计算机研究与发展
Journal of Computer Research and Development

ISSN 1000-1239/CN 11-1777/TP
44(11): 1920～1926, 2007

图像中多语种文本提取的高斯混合建模方法

付 慧[1,2] 刘峡壁[2] 贾云得[2]
[1](北京林业大学信息学院 北京 100083)
[2](北京理工大学计算机科学与技术学院 北京 100081)
(fuhuir@bjfu.edu.cn)

Gaussian Mixture Modeling of Neighbor Characters for Multilingual Text Extraction in Images

Fu Hui[1,2], Liu Xiabi[2], and Jia Yunde[2]
[1](*School of Information Technology, Beijing Forestry University, Beijing* 100083)
[2](*School of Computer Science and Technology, Beijing Institute of Technology, Beijing* 100081)

Abstract A new method based on the Gaussian mixture modeling of neighbor characters is proposed to extract multilingual texts in images. In the training phase, the Gaussian mixture model of three neighbor characters is trained from the examples. Then the texts in an input image are extracted in the following steps. Firstly, the image is binarized using the edge-pixel clustering method and the morphological closing operation is performed on the binary image, in order that each character in it can be treated as a connected component. Secondly, the neighborhood of connected components is established according to the Voronoi partition of the image. Three connected components neighboring with each other constitute a neighbor set. For each neighbor set, a posteriori pseudo-probability is computed based on the Gaussian mixture model of three neighbor characters and used to classify the neighbor set as the case of three neighbor characters. Finally, the text extraction is completed by labeling the connected components as characters or non-characters with the following rule: if a connected component is included in at least one neighbor set classified as the case of three neighbor characters, then the connected component is labeled as a character, or else as a non-character. The proposed method are tested in the applications of Chinese and English text extraction. In the experiments, the expectation-maximization algorithm is employed to train the Gaussian mixture model of three neighbor characters. The experimental results of text extraction show the effectiveness of the method.

Key words document analysis; optical character recognition (OCR); text extraction; image retrieval; Gaussian mixture modeling (GMM)

摘 要 建立了相邻字符区域的高斯混合模型,用于区分字符与非字符.在此基础上,提出了一种从图像中提取多语种文本的方法.首先对输入图像进行二值化,并执行形态学闭运算,使二值图像中每个字符成为一个单独的连通成分.然后根据各连通成分重心的 Voronoi 区域,形成连通成分之间的邻接关系;最后在贝叶斯框架下,基于相邻字符区域的高斯混合模型计算相应的伪概率,以此为判据将每个连通成分标注为字符或非字符.利用所提出的文本提取方法,进行了复杂中英文文本的提取实验,获得大于97%的准确率和大于80%的召回率,证实了方法的有效性.

收稿日期:2006-05-26;修回日期:2007-07-02
本文通讯作者:刘峡壁;E-mail:liuxiabi@bit.edu.cn
基金项目:国家自然科学基金项目(60473049);国家"九七三"重点基础研究发展规划基金项目(2006CB303105);北京理工大学优秀青年教师资助计划基金项目(2006Y1202)

关键词 文档分析;光学字符识别(OCR);文本提取;图像检索;高斯混合模型

中图法分类号 TP391.4

文本提取是文本图像分析与识别领域中的一个重要问题,在基于内容的图像检索、自动视频记录、光学字符识别(OCR)等许多方面起着关键性的作用.由于图像中的文本区域存在很多变化,如方向不定、颜色不一致、排列不规则等等,因此实现文本自动提取相当困难.

目前已有文本提取方法大致可以分为两类:基于区域的方法和基于纹理的方法[1-2].基于区域的方法利用文本区域与背景区域在颜色或灰度特性上的不同来进行提取,通常采用自底向上的策略,即首先辨识子结构,然后通过子结构的合并来标记文本区域.根据所采用的子结构的不同,基于区域的方法又可分为基于连通成分的方法和基于边缘的方法.基于纹理的方法利用文本区域与背景区域在纹理特性上的不同来进行提取,可以采用 Gabor 变换、小波变换等方法实现.这些方法主要针对图像中的规则文本,如文本颜色一致[3]、文本排列成直线[4]等等.此外,目前还缺乏描述多语种文本对象的统一模型,提取不同语种文本须采用各自不同的特征和方法.最近,Zhang 和 Chang[5]提出了字符串的马尔可夫随机场(MRF)模型,用于场景图像中颜色一致的英文文本提取,取得较好效果.

