معرفی روش جدیدی برای جداسازی حروف در متون چاپی بدون توجه به نوع قلم

رضا عزمی* و احسان اله کبیر**

بخش مهندسی برق، دانشگاه تربیت مدرس

(دریافت مقاله: ۱۳۷۶/۶/۱۹ - دریافت نسخهٔنهایی: ۱۳۷۸/۴/۱۵)

چکیده - در این مقاله نتایج یک تحقیق انجام شده در زمینهٔ جداسازی حروف چاپی بدون توجه به نوع قلم ارائه می‌شود. مراحل مختلف الگوریتم شامل جداسازی خطوط متن، تعیین نوار زمینه و جداسازی حروف با استفاده از منحنی بیرونی آنهاست. در مرحلهٔ تعیین نوار زمینه روش جدیدی ارائه شده که کارایی خود را در عمل نشان داده است. نقاط جداسازی اولیه با اعمال قواعدی به صورت یک گرامر روی منحنی پیرامونی تعیین می‌شوند و سپس توسط یک الگوریتم پس پردازش نقاط اولیه جداسازی تأیید یا تصحیح می‌شوند.

الگوریتم بالا در مورد بیست نوع قلم متداول و چندین نمونه از قلمهای قدیمیتر که در چاپ و نشر کتب و مجلات به کار رفته‌اند آزموده شده است. از چندین صفحهٔ متن چاپی شامل حدود ۱۱ هزار حرف با دقت ۲۰۰ نقطه در اینچ تصویر برداری شده است. ۷۱٪ از حروف این مجموعه متصل بوده‌اند. با استفاده از این الگوریتم حدود ۹۹ 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 Modarres 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 infomation. 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.

۱- مقدمه

بازشناسی حروف یا OCR یکی از شاخه‌های مورد توجه بازشناسی الگوست که با پیشرفت تکنولوژی رایانه کاربرد

---

* - دانشجوی دکترا ** - استادیار

فهرست علائم

d	برچسب پایین		کانتور بیرونی با کد۴	ubl	لبه بالایی نوار زمینه
ep	نقطهٔ انتهایی کانتور بالایی	lbl	لبهٔ پایینی نوار زمینه	w	پهنای تصویر
h	ارتفاع تصویر	m	برچسب وسط	x(p)	طول نقطهٔ p
h۰(n)	هیستوگرام ارتفاع نقاط روی کانتور بیرونی با کد ۰	p	نقطه روی کانتور	y(p)	عرض نقطهٔ p
		pt	پهنای قلم		
h(n)	هیستوگرام ارتفاع نقاط روی	sp	نقطهٔ شروع کانتور بالایی		

بیشتری پیدا کرده است. دربارهٔ بازشناسی اتوماتیک حروف لاتین، چینی و ژاپنی تحقیقات زیادی صورت گرفته و سیستمهای تجاری زیادی نیز ارائه شده‌اند [۱و۲]. برای بازشناسی حروف فارسی و عربی نیز تحقیقاتی صورت گرفته است، [۳، ۴، ۵، ۷، ۸ و ۱۰].

وجود ویژگیهای خاص در نگارش فارسی مانند اتصال حروف به یکدیگر، تغییر شکل حروف با توجه به موقعیت آنها در کلمه، وجود نقاط و علائم در بالا یا پایین حروف و همچنین همپوشانی آنها باعث می‌شود که به کارگیری مستقیم روشهای بازشناسی متون لاتین برای خط فارسی ممکن نباشد.

دو رویکرد کلی برای بازشناسی متون چاپی وجود دارد. در رویکرد اول فرض می‌شود که جداسازی و بازشناسی دو مرحله مستقل از هم هستند [۳ و۶]. در رویکرد دوم شناسایی در یک مرحله صورت گرفته و از مرحله جداسازی صرف نظر می‌شود [۷و۸]. هر کدام از این دیدگاهها مزایا و معایب مربوط به خود را دارند. دیدگاه دیگری را نیز می‌توان به این صورت در نظر گرفت که مرحلهٔ جداسازی و شناسایی دو مرحلهٔ کاملاً مستقل نبوده و می‌توانند به منظور کاهش خطای یکدیگر استفاده شوند. آنچه در این مورد می‌توان گفت این است که با توجه به اینکه کلمات چاپی از اتصال حروف به یکدیگر ایجاد شده‌اند، وجود مرحلهٔ جداسازی نه تنها زاید نیست بلکه توجه به این خاصیت ساختاری می‌تواند در افزایش سرعت و دقت بازشناسی کاملاً مؤثر باشد.

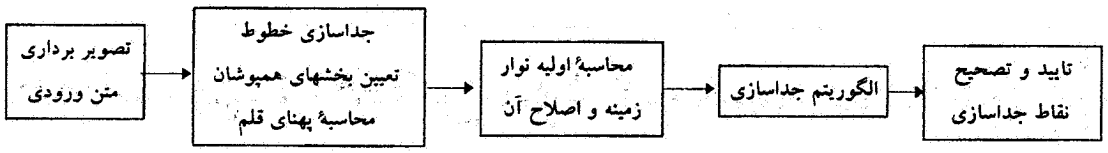
هدف از این مقاله ارائهٔ روشی برای جداسازی متون چاپی فارسی بدون وابستگی به نوع و اندازهٔ قلم به کار برده شده است. متن مورد نظر می‌تواند شامل اکثر فونتهای معمول نگارش فارسی باشد. الگوریتم مورد نظر با پردازش متن ورودی نقاط جداسازی را معرفی می‌کند.

