

الگوریتم اصلاح شده جداسازی حروف در متون چاپی با برجسب زدن به کانتور بالایی کلمات

حسین نظام‌آبادی‌پور* ، احسان اله کبیر** و رضا عزمی***
بخش مهندسی برق، هسته پژوهشی پردازش تصویر، دانشگاه شهید باهنر کرمان
بخش مهندسی برق، دانشگاه تربیت مدرس
بخش مهندسی کامپیوتر، دانشگاه الزهرا

(دریافت مقاله: ۸۰/۴/۳۰ - دریافت نسخه نهایی: ۸۲/۶/۲۴)

چکیده - در این مقاله با اصلاح الگوریتم عزمی که مبتنی بر کانتور بالایی کلمات است، الگوریتم جداسازی مناسبی برای متون با کیفیت چاپی پایین ارائه شده است. برای حل مشکل نایکنواختی نوار زمینه خط، روش مناسبی برای تعیین نوار زمینه محلی پیشنهاد شده و با اصلاح روش برجسب زنی کانتور بالایی و تکمیل قواعد جداسازی، دقت الگوریتم افزایش داده شده است. نرخ جداسازی درست حروف ۹۷٪ است. بر اساس نتایج به دست آمده، بررسی دقیقی درباره علل خطاها ارائه شده است که می‌تواند راهگشای تحقیقات بعدی در این زمینه باشد.

واژگان کلیدی: جداسازی حروف، متون چاپی فارسی، نوار زمینه محلی، کانتور بالایی، هیستوگرام، کد زنجیره‌ای، برجسب زنی

A Modified Character Segmentation Algorithm for Farsi Printed Text Using Upper Contour Labelling

H. Nezamabadi-Pour , E. Kabir and R. Azmi

Image Processing Center, Department of Electrical Engineering, Shahid Bahonar University of Kerman
Department of Electrical Engineering, Tarbiat Modarres University
Department of Computer Engineering, Alzahra University

Abstract: *In this paper, a modified segmentation algorithm for printed Farsi words is presented. This algorithm is based on a previous work by Azmi that uses the conditional labeling of the upper contour to find the segmentation points. The main objective is to improve the segmentation results for low quality prints. To achieve this, various modifications on local baseline detection, contour labeling and segmentation rules have been applied. In an experiment, the correct segmentation rate was 97%. Based on the results obtained, a detailed error analysis is presented which should be useful for further research on this topic.*

Keywords: *Character Segmentation, Farsi Printed text, Local baseline, Upper contour, Histogram, Segmental code, Labeling*

این تحقیق در چهارچوب طرح ملی "بازشناسی متون چاپی و حجم محدود کلمات دستنویس" انجام شده است.

lbl	لبه پایینی نوار زمینه	d	برچسب پایین
m	برچسب وسط	$h_0(n)$	هیستوگرام ارتفاع نقاط
p	یک نقطه روی کانتور		روی کانتور بیرونی با کد ۰
pt	پهنای قلم محلی	$h_{20}(n)$	هیستوگرام هموار شده ارتفاع نقاط روی کانتور بیرونی با کد ۰
ptg	پهنای قلم	$h_4(n)$	هیستوگرام ارتفاع نقاط روی کانتور بیرونی با کد ۴
ubl	لبه بالایی نوار زمینه	$h_{24}(n)$	هیستوگرام هموار شده ارتفاع نقاط روی کانتور بیرونی با کد ۴
u	برچسب بالا		
$x(p)$	طول نقطه p		
$y(p)$	عرض نقطه p		

۱- مقدمه

بازشناسی حروف، محوریت‌ترین بخش در پردازش مستندات است. در متون چاپی انگلیسی، مسئله جداسازی تنها در چسبیدگیهای ناخواسته حروف مطرح می‌شود که ناشی از کیفیت پایین چاپ یا تصویربرداری است [۱]. در متون دستنویس انگلیسی که حروف ممکن است به شکل پیوسته نوشته شوند، تحقیقات زیادی در مورد جداسازی کلمات به حروف و زیر حروف انجام شده است که بعضی از ایده‌های مطرح در آنها را می‌توان برای جداسازی حروف چاپی فارسی و عربی نیز به کار گرفت [۲]. در بازشناسی متون چاپی فارسی و عربی سه رویکرد کلی قابل تصور است [۳]. در رویکرد اول جداسازی و بازشناسی حروف به طور مستقل از یکدیگر انجام می‌شود. در رویکرد دوم که برای مجموعه محدودی از کلمات مناسب است، از جداسازی حروف صرف‌نظر می‌شود و کل کلمه یا زیر کلمه به عنوان یک الگوی واحد بازشناسی می‌شود. در رویکرد سوم جداسازی و بازشناسی حروف به شکل ترکیبی انجام می‌شود. در رویکردهای اول و سوم جداسازی حروف یکی از اساسی‌ترین مراحل است و نتایج بازشناسی نهایی به شدت وابسته به آن است.

در زمینه جداسازی حروف چاپی فارسی و عربی تحقیقات قابل توجهی انجام شده است. در بعضی از این تحقیقات، با استفاده از قواعد حاکم بر یک قلم خاص روشهایی برای جداسازی حروف نوشته شده با آن قلم ارائه شده است [۴ و ۵]

برای جداسازی حروف از ویژگیهای مختلفی می‌توان استفاده کرد. در بعضی روشها از کانتور کلمات استفاده شده است [۶ و ۷]. در هر ستون تصویر یک کلمه، فاصله میان دورترین نقاط سیاه حساب می‌شود. اگر این مقدار از یک حد آستانه کمتر باشد، آن ستون به عنوان نامزدی برای جداسازی در نظر گرفته می‌شود [۶]. در برخی تحقیقات، جداسازی حروف براساس هیستوگرام عمودی نقاط سیاه انجام شده است [۸-۱۱]. در اینجا فرض شده است که مقدار هیستوگرام عمودی در نقاط اتصال حروف از مقدار متوسط این هیستوگرام در کلمه کمتر است. در الگوریتم دیگری، از پروفیل بالایی کلمات برای جداسازی حروف آنها استفاده شده است [۱۲ و ۱۳]. در این الگوریتم به نقاط پروفیل بالایی کلمه با توجه به موقعیت آنها نسبت به نوار زمینه بر چسبهای مناسبی زده می‌شود. سپس محل‌های اتصال حروف با اعمال قواعد خاصی پیدا می‌شوند. در تحقیق دیگری از نازکسازی استفاده شده است [۱۴]. در این تحقیق، ابتدا از تصویر نازک شده زیر کلمه استخراج می‌شود. سپس کدهای فریمن مربوط به آن برای تعیین نقاط جداسازی به یک درخت تصمیم دودویی^۱ سپرده می‌شوند.

در روش دیگری، از برچسب زدن به کانتور بالایی کلمات استفاده شده است [۱۵، ۱۶، ۳]. این روش نسبت به سایر روشها مزایایی دارد که در اینجا به مهمترین آنها اشاره می‌شود: الف) به همپوشانی عمودی حساسیت ندارد. ب) به پهنای قلم حساس نیست. ج) برای قلمهای معمول و متداول، به نوع قلم حساس

با پهنای قلم که بیشترین تعداد عناصر سیاه در تصویر یک خط متن را در خود داشته باشد، شکل (۲). برای تعیین نوار زمینه کلی، ابتدا هیستوگرام افقی عناصر سیاه یک خط متن را محاسبه می‌شود. سپس ردیفهای متوالی این هیستوگرام به اندازه پهنای قلم، با یکدیگر جمع شده و هیستوگرام هموار شده‌ای فراهم می‌شود. محل ماکزیمم این هیستوگرام، نوار زمینه کلی برای آن خط از متن را مشخص می‌کند. به عبارت دیگر می‌توان گفت: نوار زمینه نواری است که کلمات روی آن نوشته می‌شوند. بعد از این تصویر هر خط متن بطور مجزا پردازش می‌شود.

