E2E-MLT - an Unconstrained End-to-End Method for Multi-Language Scene Text

Yash Patel[1,2*] Michal Bušta[1*]
Jiri Matas[1]
[1]Center for Machine Perception, Department of Cybernetics
Czech Technical University, Prague, Czech Republic
[2] Robotics Institute, Carnegie Mellon University
yashp@andrew.cmu.edu, bustam@fel.cvut.cz, matas@cmp.felk.cvut.cz

Abstract

An end-to-end method for multi-language scene text localization, recognition and script identification is proposed. The approach is based on a set of convolutional neural nets.

The method, called E2E-MLT, achieves state-of-the-art performance for both joint localization and script identification in natural images and in cropped word script identification. E2E-MLT is the first published multi-language OCR for scene text. The experiments show that obtaining accurate multi-language multi-script annotations is a challenging problem.

1. Introduction

Scene text localization and recognition, a.k.a. photo OCR or text detection and recognition in the wild, is a challenging open computer vision problem. Applications of photo OCR are diverse, from helping the visually impaired to data mining of street-view-like images for information used in map services and geographic information systems. Scene text recognition find its use as a component in larger integrated systems such as those for autonomous driving, in-door navigations and visual search engines.

The growing cosmopolitan culture in modern cities often generates environments where multi-language text co-appears in the same scene (Fig. 1), triggering a demand for a unified multi-language scene text system. The need is also evident from the high interest in the ICDAR competition on multi-language text [31].

The recent advances in deep learning methods have helped in improving both the text localization [15, 29, 45, 25] and text recognition [19, 39, 6] methods significantly. However, these improvements are limited to English text and existing methods do not scale well to other languages.

Figure 1: Text in multiple languages appearing in a scene. The proposed E2E-MLT method localizes words, predicts the scripts and generates a text transcription for each bounding box.

Multi-language scene text has specific challenges. Firstly, the data currently publicly available for non-English scene text is insufficient for training deep architectures. Individual languages have specific challenges, for example, Chinese and Japanese have a high number of characters, Bangla scene text is mostly hand written.

In this paper, we introduce E2E-MLT, an end-to-end multi-language scene text multi-purpose method. E2E-MLT addresses multi-language scene text localization, script identification and text recognition. The method has been trained for the following languages: Arabic, Bangla, Chinese, Japanese, Korean, Latin. Its OCR is capable of recognizing 7,800 characters (compared to less than 100 in Latin [6, 39, 18, 35]).

The first major contribution of this paper is an efficient method for script identification which demon-

*These authors contributed equally to this work

strates state-of-the-art performance on the ICDAR RRC-MLT 2017 [31] dataset. The approach learns discriminative holistic representation for cropped-word images and emphasizes on preserving aspect-ratio at cropped-word level.

We further demonstrate that an existing detection framework [48] when combined with our script identification method achieves state-of-the-art performance on joint detection and script identification in natural scene images on the ICDAR RRC-MLT 2017 [31] data.

Our main contribution is an open dictionary multi-language text recognition method for natural scene images. Learning a multi-language OCR is challenging as characters from different scripts may not only co-appear within an image, but also within the same word. Our OCR does not require any language or script specific information. To our knowledge, we are first to present a unified OCR for multiple-languages.

As an auxiliary contribution, we publicly release two large scale synthetically generated datasets for training multi-language scene text detection and recognition methods.

2. Related Work

2.1. Scene Text Localization

Scene text localization is the first step in standard text-spotting pipelines. Given a natural scene image, objective is to obtain precise word level bounding boxes or segmentation maps.

Conventional methods such as [32, 33, 9] generally seek character candidates via extremal region extraction or edge detection. Deep learning based method [19] make use of a CNN [23] for image patches to predict text/no-text score, a character and a bi-gram class.

Jaderberg *et al.* [18] proposed a multi-staged method where horizontal bounding box proposals are obtained by aggregating the output of Edge Boxes [49] and Aggregate Channel Features [8], The proposals are filtered using a Random Forest [5] classifier. As post-processing a CNN regressor is used to obtain fine-grained bounding boxes. Gupta *et al.* [15] proposed a fully-convolutional regression network trained on synthetic data for performing detection and regression at multiple scales in an image.

Tian *et al.* [45] use a CNN-RNN joint model to predict the text/no-text score, the y-axis coordinates and the anchor side-refinement. A similar approach [25] adapts the SSD object detector [28] to detect horizontal bounding boxes. Ma *et al.* [29] detects text of different orientations by adapting the Faster-RCNN [11] architecture and adding 6 hand-crafted rotations and 3 aspects. A similar approach was presented in [6], where the rotation is a continuous parameter and optimal anchor box dimensions are found using clustering on training set.

