

# A High Accuracy Rate Commercial Flight Coupon Recognition System

Shanheng Zhao, Zhiyan Wang, Member, IEEE  
South China University of Technology  
szhao@ieee.org wzhyan@ieee.org

## Abstract

*In this paper we introduce a practical flight coupon automatic processing system for scanning and recognition. We discuss the coupon classification, character location and binarization. And emphasize a high performance character segmentation and recognition engine, which are proved very effective. The results of experiment and commercial running applying the system are presented.*

## 1. Introduction

Airway companies have showed great interest to invest in coupon recognition automation, because it would reduce their long-term capital investment for the traditional coupon storage and searching. It is for a long time that the airway companies must store their coupons for more than 5 years in a storage-house according to the rule set by the administration. It is very difficult and needs lots of clerks to find a very coupon in millions of rough-sorted coupons. It also involves a lot of labor to input the coupons data into the computer. For example, in Southern China Airways, there are more than 200 people working on that job. However, errors frequently happen during the process. So, it is urgent to shift this coupon managing method to a new, e-based method, with the kernel part known as Optical Character Recognition (OCR).

Traditionally, OCR is divided into three parts: character location, character segmentation and single character recognition. The character location analyses the document and finds out the zones with characters and passes them to the character segmentation part. The segmentation procedure then cuts out the characters by line segmentation and character segmentation, sequentially. Finally, the character recognizer receives the character from the segmentation procedure and recognizes it in particular or combined methods. The conventional methods include Stroke Tracking, Pattern Matching, Hidden Markov Model (HMM), Artificial Neural Network and etc.

But most traditional OCR softwares do not suit for this task, because the coupons to be recognized are very

ugly (See Fig. 1 below), with different layout styles and character styles, while they require a very high recognition rate. Besides, traditional softwares can hardly know the characters' meanings, so they can hardly store them into database properly. Furthermore, the background texture, preprinted characters, form lines and seals and strokes written by humans can greatly reduce the correct segmentation and recognition rate. Flight coupons have different styles and patterns all over the world, even in one company. For example, Cathay Pacific Airways in Hong Kong has more than twenty patterns. So, it's necessary to develop a particular system to adapt this task.



Fig. 1 Typical Flight Coupon

## 2. Difficulty Encountered

At least for the next ten to twenty years, paper-based coupons will still be widely used. So, the arduous inputting process, the formidable finding process and the huge storage-house are inescapable, if using the traditional coupon management method. One can see hundreds of employees are engaged in typing keyboards in front of huge amount of coupons in the airline companies' offices. And it takes them a long time to find a coupon in a huge storage-house.

So, it is urgent to develop a special system to supersede the inconvenient old one by scanning and reading coupons automatically. By this method, great amount of manpower and expenses would be reduced.

Coupons, including those printed by the airway companies themselves and those by agents, are generally classified into processing categories by flights and then delivered to the processing center. Because of the different printing quality, the coupons are disparate.

Sometimes the gray level is deeper, sometimes the characters are unclear, sometimes even broken or merged. These problems bring a big trouble to the OCR system.

An OCR system for industrial purpose has different constraints and problems. In this case, the input device is a high-speed scanner with CCD, which grabs a grayscale image from the coupon. Conditions for the image grabbing are often different. The blowing problem inevitably during the high-speed scanning may need to be rotation corrected. In most cases, coupons' backgrounds are often disturbed by preprinted characters, form lines and textures.

After scanning the coupon images into the computer, the coupon number must be recognized in order to match the image with the sales data in the background database. Information to be recognized include Coupon Number (CPN, 1 digit), Airline Code (AC, 3 digits), Form and Serial Number (SN, 10 digits) and Check Digit (CK, 1 digit), respectively from left to right in Fig. 2 below.

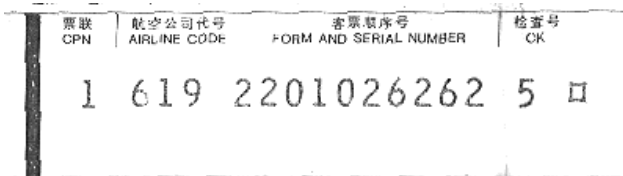


Fig. 2 Flight Coupon Numbers

But not all coupons are as clear as Fig. 2. Most of them look like Fig. 3a, or even worse, as Fig. 3b, or even worse. And the serial location is not determined. They may be in the left bottom, center of the bottom or right bottom. We've tested many famous OCR softwares in the market, but the result is disappointing. The recognition rate is very low and they can hardly process different types of coupons at the same time.