اولین مرحله از پردازش خطوط تصویر متن، پیدا کردن بخشهای همپوشان عمودی است. بخشهای همپوشان به مجموعه‌ای از زیرکلمات گفته می‌شود که هیستوگرام عمودی آنها پیوسته است، شکل (۲). بعد از جداسازی بخشهای همپوشان عمودی، پهنای قلم به صورت محلی برای هر بخش محاسبه شده و کانتور زیرکلمات آن تعیین می‌شود. با استفاده از اطلاعات کانتور، نوار زمینه محلی برای هر بخش همپوشان تعیین می‌شود. در ادامه، نقاط کانتور بالایی بر اساس فاصله آنها با نوار زمینه محلی و شیب منحنی برجسته زده می‌شوند. سپس کانتور بالایی به پاره مسیرهایی تجزیه می‌شود که برجسته خاص خود را می‌گیرند. نقاط جداسازی با استفاده از قواعدی خاص آشکار می‌شود.

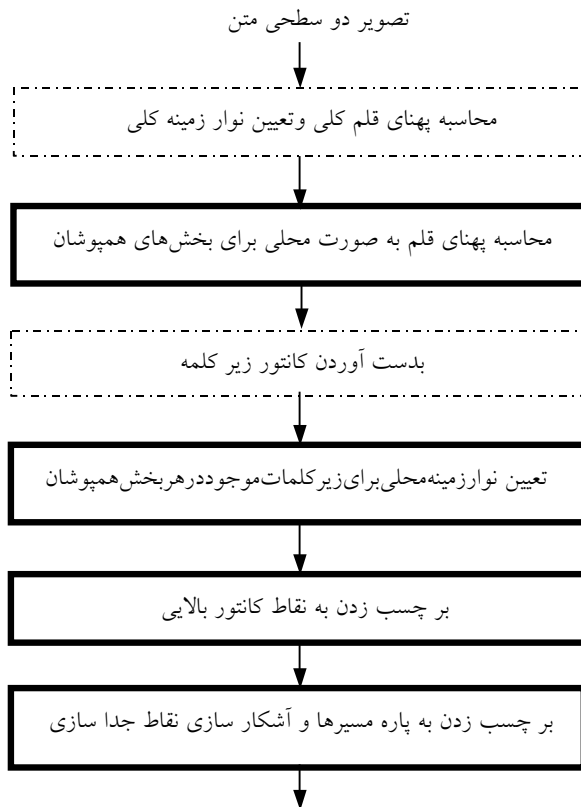
برای جداسازی حروف بهتر است مرز بین حروف تک دندان به عنوان نقطه جداسازی تعیین شود و حروف "س"، "ش"، "ص" و "ض" با وجود داشتن دندان به زیر حروف شکسته می‌شوند. در متون با کیفیت مناسب، می‌توان در مرحله پس پردازش مشکل شکسته شدن این حروف به زیر حروف را برطرف کرد. این کار برای متون با کیفیت نامناسب، به سادگی میسر نیست. بر خلاف تحقیق قبلی [۳]، در اینجا سعی می‌شود دندانها بدون توجه به اینکه متعلق به چه حرفی هستند، جدا شوند و کارایی این رویکرد نسبت به رویکرد قبلی بررسی شود. تاکید بر تعیین دقیق نوار زمینه در این مقاله، برای رسیدن به دقت بالا در جداسازی حروف به خصوص دندانهاست.

نیست. د) بعضی از حروف در مرحله جداسازی، بازشناسی هم می‌شوند.

در این مقاله اصلاح الگوریتم مرجع [۳] برای جداسازی حروف در متون با کیفیت چاپی نامناسب مورد نظر است [۱۶]. عدم توانایی این الگوریتم در رویارویی با مشکلات متون با کیفیت چاپی نامناسب از جمله یکنواخت نبودن پهنای قلم در یک خط و حتی یک زیر کلمه، بالا و پایین نوشته شدن زیر کلمات و نرم و کوتاه بودن دندانها از ضعفهای آن محسوب می‌شود. در الگوریتم اصلاح شده، بسیاری از نقاط ضعف الگوریتم عزمی بر طرف شده است که در بخشهای بعدی به آن پرداخته می‌شود. لازم به ذکر است که در این تحقیق از متون ساده استفاده شده است که شامل شکل، جدول یا گراف نیستند. بخش دوم این مقاله به کلیات الگوریتم جداسازی می‌پردازد. بخش سوم به پردازشهای اولیه اختصاص دارد. در بخش چهارم، چگونگی تعیین نوار زمینه محلی بیان می‌شود. بخش پنجم نحوه برجسته زنی به نقاط کانتور بالایی را نشان می‌دهد. در بخش ششم روش برجسته زنی پاره مسیرهها وقوانین جداسازی ارائه می‌شود. بخش هفتم به نتایج و بررسی دقیق انواع خطاها می‌پردازد. در بخش هشتم اثرات درجه تفکیک بررسی می‌شود و نتیجه گیری نهایی در بخش نهم مطرح می‌شود.

۲- کلیات الگوریتم جداسازی

نمودار جعبه‌ای عمومی الگوریتم جداسازی، در شکل ۱ آمده است. چار چوب اصلی این الگوریتم از مرجع [۳] گرفته شده است. ورودی الگوریتم، تصویر متن به صورت دو سطحی است. در اولین مرحله خطوط متن از یکدیگر جدا می‌شوند. سپس برای هر خط متن پهنای قلم کلی محاسبه شده و موقعیت نوار زمینه تعیین می‌شود. برای محاسبه پهنای قلم، هر خط از متن ورودی در راستاهای عمودی و افقی جاروب می‌شود و قطعات متصل از عناصر سیاه در هر ستون و ردیف بر حسب اندازه شمرده می‌شوند. اندازه‌ای که بیشترین فراوانی را داشته باشد به عنوان پهنای قلم در نظر گرفته می‌شود. نوار زمینه، نواری است

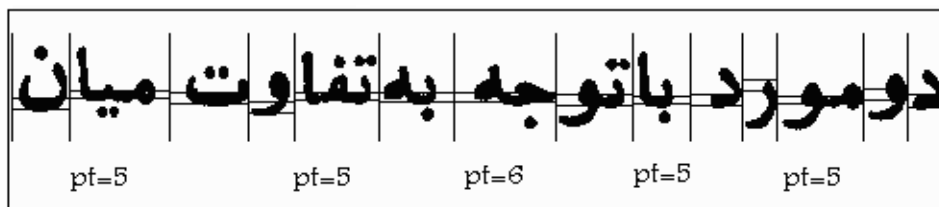


حروف جدا شده

شکل ۱- در نمودار جعبه‌ای عمومی الگوریتم جداسازی، جعبه‌هایی که با خط پر مشخص شده‌اند، نسبت به آنچه در مرجع [۳] آمده است تغییراتی انجام شده است.



(الف)



(ب)

شکل ۲- اثر اصلاح پهنای قلم به صورت محلی

الف) برای هر بخش همپوشان نوار زمینه باتوجه به پهنای قلم کلی محاسبه شده است (بخشهای همپوشان توسط خطوط عمودی از یکدیگر جدا شده‌اند)
ب) برای هر بخش همپوشان پهنای قلم محلی محاسبه شده است که در زیر آن نوشته شده و برای بخشهایی که در شکل نوشته نشده، $pt=7$ است.

۳ - پردازش‌های اولیه

مجموعه تصاویر استفاده شده در این آزمایش، از کتابهای مختلف تهیه شده است. در انتخاب این مجموعه سعی شده است که از متونی با قلم‌های متفاوت استفاده شود. از هر صفحه متن بادرجه تفکیک ۳۰۰ dpi به صورت دو سطحی تصویر برداری شده است. تصویرها با یک روبشگر Scan Jet 6300HP با تنظیم پیش فرض گرفته شده است.

