# A Copula Based Statistical Model for Text Extraction from Scene Images

Ranjit Ghoshal[1], Anandarup Roy[2], and Swapan K. Parui[2]

[1] St. Thomas' College of Engg. and Technology
Kolkata- 700023, India
{ranjit.ghoshal.stcet,roy.anandarup}@gmail.com
[2] CVPR Unit
Indian Statistical Institute, India
swapan@isical.ac.in

**Abstract.** This article proposes a scheme for automatic text extraction from scene images. The work is composed of two steps. In the first step, we apply a model based color segmentation procedure in the LCH color space. This step produces certain homogenous connected components (CCs) from the image. In the next step, these CCs are examined in order to identify possible text components. A number of features that distinguish between text and non-text components, are defined. Further, during learning, these features are considered independently and approximated using parametric distribution families. Finally, the joint distribution of the features are constructed using a multivariate Gaussian copula. Consequently, we obtain two copula based class distributions for the two classes (text and non-text). Afterwards, during testing, a CC belongs to the class that produces the highest class distribution probability. Our experiments are on the database of ICDAR 2003 Robust Reading Competition. The experimental results are satisfactory.

## 1 Introduction

Automatic recognition of text portions in a scene image is useful to blind and foreigners with language barrier. Such a recognition methodology should also employ an extraction of text portions from the scene images. Segmentation of such text portions have a crucial impact on document processing, content based image retrieval, robotics etc. There have been several studies on text segmentation in the last few years. Wu et al. [1] use a local threshold method to segment texts from gray image blocks containing texts. images. Recently, Jung et al. [2] employed a multi-layer perceptron classifier to discriminate between text and non-text pixels. Recently Ghoshal et al. [3] proposed a scheme based on analysis of connected components (CCs) for recognition of Bangla text from scene images through perspective correction. Also a few criteria for robust filtering of text components have been studied.

We first employ a statistical mixture model based clustering algorithm on the input color image. With the assumption that text portions are homogeneous in

P. Maji et al. (Eds.): PReMI 2013, LNCS 8251, pp. 489–494, 2013.

color and lightness, different clusters may contain text portions as different CCs. We further study these CCs and define various features that are used to distinguish between text and non-text components. We consider the text identification as a two class problem. Each class (i.e., text and non-text) is approximated by a copula based distribution whose margins are the individual feature distributions. So, here the features are allowed to follow different parametric families. In the testing phase, we compare the probability of each CC against these two classes. The CC is classified in the class that has higher probability. Here, we use the public database of ICDAR 2003 Robust Reading Competition.

## 2 Color Image Segmentation

The primary task in this section is segmentation of the input color image. For this task we consider LCH color space. The hue component is angular, whereas chroma and the lightness components are linear. Hue can be represented by a random variable $\Theta \in [0, 2\pi)$. Thus, the LCH space has a circular linear characteristics. The random variable $\Theta$, here, is assumed to follow a wrapped Gaussian distribution [4]. To accommodate the circular-linear characteristics, a suitable modeling should employ a distribution that is wrapped in the angular dimension and non-wrapped in the linear dimensions. The semi-wrapped Gaussian distribution [5] serves this purpose well enough. Roy et al. [6] discussed a statistical mixture model (SWGMM) with semi-wrapped Gaussian distributions. In this paper, we apply SWGMM in order to segmentation of an input image.

## 3 Connected Component Analysis

The image segmentation produces a number of CCs with one cluster producing one or more CCs. These CCs include the possible text portions. So, we analyze these CCs to identify text portions. We assume that a single text component is homogeneous in terms of color and lightness. This assumption ensures that a single text component is not broken after clustering. Now, after segmentation, the text parts may make one single cluster. However, more generally, one cluster contains non-text components along with some text components. In order to separate them we proceed as follows. We first remove sufficiently small and large components. Further, we extract the following CC based features to distinguish between text and non-text portions.