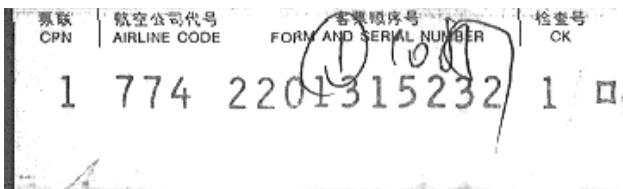


Fig. 3a Unclear Flight Coupon 1



Fig. 3b Unclear Flight Coupon 2

Previous researches have been worked on this field [1]. But their system is based on a very small test set (110 coupon images), and the coupons they processed are much more regular than those used in China. So, developing a high accuracy rate commercial system is necessary.

### 3. Scanning and Deskewing

Because of the wide variation in coupon qualities, we need to adjust the image captured from the scanner to obtain a high quality one for the highest recognition rate, such as image deskewing, brightness adjusting and image format converting.

In our case, the scanning software bind with the scanner can only provide a binary image. We have to develop a grayscale image scanning software using TWAIN. The software can deal with various grayscale processing and skewed emendation.

### 4. Coupon Classification

Typically, there are two types of coupons, named BSP (Fig. 4) and NonBSP (Fig. 2). Their main difference is the Airline Code: AC of BSP is sealed, while NonBSP is printed. One can find that ACs in BSP coupons are much ugly, with more noise than those in NonBSP coupons.

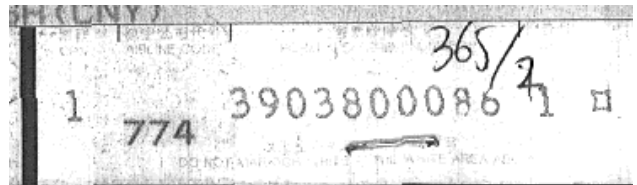


Fig. 4 BSP Type Flight Coupon

Classifying the coupons before recognizing them is important. But it is hard to recognize whether the AC is sealed or printed. So, we try to locate the barcode in the coupon. Usually, there are barcodes in BSP coupons, and none in NonBSPs.

We found that the horizontal linear filter proposed by F. LeBourgeois [2] is not so suitable with character extraction in flight coupons, but it's very good for barcode location.

- F(I,j) : Initial Image
- G(I,j) : Filter response
- T : The average slope of character contours
- W : Weight correction of the local contrast
- S : Dilation of Smearing Coefficient

$$G(i, j) = \frac{1}{S \cdot W} \sum_{k=-S/2}^{k=S/2} |F(i, j - t/2 + k) - F(i, j + t/2 + k)|$$

In our case,  $T=2$ ,  $W=1$ ,  $S=40$ .

We use the linear filter to get a response image of the origin flight coupon. Then, a fixed thresholding binarization is applied with threshold of 50. After that, we use a seed fill algorithm with stuffing rate checking to locate the barcode region. (See Fig. 5)

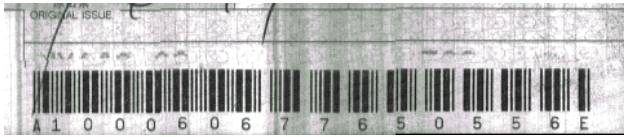


Fig. 5(a) Origin Image



Fig. 5(b) Filtered Image



Fig. 5(c) Binarized Image

If the barcode can be located, then the flight coupon is in no doubt BSP one. And the recognition result of barcode (it is not so difficult to recognize) can be redundant verified with the serial recognition result later. If the barcode cannot be located, it should be NonBSP type in most cases. But due to the disturbance, sometimes, though very few, barcodes cannot be located. So, adjusting and correcting in later recognition process is needed.

## 5. Serial Zone Location

As mentioned above, the serial zone can appear in different locations in the flight coupons. But there's a unique feature to all these coupons: a black bar in the left side of the zone. So, if we can locate the black bar, we can locate the serial zone.

There may be many methods to accomplish this task. But it is unwise to apply complex algorithm and waste too much time on this tiny job, for this is a commercial system and it requires high speed. So, we introduce a very simple, but practical algorithm.

The algorithm scans the coupon image from the bottom to 1/3 from the bottom, using horizontal lines. It grades the last  $n$  ( $n = 1, 2, \dots$ ) lines respectively by the following steps.

1. Initial the grade array with zero. The array size equals to the width of the coupon image.
2. Calculate the horizontal projection array of last  $n$  lines.
3. Find the minimum(s) of the array. Then add 1(s) to the corresponding column in the grade array.
4.  $n=n+1$  and back to step 2 until  $n$  reaches 1/3 of the height of the image.

After that, the continuous columns with highest scores are the location of black bar. This algorithm is very simple, but successful. Most important, it's very fast.

Since the black bar has been located, it is easy to locate the serial zone at the right side.