۱-۳- محاسبه پهنای قلم و تعیین نوار زمینه کلی

الگوریتم جداساز، تصویر متن را به عنوان ورودی دریافت می‌کند. در اولین مرحله، خطوط متن با استفاده از هیستوگرام افقی نقاط سیاه از یکدیگر جدا می‌شوند. سپس پهنای قلم کلی، ptg ، برای هر خط متن محاسبه می‌شود. مرحله بعد، تعیین نوار زمینه کلی براساس هیستوگرام افقی است. سپس بخشهای همپوشان با استفاده از هیستوگرام عمودی نقاط سیاه از یکدیگر جدا می‌شوند [۳].

۲-۳ محاسبه پهنای قلم به صورت محلی برای بخشهای همپوشان

در متون با کیفیت پایین مثل متون قدیمی و متونی که با دستگاه کپی نامناسب تهیه شده اند، مواردی مشاهده می‌شود که پهنای قلم در تمام متن ثابت نیست و در بخشهایی از متن تجمع جوهر و در بخشهایی دیگر خوردگی وجود دارد. با توجه به اینکه الگوریتم جدا ساز وابستگی زیادی به محل دقیق نوار زمینه دارد و پهنای قلم در تعیین عرض نوار زمینه اثر دارد، لازم است پهنای قلم به صورت محلی تعیین شود.

نحوه محاسبه پهنای قلم محلی، pt ، مانند محاسبه ptg است، با این تفاوت که pt برای یک بخش همپوشان به دست می‌آید. اگر پهنای قلم محلی از $0/65 ptg$ بزرگتر و از $1/1 ptg$ کوچکتر باشد، تأیید می‌شود. در غیر این صورت pt برای آن بخش همپوشان همان ptg در نظر گرفته می‌شود. این ضرایب بر اساس تجربه به دست آمده‌اند. شکل (۲) پهنای قلم کلی را برای

یک عبارت و پهنای قلم محلی را برای بخشهای همپوشان آن نشان می‌دهد.

۳-۳ به دست آوردن کانتور زیر کلمه

برای به دست آوردن کانتور زیر کلمه، سمت راست ترین نقطه در هر زیر کلمه را پیدا کرده و کانتور آن را با پیمایش پاد ساعتگرد به شکل کد فریم با همسایگی ۸ استخراج می‌کنیم.

۴- تعیین نوار زمینه محلی

تعیین دقیق نوار زمینه نقش مهمی در الگوریتم جداسازی دارد. اگر به دلیل تصویر برداری با درجه تفکیک پایین، پهنای قلم کم باشد، خطا حتی در حد یک پیکسل نیز می‌تواند باعث مشکلاتی در جداسازی شود. همچنین با توجه به اینکه در متن‌های قدیمی، نوار زمینه خط ثابت نیست و از کلمه به کلمه ممکن است جابه‌جا شود، به نوار زمینه کلی نمی‌توان اعتماد زیادی داشت. در این بخش روشهایی برای تعیین دقیقتر نوار زمینه به صورت محلی ارائه می‌شود.

۴-۱- تعیین نوار زمینه محلی با استفاده از کانتور زیر کلمات

توضیحات کامل مربوط به این روش در بخش سوم مرجع [۳] آمده است. به طور خلاصه می‌توان چنین گفت که برای هر زیر کلمه، نواری به پهنای $2pt$ در حوالی نوار زمینه کلی جستجو می‌شود. برای هر ردیف تعداد تکرار نقاط کانتور بالایی با کد فریم 4 که از فاصله ذکر شده قرار داشته باشند شمرده شده و هیستوگرام این نقاط، $h_4(n)$ ، محاسبه می‌شود که n متغیر مربوط به شماره ردیفهاست. ردیفی از تصویر که متناظر با حداکثر این هیستوگرام است به عنوان لبه بالایی نوار زمینه 2 ، abl ، تعیین می‌شود. کد 4 جهت از راست به چپ در راستای افق را نشان می‌دهد. به همین ترتیب با استفاده از هیستوگرام کد فریم 0 ، $h_0(n)$ لبه پایینی نوار زمینه 3 ، dbl ، مشخص می‌شود. شکل (۳) نمونه‌ای از تعیین نوار زمینه محلی را نشان می‌دهد.

لازم به ذکر است که پهنای نوار جستجو در مرجع [۳]، $2pt$ در نظر گرفته شده است، ولی به دلیل اینکه در متون قدیمی

من برای آنکه مستوحیرانت کنم نیکو شنو

شکل ۳ - تعیین نوار زمینه محلی برای بخشهای همپوشان

$$|bl|_1 = n_2 \left| h_{20}(n_2) = \max_n \{h_{20}(n)\} \right. \quad (6)$$

$$|bl|_2 = n_2 + 1 \quad (7)$$

$$|bl| = \begin{cases} |bl|_1 & \text{if } h_0(n_2) \geq h_0(n_2 + 1) \\ |bl|_2 & \text{if } h_0(n_2) < h_0(n_2 + 1) \end{cases} \quad (8)$$

همانطور که از این معادلات پیداست، از بین دو ردیف مجاوری که ماکزیمم کدهای فریمین با شماره ۴ در آنها اتفاق افتاده است، ردیفی که تعداد این کدها در آن بیشتر است به عنوان ubl برگزیده می شود. به همین ترتیب از بین دو ردیف مجاوری که ماکزیمم کدهای فریمین با شماره صفر در آنها اتفاق افتاده است، ردیفی که تعداد این کدها در آن بیشتر است به عنوان lbl انتخاب می شود.

اگر پهناي نوار زمینه به دست آمده بیش از ۲۵ درصد با ptg اختلاف داشته باشد، نوار زمینه محلی مانند روش پیشنهادی در بخش (۳) از مرجع [۳] تعیین می شود.

۳-۴- تصحیح نوار زمینه محلی با استفاده از ماکزیممهای اول و دوم هیستوگرامهای $h_{20}(n)$ و $h_{24}(n)$ و موقعیت نوار زمینه کلی

تعیین نوار زمینه براساس ماکزیمم هیستوگرامهای $h_{24}(n)$ و $h_{20}(n)$ در قسمت قبل توضیح داده شد. خطا در تعیین نوار زمینه با این روش بسیار کم است، اما به ندرت خطاهایی ایجاد می شود. در بررسی زیر کلماتی که در آنها خطا رخ داده بود، معلوم شد که در بیشتر آنها $l|bl$ و واقعی ماکزیممهای دوم هیستوگرامهای $h_{20}(n)$ و $h_{24}(n)$ هستند. این موضوع در مورد هیستوگرامهای $h_0(n)$ و $h_4(n)$ نیز اتفاق می افتد. برای رفع این

جابه جایی نوار زمینه کلمات بیشتر است، در این تحقیق پهناي نوار جستجو $3pt$ در نظر گرفته شده است. اگر پهناي نوار زمینه محلی، بیش از ۲۵ درصد با ptg اختلاف داشته باشد، با روشی که در [۳] ارائه شده است، اصلاح می شود.

