Open-Set Text Recognition via Character-Context Decoupling

Chang Liu¹ Chun Yang¹,∗ Xu-Cheng Yin¹,²,∗
¹ School of Computer and Communication Engineering, University of Science and Technology Beijing
² Institute of Artificial Intelligence, University of Science and Technology Beijing, China
lasercat@gmx.us, {chunyang, xuchengyin}@ustb.edu.cn

Abstract

The open-set text recognition task is an emerging challenge that requires an extra capability to cognize novel characters during evaluation. We argue that a major cause of the limited performance for current methods is the confounding effect of contextual information over the visual information of individual characters. Under open-set scenarios, the intractable bias in contextual information can be passed down to visual information, consequently impairing the classification performance. In this paper, a Character-Context Decoupling framework is proposed to alleviate this problem by separating contextual information and character-visual information. Contextual information can be decomposed into temporal information and linguistic information. Here, temporal information that models character order and word length is isolated with a detached temporal attention module. Linguistic information that models n-gram and other linguistic statistics is separated with a decoupled context anchor mechanism. A variety of quantitative and qualitative experiments show that our method achieves promising performance on open-set, zero-shot, and close-set text recognition datasets.

1. Introduction

Text recognition is a well-studied task and has been widely applied in various applications [7]. Most existing text recognition methods assume characters in the testing set are covered by the training set. Moreover, consistency of contextual information between the training set and the testing set is also assumed. These methods are not adaptable to recognize unseen characters without retraining the model. However, as the language evolves, novel ligatures (e.g., rare characters, emoticons, and foreign characters) can be frequently used in a region during a certain period. For example, foreign characters can be seen frequently in scene text images as a result of globalization. Hence, it is unfeasible if the model needs to be retrained whenever a “new character” emerges. This task is defined as the open-set text recognition task [23], as a specific field of open-set recognition [33] and a typical case of robust pattern recognition [54]. Currently, a few visual-matching-based text recognition methods are capable to recognize novel characters in text lines [16, 23, 52].

However, these open-set text recognition methods tend to be affected by contextual information captured from the training set. This phenomenon can be seen in the salience map (Fig. 1)¹, and is also observed in [41]. In such cases, feature representation for each character is always mixed with linguistic information. This could benefit close-set scenarios where the contextual information bias between training and evaluation is negligible, as some characters (e.g. ‘0’ and ‘O’) are hard to separate only by character visual information (glyph shapes). However, under open-set scenarios, contextual information could be severely biased from the training set. Consequentially, existing models may mistakenly “correct” a character into a wrong one that fits “better” in the context according to the training set [41].

To alleviate the impact of contextual information over open-set text recognition, we propose a character-context decoupling framework allowing explicit separation of character visual information and contextual information. Con-
textual information is further decomposed into temporal information and linguistic information. In general, temporal information models the number and order of characters in a word, while linguistic information models n-gram and other linguistic statistics. Accordingly, a Detached Temporal Attention module (DTA) is introduced to model temporal information and isolate it from visual features. Also, a Decoupled Context Anchor mechanism (DCA) is proposed to “explain away” the linguistic information from character visual information. In summary, our framework reduces the confounding effect of training-set contextual information on visual features, making it less vulnerable to the intractable contextual information bias under open-set scenarios.

The main contributions of this paper are summarized as follows:

1. Proposing a Character-Context Decoupling Framework that improves word-level open-set text recognition by reducing the effect of contextual information on the visual representation of novel characters in word-level samples.

2. Proposing a Detached Temporal Attention module that reduces the impact of temporal information over the visual feature extractor.

3. Proposing a Decoupled Context Anchor mechanism that enables the separation of linguistic information from the visual feature extractor.

2. Related Work

Open-set text recognition, as a specific field of open-set recognition [13,33], is a task that requires the model to recognize testing-set words that may contain novel characters unseen in the training set [23]. A few methods [23,52] have been proposed to address this task. Wan et al. [52] proposed to match the visual features of the word image with the glyph image, and to sample the matching results with a class aggregator. Their method does not scale well on large-scale character sets due to the size growth of the glyph images and the similarity maps. On the other hand, OSOCR [23] generates class centers from individual glyphs with a ProtoCNN and matches the class centers with serialized visual features of the word image. The character-based prototype generating design allows reducing the training cost by mini-batching the label set, thus can be applied to larger label-sets. However, these methods [23,52] do not provide effective approaches to separate contextual information, limiting the performance of open-set word-level recognition. Impacts of contextual information are also studied in [41], which suggests that RNN-free methods are also prone to contextual information bias. Hence, we propose a framework that decouples and isolates contextual information from character visual information to improve open-set visual-matching accuracy.

The conventional close-set text recognition tasks can be considered as a special case where the testing set has zero novel characters. In most conventional text recognition methods [3,8,12,20,34,44,48,55], class centers are mostly modeled as weights in linear classifiers, while visual information and contextual information are modeled together without explicit separation. Recently, more methods opt to adopt dedicated post-processing fashioned modules [11,49] to model contextual information.

The zero-shot character recognition task is another special case of open-set text recognition. Many methods [4,6,16,43,45] propose to encode each character with a unique structural representation (e.g., radical or stroke sequences) for prediction. Recently, a few methods demonstrate capabilities for Korean character recognition [6] and whole word recognition [16]. Despite performing reasonably well with large label sets, these methods require language-specific structural representations of characters, thus limit them to corresponding languages. In contrast, structure-free methods like [1,23] only require a template from a font (or a printed sample) for each character. This approach benefits scenarios where little prior knowledge of character composition can be given, for example, ancient writings of the Oracle characters. Our method follows the structure-free scheme and further achieves reliable word recognition capability by introducing character-context decoupling.