**AR:** The aspect ratio $AR = (height/width)$ of a non-text component is either very small or very large compared to text components.

**OBR:** The object to background pixels ratio (OBR) is computed by taking the bounding box. Due to the elongated nature of texts, only a few object pixels fall inside text bounding box. This ratio is usually larger for non-texts.

**ER:** The text like patterns are usually elongated. We use a measure of elongatedness (ER) of a component defined by Roy et al. [7].

**TH:** Thickness (TH) of a CC is computed from Ghoshal et al. [3].

**H:** Usually text like patterns contain less number of holes than that of non-text patterns. Using the Euler number we calculate the number of holes (H) inside a component.

Combining these features, we construct the 5-dimensional feature vector $\boldsymbol{\Upsilon} = (AR, OBR, ER, TH, H)$ for a CC.

## 4 Text Identification Method

We manually group the CCs into two classes, the text class ($C_{text}$) and non-text class ($C_{non-text}$). We now intend to obtain the feature distributions for these two classes separately. In this regard, the most popular model is the multivariate Gaussian distribution. However, the features under consideration here may not follow a Gaussian distribution. Hence, a non-Gaussian multivariate distribution may be a suitable choice to approximate the distribution of the present features. However, a drawback of using such a multivariate distribution is that all the margins usually follow distributions from the same family. For example, all the margins are Gaussian in case of a multivariate Gaussian distribution. In real situations, it is possible that the individual features may follow different families of distributions, independently. A suitable multivariate distribution should allow the marginal distributions to follow different families of distributions. The known families of multivariate distributions (for example, Gaussian or Dirichlet distributions) do not have this characteristic. However, there is an alternative construction of a multivariate distribution. In multivariate statistics, the copula approach is often taken to model the dependence between two or more random variables. Concerning the bivariate case, the copula approach to dependence modeling was first stated in a theorem due to Sklar [8]. Let $H$ be a joint distribution function with marginal distributions $F$ and $G$. Then there exists a copula $C$ such that for all $x, y \in [-\infty, \infty]$,

$$H(x, y) = C(F(x), G(y)). \tag{1}$$

Accordingly, the probability density function (pdf) of $H(x, y)$ is defined as $h(x, y) = \frac{\delta^2 H(x,y)}{\delta x \delta y}$. Here, $C(u, v)$ is a mapping $[0, 1] \times [0, 1] \rightarrow [0, 1]$, termed as copula in the sense that it couples the random variables $X$ and $Y$. The advantage of copula is that using the knowledge of the margins only, one can construct the joint distribution, accommodating however complex form of dependence structure, on the basis of different types of copulas. Sklar's theorem, we can construct the multivariate distribution incorporating different margin families. Let us start by specifying the marginal distributions. We consider each feature of a class, independently. Then we fit a series of parametric distributions on each feature. The best fitted distribution that maximizes the likelihood value, is selected for that feature. Here, we use an archive of distributions consisting of Gaussian, Gamma, Log-Normal and generalized extreme value (abbreviated by gextreme) distributions. Note that we obtain these distributions for each of the two classes

(namely, text and non-text) separately. To combine individual feature distributions, we use the multivariate Gaussian Copula [8]. Let $\boldsymbol{u} = \{u_1, \ldots, u_d\}$ be a $d$-dimensional (here $d = 5$) vector consisting of the distribution functions of all the margins. Here $u_i$ denotes the distribution function of the $i^{th}$ feature. Then $\boldsymbol{u} \in [0,1]^d$ according to the probability integral transform. The $d$-dimensional Gaussian copula is defined over $[0,1]^d$. It is constructed from a multivariate Gaussian distribution over $\Re^d$ by using the probability integral transform. Given the correlation matrix $\Sigma \in \Re^{d \times d}$, the density of a Gaussian copula can be written as