۴-۲- تعیین نوار زمینه محلی با استفاده از هیستوگرامهای $h_{24}(n)$ و $h_{20}(n)$

در مواردی متعددی مشاهده می شود که نوار زمینه محلی به دست آمده با استفاده از هیستوگرامهای $h_0(n)$ و $h_4(n)$ به علت ناهموار بودن کانتور زیرکلمات در اثر نویز، به اشتباه محاسبه می شود. برای برطرف کردن این اشتباهات از هیستوگرام هموار شده $h_0(n)$ و $h_4(n)$ که $h_{20}(n)$ و $h_{24}(n)$ نامیده شده اند، استفاده می شود، معادلات (۱) و (۲).

$$h_{20}(n) = h_0(n) + h_0(n+1) \quad (1)$$

$$h_{24}(n) = h_4(n) + h_4(n+1) \quad (2)$$

ایده استفاده از هیستوگرامهای $h_{20}(n)$ و $h_{24}(n)$ برای تعیین نوار زمینه محلی، از همان ایده هیستوگرامهای $h_0(n)$ و $h_4(n)$ گرفته شده است، با این نگاه تازه که با استفاده از جمع ۲ تایی ردیفهای هیستوگرامهای $h_0(n)$ و $h_4(n)$ سعی شده است تا این هیستوگرامها هموار شده و اثر نویزهای احتمالی از بین برود. محل نوار زمینه در این روش با معادلات زیر تعیین می شود.

$$|bl|_1 = n_1 \left| h_{24}(n_1) = \max_n \{h_{24}(n)\} \right. \quad (3)$$

$$|bl|_2 = n_1 + 1 \quad (4)$$

$$|bl| = \begin{cases} |bl|_1 & \text{if } h_4(n_1) \geq h_4(n_1 + 1) \\ |bl|_2 & \text{if } h_4(n_1) < h_4(n_1 + 1) \end{cases} \quad (5)$$

محلی ونسبی صعودی یا نزولی بودن کانتور در هر نقطه، شرایط بهتری برای پیدا کردن نقاط جداسازی فراهم شود. همچنین در این الگوریتم برای اینکه دندانه‌های کوتاه تا حد امکان تشخیص داده شوند، تبدیل برچسب بالا به وسط در $ubl-pt/2$ ولی تبدیل برچسب وسط به بالا در ubl انجام می‌شود، شکل (۵).

این موضوع را نیز باید در نظر داشت که در بعضی متون احتمال بالا رفتن یا پائین آمدن آرام کانتور در دندانه‌ها وجود دارد و ممکن است این قبیل دندانه‌ها از دید الگوریتم جا بیفتند. برای جلوگیری از چنین مواردی، شرایط مناسب به صورت OR، در A1 و A2 افزوده شده است، شکل (۵).

۶- برچسب زدن به پاره مسیرها و آشکارسازی نقاط جداسازی

نقاط کنار هم در کانتور بالایی که برچسب یکسانی دارند، پاره مسیری را ایجاد می‌کنند که برچسب آن نقاط را به خود می‌گیرد. طول هر پاره مسیر تعداد نقاط تشکیل دهنده آن است. برای برطرف کردن خطاهای احتمالی در برچسب زدن نقاط به دلیل وجود نویز در کانتور بالایی، پاره مسیرهای کوتاه تر از $pt/3+1$ برچسب پاره مسیر قبلی را به خود می‌گیرند. بعد از ادغام یک پاره مسیر کوتاه با پاره مسیر قبلی، اگر دو پاره مسیر که دارای یک برچسب هستند به یکدیگر برسند، پاره مسیر بزرگتری با همان برچسب تشکیل می‌دهند که طول آن مجموع طولهای دو پاره مسیر قبلی است.

آشکارسازی نقاط جداسازی آخرین مرحله از الگوریتم ارائه شده است. قواعد جداسازی از مرجع [۳] اقتباس شده‌اند ولی با توجه به اهداف خاص این تحقیق تغییراتی در آنها داده شده است.

نقطه جداسازی آخرین نقطه از یک پاره مسیر با برچسب وسط است که در یکی از معادلات زیر صدق کند.

$$p: PU \text{ AND } NUNE \quad (9)$$

$$q: PU \text{ AND } NUELHW \quad (10)$$

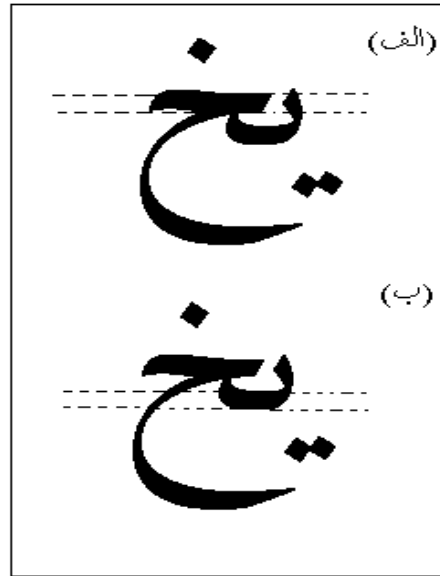
$$r: CL \text{ AND } PU \text{ AND } NDNEL \quad (11)$$

خطاها نوار زمینه محلی با در نظر گرفتن ماکزیممهای اول و دوم هیستوگرامها تعیین می‌شود. اگر ماکزیمم دوم از $0/85$ ماکزیمم اول بزرگتر و به نوار زمینه کلی نزدیکتر باشد، موقعیت آن به عنوان نوار زمینه محلی انتخاب می‌شود. در غیر این صورت موقعیت ماکزیمم اول هیستوگرام، تعیین کننده نوار زمینه محلی است [۱۶]. شکل (۴) نمونه ای از تصحیح نوار زمینه محلی را با استفاده از این روش نشان می‌دهد.

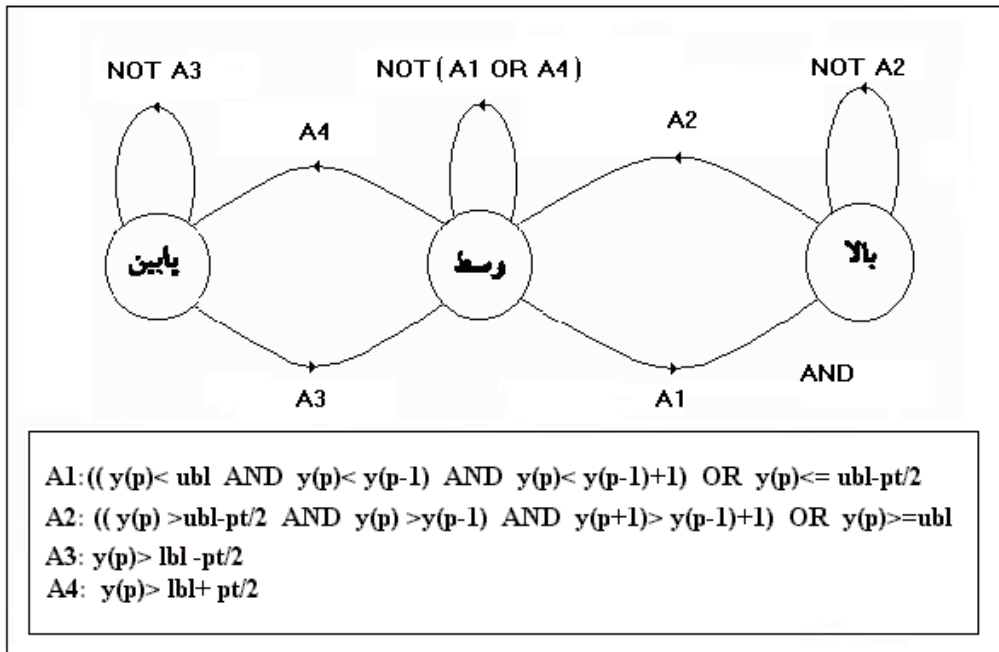
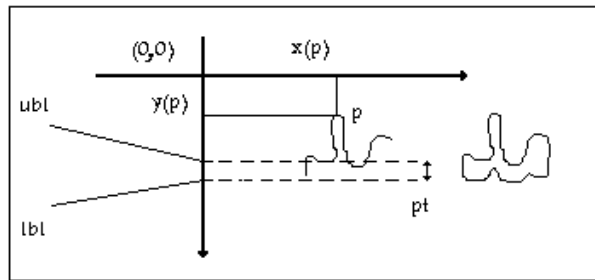
۵- برچسب زدن به نقاط کانتور بالایی

قوانین جداسازی بر اساس کانتور بالایی زیر کلمات بنا شده است. کانتور بالایی زیر کلمات از کانتور پیرامونی آنها استخراج می‌شود [۳]. سمت راست ترین نقطه کانتور برچسب بالا می‌گیرد و نقاط بعدی کانتور طبق شکل (۵) برچسب می‌خورند. به عنوان نمونه، شرط عوض شدن برچسب از بالا به پایین که در شکل (۵) با A_2 نمایش داده شده است، آورده می‌شود. اگر نقطه قبلی برچسب بالا داشته باشد و عرض نقطه فعلی بیشتر از $ubl-pt/2$ باشد و از عرض نقطه قبلی حداقل یکی بیشتر باشد و عرض نقطه بعدی از عرض نقطه قبلی حداقل دو تا بیشتر باشد یا اینکه عرض نقطه فعلی، برابر یا بزرگتر از ubl باشد، نقطه فعلی برچسب وسط می‌گیرد.