3. Proposed Method

In this work, we propose a character-context decoupling framework (shown in Fig. 2) to reduce the impact of contextual information bias under open-set scenarios, by separating and isolating character visual information and contextual information with the Detached Temporal Attention module and the Decoupled Context Anchor mechanism. The framework and its optimization are first formulated in Section 3.1. Then, a detailed explanation of the less intuitive Decoupled Context Anchor mechanism is presented in Section 3.2. Finally, the Open-set Character-Context Decoupling network (OpenCCD) is given as an example implementation of our framework in Section 3.3.

3.1. Character-Context Decoupling Framework

The framework takes a sample (word-level image) $img$ and a character set $E$ as input, and outputs the predicted word $\hat{y} = (\hat{y}_{[0]}, ..., \hat{y}_{[l]})$ with the maximum probability given the sample and character set,

$$\hat{y} = \arg \max_y P(y|x, E; \theta),$$

where $x$ is the visual feature representation of all characters in the sample. We omit the $E$ and $\theta$ in the following part for writing convenience. In our framework, we expand $P(y|x)$ with a predicted length $l$, using the law of total probability,

$$P(y|x) = \sum_{l=1}^{maxL} P(l|x)P(y|x, l),$$

where
where \( \text{maxL} \) is the maximum length of a word. Different from most existing text recognition frameworks using end-of-speech [35, 44], segmentation [21, 40], or blanks [9, 34] to handle lengths, our framework explicitly predicts the length. \( P(y|x, l) \) can be further decomposed to contextual prediction and visual prediction via the proposed Decoupled Context Anchor mechanism (detailed in Section 3.2),

\[
P(y|x, l) = \prod_{t=1}^{l} P(y[t]|x[t]) \prod_{t=1}^{\text{len}} P(y[t]|c) P(c|x, l).
\]

(3)

Here, \( c \) is the common “context” (linguistic information) of characters, \( x \) models the visual information of all characters in the input image, and \( x[t] \) corresponds to the character visual information of the \( t \)th character. Hence, the optimization goal would be maximizing the log-likelihood \( \log P(y^*|x) \) of the ground truth label sequence \( y^* \),

\[
\log P(y^*|x) = \log(\sum_{l=1}^{\text{maxL}} P(l|x)P(y^*|x, l))
\]

(a)

\[
= \log P(l^*|x) + \log P(y^*|x, l^*)
\]

\[
= \log P(l^*|x) + \sum_{t=1}^{l^*} \log(\sum_{c \in C[t]} P(y[t]|c) P(c|x, l^*))
\]

\[
+ \sum_{t=1}^{l^*} \log(\int_{c \in C(t)} P(y[t]|c) P(c|x, l^*))
\]

\[
= - (L_{\text{len}} + L_{\text{vis}} + L_{\text{ctx}}),
\]

where \( L_{\text{len}} \), \( L_{\text{vis}} \), and \( L_{\text{ctx}} \) are the corresponding cross-entropy losses of the three loglikelihood terms. Step \( (a) \) holds because the correct label can only be predicted when the length is correctly predicted.

### 3.2. Decoupled Context Anchor Mechanism

In this work, we propose a Decoupled Context Anchor mechanism to model and separate the effect of linguistic information \( c \) over character \( y[t] \) at each timestamp \( t \).

**Assumption 1 (A1)** We assume the linguistic information functions as a common-cause of the input visual information and the prediction outputs at all timestamps (See Fig. 3). We model the sample image as a “rendered” result of the label \( y \). Also, we assume the label (words) is generated according to linguistic information \( c \), making the label \( y \) a causal result of \( c \). Hence, linguistic context \( c \) and the character-level visual information \( x[t] \) are the only two direct factors affecting the probability of \( y[t] \) at timestamp \( t \),

\[
P(y[t]|x[t], x, y[t-1]...y[0], l, c) = P(y[t]|x[t], c),
\]

(5)

**Assumption 2 (A2)** The shape (character visual information) of a character and its context (linguistic information)
are independent given the character $y_{[t]}$, i.e.,
\begin{align}
P(x_{[t]} | y_{[t]}, c) &= P(x_{[t]} | y_{[t]}) \\ \iff P(x_{[t]} | c, y_{[t]}) &= P(x_{[t]} | y_{[t]})P(c | y_{[t]}).
\end{align}

This assumption implies that the linguistic information does not affect the “style” (font face, color, background, etc.) of the word, which generally holds in most synthetic datasets where styles and contents are randomly matched.

**Theorem 1: The Anchor Property of Context**

Given assumption A1, the probability of a predicted word $y$ given image $x$ and its length $l$, $P(y | x, l)$, can be written as the product of the “anchored predictions” of all timestamps, i.e.,
\begin{equation}
P(y | x, l) = \prod_{t=1}^{l} \int_{c \in C_{[t]}} P(y_{[t]} | x_{[t]}, c)P(c | x, l),
\end{equation}
and the proof is detailed in Appendix A. Here, the integral term can be interpreted as an ensemble of “anchored prediction” $P(y_{[t]} | x_{[t]}, c)$ over all possible contexts $c$, which is similar to the hidden anchor mechanism [15]. Hence, we call this theorem the anchor property of context.

