

Data Augmentation for Scene Text Recognition

Rowel Atienza

Electrical and Electronics Engineering Institute
 University of the Philippines

rowel@eee.upd.edu.ph

Abstract

Scene text recognition (STR) is a challenging task in computer vision due to the large number of possible text appearances in natural scenes. Most STR models rely on synthetic datasets for training since there are no sufficiently big and publicly available labelled real datasets. Since STR models are evaluated using real data, the mismatch between training and testing data distributions results into poor performance of models especially on challenging text that are affected by noise, artifacts, geometry, structure, etc. In this paper, we introduce STRAug which is made of 36 image augmentation functions designed for STR. Each function mimics certain text image properties that can be found in natural scenes, caused by camera sensors, or induced by signal processing operations but poorly represented in the training dataset. When applied to strong baseline models using RandAugment, STRAug significantly increases the overall absolute accuracy of STR models across regular and irregular test datasets by as much as 2.10% on Rosetta, 1.48% on R2AM, 1.30% on CRNN, 1.35% on RARE, 1.06% on TRBA and 0.89% on GCRNN. The diversity and simplicity of API provided by STRAug functions enable easy replication and validation of existing data augmentation methods for STR. STRAug is available at <https://github.com/roatienza/straug>.

1. Introduction

Humans use text to convey information through labels, signs, tags, logos, billboards and markers. For instance, a road sign with "Yield" informs drivers to wait for their turn. An "EXIT" sign points to the way going out. In the supermarket, sellers use labels or tags to inform buyers about the price and the quantity of products. Therefore, machines that read text in natural scenes can perform smarter decisions and actions.

The practical applications of scene text recognition (STR) have recently drawn interest in the computer vision community. Unfortunately, majority of the focus has always







Model	Input Text Image	Baseline Prediction	+STRAug Prediction	%Acc Gain
CRNN [33]		Pyice	Price	1.30
R2AM [22]		OLDTOW	OLDTOWN	1.48
GCRNN [36]		TiMmES	Times	0.89
Rosetta [5]		eizu	eBizu	2.10
RARE [34]		Washii1	Washing	1.35
TRBA [3]		insiid	inside	1.06

Figure 1. STRAug data augmentation significantly improves the overall accuracy of STR models especially on challenging input text images. We follow the evaluation protocol used in most STR models of case sensitive training and case insensitive validation.

been on refining model architecture and training algorithm to improve the text recognition performance. While there is nothing wrong with this, STR can also benefit from the improvement in data for training. In the absence of sufficiently large and publicly available labelled datasets, the advancement of STR relies on huge collections of automatically annotated synthetic text images for training such as MJSynth or Synth90k [18], SynthText [14], Verisimilar [43], and UnrealText [26]. Trained models are then evaluated on much smaller and fragmented real datasets such as IIIT5K (IIIT) [29], Street View Text (SVT) [37], ICDAR2003 (IC03) [27], ICDAR2013 (IC13) [21], ICDAR2015 (IC15) [20], SVT Perspective (SVTP) [31] and CUTE80 (CT) [32]. As a result, STR suffers from the typical problem of distribution shift from the training data to the evaluation data. STR models perform poorly especially in under represented or long tail samples similar to that can be found in the test data.

In deep learning, popular approaches to address distribution shift include domain adaptation, causal representation learning, regularization, information bottleneck and data augmentation. In this paper, our focus is on the relatively straightforward approach of data augmentation. In STR, data augmentation has not been rigorously explored.

Typical augmentation methods used include rotation, perspective and affine transformations, Gaussian noise, motion blur, resizing and padding, random or learned distortions, sharpening and cropping [39, 23, 24, 28, 1]. Proponents of STR methods select a subset of these augmentation methods to improve their models. To the best of our knowledge, there has been no comparative study on the performance of each of these methods. Furthermore, there are other possible text image augmentation functions that have not been used and fully explored in the existing STR literature.

In this paper, we attempt to formulate a library of data augmentation functions specifically designed for STR. While data augmentation algorithms are more developed in object recognition, they are not necessarily applicable in STR. For example, CutOut [11], CutMix [40] and MixUp [44] can easily remove one or more symbols from the text image resulting in the total loss of useful information for training. In object recognition, there is generally only one class to predict. In STR, there are one or more characters each occupying a small region in the image. Removing a region or mixing two images will annihilate one or more characters in the text image. The correct meaning of the text could be altered.

STRAug proposes 36 augmentation functions designed for STR. Each function has a simple API:

```
img = op(img, mag=mag, prob=prob).
```

An STRAug function `op()` transforms an image `img`, with magnitude `mag` and with probability `prob`. Each function has 3 levels or magnitudes of severity or intensity that can manifest in capturing text in natural images. In order to avoid the combinatorial explosion in evaluating the effect of each augmentation function, we propose 8 logical groups based on the nature, origin or impact of these methods. The 8 groups are: 1) *Warp*, 2) *Geometry*, 3) *Noise*, 4) *Blur*, 5) *Weather*, 6) *Camera*, 7) *Pattern* and 8) *Process*. Using RandAugment [10], we demonstrate overall significant positive increase in text recognition accuracy of baseline models on both regular and irregular text datasets as shown in Figure 1. The simplicity of API and the number of functions supported by STRAug enable us to easily replicate and validate other data augmentation algorithms.

2. Related Work

Scene text recognition (STR) is the challenging task of correctly reading a sequence of characters from natural images. STR models using deep learning [39, 3, 42, 5, 25] have superseded the performance of algorithms with hand-crafted features [30, 38]. Chen et al. [7] presented a comprehensive review and analysis of different STR methods. The problem with deep learning models is that they require a large amount of data to automatically learn features. For STR, there are no publicly available large labelled real datasets. Collecting and annotating huge amount of real text

data is a very costly and time consuming task. Thus, the advancement of STR relies on large synthetically generated and automatically annotated datasets.