الگوریتم برچسب زدن به نقاط کانتور، با آنچه در مرجع [۳] آمده، تفاوتی دارد. چون در اینجا هدف پیدا کردن و جداسازی دندانه‌هاست و گم کردن دندانه‌ها در مواقعی است که ارتفاع آنها کم باشد، از این رو حد آستانه فاصله مطلق تا نوار زمینه برای تبدیل برچسب وسط به بالا افزایش داده شده است ($pt/2$ در شرط A1) و شرط محلی ونسبی صعودی بودن کانتور در نقطه مورد نظر اضافه شده است، شکل (۵). برای گم نکردن دندانه‌هایی که کمی بالاتر از نوار زمینه نوشته شده‌اند، شرط تبدیل برچسب بالا به وسط با کاهش حد آستانه ($pt/2$) در شرط A2) و اضافه کردن شرط نزولی بودن کانتور اصلاح شده است شکل (۵). به عبارت دیگر سعی شده است که با بهینه کردن شرایط تبدیل برچسب‌ها و نیز استفاده از اطلاعات



شکل ۴- تصحیح نوار زمینه محلی با استفاده از ماکزیمم دوم هیستوگرامهای $h_{20}(n)$ و $h_{24}(n)$
 الف- نوار زمینه محلی اولیه ب- نوار زمینه محلی اصلاح شده



شکل ۵- نمودار حالت برای تعیین بر چسبهای کانتور بالایی زیر کلمه

الف) نوار زمینه محلی این تک حرفها حداقل به اندازه 2/pt پائین تر از زمینه کلی قرار می گیرد.

ب) کانتور بالایی این تک حرفها از سه پاره مسیر بالا، وسط و بالا تشکیل می شود.

ج) طول پاره مسیر وسط این تک حرفها از 2/5pt بیشتر است.

د) ارتفاع دو پاره مسیر بالا از 2pt بیشتر است.

ه) پهنای پاره مسیر بالایی که بعد از پاره مسیر وسط قرار دارد، از 2pt کمتر است.

۷- نتایج آزمایش، معرفی خطاها و بررسی علل آنها

مجموعه تمرین از 24 نوع قلم متفاوت، شامل 2428 حرف تشکیل می شود، جدول (1) و مجموعه آزمایش از 6 متن، شامل 6415 حرف انتخاب شده است [17] (متون شماره 1 تا 6 در جدول 1). یک متن برای آزمودن توانایی الگوریتم روی متون با کیفیت چاپ مناسب انتخاب شده است، جدول (1). نتایج اجرای الگوریتم جداسازی حروف را نشان می دهد. در این جدول خطاهای 6 گانه که در ادامه تعریف می شوند به دو گروه دسته بندی شده اند. خطاهایی که از آشکار نشدن نقاط جداسازی ناشی می شوند در گروه اول قرار دارند و خطاهایی که در اثر شکستن حروف در غیراز نقاط جداسازی ایجاد می شوند در گروه دوم قرار دارند. در جدول (1) درصد خطاها با در نظر گرفتن زیر حروف در بدنه های "س" و "ص" محاسبه شده است، بخش (2 مقاله). در ادامه انواع خطاها معرفی و بررسی می شوند.

خطای نوع 1- آشکار نشدن دندان اول بدنه "س"

یکی از شایعترین خطاهای جداسازی، آشکار نشدن دندان اول در سین و شین است. این خطا، از این جا ناشی می شود که در بعضی از قلم ها، مخصوصا در نگارشهای قدیمی، دندان اول حروف سین و شین کمی بالاتر از سایر دندانهای آنها نوشته می شود و کمی بالاتر از نوار زمینه قرار می گیرد. بنابراین نقطه جداسازی آن آشکار نمی شود، شکل (8).

(12) s: CL AND PU AND NDEL

(13) t: CL AND PD AND NU

که در آنها

به NU,PD,NDEL,NDNEL,CL,NUELHW,NUNE,PU

ترتیب زیر تعریف می شوند.

PU: پاره مسیر قبلی برچسب بالا دارد.

NUNE: پاره مسیر بعدی برچسب بالا دارد و آخرین پاره مسیر کانتور بالایی نیست.

NUELHW: پاره مسیر بعدی برچسب بالا دارد، آخرین پاره مسیر کانتور بالایی است و طول آن از 2pt بیشتر است. این پاره مسیر یکی از دو شرط زیر را نیز دارد.

1- ارتفاع آن، اختلاف بین عرض دو نقطه پاره مسیر که دارای بیشترین و کمترین عرض از مبدأ هستند، از 2pt بیشتر است.

2- پهنای آن، اختلاف بین طول دو نقطه پاره مسیر که دارای بیشترین و کمترین طول از مبدأ هستند، از 2/5pt بیشتر است.

CL: طول پاره مسیر فعلی از 2pt بیشتر است.

NDNEL: پاره مسیر بعدی برچسب پایین دارد، آخرین پاره مسیر کانتور بالایی نیست و طول آن از 3pt بیشتر است.

NDEL: پاره مسیر بعدی برچسب پایین دارد، آخرین پاره مسیر کانتور بالایی است و طول آن از 4pt بیشتر است.

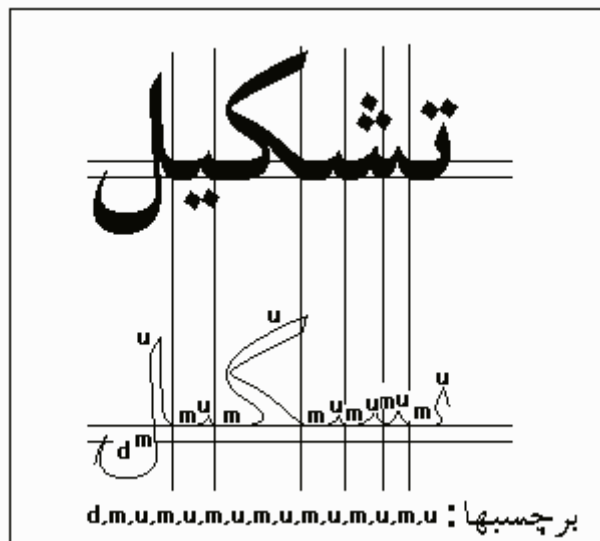
PD: پاره مسیر قبلی برچسب پایین دارد.

NU: پاره مسیر بعدی برچسب بالا دارد.

شرط NUELHW از بند q از شکستن شدن حروفی چون

"ت، پ، ب، ث، د، ذ" در آخر زیر کلمه جلوگیری می کند. شکل (6) پاره مسیرها و نقاط جداسازی یک کلمه را نشان می دهد. نمونه ای از کلمات جدا شده با استفاده از قواعد بالا در شکل (7) آمده است.

برای جلوگیری از شکسته شدن حروف "ی، ن، ل" در هنگامی که به تنهایی یک زیر کلمه را تشکیل می دهند، از چند قاعده ساده استفاده می شود.