**Theorem 2: The Separable Property of Linguistic Information and Character Visual Information**

Given Assumption A2, the effect of character visual information over the label $P(y_{[t]} | x_{[t]})$ and the effect of linguistic information $P(y_{[t]} | c)$ is separable from contextual prediction $P(y_{[t]} | x_{[t]}, c)$,
\begin{equation}
P(y_{[t]} | x_{[t]}, c) \propto \frac{P(y_{[t]} | x_{[t]})P(y_{[t]} | c)}{P(y_{[t]})}.
\end{equation}
Here, $P(y_{[t]} | x_{[t]})$ represents the predicted probability of $y_{[t]}$ with regard to character visual information $x_{[t]}$, $P(y_{[t]} | c)$ models the effect caused by linguistic information, and $P(y_{[t]})$ models the character frequency on the training set. The proof of this theorem is given in Appendix C. This theorem suggests that the effect of character visual information and linguistic information over the prediction can be explicitly separated under specific conditions.

Intuitively, $P(y_{[t]} | c)$ “explains away” [47] the linguistic information from the visual-based prediction $P(y_{[t]} | x_{[t]})$, This behavior happens in the backpropagation pass of our framework during training, where the gradients of $L_{ctx}$ and $L_{vis}$ are accumulated to update the feature extractor. This is the reason that $L_{ctx}$ needs to be backpropagated, and also makes $L_{ctx}$ a regularization term, in terms of enforcing certain properties of the network via backpropagation. This property differentiates it from the “look-twice” mechanisms [11,49] that cut gradients.

**Theorem 3: Decoupled Context Anchor Mechanism**

Combining Theorem 1 and Theorem 2, we have the Decoupled Context Anchor mechanism,
\begin{equation}
P(y | x, l) = \prod_{t=1}^{l} P(y_{[t]} | x_{[t]}) \prod_{t=1}^{l} \int_{c \in C_{[t]}} P(y_{[t]} | c)P(c | x, l).
\end{equation}

Proof of this theorem can be found in Appendix B. The mechanism further allows explicit separation of linguistic information and character visual information on the word level, which provides a way to model and separate linguistic information learned on the training set, resulting in a feature extractor focusing more on character visual information and less affected by the training set linguistic information. Considering the anchor property revealed in Theorem 1 and the decoupling nature of Theorem 2, we call this mechanism the Decoupled Context Anchor mechanism.

**3.3. OpenCCD Network**

In this section, the Open-set Character-Context Decoupling network (OpenCCD, Fig. 2) is given as an example implementation of our proposed framework. Here, character set $E : (E_v, E_c)$ consists of glyphs from the Noto font $E_v$ and semantic embeddings of the characters $E_c$. The network first extracts visual features of the word images $img$ and the glyphs $E_v$ with the 45-layer ResNet built with DSBN [5] layers (Res45-DSBN). It shares the convolutional layers between the glyphs and word images, while keeping task-specific batch statistics. Three levels of word features $(F_l, F_m, F_h)$ and the latest feature map $F_h^n$ of glyphs are used. The prototypes (class centers) $W_n$ are generated by applying geometric attention to $F_h^n$. During training, we mini-batch $E_v$ at each iteration to achieve a reasonable training speed. During the evaluation, the visual prototypes $W_n$ for the whole dataset are cached beforehand, hence prototype generation yields little extra costs.

Next, the Detached Temporal Attention (DTA) module is used to predict the length of the word $P(l | x)$, and the max probable length is denoted as $\hat{l}$. Then the DTA module samples ordered character-level visual features $x : (x_{[0]}, ..., x_{[\hat{l}]}$) from the feature map $F_h$.

The visual-based prediction $P(y_{[t]} | x_{[t]})$ is then produced by the open-set classifier. For close-set scenarios, the linguistic information oriented prediction $P(y_{[t]} | c)$ is produced via the Decoupled Context Anchor (DCA) module. For open-set scenarios where linguistic information is intractable, $P(y_{[t]} | c)$ is treated as a uniform distribution, which is equivalent to only using the visual prediction.

**Detached Temporal Attention** In OpenCCD, the detached temporal attention module (Fig. 4) is proposed to
Figure 4. The proposed detached temporal attention module. We isolate sequence modeling within the Temporal Attention module, and zero the gradient of convolution features, w.r.t., the temporal attention map. Here, GAP indicates a global-average pooling.

predict the sequence length \( P(l|x) \). It also sort and sample character in feature map \( F_h \) via the attention map \( A \). The module utilizes an FPN to model global temporal information from the input feature maps, and decodes them into the most probable length \( l \) for individual characters according to attention map \( A \) and the most probable length \( l \), allowing only character visual information backpropagating to \( F_h \) via \( A \).

**Open-Set Classifier** In OpenCCD, \( P(y_t|x_t) \) is produced by comparing the prototypes with character-level features \( x_t \),

\[
P(y_t|x_t) \propto \begin{cases} 
\alpha|x_t| & y_t \text{is [UNK]} \\
|x_t| Sim(x_t, y_t) & \text{otherwise}, 
\end{cases}
\]

where \( |x_t| \) is the L2-Norm of \( x_t \), “[UNK]” indicates unknown characters, and \( \alpha \) is a trainable similarity threshold for rejection. \( Sim(x_t, y_t) \) is defined as

\[
Sim(x_t, y_t) := \max_{w_v \in \psi(y_t)} (\cos(w_v, x_t)),
\]

where \( \psi \) returns all prototypes \( \psi(y_t) \subset W_v \) associated with label \( y_t \), and each individual prototype \( w_v \) corresponds to a “case” of character \( y_t \).