Zhang 和 Chang 的工作表明,基于字符相互关系模型的方法具有较好的推广性和较高的鲁棒性,因此本文也利用字符区域之间的相互关系,实现文本提取,所提出的文本提取方法可用于处理多语种文本和复杂文本,如混色文本和成曲线排列的文本等等.我们将 3 个相邻字符之间的关系特征建模为高斯混合模型(GMM),并根据该模型区分字符与非字符.在此基础上,通过以下 3 步完成文本提取:1)对输入图像进行文本区域初始定位和二值化[6],然后对每个候选文本块的二值图像做形态学闭运算,以使每个字符对应一个连通成分;2)检测二值图像中所有连通成分,按照每个连通成分重心的 Voronoi 区域建立连通成分之间的邻接关系;3)在贝叶斯框架下,根据所建立的相邻字符区域的高斯混合模型,计算相应伪概率,以此为判据将每个连通成分标注为字符或非字符.与 Zhang-Change 方法[4]相比,本文方法中所采用的 GMM 模型与 MRF 模型一样有效,但更简单,学习和提取效率更高.同时,本文利用 Voronoi 分割确定图像区域之间的邻接关系,避免了经验性的邻域选择方法,具有更高的鲁棒性.为了测试本文所提出的文本提取方法,我们针对图像中的复杂中英文文本进行了提取实验,获得大于 97% 的准确率和大于 80% 的召回率,表明本文所提出的方法是有效的.

1 用于区分字符与非字符的高斯混合建模方法

文本提取的关键在于如何区分字符区域与非字符区域.以往工作多根据文本区域本身的特性来进行区分,本文则利用相邻字符区域之间的关系特征来解决这一问题.

经过形态学闭运算,可以使二值图像中的每个字符对应一个单独的连通成分(有关计算方法请参见第 2 节),则图像中的连通成分或者对应于一个字符区域,或者对应于一个非字符区域.我们按如下方式定义连通成分之间的相邻性:如果两个连通成分的覆盖范围是连通的,则它们是相邻的.这里,连通成分的覆盖范围是指由距离该连通成分最近的所有像素点构成的子图像.根据上述相邻连通成分的定义考察图像中的字符区域与非字符区域,发现相邻字符区域具有以下关系特征.

1) x_1:字符重心间距的一致性

图像中的大多数文本串,无论是成直线排列还是成曲线排列,相邻字符之间的距离是近似相等的.设 $\{A,B,C\}$ 表示 3 个相邻字符的重心,不失一般性,我们假设 $\|A-B\| \leqslant \|A-C\|$,并且 $\|B-C\| \leqslant \|A-C\|$,则 3 个相邻字符之间的间距一致性可以度量为

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

2) x_2:字符区域面积的一致性

与相邻字符重心间距的一致性相似,相邻字符区域的面积常常也是近似相等的.设 $Area_A$,$Area_B$ 和 $Area_C$ 分别表示 3 个相邻字符的区域面积,则 3 个相邻字符之间的面积一致性可以度量为

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

3) x_3:区域密度

字符区域密度定义为区域内的前景像素与区域

面积的比值. 字符区域的密度与非字符区域的密度之间存在差异. 我们计算 3 个相邻字符的平均区域密度作为第 3 维特征:

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

综合起来,采用 $\boldsymbol{x} = \{x_1, x_2, x_3\}$ 作为 3 个相邻区域对应的特征矢量. 我们假设 3 个相邻字符区域的 $\boldsymbol{x}$ 服从高斯混合分布. 高斯混合模型(Gaussian mixture models)是一种半参数化的密度估计方法,它结合了参数化和非参数化方法的优点,能够逼近具有有限间断点的任意连续密度[7],是模式识别领域中经常采用的统计模型之一. 设 C 表示 3 个字符相邻的情况,$\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, w_k$ 分别表示第 k 个高斯成分的均值向量、协方差矩阵和在高斯混合模型中的权值,则有

$$p(\boldsymbol{x} \mid C) = \sum_{k=1}^{K} w_k N(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \quad (4)$$

