Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition 41 (2008) 484–493
PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/pr

Gaussian mixture modeling and learning of neighboring characters for multilingual text extraction in images

Xiabi Liu [a,*], Hui Fu [a,b], Yunde Jia [a]

[a] School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China
[b] School of Information Technology, Beijing Forestry University, Beijing 100083, China

Received 27 August 2006; received in revised form 22 March 2007; accepted 15 June 2007

Abstract

This paper proposes an approach based on the statistical modeling and learning of neighboring characters to extract multilingual texts in images. The case of three neighboring characters is represented as the Gaussian mixture model and discriminated from other cases by the corresponding 'pseudo-probability' defined under Bayes framework. Based on this modeling, text extraction is completed through labeling each connected component in the binary image as character or non-character according to its neighbors, where a mathematical morphology based method is introduced to detect and connect the separated parts of each character, and a Voronoi partition based method is advised to establish the neighborhoods of connected components. We further present a discriminative training algorithm based on the maximum–minimum similarity (MMS) criterion to estimate the parameters in the proposed text extraction approach. Experimental results in Chinese and English text extraction demonstrate the effectiveness of our approach trained with the MMS algorithm, which achieved the precision rate of 93.56% and the recall rate of 98.55% for the test data set. In the experiments, we also show that the MMS provides significant improvement of overall performance, compared with influential training criterions of the maximum likelihood (ML) and the maximum classification error (MCE).

Keywords: Text extraction; Document analysis; Character recognition; Image retrieval; Gaussian mixture models; Discriminative training; EM algorithm

1. Introduction

Texts in images are important semantic features of images, which often reveal the contents of images, or even themselves are the contents of images. Therefore, it is significant to extracting texts in images for many applications, such as content-based image retrieval (CBIR), automatic video logging, optical character recognition (OCR), document analysis, etc. In these fields, text extraction has received a great deal of attention. However, the problem is still considered open since the direction, color, arrangement, and background of texts are subject to wide variations in images.

Traditional text extraction methods can be classified into two categories: region-based and texture-based [1]. Region-based methods extract texts according to the difference between text regions and other regions. These methods often work with a bottom-up strategy, where substructures are identified and merged to mark text regions [2,3]. Texture-based methods employ distinct textural properties of texts compared with their backgrounds. In these methods, the textural properties of an image are often detected by using techniques of Gabor filters, wavelet, etc. [4,5]. People further fused the region-based and texture-based strategies for overcoming their own disadvantages [6]. Recently, Zhang and Chang proposed a part-relation based approach for scene text detection, and reported their experiments on English texts [7]. Although much progress has been made, the wide variations of texts in font, color, arrangement, and background are still obstacles to satisfactory results of text extraction. Besides, multilingual text cannot be extracted by using a single method.

In this paper, we propose a new text extraction approach to handle multilingual text and complicated cases such as multicolor text, text arranged in a curve, etc. We model the feature vector of three neighboring characters to be of the

* Corresponding author. Tel./fax: +86 10 86343158.
E-mail addresses: liuxiabi@bit.edu.cn (X. Liu), fuhuir@bjfu.edu.cn (H. Fu), jiayunde@bit.edu.cn (Y. Jia).

0031-3203/$30.00 © 2007 Pattern Recognition Society. Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2007.06.004

distribution of Gaussian mixture model (GMM). We then define and apply the 'pesudo-probability' under Bayes framework to discriminate between the case of three neighboring characters and other cases. Accordingly, the text in an input image is extracted in three steps. Firstly, the separated parts of each character in the binary image is connected by a mathematical morphology based method, in order that each character can be treated as a connected component (CC); secondly, all CCs in the binary image are detected, and their neighborhoods are established by partitioning the image into Voronoi regions of centroids of CCs; finally, each CC is label as character or not according to the pseudo-probabilities of all three neighboring characters which is involved in. We further develop a discriminative training algorithm based on the maximum–minimum similarity (MMS) criterion [8] to estimate the parameters in the proposed text extraction approach. In order to test the effectiveness of our approach trained with the MMS algorithm, we apply it to extract Chinese and English text in images. In the experiments, the precision rate of 93.56% and the recall rate of 98.55% were achieved for the test data set, which show that our approach is effective and promising. Another two training criterions, including the conventional informative training criterion of the maximum likelihood (ML) and the commonly used discriminative training criterion of the minimum classification error (MCE) [9], were used in the same experiments. The ML criterion was implemented by the expectation maximization (EM) algorithm [10]. Among three training criterions, the MMS behaved best. It provides 15.4% increase in the recognition rate, compared with the ML criterion, and 13.9% increase in the recognition rate, compared with the MCE criterion. To some extent our idea of exploring the relation between neighboring characters is similar to that in the fuzzy curve-tracing (FCT) algorithm of Yan [11,12]. The FCT algorithm can be used to detect curved text paths in noisy scanned text string images. However, the target of the FCT algorithm is the curve path along which the characters are placed. It does not deal with the problem of discriminating characters from non-characters.

