String Extraction Based on Statistical Analysis Method in Color Space

Yan Heping, Zhiyan Wang, and Sen Guo

School of Computer Science & Engineering, South China University of Technology, Guangzhou 510640, China
gzyhp@126.com, wzhyan@ieee.org

Abstract. A method based on statistical characteristics and color space consistent with human visual perception for pixels classification is brought forward in this paper. In the airline coupon color design, we use colors to distinguish different object, the idea is embodied in this method. The marked characteristics suitable for object pixels classification have been found by analysis the statistic characteristics of all sorts of pixels. The experiments have proved that this method is simpler, more efficacious and can support data analysis for the whole coupon project.

1 Introduction

Our airline coupon project group [1] developed a recognition and management system. It has processed millions of coupons for years. Its mean recognition rate is about 95%, and we wish to improve it. So we have studied on how to convert image character into graphic one to recognize it. In this research we found it is very important to extract pixels of character perfectly.

Researches [2] regard pixels extraction as classification of pixels in the coupon image; pixels on (belong to) characters which are to be recognized are classified to the foreground pixels (character objects, see Fig. 1(a)) and the others are classified to the background. In this research, the HSV space is used to extract pixel features and a neural network (NN) method which is based on the Principal Components Analysis (PCA) is used as a pixel classifier to classify all pixels into some foreordain sorts. Experiment result shows a good effect is reached.

1.1 Problems

But, there are still some problems in the segmentation result images, which include (1) some background pixels which have distinct visual perception with object pixels, see Fig.1(a)., are classified into object pixels set by mistake. (2) in some cases, some unknown color sorts are appeared in the coupon image, such as manual characters, see Fig.1(b).; in this instance, the classifier will not work normally; it will classify these color pixels into an adjacent sort; obvious, it is not reasonable.

W. Liu and J. Lladós (Eds.): GREC 2005, LNCS 3926, pp. 173 – 181, 2006.

1.2 Reasons

For pixel classification task, color features are used to simulate the classification process of human visual perception. So, if the color space is more uniform and more consistent with human visual perception, the classification effect is better.

(a) (b)

Fig. 1. (a) The obvious error in pixels classification; (b) Incorrect pixels classification for unknown color class. In the upper: object pixels; In the lower: the original image; Within the ellipse: incorrect object pixels.

By data analysis in the HSV color space, one reason for the problem (1) is about uniformity and consistency of HSV space. There are many existing system for arranging and describing color, such as RGB, YUV, HSV, LUV, CIEXYZ, CIELAB, Munsell system, etc [3-4]. But, most of them are usually different from human perception. Among all the existing color systems, the Munsell color system is the best in simulation human color vision [4]. So, we select the Munsell color system.

The data analysis results show another reason for problem (1): there are over-learning or lack-learning in the NN training process. The color number in a color space is tremendous, and just some sorts of color pixels are selected to train the NN; so it is inevitable to classify a color into a color sort, which maybe farther from another sort in human visual perception.

To problem (2), it is a certain problem of this NN method. In NN design, the number of color sorts is foreordained. When an unknown color sort appeared, the classifier works abnormally is reasonable.

Additional problems are still existed. For example, the selection of suitable samples is difficult, and the classifier is complex in practical work.

1.3 Ideas for Solving Problems

Firstly, we should solve the color space problem. As said above, the Munsell color system is a better selection. Analysis from the design of coupon, we can see different objects are in different colors. The unchangeable colors are including 5 sorts as analyzed in [2]. And the casual unknown color sorts, most in the handwriting are also in a visual different color. So we can say the design of coupon is using the striking contrast colors to distinguish the different sorts of objects in coupon.

In classifier design, because of the limitation of NN and the demand of data analysis, we consider to reduce complexity and improve generalization ability. By analysis the background of this problem, it is a statistical problem in essential. Considering an

ideal instance, the unchangeable background only includes some single colors and the objects are in the colors. So it is a simple task to classify pixels according to the color features in any color space. As a matter of fact, in the process of coupon going to be pressed and printed, the disturbing such as the ink infiltration, outspread, especially color superimposition, these single colors will be changed; they will form a complex color distribution in any color space, see Fig.2.

But, the changing of these single colors should obey some statistical rules. So to grasp the color distribution of all kinds of pixels is the essential to design an effective classifier.