其中

$$N(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = (2\pi)^{-\frac{d}{2}} \mid \boldsymbol{\Sigma}_k \mid^{\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^{\mathrm{T}} \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_k)\right). \quad (5)$$

我们将 3 个彼此相邻的区域分为两种情况,或者它们都是字符,或者其中至少一个不是字符. 设 $P(C)$ 表示 3 个彼此相邻的区域都是字符的概率,$p(\boldsymbol{x})$ 表示 $\boldsymbol{x}$ 的分布,则根据贝叶斯公式,我们得到

$$P(C \mid \boldsymbol{x}) = \frac{p(\boldsymbol{x} \mid C) P(C)}{p(\boldsymbol{x})}. \quad (6)$$

显然,用概率 $P(C \mid \boldsymbol{x})$ 来判别 $\boldsymbol{x}$ 是否属于 3 个字符相邻的情况是合理的. 在式(6)中,$p(\boldsymbol{x} \mid C)$ 可以从 3 个相邻字符的样本中估计出来,但是由于反例(非文本)情况太复杂,以致无法得到代表所有反例的典型样本,因此学习 $P(C)$ 和 $p(\boldsymbol{x})$ 很困难. $\boldsymbol{x}$ 表示根据任意 3 个彼此相邻的区域所获得的特征向量,既包括 3 个区域都是字符的情况,也包括至少一个区域不是字符的情况. 而非字符区域中的内容可以是除字符以外的其他任何图像. 因此,$\boldsymbol{x}$ 的取值范围很广,其观测值可能等概率地出现于特征空间的任意一个点上. 这说明可以假设 $\boldsymbol{x}$ 的分布是均匀的. 于是,$P(C)$ 和 $p(\boldsymbol{x})$ 对于所有的 $\boldsymbol{x}$ 来说为恒定值,因此

$$P(C \mid \boldsymbol{x}) \propto p(\boldsymbol{x} \mid C). \quad (7)$$

式(7)说明,可以用一个以 $p(\boldsymbol{x} \mid C)$ 为自变量,光滑、单调递增,且值域在[0,1]之间的函数来估计 $P(C \mid \boldsymbol{x})$. 我们将这样的函数值称为伪概率,并采用如下的函数形式来计算伪概率:

$$\rho(p(\boldsymbol{x} \mid C)) = 1 - \exp(-\alpha p(\boldsymbol{x} \mid C)). \quad (8)$$

根据式(8),$\rho(p(\boldsymbol{x} \mid C))$ 随着 $p(\boldsymbol{x} \mid C)$ 的增加而增加,并且当 $p(\boldsymbol{x} \mid C) = 0$ 时,$\rho(p(\boldsymbol{x} \mid C)) = 0$;当 $p(\boldsymbol{x} \mid C) = +\infty$ 时,$\rho(p(\boldsymbol{x} \mid C)) = 1$. 因此,$\rho(p(\boldsymbol{x} \mid C))$ 是以 $p(\boldsymbol{x} \mid C)$ 为自变量,光滑、单调递增,且值域在[0,1]之间的函数,可以用来计算伪概率,作为对 $P(C \mid \boldsymbol{x})$ 的近似.

综上所述,对于任意 3 个彼此相邻的区域,计算其特征向量 $\boldsymbol{x} = \{x_1, x_2, x_3\}$,然后用式(8)计算相应伪概率,如果伪概率值大于 0.5,则认为这 3 个相邻区域均为字符区域,否则认为其中至少存在一个非字符区域,并不再对其中某一区域的特性进行判断.

上述 GMM 模型方法的局限在于需用于由 3 个及 3 个以上字符所组成的文本. 由于图像中的大多数文本均表示有意义的语句,其中包含的字符个数往往大于 3,因此这种局限并不会明显影响本文方法的可用性.