As mentioned earlier, all of these methods deal with English text only. Multi-language method are described in ICDAR RRC-MLT 2017 [31]. *SCUT-DLVClab* separately trains two models, the first model predicts bounding box detections and second model classifies the detected bounding box into one of the script classes or background. *TH-DL* use a modified FCN with residual connections for generating text-proposals and a Fast-RCNN [11] for detection. GoogleLeNet architecture [44] is used for script-identification.

E2E-MLT makes use of the EAST [48] text-detector combined with the proposed script-identification and OCR methods.

2.2. Script Identification

The objective of script identification is to take the cropped word images and predict the script/language of the text at hand. Existing text recognition algorithms [20, 19, 39] are language-dependent which makes script identification a vital component for multi-language scene text understanding. Detecting the script and language at hand allows the existing methods to select the appropriate language model [46].

Script identification over complex background has been studied for video overlaid text [12, 42, 36, 37, 34]. However, these methods solve a different problem than scene text and rely highly on accurate edge detection of text components.

Methods for script identification in natural scene images have been presented in [13, 41, 38, 31]. Gomez *et al.* [13] uses ensembles of conjoined networks to learn representations of discriminative stroke-parts and their relative importance in a patch-based classification scheme. Similar to [13], E2E-MLT emphasize the need to preserve the aspect-ratio at the cropped word level. Unlike [13], E2E-MLT does not involve pre-processing, post-processing steps and learns holistic representations.

Within [31], *Synthetic-ECN* uses a method proposed in [13] along with synthetically generated data, *SCUT DLVC* uses a sliding window based approach, *TNet* uses a majority-voting mechanism to determine script class, *BLCT* proposed a complex pipeline making use of BoW along with CNN features, *TH-DL and TH-CNN* use GoogleLeNet [44] based features.

Unlike, most of these approaches, script identification in E2E-MLT does not involve multiple-steps.

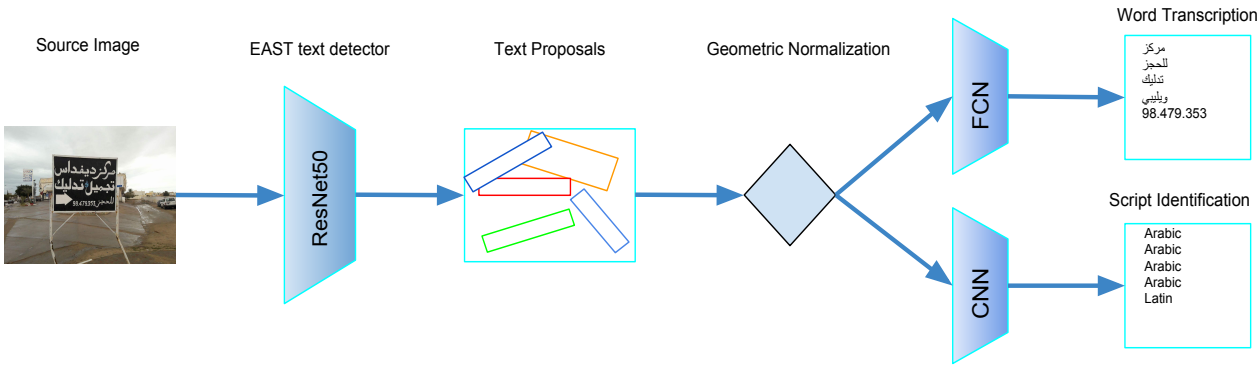


Figure 2: E2E-MLT overview : text proposals are generated and filtered by a EAST [48] based detector. Preserving the aspect ratio, each text proposal is then normalized to a fixed-height, variable-width tensor. Finally, each proposal is associated with a sequence of multi-language text and script class or is rejected as no-text.

2.3. Text Recognition

The objective of text recognition methods is to take the cropped word image and generate the transcription of the word present.

Scene text recognition has been widely studied for English text. Jaderberg *et al.* [18] train a CNN on 9 million synthetic images to classify a cropped word image as one of the words in a dictionary. The dictionary contains 90 000 English words and the words of the training and test set. Any word outside the dictionary is ignored.

Shi *et al.* [39] generates one sequence of characters per image by training a fully-convolutional network with a bidirectional LSTM using the Connectionist Temporal Classification (CTC) [14]. Unlike the OCR of proposed E2E-MLT, both [39, 18] resize the source cropped word image to a fixed-sized matrix of 100×32 pixels regardless of the number of characters present.