شکل (۱) نمودار کلی سیستم را نشان می‌دهد. همان طور که در نمودار ملاحظه می‌شود ابتدا بعد از انجام پردازشهای اولیه روی تصویر متن ورودی، خطوط متن و بخشهای همپوشان جدا شده و پهنای قلم تعیین می‌شود. در مرحلهٔ بعد مقدار اولیه نوار زمینه تعیین شده و با محاسبهٔ منحنی پیرامونی و با یک روش پیشنهادی نوار زمینه اصلاح می‌شود. مرحلهٔ بعدی الگوریتم مربوط به تعیین اولیه نقاط جداسازی بوده و در مرحلهٔ نهایی با اعمال قواعدی نقاط جداسازی نهایی تعیین می‌شوند.

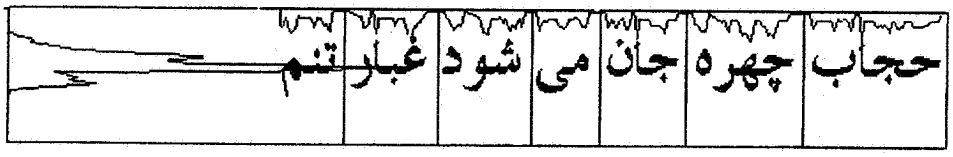
در بخش دو، پردازشهای اولیه توضیح داده شده است. بخش سه، روش تعیین نوار زمینه را ارائه می‌کند. بخش چهار، الگوریتم اصلی جداسازی را معرفی می‌کند. در بخش پنج، الگوریتم مرحله اصلاح و تأیید نقاط جداسازی توضیح داده شده است. بخش شش، نحوهٔ به کارگیری نتایج الگوریتم جداسازی را در مرحلهٔ بازشناسی بیان می‌کند. بخش هفت، به بررسی و ارائهٔ نتایج الگوریتم اختصاص دارد.

۲- پردازشهای اولیه

ابتدا تصویر متن ورودی توسط یک پویشگر نوری با دقت ۲۰۰ نقطه در اینچ به صورت یک تصویر دو سطحی در رایانه ذخیره می‌شود. تعیین دقت تصویر برداری به گونه‌ای است که پهنای قلم محاسبه شده برای متن مورد نظر حداقل سه نقطه باشد. افزایش دقت تصویر برداری باعث بهتر شدن نتیجهٔ الگوریتم می‌شود ولی این مسئله به بهای افزایش حجم اطلاعات و کم شدن سرعت بازشناسی است.



شکل ۱- نمودار کلی سیستم



شکل ۲- جداسازی بخشهای همپوشان

۲-۱- جداسازی خطوط متن ورودی

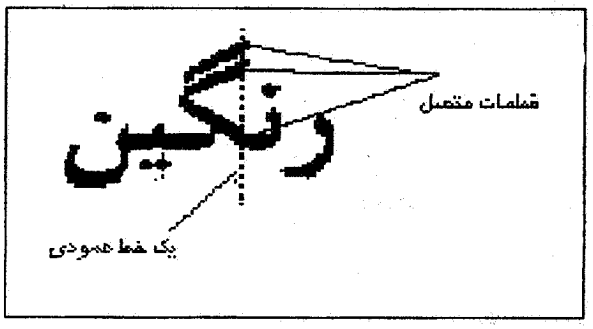
در مورد متن ورودی این فرض را می‌پذیریم که خطوط با فاصله نوشته می‌شوند. الگوریتم جداسازی خطوط از این فاصلهٔ طبیعی استفاده می‌کند. هیستوگرام افقی با شمارش عناصر سیاه تصویر محاسبه شده و نقاطی که مقدار هیستوگرام در آنها صفر یا از حد آستانه‌ای کمتر باشد به عنوان نقطهٔ جداسازی خطوط در نظر گرفته می‌شوند.

با استفاده از این روش نقاط و علائمی که بین آنها و خط اصلی ردیفهای خالی وجود دارد به عنوان یک خط جدید شناسایی می‌شوند. این خطا با در نظر گرفتن حد آستانه‌ای برای ارتفاع هیستوگرام افقی خطوط برطرف می‌شود و بخشهایی که هیستوگرام افقی آنها از حد آستانه‌ای کمتر باشد به نزدیکترین خط مجاور خود ملحق می‌شوند. همپوشانی خطوط نیز باعث خطا در جداسازی آنها می‌شود. در این حالت با استفاده از الگوریتم رشد ناحیه‌ای، عناصر همپوشان به خطوط مربوط به خود اختصاص می‌یابند.

۲-۲- جداسازی بخشهای همپوشان

بعد از محاسبهٔ هیستوگرام عمودی بخشهایی از کلمات خط ورودی که دارای هیستوگرام پیوسته‌اند به عنوان بخشهای همپوشان در نظر گرفته می‌شوند، (شکل ۲).

هر بخش همپوشان از یک یا چند زیرکلمه تشکیل شده است که معمولاً از روش برچسب زدن به مؤلفه‌ها برای تفکیک آنها استفاده می‌شود. چون در این تحقیق از منحنی پیرامونی به منظور

