指导教师：杨涛　　提交时间：2016.3.17

CVPR2015 Paper Translation

No: 1

姓名：郭浩瀚

学号：2013302356

班号：10011301

独立于笔迹的线上手写识别

Oendrila Samanta*, Anandarup Royt, Ujjwal Bhattacharya* and Swapan K. Parui*
*CVPR Unit, Indian Statistical Institute, 203 B. T. Road, Kolkata, India
Email: oendrila.payel@gmail.com.ujjwal@isical.ac.in.swapan@isical.ac.in
tEcole de Tech. Superieure,University of Quebec,Montreal, Quebec, Canada
Email: roy.anandarup@gmail.com

摘要：手写体的格式普遍会融合草体书，而且这是自动识别中最难的书写体。而目前在不同的语种例如英语，阿拉伯语，孟加拉语等中，手写体都非常受欢迎。但是，当语种的字母表包含大量字母（例如孟加拉语有350个字母）时，手写识别的难度会显著增大。目前，HMM（隐马尔可夫模型）就是用于解决相似识别问题的最流行的结构。然而如果依靠于特殊应用的基础词汇表被给出，那么任务就会变的简单。因为在这种情况下，我们可以采用全局或基于单词的识别，从而并不需要去识别单一的字母。从另一个角度去讲，当词汇表的内容增多，或者包含了较多的相似形状的字母是，识别任务就会变得复杂。在一个最近对这种相似情况的研究[1]中，全联系非同质HMM模型就被用于处理观测序列已经被明确分隔的情况。目前研究表明，我们已经发现基于HMM的语音识别模式的表现独立于笔迹和特定的智能分词策略。我们采用基于DISCRETE CURVE EVOLUTION ALGORITHM(离散曲线演化算法)[2]的文本分词模式和两个其他的已经存在的分词模式，并以英文，阿拉伯文和孟加拉文的数据库为评判标准，来得出上述结论。同事，利用统计假设测试的模式结果更加证明了上述结论。

I.介绍

无论是线下模型还是线上模型，数字化手写数据都是可取的手段。在线下模型中，数据可用图片等以矩阵为格式的方式获取，但它不能和线上模型一样保护笔移动的时间信息。而时间模型中，数据有一个序列格式用于保存以时间为轴的笔的移动坐标序列。这种在线上模型中最先进的识别速度通常要高于同等的线下模型[6]。基于各形各色的电子笔装置的欢迎程度的增强，线上识别研究最近已经获得了很大的动力。这些识别研究追求两个方面，整体性和分析性[3]。在一个分析通道中，手写单词的边界要在他在整体模式中识别之前被辨别出来，自从识别应用于单词级别，我们不再需要识别字母之间的边界。在这里，需要注意手写样例有不同的字体[3]包括正体，草体，混合正体和草体的字体等用于自动识别。在本工作中，我们主要处理线上混合字体的整体识别。

整体识别从定义上说不需要从手写单词的不同的子部分中进行特征提取。它通常计算代表单词整体形状的全局特征。然而，随着词汇表中词汇量的增长或者单词之间相似度增加，这种全局特征也许会表现的很差。实际上，全局特征识别方法是被用来对已知受限的词汇表[9]进行精确识别的方法。那么，为了维持即使在负面情况下也要保持的期望的识别精确度，目前的策略是把输入单词分割成一系列子部分，并对每一个子部分进行特征计算。这保证了两个相思形状的单词通过一个或两个字母的不同获得在观测序

列中的映射。如此就需要在单词分界的地方的分割点而不是其他可能的分割点

分隔是分析识别方法的基本组成部分［7］，而且这部分受分隔歧义影响，因为特征边界需要从分隔点的建立中辨认出来。然而，他不是全局特征识别问题的分割模型的问题。因为全局特征方法不识别独立的字母。基于滑动窗口的特征计算已经被使用于大量的手写单词分析识别的研究中。在这些方法中，明确的手写分割可以不被需要。尽管这个方法已经成功的模拟于几种字迹，例如拉丁文等，但是在对其他一些例如孟加拉文，Devanagari 等的字体并不是一个有效的方法，因为这里出现一些经常发生的延笔的问题。在这些笔迹中，不同的字母经常在手写的时候被放在单词上面或者下面插入，或用 \ / 的方法插入。