The aforementioned methods only deal with English text, where the number of characters is limited. Methods like [18, 19] approach the problem from close dictionary perspective where any word outside the dictionary is ignored. Such setting is not applicable to multi-language scenario where the number of characters and possible set of words are very high. Text recognition in E2E-MLT is open-dictionary and does not require language specific information.

3. Method

Given a natural scene image containing multi-language text, E2E obtain text localizations, generate text transcription and script class for each detected region. Overview of E2E-MLT is provided in Fig. 2.

3.1. Multi-Language Synthetic Data Generation

For Synthetic data generation, we adapt the framework proposed by Gupta *et al.* [15] to a multi-language setup. The framework generates realistic images by overlaying synthetic text over existing natural background images and it accounts for 3D scene geometry.

[15] proposed the following approach for scene-text image synthesis :

- Text in real-world usually appears in well-defined regions, which can be characterized by uniform color and texture. This is achieved by thresholding gPb-UCM contour hierarchies [2] using efficient graph-cut implementation [3]. This gives us prospective segmented regions for rendering text.
- Dense depth map of segmented regions is then obtained using [27] and then planer facet are fitted to them using RANSAC [10]. This way normals to the prospective regions for text rendering is estimated.
- Finally, the text is aligned to prospective image region for rendering. This is achieved by warping image region to frontal-parallel view using estimated region normals, then a rectangle is fitted to this region, then text is aligned to larger side of this rectangle.

Details of the experimental setup are presented in Section 4.1.

3.2. Multi-Language Text Localization

We make use of the EAST [48] text detector for text localization in a multi-language setup. EAST [48] follows the general design of DenseBox [17]. The image is fed into the fully convolutional network (FCN) and multiple levels of pixel-level text-scores and geometries are generated. As a post-processing step, threshold-

ing on text-scores and NMS on predicted geometries is performed.

ResNet50 [16] architecture is used in this paper. The model is trained with Adam optimizer [22] with a base learning rate of 0.001. The overall loss is given bt $L = L_{geo} + L_{dice}$ where L_{geo} is IoU loss proposed in [48] and L_{dice} is dice loss proposed in [30] The batch size is set to 16. The model is trained only on ICDAR MLT-RRC 2017 dataset [31] and convergence is observed after $70K$ iterations.

3.3. Holistic Cropped-Word Script Identification

We formulate script identification as an image classification problem at the cropped-word level. Our main idea is to preserve the aspect ratio of input cropped-word images during both training and testing, thus we resize the images into fixed-height ($H' = 64$ pixels) and variable width ($\overline{W} = \frac{wH'}{h}$) tensors.

Given a cropped-word image, our method learns a holistic representation. We make use of convolution layers form VGG-16 [43] with ImageNet [7] initialized weights along with Global-Average-Pooling [26] layer after the final convolution layer, followed by two fully-connected layers. Detailed description of architecture is in Tab. 1.

For learning script identification we minimize categorical-cross-entropy loss on cropped word image dataset of [31]. We use a Stochastic Gradient Descent (SGD) optimizer, with a base learning rate of 0.001, multiplied by 0.1 every 5 epochs, and momentum of 0.9. The batch size is set to 128. With these settings the network converges in 16 epochs.

Since the weight-initialization for convolutional layers is done using pre-trained VGG-16 [43] on ImageNet [7] data, the layers are tuned for objects and not text. Thus, we update both the convolution and fully-connected layers during back-propagation. Details of the dataset and experimental results are provided in Section 4.2.

3.4. Multi-Language Text Recognition

The proposed OCR in E2E-MLT works forArabic, Bangla, Chinese, Japanese, Korean and Latin. We adapt the OCR module presented in [6], and extend it for a multi-language setup. We select this model because of its simplicity, generality and relatively fast training time. Further, this model can be easily extended by standard tricks such as stacking LSTM modules [24], using models ensemble [47].

The E2E-MLT OCR is a fully-convolutional neural network, which takes a variable-width feature tensor $\overline{W} \times H' \times C$ as an input ($\overline{W} = \frac{wH'}{h}$) and outputs a matrix $\frac{\overline{W}}{4} \times |\hat{\mathcal{A}}|$, where $\mathcal{A}$ is the alphabet (all characters in set of languages $= 7800$ log-Softmax output). The full network definition is provided in Tab. 2.