**Decoupled Context Anchor** Instead of implementing a Variational Auto Encoder [19] to estimate the distribution of linguistic information and estimate the integral with Monte-Carlo, we approximate the integral with predicted context \( \hat{c} \), which is similar to the conventional anchor mechanisms using only the anchor with maximum prediction likelihood [30, 31],

\[
\int_{c \in C[\hat{c}]} P(y_t|c)P(c|x, l) \approx P(y_t|\hat{c}).
\]

Combined with Eq. 9, the probability of a predicted character at timestamp \( t \) can be approximated as,

\[
P(y_t|x) \approx P(y_t|x_t) P(y_t|\hat{c}).
\]

As linguistic information is mostly related to labels, we estimate the linguistic information \( \hat{c} \) from the predicted label instead of the feature map. More specifically, the module reuses the estimated character probability distribution \( Y \in (0, 1)^{M} : (P(Y_0|x_t), \ldots, P(Y_M|x_t)) \) with regard to character visual information at each timestamp \( t \), and \( P(Y_t|x_t) : (P(y_t^0|x_t), \ldots, P(y_t^M|x_t)) \) is the probability distribution of all characters at timestamp \( t \). Then \( \hat{c} \) is estimated with a 4-layer transformer encoder [39] applied on the expectation of character embeddings,

\[
\hat{c} = \text{Trans}(YE_c),
\]

where \( E_c \in R^{M \times C} \) is the semantic embedding of seen characters in the training set, hence \( YE_c \) interprets as expectation. \( \text{Trans} \) indicates the 4-layer transformer encoder. Finally, \( P(Y_t|\hat{c}) \) is estimated by comparing character embedding \( E_c \) to \( \hat{c} \),

\[
P(Y_t|\hat{c}) = \sigma(\hat{c}E_c^T)_{[t]},
\]

where \( \sigma \) is the softmax function.

**Optimization** With Eq.12 reducing the integral down to a standard classification problem, \( L_{cls} \) in Eq. 4 can be implemented as a cross-entropy loss like \( L_{ten} \) and \( L_{vis} \). Hence, OpenCCD can be optimized with the three equally-weighted cross-entropy losses.

4. Experiments

The work is based on the OSOCR [23], our code\(^2\) and datasets\(^3\) are released. We conduct experiments on benchmarks for all three scenarios: open-set word-level recognition, zero-shot character recognition, and the conventional close-set word-level recognition benchmarks. Moreover, the ablative studies for open-set word-level recognition are also performed. We use the AdaDelta optimizer, the learning rate is set to \( 10^{-2} \), and decreases by every 200k iterations. For word recognition tasks, we provide a “large” network for an alternative speed-performance trade-off profile closer to SOTA methods, where the large network has more latent channels in the ResNet45-DSBN backbone.

4.1. Open-Set Text Recognition

We use a collection of Chinese text recognition datasets [10, 28, 36, 38, 50] as the training set and the

\(^2\)https://github.com/lancercat/VSDF
\(^3\)https://www.kaggle.com/vsdf2898kaggle/osocrtraining
Table 1. Detailed performance analysis on the open-set text recognition dataset. Performance data listed in Character Accuracy (top) / Line Accuracy (bottom) manner.

| Method   | OSOCR [23] | OSOCR Large [23] | Ours       | Ours Large |
|----------|------------|------------------|------------|------------|
| Kana     | 18.75      | 0.10             | 43.55      | 11.17      |
| Shared   | 79.86      | 73.81            | 76.48      | 74.15      |
| Kanji    | 71.74      | 34.08            | 77.50      | 58.33      |
| All      | 75.33      | 51.64            | 77.06      | 65.42      |
| Overall  | 47.89      | 29.08            | 62.16      | 41.31      |

Japanese subset of MLT [28] as the testing set following OSOCR [23], and all models are trained for 200k iterations. Quantitative performances are shown in Table 1 along with SOTA methods, and qualitative samples can be found in Fig 5. The results show overall significant performance improvement over OSOCR [23]. Details suggest the performance gain comes from recognizing unseen characters. The model shows some extent of robustness over novel characters (text with yellow color in Fig. 5) like unique Kanjis and Kanas.

Results indicate that characters having close shapes are the most significant source of mistakes. The reason for this phenomenon could be pushing all negative classes alike with hard labels (in contrast to soft-label), which is also mentioned in fine-grain classification [53]. Blur, text art can be another major cause for the failure cases, which is expectable as linguistic information is intractable under open-set scenarios, consequentially cannot be used to recover the visually indistinguishable characters.

4.2. Ablative Study

We conduct ablative studies on the open-set text recognition challenge to validate the effect of decoupling character visual information and contextual information. In this section, we train all ablative models on the same server to minimize the confounding factors, the results (Fig. 6) show that isolating temporal information with the Detached Temporal Attention module can improve the open-set recognition performance. Also, further separating linguistic information with the Decoupled Context Anchor mechanism is proved to yield more improvement.

Intuitively, test-set accuracy curves show that both proposed approaches introduce a steady performance improvement on most iterations. The instability of the curves is caused by the Line Accuracy metric where one wrong character can compromise the whole line. We further perform paired t-tests to quantitatively validate the robustness of the performance improvements. Separating temporal information with the DTA module shows a 2.00 t-value and 0.06 p-value, while using DCA to separate linguistic information gives a 5.87 t-value and 1.54 × 10^{-5} p-value. The p-values suggest we can reject the “two-sided” (no improvement) null hypothesis for both approaches. Hence, there is strong evidence that both DTA and DCA can robustly improve the open-set recognition performance.

