An Advanced Pyramid Network Technology for Optical Character Recognition

Haoruo Zhang and Zhongquan Cheng
Shanghai Southwest Weiyu Middle School, Shanghai, China.
Email: sadojijzoxc@163.com
SMIC Private School, Shanghai, China.
Email: petercheng0630@163.com

Abstract. Optical Character Recognition (OCR) technology can quickly convert text and digital information in pictures into text information, which has been widely used in actual scenes. However, in some complex scenes, such as different light, angle or occlusion, the existing OCR systems still cannot reach the required accuracy. This paper proposes an advanced pyramid network structure, which uses multiple different scales and parallel pyramid network structures to deal with the problem of different character scales and misalignment. Each pyramid network structure has at least four convolutional layers. At the same time, in each pyramid network connection, the proportionally discarding parameters is introduced to increase the calculation speed further. The advantage of this network structure is very robust to more differently sized characters and increases the number of valid parameters obtained through training. Experiments on open source data sets show that the method has better recognition accuracy and speed.

1. Introduction
In any environment, such as street scenes or indoor scenes such as supermarkets and shopping malls, the detection and identification of text in the scene (usually classified into the field of computer vision) is very challenging. It carries a wealth of high-level semantic information, which is very important for image understanding. Identifying text in an image can help with many practical applications, such as drone navigation, unmanned car navigation, and machine translation.

The traditional optical character recognition (OCR) system in the usual impression is to find and recognize text content from electronic files. In the early days of pattern recognition research, almost everyone accepted the theme of OCR [1]. One reason is that it is very convenient to handle and considered an easy problem to solve. Engineers believe that OCR is a possible reality. Nevertheless, there are significant limitations in terms of quantity and complexity. The basic reduction of hardware achieved by the following methods, from two-dimensional information to one. The basic reduction in complexity is through the project from two-dimensional information to one. In 1956, the magnetic shift register used to scan the input characters at the top. The bottom is a slit that reflects light. The print input paper is given to the photodetector. This is a simple calculation that uses algebraic addition to get a. This value is proportional to the area of the black portion. Split the input characters in the slit. The sampled value is sent to a register to convert the analog signal value to be digitized. Template matching is done by summing the differences between each sampled value and the corresponding template value, each normalized.

However, after some initial progress, the difficulty of solving this problem surfaced. The theme of OCR is not so special, but rather universal, because it contains the basic problems of pattern
recognition common to all other topics. For example, text obtained from natural scenes is more challenging due to distortion, occlusion, blurred directions, cluttered backgrounds or different perspectives [2]. Besides, Field text may have an irregular shape. For example, some scene text is perspective text, which is introduced by the side view camera angle; some have a curved shape, which means that its character is placed along a curve rather than a straight line. This type of text is called irregular text, in contrast to regular text (horizontal and positive).

The Spatial Transformer module has been proposed [3], which can be included in the standard neural network architecture to provide space conversion functions. The motion of the spatial transformer depends on the individual data samples and the appropriate behavior is learned during training. The spatial transformer module is a dynamic mechanism that actively spatially transforms an image by generating an appropriate transform for each input sample. The transformation is then performed on the entire feature map. This allows the network including the spatial transformer to not only select the regions of the image that are of most interest, but also transform those regions into the intended poses of the specification to simplify the identification in the following layers.

In general, text recognizers work best when their input image contains tightly bounded regular text [4]. It is therefore necessary to apply a spatial transformation prior to recognition in order to correct the input image to an image that is more "readable" by the recognizer. A recognition method that is robust to irregular texts is proposed, and a deep neural network is constructed, which combines spatial converter network (STN) and sequence identification network (SRN). STN is an image that produces more regular text than the original text, using a nonlinear thin plate-spline (TPS) transform to correct irregular text. On the other hand, the text in the image carrying the text is usually arranged horizontally, which is somewhat similar to the sequential signal. Therefore, SRN uses a attention-based model and uses sequence recognition methods to identify these texts. The SRN consists of an encoder and a decoder by generating an ordered feature representation by an encoder, the decoder functioning to generate a sequence of characters over the input sequence cyclically.