Type	Size/Stride	Dim × Chn
input	-	$\overline{W} \times 64 \times 1$
Conv, ReLU ×2	3×3	$\overline{W} \times 64 \times 64$
MaxPool	$2 \times 2/2 \times 2$	$\overline{W}/2 \times 32 \times 64$
Conv, ReLU ×2	3×3	$\overline{W}/2 \times 32 \times 128$
Maxpool	$2 \times 2/2 \times 2$	$\overline{W}/4 \times 16 \times 128$
Conv, ReLU ×3	3×3	$\overline{W}/4 \times 16 \times 256$
Maxpool	$2 \times 2/2 \times 2$	$\overline{W}/8 \times 8 \times 256$
Conv, ReLU ×3	3×3	$\overline{W}/8 \times 8 \times 512$
Maxpool	$2 \times 2/2 \times 2$	$\overline{W}/16 \times 4 \times 512$
Conv, ReLU ×3	3×3	$\overline{W}/16 \times 4 \times 512$
Maxpool	$2 \times 2/2 \times 2$	$\overline{W}/32 \times 2 \times 512$
Global-avg-pool, Relu	-	512
Fully-con, Relu	512	512
Dropout (0.5)	-	-
Fully-con, Softmax	7	7

Table 1: E2E-MLT : Convolutional Neural Network for Script Identification.

Type	Chn.	Size/Stride	Dim
input	C	-	$\overline{W} \times 32$
C,Norm,ReLU	32	3×3	
C,Norm,ReLU	32	3×3	
maxpool		$2 \times 2/2$	$\overline{W}/2 \times 16$
C,Norm,ReLU	64	3×3	
C,Norm,ReLU × 2	64	3×3	
maxpool		$2 \times 2/2$	$\overline{W}/4 \times 8$
C,Norm,ReLU	128	3×3	
C,Norm,ReLU × 2	128	3×3	
maxpool		$2 \times 2/2 \times 1$	$\overline{W}/4 \times 4$
C,Norm,ReLU	256	3×3	
C,Norm,ReLU × 2	256	3×3	
maxpool		$2 \times 2/2 \times 1$	$\overline{W}/4 \times 2$
C,Norm,ReLU	512	3×3	
C,Norm,ReLU × 2	512	3×3	
Dropout (0.2)			
C	$\lvert\hat{\mathcal{A}}\rvert$	1×1	$\overline{W}/4 \times 1$
log softmax			

Table 2: E2E-MLT OCR: Convolutional Neural Network for Text Recognition.

The model is trained with Adam optimizer [22] (base learning rate = 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay = 0) and CTC loss function [14].

An union of the ICDAR RRC-MLT 2017 [31] train dataset, the ICDAR RCTW 2017 [40] train dataset and the Synthetic Multi-Language in Cropped Word Dataset for Text Recognition dataset 4.1 is used for training.

During all experiments we use greedy decoding of network output. Alternative could be the use of, task and language specific techniques such as prefix decoding or decoding with language models (for word spot-

ting task). However, our OCR is generalized, open-dictionary and language independent.

4. Experiments

4.1. Synthetic Datasets

The method for synthetic data generation is explained in Section 3.1. Two datasets were generated, both covering the same set of language classes as in ICDAR RRC-MLT 2017 [31] : Arabic, Bangla, Chinese, Japanese, Korean, Latin.

The **Synthetic Multi-Language in Natural Scene Dataset** for Text Detection contains text rendered over natural scene images selected from the set of $8,000$ background images collected by [15]. Annotations include word level and character level text bounding boxes along with the corresponding transcription. The dataset has $206,000$ images with thousands of images for each language. Sample examples are shown in Fig. 3.

The **Synthetic Multi-Language in Cropped Word Dataset** for Text Recognition (3.8M images) consists of cropped word images, language class and the corresponding text-transcriptions. The multi-language text is rendered over randomly generated plain backgrounds. Sample examples are shown in Fig. 4.

4.2. Script Identification on ICDAR RRC-MLT

The dataset [31] comprises $68,613$ training, $16,255$ validation and $97,619$ test image cut-out images and deals with the following 6 languages: Arabic, Latin, Chinese, Japanese, Korean, Bangla. Additionally, punctuation and some math symbols sometimes appear as separate words and are assigned a special script class called Symbols, hence 7 script classes are considered. The ground truth annotation provides the script class and the corresponding text transcription. The dataset has high class imbalance, Fig. 5 shows the number of images in each class.

The E2E-MLT method for script identification is described in Section 3.3. The training is done only on the real images provided by ICDAR MLT-RRC 2017 [31]. Tab. 3 shows that the E2E-MLT methods outperforms all entries

4.3. Joint Multi-Language Text Localization and Script Identification on ICDAR RRC-MLT

The dataset [31] comprises $7,200$ training, $1,800$ validation and $9,000$ testing natural scene images. The ground truth annotations includes bounding box coordinates, the script class and text-transcription.

