2004 International Symposium on Nonlinear
Theory and its Applications (NOLTA2004)
Fukuoka, Japan, Nov. 29 - Dec. 3, 2004

Character string extraction from color documents using supervised learning

Takeshi Ogura†, Fumihiro Hasegawa‡, Shigeyuki Oba†, Toshio Miyazawa‡ and Shin Ishii†

†Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayama-cho, Ikoma-shi, Nara 630-0101, Japan
‡Software R&D Group, RICOH Company Ltd.
1-1-17 Koishikawa, Bunkyo-ku, Tokyo 112-0002, Japan
Email: { takesi-o, shige-o, ishii }@is.naist.jp , { fumihiro.hasegawa, toshio.miyazawa }@nts.ricoh.co.jp

Abstract—In this study, we deal with a pattern classification problem for extracting character string regions from digital color documents that contain both character images and other types of images. We applied supervised classification methods to data set represented in the feature space.

By utilizing the characteristics that especially the negative data distribute as to have a cluster structure, we devised a method using a mixture of classifiers. Additionally, we propose a parametric feature that allows to improve the discriminative performance by optimizing its own parameters.

As a consequence, we have achieved a high classification accuracy, which is as high as 96.5%, when applied to various color documents even with complicated layouts.

1. Introduction

In this study, we deal with a pattern classification problem for extracting character string regions from digital color documents which contain both character images and other types of images.

Recently, several algorithms to extract character strings from color documents have been proposed. Kasuga et al.[1] extracted a region of character strings as a cluster of blocks in the color space, based on the fact that blocks constituting a single character region are likely to have similar colors. However, this study was not of the supervised classification; namely, they did not consider the decision problem of the extracted character string regions. In addition, documents treated in their work were as small as a poster card and had relatively simple layouts. Similarly, Hase et al.[2] extracted character strings from color documents like cover pages of magazines. However, they were not successful in accurately discriminating character strings from background noises.

A system which is applicable to color documents with complicated layouts and is able to detect character string regions in high accuracy is therefore required.

2. Feature extraction and feature distribution

2.1. Feature extraction

We had produced a training data set and a test data set for detecting character string regions from various digital color documents.

We first extracted groups of connected pixels with similar colors in the horizontal and vertical directions, i.e., rectangles each of whose pixels have similar colors. These rectangles are candidates for character string regions. A human labeled each candidate in the sets, as positive or negative, when he regarded it as a character string or not, respectively. Although there may be mis-labels, we call a true character string a positive datum or a negative datum otherwise throughout this article.

Each candidate is next transformed into a 12-dimensional feature vector each of whose components corresponds to the image contrast, sparseness of connected pixels in circumscribed rectangles, the number of circumscribed rectangles, or so on. Here, a character string region is composed by several circumscribed rectangles.

2.2. Feature distribution

We applied principal component analysis (PCA) in order to visualize the distributions of positive and negative data in the 12-dimensional feature space. We randomly selected 2,000 samples from each of the positive and negative data sets, and normalized them so that each feature has the average of 0 and the variance of 1 in each of the subsets.

We then applied PCA in two ways.

1. Linear transformation by PCA is obtained based only on positive data and then both of the positive and negative data are projected by the transformation.

2. Linear transformation by PCA is obtained based only on negative data and then both of the positive and negative data are projected by the transformation.

Figure 1 shows cumulative contribution ratio of principal components obtained in the two ways above.

The first four principal components of positive data can explain about 80% of the distribution in the feature space,