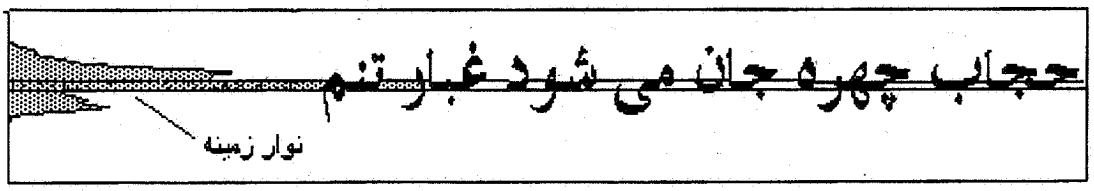


شکل ۳- نحوهٔ محاسبهٔ قطعات عمودی متصل برای تعیین پهنای قلم

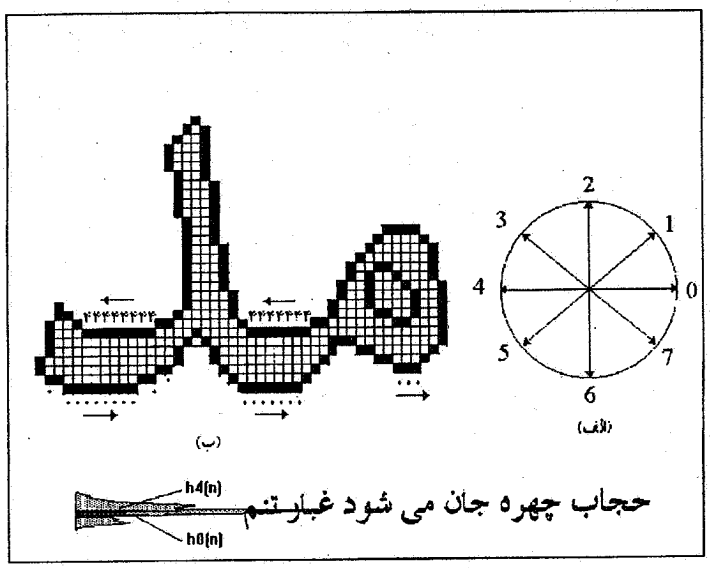
جداسازی حروف استفاده شده، محاسبه این منحنی خود به خود باعث تفکیک بخشهای مورد نظر می‌شود و بنابر این نیاز به پردازشهای اضافه و به کارگیری روش برچسب زدن به مؤلفه‌ها نیست.

۲-۳- محاسبهٔ پهنای قلم

برای محاسبهٔ پهنای قلم، هر خط از متن در جهت عمودی جاروب شده و قطعات متصل از عناصر سیاه برحسب اندازه شمرده می‌شوند و اندازه‌ای که بیشترین فراوانی را داشته باشد به عنوان پهنای قلم در نظر گرفته می‌شود، شکل (۳). برای اینکه دقت بیشتری داشته باشیم می‌توانیم این جاروب را در جهت افقی نیز انجام دهیم اما در اکثر اوقات نتیجهٔ حاصل تغییر چندانی نمی‌کند، بنابر این به منظور افزایش سرعت از آن صرف نظر می‌شود.



شکل ۴- تعیین نوار زمینهٔ اولیه



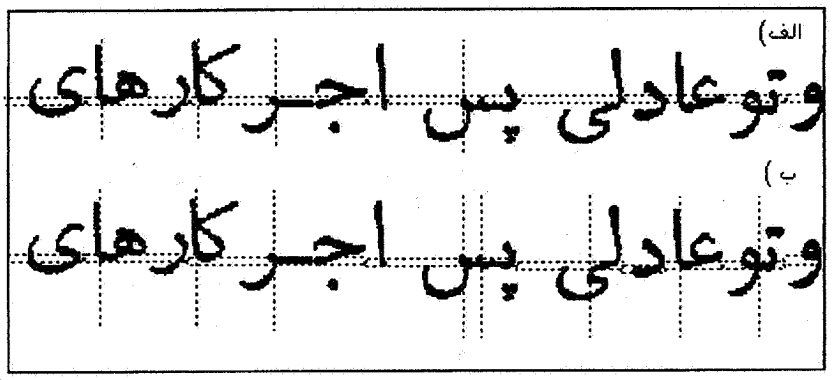
شکل ۵- روش تعیین نوار زمینهٔ نهایی
الف- کدهای فریمن ب- نقاط انتخاب شدهٔ کانتور
ج- نمونه‌ای از تعیین نوار زمینهٔ نهایی با استفاده از هیستوگرامهای $h_0(n)$ و $h_4(n)$

۳- تعیین نوار زمینه

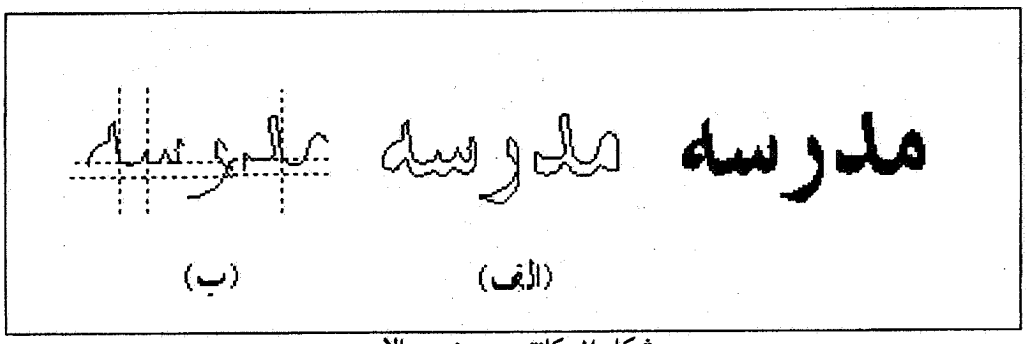
تعیین دقیق نوار زمینه نقش مهمی در انجام صحیح مرحله جداسازی دارد. به ویژه برای قلمهایی با پهنای کم، خطا در حد یک عنصر تصویر نیز می‌تواند باعث مشکلاتی در مرحلهٔ جداسازی شود. برای تعیین مقدار اولیه نوار زمینه آن را به صورت زیر تعریف می‌کنیم: «نوار زمینه نواری افقی است با پهنای قلم که بیشترین تعداد عناصر سیاه در تصویر یک خط از متن را در خود داشته باشد، شکل(۴).

در صورتی که احتمال کج شدگی در متن ورودی وجود دارد یا متن چاپی دارای کرسی ثابتی نیست، بهتر است خط ورودی را به بخشهای کوچکتری تقسیم کرده و نوار زمینه را برای هر بخش مستقلاً محاسبه کنیم تا خطای حاصل از کج شدگی کاهش یابد. بعد از محاسبهٔ مقدار اولیه نوار زمینه روش ابداعی زیر برای تعیین مقدار نهایی آن به کار برده شده است.