At the same time, OCR’s latest task is not only to transcribe all the text in a given image, but also to focus on more issues, such as how to focus on important parts of a chaotic scene and extract useful information. Combined with Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN) and Spatial Annotations was proposed [5]. Many of these OCRs are based on an important data set: Google French Street Name Signs (FSNS) dataset (http://rrc.cvc.uab.es/?ch=6), which converts different views of the same symbol into batch dimensions, where multiple multi-layer LSTMs are designed to process each line of text separately and use CTC loss. Structured information extraction models based on attention are simpler, more accurate, and have fewer assumptions about the data. The model was evaluated on FSNS and Street View business name datasets. This literature also studies the accuracy and speed of three different CNN-based feature extractors as input to our attention model. The ablated versions of these models use fewer layers to meet the speed requirements.

Convolutional neural networks (CNN) are currently the main deep learning method for image recognition problems. However, many of the problems have arisen with the need for models that can analyze spherical images, including omnidirectional vision of drones, robots and self-driving cars, molecular regression problems, and global weather and climate models. Simply applying a convolutional network to a planar projection of a spherical signal is bound to fail because the spatially varying distortion introduced by such projection will invalidate the translation weight sharing. The construction module of spherical CNN is proposed in literature [6], which gives the definition of spherical cross-correlation, which has both expressiveness and rotation and other denaturation. Spherical correlation satisfies the generalized Fourier theorem, which allows efficient calculation of spherical cross-correlation using a generalized Fast Fourier transform (FFT) algorithm. The computational efficiency, numerical accuracy and effectiveness of spherical CNN applied to 3D model recognition and atomization energy regression are obvious.

Although the latest technologies and systems are constantly emerging in the OCR field, and the speed of technology update is accelerating, the OCR technology in the chaotic scene still has performance degradation caused by interference. The shooting angle, light intensity, different fonts, and different weather may effect on the system and reduces the recognition accuracy, so the system could not meet the user requirements. This paper presents an advanced pyramid neural network OCR
detection and recognition technology. This method trained and tested on the scene OCR recognition open source dataset. Compared with the above existing methods, it has certain improvement effects on the test dataset.

This paper is organized as follows: The second part introduces the basic OCR methods and the hourglass principle; then gives the new pyramid neural network structure and OCR application; the fourth part details the experimental results on the standard datasets; finally, the conclusion is given.

2. OCR Methods and Pyramid Network

2.1. OCR Methods
Since the 1950s, researchers have concentrated on OCR research, mainly using template matching and structural analysis methods. NCR, Farrington, and IBM have developed their own OCR software. These software usually only recognize the numbers, English letters and some symbols of the printed matter in the specified font format. OCR began to focus on handwritten text recognition, initially with handwritten digit recognition for handwritten zip codes, bank handwritten digit recognition, etc. These applications are popular in some organizations until now. In recent years, OCR products have applied for identifying documents with a lot of noise and rich character sets.

Essract is an open source OCR engine that was developed at HP between 1984 and 1994 [7]. Tesseract assumes that its input is a binary image in which an optional polygonal text area is defined. The first step is connected component analysis, which outlines the component storage. At this stage, the outline is gathered into the blob purely through nesting. Blobs are organized into lines of text and analyze rows and regions for fixed pitch or proportional text. Depending on the type of character spacing, text lines are divided into words differently. Fixed pitch text is immediately cut off by character cells. Decompose proportional text into words using deterministic space and fuzzy space. Then, the identification is performed in a two-pass process.

In the last decade, deep neural network models, especially deep convolutional neural networks (DCNN), have been widely used in various computer vision applications, such as detection or classification of object categories for images. In the real world, stable visual objects, such as scene text, handwriting, and music, tend to appear in a sequence rather than in isolation. Unlike general object recognition, identifying such sequences of similar objects typically requires the system to predict a series of object tags rather than a single tag. Therefore, the recognition of these objects can naturally be used as a sequence identification problem [8]. Another unique property of sequential objects is that their length can vary greatly. The most popular depth models, such as DCNN, cannot be directly applied to sequence prediction because DCNN models typically operate on inputs and outputs with fixed dimensions and therefore cannot produce variable length tag sequences. Even with the method of first detecting a single character and then using the DCNN model to recognize these detected characters, or treating the scene text recognition as an image classification problem, it is difficult to directly use image-based sequence recognition. The reason is that a well-trained model has a large number of classes that are difficult to generalize to other types of sequenced objects.