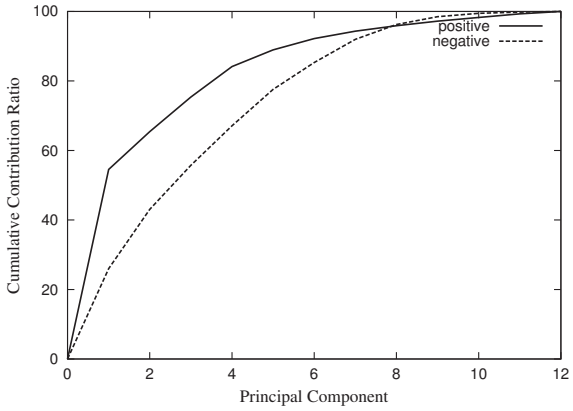

Figure 1: Cumulative contribution ratio

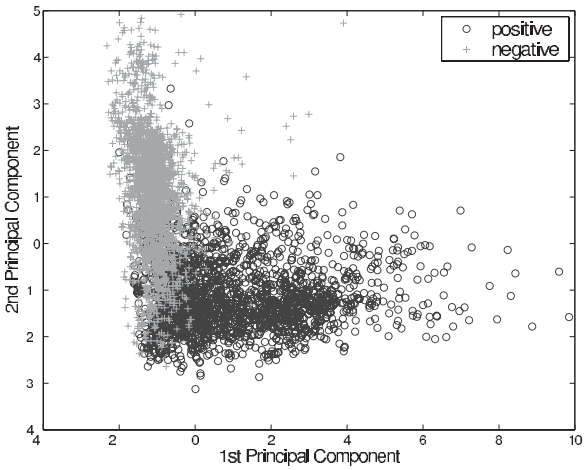

Figure 2: Scatter plot of the positive data and negative data

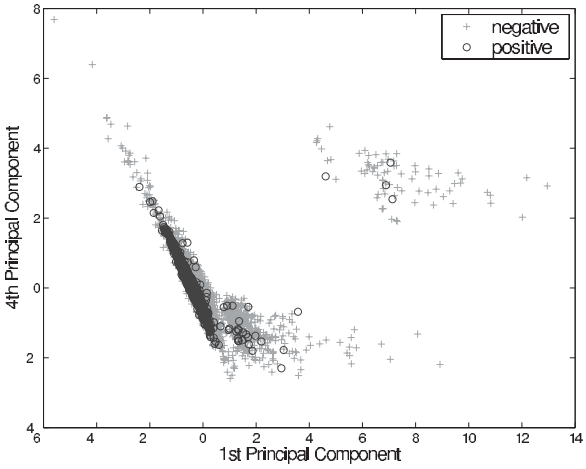

Figure 3: Scatter plot of the positive data and negative data

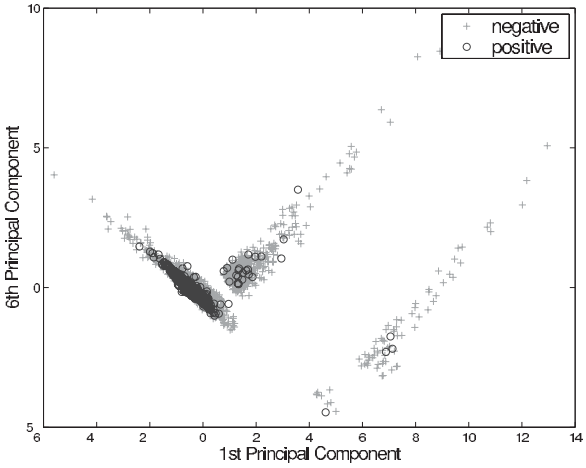

Figure 4: Scatter plot of the positive data and negative data

while six principal components are needed to explain the comparative amount of negative data. This implies that the distribution of the positive data has a smaller number of effective dimensionality than that of the negative data.

In addition, when we employed the projection plane consisting of the first and second principal components, positive and negative data had different distributions (Figure 2). The cumulative contribution ratio of the first two principal components was 71.9%.

Figure 3 shows the scatter plot of the positive and negative data projected on the two-dimensional plane consisting of the first and fourth principal components obtained based on the negative data. Similarly, Figure 4 shows the projection onto a plane consisting of the first and sixth principal components.

From these figures, it is obvious that there are three clusters, and each of the clusters has different distribution between the positive and negative data sets. By examining factor loading values (data not shown), we found that the three clusters have different characteristics of rectangular regions, namely, they correspond to horizontally-long rectangles, vertically-long rectangles and squares. In addition, we can see that positive data are mainly either of horizontally-long or vertically-long rectangles. We will later show a classification method utilizing these facts for the data distributions in the feature space.

3. Supervised Learning

First, we applied supervised classification methods by ignoring the cluster structures found in section 2.2.