Text localization and script identification approaches for E2E-MLT are explained in Section 3.2 and Section 3.3 respectively. As shown in Tab. 4, E2E-MLT achieves state-of-the-art performance on joint text localization and script identification on the ICDAR MLT-RRC 2017 [31] dataset.

Method	Accuracy
E2E-MLT	**88.54%**
CNN-based method	88.09%
SCUT DLVC	87.69%
BLCT	86.34%
TH-DL	80.72%
Synthetic-ECN [13]	79.20%
TNet	48.33 %
TH-CNN	43.22 %

Table 3: Script Identification accuracy on the ICDAR MLT-RRC 2017 test data [31].

Method	F-Measure	Recall	Precision
E2E-MLT	**58.69%**	**53.77%**	64.61%
SCUT-DLVClab2	58.08%	48.77%	**71.78%**
TH-DL	39.37%	29.65%	58.58%

Table 4: Joint text localization and script identification on the ICDAR RRC-MLT [31] test data.

4.4. Multi-Language Text Recognition

E2E-MLT approach for text recognition is explained in Section 3.4. First we run the analysis of scripts co-occurrence in individual images (Tab. 5) and scripts co-occurrence in words (Tab. 6). The script of the character is defined by Unicode table [1]. Each character has its unique name (for example character 'A' has unicode name 'Latin Capital Letter A' therefore its script is Latin). The scripts which occur in the ICDAR MLT 2017 dataset[31] are Latin (Lat), Arabic (Ara), Bengali (Beng), Hangul (Hang), CJK, Hiragana (Hir), Katakana (Kat) and Digit (Dig). The rest of characters are considered to be Symbols (Sym). The abbreviation CKH marks the group of CJK, Hiragana and Katakana scripts. Tab. 6 shows that script co-occurrence is non-trivial even on word level. The OCR module in practical application should satisfactorily handle at least the common combination of scripts of non-Latin script and Latin, Digit, and Symbols script.

OCR accuracy on cropped words of E2E-MLT is shown in Tab. 7 and the confusion matrix for individual script in Tab. 8. In this evaluation, the ground truth for a word is defined as the most frequent script. Tab. 8 shows that E2E-MLT does not make make many mistakes due to confusing script confusion. To confirm this observation, we conducted the following experiment the on ICDAR 2013 and

(a) Arabic Scene Text (b) Bangla Scene Text (c) Japanese Scene Text

Figure 3: Images from the Synthetic Multi-Language in Natural Scenes Dataset for Text Detection.

(a) Bangla (b) Korean (c) Japanese (d) Chinese (e) Arabic

Figure 4: Images from the Synthetic Multi-Language Cropped Word Dataset for Text Recognition.

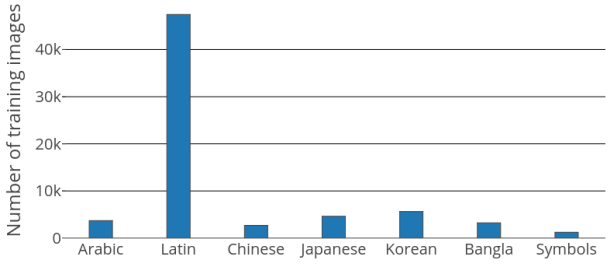


Figure 5: The number of training images for each class in ICDAR MLT-RRC 2017 [31] cropped word dataset.

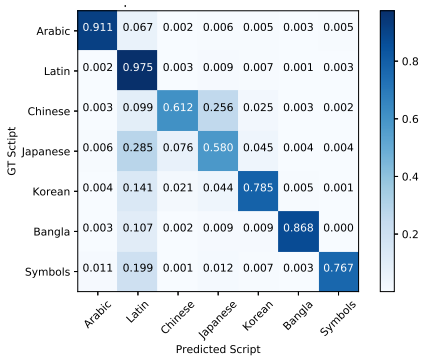


Figure 6: E2E-MLT confusion matrix for Script identification on the ICDAR MLT-RRC 2017 [31] test data.

	Sym	Dig	Lat	Ara	Beng	Hang	CJK	Hir	Kat
Sym	4361	2285	3264	400	482	652	903	378	312
Dig	2285	2838	2166	205	136	460	758	274	219
Lat	3264	2166	5047	501	150	443	876	299	258
Ara	400	205	501	797	0	0	0	0	0
Beng	482	136	150	0	795	0	0	0	0
Hang	652	460	443	0	0	847	81	32	28
CJK	903	758	876	0	0	81	1615	447	355
Hir	378	274	299	0	0	32	447	462	300
Kat	312	219	258	0	0	28	355	300	374

Table 5: Script co-occurrence on the ICDAR MLT-RRC 2017 [31] validation dataset. The row-column entry is incremented for all pairs of scripts present in an image.