2 文本提取方法

为了将第 1 节中所述相邻字符区域的 GMM 模型用于文本提取,需要获得图像中每个字符对应的区域. 因此,首先对输入图像进行文本初始定位和二值化. 这里采用了基于边缘像素聚类的文本区域初始定位和二值化方法[6],可以保证二值图像中文本的完整性,但其中也存在许多非文本区域,即方法准确性较低. 然后,对所获得的二值图像做形态学闭运算,以使得二值图像中的每个字符对应一个单独的连通成分. 考虑到字符在图像中可能出现的多种形态,我们采用了具有不同大小和方向的结构元素执行闭运算,从而获得多个新二值图像. 图 1 显示了在不同尺度下进行形态学变换的相应结果.

在这些新二值图像和原始二值图像中,基于第 1 节所述的相邻字符区域的 GMM 模型将连通成分标注为字符或非字符,从而完成文本提取. 如果新二值图像中的连通成分被标注为字符,那么在原始二值图像中确定该连通成分对应的区域,并将原始二值图像中这一区域内的所有连通成分都标注为字符. 如图 2 所示,图 2(a)是原始二值图,图 2(b)是根据图 2(a)得到的一幅新二值图. 图 2(b)中的方框表示在该二值图像中所标定的一个文本区域. 图 2(a)中的灰色像素表示这一文本区域在原始二值图中所对应的连通成分,这些连通成分都被标注为字符.

(a) (b) (c)

Fig. 1 Some results of closing operations with different scales. (a) Original images; (b) The results of closing operations with scale 1; and (c) The results of closing operations with scale 2.

图 1 不同尺度的形态学闭运算结果. (a) 原始图像;(b) 对应于尺度 1 的形态学闭运算结果;(c) 对应于尺度 2 的形态学闭运算结果

(a) (b)

Fig. 2 The illustration of the correspondence between the labeling result in the new binary image and that in the original binary image. (a) The original binary image, in which three connected components of Chinese character '江' are labeled as characters and shown in light grey, because they are contained in the region corresponding with the rectangle in Fig.2(b) and (b) The new binary image, in which the connected component of Chinese character '江' is labeled as a character and indicated by the rectangle.

图 2 原二值图和新二值图连通成分标注对应关系. (a) 原二值图,其中"江"用灰色像素显示,表示这些连通成分位于图 2(b)中的方框所对应的区域,都被标注为字符区域;(b) 新二值图,其中方框表示一个被标注为文本的连通成分

对于每个二值图像,连通成分的标注包括两步:1) 计算连通成分之间的邻接关系;2) 基于相邻字符区域的 GMM 模型标注每个连通成分. 下面分别进行详细的阐述.

2.1 计算连通成分之间的邻接关系

连通成分之间的邻接关系根据连通成分重心的 Voronoi 区域来确定. 实际上,我们计算 Voronoi 分割的对偶问题,即重心集合的 Delaunay 三角剖分来建立邻接关系[8],以更为直观地反映连通成分之间的邻接关系. 在 Delaunay 三角剖分中,对应一个三角形的 3 个连通成分构成一个邻域集. 图 3 和图 4 给出两个计算邻接关系的例子. 在两个图中,图3(a)与图 4(a)是完成闭运算后的文本图像,其中粗黑点表示连通成分的重心;图 3(b)与图 4(b)显示了重心的 Delaunay 三角剖分,其中每个三角形对应一个邻域集. 然而,Delaunay 三角剖分不能完全反映字符之间的邻接关系,存在两种特殊情况,我们分别在图 3(c)和图 4(c)中显示这两种情况以及相应的解决方法.

根据图 3(b)中的 Delaunay 三角剖分结果,可以得到 3 个邻域集,其中没有反映出文本串中间 3 个相邻字符的情况,即缺失 $\triangle BCD$. 为了解决这类问题,我们将重心点集对应凸壳上所有相邻 3 个点均作为邻域集,如图 3(c)中的点线所示.