## 6. Binarization

Though many gray-level recognition methods have been proposed, the recognition kernel we used is based on binarized image, because the gray level varies greatly in different flight coupons. So, the separated serial zone needs to be binarized.

We have tried several binarization methods, and finally use a modified local adaptive thresholding algorithm proposed by Niblack [3]. The algorithm gives a very good performance, except that it is relative slow and the binarized image may have burrs (Fig. 6).

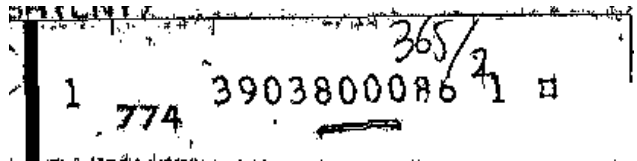


Fig. 6 Binarized Coupon Number Image

## 7. Character Extraction

After pre-processing, characters need to be extracted for recognition.

In the segmentation, we use seed fill algorithm again with pyramid algorithm to extract possible character blocks in the zone. Then all extracted blocks are clustered line by line.

The seed fill algorithm is a good idea to extract the characters. But it has a weakness: it is sensitive to stroke disturbance. As mentioned before, many flight coupons have irregular strokes on them. If these strokes touch the

character, the algorithm may find a very big block. So a size filter must be introduced to delete the irregular block. For instance, the digit “6” in Fig. 5 is touched by a stroke and will be deleted after block size checking.

In line clustering, the vertical center of each block is calculated. Then, blocks with similar vertical centers will be clustered in the same line. The line with maximum blocks will be seen as the serial line.

The block deletion will arouse character deletion. So, we introduce an intelligent detector to recover the deleted characters. We can find that the serial line appears like 1-3-10-1, which means 1 digit Coupon Number, 1 gap, 3 digits Airline Code, 1 gap, 10 digits Form and Serial Number, 1 gap and then 1 digit Check Digit. When we successfully locate some characters in this 1-3-10-1 pattern, we can recover the remaining by determining the relative positions of the characters. The main difficulty in this method is how to know the “blank block” is a deleted character or a real blank, or the “character block” is a real character or noise. So, the intelligent searching, matching and grading is inseparable. To every flight coupon, we give all possible recovered digits and gaps sequence, and then rank them by a grading system. Only the sequence with the highest grade is selected as the serial sequence. After that, the character blocks in the serial sequence include the exact border of each digit of the flight coupon serial.

The seed fill algorithm with intelligent matching can greatly reduce touching characters or broken characters. In our test, more than 99% in a test set of 7118 flight coupons can be successfully separated.

## 8. Character Recognition

There are numerous algorithms in single character recognition, such as Pattern Matching, Experts System, Hidden Markov Model (HMM) and Artificial Neural Network. We use Back Propagation Neural Network (BPNN) as one recognition kernel in our system, for the relatively higher recognition rate and robustness. BPNN was advanced by Rumelhart in 1987 and was one of the most useful neural network algorithms [4].

The network we used has one input layer, one hidden layer and one output layer. The input layer is the pixel array of isolated character. The number of neural cells in output layer is 10, which indicates 10 digits. Because the characters are extracted almost exactly, we do not face the problems encountered in similar recognition systems, such as character overlapping, and mis-segmentation.

The network was trained with 12000 input patterns from 800 coupons. We use gradient-based learning algorithm to update weights in the network.

$$W_k = W_{k-1} - \varepsilon \frac{\partial E^{Pk(W)}}{\partial W}$$

We use uniform distribution between  $-2.4/\text{FanIn}$  and  $2.4/\text{FanIn}$  as the network initial parameters to accelerate the training process [5]. After recognition, the result and confidence of each input is then passed to the check module.

Besides the BPNN recognition engine, a dual-channel Convolutional Neural Network (CNN) is used as another recognizer (see Fig. 7). This recognizer is an implementation of the work in Ref. [6].