شکل ۶- پاره مسیره‌های بر چسب خورده و نقاط جداسازی بالا (u) ، وسط (m) و پایین (d)

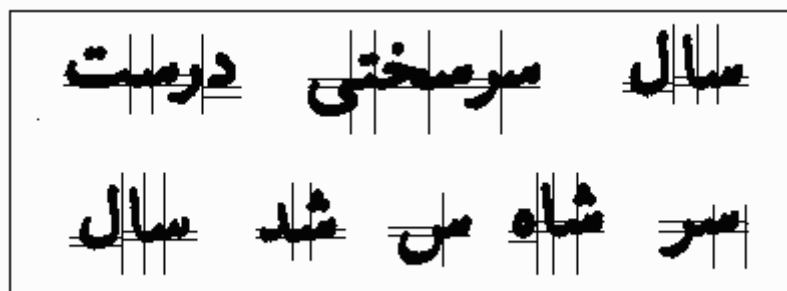
جدول ۱ - مشخصات مجموعه‌های تمرین و آزمایش و تعداد خطاها بر حسب نوع و گروه آنها

خطای گروه اول: آشکارسازی نقاط جداسازی، خطای گروه دوم: شکستن حروف در غیر از نقاط جداسازی

متن با کیفیت چاپ مناسب	مجموعه تمرین	مجموعه آزمایش	مجموعه آزمایش به تفکیک						
			۶	۵	۴	۳	۲	۱	
۱۰۸۱	۲۴۲۸	۶۴۱۵	۱۶۱۳	۱۳۷۸	۷۱۸	۵۴۴	۶۷۵	۱۴۸۷	تعداد کل حروف
۸۵۶	۱۸۸۳	۴۷۹۰	۱۲۵۷	۹۷۹	۵۲۵	۴۴۳	۵۱۲	۱۰۷۴	تعداد حروف به هم چسبیده
۲۲۵	۵۴۵	۱۶۲۵	۳۵۶	۳۹۸	۱۹۳	۱۰۱	۱۶۳	۴۱۳	حروف مجزا
۸	۲۲	۳۱	۸	۳	۳	۳	۳	۱۱	تعداد ص و ض
۵۹	۱۱۳	۳۱۹	۷۰	۷۷	۴۳	۳۶	۲۴	۶۹	تعداد س و ش
۹۸۲	۲۱۳۱	۵۴۴۹	۱۴۰۵	۱۱۲۶	۶۱۴	۵۱۸	۵۶۳	۱۲۲۳	تعداد نقاط جداسازی
-	۱۴	۴۴	۴	۵	۴	-	۷	۲۴	خطای نوع ۱
-	-	۴	-	۱	-	-	-	۳	خطای نوع ۲
-	-	۱۱	-	-	۵	-	-	۶	خطای نوع ۳
۱	۶	۱۴	۹	۲	-	۲	۱	-	خطای نوع ۴
۴	۱۴	۲۹	۱۳	۶	۵	۲	۱	۲	خطای نوع ۵
-	۶	۶۳	۲۹	۴	۷	۱۳	۴	۶	خطای نوع ۶
۱	۲۲	۱۳۳	۴۰	۱۱	۱۷	۱۴	۱۲	۳۶	خطای گروه اول
۴	۱۸	۳۲	۱۵	۶	۵	۳	۱	۲	خطای گروه دوم
۰/۱	۱/۰۳	۲/۴۲	۲/۸۴	۰/۹۷	۲/۷۶	۲/۷	۲/۱۳	۳/۱۸	درصد خطای گروه اول
۰/۴۱	۰/۸۴	۰/۵۸	۱/۰۶	۰/۵۳	۰/۸۱	۰/۵۸	۰/۱۷	۰/۱۶	درصد خطای گروه دوم



شکل ۷- نمونه هایی از جداسازی حروف با استفاده از الگوریتم ارائه شده



شکل ۸- نمونه هایی از خطاهای نوع اول

خطای نوع ۲- چسبیدگی ناخواسته حروف به یکدیگر

بعضی اوقات حروف علاوه بر آنکه روی نوار زمینه به یکدیگر می چسبند، در اثر پراکندگی جوهر در بالای نوار زمینه نیز با یکدیگر تماس پیدا می کنند. از آنجایی که نقطه جداسازی همواره در پایان یک پاره مسیر با بر چسب وسط قرار دارد، در مواردی که چسبیدگی حروف، بالای نوار زمینه اتفاق می افتد، این قسمت از کانتور بر چسب بالا خواهد گرفت و نقطه جداسازی را می پوشاند، شکل (۹). البته این نوع خطا در متون با کیفیت پایین اتفاق می افتد.

خطای نوع ۳- اشتباه در تشخیص نوار زمینه

بعضی از نقاط جداسازی بر اثر اشتباه در به دست آوردن نوار زمینه، آشکار نمی شوند. این نوع خطا، ممکن است در اثر عوامل مختلفی بروز کند. بعضی از این خطاها در زیرکلمات همپوشان روی می دهد. بدین ترتیب که زیر کلمات همپوشان، نسبت به یکدیگر جابه جایی کرسی دارند ولی در الگوریتم ارائه شده، برای این زیر کلمات، تنها یک نوار زمینه به دست می آید. این موضوع باعث می شود که بعضی از نقاط جداسازی آشکار نشوند. به نظر می آید که راه گریز از این خطا، به دست آوردن

خطای نوع ۶- پایین بودن کیفیت چاپ

بعضی از خطاها در اثر کیفیت خیلی پایین چاپ ایجاد می‌شوند. در مجموعه آزمایش ما، این نوع خطا بیشترین سهم را به خود اختصاص داده است. در این موارد کیفیت متن به گونه‌ای است که اجتناب از خطا مشکل به نظر می‌رسد. خطاهایی که در این دسته قرار دارند، ناشی از عوامل مختلفی‌اند که به بررسی آنها می‌پردازیم، شکل (۱۳).

الف) دندان‌های کوتاه و نرم: از مهمترین عوامل تولید خطا در این دسته، دندان‌های کوتاه و نرم‌اند. الگوریتم جداسازی حروف این توانایی را دارد که دندان‌های کوتاه را آشکار کند. اما دندان‌ها در بعضی از متون به حدی کوتاه‌اند که آشکارسازی آنها بسیار مشکل است. اگر بخواهیم پارامترهای الگوریتم را به نحوی تنظیم کنیم که این دندان‌ها را نیز تشخیص دهد، آن‌گاه هر نویز پیش آمده روی نوار زمینه را به عنوان دندان آشکار می‌کند و نرخ تولید خطا افزایش می‌یابد. دندان‌های نرم نیز از عوامل تولید خطا هستند. این دندان‌ها از دندان‌های کوتاه، کمی بلندترند، اما به حدی آرام بالا می‌روند که الگوریتم را برای تبدیل برچسب کانتور بالایی از وسط به بالا تحریک نمی‌کنند. بنابراین بعضی از این دندان‌ها نیز آشکار نمی‌شوند، شکل (۱۳-الف).

ب) نویزهای موجود در لبه بالایی نوار زمینه: از آنجا که متون با دندان‌های کوتاه فراوانند، برای آشکارسازی این دندان‌ها، پارامترهای الگوریتم را تا حد امکان حساس کرده‌ایم. این حساسیت باعث شده است که بعضی از نویزهای شدید در لبه بالایی نوار زمینه به عنوان دندان آشکار شوند. به نظر می‌رسد این خطاها نیز در قسمت پس پردازش قابل تصحیح‌اند، شکل (۱۳-ب).

ج) قرار نگرفتن حروف زیر کلمه در یک راستا: در بعضی از متون چاپی حروف یک زیر کلمه در یک راستای افقی قرار نمی‌گیرند. بعضی از حروف بالاتر و بعضی دیگر پایینتر نوشته می‌شوند. از آنجایی که الگوریتم ما برای کل زیر کلمه یک نوار



شکل ۹- نمونه‌هایی از خطاهای نوع دوم

نوار زمینه محلی برای هر زیر کلمه به صورت مجزا باشد. اما محدود کردن نوار زمینه محلی به زیر کلمات، می‌تواند خطاهای زیادتری در تولید نقاط جداسازی اضافی، ایجاد کند، شکل (۱۰)، کلمات "جوش" و "سرمست".