	Sym	Dig	Lat	Ara	Beng	Hang	CJK	Hir	Kat
Lat	1046	635	52050	0	0	0	36	0	0
Ara	20	13	0	4881	0	0	0	0	0
Beng	63	5	0	0	3688	0	0	0	0
Hang	118	84	0	0	0	3767	0	0	0
CKH	416	499	21	0	0	0	7265	1424	1037

Table 6: Script co-occurrence in words in the ICDAR MLT-RRC 2017 validation dataset [31]. Column: script of a character. Row: script / script group of the word the character appeared in. If multiple scripts are present in the word, the row-column entry is incremented for each script.

ICDAR 2015 word recognition task [21]. We evaluated two variants of E2E-MLT: full-softmax (7800 classes) and a soft-max over a subset including Latin-Digit-Symbol (230 classess). Note that the latin subset we recognise is much wider than the English character set used in ICDAR tasks; we include the following: ßàáâãäåæçèéêëìíîïðñòóôõöøùúûüýþÿāăąćčďđēęěğħīŁł ńňōőŒœřŚśşštūůŷźżŽžfŭűğş ḥṃṇṛṣẽệùfffifl.

Script	Acc	$\frac{Edits}{len(GT)}$	Character Instances	Images
Symbol	0.442	0.506	890	530
Digit	0.685	0.189	6647	1862
Latin	0.756	0.101	52554	9285
Arabic	0.268	0.361	4892	951
Bengali	0.264	0.390	3752	673
Hangul	0.536	0.284	3839	1168
CJK	0.402	0.378	7392	1376
Hiragana	0.274	0.310	1663	230
Katakana	0.147	0.441	1010	177
Total	0.629	0.183	82639	?

Table 7: E2E-MLT OCR accuracy on the ICDAR MLT-RRC 2017 validation dataset [31]

	Sym	Dig	Lat	Ara	Beng	Hang	CJK	Hir	Kat
Sym	338	85	90	5	2	1	8	1	0
Dig	57	1695	87	9	4	1	6	1	2
Lat	172	92	8946	27	2	4	36	3	3
Ara	28	13	61	843	2	1	2	1	0
Beng	17	9	19	4	615	4	5	0	0
Hang	47	27	37	1	6	1013	34	2	1
CJK	47	13	14	0	0	3	1281	8	10
Hir	19	5	12	2	0	3	25	159	5
Kat	12	2	7	0	0	1	18	6	131

Table 8: Confusion matrix of E2E-MLT OCR on the ICDAR MLT-RRC 2017 [31] validation dataset. GT script is in row, the recognized script in columns

Method	TED case insensitive	C.W.R case insensitive
TencentAILab*	39.35	0.953
Baidu IDL*	57.53	0.899
...	...	...
E2E-MLT OCR Lat. Softmax	91.00	0.859
E2E-MLT OCR	95.33	0.850
PhotoOCR (2013) [4]	109.90	0.853

Table 9: E2E-MLT OCR evaluated on the ICDAR 2013 task 2 dataset [21]. Methods marked by an asterisk are unpublished.

Qualitative evaluation. The ICDAR MLT dataset includes a considerable number of vertical texts, see Tab. 11. Two types are present: text written in a horizontal direction – (a), (b), (c), (d), and text written in vertical direction – (e), (f). The first type can be detected by grouping detections as a post-processing step (see Tab. 16, first image, the vertical text is detected as single characters) and the second type just by reading rotated image.

Mistakes in the ground truth annotations add to the challenge, see Tab. 12. For Latin-script native readers, they are hard to identify. Another common source of error is caused by GT bounding boxes that are incorrectly axis-aligned bounding. Such images often contain more lines of text, confusing the recongiser assuming a single line, see Tab. 13.

Method	TED case insensitive	C.W.R case insensitive
Dahua OCR*	179.26	0.859
Baidu-IDL*	298.80	0.709
...	...	...
E2E-MLT OCR Lat. Softmax	633.00	0.606
E2E-MLT OCR	639.67	0.603
MAPS2015	1,068.72	0.339

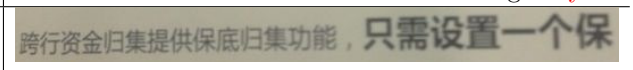
Table 10: E2E-MLT OCR evaluated on the ICDAR 2015 task 2 dataset [21], methods marked with asterisk are unpublished

(a) (b) (c) (d) (e) (f)