在图 4 所示例子中,由于图 4(a)中上下两条曲线的干扰,该图中所有字符邻接关系都没有反映在图 4(b)中的 Delaunay 三角剖分结果中. 我们通过定义二阶邻接关系来处理这种情况. 首先将上述处理后得到的所有邻域集统称为一阶邻接关系,则二阶邻接关系定义如下:以图 4(b)中的点 A 为例,它的一阶邻域集是 $\{A, D, E\}$. 我们将 A 的邻接关系

分别扩展到点 D 和 E 的一阶邻接关系,从而得到点 A 的二阶邻域集 $\{A,B,C\}$ 和 $\{A,C,F\}$. 其他连通成分的二阶邻域集采用同样方法计算. 图 4(c)用点线表示了在图 4(b)基础上扩展的所有二阶邻域集.

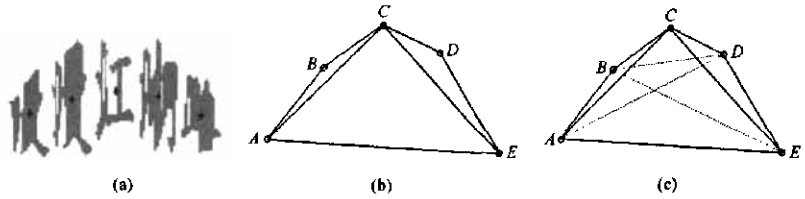

Fig. 3 Illustration of the first-order neighborhood computation. (a) The text image after closing operation; (b) The Delaunay triangulation of centroids; and (c) The expanded neighbor sets through joining one by one in the convex hull of the centroid set.

图 3 一阶邻接关系计算示意图. (a) 执行闭运算后的文本图像;(b) 重心的 Delaunay 三角剖分;(c) 通过将重心点集对应凸壳上所有相邻 3 个点均作为邻域所获得的扩展邻域集

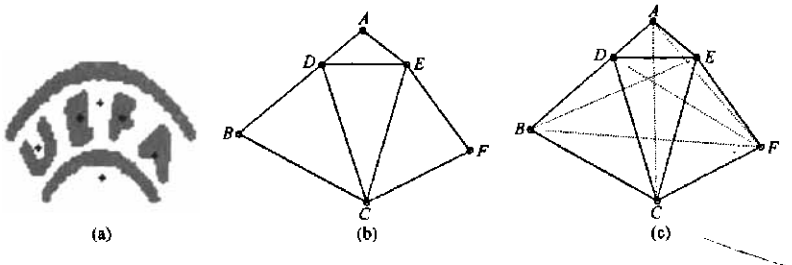

Fig. 4 Illustration of the second-order neighborhood computation. (a) The text image after closing operation; (b) The Delaunay triangulation of centroids; and (c) The expanded neighbor sets after the second-order neighborhood computation.

图 4 二阶邻接关系计算示意图. (a) 执行闭运算后的文本图像;(b) 重心的 Delaunay 三角剖分;(c) 计算二阶邻域后所获得的扩展邻域集

2.2 连通成分标注

获得二值图像中所有连通成分对应的一阶和二阶邻域集之后,根据第 1 节介绍的高斯混合模型方法判断每个邻域集是否对应于 3 个字符相邻的情况.如果一个连通成分至少属于一个对应于 3 个相邻字符的邻域集中,则标注该连通成分为字符,否则标注它为非字符.

3 实验结果与分析

为了测试本文所提出的文本提取方法,我们应用它提取图像中的中英文文本. 实验中,采用 50 幅英文文本图像和 40 幅中文文本图像作为训练集,另外 20 幅图像作为测试集. 以上图像中,10 幅测试图像来自 ICDAR 2003 测试集[9],其余的训练和测试图像均来自自己建立的数据库.

对于每幅训练图像,手工确定其中所有 3 个字符相邻的情况,总共得到 343 个英文训练样本和 167 个中文训练样本. 相邻字符区域的 GMM 模型通过期望最大化(EM)算法[10]从这些样本中学习得到,其中高斯成分个数 K 通过实验方式确定为 3. 在提取实验中,我们采用了两个方向(垂直和水平)和 5 种尺寸的结构元素执行二值图像形态学闭运算.

