معرفی روش جدیدی برای جداسازی حروف در متن چپی

بی‌توجهی به نوع قلم

رضا عزیزی و احسان الهی

یکش مهندسی برق، دانشگاه تربیت مدرس

(درباره مقاله: 1976/6/19 - دریافت نسخه نهایی: 1378/2/5)

چکیده - در این مقاله تایپ یک تحقیق انجام شده در زمینه جداسازی حروف چپی بدون توجه به نوع قلم ارائه می‌شود. مراحل مختلف گروه‌بندی شامل جداسازی خطوط متن، تعیین نوار زمینه و جداسازی حروف با استفاده از متن‌های پری‌نی آماده است. در مرحله تعیین نوار زمینه روش جدیدی ارائه شده که با کارایی خوبی را در عمل داشت. از اعمال قواعدی به صورت یک گرامر روش توسط پیشرفت می‌شوند و سپس توسط یک الگوریتم پس‌پردازش قطاع اولیه جداسازی تایید یا تصحیح می‌شوند.

الگوریتم بالا در مورد بیست نمای قلم متن اغلب و رنگ نمونه از قلم‌های قلمی که در تایپ آزمون‌های در مقطع دارای قدرت است که پسپردازش قطعه اولیه است. در این حرف از جمله که مجزا می‌شوند با استفاده از الگوریتم جدید 99 درصد از این حروف به درستی جداسازی می‌شوند.

A New Segmentation Technique for Omnifont Farsi Text

R. Azmi and E. Kabir
Department of Electrical Engineering, Tarbiat Modares University

ABSTRACT- In this paper a segmentation technique for omnifont Farsi text is presented. The upper contour of the word is traced and a set of proper rules is applied to find the presegmentation points.

A pre-processing step is introduced which adjusts the base line using chain code information. In a post-processing step, a set of heuristic rules is used to find segmentation points.

The Algorithm was tested on a data set of printed text with about 11,000 characters in 20 different fonts. The scanning resolution was 200 dpi. In this set, 71% of characters were connected, 98.5% of them being correctly segmented.

1- مقدمه

 بازشناسی الگوست که با پیشرفت تکنولوژی رایانه کاربرد

بازشناسی حروف با OCR

یکی از شاخصه‌های مورد توجه

استاد بزرگ در تکنولوژی رایانه کاربرد

- داشجوی دکتر

- استادیار

استقلال، سال 1، شماره 3، اسفند 1378
به‌شماری پیدا کرده است. درباره پژوهش‌های اتوماتیکی حروف‌لاین، چینی و ایننه تحقیقات زبانی صورت گرفته و سیستم‌های تجاری زبان‌آموزی نیاز ارائه شده‌اند [۱،۲]. برای پژوهش‌های صورت‌گرفته و ارائه نیز تحقیقات صورت گرفته است [۳،۴،۵،۶،۷،۸،۹،۱۰].

و جدول یکی چالش در نگارش فارسی مانند اتصال حروف به یکدیگر، تغییر شکل حروف با توجه به موقعیت آنها در کلمه، وجود حروف و علائم در بالا یا پایین حروف و همچنین به‌روش‌های آن‌ها ابعاد صورت‌های که به یکدیگر مستقیم و به‌روش‌های متفاوت نیستند، خلاصه شده است.

در بخش دوم، پژوهش‌های اولیه توضیح داده شده است. بخش سوم، روش تغییر نوار زمینه و ارائه‌ای مکانیکی بخش چهارم، الگوریتم‌های جداسازی را معرفی می‌کند. بخش پنجم، الگوریتم جداسازی در داده‌های درون‌کنشی بخش ششم، نحوه به‌کارگیری تناوب الگوریتم جداسازی را در مرحله پایانی الگوریتم‌های الگوریتم‌های اختصاصی دارد.

