 84–89.