利用 EM 算法学习结果,我们分别进行了中英文文本提取的开放测试和封闭测试,总体实验结果见表 1. 表中采用召回率 R 和准确率 P 评估算法性能. 设 L 表示图像中真实字符对应的连通成分集合,I 表示提取方法所确定的对应于字符的连通成分集合,则召回率和准确率计算公式如下[11]:

$$p=\frac{|L\cap I|}{|I|},\ R=\frac{|L\cap I|}{|L|}, \tag{9}$$

此外,表 1 中 NT 表示图像中真实非文本连通成分的集合.

Zhang 和 Chang 针对英文文本对他们所提出的基于马尔可夫随机场字符串模型的文本提取方法进行了实验,所报告的最佳结果约为准确率 90%,召回率 85%[4]. 如表 1 所示,我们在测试数据上得到了 97.16% 的准确率和 80.54% 的召回率. 从准确率和召回率上看,本文所提出的文本提取方法与 Zhang-Chang 方法性能基本相当,但 Zhang 和 Chang

仅测试了颜色一致的英文文本的提取,而我们的实验结果是针对中英文两种文本的,并且其中包括混色文本、成曲线排列的文本、嵌入在图形中的文本等复杂情况.

Table 1 The Experimental Results

表 1 实验结果

Images for Text Extraction	$\|L\|$	$\|J\|$	$\|L\cap J\|$	$\|NT\|$	$P(\%)$	$R(\%)$
Training Images(English)	445	384	375	29	97.66	84.27
Training Images(Chinese)	251	224	220	30	98.21	87.65
Test Images	298	247	240	88	97.16	80.54

图 5 显示了部分测试图像及其提取结果,其中矩形框包含部分为提取结果. 从图中可以看出,本文所提出的方法具有处理多语种文本和复杂文本的能力. 但是,实验结果也反映出本文方法不能有效处理相邻字符区域之间面积差异显著的情况,这是本文方法召回率不够理想的主要原因,也是下一步将要努力的方向.

Fig. 5 The extraction results for some test images.

图 5 部分测试图像及其文本提取结果

4 结 论

本文提出了一种基于相邻字符区域高斯混合模型的图像中文本提取方法,可用于处理多语种文本和复杂文本. 我们用高斯混合模型表示相邻字符区域之间关系特征的统计规律,并通过 EM 算法从样本中学习得到. 对于输入图像,首先进行文本区域初始定位和二值化;然后对二值图像进行形态学闭运算,使图像中每个字符成为一个单独的连通成分;进而利用 Delaunay 三角剖分建立连通成分之间的邻接关系;最后根据相邻字符区域的高斯混合模型和在贝叶斯框架下定义的伪概率,将每个连通成分标注为字符或非字符. 我们针对图像中复杂中英文文本的提取对本文方法进行了测试,实验结果显示本文方法是有效的. 在后续工作中,我们将研究更为稳定的关系特征,以增强方法的适应性. 同时,准备收集更多训练样本和采用判决学习方法[12],以进一步提高准确率和召回率,特别是召回率.

参 考 文 献

[1] Keechul Jung, Kwang In Kim, Anil K Jain. Text information extraction in images and video: A survey [J]. Pattern Recognition, 2004, 37(5): 977-997

[2] Mi Congjie, Liu Yang, Xue Xiangyang. Video texts tracking and segmentation based on multiple frames [J]. Journal of Computer Research and Development, 2006, 43(9): 1523-1529 (in Chinese)

(密聪杰，刘洋，薛向阳. 基于多帧图像的视频文字跟踪和分割算法[J]. 计算机研究与发展，2006，43(9)：1523-1529)

[3] H Hase, T Shinokawa, M Yoneda, *et al*. Character string extraction from color document [J]. Pattern Recognition, 2001, 34(7): 1349-1365

[4] V Wu, R Manmatha, E M Riseman. TextFinder: An automatic system to detect and recognize text in images [J]. IEEE Trans on Pattern Analysis Machine Intelligence, 1999, 21(11): 1224-1229

[5] Dong-Qing Zhang, Shih-Fu Chang. Learning to detect scene text using a higher-order MRF with belief propagation [C]. In: Proc of the IEEE Conf on Computer Vision and Pattern Recognition Workshops (CVPRW'04). Los Alamitos: IEEE Computer Society Press, 2004. 101-108