پژوهش‌های اولیه

این اندکی تصور می‌تواند و توسط یک پویشگر نوری با دقت و پر کرده باشد. اینکه به‌جای صورت یک پیوست مطلقی در راک ساخته و یک الگوریتم باید جویانه است که به‌جای الگوریتم محاسبه که برای مندر نظر حداکلی به‌جای باشد. اینکه پیوست صورت برداری باعث بهتر شدن روند انجام الگوریتم می‌شود ولی این مسئله به‌جای افزایش حجم اطلاعات و کم شدن سرعت پژوهش‌های است.
شکل ۱- نمودار کل سیستم

شکل ۲- جداسازی بخش‌هایی می‌تواند

۲-۱ جداسازی خطوط مرن و روی

در مورد مرن و روی یک فرصت را به دریافت یک خطوط با فاصله
نوشته می‌شود. اگر یک جداسازی خطوط از این فاصله بتوان
استفاده کنید، هر خطوط یک شمارش سیاه برای تصویر
محاسبه شده و تفاوتی که مقدار هیستوگرام در آن‌ها است
آسان‌تری کنید باعث نمایش حساسیت جداسازی خطوط در نظر
گرفته می‌شود.

با استفاده از این روش فقط و علائمی که بین آنها و خط اصلی
رده‌پذیری خالی وجود دارد به عنوان یک خط جدید شناسایی
می‌شود. این خط با در نظر گرفتن حد آن‌ها برای ارتفاع
هیستوگرام افزایش یک خطوط بر طرف می‌شود و بخش‌هایی که هیستوگرام
افقت آنها از حد آسان‌تری کنید باعث به مدتکنین خط مجاری خود
ملحق می‌شود. همچنین خطوط نیز باعث خطای جداسازی آنها
می‌شود. این حال با استفاده از الگوریتم رش ناحیه‌ای، عنصر
هم‌پوشانی به خطوط مربوط به خود اختصاص می‌یابد.

۲-۲ جداسازی بخش‌هایی می‌تواند

بعد از محاسبه هیستوگرام عمده‌ای بخش‌هایی از کلمات خط
ورودی یک هیستوگرام پیوسته با عنوان بخش‌هایی می‌تواند
در نظر گرفته می‌شود. (شکل ۲.)

هر جداسازی از یک با یک نعل برای تشکیل شده است که
کلمات خطوط از روی بررسی جداسازی می‌تواند با مولفه‌ها برای تفکیک آنها استفاده
می‌شود. چون در این تحقیق از مسجد پیرو می‌باشد، به منظور

شکل ۳- نحوه محاسبه قطعات عمومی متصل برای تعیین پهنای قلم

۲-۲ جداسازی پهنای قلم

برای محاسبه پهنای قلم، به خط از مرن در جهت عمودی
جاری شده، و قطعات متصل از عناصر سیاه برای انتخاب شرده
می‌شود و اندام‌هایی که بیشترین قرارگیری را داشته باشد به عنوان
پهنای قلم در نظر گرفته می‌شود. شکل (۲)، برای اینکه دقت بیشتری
داشته باشیم می‌توانیم این جراثیب را در جهت افقی نشان دهیم
اما اگر آنها نتایج حساسیت از چندین دیده، با بارب‌این به
منظور افزایش سرعت از این صورت نظر می‌شود.
3- تعیین نواز زمینه

تعیین دقیق نواز زمینه نقش مهمی در انجام صحیح مرحله جداسازی دارد. به ویژه برای فلزهایی با پهنای کم خطا در حد یک عصر تصویر نیز می‌تواند باعث مشکلاتی در مرحله جداسازی شود.

برای تعیین مقدار اولیه نواز زمینه آن را به صورت زیر تعیین می‌کنیم:

نواز زمینه نواری افقی است با پهنای قلم که پیشترین تعداد عناصر سیاه در تصویر یک خط از میان را در خود داشته باشد، شکل (4).

در صورتی که احتمال چگالی در میان ورودی وجود دارد یا متن چایی در عاری کردن تابی تیست، بهتر است خروجی را به
بخش‌های کچکتری تقسیم کرده و نواز زمینه از برای هر بخش

بستگی محاسبه کنید تا خاطر حاصل از چگالی کاهش یابد. بعد

از محاسبه مقدار اولیه نواز زمینه روشن ادعا زیر برای تعیین مقدار

نهاپی آن به کار برده شده است.
شکل 7- کانال پرتویه و بالایی
الف - کانال پرتویه ب - کانال بالایی

شکل 6- اثر اصلاح نوار زمینه در تعیین نقطه جداسازی
الف - استفاده از نوار زمینه اولیه ب- استفاده از نوار زمینه اصلاح شده

مدیر مدرسه مهدیه
(پ.)

نمونه‌ای از اثر اصلاح‌های این روش را نشان می‌دهد.