Table 11: Vertical text instances in the ICDAR MLT-RRC 2017 dataset [31]

GT	تمريات (a)	힝으로 (b)	기와이츠	(c) 베	(d) 4	(e)
Rec	تمريات	항으로	기와미츠	제	다	

GT	PUII	(f) 365원	(g) ALBERTO	(h) königsleig	(i)	
Rec	PULI	365 일	CRISTINA	Konigsbery		

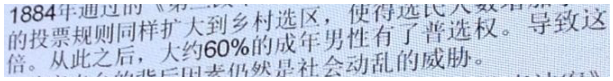


GT	跨行资金归集提供保底归集功能，missing text	(j)
Rec	烤行资金日集提供保定日能，只需设置一个保	

Table 12: Errors in the GT of the ICDAR MLT-RRC 2017 validation dataset [31]. Incorrect transcriptions are highlighted in red. Note that some errors lead to very large edit distances, e.g. in (f) and (h). GT errors effect both training and evaluation of the method. We estimate that at least 10% of errors on non-latin words reported for E2E-MLT on the ICDAR MLT dataset are due to GT mistakes.

4.5. End-to-End recognition

Quantitative evaluation of end-to-end recognition (localization and recognition) on the validation set of ICDAR MLT 2017 [31] data is shown in Tab. 15. Qual-



GT 倍。从此之后, 大约 60% 的成年男性有了普选权。..
Rec 自湖光程单用带头大 68 水城个加有心请连权导饮

Table 13: Multi-line text in ICDAR MLT-RRC 2017 [31] GT

Table 14: Difficult cases (for Latin-script native readers). Transcription errors, shown in red, which require close inspection - (a), (g), (h), (i). Note that for (i), the error is also in the ground truth. We were not able to establish a clear GT for (e) and (f). For (b), th transcription is 70004 in Bangla. In the context of Latin scripts, this same image will be interpreted as 900 08. Note the errors related to “:” and —, there are multiple types of colons and dashes in UTF.

Text Length	E2E Recall	Precision	E2E Recall ED1	Loc. Recall IoU 0.5
3+	0.319	0.420	0.407	0.613
2+	0.327	0.405	0.420	0.623

Table 15: E2E-MLT end-To-end recognition results on ICDAR MLT 2017 [31] validation set.

itative results are demonstrated in Tab. 16.

5. Conclusion

E2E-MLT, an end-to-end multi-language method, has been proposed. It achieves state-of-the-art performance for both joint localization and script identification in natural images and in cropped word script identification. E2E-MLT is the first published multi-language OCR for scene text. The implementation

Table 16: Example E2E-MLT results on the ICDAR MLT 2017 dataset[31]

along with trained models are publicly released here: https://github.com/yash0307/E2E-MLT.

The current mistakes in ground truth of ICDAR

RRC-MLT 2017 data [31] hurts both our training and evaluation. We will engage with ICDAR MLT competition organizers to improve quality of Ground Truth and Training data.

In our future work, we will train the proposed text localization and script identification approaches on the introduced synthetic datasets.

References

[1] Unicode table. `https://unicode-table.com/`. Accessed: 2017-09-30.

[2] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 2011.

[3] P. Arbeláez, J. Pont-Tuset, J. T. Barron, F. Marques, and J. Malik. Multiscale combinatorial grouping. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014.

[4] A. Bissacco, M. Cummins, Y. Netzer, and H. Neven. Photoocr: Reading text in uncontrolled conditions. In *Proceedings of the IEEE International Conference on Computer Vision*, 2013.

[5] A. Bosch, A. Zisserman, and X. Munoz. Image classification using random forests and ferns. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, 2007.

[6] M. Busta, L. Neumann, and J. Matas. Deep textspotter: An end-to-end trainable scene text localization and recognition framework. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*.

[8] P. Dollár, R. Appel, S. Belongie, and P. Perona. Fast feature pyramids for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014.

[9] B. Epshtein, E. Ofek, and Y. Wexler. Detecting text in natural scenes with stroke width transform. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, 2010.

[10] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 1981.

[11] R. Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2015.

[12] J. Gllavata and B. Freisleben. Script recognition in images with complex backgrounds. In *Signal Processing and Information Technology, 2005. Proceedings of the Fifth IEEE International Symposium on*, 2005.

[13] L. Gomez, A. Nicolaou, and D. Karatzas. Improving patch-based scene text script identification with ensembles of conjoined networks. *Pattern Recognition*, 2017.

[14] A. Graves, S. Fernández, F. Gomez, 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.