We prepared 208 digital color documents that included 19,588 positive data and 72,027 negative data; these data sets were used for training. We also prepared 41 digital color documents that included 2,730 positive data and 35,495 negative data; these data sets were used for test (see

for training		for test	
208 documents		41 documents	
positive	negative	positive	negative
19,588	72,027	2,730	35,495

Table 1: Number of documents and data

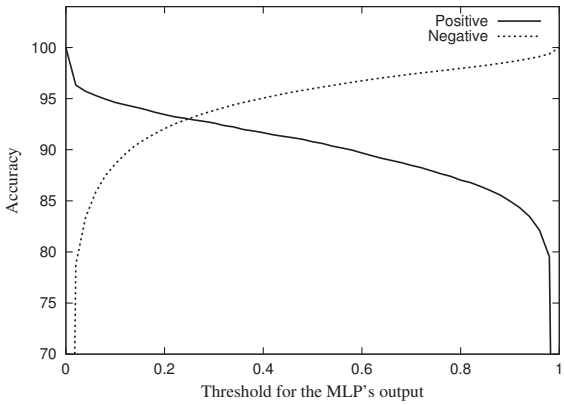

Figure 5: Multi-layered perceptron's ROC curve

Table 1). We normalized each feature so to have the average of 0 and the variance of 1 as a preprocess.

3.1. Multi-layered perceptron

We applied a multi-layered (two-layered) perceptron (MLP) possessing one hidden layer.

The number of hidden units was determined by repeating cross-validation ten times, in each of which training data set consisted of 15,000 positive data and 15,000 negative data selected from the training data set.

When we set 0.5 to the threshold for the MLP's output, the classification accuracy was 90.8%, 96.0%, and 95.6% for positive test data, negative test data, and both test data, respectively.

By changing the threshold value, the accuracy for the positive and negative data varied, which is summarized as the receiver operating characteristic (ROC) curve (Figure 5). From the ROC curve, one can see that when 95% accuracy for positive data is required, the accuracy for negative data becomes 87%.

3.2. Support Vector Machines

We also applied ν Support Vector Classifier (νSVC) [5][4] with a polynomial kernel:

$$k(x, x') = \langle x, x' \rangle^d. \quad (1)$$

Degree of the polynomial kernel and the smoothness hyper-parameter ν were determined by 10-fold cross validation as $d = 5$ and $\nu = 0.12$. As a result, the classification accuracy was 91.8%, 95.8%, and 95.5% for positive test data, negative test data, and both test data, respectively.

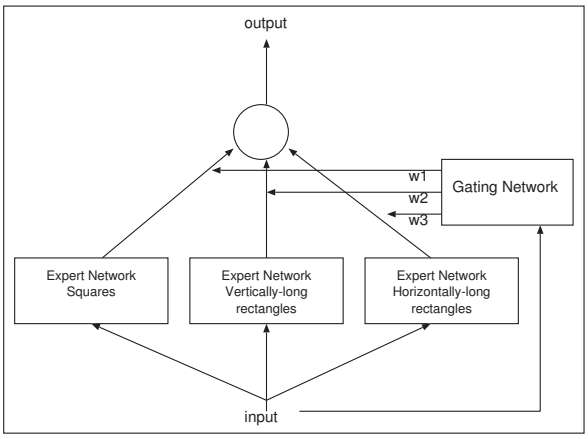

Figure 6: Structure of mixture of experts

3.3. Mixture of Experts

We have shown in section 2.2 that character string regions can be separated into three categories; horizontally-long rectangles, vertically-long rectangles, and squares. Because positive and negative data have different distributions in each category, these characteristics can be used for constructing a classifier with higher accuracy.

By examining error cases, we found that the supervised classifiers, MLP and νSVC, produced many misclassificaion cases for square regions, because they likely try to achieve high accuracy for vertical and horizontal rectangles due to their prominent features.

In order to achieve higher accuracy, we introduced a Mixture of Experts (MoE) model [3]. MoE is a divide-and-conquer algorithm in which the feature space is divided into several subspaces and each classifier is trained to be highly efficient in the subspace. In this study, we used an architecture composed of three expert networks and a gating network (Figure 6); the gating network is tuned by hard with considering the three clusters described in section 2.2.

Let H denote the height of a character string region and W the width. If $(1 - H/W)^2 < 0.1$ then the datum is processed by a classifier (expert network) to discriminate squares, if $H/W > 1$ then the datum is processed by a classifier for vertical rectangles, or by the one for horizontal rectangles otherwise.

4. Parametric feature

4.1. Parametric feature

Ordinarily, feature variables are selected arbitrarily so that each of them has different distributions for positive and negative samples. However, effectiveness of the variable will not appear until the learning is carried out for a certain supervised classifier. In particular, it is difficult to find out an efficient feature for the samples which have been incorrectly classified by the current classifier.

We therefore propose a parametric feature which possesses parameters as a flexible part of the feature. In this study, we determined the parameter values by a random search algorithm with cross-validation evaluation.