不同的线上手写的整体识别方法可见于［5］［8］［10］［11］［12］。他们使用了不同的工具例如神经网络，支持向量机，HMM 等方法作为辨别装置。应用于全局研究的手写数据包括了几组有相似形状的单词和由于相思问题大面积出错的单词[8]。如此，尽管全局识别任务比分析识别要简单，他们仍需要关注于这种复杂情况。［1］中基于 HMM 的方法被不断的应用于去处理和别的方法一样要处理的的相似单词组。

在线上手写识别中使用的 HMM 的动机源于它们在语音识别研究中获得的成功。跟随他在语音识别领域的实践，通常只允许两种状态转移（自身和右边）的简单的从左至右的 HMM 已经大面积的应用于手写识别研究中。他们不仅仅允许无状态跳转，还允许单状态转移。对于更多 HMM 的产生模型很难在研究文献中见到。对于这种情况，一种可能的原因就是没有获取为实施更多归纳的 HMM 模型时而需要的工具。然而，在一个最近的研究中，全联系非同质的 HMM 模型被用来给

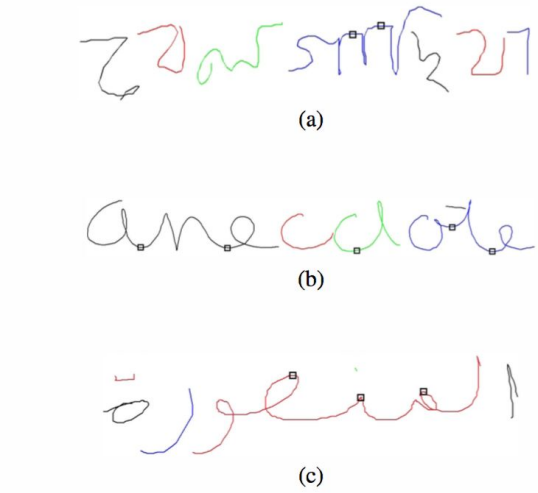

Fig. 1. Unconstrained handwriting samples of (*a*) Bangla, (*b*) English and (*c*) Arabic words. Expected segmentation points are shown by small squares.

每个单词类建模，而且该拟合结果表示这些 HMM 模型在处理一些复杂情况时是有效的。

在这篇文章中，我们描述了一个应用于线上手写样例分词的新的策略。并且，我们评估了基于 HMM 的识别策略（见［1］）和不同分词策略对不同笔迹的性能。研究中我们找到了一个有意思的发现，对于任何不需要忍受分词级别以下的问题的特殊分词策略，在最终的识别精确度上都没有很大的成效。那么，这篇文章，一方面，提出了一个新颖的对于线上自由手写单词的分词策略，另一方面，它探索了基于三种不同笔迹的多种不同分词策略的效果。

本文剩余部分由如下部分组成。第二部分列举了研究中使用的预处理工作。共三种分词策略详细记述在第三部分。基于 HMM 的识别模式祥述在第四部分。第五部分展示了实验结果。最后，我们在第六部分进行了总结。

II. 预处理

一个输入样例要经过如下预操作序列：（i）消除重复点，（ii）用三点滑动平均法

平滑化，（iii）重采样，（iv）使单词的核心部分的高度通过上下对称调整的方法正常化。最后，输入样例的点转换成以 y 轴为方向使得核心部分在 x 轴方向上对齐。

III. 分词方法

就我们在第一部分所讨论的，现在的基于 HMM 的识别方法包括一个明确的可以产生一系列子块的分词模块。这些字块的特征表现组成了 HMM 所用到的观测序列。在结果的分割后单词中，任意两个字母没有在字块中重合。一个自由的带有英语，孟加拉语，阿拉伯语字迹和他们被期望的分词点的手写单词样例展示在图 1 。在这里，需要注意的是，相邻字母的联络性是发生在字母的核心部分的顶端（孟加拉文和阿拉伯文）或者字母的低端（英文）上。那么，输入单词的字迹轮廓是英文，那么它在使用分词模块之前首先转换成水平轴，而且该模块的输出需要再次转换以获取分词后的英语单词。