$$c_\Sigma(\boldsymbol{u}) = \frac{1}{\sqrt{|\Sigma|}} \exp\left( -\frac{1}{2} \begin{pmatrix} \Phi^{-1}(u_1) \\ \vdots \\ \Phi^{-1}(u_d) \end{pmatrix}^T (\Sigma^{-1} - \boldsymbol{I}) \begin{pmatrix} \Phi^{-1}(u_1) \\ \vdots \\ \Phi^{-1}(u_d) \end{pmatrix} \right), \tag{2}$$

where $\Phi^{-1}$ is the inverse cumulative distribution function of a standard Gaussian distribution and $\Phi_\Sigma$ is the joint cumulative distribution function of a multivariate Gaussian distribution with mean vector zero and covariance matrix $\Sigma$. $\boldsymbol{I}$ is the identity matrix. We now combine all the five distributions for a class (say, for $C_{text}$ class) and define the class distribution as:

$$f(C_{text}) = c_\Sigma(u_1, \ldots, u_5) \prod_{i=1}^{5} f_i. \tag{3}$$

Here, $f_i$ is the density of the $i^{th}$ feature. In a similar fashion, the probability distribution function $f(C_{non-text})$ for the non-text class can be computed. Finally, we classify a component to text class if $f(C_{text}) > f(C_{non-text})$.

## 5 Results and Discussion

Let us now present the segmentation and text extraction results. Here, we use the public database of ICDAR 2003 Robust Reading Competition. For the experiments, we select 200 images, randomly from this database. Further, we apply the SWGMM for segmenting the images into a set of homogenous CCs. A few sample images with segmentation results are presented in Table 3. We observe that our segmentation approach could preserve the text like components. For the purpose of training, we manually label the CCs of the training images to construct $C_{text}$ and $C_{non-text}$. Our training set has 4000 samples of text components and 10000 samples of non-text components. We first perform a five-fold cross validation over the training data to assess the performance of our model. We get the cross validation accuracy 66.15% for $C_{text}$ and 68.33% for $C_{non-text}$. Later, we obtain the distributions for the text and non-text classes. The individual features and their corresponding distributions are presented in Table 1. Note that most of the time we obtain the generalized extreme value family as a suitable distribution. In Table 1, we elaborate the parameter values of the distributions. In gextreme distribution, the first parameter (i.e., $\xi$) governs the tail behavior. The sub-family distributions corresponding to $\xi > 0$ and $\xi < 0$

are Fréchet and Weibull families respectively ($\xi = 0$ does not occur here). The other gextreme parameters $\mu$ and $\sigma$ control the location and scale of the sub-family distribution. We also obtain Log-Normal distribution for the feature TH for the class $C_{text}$. The Log-Normal parameters are $\alpha$ and $\beta$ that control the location and the shape of the distribution respectively. Then, we estimate the

**Table 1.** Parametric distribution correspond to individual features

| Feature | $C_{text}$ | $C_{non-text}$ |
|---|---|---|
| AR | gextreme $(\xi = 0.1812, \sigma = 0.4743, \mu = 1.1071)$ | gextreme $(\xi = 0.5747, \sigma = 0.5119, \mu = 0.5397)$ |
| OBR | gextreme $(\xi = 0.3046, \sigma = 0.5304, \mu = 0.7791)$ | gextreme $(\xi = 0.4825, \sigma = 0.8545, \mu = 0.9742)$ |
| ER | gextreme $(\xi = -0.0781, \sigma = 2.6221, \mu = 9.9923)$ | gextreme $(\xi = 0.2138, \sigma = 2.5187, \mu = 8.5799)$ |
| TH | Log-Normal $(\alpha = 3.4281, \beta = 0.5264)$ | gextreme $(\xi = 0.5102, \sigma = 7.5015, \mu = 10.9475)$ |
| H | gextreme $(\xi = 4.2755, \sigma = 0.0242, \mu = 0.0057)$ | gextreme $(\xi = 4.8188, \sigma = 0.1585, \mu = 0.0329)$ |