Since STR models are evaluated on real, small and fragmented datasets, the bad side effects of data distribution shift are apparent especially on natural text images that are sometimes curved, noisy, distorted, blurry, under perspective transformation or rotated. In this paper, we believe that data augmentation can partially address the problem of data distribution shift in STR. Data augmentation automatically introduces certain transformations that can be found in the test datasets or natural scenes but under represented in the training datasets. Data augmentation can help in narrowing the gap between training and evaluation distributions.

To the best of our knowledge, there has been no comprehensive study and empirical evaluation on different data augmentation functions that are helpful for STR models in general. Luo et al. [28] proposed *Learn to Augment* to train an STR model to learn difficult text distortions. Experimental results on irregular text datasets such as ICDAR2015 (IC15) [20], SVT Perspective (SVTP) [31] and CUTE80 (CT) [32] demonstrated significant performance improvement. The disadvantage of *Learn to Augment* is it requires additional agent and augmentation networks that must be trained with the main STR network. This results to a more complex setup, additional 1.5M network parameters, difficult to reuse algorithm and a longer training time. Furthermore, *Learn to Augment* is only focused on distorted text, one of the many causes of data distribution shift in STR.

In the STR literature, data augmentation has been treated more of an after thought when proposing a new algorithm. Litman et al. [24] used random resizing and distortion to improve SCATTER. Yu et al. [39] applied random resizing plus padding, rotation, perspective transformation, motion blur and Gaussian noise to gain additional performance for its semantic reasoning network (SRN). Lee et al. [23] used random rotation with a normal distribution to train its transformer-based SATRN model to improve irregular text recognition. Du et al. [12] used a combination of techniques in *Learn to Augment* and SRN to improve PP-OCR. While there is a strong evidence of improvement in performance, there is a lack of focused study in this area. In this paper, we attempt to address this problem by designing 36 data augmentation functions, creating 8 logical groups, analyzing the effect of each group, and systematically combining all groups to maximize their overall positive impact.

3. Data Augmentation for STR

Text in natural scenes can be found in various unconstrained settings such as walls, shirts, car plates, book covers, signboards, product labels, price tags, road signs, markers, etc. The captured text images have many degrees of variation in font style, orientation, shape, size, color, ren-

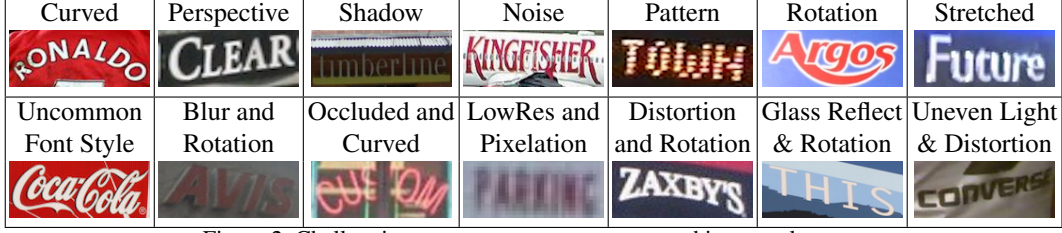


Figure 2. Challenging text appearances encountered in natural scenes

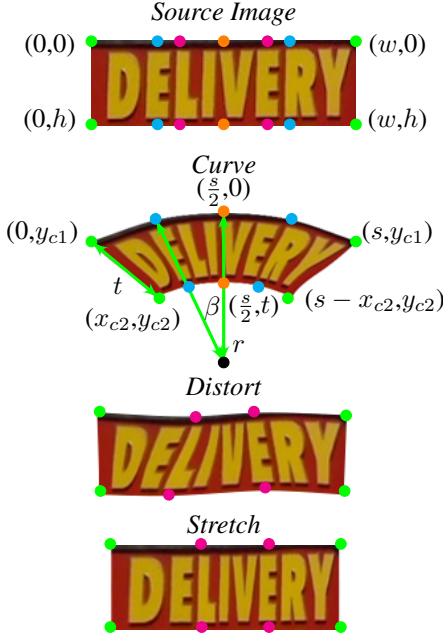


Figure 3. Source and destination control points used in TPS image warping transformation for *Curve*, *Distort* and *Stretch* data augmentation.

dering, texture and illumination. The images are also subjected to camera sensor orientation, location and imperfections causing image blur, pixelation, noise, and geometric and radial distortions. Weather disturbances such as glare, shadow, rain, snow and frost can also greatly affect the appearance of text. Figure 2 shows that real-world text appearances are challenging for machines to read. In fact, text images may be simultaneously altered by several factors. In the following, we discuss the 36 functions classified into 8 groups that attempt to mimic the issues in capturing text in natural scenes.

3.1. Warp

Curved, distorted and stretched text styles are found in natural scenes but are usually not well represented in train datasets. The *Warp* group includes *Curve*, *Distort* and *Stretch*. Curved text images are found in logos, seals, coins, product labels, emblems and tires. Distorted text can be