ابتدا کانتور بیرونی کلمات با پیمایش در خلاف جهت عقربه‌های ساعت استخراج شده و با استفاده از کدهای فریمن با معیار همسایگی ۸ نشان داده می‌شود، شکل (۵ - الف). برای محاسبهٔ لبهٔ بالایی نوار زمینه (ubl) تنها نقاطی از کانتور بیرونی کلمات را در نظر می‌گیریم که کد فریمن مربوط به آنها ۴ بوده و در محدودهٔ مشخصی از لبهٔ بالایی نوار زمینهٔ اولیه قرار داشته باشند، شکل (۵ - ب). در مرحلهٔ بعدی هیستوگرام ارتفاع این نقاط را محاسبه کرده و آن را $h_4(n)$ می‌نامیم. ردیفی از تصویر که متناظر با مقدار بیشینهٔ این هیستوگرام است به عنوان مرز بالایی نوار زمینه (ubl) انتخاب می‌شود، معادلهٔ (۱). به طور مشابه، برای محاسبهٔ لبه پایینی نوار زمینه (lbl) تنها نقاطی از کانتور را در نظر می‌گیریم که کد فریمن آنها صفر بوده و در محدودهٔ مشخصی از لبه پایینی نوار زمینه اولیه قرار داشته باشند، شکل (۵- ب). آن گاه هیستوگرام ارتفاع این



شکل ۶- اثر اصلاح نوار زمینه در تعیین نقاط جداسازی
الف - استفاده از نوار زمینهٔ اولیه ب - استفاده از نوار زمینهٔ اصلاح شده



شکل ۷- کانتور بیرونی و بالایی
الف - کانتور بیرونی ب - کانتور بالایی

نقاط را محاسبه کرده و آن را با $h_0(n)$ نشان می‌دهیم. ردیفی از تصویر که متناظر با مقدار بیشینهٔ این هیستوگرام‌است به عنوان مرز پایینی نوار زمینه (ubl) انتخاب می‌شود، معادلهٔ (۲). شکل (۵-ج) نمونه‌ای از محاسبهٔ نوار زمینه را بر اساس روش بالا نشان می‌دهد.

$$ubl=n_1 | h_4(n_1) = \max_n \{h_4(n)\} \quad (1)$$

$$lbl=n_2 | h_0(n_2) = \max_n \{h_0(n)\} \quad (2)$$

بعد از محاسبهٔ دقیق نوار زمینه، اگر پهنای این نوار بیش از ۲۵ درصد با پهنای قلم اختلاف داشته باشد، یکی از مرزهای بالایی یا پایینی که مقدار متناظر آن در هیستوگرام مربوطه بیشتر است تثبیت می‌شود و مرز دیگر با توجه به پهنای قلم تعیین می‌شود.

به کارگیری روش بالا به منظور محاسبهٔ نوار زمینه، همراه با تعیین نوار زمینه به صورت محلی بسیاری از خطاها و مشکلاتی را که می‌بایست با پس پردازشهای تجربی حذف کرد بر طرف می‌کند.

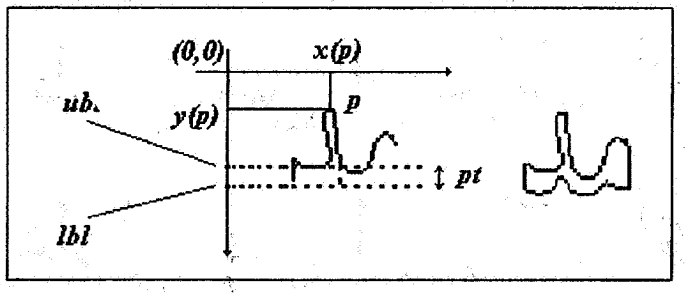
شکل (۶) نمونه‌ای از اثر اصلاحی این روش را نشان می‌دهد.

۴- الگوریتم جداسازی

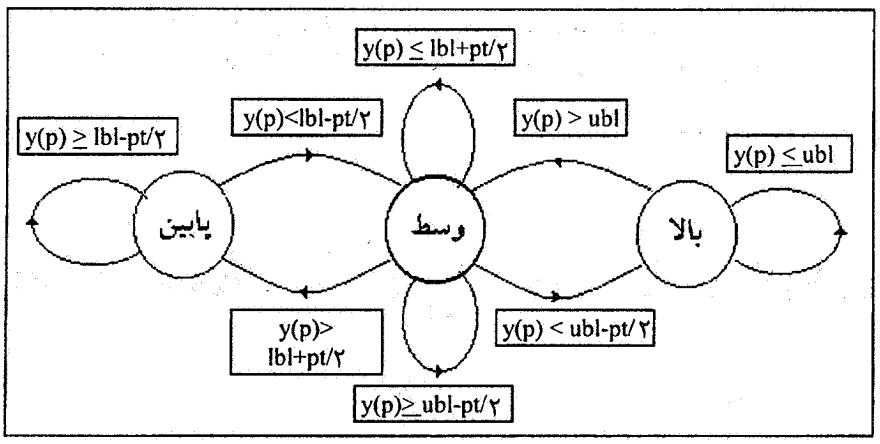
قوانین جداسازی بر اساس کانتور بالایی زیر کلمات پایه گذاری شده است. نمونه‌ای از این کانتور در شکل (۷) نشان داده شده است. بعد از محاسبهٔ کانتور از قواعد فو[۹] برای حذف نویز استفاده شده است. در زیر به شرح مختصر مراحل مختلف الگوریتم جداسازی می‌پردازیم.