4.2. Rectangular Score

We devised a parametric feature called rectangular score which is calculated with respect to circumscribed rectangles constituting a character string region.

Let h denote the height of a circumscribed rectangle and w the width, in a character string region with the height of H, the width of W, and the area of A. Let a denote the number of connected pixels in the circumscribed rectangle.

With the notations: $v_1 = h/H, v_2 = h/w, v_3 = w/h, v_4 = w/W, v_5 = a/A, v_6 = A/a$, the rectangular score is defined as the sum of weighted value:

$$S = v_1w_1 + v_2w_2 + v_3w_3 + v_4w_4 + v_5w_5 + v_6w_6 \qquad (2)$$

accumulated over all circumscribed rectangles in the character region. The rectangular score is used as an additional feature for a character string region. The parameters in the new feature, $w_1, \ldots, w_6$, were determined by random searching with 4-fold cross validation so to have the highest accuracy. After the optimization of the parametric feature, training data sets were provided to a ν SVC with hyperparameters, d and ν, which were optimized by 10-fold cross validation.

In the recognition phase, the gating network categorized each test datum, and the expert network (ν SVC) for the corresponding category output the predicted label.

As a result, the classification accuracy was 91.6%, 96.8%, and 96.5% for positive test data, negative test data, and both test data, respectively.

Figure 7 shows examples of the character string regions extracted by our MoE model after training. Accordingly, highly accurate recognition of character string regions in color documents with complicated layouts has been achieved by our method.

5. Conclusion

In this study, we developed a supervised classification method to extract character string regions with high accuracy from color documents with complicated layouts. Because we found that the data constitute a cluster structure in the feature space, we employed an architecture consisting of three classifiers, i.e., a mixture of classifiers.

In addition, we devised a parametric feature whose parameters were optimized by cross-validation. By utilizing this new architecture and new feature, the classification accuracy has become as high as 96.5%.

We are planning to devise some other parametric features to achieve further high accuracy in our future work. The automatic acquisition of the gating network is in addition remained as our future work.

	Positive	Negative	ALL
MLP	90.8 %	96.0 %	95.6 %
SVM	91.5 %	95.8 %	95.5 %
MoE	91.6 %	96.8 %	96.5 %

Table 2: Results of supervised learning

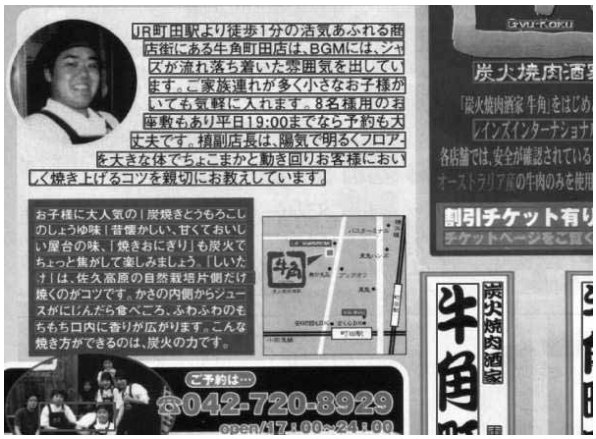

Figure 7: Examples of recognition from color documents

References

[1] H. Kasuga, M. Okamoto and H. Yamamoto, "Extraction of characters from color documents", *Proceedings of the SPIE - The International Society for Optical Engineering*, Vol.3967, pp.278-85, 2000.

[2] H. Hase, T. Shinokawa, M. Yoneda and C. Y. Suen, "Character string extraction from color documents", *Pattern Recognition* Vol.34, pp.1349-1365, 2001.

[3] R. A. Jacobs, M. I. Jordan, S. J. Nowlan and G. E. Hinton "Adaptive Mixtures of Local Experts", *Neural Computation*, Vol.3, pp.79-87, 1991.

[4] C. C. Chang and C. J. Lin "LIBSVM: a library for support vector machines", 2001. Software available at http://www.csie.ntu.edu.tw/˜cjlin/libsvm

[5] B. Schölkopf, A. Smola, R. C. Williamson and P. L. Bartlett, "New support vector algorithms", *Neural Computation*, Vol.12, pp.1207-1245, 2000.

[6] R. O. Duda, P. E. Hart and D. G. Stork "Pattern Classification", Wiley Interscience, 2000, 2nd edition.