Supplementary Material of SCAN

To give a more comprehensive understanding of SCAN (the abbreviation of Sliding Convolutional Attention Network), we further conduct the extended visualization experiments. First, we vividly give a step-by-step introduction of the recognition process in detail. Then, we give more visualization examples to demonstrate the superiority of SCAN.

1 An Example of Recognition Process

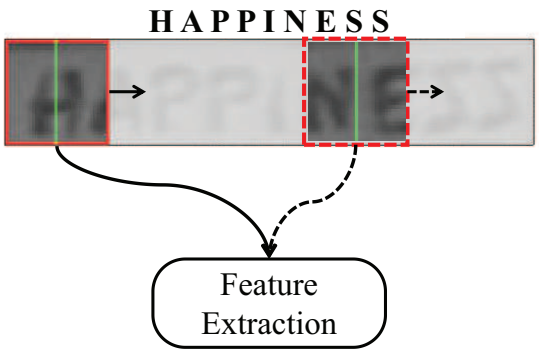
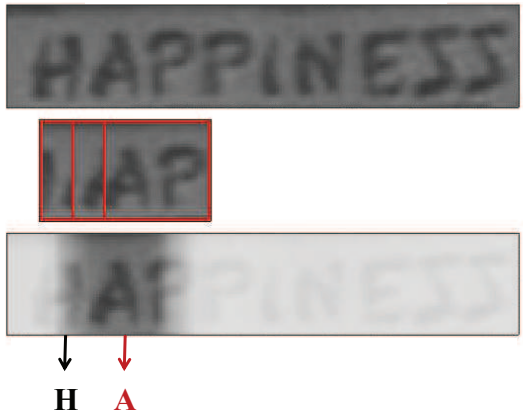
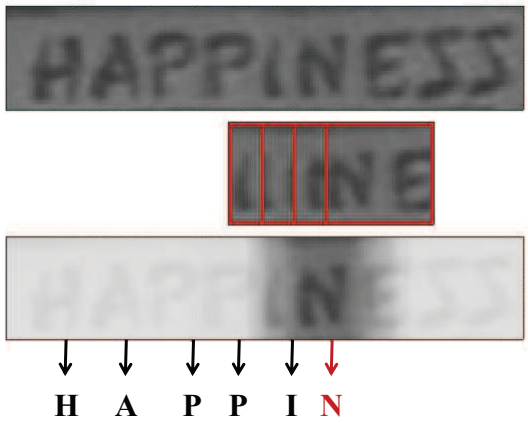


Figure 1: Sliding window and feature extraction process for the textline image "HAPPINESS".

The SCAN recognition system consists of three major parts, namely a sliding window layer, a convolutional feature extractor, and a convolutional sequence network. Firstly, the sliding window layer splits the textline into overlapped windows. On the top of sliding window, a convolutional feature extractor is built to extract the discriminative features. These two steps are shown in Fig. 1 for the textline image "HAPPINESS". Then, based on the extracted feature sequences, a convolutional sequence learning network is adopted to map the input to the output result. In Fig. 2(a), the character "A" is emitted according to the most relevant windows when considering the emitted previous characters and the whole feature sequence. Similar case is shown in Fig. 2(b), where the character "N" is emitted. The behavior of SCAN is very similar to the acuity of foveal vision in human reading. The dynamic process of SCAN can be found in the fold **IMG** containing two images of gif format (EG1.gif and EG2.gif).



(a) The process of emitting the character "A"

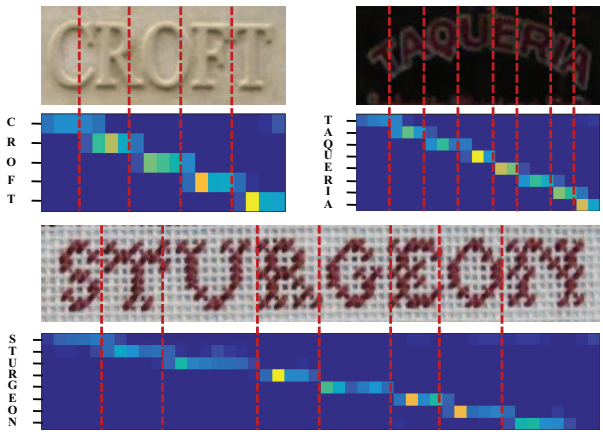


(b) The process of emitting the character "N"

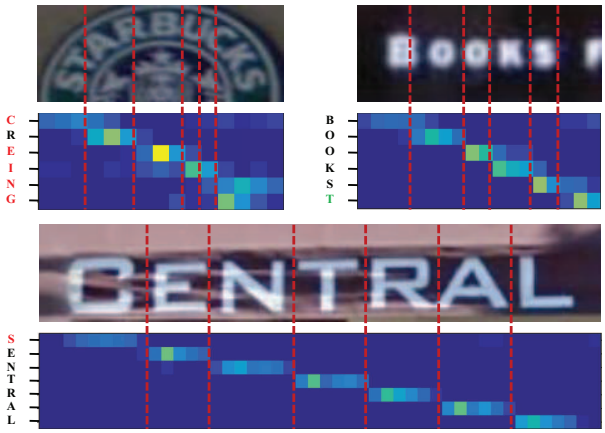
Figure 2: The attention mechanism of SCAN based on the input feature sequence.

2 Visualization of Recognition Results

More visualization examples of the recognition results are shown in Fig. 3, where each one is equipped with a corresponding attention heatmap. Fig. 3(a) shows some correct recognition samples. We can see that SCAN can not only transcribe texts in slightly blurred or curved images robustly in most cases, but also locate the boundary of each character. Therefore, SCAN can accurately attend on the most relevant windows to give the final recognition results. We find that in some cases (such as the word image "CROFT"), some attention weights far away from the specified character ("C" in this case) also has certain responses in the heatmap. This phenomenon indicates that SCAN may have implicitly modeled the between-character relationship because of the long context information the convolutional model can capture. Moreover, we could make use of some heuristic rules to remove these weights if we need to locate the position of each character precisely, since they are usually restricted to a very small range with relatively low responses.



(a) Correct recognition samples



(b) Incorrect recognition samples. The ground truths of the three images are"**STARBUCKS**","**BOOKS**" and "**CENTRAL**", respectively.

Figure 3: Recognition examples of SCAN.

On the other hand, some incorrect recognition samples are shown in Fig. 3(b). It can be seen that SCAN are still unable to deal with severely blurred or curved word images, although it has powerful feature extractor and sequence learning network. The attention of SCAN is usually drifted in images containing insertion and deletion errors. It is an alternative approach to utilize better structure for feature extraction (such as DenseNet) to enhance the discriminant of features. For those irregular text images, we may first acquire the center curve of the text line by some detection[1] or curve fitting techniques, and then slide along the curve to obtain the window sequence. In this way, we may further improve the performance of SCAN.

[1]Nowadays, it is a trend to use direct regression for multi-oriented scene text detection.