A Dataset and Evaluation Metric for Coherent Text Recognition from Scene Images

Name of author

Address - Line 1
Address - Line 2
Address - Line 3

Abstract

In this paper, we deal with extraction of textual information from scene images. So far, the task of Scene Text Recognition (STR) has only been focusing on recognition of isolated words and, for simplicity, it omits words which are too short. Such an approach is not suitable for further processing of the extracted text. We define a new task which aims at extracting coherent blocks of text from scene images with regards to their future use in natural language processing tasks, mainly machine translation. For this task, we enriched the annotation of existing STR benchmarks in English and Czech and propose a string-based evaluation measure that highly correlates with human judgment.

Keywords: scene text recognition, machine translation

1. Introduction

Scene Text Recognition (STR) is a subfield of artificial intelligence that has been studied for a long time (Gómez and Karatzas, 2014) with a recent advances achieved by employing deep learning methods(Jaderberg et al., 2014b). With the increasing volume of pictures taken by hand-held devices, scene text (ST) became an interesting potential source of text for processing by Natural Language Processing (NLP) methods. Nevertheless, most of the previously published work strictly focus on recognition of isolated words and do not view the recognized words as utterances that belong into a particular language context. Another drawback of the state-of-the-art STR methods is that the benchmarks usually omit short (mostly function) words, for which they claim there is not enough visual evidence to be recognized. Most NLP methods usually deal with either text that can be split into sentences or directly with text on sentence level. ST on the other hand, consists of rather short chunks, such as proper names, isolated noun phrases or very short sentences. To machine-translate the ST we need to be able to recognize these chunks properly.

The only work mentioning Machine Translation (MT) of ST (Bijalwan and Aggarwal, 2014) we are aware of uses only simple rules for forming coherent text and pass the text to a statistical MT system. No systematic evaluation of the process is given. There exist few mobile applications for MT of ST[1] which very likely work similarly. As far as we know, no one approached this problem more systematically. In the next section, we briefly summarize the state of the art in ST localization and recognition. Section 3. brings a syntactic definition of connected text blocks and introduces a dataset for training automatic coherent text recognition from scene images by enriching existing STR benchmarks. In Section 4., we propose an automatic evaluation metric that allows fast comparison of methods.

[1] *Google Goggles* (`http://www.google.com/mobile/goggles`) and *Bing Translator* for Windows Phones (`http://www.bing.com/translator/phone/`) are the applications we know about.

Figure 1: Examples of ST images from the ICDAR Focused Scene Text Dataset.

2. Scene Text Localization and Recognition

Unlike the well-solved problem of optical character recognition, STR is a more challenging and still not satisfiably solved task. The text is usually placed on heterogeneous background with many distortions including shadows, reflections, and deformations (see the examples in Figure 1). Text extraction from scene images is usually divided into separate steps of text localization and recognition. Even though methods for unbounded recognition (Bissacco et al., 2013; Jaderberg et al., 2014a) exist, the recognition typically uses a limited vocabulary (Roy et al., 2014; Jaderberg et al., 2014b). The state-of-the-art methods are summarized, e.g., in the 2015 ICDAR Robust Reading Competition results (Karatzas et al., 2015).

3. Coherent Text Reading

Our goal is to indetify blocks of coherent text which can be further used in NLP tasks. We thus want to find *the minimum coherent text blocks closed on syntactic dependencies* (as perceived by an annotator, not automatically computed). Our original idea of the coherent text blocks was semantically motivated. We observed that ST frequently has a hierarchical nature. Signboards often contain lists of offered goods or services (coordinated on the same level), with a name of a venue as a kind of headline of the list on which the items depend. This hierarchy induces a natural order in which readers read the words. Annotating this would be

Figure 2: Example of an image with focused ST (left) and image with incidental ST (right). The word bounding boxes are highlighted by colorful boxes.

very laborious. Moreover, most of the hierarchies in the existing STR data are very flat, so such complex annotation would not pay off.

This is related to a problematic syntactic phenomena for the block definition which are coordinations indicated entirely by visual means where a coordination token is missing. Other ellipsis could be identified also on the pragmatic level (e.g., missing 'this shop offers:' on a signboard).

To avoid these problems we disregard all dependencies that are not explicitly present in the text. Unlike the standard STR benchmarks, we do not rely on the visual evidence only and also include cases where the text is obvious from the language context.