In this research, we analyzed the statistical characteristics of all sorts of pixels in the HVC color space. Then, based on these analysis results, we designed a simple classifier to extract the object strings. It is effectively proved by practical work.

This paper is organized into 5 sections. In section 2, we introduce the color space and color distance used in our research. The 3rd section analysis the statistical characteristics of the fixed color sorts, mainly in the object sort, sort 1; then based on these analysis results, we propose a method to separate the object pixels from others. In section 4, we give an experimental results compared to method in [2]. At the last section, a conclusion is given.

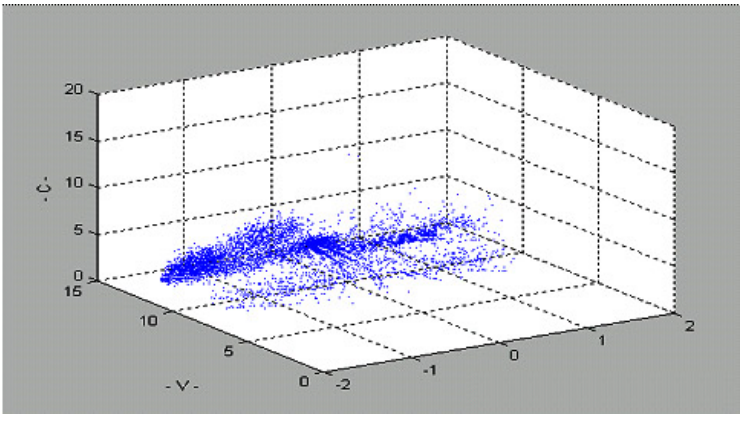


Fig. 2. Color distribution in the HVC color space

2 Color Space Selection

Although there are so many color-order systems, most of them are inconvenient to apply in segmentation because the color expressed is usually different from human perception. Among all the existing color systems, the Munsell color system is the closest to the human color vision [4].

The Munsell System describes all possible colors in terms of its three coordinates, Munsell Hue, Munsell Value, and Munsell Chroma [4]. A color in the Munsell color system can be written as HV/C. But, it is impossible to calculate color difference of two colors by such representation. We have to convert such representation into real

numbers. For this reason, we use the hue, value, chroma space(HVC color space) instead of the Munsell colors in terms of tri-attributes of human color perception[5]. In essential, the HVC color space is same as the Munsell color system, so the HVC color space also gives the best performance for experiments with variety of color spaces [6].

Before using the HVC color system to deal with images, we have to transform images from the RGB space to the HVC space. There are many ways to transform the images between the RGB space and HVC space. Here we use the improved mathematical transform of RGB coordinates to the HVC color system described in Gen et al[7]. Suppose *r*, *g*, *b* represent the three components red, green and blue in RGB color space. See [7] for the color transformation of RGB to HVC.

To calculate the color difference in the HVC color space, we make use of National Bureau of Standards (NBS) color distance [8] instead of the Euclidean distance measure.

It is found that in the HVC color space, the human color perception has close relation to the NBS color distance. The relation of human color perception and NBS distance is shown in Tab.1. From the table, we know that if the NBS color distance of two colors below 3.0, human beings will regard the colors almost the same.

Table 1. The correspondence between the human color perception and the NBS distance