Source Pt	Destination Pt
<i>Curve</i>	
$(0, 0)$	$(0, y_{c1})$ s.t. $r = \text{rand}(r_{min}, r_{max})h$ $x_1 = (r^2 - \frac{s^2}{4})^{\frac{1}{2}}, y_{c1} = r - x_1$
$(\frac{s}{4}, 0)$	$(x_2, y_2) = (\frac{s}{2} - r \sin \beta, r(1 - \cos \beta))$ s.t. $\sin \beta = (\frac{1}{2} - \frac{x_1}{2r})^{\frac{1}{2}}, \cos \beta = (\frac{1}{2} + \frac{x_1}{2r})^{\frac{1}{2}}$
$(\frac{3s}{4}, 0)$	$(s - x_2, y_2)$
$(s, 0)$	(s, y_{c1})
$(\frac{s}{4}, s)$	$(x_3, y_3) =$ $(\frac{w}{2} - (r - t) \sin \beta, r - (r - t) \cos \beta)$
$(\frac{3s}{4}, s)$	$(s - x_3, y_3)$
$(\frac{s}{2}, 0)$	$(\frac{s}{2}, 0)$
$(\frac{s}{2}, s)$	$(\frac{s}{2}, t)$ s.t. $t = \frac{s}{2} \text{rand}(0.4, 0.5)$
$(0, s)$	(x_{c2}, y_{c2}) s.t. $x_{c2} = \frac{st}{2r}, y_{c2} = y_{c1} + \frac{tx_1}{r}$
(s, s)	$(s - x_{c2}, y_{c2})$
<i>Distort and Stretch</i>	
$(0, 0)$	$(\frac{w}{3} \text{rand}(0, k), \frac{h}{2} \text{rand}(0, k))$
$(\frac{w}{3}, 0)$	$(\frac{w}{3}(1 + \text{rand}(-k, k)), \frac{h}{2} \text{rand}(0, k))$
$(\frac{2w}{3}, 0)$	$(\frac{w}{3}(2 + \text{rand}(-k, k)), \frac{h}{2} \text{rand}(0, k))$
$(w, 0)$	$(w - \frac{w}{3} \text{rand}(0, k), \frac{h}{2} \text{rand}(0, k))$
$(0, h)$	$(\frac{w}{3} \text{rand}(0, k), h - \frac{h}{2} \text{rand}(0, k))$
$(\frac{w}{3}, h)$	$(\frac{w}{3}(1 + \text{rand}(-k, k)), h - \frac{h}{2} \text{rand}(0, k))$
$(\frac{2w}{3}, h)$	$(\frac{w}{3}(2 + \text{rand}(-k, k)), h - \frac{h}{2} \text{rand}(0, k))$
(w, h)	$(w - \frac{w}{3} \text{rand}(0, k), h - \frac{h}{2} \text{rand}(0, k))$

Table 1. Formula for source and destination control points used by TPS image warping. For *Curve*, the image is first resized to a square with side s before the TPS transformation is applied. Afterward, the image is returned to its original dimensions (w, h) . r decreases with the level of severity. *Distort* and *Stretch* share the same set of formula. For *Stretch*, the y coordinate is not randomized. k increases with the level of intensity.

seen on clothing, textiles, candy wrappers, thin plastic packaging and flags. Stretched text can be observed on elastic packaging materials and balloons. In this paper, we use stretch to refer to elastic deformation which may include its literal opposite word meaning, contract. Both *Distort* and *Stretch* are also used in certain artistic styles or can be caused by structural deformation and camera lens radial distortion (e.g. fish eye lens).

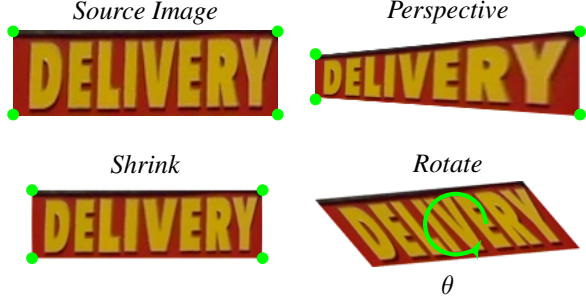


Figure 4. Image transformation under *Geometry* group.

Figure 3 shows the control points used in smooth deformation of a horizontal text image into *Curve*, *Distort* and *Stretch* versions. We use Thin-Plate-Spline (TPS) [4] to produce a warped version of the original image by moving pixels at source control points to their destination coordinates. All neighboring pixels around source control points are also re-positioned while following the smooth deformation constraints. With proper values of destination control points, various realistic deformations can be approximated such as curved, distorted and stretched. Table 1 lists the source and destination control points for our *Warp* data augmentation. An alternative algorithm to TPS is moving least squares as used in STR by Luo et al. [28].

In the *Curve* image warping, the text image is first resized to a square with side s . Then, random vertical flip is applied to get either concave or convex text shape. After the TPS smooth deformation, the upper half of the image is cropped since the lower half is covered by blank filler color. Then, the original image dimension is restored. We used 8 control points for warping. Two optional control points at the mid point of each side of the source image help improve the straightness of the edges. As the magnitude of augmentation increases, the radius of curvature r decreases. For *Distort* and *Stretch*, the extent of distortion k increases with the level of severity.

3.2. Geometry

When viewing natural scenes, perfect horizontal frontal alignment is seldom achieved. Almost always there is some degree of rotation and perspective transformation in the text image. Text may not also be perfectly centered. Translation along x and/or y coordinates is common. Furthermore, text can be found in varying sizes. To simulate these real-world scenarios, the *Geometry* group includes *Perspective*, *Shrink* and *Rotate* image transformations. Figure 4 shows these data augmentations while Table 2 lists the source and destination control points. For *Rotate*, there is only θ as the degree of freedom.

For *Perspective*, the horizon can be at the left or right side of the image. For simplicity, Figure 4 and Table 2 show

Source Pt	Destination Pt
<i>Perspective</i>	
$(0, 0)$	$(0, h \cdot \text{rand}(k, k + .1))$
$(w, 0)$	$(w, 0)$
$(0, h)$	$(0, h \cdot \text{rand}(0.9 - k, 1 - k))$
(w, h)	(w, h)
<i>Shrink</i>	
$(0, 0)$	$(\Delta w, \Delta h)$ $\Delta w = \frac{w}{3} \text{rand}(k, k + .1)$ $\Delta h = \frac{h}{2} \text{rand}(k, k + .1)$
$(w, 0)$	$(w - \Delta w, \Delta h)$
$(0, h)$	$(\Delta w, h - \Delta h)$
(w, h)	$(w - \Delta w, h - \Delta h)$
<i>Rotate</i> , $\theta = \text{rand}(\theta_{\min}, \theta_{\max})$	

Table 2. Formula for source and destination control points used in *Perspective* and *Shrink* data augmentations. k increases with the level of intensity. For *Rotate*, θ is sampled from a uniform distribution. θ_{\min} and θ_{\max} values increase with the magnitude of data augmentation.