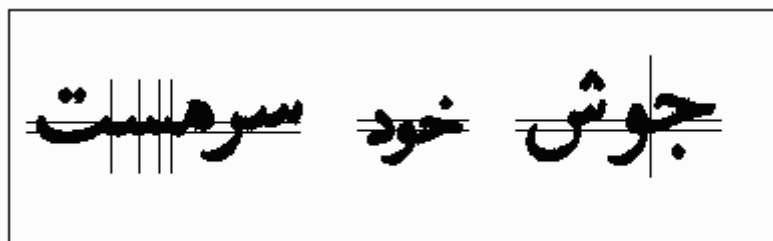
یکی دیگر از عوامل خطا در تعیین نوار زمینه این است که نقاط جداسازی حروف در بعضی از زیر کلمات روی نوار زمینه‌ای که بر اساس ماکزیمم هیستوگرامهای h_{20} و h_{24} به دست می‌آید، قرار نمی‌گیرند (بخش ۴-۳)، (شکل ۱۰، کلمه "خود").

خطای نوع ۴- ضعف قواعد جداسازی

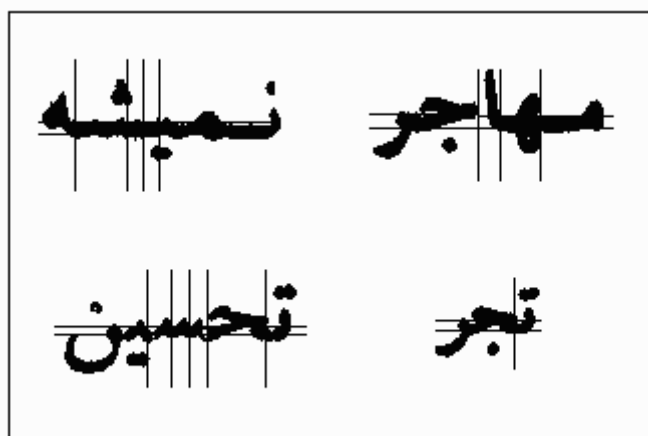
بعضی از نقاط جداسازی بر اثر ضعف الگوریتم جداسازی آشکار نمی‌شوند. این خطاها در اثر کم یا زیاد بودن بعضی از آستانه‌های به کار رفته، ایجاد می‌شوند. اما باید به این نکته توجه داشت که حدود آستانه با استفاده از یک مجموعه متنوع و روش سعی و خطا انتخاب شده‌اند و تا حد امکان، سعی شده است که کمترین خطا به وجود آید. برطرف کردن خطاهایی که از این دسته هستند، برای یک قلم خاص با تنظیم حدود آستانه برای آن قلم ممکن است، اما باعث می‌شود که خطاهای دیگری در جداسازی حروف قلمهای دیگر ایجاد شود، شکل (۱۱).

خطاهای نوع ۵- شکستن بعضی از حروف به زیر حروف

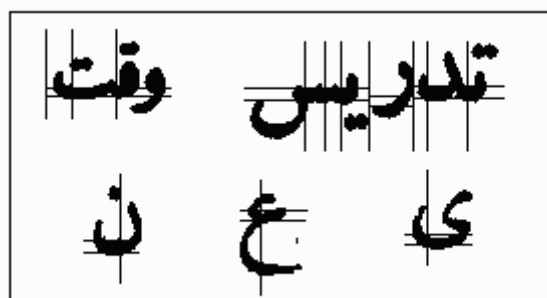
یکی دیگر از انواع خطاها که در جداسازی حروف اتفاق می‌افتد. شکستن ناخواسته حروف "ی، ن، ل، د، ذ، ب، ت، پ، ث، ع، غ" به زیر حروف است، شکل (۱۲). خوشبختانه این نوع خطاها، به راحتی در مرحله پس پردازش قابل شناسایی و تصحیح هستند [۳، ۱۲].



شکل ۱۰- نمونه‌هایی از خطاهای نوع سوم



شکل ۱۱- نمونه‌هایی از خطاهای نوع چهارم



شکل ۱۲- نمونه‌هایی از خطاهای نوع پنجم

حداکثر هیستوگرامهای h_{20} و h_{24} در این نوار اتفاق می افتد، بخش (۴-۴). اما در بعضی از نگارشها، حروف به حدی زاویه دارند که تعیین یک نوار زمینه افقی برای آنها ممکن نیست. در نهایت در این قبیل کلمات نوار زمینه در هر جا که قرار داده شود، چند نقطه جداسازی آشکار نمی شود. برای حل مشکل این قبیل خطاها باید یک نوار زمینه شیبدار تعریف کرد. در

زمینه پیدا می کند، بعضی از نقاط جداسازی آشکار نمی شوند، شکل (۱۳-ج) (د) شیبدار نوشته شدن حروف در یک زیر کلمه: یکی دیگر از عوامل تولید خطا در آشکار سازی نقاط جداسازی، شیبدار نوشته شدن حروف یک زیر کلمه است، شکل (۱۳-د). الگوریتم ارائه شده، نوار زمینه را با این تعریف تعیین می کند که



(ب)



و خلا

شت

و خر

(د)

ب- مثالهایی از خطاهای نوع ۶- ب

د- نمونه‌هایی از خطاهای نوع ۶- د

تشکیل می‌دهند، به دلیل ضعف قواعد جداسازی و نحوه تعیین نوار زمینه به وجود آمده‌اند، جدول (۲).

۸- اثر افزایش درجه تفکیک

برای تعیین اثر درجه تفکیک، از متن شماره ۶ با درجه تفکیک ۴۰۰ dpi نیز تصویر برداری شد و الگوریتم جداسازی روی تصویر جدید اعمال شد. در مقایسه با نتایج به دست آمده برای ۳۰۰ dpi (ستون مربوط به متن شماره ۶ در جدول ۱)، تعداد خطاهای گروه اول از ۴۰ به ۲۳ و تعداد خطاهای گروه دوم از ۱۵ به ۱۰ کاهش یافت. به عبارت دیگر حدود نیمی از خطاهای مربوط به آشکار نشدن نقاط جداسازی برطرف شدند.



(الف)



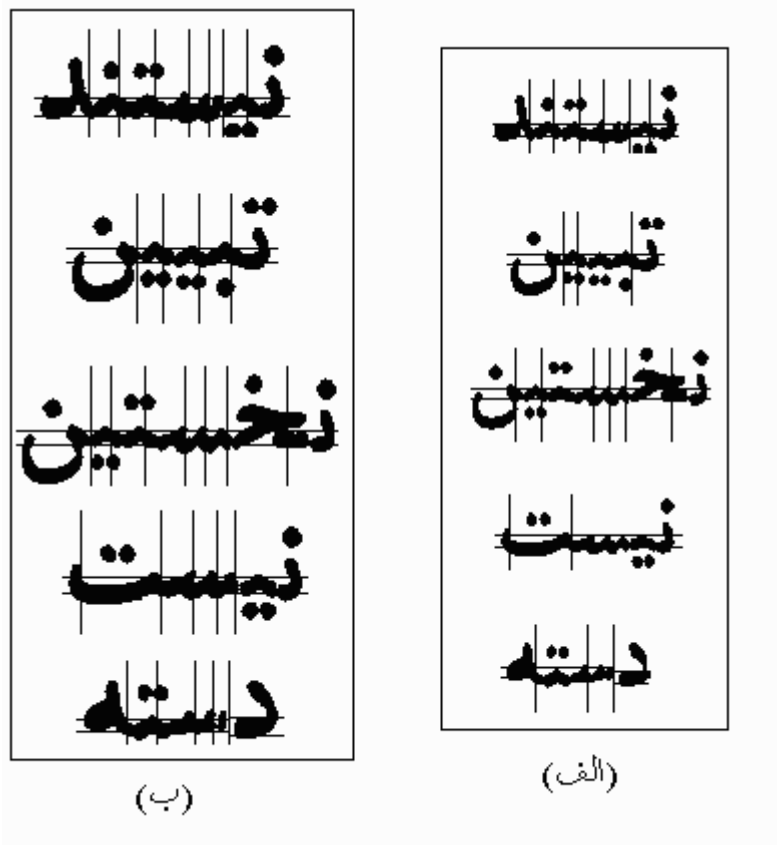
(ج)

شکل ۱۳- الف- مثالهایی از خطاهای نوع ۶- الف

ج- نمونه‌هایی از خطاهای نوع ۶- ج

ضمن قابل ذکر است که سیستم موجود نسبت به کجی خط ناشی از تصویر برداری حساس است. چنانچه اختلاف بین بالاترین و پایینترین نوار زمینه محلی در یک خط متن کمتر از دو برابر پهنای قلم باشد، الگوریتم می‌تواند آن را پردازش کند. معمولاً برای کجی بیشتر، از روشهای رفع کجی متن استفاده می‌شود.