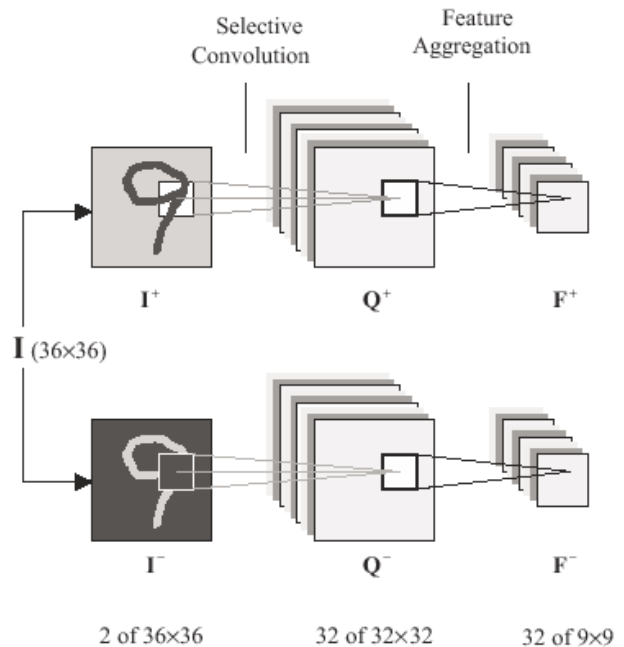


Fig. 7 The Feature Extraction Process of Convolutional Neural Network

Here **I** denotes the size normalized single digit character image. **I**<sup>+</sup> is copied from **I**, while **I**<sup>-</sup> is the reversed image of **I**. **Q**<sup>+</sup> and **Q**<sup>-</sup> denote the convolution layer: **F**<sup>+</sup> and **F**<sup>-</sup> are used to detect the presence of a feature in each window. There are totally 16 different features in the network. As for the detail of this character recognition method, please refer to Ref. [6].

The two recognizers are applied simultaneous. To get better effect, the two networks are trained by different set of flight coupons. Training set of CNN is shifted from top-left to right-bottom and produce more training images

to tolerate the shifting of the test set, while the training set of BPNN hasn't been shifted.

## 9. Check Module

The check module first checks the Coupon Number, Form and Serial Number with the Check Digit. The check digit can check one digit error in the Coupon Number and Form and Serial Number. But the modulus 7 based checking cannot check the errors greater or less by 7, such as 0 and 7, 1 and 8, 2 and 9.

The airline companies require a very high correcting rate in this commercial system. That is: those the system believes to be correct MUST be correct. So, besides the Check digit, we inspect the top two choices from the neural network. Only if the confidence of the top choice is greater than 90% and the second is lower than 50% (it means confidence of all other choices are all under 50%), we consider this digit to be reliable. Otherwise, it is unreliable.

The Airline Code has no check equation. So we try to establish a maintainable database to store all possible ACs. The top three choices from the network are inspected. So, three digits of ACs should have 27 different combinations. Only the most reliable combination is selected as the correct AC.

To some cases, BPNN has relative higher discriminability than CNN, and some cases in the contrary. Only both recognizers give same reliable result, the digit character is believed as correct. Otherwise, it is doubtful and warnings will be raised.

As to BSP type flight coupons, the barcode recognition result can also serve as verified information. If the barcode recognition result and the results from both BPNN and CNN character recognizer give the same number serial, the result can be treated as reliable, in spite of some single character unreliable.

## 10. Performance and Conclusion

In a performance test with 7118 real coupons from several hundred different flights. The system got a recognition rate of 96.76%, with 0.27% location failure, 1.03% check failure, and 1.94% unreliable coupons. Usually, those unreliable coupons are in fact correct.

This system also used for several months in Shanghai Airline in China. It has processed more than half a million coupons, while gets a recognition rate more than 93%. The reason that the recognition rate is lower than the test is due to the coupon quality and the scanning quality. Some passengers pasted their VIP sheet just on the serial zone. It is impossible for the system to recognize the coupon in this case. However, the performance of the system is very good and it is very fast: it can process more than 5 coupons with image compressing after recognition in an Intel Pentium III machine.

## 11. Acknowledgment

Our research work was supported by Guangdong Provincial Natural Science Funding Project B6-109-497.

## References

- [1] Jianchang Mao, Raymond Lorie and K. Mohiuddin, "A System for Automatically Reading IATA Flight Coupons", *Proceedings of the 4th International Conference on Document Analysis and Recognition (ICDAR '97)*, pp. 153-157, Ulm, Germany, August 18-20, 1997
- [2] F. LeBourgeois, "Robust Multifont OCR System from Gray Level Images", *Proceedings of the 4th International Conference on Document Analysis and Recognition (ICDAR '97)*, pp. 1-5, Ulm, Germany, August 18-20, 1997
- [3] W. Niblack, *An Introduction to Digital Image Processing*, pp. 115-116, Prentice Hall, Englewood Cliffs, NJ, 1986
- [4] D.E. Rumelhart, G.E. Hinton and R.J. Williams, *Parallel Distributed Processing*, vol. 1, MIT Press, Cambridge, MA, 1987
- [5] Y. LeCun, P. Haffner, L. Bottou, and Y. Bengio, "Object Recognition with Gradient-Based Learning", in *Feature Grouping*, (D. Forsyth, ed.), 1999. 28 pages.
- [6] Loo-Nin Teow and Kia-Fock Loe, "Robust Vision-based Features and Classification Schemes for Off-line Handwritten Digit Recognition", *Pattern Recognition*, The Journal of the Pattern Recognition Society, 2002, vol. 35, pp. 2355-2364.