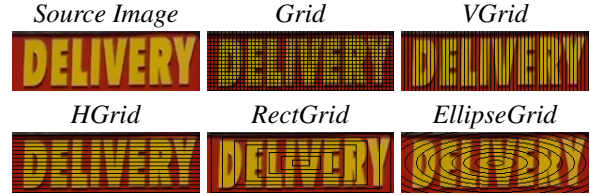


Figure 5. Examples of images affected by *Pattern* data augmentation.

left horizon only. For both *Perspective* and *Shrink*, k increases with the magnitude of augmentation. We also use TPS to perform *Shrink* deformation. As mentioned earlier, text may not be necessarily centered at all times. Therefore for *Shrink*, we randomly translate along horizontal or vertical axis. To avoid unintentional cropping of text symbols, the maximum horizontal translation is set to Δw (maximum of Δh for vertical translation).

For *Rotation*, θ is uniformly sampled from θ_{\min} to θ_{\max} . The magnitude of data augmentation increases with θ_{\min} and θ_{\max} . Clockwise rotation is supported by flipping the sign of θ with 50% probability.

3.3. Pattern

Regional dropout data augmentation methods such as CutOut [11], MixUp [44] and CutMix [40] are not suitable in STR since one or more symbols may be totally removed from the image. Inspired by GridMask [6], we designed 5 grid patterns that mask out certain regions from the image while ensuring that text symbols are still readable. For the *Pattern* group, we introduce 5 types of Grid: *Grid*, *VGrid*, *HGrid*, *RectGrid* and *EllipseGrid* as shown in Figure 5. The distance between grid lines decreases with the magnitude of data augmentation. Text with grid like appearance

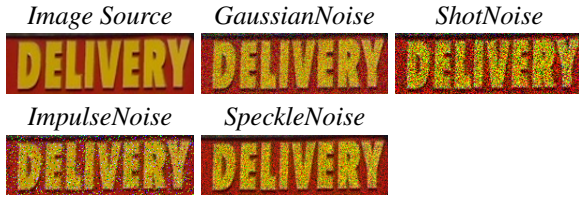


Figure 6. Examples of images affected by *Noise* data augmentation.



Figure 7. Example images affected by *Blur* data augmentation.

can be found in certain electronic displays and billboards, dot-matrix printer type of fonts and signs behind a meshed fence.

3.4. Noise

Noise is common in natural images. For STRAug, we lumped together different *Noise* types: 1) *GaussianNoise*, 2) *ShotNoise*, 3) *ImpulseNoise* and 4) *SpeckleNoise*. Figure 6 shows how each type of noise affects the text image. Gaussian noise manifests in low-lighting conditions. Shot noise or Poisson noise is electronic noise due to the discrete nature of light itself. Impulse noise is a color version of salt-and-pepper noise which can be caused by bit errors. For the *Noise* group, we adopted the implementation by Hendrycks and Dietterich [17] but using only half of the levels in order to ensure that the text in the image is still human readable. The amount of noise corruption increases with the level of severity of data augmentation.

3.5. Blur

Similar to noise, blur is common in natural images. Blur may be caused by unstable camera sensor, dirty lens, relative motion between the camera and the subject, insufficient illumination, out of focus settings, imaging while zooming, subject behind a frosted glass window, or shallow depth of field. The *Blur* group includes: 1) *GaussianBlur*, 2) *DefocusBlur*, 3) *MotionBlur*, 4) *GlassBlur* and 5) *ZoomBlur*. Figure 7 shows resulting images due to *Blur* functions. The degree of blurring increases with the level of severity of data augmentation. Except for *GaussianBlur*, we adopted the implementation by Hendrycks and Dietterich [17].

3.6. Weather

Scene text may be captured under different weather conditions. As such, we simulate these conditions under



Figure 8. Example images affected by *Weather* data augmentation.



Figure 9. Example images affected by *Camera* data augmentation.

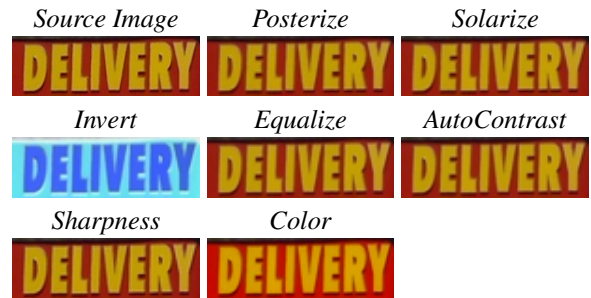


Figure 10. Example images affected by *Process* data augmentation.

Weather group: 1) *Fog*, 2) *Snow*, 3) *Frost*, 4) *Rain* and 5) *Shadow*. Figure 8 shows how a text image is affected by different weather conditions. As the magnitude of data augmentation increases, the severity of weather condition is increased. For example, as the magnitude of augmentation increases, the number of rain drops increases or the opacity of the shadow increases. The weather conditions around the world are extremely varied that it may not be possible to cover all possible scenarios. *Weather* simulates some common conditions only.

3.7. Camera

Camera sensors have many imperfections and tunable settings. These are grouped under *Camera*: 1) *Contrast*, 2) *Brightness*, 3) *JpegCompression* and 4) *Pixelate*. *Contrast* enables us to distinguish the different objects that compose an image. This could be the text against background and other artifacts. *Brightness* is directly affected by scene luminance. *JpegCompression* is the side effect of image compression. *Pixelate* is exhibited by increasing the resolution of an image. The severity of camera effect increases with the level of data augmentation. Figure 9 illustrates the effect of *Camera* data augmentation.