2.2. Pyramid Network Structure
Image synthesis methods based on Generating Against Networks (GAN) are very effective for synthesizing MNIST numbers and small size images by training multiple individual GANs, one for each level in the Laplacian pyramid [9]. Each model is trained independently to synthesize the details of its scale. Assembling individual training models in this way allows the author to synthesize smoother images and increase resolution. Multi-scale refinement is a core feature of this kind of release, training individual models end-to-end to directly synthesize output images, and does not use confrontation training.
Figure 1. Cascaded Pyramid Network.

The Cascading Pyramid Network (CPN) network structure [10] is used for multi-person pose estimation, which is shown in Fig.1. Multi-person pose estimation is the key point for identifying and locating all people in the image. This is the basic research topic of many visual applications. The CPN network architecture is divided into two phases: GlobalNet and RefineNet. GlobalNet learn a good feature representation based on the feature pyramid network. More importantly, pyramid feature representations can provide sufficient contextual information, which is inevitable for occlusion and invisible joints. Based on pyramid features, RefineNet clearly addresses certain joints based on online hard key point mining losses. The multi-person pose estimation problem is solved based on the top-down pipeline of the cascade pyramid network. The human body detector is first used to generate a set of human bounding boxes and then use CPN to locate key points in each bounding box.

3. Advanced Pyramid Network Structure

A method of combining CRNN with CTC mentioned in literature [8] is useful in the OCR method, which mainly includes three parts: Convolutional Layers, which is a common CNN network to extract Convolutional feature maps of input images. Recurrent Layers, a deep two-way LSTM network, continues to extract text sequence features based on convolutional features. Transcription Layers, which converts the RNN output to softmax and converts it to characters. In fact, in Recurrent Layers to calculate Softmax Loss, the output here corresponds to the characters one by one. Such processing has many requirements for the labeling of training data. Only very high-quality labels, including the position of each character, and the label of each character, can be properly trained at this level. Connectionist Temporal Classification (CTC) could solve this difficulty, which is a Loss calculation method that gives all possible output distributions for the input sequence and calculates the probability of one of the most likely outputs. Each time slice takes the node with the highest probability of the time slice as the output, as shown in Equation (1):

$$A = \arg \max_d \prod_t t^T p_t (a_t | X)$$

(1)

The above CTC scheme has the disadvantage that it ignoring that one output may correspond to multiple alignments, and there is also a problem that the calculation cost is high.

Therefore, we propose an advanced pyramid network structure, as shown in Figure 2. By adopting M different scale and parallel pyramid network connection structures, the problem that the character scales are different and the alignment cannot be aligned is dealt with. Each of the pyramidal network connection structures has at least 4 convolutional layers. At the same time, in order to cope with the problem of excessive calculation, in each pyramidal network connection structure, the parameters are discarded according to the following proportions: 10%, 20%, 30%, 50%, that is, the proportion of the discarding parameters is gradually increased.

By this kind of network structure, on the one hand, the system can accommodate more size characters to perform feature extraction in the network, and on the other hand, the proportion of effective parameters obtained through training becomes larger, and the calculation amount decreases accordingly.
4. Experimental Results and Evaluation

In order to prove the performance of advanced pyramid network structure, the experiment was performed under pytorch, configuring the relevant Python environment. The data selects part of the data of ICDAR2017 Competition on Reading Chinese Text in the Wild (RCTW-17) of ICDAR2017 (http://u-pat.org/ICDAR2017/index.php), as shown in Figure 3 for an example of data [11].

**Figure 2.** Advanced Pyramid Network.

**Figure 3.** Example images and annotations of the CTW-12k dataset.

**Figure 4.** Experimental results of baseline and advanced pyramid networks under five subsets.

**Figure 5.** Experimental results of baseline and advanced pyramid networks under five subsets.
The data set consists of 12,263 pictures of natural scenes containing Chinese, most of which are taken directly by the camera or mobile phone, and each image contains at least one line of Chinese. The labeling of the data is annotated by hand, using a quad to mark the text. There are font differences, layout differences, and language differences in the data set. The training set and the verification set totaled 8034, and the test set totaled 4229 pictures. Our experiment reselected five subsets of the data set as the training set and the verification set and test set respectively. Some image preprocessing methods, such as image moderate rotation, special character substitution, as a means to use in experiments.