۴- الگوریتم جداسازی

قوانین جداسازی بر اساس کانال بالایی و کانال پایین کلیمات پایه‌گذاری
شده است. نمونه‌ای از این کانال‌ها در شکل (۷) نشان داده شده است.
بعد از محاسبه کانالات از نوع‌های فرآیند برای حذف نویز استفاده شده است. در زیر به شرح مختصر مراحل مختلف الگوریتم جداسازی
صریحاً پراکنده گردید.

۱-۶- مرحله اول: تعیین کانال بالایی

سمت راست ترین نقطه کانال پیرامونی را به عنوان نقطه شروع در نظر گرفته‌ایم (نقطه sp در شکل ۹). ان گاه منحنی پیرامونی را در
جهت خلاف وقوع‌های ساعت دنیال می‌کنیم تا به سمت چپ‌ترین
نقطه کانال پیرامونی برسیم (نقطه sp در شکل ۹).


tbl=n1 | h1(n1) = maxn ( h1(n))
(1)

ubl=n1 | h1(n1) = maxn ( h1(n))
(2)

بعد از محاسبه دقیق نوار زمینه، اگر پیمانی این نوار بیش از ۷۵ درصد با پیمان فلز اختلاف داشته باشد، یکی از مراحل بالایی یا
پایینی که مقدار منفی‌تر از یک هیستوگرام مربوط به این است تبیین
می‌شود و مراز دیگر با توجه به پیمان قلم تعیین می‌شود.

به کانال جنوبی، روش بالا به معنی محاسبه نوار زمینه، حفره با
تعیین نوار زمینه به صورت محلی بسیاری از خطاهای و مشکلاتی را
که می‌پیشده با پیش‌داده‌های جریبی حذف کرد.
شکل 8- نمودار حالت مرتبه بی‌درجه کاتانور بالایی زیر کلمه
الف- تعریف پارامترهای pt, ubl, y(p), x(p), p و ubl, y(p) ≤ ubl, y(p) > ubl
طول نقطه عرض نقطه عرض بالایی، لبه بالایی نوار زمینه لبه بالایی نوار زمینه و پنجم قائم
ب- نمایی نمودار حالت

4- مرحله دوم: تعیین برحسب نقاط کاتانور بالایی
با استفاده از قواعد زیر برحسب کلیه نقاط کاتانور بالایی تعیین می‌شود، شکل (8):

1- اولین نقطه کاتانور برحسب بالا (u) می‌گردد.
2- اگر نقاط قبیل برحسب بالا داشته باشند و عرض نقطه فعلی بیشتر از ubl باشد، نقطه فعلی برحسب وسط می‌گردد. در غیر این صورت همان برحسب نقطه قبلی تعیین برحسب بالا را خواهد گرفت.
3- اگر نقاط قبیل برحسب وسط داشته باشند، در صورتی که عرض نقطه فعلی برحسب بالا (u) ubl > pt/2 باشد، برحسب بالا را می‌گیرد.
4- اگر حدف یک پاره مسیر با استفاده از قاعدة بالا باعث شود که منجر به اجرای یک پاره مسیر توسط کاتانور شود.
جدول 1 - خلاصه تایپی از مجموعه تایپی و حروف جداسازی (قبل از مرحله تایپ و تصویر)

<table>
<thead>
<tr>
<th>تعداد حروف متصول</th>
<th>تعداد حروف حرف متصل</th>
<th>تعداد کل جداسازی</th>
</tr>
</thead>
<tbody>
<tr>
<td>91</td>
<td>8056</td>
<td>11327</td>
</tr>
</tbody>
</table>

معلوم در دقتهای پایین تصویر برداری یا پایین کم قلم مسئولیت داشته باشد. شکل (9) مراحل اجرای الگوریتم جداسازی و نشان می‌دهد.

الگوریتم جداسازی بالای روند مختلف که با 20 قلم می‌عمل و چند نمونه قلمه‌های مشترک چاب‌شده آزموده شده است. خلاصه نتایج در جدول (1) آمده است. آزمون بالا کاربردی مناسب الگوریتم جداسازی را نشان می‌دهد. حروف الگوریتم نقطه اولیه جداسازی‌آند در مرحله بعدی الگوریتم تایپ و تصویر می‌شوند.