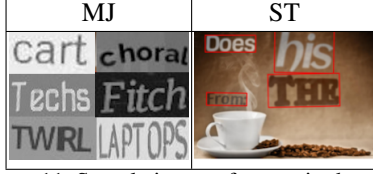


Figure 11. Sample images from train datasets.

3.8. Process

All other image transformations used in object recognition data augmentation literature that may be applicable in STR are grouped together in *Process*: 1) *Posterize*, 2) *Solarize*, 3) *Invert*, 4) *Equalize*, 5) *AutoContrast*, 6) *Sharpness* and 7) *Color*. Figure 10 demonstrates the effect of *Process*. These functions were used in AutoAugment [9] and can help STR models learn invariant features of text in images. The functions are image processing routines that change the image appearance but not the readability of the text. This is done through bit-wise or color manipulation. For example, *Invert* and *Color* can drastically change the color of the image but the readability of the text remains. *Invert*, *AutoContrast* and *Equalize* support 1 level of intensity only.

4. Experimental Results and Discussion

We evaluated the impact of STRAug on different strong baseline STR methods using the framework developed by Baek et al. [3]. We first describe the train and test datasets. Then, we present and analyze the empirical results.

4.1. Train Dataset

The framework uses 1) MJSynth (MJ) [18] or also known as Synth90k and 2) SynthText (ST) [14] to train STR models. Figure 11 shows sample images from MJ and ST. We provide a short description of the two datasets.

MJSynth (MJ) is a synthetically generated dataset made of 8.9M realistically looking word images. MJSynth was designed to have 3 layers: 1) background, 2) foreground and 3) optional shadow/border. It uses 1,400 different fonts, different background effects, border/shadow rendering, base colors, projective distortions, natural image blending and noise.

SynthText (ST) is a synthetically generated dataset made of 5.5M word images. SynthText blends synthetic text on natural images. It uses the scene geometry, texture, and surface normal to naturally blend and distort a text rendering on the surface of an object. The text is then cropped from the modified natural image.

4.2. Test Dataset

The test dataset is made of several small publicly available STR datasets of text in natural images. These datasets

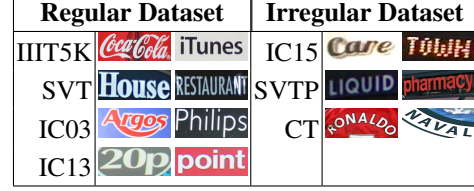


Figure 12. Sample text images from test datasets.

Table 3. Train conditions

Train dataset: 50%MJ + 50%ST	Batch size: 192
Iterations: 300,000	Parameter init: He [15]
Optimizer: Adadelata [41]	Learning rate: 1.0
Adadelata ρ: 0.95	Adadelata ϵ: $1e^{-8}$
Loss: Cross-Entropy/CTC	Gradient clipping: 5.0
Image size: 100×32	Channels: 1 (grayscale)

are generally grouped into two: 1) Regular and 2) Irregular.

The regular datasets have text images that are mostly frontal with a minimal amount of rotation or perspective distortion. IIIT5K-Words (IIIT) [29], Street View Text (SVT) [37], ICDAR2003 (IC03) [27] and ICDAR2013 (IC13) [21] are considered regular datasets. **IIIT5K** has 3,000 test images. These images are mostly from street scenes such as sign boards, brand logos, house number or street signs. **SVT** has 647 test images. The text images are cropped from Google Street View images. **IC03** has 1,110 test images from ICDAR2003 Robust Reading Competition. Images were captured from natural scenes. Both versions, 860 and 867 test images, are used. **IC13** is an extension of IC03 and shares similar images. IC13 was created for the ICDAR2013 Robust Reading Competition. Both versions, 857 and 1,015 test images, are used.

Meanwhile, irregular datasets are made of text with challenging appearances such as curved, vertical, under perspective transformation, low-resolution or distorted. IC-DAR2015 (IC15) [20], SVT Perspective (SVTP) [31] and CUTE80 (CT) [32] belong to irregular datasets. **IC15** has text images from the ICDAR2015 Robust Reading Competition. Many images are blurry, noisy, rotated, and sometimes of low-resolution, perspective-shifted, vertical and curved. Both versions, 1,811 and 2,077 test images, are used. **SVTP** has 645 test images from Google Street View. Most are images of business signage. **CT** focuses on curved text images captured from shirts and product logos. The dataset has 288 test images.

Figure 12 shows samples from both regular and irregular datasets. For both datasets, only the test splits are used in evaluating STR models.

Table 4. Individual group absolute percent gain in accuracy for the RARE [34] model.

Group	IIIT 3,000	SVT 647	IC03 860 867		IC13 857 1,015		IC15 1,811 2,077		SVTP 645	CT 288	Acc %
<i>Warp</i>	0.34	0.03	0.25	0.21	-0.12	0.08	0.78	0.51	-0.13	0.52	0.33
<i>Blur</i>	-0.08	0.21	0.54	0.37	-0.10	-0.10	2.05	1.76	0.18	0.29	0.67
<i>Noise</i>	-0.31	0.53	0.56	0.63	0.12	0.00	1.49	1.29	1.10	-1.38	0.52
<i>Geometry</i>	0.17	-0.18	0.10	0.13	0.39	0.33	0.99	0.87	-0.13	0.41	0.41
<i>Camera</i>	-0.09	-0.18	0.03	0.03	-0.23	-0.06	1.16	0.93	0.26	-2.56	0.24
<i>Weather</i>	0.26	0.41	0.10	0.33	-0.54	-0.47	1.03	0.80	0.98	-0.13	0.38
<i>Pattern</i>	-0.35	-0.95	0.00	-0.07	-0.33	-0.53	0.87	0.72	-0.45	-0.34	0.01
<i>Process</i>	-0.48	-0.57	0.27	0.18	-0.12	-0.07	1.05	0.76	0.31	0.64	0.19

4.3. Experimental Setup

The training configurations used in the framework are summarized in Table 3. We reproduced the results of 6 strong baseline models: CRNN [33], R2AM [22], GCRNN [36], Rosetta [5], RARE [34] and TRBA [3]. Each model is differentiated by 4 stages [3]: 1) *Image Rectification*: TPS [19] or None, 2) *Feature Extractor*: VGG [35], ResNet [16] or RCNN [22], 3) *Sequence Modelling*: BiLSTM [33] or None, and 4) *Prediction*: Attention [34, 8] or CTC [13]. We trained all models from scratch for at least 5 times using different random seeds. The best performing weights on the test datasets are saved to get the mean evaluation scores.