Qualitatively, we show the result comparison between our model and the base model in Fig. 7. Our framework demonstrates decent robustness improvement against the linguist information bias compared to the base model by separating linguistic information and character visual information.
Table 2. Zero-shot character recognition accuracy on HWDB and CTW datasets. * indicates “online trajectory” data required.

| Method       | Venue     | HWDB      | CTW      |
|--------------|-----------|-----------|----------|
|              |           | # characters in training set | # characters in training set |
|              |           | 500 | 1000 | 1500 | 2000 | 500 | 1000 | 1500 | 2000 |
| CM* [1]      | ICDAR ’19 | 44.68 | 71.01 | 80.49 | 86.73 | -   | -   | -   | -   |
| DenseRan [46]| ICFHR ’18 | 1.70 | 8.44  | 14.71 | 63.8  | 0.12 | 1.50 | 4.95 | 10.08 |
| FewRan [43]  | PRL ’19   | 33.6 | 41.5  | 63.8  | 70.6  | 2.36 | 10.49 | 16.59 | 22.03 |
| HCCR [4]     | PR ’20    | 33.71| 53.91 | 66.27 | 73.42 | 23.53| 38.47 | 44.17 | 49.79 |
| OSOCR [23]   | -         | 46.67| 72.19 | 79.82 | 84.31 | 27.94| 48.23 | 58.56 | 63.77 |
| Ours         | -         | 90.93| 94.10 | 94.58 | 95.55 | 58.22| 68.56 | 74.45 | 77.18 |

Figure 7. Comparison between our method and the base method. Green indicates correct predictions, red indicates wrong predictions. Yellow indicates novel characters and white indicates seen characters.

4.3. Conventional Benchmarks

Due to the lack of open-set text recognition benchmarks, we adopt two well-studied special cases to give referenced comparisons on generalization capability and word recognition capability. Here, we stick to the most applied protocols in each corresponding community to train, evaluate, and measure the performances.

Zero-Shot Character Recognition Following the common protocol in the community [1,4,23,43], we perform the zero-shot Chinese character recognition benchmarks on the HWDB [24] and the CTW [50] dataset following [4,23,43]. The model is trained for 50k iterations due to the small size of the training set. As shown in Table 2, our method shows a significant performance advantage over existing methods. Qualitative samples in Fig. 8 show some robustness over style diversity, slight blur, and other confounding factors. This also suggests that some degenerates like blur or low-contrast do not necessarily yield permanent information loss and could be inverted with sufficient well-distributed training data. We owe part of the robustness difference between open-set word recognition and this challenge to the potential language-specific “image style” bias, caused by factors including different cameras, picture-taking habits in different corresponding region.

These experiments demonstrate reasonable generalization capability compared to the SOTA zero-shot character recognition methods. This also justifies the choice of the data-driven latent representation against model-driven representations like the radical sequences [4, 6]. Our method does not require structural knowledge of characters, which enables potential use to recognize Oracles and other ligatures where such knowledge is unknown or not applicable.

Close-Set Benchmarks Finally, we perform experiments on the conventional close-set benchmarks, where the method is compared to SOTA text recognition methods performance-wise and speed-wise. We report both dictionary-free performance (Table 3) and dictionary-based
Table 3. Performance on conventional close-set benchmarks. * indicates character-level annotation and + for multi-batch evaluation.

| Methods     | Venue     | Training Set | RNN | FPS | IIIT5K | SVT | IC03 | IC13 | CUTE |
|-------------|-----------|--------------|-----|-----|--------|-----|------|------|------|
| Comb.Best [2] | ICCV’19   | MJ+ST        | Y   | 36.23 | 87.9 | 87.5 | 94.4 | 92.3 | 71.8 |
| SAR [20]     | AAAI’19   | MJ+ST        | Y   | -    | 91.5  | 84.5 | -    | -    | 83.3 |
| ESIR [51]    | CVPR’19   | MJ+ST        | Y   | -    | 93.3  | 90.2 | -    | -    | 83.3 |
| SCATTER [22] | CVPR’20   | MJ+ST+Extra  | Y   | -    | 93.7  | 92.7 | 96.3 | 93.9 | 87.5 |
| SEED [29]    | CVPR’20   | MJ+ST        | Y   | -    | 93.8  | 89.6 | -    | 92.8 | 83.6 |
| DAN [44]     | AAAI’20   | MJ+ST        | Y   | -    | 94.3  | 89.2 | 95.0 | 93.9 | 84.4 |
| Rosetta [3]  | KDD’18    | MJ+ST        | N   | 212.76 | 84.3 | 84.7 | 92.9 | 89.0 | 69.2 |
| CA-FCN* [21] | AAAI’19   | ST           | N   | 45   | 92.0  | 82.1 | -    | 91.4 | 78.1 |
| TextScanner* [40] | AAAI’20 | MJ+ST+Extra  | N   | -    | 93.9  | 90.1 | -    | 92.9 | 83.3 |
| Ours-Large   |           | MJ+ST        | N   | 66.91/255+ | 91.90 | 85.93 | 92.38 | 92.21 | 83.68 |

Table 4. Experiments on lexicon-based close-set benchmarks. * indicates close-set methods and + indicates datasets other than MJ and ST are used.