۴-۱- مرحله اول: تعیین کانتور بالایی

سمت راست‌ترین نقطهٔ کانتور پیرامونی را به عنوان نقطهٔ شروع در نظر می‌گیریم (نقطهٔ sp در شکل ۹). آن گاه منحنی بیرونی را در جهت خلاف عقربه‌های ساعت دنبال می‌کنیم تا به سمت چپ‌ترین نقطهٔ کانتور بیرونی برسیم (نقطهٔ ep در شکل ۹).



(الف)



(ب)

شکل ۸- نمودار حالت مربوط به مرحلهٔ تعیین برچسبهای کانتور بالایی زیر کلمه

الف- تعریف پارامترهای p، x(p)، y(p)، ubl، lbl و pt که به ترتیب نشان دهندهٔ یک نقطه روی کانتور، طول نقطه، عرض نقطه، لبهٔ بالایی نوار زمینه، لبهٔ پایین نوار زمینه و پهنای قلم‌اند.

ب- نمایش نمودار حالت

۴-۲- مرحلهٔ دوم: تعیین برچسب نقاط کانتور بالایی

با استفاده از قواعد زیر برچسب کلیه نقاط کانتور بالایی تعیین می‌شود، شکل (۸):

۱- اولین نقطهٔ کانتور برچسب بالا (u) می‌گیرد.

۲- اگر نقطهٔ قبلی برچسب بالا داشته باشد و عرض نقطهٔ فعلی بیشتر از ubl باشد، نقطهٔ فعلی برچسب وسط می‌گیرد. در غیر این صورت همان برچسب نقطهٔ قبلی یعنی برچسب بالا را خواهد گرفت.

۳- اگر نقطهٔ قبلی برچسب وسط داشته باشد، در صورتی که عرض نقطهٔ فعلی کمتر از ubl-pt/۲ باشد این نقطه برچسب بالا (u) می‌گیرد. در صورتی که عرض آن بیشتر از lbl+pt/۲ باشد، برچسب پایین خواهد گرفت و در غیر این دو صورت، همان برچسب نقطهٔ قبلی یعنی برچسب وسط (m) را خواهد گرفت.

۴- اگر نقطهٔ قبلی برچسب پایین داشته باشد، در صورتی که عرض نقطه فعلی کمتر از lbl-pt/۲ باشد این نقطه برچسب وسط می‌گیرد.

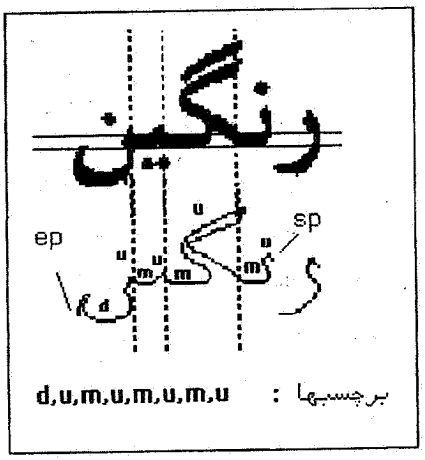
۴-۳- مرحلهٔ سوم: تعیین پاره مسیرهای برچسب خورده

نقاط همسایه در کانتور که برچسب یکسانی داشته باشند، پاره مسیرهایی را تشکیل می‌دهند که برچسب نقاط خود را می‌گیرند و طول هر کدام برابر تعداد نقاط تشکیل دهندهٔ آنهاست. در تعیین این پاره مسیرها قواعد زیر را در نظر داریم:

- اگر اندازهٔ یک پاره مسیر کمتر از pt/۲+۱ باشد، آن را به عنوان ادامهٔ پاره مسیر قبلی در نظر می‌گیریم.
- اگر حذف یک پاره مسیر با استفاده از قاعدهٔ بالا باعث شود که

جدول ۱- خلاصهٔ نتایج الگوریتم جداسازی (قبل از مرحلهٔ تأیید و تصحیح)

تعداد حروف	تعداد حروف متصل	درصد جداسازی حروف متصل	درصد کل جداسازی
۱۱۳۴۷	۸۰۵۶	۹۱	۹۳



شکل ۹- پاره مسیرهای برچسب خورده و نقاط جداسازی بالا (u)، وسط (m) و پایین (d)

دوپاره مسیر همنوع درکنار یکدیگر قرارگیرند، آنها را نیز به یکدیگر متصل کرده و یک پاره مسیر بزرگتر تشکیل می‌دهیم.

- اگر یک پاره مسیر پایین با اندازهٔ بزرگ در جایی جز انتهای کانتور بالایی داشته باشیم، به معنی اتصال دو کلمه است که برای رفع آن، انتهای پاره مسیر پایین را به عنوان انتهای کلمه اول در نظر گرفته و ادامهٔ کانتور را به عنوان یک کلمهٔ جدید در نظر می‌گیریم. این قاعده باعث رفع برخی اتصالات بین کلمات می‌شود.

۴-۴- مرحلهٔ چهارم: تعیین نقاط جداسازی

نقطهٔ جداسازی آخرین نقطه از یک پاره مسیر با برچسب وسط است که در شرایط زیر صدق می‌کند (pt پهنای قلم است):

الف- طول این پاره مسیر از مقدار pt+۱ بیشتر باشد.

ب- پاره مسیر قبلی برچسب بالا داشته و طول آن از مقدار pt کمتر نباشد.

ج- طول پاره مسیر بعدی اگر برچسب آن بالاست از ۱/۵pt بزرگتر بوده و در صورتی که برچسب آن پایین است از ۲/۵pt بیشتر باشد.

این مقادیر آستانه با توجه به قلمهای مختلف معمول در نگارش فارسی تعیین شده و باعث می‌شوند که الگوریتم نسبت به نویزهای معمول در دقتهای پایین تصویر برداری و پهنای کم قلم مصونیت داشته باشد. شکل (۹) مراحل اجرای الگوریتم جداسازی را نشان می‌دهد.