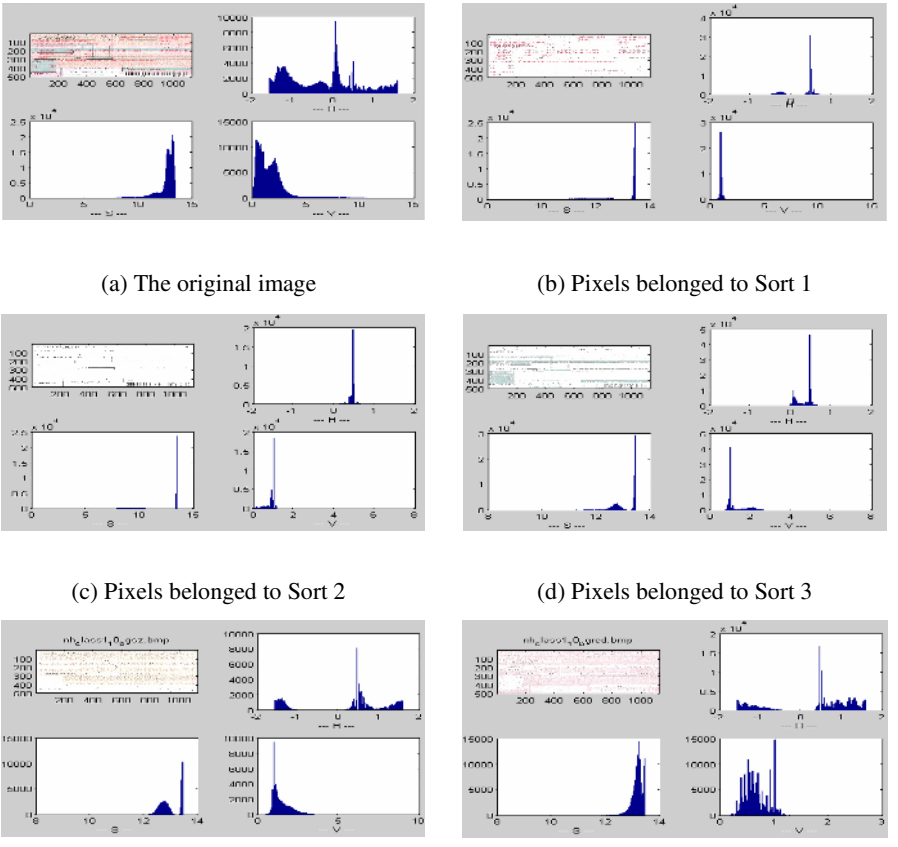
NBS Value			Human Perception
0	-	1.5	Almost the same
1.5	-	3.0	Slightly different
3.0	-	6.0	Remarkably different
6.0	-	12.0	Very different
12.0	-		Different color

3 Statistical Analysis for Pixels Color features

The original image to segment in our application is shown in Fig.3(a).

By analysis the colors appeared in the image, we classify most of the pixels into 5 fixed classes: red(object characters), black (background characters and form lines), green (background), yellow (background characters), and low red (background). There may be some uncertain color pixels in it, such as other smear spot or manual handwriting in it. Our object is to extract pixels in sort 1 and eliminate all other pixels.

In order to analysis the statistical characteristics of all kinds of pixels, we classified all these pixels in the image to the five sorts manually, as shown in Fig.3(b). to Fig.3(f).. After the transformation of the pixel values from RGB to HVC, the histograms of three coordinates of all sorts are showed in the corresponding parts in Fig.3. In the H parts, as shown in Fig.3, different sorts are located in different sections obviously. But in the V and C parts, this characteristic is not so obvious. This result is consistency with the design idea of airline coupon as we analyzed before, which use



(a) The original image

(b) Pixels belonged to Sort 1

(c) Pixels belonged to Sort 2

(d) Pixels belonged to Sort 3

(e) Pixels belonged to Sort 4

(f) Pixels belonged to Sort 5

Fig. 3. All sorts of pixels and histograms of their tri-attributes in a typical image

striking contrast colors to distinguish the different objects. So, the colors are mainly represented by hues.

Observing the H value distribution, we found it include two parts separated by 0. We classify the pixels into two parts by this threshold. By zooming in the image which just include two kinds of pixels, and observing carefully we can find that pixels in the second part whose H value is large than 0 is not the red object pixels indeed. The two kinds of pixels are in a much color difference. It means that the other sort pixels are classified into the object sort by mistake. By calculating means of the two groups of pixels, and calculating their NBS distance, the result shows that they are in a biggish difference, in slightly different level. So we are sure that the H value of sort 1 is in the section less then 0. We can also use this method to analysis data; it shows the advantage of our method. Further analysis in all coordinates reveals that the histogram of H value has obvious characteristics: the cohesion within sort 1 is strong and the distances between sorts are far. For example, we can separate sort 1 from sort 2 and 3 directly. But this characteristic is not so obvious for the V value and C value.

3.1 The Statistical Characteristic Analysis

Based on the analysis above, if we can find the probability distribution of the H-attribute of all sorts of pixels, then we can separate them by using the Bayesian Theorem.