شکل 9 - پایه مسیرها برای خروج و نقطه جداسازی بالا (u) و پایین (d)

دوام مسیر همیشه در کنار هم یک دیگر قرار گرفته، خنثی نیز به هر دوی می‌رساند و یک دیگر شکل می‌دهد.

- اگر یک پایه مسیر پایین با اندازه غیر به جای جنگ ونتون مناسب کتابی‌ای که بیشتر باشد، به معنی تغییر در کل مسیر است که پایین رفع آن، پایه مسیر پایین را به عنوان نقطه انتهایی کل مسیر اول در نظر گرفته و ادامه کتاب‌یکه به عنوان نقطه جدید در نظر گرفته و دوام مسیر باعث برخی تغییرات بین کلمات می‌شود.

- مرحله چهارم: تغییر نقطه جداسازی نقطه جداسازی آخرین نقطه از پایه مسیر با برچسب وسط است که در شرایط دو سطحی می‌کنند (pt پایه ایم است):

الف - طول این پاره مسیر از مقدار 1pt پاییز باشد.
ب - پاره مسیر قبلی برچسب بالا داشته و طول آن از مقدار pt کمتر باشد.
ج - طول پاره مسیر بعدی اگر پایین باشد از 1/3در میانگین یک پاره مسیر جداسازی بالا انجام می‌گیرد و عنوان یک حرف مجزا جداسازی می‌شود. این پاره مسیر انتهایی با پایین کم آیکت از پایه قلم شناسایی شده.
جدول ۲- نمونه‌هایی از اثر الگوریتم تایید و تصحیح نهایی روی نقاط جداسازی

<table>
<thead>
<tr>
<th>شماره</th>
<th>مقدار</th>
<th>تکنیک</th>
<th>مقاومت</th>
</tr>
</thead>
<tbody>
<tr>
<td>۱</td>
<td>کندن</td>
<td>۱۵۰</td>
<td>۱۵۰</td>
</tr>
<tr>
<td>۲</td>
<td>کم</td>
<td>۱۵۰</td>
<td>۱۵۰</td>
</tr>
<tr>
<td>۳</td>
<td>کمصور</td>
<td>۹۰</td>
<td>۹۰</td>
</tr>
<tr>
<td>۴</td>
<td>فصل نص</td>
<td>۲۰۰</td>
<td>۲۰۰</td>
</tr>
<tr>
<td>۵</td>
<td>پیتشر</td>
<td>۱۵۰</td>
<td>۱۵۰</td>
</tr>
<tr>
<td>۶</td>
<td>جلد</td>
<td>۲۰۰</td>
<td>۲۰۰</td>
</tr>
</tbody>
</table>

و به‌یژه، حرف قبلی متصل می‌شود، جدول (۲) نمونه (۲) 

حرف "م" در بعضاً از رسوم الخط‌ها در انتهای کلمه به دو زیر حرف بشکسته می‌شود که حرف انگیزه باید با یک حرف افزایش یابد، انتهای شناسایی شده و به حرف اصلی متصل می‌شود، جدول (۲) نمونه (۲) 

دندانهای "س" و "ش" نیز در بسیاری از موارد باعث شکسته می‌شوند که حروف به هم منفی باشند. برای رفع این نوع خطای الگوریتمی نوشته شده که به این نوام محل و جایگذاری می‌شود، جدول (۲) نمونه (۲) 

اگر یک نقاط ماهیت داشته باشد که روی آن چه قطعاتی نیست، آن را به حرف قبلی خود متصل می‌کنیم (حرف "ص" یا "ض")
جدول 3- خلاصه نتایج نهایی الگوریتم جداسازی

<table>
<thead>
<tr>
<th>درصد کل جداسازی</th>
<th>درصد جداسازی حروف مفصل</th>
<th>تعداد حروف مفصل</th>
<th>تعداد حروف</th>
</tr>
</thead>
<tbody>
<tr>
<td>98/9</td>
<td>98/5</td>
<td>8056</td>
<td>11347</td>
</tr>
</tbody>
</table>

(الف) ب- الگوریتم‌های جداسازی برای مسئله‌های مختلف
(ب) اثر کشی‌گذاری حروف
(الف) شکل 10- نموداری از اثر کشی‌گذاری در تغییر نقاط جداسازی توسعه الگوریتم‌های مختلف

به کارگری الگوریتم میرود تا ایجاد ارتباطی، خطاهای اولیه الگوریتم را تا حد زیادی کاهش می‌دهد. جدول 3 (خلاصه نتایج نهایی الگوریتم) را نشان می‌دهد.