| Method      | Venue     | IIIT5k (small/middle) | IIIT5k (full) | IC03 (50) | SVT | SVT |
|-------------|-----------|------------------------|---------------|-----------|-----|-----|
| AON* [9]    | CVPR’18   | 99.6/98.1              | 96.7          | 96        |    |    |
| ESIR* [51]  | CVPR’19   | 99.6/98.8              | -             | 97.4      |    |    |
| CA-FCN* [21]| AAAI’19   | 99.8/98.9              | -             | 98.5      |    |    |
| Zhang et al. [52] | ECCV’20 | 96.2/92.8              | 93.3          | 92.4      |    |    |
| OSOCR-L [23] |           | 99.5/98.6              | 96.7          | 96.7      |    |    |
| Ours-L      |           | 99.8/99.0              | 96.9          | 97.9      |    |    |

performance [52]. More specifically, the model is trained on MJ [17] and ST [14] following the mainstream technique of SOTA methods. For evaluation, IIIT5k [27], SVT [42], ICDAR 2003 [25], ICDAR 2013 [18], and CUTE [32] are used. Our model is trained for 800k iterations due to the significantly larger training set.

We first compare our method to other open-set text recognition methods that report their performances on lexicon-based benchmarks in Table 4, together with some popular close-set recognition methods. Results show our method retains reasonable close-set performance compared to other open-set methods. Our method also reaches close performance against SOTA close-set methods on this benchmark. Second, comparisons using the dictionary-free protocol are shown in Table 3. Our method shows competitive performance against lightweight text recognition methods. Following community convention [2, 21], running speed is adopted to measure the cost of the method. Our method can reach 67 FPS single batched and 255 FPS multi-batched on a laptop with an RTX 2070 Mobile GPU (7 TFlops), while only using 2.5 GiB Vram. This justifies our model as a competitive light-weight method for conventional tasks.

5. Limitations

Despite showing reasonable performances on all tested scenarios, our method still has some limitations. Framework-wise, we made a few strong assumptions. First, we assume the visual feature extractor can be generalized to a new language. Despite showing better intra-language transferring capability than radical-based methods, it is a little bit too strong to assume robust inter-language transferring capability. These limitations could be causing the performance gap between Kanas and Unique kanjis. Implementation-wise, our method uses a small input (32 * 128 patches) and lacks effective rectification modules [26, 35]. This leads to a very small effective text area, hence limiting the performance of skewed and curvy samples. We will discuss how to address these limitations in our next work.

6. Conclusion

In this paper, we propose a Character-Context Decoupling Framework for open-set text recognition, which is theoretically sound and experimentally feasible. Specifically, the ablative studies and comparative experiments verify that our implementation is an effective open-set text recognition method and a production-ready lightweight text recognition method under close-set scenarios.

7. Acknowledgement

The research is partly supported by the National Key Research and Development Program of China (2020AAA09701), The National Science Fund for Distinguished Young Scholars (62125601), and the National Natural Science Foundation of China (62006018, 62076024).
References

[1] Xiang Ao, Xu-Yao Zhang, Hong-Ming Yang, Fei Yin, and Cheng-Lin Liu. Cross-modal prototype learning for zero-shot handwriting recognition. In ICDAR, pages 589–594, 2019. 2, 7

[2] Jeonghun Baek, Gyeok Su Kim, Junyeeop Lee, Sungrae Park, Dongyoon Han, Sangdoo Yun, Seong Joon Oh, and Hwalsuk Lee. What is wrong with scene text recognition model comparisons? dataset and model analysis. In ICCV, pages 4714–4722, 2019. 8

[3] Fedor Borisyuk, Albert Gordo, and Viswanath Sivakumar. Rosetta: Large scale system for text detection and recognition in images. In KDD, pages 71–79, 2018. 2, 8

[4] Zhong Cao, Jianguo Lu, Sen Cui, and Changshui Zhang. Zero-shot handwritten Chinese character recognition with hierarchical decomposition embedding. Pattern Recognition, 107:107488, 2020. 2, 7

[5] Woong-Gi Chang, Tackgeun You, Seonguk Seo, Suha Kwak, Jeonghun Baek, Geewook Kim, Junyeop Lee, Sungrae Park, Xiang Ao, Xu-Yao Zhang, Hong-Ming Yang, Fei Yin, and Mingshuai Dong. Transformer text recognition with deep learning algorithm. In IJCAI, pages 615–621, 2021. 2, 7

[6] Xiao-xue Chen, Lianwen Jin, Yuanzhi Zhu, Canjie Luo, and Tianwei Wang. Text recognition in the wild: A survey. ACM Comput. Surv., 54(2):42:1–42:35, 2021. 1

[7] Yuhao Huang, Lianwen Jin, and Dezhi Peng. Zero-shot Chinese text recognition via matching class embedding. In ICDAR, volume 12823, pages 127–141, 2021. 1, 2

[8] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Synthetic data and artificial neural networks for natural scene text recognition. CoRR, abs/1406.2227, 2014. 8

[9] Simon M. Lucas, Alex Panaretos, Luis Sosa, Anthony Tang, Shirley Wong, Robert Young, Kazuki Ashida, Hiroki Nagai, Masayuki Okamoto, Hiroaki Yamamoto, Hitoshi Miyao, Tomeki Ogata, and Bohyung Han. Domain-specific batch normalization for unsupervised domain adaptation. In CVPR, pages 7354–7362, 2019. 3, 4