3.1. Original STR Data

The most frequently used benchmarks in STR come from the ICDAR Robust Reading Competition (Karatzas et al., 2015). For every competition, the annotation and evaluation protocol slightly differ. In the 2015 competition, all words in the images were localized in quadrilaterals and most of them were accompanied with a transcription. Words that are not readable or are shorter than 3 characters are marked as "not-care" words. *Focused ST* and *incidental ST* (see Figure 2 for examples) are distinguished as separate categories.

In the focused ST dataset, the main purpose of taking the pictures was the text. The pictures usually capture signboards and notices from an urban environment together with a few book covers and signs of electronics. The dataset consists of 229 training and 223 test images. On average, there are 6 words in an image out of which less than 3 are the "not-care" words. Most of the text is in English, with a few images containing signboards with a text in German.

The dataset of incidental ST consists of 1,000 training images and 500 test images taken in streets, shopping centers, and public transport of Singapore. The images capture complete urban scenes with a lot of text which often suffer from being out of focus and motion-blurred. There is, on

	pilot		final	
dataset	F	acc	F	acc
English focused	.820	.705	.943	.917
English incidental	.533	.190	—	
Czech focused	.853	.600	.962	.900

Table 1: Average inter-annotator agreement for both the pilot and final annotation.

average, 12 words in each image out of which 7 are "not-care" words. Most of the text is in English with some signs in non-Latin scripts which are localized but not transcribed.

The benchmarks only expect words from certain vocabularies to be recognized. For that purpose, sets of 50, 1k and 90k words are provided. Even though, the biggest lexicon may seem big enough for English, it may not be sufficient for languages with rich inflection or compounding. In addition, we use 81 images of Czech focused text (Hadáček, 2014) with 16 words per image.

Apart from the mentioned datasets, there exist other datasets worth mentioning. The *KAIST Scene Text Database* (Jung et al., 2011) consists of 3k images with focused texts in English and Korean. The *NEOCR* dataset (Nagy et al., 2012) is a set of 659 real world images with more than 13k words annotated on line level instead of word level.

3.2. Annotation Process

We annotated the coherence by explicitly marking chains of words in the images. Initially, we did a pilot annotation of 20 images from both ICDAR 2015 focused and incidental datasets and the Czech focused text dataset. Five annotators were provided with a simple definition of the task with little further details. They were asked to add transcription of "not-care" words if possible and to mark cases where a single word has been falsely split into multiple bounding boxes. An example of the annotation is in Figure 3.

We measured the inter-annotator agreement by *mutual accuracy* defined as a proportion of images that have been equally annotated and *mutual F-score* defined as a harmonic mean of the precision of the first annotator given the second one and vice versa. Values are tabulated in Table 3.

During the pilot annotation we experienced some problems with guessing the text. Different annotators set themselves different thresholds when they are certain about a word. The incidental text dataset was acquired in Singapore with a high density of shops. One annotator familiar with the fashion brands was able to transcribe much more signs than the others. Another annotator admitted he searched the Internet to find unreadable titles of books whose covers were in the dataset claiming that the image provided him with enough information to find out what the rest of the text is. The low agreement in the incidental dataset was mostly because the annotators were inconsistent in deciding what is readable in the images and what is not. With 10 seconds per word on average, the incidental text took more than twice as long as in the case of focused text annotation.

Based on the pilot annotation, we decided to only annotate the focused ST images. The annotation guidelines were refined to cover the most frequent inconsistent cases. These

image

STR annotation

```
35,120,390,120,390,222,35,222,HSBC
39,256,131,256,131,295,39,295,The
151,259,327,259,327,298,151,298,world's
350,261,461,261,461,300,350,300,local
484,262,602,262,602,301,484,301,bank
```

coherent text

```
1(The),2(world's),
3(local),4(bank);
0(HSBC)
```

Figure 3: An example of annotation of an image with focused ST. Numbers preceding the brackets refer to the line numbers in the original annotation, blocks are separated by semicolons.

were: a headline is a separate chunk; if a new line in the text is a substitute for a punctuation mark, it is a block separator; an address should be segmented as on an envelope; ignore characters which are not text (e.g., *P* for parking place); searching for additional knowledge is not allowed. An ex-post standardization was done on the annotation of rare punctuation (trade-marks, bullets, and vertical bars) increasing the mutual accuracy by 10 percentage points. The inter-annotator agreement on the final annotation is tabulated in Table 3.