Our research is based on the same sorts of airline coupons, which are chosen by visual quality. In this statistical analysis, our samples are came from the 5 sorts of color pixels, a sort of color pixels in an image form a sample, and there are 5 sorts of samples. We set the confidence in 0.05.

By hypothesis testing, we get results as follows:

1. Samples in sort *i(i=1,...,5)* are in the same distribution; this is the foundation to separate them with others.

2. The distribution samples in sort 1 do not obey the normal distribution.

According to [9], the color distribution in printing paper maybe obeys the normal distribution or Passion distribution. But in our research we find it is not suitable to our case. Seeing from the fitted curve, we can see that the curve is too steep and the distribution is too concentrated in the mean point. The deflection and steepness is not conformed to the characteristics with normal distribution. See Fig.4.

3. The distribution of samples in sort 1 is similar with the log normal distribution.

By the analysis and observation above, we think that the distribution of samples in sort 1 should obey the log normal distribution. But the K-S testing and the rank testing denied this hypothesis. So, we think it is not a standard distribution, just similar with the log normal distribution.

4. The samples in sort 1 can be separated from the other sorts by a threshold.

3.2 The Extraction of Pixels in Sort 1

The distributions of samples in sort 1 and sort 2, sort 3 is not overlapped. We can separate them directly. But the distribution of samples in sort 1 and sort 4, sort 5 have a little overlap, see Fig.5.

Seeing from the overlapped curves, we can separate samples in sort 1 from sort 4 and sort 5 by select a threshold. In practical, we select the intersection of the red curve

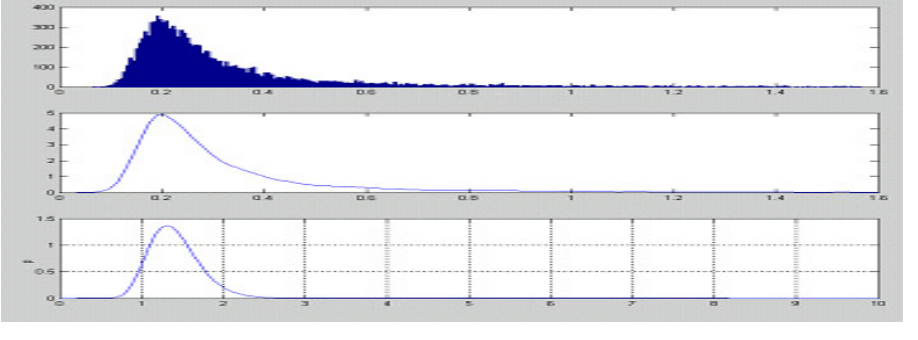


Fig. 4. Comparison of the estimative density distribution curve. Note: curves are symmetric flipped; Upper: the histogram of sort 1 pixels; Middle: the estimative density distribution curve of sort 1; Lower: the log normal distribution curve.

and the green curve as the theshold (T) to extract pixels in sort 1. The experiential value is -0.63 in our experiment, and the misclassification rate is under 0.05.

So, this method includes two parts, the statistical analysis part and classification part, which are describe as following:

Part 1: statistical analysis

1. Select a sort of airline coupon images, in which the pixels in sort 1 are in the same distribution; these images are come from a batch of coupon tickets.
2. Collect pixels in sort 1, 4 and 5 by manual visual observation.
3. Use the three sorts of samples to estimate their probability distribution curves, then determinate the threshold (T) by their intersection.

Part 2: Classification

1. Input a coupon image.
2. Calculate the H value (H_{ij}) in HVC space for each pixel P_{ij}.
3. Classify P_{ij} into foreground if $P_{ij}>T$; otherwise, classify P_{ij} into background.
4. Output the foreground image.

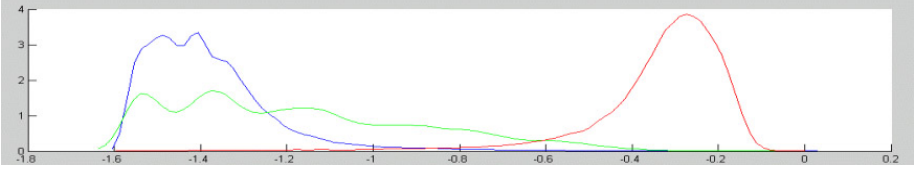


Fig. 5. The estimative density distribution curve of samples in sort 1(red), sort 4(blue) and sort 5(green)

3.3 Performance

In practical operation, different from the method in [2], which needs a complex training process, we just need to classify all pixels into two sorts: sort 1 pixels and the other. So it has the time complexity of $O(n)$, n is all the pixels in the original image, and need no extra space. It is a simple and effective method to extract strings.