[10] Jingye Chen, Bin Li, and Xiangyang Xue. Zero-shot Chinese character recognition with stroke-level decomposition. In ICCV, pages 7089–7107, 2021. 2, 4

[11] Canjie Luo, Lianwen Jin, and Zenghui Sun. MORAN: A multi-object rectified attention network for scene text recognition. In ICDAR, pages 8610–8617, 2019. 2, 8

[12] Hui Li, Peng Wang, Chunhua Shen, and Guyu Zhang. Show, attend and read: A simple and strong baseline for irregular text recognition. In AAAI, pages 8610–8617, 2019. 2, 8

[13] Minghui Liao, Jian Zhang, Zhaoyi Yan, Fengming Xie, Junjun Liang, Pengyuan Lyu, Cong Yao, and Xiang Bai. Chinese scene text recognition from two-dimensional perspective. In AAAI, pages 8714–8721, 2019. 3, 8

[14] Ron Litman, Oron Anschel, Shahar Tsiper, Roei Litman, Shai Mazor, and R. Manmatha. SCATTER: Selective context attentional scene text recognizer. In CVPR, pages 11959–11969, 2020. 8

[15] Cheng-Lin Liu, Fei Yin, Da-Han Wang, and Qiu-Feng Wang. CASIA online and offline Chinese handwriting databases. In ICDAR, pages 37–41, 2011. 7

[16] Zhi Qiao, Yu Zhou, Dongbao Yang, Yucan Zhou, and Weipeng Sun, Chun Chet Ng, Canjie Luo, Zihan Ni, Chuanhong Mao, and Shuang Peng. Transformer text recognition via transformer. Comput. Commun., 178:153–160, 2021. 2

[17] Shancheng Fang, Hongtao Xie, Yuxin Wang, Zhendong Mao, and Yongdong Zhang. Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition. In CVPR, pages 7098–7107, 2021. 2, 5

[18] Xinjie Feng, Hongxun Yao, Yuankai Qi, Jun Zhang, and Shengping Zhang. Scene text recognition transformer. CoRR, abs/2003.08077, 2020. 2

[19] Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. Synthetic datasets for text localisation in natural images. In CVPR, pages 2315–2324, 2016. 8

[20] Yuhao Huang, Lianwen Jin, and Dezhi Peng. Zero-shot Chinese text recognition via matching class embedding. In ICDAR, volume 12823, pages 127–141, 2021. 1, 2

[21] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Synthetic data and artificial neural networks for natural scene text recognition. CoRR, abs/1406.2227, 2014. 8

[22] Dimosthenis Karatzas, Faisal Shaafait, Seiichi Uchida, Masakazu Iwamura, Lluis Gomez i Bigorda, Sergi Robles Mestre, Joan Mas, David Fernández Mota, Jon Almazán, and Lluís-Pere de las Heras. ICDAR 2013 robust reading competition. In ICDAR, pages 1484–1493, 2013. 8

[23] Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. Synthetic data and artificial neural networks for natural scene text recognition. CoRR, abs/1406.2227, 2014. 8

[24] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun, editors, ICLR, 2014. 5

[25] Simon M. Lucas, Alex Panaretos, Luis Sosa, Anthony Tang, Shirley Wong, Robert Young, Kazuki Ashida, Hiroki Nagai, Masayuki Okamoto, Hiroaki Yamamoto, Hitoshi Miyao, JunMin Zhu, WuWen Ou, Christian Wolf, Jean-Michel Jolion, Leon Todoran, Marcel Worring, and Xiaofan Lin. ICDAR 2003 robust reading competitions: entries, results, and future directions. Int. J. Document Anal. Recognit., 7(2-3):105–122, 2005. 8

[26] Cheng-Lin Liu, Tianwei Wang. Text recognition in the wild: A survey. Pattern Recognition, 73:62, 2019. 3, 4

[27] Cheng-Lin Liu, Fei Yin, Da-Han Wang, and Qiu-Feng Wang. CASIA online and offline Chinese handwriting databases. In ICDAR, pages 37–41, 2011. 7

[28] Simon M. Lucas, Alex Panaretos, Luis Sosa, Anthony Tang, Shirley Wong, Robert Young, Kazuki Ashida, Hiroki Nagai, Masayuki Okamoto, Hiroaki Yamamoto, Hitoshi Miyao, JunMin Zhu, WuWen Ou, Christian Wolf, Jean-Michel Jolion, Leon Todoran, Marcel Worring, and Xiaofan Lin. ICDAR 2003 robust reading competitions: entries, results, and future directions. Int. J. Document Anal. Recognit., 7(2-3):105–122, 2005. 8

[29] Canjie Luo, Lianwen Jin, and Zenghui Sun. MORAN: A multi-object rectified attention network for scene text recognition. Pattern Recognition, 90:109–118, 2019. 8

[30] A. Mishra, K. Alahari, and C. V. Jawahar. Scene text recognition using higher order language priors. In BMVC, pages 127.1–127.11, 2012. 8

[31] Simon M. Lucas, Alex Panaretos, Luis Sosa, Anthony Tang, Shirley Wong, Robert Young, Kazuki Ashida, Hiroki Nagai, Masayuki Okamoto, Hiroaki Yamamoto, Hitoshi Miyao, JunMin Zhu, WuWen Ou, Christian Wolf, Jean-Michel Jolion, Leon Todoran, Marcel Worring, and Xiaofan Lin. ICDAR 2003 robust reading competitions: entries, results, and future directions. Int. J. Document Anal. Recognit., 7(2-3):105–122, 2005. 8
framework for scene text recognition. In CVPR, pages 13525–13534, 2020.