In total, 81 images of Czech and 452 images of English focused ST were annotated. The images contain 3.6 blocks per image on average with the average length of 3.3 words. The images with the Czech focused text contain on average 4.9 blocks per image with the average length of 2.7 words. The dataset is relatively small. We expect the training part of the ICDAR Focused Scene Text can to be used for training postprocessing of the STR results. The Czech data and the test part of the ICDAR dataset will be used for testing.

4. Evaluation Metric

For training and comparing automatic methods for coherent text recognition, an automatic evaluation measure is needed. The standard STR evaluation metric (Karatzas et al., 2015) is a conjunction of the localization and string correctness. With coherent text recognition, we would like to have a measure that captures how comprehensible text would be if we did not have an access to the image. We believe we can disregard the text location and evaluate the transcription purely based on text similarity because for further text processing the text location does not matter at all. We first tried to explore the human perception of the recognition errors and based on that we designed an evaluation measure. We then explored different configurations of the measure and selected one that agreed the most with the human judgment.

4.1. Experiments

We asked annotators to evaluate erroneous transcriptions of the ST. It was done by three annotators who participated in the pilot annotation (were familiar with the task) but not in the main annotation (were not biased by already having seen images).

We generated two artificial erroneous transcriptions for each of the images that were previously unseen by the annotators. They were asked to imagine they are receiving the blocks in a random order and should translate them to a different language without seeing the image. Then they chose the one they think would lead to better translation.

error type	weight	
	human	machine
character insertion	12.8	3.6
character deletion	12.4	5.9
character substitution	12.8	6.0
block join	20.5	34.8
block split	24.0	34.7
block permutation	17.6	15.0

Table 2: Comparison of the estimated error weights for human annotators and the best fitting automatic measure.

The transcription errors were: character insertions, deletions, and substitutions, joining two blocks, splitting a block into two, permuting words within a block. The edit operations were sampled randomly from the distribution of edit changes obtained from running the TextSpotter STR tool (Neumann and Matas, 2012) on the same dataset. The annotators evaluated three different pairs of transcriptions for each image. One third of them was common for all annotators and was used to measure the inter-annotator agreement. The average-pairwise agreement was 0.670 with Cohen's kappa equal to 0.341.

To roughly estimate the importance of different error types for the annotators, we can view their decisions as a result of a linear combination of the error counts in each image. We do the estimation by fitting a logistic regression model. Normalized weights obtained from the model are tabulated in Table 2.

The model shows that the annotators consider joining or splitting blocks to be more serious errors than the character edit operations that all received similar weights.

4.2. Automatic Measure Description

Because the blocks can be recognized in a random order, we need to match the transcription and reference chunks before measuring their similarity. Expecting a reasonable quality of the underlying STR, we can match the reference with the recognition using entirely by the string similarity, disregarding their spacial position.

Formally, let $\mathbf{b} = (b_1, \ldots, b_n)$ be a machine-generated blocks, $\hat{\mathbf{b}} = (\hat{b}_1, \ldots, \hat{b}_m)$ its reference blocks and G a complete bipartite graph with $\mathbf{b}$ and $\hat{\mathbf{b}}$ its partite sets. The edges are weighted by string similarity $\mathrm{sim}(b_i, \hat{b}_j) \in [0, 1]$. A chunk matching $M \subset \mathbf{b} \times \hat{\mathbf{b}}$ is obtained as the minimum weighted maximum bipartite matching (Munkres, 1957) in the one-to-one case or as a minimum weighted edge cover (Schrijver, 2003) in case of many-to-many matching. We

	MM	MWEC
norm. Levensthein	.723	.722
Marzal-Vidal	**.726**	.724
Jaro-Winkler	.701	.715
PER	.645	.646
3-gram prec.	.685	.682
4-gram prec.	.696	.690
5-gram prec.	.696	.688

Table 3: Average agreement of the different configurations of thea automatic measure with the human judgment.

define the evaluation measure as:

$$m(\mathbf{b}, \hat{\mathbf{b}}) = \frac{\sum_{(b_1,b_2)\in M} 1 - \mathrm{sim}(b_1, b_2)}{\max(|\mathbf{b}|, |\hat{\mathbf{b}}|)} \qquad (1)$$

We explored the following string similarity measures: normalized Levenshtein distance, Marzal-Vidal distance (Marzal and Vidal, 1993), Jaro-Winkler distance (Winkler, 1990), position independent word error rate (Tillmann et al., 1997), and character n-gram precision as defined by Papineni et al. (2002). The last two measures are non-symmetric. Marzal-Vidal distance is the only one satisfying the triangular inequality, thus combined with the maximum matching algorithm, yields a distance metric.