4.4. Individual Group Performance

After establishing the baseline scores, each STRAug group was used as a data augmentation method during training in order to understand the individual gain in accuracy. We performed an ablation study using the RARE model since it is the smallest model with all 4 stages present. Data augmentation is randomly applied with 50% probability. The magnitude of data augmentation is randomly drawn from (0, 1, 2). Table 4 shows that the biggest gain of 0.67% in absolute accuracy is from *Blur*, followed by *Noise* 0.52%, *Geometry* 0.41%, *Weather* 0.38%, *Warp* 0.33%, *Camera* 0.24%, *Process* 0.19% and *Pattern* 0.01%. *Blur* has the biggest gain on IC15 since the dataset has a substantial number of low resolution, low-light and blurry images. As expected, *Warp* improves curved text that can be found in CT. Both *Warp* and *Geometry* improved the model performance on IC15 and IC03. Both datasets have rotated and distorted text. Surprisingly, SVTP did not improve with *Warp* and *Geometry*. We believe that while SVTP is from Google Street View, the amount of perspective distortion and rotation is not that significant unlike in IC15 and IC13. *Noise* has performance gains across all datasets except for CT and IIIT. These two datasets are dominated by clean text images. *Weather* has accuracy gains on both IC15 and SVTP. Both datasets have a substantial number of outdoor scenes. *Process* improved the model performance on IC15, SVTP, CT and IC13. These datasets are characterized by highly varied color and texture. Overall, *Pattern* has a neg-

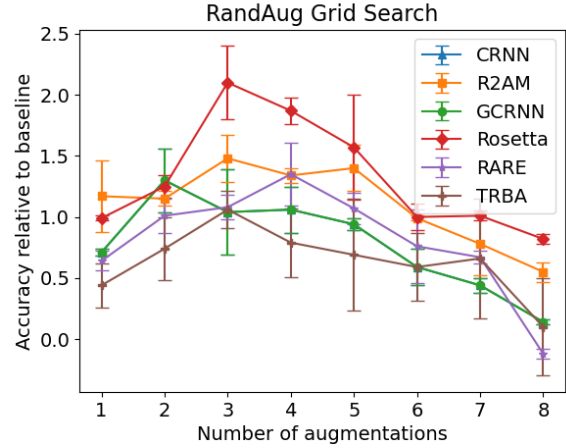


Figure 13. Overall absolute percentage accuracy increase versus the number of data augmentations when STRAug is used in different baseline STR models.

ligible positive impact but it has a significant contribution in IC15 which contains text images with patterns. An example is the word TOWN in Figure 12.

4.5. Combined Group Performance

The individual group performance gains may not appear impressive. However, combining all groups significantly pushes the accuracy higher. We used RandAugment [10] as a policy to randomly select N data augmentation groups each with a random magnitude M to apply during training. Unlike AutoAugment [9], RandAugment has a comparable performance and is easy to optimize using a simple grid search. Figure 13 shows the absolute percent accuracy gain versus N for different baseline models. Unlike in object recognition where the accuracy increases with the number of augmentations, in STR the peak is between $N = 2$ and $N = 4$. Table 5 shows the gains per dataset. When compared with the results from applying data augmentation on an individual group basis, the gains using a mixture of groups are significantly higher. For example, in evaluating CT with the RARE model the highest group gain is only 0.64% while it is 2.95% or about $4.6\times$ for the combined.

Table 5. Reproduced baseline scores and absolute accuracy increase using different data augmentation methods.