在该研究中，我们已经开发了一种新颖却很简单的分词模式，叫 DCE 分词算法，基于知名的离散曲线演化（DCE）策略的一种算法。DCE 算法被用于分解 2D 物体为有意义的视觉部分的研究。他后来被 Bai 应用于对由任何标准骨骼生成算法处理的图片的骨骼形状的剪枝和虚假部分的感知。DCE 的效果依靠于它可以辨认出保护物体的视觉形状的最好的轮廓点。传统上，DCE 被用来处理每个物体可以在有限的分解下被有限的多边形分清边界的图片。多边形的顶点来源于通过 DCE 对构成连续物体的基础的边界的采样，而且它提供了对给予的轮廓最能表现其形状的顶点的建立。由于一个线上手写样例相当于一个对应的厚物体的轮廓，我们可以使用这些由 dce 产生的轮廓上的有鉴别力的点去分割输入的手写单词成一些子部分。依靠于确定的相关参数值，DCE 也许会产生过度分割单词样例的现象，而且大多数导致的子部分可能没有可分辨形状。那么，一个聪明的对于鉴别点的建立的选择就变得很有必要，而且据观测发现那些区域内拥有最小值的鉴别点也是候选分割点。算法 1 用有条理的方式，多个步骤实现了提出的分词方法。

Algorithm 1 DCEseg Segmentation Algorithm

Input: $W = \{P_1, \ldots, P_n\}$ where $P_i = (x_i, y_i, \delta_i)$, $\delta_i \in$ {pen-down, pen-up}
Output: Segmented word sample W'
1: Copy W to W'
2: Obtain $C = \{Q_1, \ldots, Q_m\}$, the set of critical points of W produced by the DCE algorithm [13].
3: Let $U \subseteq C$ be the set of all $Q_k \in C$ such that if $Q_k = P_i$, then $y_i < y_{i-1}$ and $y_i < y_{i+1}$ and $y_i \leq 50$. Thus, U consists of the critical points which are also local minima lying in the upper half of the core region of W.
4: **if** $U \neq \phi$, i.e., U contains at least one point of W **then**
5: Reset δ_k = pen-up in W' corresponding to each $Q_k \in U$, (i.e., W is segmented at each point of U)
6: **end if**
7: **return** W'

除了提出的 DCE 分割算法，我们模拟了两个其他已经存在的分割策略，我们指认他们为 MINseg 和 ANGseg 算法。MINseg 算法是首先由 [4] 提出，而且该算法的提升版本后来被应用于 [1]。它发现一系列最小值点处被输入样例的核心部分的顶端的两边拓展的水平轴。这个水平轴选的很聪明。这个分词算法步骤见于算法 2 。

Algorithm 2 MINseg Segmentation Algorithm

Input: $W = \{P_1, \ldots, P_n\}$ where $P_i = (x_i, y_i, \delta_i)$, $\delta_i \in$ {pen-down, pen-up}
Output: Segmented word sample W'
1: Copy W to W'
2: Set $L_1 = \frac{100}{3}, L_2 = \frac{100}{2}, L_3 = \frac{100}{3}$
3: Obtain $U = \{Q_1, \ldots, Q_m | Q_k$ is a local minimum and $Q_k \in$ core region of W or above it$\}$.
4: **for** each point $Q_k \in U$ **do**
5: Compute d_{Q_k} = distance of Q_k from the top of core region of W.
6: Obtain $h_{Q_k,1}$ and $h_{Q_k,2}$ which are respectively the vertical distances between the top-most and the bottom-most points of the two parts of its stroke on both sides of Q_k.
7: **end for**
8: Let $V \subseteq U$ be the set of all $Q_k = (x_k, y_k, \delta_k) \in U$ such that $\min(h_{Q_k,1}, h_{Q_k,2}) \geq L_3$ and if Q_k lies above the core region, then the distance $d_{Q_k} < L_1$ or if Q_k lies inside the core region, then the distance $d_{Q_k} < L_2$.
9: **if** $V \neq \phi$, i.e., V contains at least one point of W **then**
10: Reset δ_k = pen-up in W' for each $Q_k \in V$, i.e., W is segmented at each point of V
11: **end if**
12: **return** W'

ANGseg 是第三种也是最后一种目前研究使用的的分词算法工具。他的第一个版本可见于［14］，后者修改见于［5］。这个算法跟从了每一个独立笔画的书写轨迹，如果一个点超出阈值，这个笔画就是分割处。实际上这种分词方法把复杂参数分割成大量拥有简单形状的子笔画。他的第一个版本（可见于［14］）生成了大量不同的字块，而后来的版本（见于［5］）产生了合适数量的字块。在现在的工具中，分割后单词最后由 20 个字块。该算法的步骤可见于算法 3.