The rest of this paper is organized as follows. Section 2 introduces the GMM based method of discriminating characters from non-characters. Section 3 describes the text extraction process in detail. Section 4 presents the MMS training algorithm for our text extraction approach. Finally, we provide the experimental results and conclusions in Sections 5 and 6, respectively.

2. Discriminating characters from non-characters

The key of text extraction is to discriminate the character regions from non-character regions. In this work, we employ the relation between three neighboring characters to solve this problem. The case of three neighboring characters is found to be distinguished from other cases by the following features:

x_1: the consistency of distances between centroids of characters. For most text strings in images, no matter characters are arranged in a line or in a curve, distances between adjacent characters are approximately equal. Let $\{A, B, C\}$ denote the centroids of three neighboring characters. Without lose of generality, we assume $\|A-B\| \leqslant \|A-C\|$, and $\|B-C\| \leqslant \|A-C\|$, where $\|\cdot\|$ is Euclidean norm. Then the distance consistency of three neighboring characters is measured as

$$x_1 = \frac{\|A-B\|}{\|B-C\|}. \tag{1}$$

x_2: the consistency of region areas of characters. Similar with distance between characters, region areas of neighboring characters are often approximately equal. We measure the area consistency of three neighboring characters by the ratio

$$x_2 = \frac{\max(Area_A, Area_B, Area_C)}{\min(Area_A, Area_B, Area_C)}, \tag{2}$$

where $Area_A$, $Area_B$ and $Area_C$ are region areas of three characters, respectively.

x_3: the region density which is defined as the ratio of the foreground pixels to the background pixels. The region density of a character is usually different from that of other objects. We compute the mean region density of three neighboring characters as the third feature, i.e.

$$x_3 = \frac{(density_A + density_B + density_C)}{3}. \tag{3}$$

Then the feature vector $\mathbf{X} = \{x_1, x_2, x_3\}$ of three neighboring characters is assumed to be of the distribution of GMM. The GMM is a general model for estimating an unknown probability density function and under regular conditions it may approximate any continuous function having a finite number of discontinuities [13]. So the GMMs are among the most widely used statistical models in the pattern recognition community. Let C denote the case of three neighboring characters, $\boldsymbol{K}$ be the number of components in GMM, w_k, $\boldsymbol{\mu_k}$ and $\boldsymbol{\Sigma_k}$ be the weight, the mean vector and the covariance matrix of the kth Gaussian component, respectively, $\sum w_k = 1$, then we have

$$p(\mathbf{X}|C) = \sum_{k=1}^{K} w_k N(\mathbf{X}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{4}$$

where

$$N(\mathbf{X}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = (2\pi)^{-d/2} |\boldsymbol{\Sigma}_k|^{-1/2} \times \exp(-\tfrac{1}{2}(\mathbf{X}-\boldsymbol{\mu}_k)^{\mathrm{T}} \boldsymbol{\Sigma}_k^{-1} (\mathbf{X}-\boldsymbol{\mu}_k)). \tag{5}$$

The use of a GMM with full covariance matrices leads to a huge number of parameters and presents the risk of over-fitting. Therefore, the covariance matrices are often constrained to be spherical or diagonal [14]. In this paper the diagonal covariance matrices are employed, i.e. $\boldsymbol{\Sigma_k} = [\sigma_{kj}]_{j=1}^{3}$. The experimental results reported in Section 5 confirm that the GMM with diagonal covariance matrices is adequate for our problem.

Using Bayes' formula, we get

$$P(C|\mathbf{X}) = \frac{p(\mathbf{X}|C)P(C)}{p(\mathbf{X})}. \tag{6}$$