الگوریتم بالا روی متون مختلفی که با ۲۰ قلم معمول و چند نمونه قلمهای قدیمیتر چاپ شده‌اند آزموده شده است. خلاصهٔ نتایج در جدول (۱) آمده است. آزمون بالا کارایی مناسب الگوریتم جداسازی را نشان می‌دهد. خروجی الگوریتم نقاط اولیه جداسازی‌اند که در مرحلهٔ بعدی الگوریتم تأیید یا تصحیح می‌شوند.

۵- تأیید و تصحیح نقاط جداسازی

برخی از خطاهای الگوریتم تا قبل از مرحلهٔ تأیید و تصحیح نهایی نقاط جداسازی همراه با نتیجهٔ اصلاح شده آنها در جدول (۲) نشان داده شده‌است. با توجه به ویژگیهای خاص نگارش فارسی و نحوهٔ عملکرد الگوریتم می‌توان اکثر این خطاها را حذف کرد. در زیر به بررسی انواع مختلف این خطاها و راه حلهای در نظر گرفته شده در الگوریتم تأیید و تصحیح نقاط جداسازی می‌پردازیم.

- حروف "ب"، "پ"، "ت"، "ث"، "ک" و "گ" وقتی که در انتهای کلمه قرار بگیرند، در انتهای خود یک پاره مسیر بالا ایجاد می‌کنند و باعث می‌شوند که انتهای این حروف به عنوان یک حرف مجزا جداسازی شود. تنها حروفی که می‌توانند مشابه این حالت را ایجاد کنند حروف "ا" و "ه" در انتهای زیر کلمه‌اند. حرف "ا" به خاطر ارتفاع زیاد آن به راحتی قابل بازشناسی است و حرف "ه" نیز اگر تصویربرداری دارای کیفیت مناسب باشد، به خاطر وجود حفره شناسایی می‌شود و در ضمن پهنای پاره مسیر ایجاد شده در این حالت بیشتر از حالت انتهایی ایجاد شده در حروف مذکور است، جدول (۲) نمونهٔ (۱).

- انتهای حروف "د" و "ذ" در بعضی از رسم‌الخطها یک پاره مسیر بالا ایجاد می‌کند و به عنوان یک حرف مجزا جداسازی می‌شود. این پاره مسیر انتهایی با پهنای کم آن (کمتر از پهنای قلم) شناسایی شده

جدول ۲- نمونه‌هایی از اثر الگوریتم تأیید و تصحیح نهایی روی نقاط جداسازی

شماره نمونه	قبل از اجرای الگوریتم تایید و تصحیح نهایی	بعد از اجرای الگوریتم تایید و تصحیح نهایی
۱	مقاومت تکنیک	مقاومت تکنیک
۲	کمند	کمند
۳	کم	کم
۴	کشور س	کشور س
۵	فصل نص	فصل نص
۶	پیر جلب	پیر جلب

و به زیر حرف قبلی متصل می‌شود، جدول (۲) نمونهٔ (۲).

- حرف "م" در بعضی از رسم‌الخطها در انتهای کلمه به دو زیر حرف شکسته می‌شود، که زیر حرف انتهایی به خاطر ارتفاع زیاد آن شناسایی شده و به حرف اصلی متصل می‌شود، جدول (۲) نمونهٔ (۳).

- دندانه‌های "س" و "ش" نیز در بسیاری از موارد باعث شکسته شدن این حروف به چند زیر حرف می‌شوند. برای رفع این نوع خطا الگوریتمی نوشته شده که بتواند نوع و محل نقاط و وجود دانداه را شناسایی کند، جدول (۲) نمونهٔ (۴). اگر پهنا و ارتفاع تصویر نقاط و پهنای قلم را به ترتیب با w و h و pt نشان دهیم، الگوریتم شناسایی نوع نقاط را می‌توان به صورت زیر خلاصه کرد:

اگر ۰/۸ > h/w > ۱/۳ و ۰/۷ > w/pt > ۱/۶ باشد آن گاه تصویر مورد نظر یک تک نقطه است.

اگر ۰/۳ < h/w < ۰/۷ و ۱/۵ < w/pt < ۳/۵ باشد آن گاه تصویر مورد نظر یک دو نقطه است.

اگر ۰/۷ < h/w < ۱/۳ و ۱/۵ < w/pt < ۳/۵ باشد آن گاه تصویر مورد نظر یک سه نقطه است.

اگر دو تک نقطهٔ منفرد داشته باشیم که فاصلهٔ طولی آنها کمتر از pt/۲+۱ است آن دو را به عنوان یک دو نقطه در نظر می‌گیریم.

اگر یک تک نقطه داشته باشیم که در زیر آن یک دو نقطه وجود دارد، آنها را با هم به‌عنوان یک سه نقطه در نظر می‌گیریم.

دندانه را نیز به صورت زیر تعریف می‌کنیم:

هرگاه در کانتور بالایی یک پاره مسیر بالا وجود داشته باشد که اندازهٔ آن کمتر از ۱/۵pt است و در کانتور پایین کلمه در همان محل یک فرورفتگی وجود داشته باشد. آن پاره مسیر بالا را به عنوان یک دندانه در نظر می‌گیریم.

بعد از شناسایی کامل نقاط و دندانه‌ها در کلمه با اجرای قواعد زیر شکستگی حروف "س" و "ش" را اصلاح می‌کنیم:

- اگر در ابتدای کلمه بعد از یک پاره مسیر بالا دو دندانه متوالی داشته باشیم که روی آنها هیچ نقطه‌ای نیست یا یک سه نقطه‌ای در بالای آن قرار دارد، دو دندانه مورد نظر را به یکدیگر متصل می‌کنیم (حرف "س" یا "ش" در ابتدای کلمه).