4.3. Agreement with Human Judgment

For each pair of transcriptions presented to the annotators, we compute the similarity with the ground truth transcription. We measure the agreement as a proportion of cases when the annotator voted for the transcription with higher similarity score with the annotation. Surprisingly, the agreement with the automatic measures is higher that between the annotators themselves. It may be because the annotators must have picked randomly in cases when it was hardly distinguishable which transcription is better. The values are tabulated in Table 3.

The asymmetric measures lead to approximately the same agreement as the annotators reached with each other. A higher agreement was achieved by using the similarity measure that counts the edit operations which corresponds to the finding that the annotators attributed approximately the same weight to all character-level edit operations. The best underlying similarity measure is the Marzal-Vidal distance.

The best measure also appears to weight the importance of the error types more similarly to the human annotators (see Table 2), although it underestimates character edit operations.

5. Conclusions & Future Work

We introduced a task of coherent text recognition from scene images, enriched the existing STR benchmarks for this task, and proposed an automatic evaluation metric. Although the measure disregards the localization and is based entirely on text similarity, it achieves high agreement with human judgment. As a future work, we would like to machine-learn automatic procedures for this task.

6. References

Bijalwan, D. C. and Aggarwal, A. (2014). Automatic text recognition in natural scene and its translation into user defined language. In *PDGC 2014*, pages 324–329. IEEE.

Bissacco, A., Cummins, M., Netzer, Y., and Neven, H. (2013). PhotoOCR: Reading text in uncontrolled conditions. In *ICCV 2013*, pages 785–792. IEEE.

Gómez, L. and Karatzas, D. (2014). Scene text recognition: No country for old men? In *Computer Vision-ACCV 2014 Workshops*, pages 157–168. Springer.

Hadáček, J. (2014). Detection and recognition of diacritical and punctuation marks in real-world images. Master's thesis, Czech Technical University, Prague.

Jaderberg, M., Simonyan, K., Vedaldi, A., and Zisserman, A. (2014a). Deep structured output learning for unconstrained text recognition. *arXiv preprint arXiv:1412.5903*.

Jaderberg, M., Simonyan, K., Vedaldi, A., and Zisserman, A. (2014b). Synthetic data and artificial neural networks for natural scene text recognition. *arXiv preprint arXiv:1406.2227*.

Jung, J., Lee, S., Cho, M. S., and Kim, J. H. (2011). Touch TT: Scene text extractor using touchscreen interface. *ETRI Journal*, 33(1):78–88.

Karatzas, D., Gomez-Bigorda, L., Nicolaou, A., Ghosh, S., Bagdanov, A., Iwamura, M., Matas, J., Neumann, L., Chandrasekhar, V., Lu, S., Shafait, F., Uchida, S., and Valveny, E. (2015). ICDAR 2015 competition on robust reading. In *ICDAR 2015*, pages 1156–1160, Aug.

Marzal, A. and Vidal, E. (1993). Computation of normalized edit distance and applications. *IEEE PAMI*, 15(9):926–932.

Munkres, J. (1957). Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial and Applied Mathematics*, 5(1):32–38.

Nagy, R., Dicker, A., and Meyer-Wegener, K. (2012). NEOCR: A configurable dataset for natural image text recognition. In *Camera-Based Document Analysis and Recognition*, pages 150–163. Springer.

Neumann, L. and Matas, J. (2012). Real-time scene text localization and recognition. In *CVPR 2012*, pages 3538–3545, California, US. IEEE.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *ACL 2002*, pages 311–318. ACL.

Roy, U., Mishra, A., Alahari, K., and Jawahar, C. V. (2014). Scene text recognition and retrieval for large lexicons. In *ACCV 2014*.

Schrijver, A. (2003). *Combinatorial Optimization - Polyhedra and Efficiency*. Springer.

Tillmann, C., Vogel, S., Ney, H., Zubiaga, A., and Sawaf, H. (1997). Accelerated dp based search for statistical translation. In *Proceedings of the Fifth European Conference on Speech Communication and Technology*, page 2667–2670, Rhodes, Greece.

Winkler, W. E. (1990). String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage. In *Proceedings of the Section on Survey Research*, pages 354–359.