4 Experiment and Results

We use our method to extract the object strings in the airline coupon project, which separate an image into two images of the same size, one is the object level and the other is the background level. We only concern on the object image.

By the subjective evaluation, we think this classification method is effective; the string objects image is cleaner compared to the result image of method in [2], see Fig.6.To prove its performance quantitatively, we use our evaluation system[10] to evaluate it., see formula 1.

$$u_f = C_f / |\mathrm{Fs}|;\ u_b = C_b / |\mathrm{Bs}|;\ u_t = \left[C_f + C_b\right] / |\mathrm{T}| \tag{1}$$

Where, c_b represent the counts of pixels which should be classified to background pixel set but have been classified to object pixel set, c_f represent the counts of pixels which should be classified to object pixels set but have been classified to background pixels set, F_s and B_s represent the object pixel set and the background pixel set of the standard segmentation image and $T = F_s \cup B_s$; u_f is the object pixel misclassified rate, u_b is the background pixel misclassified rate, and u_t is the total misclassified rate. The two indicators u_f and u_b are mainly used to analysis the algorithm and data in detail. The indicator u_t is mainly used to evaluate the integrated performance of the algorithm. The results are shown in Tab.2.

Table 2. Results comparing to method in [1]

	μ_b	μ_f	μ_t
Our method	0.0441	0.0526	0.0456
Method in [1]	0.0931	0.0486	0.0908

Seeing from the results, we can find that the method in [2] is a little better in the indicator of u_f. This is because of that its training samples are retained as more sort 1 pixel as possible. But, our method is much better in u_b and u_t. By all counts, it is a simple and effective method to extract strings from the airline coupon images.

Fig. 6. Subject effect comparing to method in [2]; Upper: the result images of method in [2]; Middle: the original images; Lower: the result images of our method

5 Conclusion

In this paper, we propose a simple method to extract strings from the airline coupon. It is based on the Munsell color system and on the statistical analysis. It is effective proved by practical work in the airline coupon project.

In the further research, we should study the method to determinate the threshold automatically and dynamically, for the H-attribute of the object pixels maybe drifted when the sorts of coupons increased.

References

1. S. Zhao, et al, "A High Accuracy Rate Commercial Flight Coupon Recognition System", Proc. of 7th International Conf. on Document Analysis and Recognition, 2003, Edinburgh, pp. 82-86.
2. Y. Li, et al, "String Extraction in Complex Coupon Environment Using Statistical Approach", Proc of 7th International Conf. on Document Analysis and Recognition, 2003, Edinburgh, pp. 289-294.
3. X. Wand and C.C. Kuo, "A new approach to image retrieval with hierarchical color clustering", IEEE Trans. on CSVT, vol.8, no.5, Sep., 1998.
4. F.W. Billmeyer and M. Saltzman, "Principles of Color Technology", 2nd Ed., New York, Wiley, 1981.
5. S.C. Pei and C.M. Cheng, "Extracting color features and dynamic matching for image database retrieval", IEEE Trans. on CSVT, vol.9, no.3, pp.501-512, Apr., 1999.
6. J. Hafner, H.S. Sawhney, W. Equitz, M. Flickner, and W. Niblack, "Efficient color histogram indexing for quadratic form distance functions", IEEE Trans. on PAMI, vol.17, no.7, pp. 729-736, July 1995.
7. M. Bartkowialk and M. Domanski, "Vector median filters for processing of color images in various color spaces", Proc. IEE Conference on Image Processing and Its Applications, 1995, pp. 4-6.
8. Y. Gong, G. Proietti, and C. Faloutsos, "Image indexing and retrieval based on human perception color clustering", Proc. IEEE Conference on Computer Vision and Pattern Recognition, Santa Barbara, CA, 1998, pp. 578-583.
9. X. Lu, Color Science in Encapsulation, ZhenZhou University Press, 2002.