5-1 اثر کشی‌گذاری حروف

کجگین حروف از نوع اتالیک یا اترایک در پیش آمده الگوریتم‌های جداسازی می‌تواند ایجاد اشکال کند. این مثال در الگوریتم‌های که براساس هسته‌گرام عمودی یا پروفیل دید از بالای زیر کلمات عمل می‌کند، کجگین حروف می‌تواند باعث همبستگی بین حروف جمع شده یا حروف جمع شده یا حروف جمع شده در ترتیب نقاط جداسازی آنها را پنهان کند. ولی الگوریتم مطرح شده در این مقاله موجب می‌شود کنترل بالایی برای حروف جمع شده ایجاد شود.

6- به کارگری نتایج الگوریتم جداسازی در مرحله

پژوهشی

الگوریتم مطرح شده در این مقاله با الگوریتم C روی یک رایانه

پردازنده یا یک رایانه

687.5 مگاهرتز با پایه سازی دو زبان است. سرعت متغیر جداسازی در حد 3743 کاراکتر در ثانیه است.
بانا جداسازی کرد. و آنها را برای مرحله باندیژاسی آماده می‌کند [۱۰]. استفاده از الگوریتم تایید و تصویب مطرح شده در بخش قابلی به کارگیری اطلاعات حاصل از نوع و محتوای نقاطه، تعداد حفره‌ها، یکتا و ارتفاع پارامترهای ایجاد شده، پریخ از حروف زیر کلمه‌ها نیز شناختی می‌کند. این حروف و نحوه شناسایی این حروف در ازای هر می‌شود.

7- نتیجه‌گیری
هدف از طراحی الگوریتم مشخص کردن نقاط جداسازی است. حروف جدا شده در مرحله بعد به الگوریتم باندیژاسی آماده شده و در آنجا شناسایی نیازمندی می‌شوند. بررسی خطاها باید بینانی که است که از دو کیفتی پایین تصویربرداری می‌شود. بدین ترتیب در انجام پرای برخی از پلمب و کج شدنی باشد و حتما در زمان تصویربرداری است. الگوریتم علاوه بر جداسازی حروف با بیان برخی از حروف را نیز باندیژاسی می‌کند.

میزبان مشکلات را که در جداسازی ابتدایی مشخص نمی‌کند. همچنین حفظ خصوصیات و نقاط و اطلاعات در مرحله جداسازی مشخص داده می‌شود. الگوریتم ارائه شده در این مقاله اطلاعات را به شکل مناسب برای محاسبه دقیق تازه‌گیری انجام جداسازی با حفظ کم و باندیژاسی مستقیم پریخ از حروف بدون نیاز به مرحله باندیژاسی به کار گرفته است. هر این که مشابه عملی یا پردازش دید از بالایی کلمات استفاده می‌کند این مزایا را ندارند.


مراجع

حرف در رسم‌الخط‌های مختلف، "مجموعه مقالات سومین کنفرانس الکترونیک، ص ۹۳-۱۰۹، دانشگاه شیراز، مهر ۱۳۷۷.
7. هیلی، ج. و تیمرساردی، ب. "باندیژاسی حروف کلمات تایید فارسی با استفاده از روش مرحله‌ای." مجموعه مقالات اولین کنفرانس سومین کریستال، جلد ۲، ص ۲۷۸-۲۷۹، دانشگاه صنعتی امیرکبیر، تهران ۱۳۷۵.
8. میرزاپور، ح. و فاطمی، ک. "روش نوین در شناسایی متون فارسی به کمک شبکه‌های عصبی," مجموعه مقالات سومین کنفرانس الکترونیک، ص ۳۹-۴۹، دانشگاه شیراز، مهر ۱۳۷۵.
10. عزیزی، ر. و یوسفزاده، آ. "آزمایش و بررسی الگوریتم باندیژاسی حروف فارسی." مجموعه مقالات سومین کنفرانس بین‌المللی انجمن کامپیوتر ایران، ص ۱۹۱-۱۹۷، دانشگاه علم و صنعت ایران، تهران ۱۳۷۶.