[30] Joseph Redmon and Ali Farhadi. YOLO9000: Better, faster, stronger. In CVPR, pages 6517–6525, 2017.

[31] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell., 39(6):1137–1149, 2017.

[32] Anhar Risnumawan, Palaihnakote Shivakumara, Chee Seng Chan, and Chew Lim Tan. A robust arbitrary text detection system for natural scene images. Expert Syst. Appl., 41(18):8027–8048, 2014.

[33] Walter J. Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E. Boult. Toward open set recognition. IEEE Trans. Pattern Anal. Mach. Intell., 35(7):1757–1772, 2013.

[34] Baoguang Shi, Xiang Bai, and Cong Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE Trans. Pattern Anal. Mach. Intell., 39(11):2298–2304, 2017.

[35] Baoguang Shi, Mingkun Yang, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. ASTER: an attentional scene text recognizer with flexible rectification. IEEE Trans. Pattern Anal. Mach. Intell., 41(9):2035–2048, 2019.

[36] Baoguang Shi, Cong Yao, Minghui Liao, Mingkun Yang, Pei Xu, Linyan Cui, Serge J. Belongie, Shijian Lu, and Xiang Bai. ICDDR2017 competition on reading Chinese text in the wild (RCTW-17). In ICDAR, pages 1429–1434, 2017.

[37] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In ICLR, 2014.

[38] Yipeng Sun, Dimosthenis Karatzas, Chee Seng Chan, Lianwen Jin, Zihan Ni, Chee Kheng Chng, Yuliang Liu, Canjie Luo, Chun Chet Ng, Junyu Han, Errui Ding, and Jingtuo Liu. ICDAR 2019 competition on large-scale street view text with partial labeling (RRC-LSVT). In ICDAR, pages 1557–1562, 2019.

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In NeurIPS, pages 5998–6008, 2017.

[40] Zhaoyi Wan, Minghang He, Haoran Chen, Xiang Bai, and Cong Yao. Textscaler: Reading characters in order for robust scene text recognition. In AAAI, pages 12120–12127, 2020.

[41] Zhaoyi Wan, Jielei Zhang, Liang Zhang, Jiebo Luo, and Cong Yao. On vocabulary reliance in scene text recognition. In CVPR, pages 11422–11431, 2020.

[42] Kai Wang, Boris Babenko, and Serge J. Belongie. End-to-end scene text recognition. In ICCV, pages 1457–1464, 2011.

[43] Tianwei Wang, Zecheng Xie, Zhe Li, Lianwen Jin, and Xiangge Chen. Radical aggregation network for few-shot offline handwritten Chinese character recognition. Pattern Recognition Letters, 125:821–827, 2019.

[44] Tianwei Wang, Yuanzhi Zhu, Lianwen Jin, Canjie Luo, Xiaoxue Chen, Yaqiang Wu, Qianying Wang, and Mingxiang Cai. Decoupled attention network for text recognition. In AAAI, pages 12216–12224, 2020.

[45] Wenchao Wang, Jianshu Zhang, Jun Du, Zi-Rui Wang, and Yixing Zhu. Denseran for offline handwritten Chinese character recognition. In ICFHR, pages 104–109, 2018.

[46] Wenchao Wang, Jianshu Zhang, Jun Du, Zi-Rui Wang, and Yixing Zhu. Denseran for offline handwritten Chinese character recognition. In ICFHR, pages 104–109, 2018.

[47] M.P. Wellman and M. Henrion. Explaining ‘explaining away’. IEEE Trans. Pattern Anal. Mach. Intell., 15(3):287–292, 1993.

[48] Zecheng Xie, Xiaoxiao Huang, Yuanzhi Zhu, Lianwen Jin, Yuliang Liu, and Lele Xie. Aggregation cross-entropy for sequence recognition. In CVPR, pages 6538–6547, 2019.

[49] Deli Yu, Xuan Li, Chengquan Zhang, Tao Liu, Junyu Han, Jingtuo Liu, and Errui Ding. Towards accurate scene text recognition with semantic reasoning networks. In CVPR, pages 12110–12119, 2020.

[50] Tai-Ling Yuan, Zhe Zhu, Kun Xu, Cheng-Jun Li, Tai-Jiang Mu, and Shi-Min Hu. A large Chinese text dataset in the wild. J. Comput. Sci. Technol., 34(3):509–521, 2019.

[51] Fangneng Zhan and Shijian Lu. ESIR: end-to-end scene text recognition via iterative image rectification. In CVPR, pages 2059–2068, 2019.

[52] Chuhan Zhang, Ankush Gupta, and Andrew Zisserman. Adaptive text recognition through visual matching. In ECCV, pages 51–67, 2020.

[53] Chang-Bin Zhang, Peng-Tao Jiang, Qibin Hou, Yunchao Wei, Qi Han, Zhen Li, and Ming-Ming Cheng. Delving deep into label smoothing. IEEE Trans. Image Process., 30:5984–5996, 2021.

[54] Xu-Yao Zhang, Cheng-Lin Liu, and Ching-Y. Suen. Towards robust pattern recognition: A review. Proceedings of IEEE, 108(6):894–922, 2020.

[55] Yiwei Zhu, Shilin Wang, Zheng Huang, and Kai Chen. Text recognition in images based on transformer with hierarchical attention. In ICIP, pages 1945–1949, 2019.