Algorithm 3 ANGseg Segmentation Algorithm

Input: $W = \{P_1, \ldots, P_n\}$ where $P_i = (x_i, y_i, \delta_i)$, $\delta_i \in$ {pen-down, pen-up}
Output: Segmented word sample W'

Copy W to W'
Set $i = 1; j = 2; k = 0; K = 20$
for $t = 1, \ldots, n$ **do**
  Compute $\theta = \text{angle}(P_i, P_j)$
  while $\delta_j \neq$ pen-up and $t < n$ **do**
    Compute $\alpha_j = \text{angle}(P_i, P_j)$, $\beta_j = \text{angle}(P_{j-1}, P_j)$
    if $(\min\{|\theta - \alpha_j|, 360° - |\theta - \alpha_j|\} < 90°$ and $\min\{|\beta_{j-1} - \beta_j|, 360° - |\beta_{j-1} - \beta_j|\} < 90°)$ **then**
      $j = j + 1$ and $t = t + 1$
    else
      Set δ_{j-1} = pen-up in W'
      Set $i = j$; $j = j + 1$; $t = t + 1$
      break
    end if
  end while
  $k = k + 1$
end for
while $k > K$ **do**
  Compute curve length of each of k sub-strokes
  Identify a pair of consecutive sub-strokes and their segmentation point P_r for which the sum of lengths is maximum among $(k - 1)$ such sub-stroke pairs
  Set δ_r = pen-up
  $k = k - 1$
end while
return W'

IV. 特征与分类器

我们首先把子部分重采样成 25 个点，然后从中提取 26 的特征值。这些特征值包括（i）24 个从 Pt（xt，yt）到下一个点 Pt+1 移动的角产生于 x 轴正方向，（ii）子部分的长度 l=sum(lt),t=1~24，其中 lt 是 Pt 到 Pt+1 之间的欧拉距离，（iii）子块的中心部分的 y 坐标。这些特征起初有 Samanta（见［1］），前 24 个特征中的每个特征组成特征向量并且遵从 Von Mises 分布［15］，而且剩余两个成分遵从 Gauss 分布。子块的时间序列由分割单词样例获得，被对应输入样例而组成的特征向量代表。

在研究中，我们已经使用的[1]中的 HMM 框架作为分类器。与通常的实践对比，这个 HMM 是一个全联系非同质的模型，可容纳不同的奇怪的不同笔迹的自由手写数据。由于这种 HMM 的庞大的体系结构，他的训练需要大量的输入样例。这个问题可通过倒入大量由整条训练数组切割而成的子块数组解决。HMM 的状态可由加入的子块串获得。一个单词训练样例得到的串可贡献至少一个字块以作为对应 HMM 的状态。那么，对于不同的 HMM，状态的数量也会不同。这种方法的一个优点就是词汇表中一个新单词类的添加不需要新的子块串。

我们构建了两个 HMM，A1 和 A2，对于单词类 R 集合中的每个单词，A1 处理用自然循序处理观测序列，A2 用相反序列处理观测序列。这个类分辨了一个由未知测试样例 X 产生的观测序列 O（以自然顺序），Ō（以相反顺序）通过等式 1 获取。

$$c = \arg \max_{1 \leq r \leq R} \min\{Prob(O|\Lambda_1), Prob(\overline{O}|\Lambda_2)\}, \quad (1)$$

V. 结果与讨论

对应这三种分割方法的 HMM 分类器的识别结果通过三种不同的笔迹的数据集获取。这些数据集采集于（i）unipen-ICROW-2003 的 211 个英文单词的数据库，（ii）拥有 200 单词的应用在 ICAR2009 阿拉伯手写识别竞赛的 ADAB 数据库，（iii）用在［1］的拥有 100 个类的孟加拉线上识别数据库。这些数据集被以写者独立工作的方式区分为训练，验证和测试集。验证样例被用来决定一个最佳的加入子块串的数量。用于我们的模拟的训练数据和测试数据的量展示于表 1 。