[15] A. Gupta, A. Vedaldi, and A. Zisserman. Synthetic data for text localisation in natural images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[16] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[17] L. Huang, Y. Yang, Y. Deng, and Y. Yu. Densebox: Unifying landmark localization with end to end object detection. *arXiv preprint arXiv:1509.04874*, 2015.

[18] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman. Reading text in the wild with convolutional neural networks. *International Journal of Computer Vision*, 2016.

[19] M. Jaderberg, A. Vedaldi, and A. Zisserman. Deep features for text spotting. In *European conference on computer vision*, 2014.

[20] M. Jain, M. Mathew, and C. Jawahar. Unconstrained scene text and video text recognition for arabic script. In *Arabic Script Analysis and Recognition (ASAR), 2017 1st International Workshop on*, 2017.

[21] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V. R. Chandrasekhar, S. Lu, et al. Icdar 2015 competition on robust reading. In *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*, 2015.

[22] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[23] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998.

[24] C.-Y. Lee and S. Osindero. Recursive recurrent nets with attention modeling for ocr in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2231–2239, 2016.

[25] M. Liao, B. Shi, X. Bai, X. Wang, and W. Liu. Textboxes: A fast text detector with a single deep neural network. In *AAAI*, 2017.

[26] M. Lin, Q. Chen, and S. Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013.

[27] F. Liu, C. Shen, and G. Lin. Deep convolutional neural fields for depth estimation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[28] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, 2016.

[29] J. Ma, W. Shao, H. Ye, L. Wang, H. Wang, Y. Zheng, and X. Xue. Arbitrary-oriented scene text detection via rotation proposals. *arXiv preprint arXiv:1703.01086*, 2017.

[30] F. Milletari, N. Navab, and S.-A. Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *3D Vision (3DV), 2016 Fourth International Conference on*, pages 565–571. IEEE, 2016.

[31] N. Nayef. Icdar2017 competition on multi-lingual scene text detection and script identification, 2017.

[32] L. Neumann and J. Matas. A method for text localization and recognition in real-world images. In *Asian Conference on Computer Vision*, 2010.

[33] L. Neumann and J. Matas. Real-time scene text localization and recognition. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2012.

[34] A. Nicolaou, A. D. Bagdanov, L. Gómez, and D. Karatzas. Visual script and language identification. In *Document Analysis Systems (DAS), 2016 12th IAPR Workshop on*, 2016.

[35] Y. Patel, L. Gomez, M. Rusiñol, and D. Karatzas. Dynamic lexicon generation for natural scene images. In *European Conference on Computer Vision*, 2016.

[36] T. Q. Phan, P. Shivakumara, Z. Ding, S. Lu, and C. L. Tan. Video script identification based on text lines. In *Document Analysis and Recognition (ICDAR), 2011 International Conference on*, 2011.

[37] N. Sharma, R. Mandal, R. Sharma, U. Pal, and M. Blumenstein. Bag-of-visual words for word-wise video script identification: A study. In *Neural Networks (IJCNN), 2015 International Joint Conference on*, 2015.

[38] B. Shi, X. Bai, and C. Yao. Script identification in the wild via discriminative convolutional neural network. *Pattern Recognition*, 2016.

[39] 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*, 2017.

[40] B. Shi, C. Yao, M. Liao, M. Yang, P. Xu, L. Cui, S. Belongie, S. Lu, and X. Bai. Icdar2017 competition on reading chinese text in the wild (rctw-17). *arXiv preprint arXiv:1708.09585*, 2017.

[41] B. Shi, C. Yao, C. Zhang, X. Guo, F. Huang, and X. Bai. Automatic script identification in the wild. In *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*, 2015.

[42] P. Shivakumara, Z. Yuan, D. Zhao, T. Lu, and C. L. Tan. New gradient-spatial-structural features for video script identification. *Computer Vision and Image Understanding*, 2015.

[43] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[44] 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.

[45] Z. Tian, W. Huang, T. He, P. He, and Y. Qiao. Detecting text in natural image with connectionist text proposal network. In *European Conference on Computer Vision*, 2016.

[46] R. Unnikrishnan and R. Smith. Combined script and page orientation estimation using the tesseract ocr engine. In *Proceedings of the International Workshop on Multilingual OCR*, 2009.

[47] C. Yang, X.-C. Yin, Z. Li, J. Wu, C. Guo, H. Wang, and L. Xiao. Adadnns: Adaptive ensemble of deep neural networks for scene text recognition. *arXiv preprint arXiv:1710.03425*, 2017.

[48] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Liang. East: An efficient and accurate scene text detector. *arXiv preprint arXiv:1704.03155*, 2017.

[49] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *European Conference on Computer Vision*, 2014.