Model	IIT	SVT	IC03		IC13		IC15		SVTP	CT	Acc %
+Augmentation	3,000	647	860	867	857	1,015	1,811	2,077	645	288	
CRNN [33]	81.56	80.26	91.69	91.41	89.35	88.30	65.64	60.80	65.85	61.63	76.76
+SRN Aug	0.71	-0.23	0.26	0.26	0.03	0.02	2.18	1.81	1.40	2.69	0.98
+PP-OCR Aug	0.37	-0.14	0.29	0.25	-0.01	0.04	2.11	1.83	0.56	3.99	0.88
+STRAug (Ours)	0.63	0.37	0.79	0.87	0.22	0.24	2.68	2.34	2.00	2.26	1.30
R2AM [22]	83.28	80.95	91.69	91.38	90.17	88.15	68.48	63.35	70.50	64.67	78.46
+SRN Aug	0.81	2.72	0.99	1.01	0.92	0.68	1.46	1.49	1.80	2.11	1.24
+PP-OCR Aug	0.84	1.17	0.99	0.82	1.04	1.04	1.90	1.78	1.49	4.08	1.33
+STRAug (Ours)	0.84	2.92	0.60	0.59	0.73	0.62	2.34	2.10	2.67	3.15	1.48
GCRNN [36]	82.89	81.14	92.67	92.31	89.97	88.37	68.12	62.94	68.48	65.51	78.30
+SRN Aug	0.54	0.00	-0.35	-0.42	-0.08	0.53	0.35	0.42	0.21	2.20	0.31
+PP-OCR Aug	0.34	0.00	-0.23	-0.15	-0.39	0.16	0.96	0.91	1.40	3.13	0.49
+STRAug (Ours)	0.52	-0.41	-0.12	0.04	0.31	0.79	1.73	1.62	2.22	1.62	0.89
Rosetta [5]	82.59	82.60	92.60	91.97	90.32	88.79	68.15	62.95	70.02	65.76	78.43
+SRN Aug	1.06	0.14	0.46	0.49	-0.12	0.05	2.58	2.44	1.10	5.76	1.34
+PP-OCR Aug	2.14	0.97	0.69	0.76	0.39	0.28	2.58	2.44	1.46	4.72	1.74
+STRAug (Ours)	2.15	1.58	0.47	0.72	0.61	0.69	3.47	3.30	1.61	5.07	2.10
RARE [34]	85.95	85.19	93.51	93.33	92.30	91.03	73.94	68.42	75.58	70.54	82.12
+SRN Aug	0.18	0.34	0.91	0.87	0.62	0.59	2.04	1.74	0.96	1.22	0.97
+PP-OCR Aug	0.16	0.95	0.10	-0.13	0.51	0.62	2.43	2.16	2.14	2.49	1.09
+STRAug (Ours)	0.75	1.31	0.83	0.67	0.62	0.72	2.43	2.06	1.63	2.95	1.35
TRBA [3]	87.71	87.44	94.54	94.20	93.38	92.14	77.32	71.62	78.14	75.52	84.29
+SRN Aug	0.85	0.72	-0.02	0.21	0.16	0.31	2.00	1.82	1.74	0.80	1.02
+PP-OCR Aug	0.70	0.23	0.17	0.00	0.32	0.54	2.03	1.83	2.02	2.08	1.04
+STRAug (Ours)	1.23	0.58	0.35	0.52	0.64	0.69	1.35	1.08	1.94	2.60	1.06
ViTSTR-Tiny [2]	83.7	83.2	92.8	92.5	90.8	89.3	72.0	66.4	74.5	65.0	80.3
ViTSTR-Tiny+STRAug	85.1	85.0	93.4	93.2	90.9	89.7	74.7	68.9	78.3	74.2	82.1
ViTSTR-Small	85.6	85.3	93.9	93.6	91.7	90.6	75.3	69.5	78.1	71.3	82.6
ViTSTR-Small+STRAug	86.6	87.3	94.2	94.2	92.1	91.2	77.9	71.7	81.4	77.9	84.2
ViTSTR-Base	86.9	87.2	93.8	93.4	92.1	91.3	76.8	71.1	80.0	74.7	83.7
ViTSTR-Base+STRAug	88.4	87.7	94.7	94.3	93.2	92.4	78.5	72.6	81.8	81.3	85.2

Given the library of STRAug functions, it is easy to implement and validate other data augmentation algorithms. For example, SRN [39] data augmentation which is made of random resizing plus padding, rotation, perspective transformation, motion blur and Gaussian noise can be formulated as:

```

geometry = [Rotate(), Perspective(), Shrink()]
noise = [GaussianNoise()]
blur = [MotionBlur()]
augmentations = [geometry, noise, blur]
img = RandAugment(img, augmentations, N=3)

```

Similarly, using STRAug functions, we can easily implement and validate PP-OCR [12] data augmentation. PP-OCR uses the combined methods of SRN and *Learn to Augment* (i.e. random distortion). Note that the main difference of our implementation of SRN and PP-OCR data augmentations is that we further fine tuned both methods using RandAugment to maximize their potential.

Table 5 presents a comparison of the absolute accuracy increase due to SRN, PP-OCR and STRAug data augmentation techniques in the baseline models. The increase is significant especially on challenging irregular datasets such

as CT (1.62%-5.07%), SVTP (1.61%-2.67%) and IC15 (1.08%-3.47%). Figure 1 shows example text images that baseline models made correct predictions when trained with STRAug. Given that STRAug is using a more diverse set of data augmentation functions, it outperforms recent STR data augmentation methods. STRAug is also an effective regularizer on a vision transformer-based STR such ViTSTR[2]. Table 5 shows substantial gains in performance on all sizes, Tiny: +1.8%, Small: +1.6% and Base: +1.5%.

5. Conclusion

STRAug is a library of diverse 36 STR data augmentation functions with a simple API. The empirical results showed that a significant accuracy gain can be obtained using STRAug.

6. Acknowledgement

This work was funded by the University of the Philippines ECWRG 2019-2020. Thanks to CNL people: Roel Ocampo and Vladimir Zurbano, for hosting our servers.