TABLE I. STATISTICS OF THREE DATASETS

Bangla		English		Arabic	
Training	Test	Training	Test	Training	Test
No. of samples					
17624	8856	6130	1899	2679	1227
No. of substrokes by DCEseg					
242590	118914	76676	22977	38497	17737
No. of substrokes by MINseg					
154542	76197	54428	15656	28926	13081
No. of substrokes by ANGseg					
352480	177120	122600	37980	53580	24540

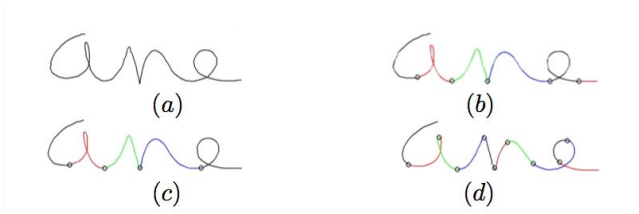

Fig. 2. (*a*) An original English sample and its segmentation points produced by (*b*) DCEseg, (*c*) MINseg and (*d*) ANGseg methods.

由不同的分词方法在不同的样例上作用产生的子块的数量可见于表 1 。这里需要注意 MINseg 算法生成了最小字块的数量，而

ANGseg 为每个样例提供了最大字块数量。我们用产生少量字块的 DCEseg 方法和 MINseg 方法观测和用 ANGseg 分词方法观测单词“ane”，从图二和图三中产生了相似的观测结果。

TABLE II. RECOGNITION ACCURACIES ON TEST SAMPLES OF DIFFERENT SCRIPTS PRODUCED BY THE HMM-BASED CLASSIFIER CORRESPONDING TO DIFFERENT SEGMENTATION APPROACHES

Segmentation Scheme	Recognition Accuracy (in %)		
	Bangla	English	Arabic
DCEseg	84.67	82.31	82.72
MINseg	85.69	83.31	84.35
ANGseg	84.85	81.67	82.31

表 2 展示了基于的 HMM 的分类器在对不同的笔迹测试时的识别精确度。我们观测到 MINseg 分词方法产生了最好的识别效果。在这三种分词方式中，对英语和阿拉伯语数据集处理精确度最低的是 ANGseg 分词方法，孟加拉语精确度最差的是 DCEseg 方法。然而，这些不同分词方法的精确度的差别并不是很大。那么，为了判断基于 HMM 的识别的不同效果依靠于不同的分词方式从统计学角度讲是非常有意义的，

TABLE III. TWO-TAILED z-TEST (SIGNIFICANCE LEVEL $\alpha = 0.05$) FOR THE EQUALITY OF RECOGNITION ACCURACIES CORRESPONDING TO DIFFERENT SEGMENTATION ALGORITHMS. NULL HYPOTHESIS H_0 IS REJECTED IF p-VALUE$(\alpha_0) < \alpha$.

Null hypothesis (H_0)	p-value (α_o)		
	Bangla	English	Arabic
$ACC_{MIN} = ACC_{DCE}$	0.0561	0.1836	0.2763
$ACC_{ANG} = ACC_{DCE}$	0.7389	0.4141	0.7892
$ACC_{MIN} = ACC_{ANG}$	0.2148	0.0608	0.1752

我们安排了两个部分组成的 Z 检验，在以＝0.05 的两个准确结果的质量无效假说。在表三中，我们列出了对应三对本研究中的字迹的 p 的值。在这里， ACCDCE，ACCMIN 和 ACCANG 表现了由基于 HMM 的对英语这三种分词方法各自的识别模式的准确度计量。表三中可以清楚地看到基于 HMM 的识别对应的三种分词算法没有显著的差别。那么，我们可以总结出来，当前的基于 HMM 手写识别模式在三中分词模式中没有太大差别。