First, CRNN+CTC is used as the baseline on five sub-data sets, and compared with the proposed advanced pyramid network structure. The experimental results are shown in Fig. 4. On the five sub-data sets, the latter has certain advantages, which can improve the correct rate of text recognition and achieve a correct rate of 71% on the fourth sub-data set. Further, the partial discard parameter method proposed by us is applied to the five sub-data sets, and the result is shown in Fig. 5. It can be observed that, assuming that the discarding parameter is not used and the test data consumption time of the network is 100%, the computational efficiency on any sub-data set is improved under any method of discarding parameters. Therefore, it can be considered that combining the pyramid network structure with the method of discarding parameters can have both good robustness and high recognition accuracy.

5. Conclusions
We propose an advanced pyramid network structure for text recognition in images. Character scales and alignment problems are handled using multiple, parallel, pyramidal network connections of different scales. At the same time, in order to cope with the problem of over-computation in each pyramid network connection structure, the parameters are discarded and compared according to the increasing proportion. Experiments show that compared with the baseline, the method has better recognition accuracy and computational efficiency, and is more suitable for practical applications.

6. References
[1] Aderhold J, Davydov V Yu, Fedler F, Klausing H, Mistele D, Rotter T, Semchinova O, Stemmer J and Graul J 2001 J. Cryst. Growth 222 701
[2] Dorman L I 1975 Variations of Galactic Cosmic Rays (Moscow: Moscow State University Press) p 103
[3] Caplar R and Kulisic P 1973 Proc. Int. Conf.on Nuclear Physics (Munich) vol 1 (Amsterdam: North-Holland/American Elsevier) p 517
[4] Szytula A and Leciejewicz J 1989 Handbook on the Physics and Chemistry of Rare Earths vol 12,ed K A Gschneidner Jr and L Erwin (Amsterdam: Elsevier) p 133
[5] Kuhn T 1998 Density matrix theory of coherent ultrafast dynamics Theory of Transport Properties of Semiconductor Nanostructures(Electronic Materials vol 4)ed E Schöll (London: Chapman and Hall) chapter 6 pp 173–214
[6] Mori S, Suen C Y, Yamamoto K. Historical review of OCR research and development. Proceedings of the IEEE, 1992, 80(7): 1029-1058.
[7] Wojna Z, Gorban A N, Lee D S, et al. Attention-based extraction of structured information from street view imagery. 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). IEEE, 2017, 1: 844-850.
[8] Jaderberg M, Simonyan K, Zisserman A. Spatial transformer networks. Advances in neural information processing systems. 2015: 2017-2025.
[9] Shi B, Wang X, Lyu P, et al. Robust scene text recognition with automatic rectification. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 4168-4176.
[10] Wojna Z, Gorban A N, Lee D S, et al. Attention-based extraction of structured information from street view imagery. 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). IEEE, 2017, 1: 844-850.
[11] Cohen T S, Geiger M, Köhler J, et al. Spherical cnns. arXiv preprint arXiv:1801.10130, 2018.

[12] Smith R. An overview of the Tesseract OCR engine. Ninth International Conference on Document Analysis and Recognition (ICDAR 2007). IEEE, 2007, 2: 629-633.

[13] Shi B, Bai X, Yao C. 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, 39(11): 2298-2304.

[14] Chen Q, Koltun V. Photographic image synthesis with cascaded refinement networks. Proceedings of the IEEE International Conference on Computer Vision. 2017: 1511-1520.

[15] Chen Y, Wang Z, Peng Y, et al. Cascaded pyramid network for multi-person pose estimation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7103-7112.

[16] Shi B, Yao C, Liao M, et al. ICDAR2017 competition on reading chinese text in the wild (RCTW-17). 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). IEEE, 2017, 1: 1429-1434.