[6] Fu Hui, Liu Xiabi, Jia Yunde. Edge-pixel clustering for text area extraction [J]. Journal of Computer-Aided Design & Computer Graphics, 2006, 18(5): 729-734 (in Chinese)
(付慧，刘峡壁，贾云得. 用于文本区域提取的边缘像素聚类方法[J]. 计算机辅助设计与图形学学报，2006，18(5)：729-734)

[7] Perry Moerland. A comparison of mixture models for density estimation [C]. In: Proc of the 9th Int'l Conf on Artificial Neural Networks (ICANN'99). London: IEE Press, 1999. 25-30

[8] J E Beasley, F Goffinet. A Delaunay triangulation-based heuristic for the Euclidean Steiner problem [J]. Networks, 1994, 24(14): 215-224

[9] S M Lucas, A Panaretos, L Sosa, *et al*. Icdar 2003 robust reading competitions [C]. In: Proc of the 7th Int'l Conf on Document Analysis and Recognition. Berlin: Springer-Verlag, 2003. 682-687

[10] J A Bilmes. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models [R]. U C Berkeley, Tech Rep: 97-021, 1998

[11] D Karatzas, A Antonacopoulos. Text extraction from Web images based on a split-and-merge segmentation method using colour perception [C]. In: Proc of the IEEE 17th Int'l Conf on Pattern Recognition (ICPR'04). Los Alamitos: IEEE Computer Society Press, 2004. 634-637

[12] Biing-Hwang Juang, Wu Hou, Chin-Hui Lee. Minimum classification error rate methods for speech recognition [J]. IEEE Trans on Speech and Audio Processing, 1997, 5(3): 257-265

Fu Hui, born in 1978. Ph. D. and lecturer. Her current research interests include image processing and pattern recognition.
付 慧，1978 年生，博士，讲师，主要研究方向为图像处理与模式识别.

Liu Xiabi, born in 1972. Ph. D. and lecturer. Member of China Computer Federation. His current research interests include pattern recognition, machine learning, and multimedia information retrieval.
刘峡壁，1972 年生，博士，讲师，中国计算机学会会员，主要研究方向为模式识别、机器学习和多媒体信息检索.

Jia Yunde, born in 1962. Ph. D. and Professor. Senior member of China Computer Federation. His current research interests include computer vision, artificial intelligence, and intelligent systems.
贾云得，1962 年生，博士，教授，中国计算机学会高级会员，主要研究方向为计算机视觉、人工智能和智能系统(jiayunde@bit.edu.cn).

Research Background

Along with the substantial increase in the use of digital cameras, Web-cams, and camera-enabled mobile devices, camera-based text and document analysis and recognition is becoming an attractive research area. The applications include content-based image indexing, automatic video logging, capturing and extracting text for storing or translating notices, transforming text to audio for the visually-impaired, for navigation by reading road signs, and many more.

In camera-based text and document analysis and recognition, text extraction is a key problem, which is very difficult because text in images has wide variations in direction, color, arrangement, and background. Therefore, most existing methods of text extraction aim at regular text such as consistent text color and line-arranged text. In order to deal with the complicated cases, character-relation based approaches seem promising.

In this paper, we introduce a novel statistical modeling approach to character-relation based multilingual text extraction. We use the Gaussian mixture modeling (GMM) of three neighbor characters to discriminate between characters and non-characters. The GMM of three neighbor characters is learned from examples by an expectation-maximization (EM) algorithm. According to this modeling, each connected component in the image is classified as character to realize the text extraction. Our approach has been applied in a demonstration system of camera-based text and document analysis and recognition. Its effectiveness is confirmed by the experimental results. This work is partially supported by the 973 Program of China (2006CB303105) and the Excellent Young Scholars Research Fund of Beijing Institute of Technology (2006Y1202).