خطاهای نوع دوم و ششم که مجموعاً ۴۰/۶۲٪ از کل خطا را تشکیل می‌دهند، در اثر کیفیت خیلی پایین چاپ پدیدار شده‌اند. خطاهای نوع پنجم که ۱۷/۵۷٪ از کل خطا را تشکیل می‌دهند با پس پردازشهای مناسب قابل تصحیح‌اند. خطاهای نوع اول، سوم و چهارم که مجموعاً ۴۱/۸۱٪ از کل خطا را



شکل ۱۴- الف) برخی از خطاهای جداسازی در ۳۰۰dpi
ب) تصحیح شدن بعضی از این خطاها در ۴۰۰dpi

جدول ۲- درصد خطاهای ۶ گانه نسبت به خطای کل در مجموعه آزمایش (متون ۶تا)

نوع خطا	۱	۲	۳	۴	۵	۶
درصدخطا به کل	۲۶/۶۶	۲/۴۲	۶/۶۷	۸/۴۸	۱۷/۵۷	۳۸/۲۰

ارائه شده است. در این قسمت سطوح آستانه در تبدیل بین برچسبهای مختلف تغییر داده شده اند و با اضافه کردن مفهوم شیب کانتور، جداسازی دندانهای کوچک و نرم با دقت بیشتری صورت گرفته است. در بخش ششم، قوانین جدا سازی برای کاربرد در متون با کیفیت چاپی پایین تکمیل شده‌اند. در بخش هفتم علل بروز انواع مختلف خطا به طور جامع بررسی شده است.

میزان جداسازی درست در مرجع [۳] قبل از پس پردازش ۹۱٪ و بعد از آن ۹۸/۵٪ است. برای الگوریتم اصلاح شده،

در حالی که حدود یک سوم از خطاهای مربوط به شکستن ناخواسته حروف تصحیح شدند، شکل (۱۴).

۹- نتیجه گیری

هدف از این تحقیق اصلاح الگوریتم جداسازی حروف در مراجع [۳، ۱۵] برای متون با کیفیت چاپی پایین بود. در بخش چهارم، مراحل مختلف اصلاح الگوریتم آشکارسازی نوار زمینه محلی مطرح شده است. در بخش پنجم مقاله تغییرات انجام شده در برچسب‌زنی نقاط کانتور بالایی نسبت به الگوریتم [۳]

متفاوت است. همچنین باید توجه داشت که در این تحقیق سعی شده است با جامعتر کردن قواعد جداسازی حتی الامکان از پس پردازشهای مطرح شده در مرجع [۳] اجتناب شود.

میزان جداسازی درست بدون پس پردازش ۹۷٪ است. البته لازم به ذکر است که مجموعه های آزمایش در این دو تحقیق کاملاً یکسان نیستند و تعریف خطا نیز با توجه به اینکه در این تحقیق جداسازی دندانهای "س" و "ص" مورد نظر بوده است، اندکی

واژه‌نامه

1. binary dicision tree
2. upper base line(ubl)
3. lower base line(lbl)

4. upper
5. middle
6. down

مراجع

1. Lu, Y., "Machine Printed Character Segmentation-an Overview," *Pattern Recognition*, vol. 28, no. 1, pp. 67-80, 1995.
2. Lu, Y. and M. Shridhar, "Character Segmentation in Handwritten Words – an Overview," *pattern Recognition*, vol. 29, no. 1, pp. 77-96, 1996.
3. عزمی، ر. و کبیر، ا. "معرفی روش جدیدی برای جداسازی حروف در متون چاپی بدون توجه به نوع قلم،" مجله استقلال، سال ۱۸، شماره ۲، صفحات ۱۰-۱، اسفند ۱۳۷۸.
4. احمد زاده، م. و کبیر، ا. "شکستن کلمات تایپ شده فارسی به حروف،" گزارش اولین کنفرانس بین المللی کامپیوتر در علوم، فنون و پزشکی ایران، صفحات ۶-۱، ۷-۵ دیماه ۱۳۷۰، اصفهان.
5. Parhami, B., and Taraghi, M., "Automatic Recognition of Printed Farsi Text," *Pattern Recognition*, vol. 14, pp. 395 – 403, 1981.
6. EL-Sheikh, T.S., and Guindi, R.M. "Computer Recognition of Arabic Cursive Scripts," *Pattern Recognition*, vol. 24, no. 4, pp. 293 – 302, 1988.
- 7- شیرعلی شهررضا، م. ح، "تشخیص کلمات و ارقام دستویس فارسی به وسیله شبکه های عصبی،" رساله دکتری، دانشکده برق، دانشگاه صنعتی امیر کبیر، ۱۳۷۴.
8. Amin, A., and Masini, "Machine Recognition of Multi Font Printed Arabic Texts," *Proc. 8th Int. Conf. on Patt. Recog.* Paris, pp. 392-395, 1986.
9. Amin, A., and Mari, J.F., "Machine Recognition and Correction of Printed Arabic Text," *IEEE Trans. Systems, Man and Cybernetic*, vol. 9, pp 1300 – 1306, 1989.
10. Amin, A., and Al- Fedaghi, S., "Machine Recognition of Printed Arabic Text Utilizing a Natural Language Morphology," *International Journal of Man - Machine Studiese*, vol. 35, pp. 759 – 768, 1991.
۱۱. یکتایی، م.، زحیح، ا. و منارد، م. "جداسازی کلمات فارسی به حروف وزیر-حروف،" دومین کنفرانس سالانه انجمن کامپیوتر ایران، دانشگاه صنعتی امیر کبیر، صفحات ۲۱۱-۲۱۶، ۴ الی ۶ دیماه ۱۳۷۵.
۱۲. رفیعی، ش. و کبیر، ا. "شکستن کلمات تایپ شده فارسی به حروف در رسم الخطهای مختلف،" مجموعه مقالات سومین کنفرانس الکترونیک، دانشگاه شیراز، صفحات ۱۰۴-۹۸، مهرماه ۱۳۷۴.
13. Kurdy, B. M., and Joukhadr, A., "Multifont Recognition System for Arabic Characters," *proc. 3rd Int. Conf. Exhibition of Multi-Lingual Computing (Arabic and Roman script)*, U.K, pp. 731 – 739, 1992.
14. Amin, A., and Al-Sadoun, H.B., "A New Segmentation Technique of Arabic Texts," *Proc. 11th IAPR Int. Conf. on Patt. Recog.* Vol. II. Conf. B: Pattern Recognition Methodology and system, pp. 441-5, 1992.
15. Azmi, R. and Kabir, E., "A New Segmentation Technique for Omnifont Farsi Text," *Pattern Recognition Letters*, vol. 22, no. 2, pp. 97-104, 2001.
۱۶. نظام آبادی پور، ح.، کبیر، ا. و عزمی، ر.، "جداسازی حروف در متون چاپی قدیمی،" مجموعه مقالات ششمین کنفرانس بین المللی سالانه انجمن کامپیوتر ایران، دانشگاه اصفهان، صفحات ۵۹۰-۵۷۹، اسفندماه ۱۳۷۹.
۱۷. نظام آبادی پور، ح. "پیش پردازش متون چاپی فارسی برای جداسازی حروف،" پایان نامه کارشناسی ارشد، بخش مهندسی برق، دانشگاه تربیت مدرس، ۱۳۷۹.