VI. 总结

本研究的目标在于调查手写识别系统的发展是否可以通过跳过手写具体分词策略设计简化。那么，在本次实验中，我们考虑了对于草书手写单词的三种分词策略和训练与测试一个基于 HMM 的对应这三种字迹的分词策略的分类器模式。基于统计假说测试结果，我们已经获得了这样一个结论，即一个为单个字迹设计后的分词策略放入另一种字迹的语音识别模式中并不会显著影响这个识别效果。本研究的发现会引领更长远的研究无论分析识别框架的观测是否站得住脚。

特别感谢：这个研究工作部分受到了 Technology Development for Indian Languages (TDIL) Programme of the Dept. of Information Technology, Govt. of India 的支持。

Segmentation Scheme	Bangla word samples	English word samples	Arabic word samples
MINseg			
ANGseg			
DCEseg			

REFERENCES

[I] O. Samanta,U. Bhattacharya,and S. K Parui. Smoothing of HMM parameters for efficient recognition of online handwriting. Pattern Recognition, Vol. 47(11),pp. 3614-3629,Nov. 2014.

[2] X. Bai, L J. Latecki, and W Liu. Skeleton pruning by contour partitioning with discrete curve evolution. IEEE Trans. Pattern Anal. Mach. Intell., Vol. 29(3),pp. 449-462,2007.

[3] S. Madhvanath, V Govindaraju, The role of holistic paradigms in handwritten word recognition,IEEE Transactions on Pattern Analysis and Machine Intelligence,Vol. 23(2),pp. 149-164,2001.

[4] U. Bhattacharya,A. Nigam,Y. S. Rawat,and S. K. Parui. An analytic scheme for online handwritten bangla cursive word recognition. In Proc. of the Int. Con! on Frontiers in Handwriting Recognition, pp. 320-325,2008.

[5] S. Mohiuddin,U. Bhattacharya,and S. K Parui. Unconstrained bangla online handwriting recognition based on MLP and SVM. In Proc. of the loint Workshop on Multilingual OCR and Analytics for Noisy Unstructured Text Data, pp. 16:1-16:6. ACM,2011.

[6] R. Plamondon and S. N. Srihari. On-line and off-line handwriting recognition: A comprehensive survey. IEEE Trans. Pattern Anal. Mach. Intell., Vol. 22(1),pp. 63-84,2000.

[7] E. Kavallieratou, N. Fakotakis, G. Kokkinakis. An unconstrained handwriting recognition system, International lournal on Document Analysis and Recognition, Vol. 4(4),pp. 226-242,2002.

[8] G. Wilfong, F Sinden, L Ruedisueli. On-line recognition of handwritten symbols, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 18(9),pp.935-940,1996.

[9] M. Liwicki,H. Bunke. Combining diverse on-line and off-line systems for handwritten text line recognition,Pattern Recognition, Vol. 42(2), pp. 3254-3263,2009.

[10] A. Srimany,S. Dutta Chowdhury,U. Bhattacharya and S. K Parui. Holistic

Recognition of Online Handwritten Words Based on an Ensemble of SVM Classifiers, Proc. of the 11th IAPR Int. Workshop on Document Analysis Systems, pp. 86-90,2014.
[II] A. R. Ahmad, C. V iard-Gaudin, M. Khalid. Lexicon-Based Word Recognition Using Support Vector Machine and Hidden Markov Model, Proc. of the Int. Conf- Doc. Anal. and Recog., pp. 161-165,2009.
[12] G. Fink,S. Vajda,U. Bhattacharya,S. K. Parui and B. B. Chaudhuri. Online Bangia word recognition using sub-stroke level features and hidden Markov models, Proc. of 12th Int. Con! on Frontiers in Handwriting Recognition, pp. 393-398,2010.
[13] L 1. Latecki and R. Lakiimper. Convexity Rule for Shape Decomposition Based on Discrete Contour Evolution,Computer Vision and Image Understanding, Vol. 73(3),pp. 441-454,1999.
[14] T Mondal,U. Bhattacharya,S. K. Parui,K. Das and V Roy. Database generation and recognition of online handwritten Bangia characters, Proc. of Multilingual OCR, Article No. 9,ACM Int. Conf. Proc. series, Spain,2009.
[15] K. V Mardia,Statistics of Directional Data, Academic Press,New York, 1972.