1. Problem Definition
Text Recognition has attracted increasingly interests in the last few years and shown great progresses in the field of Optical Character Recognition (OCR) for its numerous potential applications in image retrieval, scene understanding, visual assistance, etc.

However, due to the large variations in foreground texts and background objects, and the diverse text variabilities in shapes, colors, fonts, orientations and scales, along with the extreme illumination and occlusion, text recognition is still faced with considerable challenges.

Naturally, semantic informations from words represent the significant part of whole images’ meanings. So letting deep learning model understand these essential information is a meaningful work for computer vision researchers. To address these problems, a novel semantic-based text recognition deep learning model is proposed, which is abridged as STR.

The main improvements are as follows:
(1) We proposed an algorithm (TGA) which can group words in a sentence and paragraph together and arrange them in the correct order.
(2) We implemented seq2seq model to effectively correct the results from word recognition based on semantic information.

2. Background Investigation
(1) Text recognition
Recently, great progress has been made with the huge development of Convolutional Neural Networks (CNNs). Based on text recognition, various methodologies has been proposed to successfully recognize words in images.

State-of-the-art methodologies are to find image regions with text given the text bounding boxes ground truth and then consider one character at a time in these regions to recognize the text, such as CRNN, which converts sequences from one domain to sequences in another domain. In this framework, convolutional features are extracted at encoder stage, and then RNN or CNN network is applied to decode these features, then CTC is used to form the final word. Later they also developed an attention-based STN for rectifying text distortion, which is useful to recognize curved scene text.

(2) Sequence-to-sequence model and Text Spelling Correction
Aside from text recognition, sequence-to-sequence learning model (Seq2Seq) is also widely used in the field of Natural Language Processing (NLP). Some popular implementations of
Seq2seq model in natural language processing include dialog system, neural machine translation, automatic text summarization and so on.

Spelling error correction is a longstanding natural language processing problem. Based on a lot of work already done in sequence-to-sequence learning model, we implement text recognition model to get prediction results for word recognition, then proposed AWG algorithm groups and arranges isolated words from image together. Finally, a sequence-to-sequence learning model is proposed to improve text recognition accuracy of prediction based on semantic information from words in images.

3. Baseline Implementation
CRNN text recognition baseline methodology is implemented from:
https://github.com/meijieru/crnn.pytorch

CNN and Bidirectional LSTM are adopted to get text recognition results from text images. For the convolutional part, VGG-based convolutional layers automatically extract a feature sequence from each input image. For the recurrent part, A bidirectional LSTM model is built for making prediction for each frame of the feature sequence, getting predicted results for text recognition. Table presents configuration of the text recognition network model. The abbreviations 'k,' 's,' 'p,' 'o,' and 'w' refer to kernel size, stride size, padding size, number of output channels, and the max-pooling window size, respectively.

<table>
<thead>
<tr>
<th>Layer Type</th>
<th>Layer Configuration</th>
</tr>
</thead>
<tbody>
<tr>
<td>conv_bn_relu</td>
<td>k: [3, 3]; s: 1; p: 1; o: 64</td>
</tr>
<tr>
<td>max-pool</td>
<td>w: [2, 2]; s: 2;</td>
</tr>
<tr>
<td>conv_bn_relu</td>
<td>k: [3, 3]; s: 1; p: 1; o: 128</td>
</tr>
<tr>
<td>max-pool</td>
<td>w: [2, 2]; s: 2;</td>
</tr>
<tr>
<td>conv_bn_relu</td>
<td>k: [3, 3]; s: 1; p: 1; o: 256</td>
</tr>
<tr>
<td>batch-normalization</td>
<td>-</td>
</tr>
<tr>
<td>conv_bn_relu</td>
<td>k: [3, 3]; s: 1; p: 1; o: 256</td>
</tr>
<tr>
<td>max-pool</td>
<td>k: [2, 2]; s: [2, 1]; p: [0, 1];</td>
</tr>
<tr>
<td>conv_bn_relu</td>
<td>k: [3, 3]; s: 1; p: 1; o: 512</td>
</tr>
<tr>
<td>batch-normalization</td>
<td>-</td>
</tr>
<tr>
<td>conv_bn_relu</td>
<td>k: [3, 3]; s: 1; p: 1; o: 512</td>
</tr>
<tr>
<td>max-pool</td>
<td>k: [2, 2]; s: [2, 1]; p: [0, 1];</td>
</tr>
<tr>
<td>conv_bn_relu</td>
<td>k: [3, 3]; s: 1; p: 1; o: 512</td>
</tr>
<tr>
<td>batch-normalization</td>
<td>-</td>
</tr>
<tr>
<td>map-to-sequence</td>
<td>-</td>
</tr>
<tr>
<td>bi-directional LSTM</td>
<td>hidden units: 256</td>
</tr>
<tr>
<td>bi-directional LSTM</td>
<td>hidden units: 256</td>
</tr>
</tbody>
</table>

4. Methodologies for Improvement
(1) Text Grouping and Arranging
In order to retrieve the semantic information of a text in an image, it is essential to group and arrange isolated text or word regions into paragraphs or sentences in the correct logical order. To accomplish this, we propose the Text Grouping and Arranging (TGA) algorithm. The TGA algorithm takes geometric information of each text region from the Text Detection module, and
outputs the text of logically grouped and arranged sentences or paragraphs. It uses two processes, text grouping and text arranging, which are explained in detail below.

The core idea of the text grouping process is inspired by the flood fill algorithm. The text grouping process starts from a randomly selected text bounding box and groups neighboring bounding boxes together. We assume that these neighboring text regions belong to one sentence or paragraph.

The text arranging process is inspired by the Linked List Traversing Algorithm. We assume that the correct text arrangement starts from the left and continues to the right, and always proceeds from the top to the bottom. Therefore all the grouped text bounding boxes are first sorted according to their horizontal coordinates then their vertical coordinates.

(2) Spelling Correction
A sequence-to-sequence (or encoder-decoder) model is designed to correct spelling error by relying on semantic information in a specific domain. The backbone of the proposed model is based on a sequence-to-sequence network.

These models learn to generate a variable-length information sequence (e.g., an English sentence) from a variable-length input sequence (e.g., the corresponding Chinese sentence). The encoder of the sequence-to-sequence framework compresses the entire sequence of information into a fixed-length context vector, which cannot fully represent the information of the entire sequence, especially when long input sequences are given. Fortunately, this problem has been solved by the attention model. It generates a “focus range” to indicate which parts of the input sequence should be focused on, and then produces the next output sequences based on the region of focus. The attention-based sequence-to-sequence model enables focus on specific parts of the input automatically to help generate a more accurate output result.

(3) Experiments on dataset:
Interior Design Dataset (IDD) and UCLA Protest Image dataset are used. Accuracy as fellows:

<table>
<thead>
<tr>
<th></th>
<th>Design (IDD)</th>
<th>Protest (TPID)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>85.36 %</td>
<td>65.63 %</td>
</tr>
<tr>
<td>Proposed STR</td>
<td>90.04 %</td>
<td>71.09 %</td>
</tr>
</tbody>
</table>

(4) Conclusion and Reasons for Improvement
In this work, we proposed a new deep learning model called STR. This model can efficiently understand the context between regions of text or between words in images. By grouping and arranging text and relying on semantic information from images, the model is able to efficiently improve prediction results from text recognition and extract sentences or paragraphs from images instead of isolated text regions or words like state-of-the-art frameworks do. The
experiments demonstrate that STR achieves superior or highly competitive performance and suggest generality in that STR can handle different semantic contexts in images.