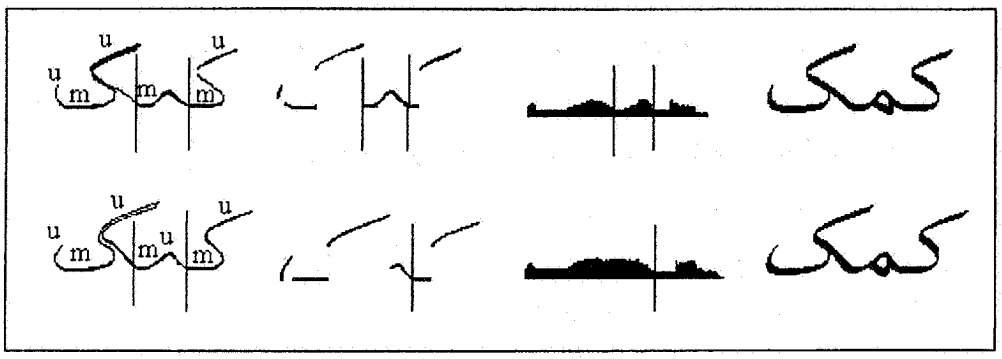
- اگر در میانه کلمه سه دندانه داشته باشیم که روی آنها هیچ نقطه‌ای نیست یا یک سه نقطه در بالای آنها قرار دارد، آنها را به یکدیگر متصل می‌کنیم (حرف "س" و "ش" در میانه کلمه).

- اگر در انتهای کلمه دو دندانه متوالی وجود داشته که روی آنها هیچ نقطه‌ای نیست یا یک سه نقطه در بالای آنها قرار دارد و پاره مسیرهای بعدی آن به ترتیب بالا و پایین بوده و اندازهٔ پاره مسیر پایین از ۳pt بزرگتر باشد، این دندانه‌ها و پاره مسیر پایین انتهایی را به یکدیگر متصل می‌کنیم (حرف"س" یا "ش" در انتهای کلمه).

- اگر یک دندانه منفرد داشته باشیم که روی آن هیچ نقطه‌ای نیست آن را به حرف قبلی خود متصل می‌کنیم (حرف "ص" یا "ض")

جدول ۳- خلاصه نتایج نهایی الگوریتم جداسازی

تعداد حروف	تعداد حروف متصل	درصد جداسازی حروف متصل	درصد کل جداسازی
۱۱۳۴۷	۸۰۵۶	۹۸/۵	۹۸/۹



(ج) (ب) (الف)

شکل ۱۰- نمونه‌ای از اثر کج شدگی در تعیین نقاط جداسازی توسط الگوریتمهای مختلف

الف- الگوریتمهای جداسازی براساس هیستوگرام عمودی

ب- الگوریتمهای جداسازی براساس پروفیل دید از بالا

ج- الگوریتم مطرح شده در این مقاله

جدول (۲) نمونهٔ (۵).

به کارگیری الگوریتم مرحلهٔ تأیید و تصحیح، خطاهای اولیه الگوریتم را تا حد زیادی کاهش می‌دهد. جدول (۳) خلاصهٔ نتایج نهایی الگوریتم را نشان می‌دهد.

۵-۱- اثر کج شدگی حروف

کج شدگی حروف از نوع ایتالیک یا ایرانیک در بسیاری از الگوریتمهای جداسازی می‌تواند ایجاد اشکال کند. برای مثال در الگوریتمهایی که بر اساس هیستوگرام عمودی یا پروفیل دید از بالای زیر کلمات عمل می‌کنند، کج شدگی حروف می‌تواند باعث همپوشانی بعضی از حروف مجاور شده و در نتیجه نقطهٔ جداسازی آنها را پنهان کند. ولی الگوریتم مطرح شده در این مقاله چون از کانتور بالایی استفاده می‌کند همپوشانی حروف روی آن اثری ندارد و چون نقطهٔ جداسازی را با توجه به برچسب پاره مسیرهای متوالی و اندازهٔ آنها تعیین می‌کند و اندازه و برچسب پاره مسیرها با کج شدن آنها تغییر چندانی نمی‌کند، کج شدگی حروف تأثیر قابل ملاحظه‌ای در نتیجه جداسازی آنها نخواهد داشت، شکل (۱۰).

۵-۲- اثر کشیدگی حروف

کشیدگی حروف باعث می‌شود که پاره مسیر وسط در نقطهٔ جداسازی بزرگتر از حد معمول باشد. این مسئله باعث خطا در جداسازی نمی‌شود. اما اشکالی که ایجاد می‌شود این است که حروف در زمان کشیدگی یک دنباله دارند که اندازهٔ آن نامشخص بوده و در مرحله بازشناسی می‌تواند ایجاد اشکال کند. برای رفع این مشکل چون کانتور بالایی حروف را به پاره مسیرهای برچسب خورده تبدیل کرده‌ایم و دنبالهٔ ایجاد شده دارای برچسب وسط (m) است، برای حذف اثر آن اندازه پاره مسیر وسطی را که در دنبالهٔ حروف جدا شده قرار دارد به پهنای قلم محدود می‌کنیم، جدول (۲) نمونهٔ (۶).

الگوریتم مطرح شده در این مقاله با زبان C روی یک رایانهٔ شخصی ۴۸۶DX۴ با فرکانس ۱۰۰ مگاهرتز پیاده سازی شده است. سرعت متوسط جداسازی در حدود ۷۴/۳ کاراکتر در ثانیه است.