图像中多语种文本提取的高斯混合建模方法

作者： 付慧， 刘峡壁， 贾云得， Fu Hui， Liu Xiabi， Jia Yunde

作者单位： 付慧,Fu Hui(北京林业大学信息学院,北京,100083;北京理工大学计算机科学与技术学院,北京,100081)， 刘峡壁,贾云得,Liu Xiabi,Jia Yunde(北京理工大学计算机科学与技术学院,北京,100081)

刊名： 计算机研究与发展 ISTIC EI PKU

英文刊名： JOURNAL OF COMPUTER RESEARCH AND DEVELOPMENT

年，卷(期)： 2007，44(11)

被引用次数： 1次

参考文献(12条)

1. Keechul Jung. Kwang In Kim. Anil K Jain Text information extraction in images and video:A survey 2004(05)

2. 密聪杰. 刘洋. 薛向阳 基于多帧图像的视频文字跟踪和分割算法[期刊论文]-计算机研究与发展 2006(09)

3. H Hase. T Shinokawa. M Yoneda Character string extrsction from color document 2001(07)

4. V Wu. R Manmatha. E M Riseman TextFinder:An automatic system to detect and recognize text in images 1999(11)

5. Dong-Qing Zhang. Shih-Fu Chang Learning to detect scene text using a higher-order MRF with belief propagation 2004

6. 付慧. 刘峡壁. 贾云得 用于文本区域提取的边缘像素聚类方法[期刊论文]-计算机辅助设计与图形学学报 2006(05)

7. Perry Moerland A comparison of mixture models for density estimation 1999

8. J E Beasley. F Goffinet A Delaunay triangulation-based heuristic for the Euclidean Steiner problem 1994(14)

9. S M Lucas. A Panaretos. L Sosa Icdar 2003 robust reading competitions 2003

10. J A Bilmes A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models[Tech Rep:97-021] 1998

11. D Karatzas. A Antonacopoulos Text extraction from Web images based on a split-and-merge segmentation method using colour perception 2004

12. Biing-Hwang Juang. Wu Hou. Chin-Hui Lee Minimum classification error rate methods for speech recognition 1997(03)

相似文献(1条)

1. 学位论文 桑伯男 中文科技文档中数学公式的抽取 2007

随着计算机和互联网的发展，越来越多的资料被以文档图像的形式存储到计算机上。通过网络进行信息的存储、查找和传播也越来越成为当前信息流通的主要渠道。如何快速、高效地将这些文档图像转化为可编辑的格式成为急需解决的问题，文档图像分析技术作为一个新的研究领域应运而生。

光学字符识别(OCR)是文档图像分析的核心技术。现有的OCR系统对打印字符已经能做到很高的识别率。而数学公式由于其存在二维结构，单纯通过扩充识别系统字库无法完全记录公式图像所含全部信息。如何将打印科技文档中的公式进行定位、识别和重组，依然是一个正在研究中的课题。虽然已经提出了多种算法，但这些算法大部分是针对英文环境下的文档。由于中英文在字库技术，字符连通体构成上的诸多不同，简单地将英文环境下算法移植到中文环境下会产生大量错误，且没有利用中文文档的特点，是不可取的。

本文首先在绪论中介绍了文档图像分析技术，以及模式识别和神经网络等相关领域的背景知识。在定位数学公式的时候，本文给出的新算法需要对数学符号进行识别。第二章主要介绍了利用Zemike距提取字符的特征，由自组织特征映射(SOFM)神经网络和BP神经网络组成多分类器进行符号识别的技术。

第三章首先回顾了当前一些应用于英文环境中的公式定位算法，提出了这些算法在应用于中文科技文当时会出现的问题，讨论了标记连通体这一当前文档分析技术中非常依赖的技术。并对中文字符的特点，中文文档排版的特点，人类阅读方式，及科技文档中普遍存在的公式分布局部性进行了讨论。

在此基础上，本文提出了一种新的算法，该算法采用输入框组并行的读入目标，并判定其是否是规则汉字，从而规避了标记连通体步骤。并且利用了公式分布的局部性，对不同密度采用速度不同的算法，从而提高了整体公式定位速度。对于算法中遇到的各种具体问题，包括输入框标准的确定，汉字的确认，排版微调造成的所占空间的小差异等等，都给出了具体的解决方法。

在本文的最后部分，分析了系统中仍然存在的问题，并讨论了新系统未来的扩展方向。

引证文献(1条)

1.黄百钢.李俊山.胡双演 基于颜色和笔画特征的文本分割算法[期刊论文]-计算机科学 2009(7)