Gaussian copula parameter, i.e., the correlation matrix $\Sigma$. We plug in the previously estimated marginal distribution into the copula to obtain a maximum likelihood estimation of the correlation matrix. In Table 2, we present the correlation matrices obtained for the text and the non-text classes. Finally, after

**Table 2.** Correlation matrices of the Gaussian copulas for (a) $C_{text}$ and (b) $C_{non-text}$

| | | | | |
|---|---|---|---|---|
| 1 | -0.19 | -0.32 | -0.37 | -0.30 |
| -0.19 | 1 | -0.13 | 0.25 | 0.04 |
| -0.32 | -0.13 | 1 | 0.63 | 0.43 |
| -0.37 | 0.25 | 0.63 | 1 | 0.32 |
| -0.30 | 0.04 | 0.43 | 0.32 | 1 |

(a)

| | | | | |
|---|---|---|---|---|
| 1 | 0.05 | -0.09 | 0.18 | 0.01 |
| 0.05 | 1 | 0.20 | 0.41 | 0.12 |
| -0.09 | 0.20 | 1 | 0.49 | 0.47 |
| 0.18 | 0.41 | 0.49 | 1 | 0.46 |
| 0.01 | 0.12 | 0.47 | 0.46 | 1 |

(b)

obtaining the distribution we now provide the test samples. We put each CC against the two classes and assign the component in class $C$ for which $f(C)$ (Eq. 3) is larger. Some of the images that are the extracted text components (third row) are shown in Table 3.

We observe from Table 3 that often some non text components are included in the text class. On the other hand, some text components are still missing. However, during segmentation, some of the text components are not separated from the non-text portions. Such text portions are essentially not included in the extracted text components. We obtain satisfactory text extraction results.

**Table 3.** Some images (first row), the corresponding segmentation results (second row) and extracted text components (third row)


## 6 Summary and Future Scope

This article provides an automatic extraction of text entities from scene images. It is based on color image segmentation followed by extraction of several features that lead towards identification of text components. One primary objective is to obtain statistical models for feature distribution. We observe that different families of distributions can better approximate the distributions of individual features. However, it is quite possible that correlation is present among the features. So, we incorporate correlation among different features using multivariate Gaussian copula. However, the Gaussian copula omits the tail dependence and is thus less ambitious in modeling complex types of correlations. Our later work will address this issue and explore the use of other copula models.

## References

1. Wu, V., Manmatha, R., Riseman, E.M.: Textfinder: An automatic system to detect and recognize text in images. IEEE Trans. Pattern Anal. Mach. Intell. 21(11), 1224–1229 (1999)
2. Jung, K., Kim, I.K., Kurata, T., Kourogi, M., Han, H.J.: Text scanner with text detection technology on image sequences. In: Proc. of Int. Conf. on Pattern Recognition, vol. 3, pp. 473–476 (2002)
3. Ghoshal, R., Roy, A., Parui, S.K.: Recognition of bangla text from scene images through perspective correction. In: Proc. ICIIP, pp. 385–390 (2011)
4. Mardia, K.V., Jupp, P.: Directional Statistics. John Wiley and Sons Ltd. (2000)
5. Bahlmann, C.: Directional features in online handwriting recognition. Pattern Recognition 39, 115–125 (2006)
6. Roy, A., Parui, S.K., Nandi, D., Roy, U.: Color image segmentation using a semi-wrapped gaussian mixture model. In: Kuznetsov, S.O., Mandal, D.P., Kundu, M.K., Pal, S.K. (eds.) PReMI 2011. LNCS, vol. 6744, pp. 148–153. Springer, Heidelberg (2011)
7. Roy, A., Parui, S.K., Paul, A., Roy, U.: A color based image segmentation and its application to text segmentation. In: Proc. of Ind. Conf. on Computer Vision, Graphics & Image Processing, pp. 313–319 (2008)
8. Nelsen, R.B.: An Introduction to Copulas. Springer (2006)