References

- [1] Aviad Aberdam, Ron Litman, Shahar Tsiper, Oron Anshel, Ron Slossberg, Shai Mazor, R Manmatha, and Pietro Perona. Sequence-to-sequence contrastive learning for text recognition. *arXiv preprint arXiv:2012.10873*, 2020. 2
- [2] Rowel Atienza. Vision transformer for fast and efficient scene text recognition. In *International Conference on Document Analysis and Recognition (ICDAR)*, 2021. 8
- [3] Jeonghun Baek, Geewook Kim, Junyeop 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 4715–4723, 2019. 1, 2, 6, 7, 8
- [4] Fred L. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *Trans on Pattern Analysis and Machine Intelligence*, 11(6):567–585, 1989. 4
- [5] Fedor Borisjuk, Albert Gordo, and Viswanath Sivakumar. Rosetta: Large scale system for text detection and recognition in images. In *Intl Conf on Knowledge Discovery & Data Mining*, pages 71–79, 2018. 1, 2, 7, 8
- [6] Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Gridmask data augmentation. *arXiv preprint arXiv:2001.04086*, 2020. 4
- [7] Xiaoxue Chen, Lianwen Jin, Yuanzhi Zhu, Canjie Luo, and Tianwei Wang. Text recognition in the wild: A survey. *arXiv preprint arXiv:2005.03492*, 2020. 2
- [8] Zhanzhan Cheng, Fan Bai, Yunlu Xu, Gang Zheng, Shiliang Pu, and Shuigeng Zhou. Focusing attention: Towards accurate text recognition in natural images. In *ICCV*, pages 5076–5084, 2017. 7
- [9] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *CVPR*, pages 113–123, 2019. 6, 7
- [10] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *CVPR Workshops*, pages 702–703, 2020. 2, 7
- [11] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 2, 4
- [12] Yuning Du, Chenxia Li, Ruoyu Guo, Xiaoting Yin, Weiwei Liu, Jun Zhou, Yifan Bai, Zilin Yu, Yehua Yang, Qingqing Dang, et al. Pp-ocr: A practical ultra lightweight ocr system. *arXiv preprint arXiv:2009.09941*, 2020. 2, 8
- [13] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *ICML*, pages 369–376, 2006. 7
- [14] Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. Synthetic data for text localisation in natural images. In *CVPR*, pages 2315–2324, 2016. 1, 6
- [15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, pages 1026–1034, 2015. 6
- [16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 7
- [17] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *ICLR*, 2018. 5
- [18] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Synthetic data and artificial neural networks for natural scene text recognition. *NIPS Workshop on Deep Learning*, 2014. 1, 6
- [19] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In *NeurIPS*, pages 2017–2025, 2015. 7
- [20] Dimosthenis Karatzas, Lluís Gomez-Bigorda, Angelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, et al. Icdar 2015 competition on robust reading. In *ICDAR*, pages 1156–1160. IEEE, 2015. 1, 2, 6
- [21] Dimosthenis Karatzas, Faisal Shafait, Seiichi Uchida, Masakazu Iwamura, Lluís Gomez i Bigorda, Sergi Robles Mestre, Joan Mas, David Fernandez Mota, Jon Almazan Almazan, and Lluís Pere De Las Heras. Icdar 2013 robust reading competition. In *ICDAR*, pages 1484–1493. IEEE, 2013. 1, 6
- [22] Chen-Yu Lee and Simon Osindero. Recursive recurrent nets with attention modeling for ocr in the wild. In *CVPR*, pages 2231–2239, 2016. 1, 7, 8
- [23] Junyeop Lee, Sungrae Park, Jeonghun Baek, Seong Joon Oh, Seonghyeon Kim, and Hwalsuk Lee. On recognizing texts of arbitrary shapes with 2d self-attention. In *CVPR Workshops*, pages 546–547, 2020. 2
- [24] Ron Litman, Oron Anshel, Shahar Tsiper, Roei Litman, Shai Mazor, and R Manmatha. Scatter: selective context attentional scene text recognizer. In *CVPR*, pages 11962–11972, 2020. 2
- [25] Wei Liu, Chaofeng Chen, Kwan-Yee K Wong, Zhizhong Su, and Junyu Han. Star-net: a spatial attention residue network for scene text recognition. In *BMVC*, volume 2, page 7, 2016. 2
- [26] Shangbang Long and Cong Yao. Unrealtext: Synthesizing realistic scene text images from the unreal world. *arXiv preprint arXiv:2003.10608*, 2020. 1
- [27] Simon M Lucas, Alex Panaretos, Luis Sosa, Anthony Tang, Shirley Wong, Robert Young, Kazuki Ashida, Hiroki Nagai, Masayuki Okamoto, Hiroaki Yamamoto, et al. Icdar 2003 robust reading competitions: entries, results, and future directions. *Intl Journal of Document Analysis and Recognition*, 7(2-3):105–122, 2005. 1, 6
- [28] Canjie Luo, Yuanzhi Zhu, Lianwen Jin, and Yongpan Wang. Learn to augment: Joint data augmentation and network optimization for text recognition. In *CVPR*, pages 13746–13755, 2020. 2, 4
- [29] Anand Mishra, Karteeek Alahari, and CV Jawahar. Scene text recognition using higher order language priors. In *BMVC*. BMVA, 2012. 1, 6
- [30] Lukáš Neumann and Jiří Matas. Real-time scene text localization and recognition. In *CVPR*, pages 3538–3545. IEEE, 2012. 2

- [31] Trung Quy Phan, Palaiahnakote Shivakumara, Shangxuan Tian, and Chew Lim Tan. Recognizing text with perspective distortion in natural scenes. In *ICCV*, pages 569–576, 2013. 1, 2, 6
- [32] Anhar Risnumawan, Palaiahankote Shivakumara, Chee Seng Chan, and Chew Lim Tan. A robust arbitrary text detection system for natural scene images. *Expert Systems with Applications*, 41(18):8027–8048, 2014. 1, 2, 6
- [33] 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. *Trans on Pattern Analysis and Machine Intelligence*, 39(11):2298–2304, 2016. 1, 7, 8
- [34] Baoguang Shi, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. Robust scene text recognition with automatic rectification. In *CVPR*, pages 4168–4176, 2016. 1, 7, 8
- [35] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015. 7
- [36] Jianfeng Wang and Xiaolin Hu. Gated recurrent convolution neural network for ocr. In *NeuRIPS*, pages 334–343, 2017. 1, 7, 8
- [37] Kai Wang, Boris Babenko, and Serge Belongie. End-to-end scene text recognition. In *ICCV*, pages 1457–1464. IEEE, 2011. 1, 6
- [38] Cong Yao, Xiang Bai, and Wenyu Liu. A unified framework for multioriented text detection and recognition. *Trans on Image Processing*, 23(11):4737–4749, 2014. 2
- [39] 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 12113–12122, 2020. 2, 8
- [40] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, pages 6023–6032, 2019. 2, 4
- [41] Matthew D Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012. 6
- [42] Fangneng Zhan and Shijian Lu. Esir: End-to-end scene text recognition via iterative image rectification. In *CVPR*, pages 2059–2068, 2019. 2
- [43] Fangneng Zhan, Shijian Lu, and Chuhui Xue. Verisimilar image synthesis for accurate detection and recognition of texts in scenes. In *ECCV*, pages 249–266, 2018. 1
- [44] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *ICLR*, 2018. 2, 4