۶- به کارگیری نتایج الگوریتم جداسازی در مرحلهٔ بازشناسی

الگوریتم مطرح شده علاوه بر اینکه حروف زیر کلمه را با دقت

بالا جداسازی کرده و آنها را برای مرحله بازشناسی آماده می‌کند [۱۰]، با استفاده از الگوریتم تأیید و تصحیح مطرح شده در بخش قبلی و به کارگیری اطلاعات حاصل از نوع و محل نقاط، تعداد حفره‌ها، پهنا و ارتفاع پاره مسیرهای ایجاد شده، برخی از حروف زیر کلمه را نیز شناسایی می‌کند. این حروف و نحوه شناسایی این حروف در زیر ارائه می‌شود

- حروف "س" و "ش" در هر جا از زیر کلمه مطابق آنچه در توضیح الگوریتم تأیید و تصحیح گفته شد.
- حرف "م" در انتهای زیر کلمه مطابق آنچه در توضیح الگوریتم تأیید و تصحیح گفته شد.
- حروف "ص" و "ض" در انتهای زیر کلمه به خاطر وجود دندانه، حفره و یک پاره مسیر پایین با اندازهٔ بزرگ در انتها.
- حرف "ا" در انتهای زیر کلمه به خاطر ارتفاع زیاد و پهنای کم.
- حرف "ل" در ابتدا و میانه به خاطر ارتفاع زیاد و پهنای کم و در انتهای زیر کلمه به خاطر داشتن یک پاره مسیر بالا و یک پاره پایین هر دو با اندازهٔ بزرگ.
- حروف "ک" و "گ" در هر جا از زیر کلمه به خاطر ارتفاع و پهنای زیاد و داشتن سرکش در بعضی از رسم‌الخطها.
- حرف "ه" در انتهای زیر کلمه به خاطر داشتن یک حفره و یک پاره مسیر بالا با پهنایی بیشتر از ۱/۵ برابر پهنای قلم.

۷- نتیجه‌گیری

هدف از طراحی این الگوریتم مشخص کردن نقاط جداسازی است. حروف جدا شده در مرحلهٔ بعد به الگوریتم بازشناسی ارائه شده و در آنجا شناسایی نهایی صورت می‌گیرد. بررسی خطاهای باقیمانده نشان می‌دهد که اکثر آنها در اثر کیفیت پایین تصویربرداری، کم بودن دقت ۲۰۰ نقطه در اینچ برای برخی از قلمها و کج شدگی بیش از حد متن در زمان تصویر برداری است. الگوریتم علاوه بر جداسازی حروف با دقت بالا برخی از حروف را نیز بازشناسی می‌کند.

مزیت عمدهٔ استفاده از کانتور بالایی این است که حروف همپوشان مشکلی را در جداسازی ایجاد نمی‌کنند. همچنین حفره‌ها، نقاط و علائم در مرحلهٔ جداسازی تشخیص داده می‌شوند. الگوریتم ارائه شده در این مقاله اطلاعات را به شکل مناسب برای محاسبهٔ دقیق نوار زمینه، انجام جداسازی با خطای کم و بازشناسی مستقیم برخی از حروف بدون نیاز به مرحله بازشناسی به کار گرفته است. روشهایی که از هیستوگرام عمودی یا پروفیل دید از بالای کلمات استفاده می‌کنند این مزایا را ندارند.

مراجع

1. Bokser, M., "Omnidocument Technologies," *Proc. of IEEE*, Vol. 80, No. 7, pp. 1066-1078, July 1992.
2. Mori, S., Suen, C.Y., and Yamamoto, K., "Historical Review of OCR Research and Development," *Proc. of IEEE*, Vol. 80, No. 7, pp. 1029-1058, July 1992.
3. Parhami, B., and Taraghi, M., "Automactic Recognition of Printed Farsi Text," *Pattern Recognition*, Vol. 14, pp. 395-403, 1981.

۴. کبیر، ا.، بهاری ک.، و احمدزاده م.، "بازشناسی متون تایپ شدهٔ فارسی،" *مجموعهٔ مقالات اولین کنفرانس مهندسی برق ایران*، جلد ۲، ص ۲۸۵-۲۹۴، دانشگاه صنعتی امیرکبیر، تهران ۱۳۷۲.

۵. فهیمی، ح.، و حاتم، ا.، "ارائهٔ یک روش ساختاری برای تشخیص حروف در متن تایپ شدهٔ فارسی،" *مجموعهٔ مقالات سومین کنفرانس مهندسی برق ایران*، ص ۱۹۰-۱۹۷، دانشگاه علم و صنعت، تهران ۱۳۷۴.

۶. رفیعی، ش.، و کبیر. ا.، "شکستن کلمات تایپ شده فارسی به حروف در رسم‌الخطهای مختلف،" *مجموعهٔ مقالات سومین کنفرانس الکترونیک*، ص ۹۸-۱۰۴، دانشگاه شیراز، مهر ۱۳۷۴.

۷. فهیمی، ح.، و تیمساری، ب.، "بازشناسی حروف در کلمات تایپ شده فارسی با استفاده از روش مرفولوژی،" *مجموعهٔ مقالات اولین کنفرانس مهندسی برق ایران*، جلد ۲، ص ۲۷۷-۲۸۴، دانشگاه صنعتی امیرکبیر، تهران ۱۳۷۲.

۸. میرزاخانی، ح. و فائز، ک. "روش نوین در شناسایی متون فارسی به کمک شبکه‌های عصبی،" *مجموعهٔ مقالات سومین کنفرانس الکترونیک*، ص ۴۹-۱۵۴، دانشگاه شیراز، ۱۳۷۴.

9. Fu, K.S., *Syntactic Methods in Pattern Recognition*, Academic Press, 1974.

۱۰. عزمی، ر.، و کبیر، ا.، "ارائهٔ دو الگوریتم برای شناسایی حروف چاپی فارسی،" *مجموعهٔ مقالات سومین کنفرانس بین‌المللی انجمن کامپیوتر ایران*، ص ۱۹۱-۱۹۷، دانشگاه علم و صنعت ایران، تهران ۱۳۷۶.