

SID



ابزارهای
پژوهش



سرویس ترجمه
تخصصی



کارگاه های
آموزشی



بلاگ
مرکز اطلاعات علمی



سامانه ویراستاری
STES



فیلم های
آموزشی

کارگاه های آموزشی مرکز اطلاعات علمی



آموزش مهارت های کاربردی در تدوین و چاپ مقالات ISI

آموزش مهارت های کاربردی
در تدوین و چاپ مقالات ISI



روش تحقیق کمی

روش تحقیق کمی



آموزش نرم افزار Word برای پژوهشگران

آموزش نرم افزار Word
برای پژوهشگران

استخراج خطوط در اسناد دستنویس فارسی مبتنی بر خوشه-

بندی سلسله مراتبی

مجید ایرانپور مبارکه^۱، علیرضا احمدی فرد^۲

^۱دانشجوی دکتری دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه صنعتی شاهرود، majid.iranpor@shahroodut.ac.ir

^۲دانشیار دانشکده مهندسی برق و رباتیک، دانشگاه صنعتی شاهرود، ahmadyfard@shahroodut.ac.ir

چکیده

استخراج خط از اسناد دستنویس یکی از مهم‌ترین مراحل پیش‌پردازش در آنالیز اسناد مانند درک اسناد تصویری، شناسایی متون دستنویس یا چاپی و جستجوی کلمه در اسناد تصویری (Word Spotting) است. تنوع در فاصله بین خطوط، فاصله بین کلمات یک خط و شیب خط و همچنین اتصال و همپوشانی بین خطوط باعث شده که این مسئله یک چالش بزرگ باقی بماند. این مشکل در زبانهایی با رسم‌الخط بهم‌چسبیده، مانند فارسی و عربی، بخاطر وجود فاصله بین زیرکلمات و همچنین تنوع در تعداد و محل نقاط و وجود سرکش بسیار پیچیده‌تر می‌باشد. در این مقاله یک رهیافت جدید برای استخراج و قطعه‌بندی خطوط در متن دستنویس فارسی ارائه شده است. یک روش خوشه‌بندی سلسله‌مراتبی (Hierarchical Clustering) براساس نزدیکترین فاصله (Single-Linkage) با یک معیار فاصله جدید که ساختار نگارش فارسی را در نظر می‌گیرد برای خوشه‌بندی اجزاء متصل (Connected Component (CC)) مورد استفاده قرار گرفته است. سپس یک سری قواعد براساس شیب خط و ساختار زبان فارسی جهت اتصال و جداسازی خوشه‌های بدست آمده اعمال شده است. پارامترهای مورد استفاده براساس سند بصورت وفقی تعیین می‌گردند. تست این روش روی دو مجموعه داده استاندارد نتایج قابل قبولی را نشان می‌دهد.

واژه‌های کلیدی

آنالیز اسناد تصویری، استخراج خط، اسناد دستنویس فارسی، خوشه‌بندی سلسله‌مراتبی.

۱- مقدمه

متصل از دو خط متوالی است؛ که با توجه به وجود نقاط و سرکش‌ها در متون فارسی و عربی این مسئله بیشتر مشکل ایجاد می‌نماید. استخراج صحیح یا غلط خطوط تاثیر مستقیم زیادی بر دقت قطعه‌بندی کلمات و حروف دارد.

روش‌های متفاوتی برای استخراج و قطعه‌بندی خطوط از اسناد دستنویس ارائه شده است [۱]. روش‌های موجود در استخراج خط در مقالات به روش‌های مختلفی دسته‌بندی شده است که بصورت زیر هستند: روش‌های مبتنی بر افکنش^۳، روش‌های مبتنی بر لکه‌دار کردن^۴،

استخراج یا قطعه‌بندی خط یکی از مراحل مهم پیش‌پردازش در بسیاری از کاربردهای آنالیز اسناد تصویری مانند شناسایی متون چاپی و دستنویس برون خط^۱، جستجوی کلمات^۲، بازیابی کلمات تصویری^۳ و شناسایی کاراکتر نوری^۴ است. با توجه به اینکه در اسناد دستنویس فاصله خطوط و کلمات در یک سند متغیر است و همچنین معمولاً راستای مستقیم برای خط مبنا رعایت نشده و خطوط یک سند دارای زوایای متفاوتی نسبت به خط مبنا هستند، استخراج خطوط در اسناد دستنویس بسیار چالش برانگیزتر است. یکی از مشکلات بزرگ دیگر در این باب امکان به هم چسبیدگی اجزاء

^۴ Optical Character Recognition (OCR)

^۵ Projection-based Methods

^۶ Smearing Methods

^۱ offline

^۲ Word Spotting

^۳ Word Image Retrieval

روش‌های مبتنی بر تبدیل هاف^۷، روش مبتنی بر نازک‌سازی^۸ و روش‌های احتمالی^۹.

روش‌های مبتنی بر افکنش معمولاً در مورد اسناد تصویری چاپی مورد استفاده قرار می‌گیرد و برای اسناد دستنویسی که خطوط همپوشانی ندارند نیز توسعه یافته است [۲، ۳]. اما این روش بعثت مشکلات فراوان، مانند در نظر نگرفتن شیب خط، در متون اسناد دستنویس تصویری، نمی‌تواند روی اسناد مستقیماً اعمال شود. به همین دلیل جهت بهبود این روش راهکارهایی مانند اعمال روش‌های مبتنی بر افکنش به برش‌های عمودی در سند و سپس ترکیب نتایج، توسعه داده شده است [۴-۷]. در این روش سند تصویری به صورت موازی به نوارهای عمودی تقسیم‌بندی می‌شود. سپس مرزهای بالقوه جداکننده خطوط در هر بخش بدست می‌آید و در مرحله بعد این بخش‌های کاندید با استفاده از هیوربستیک یا مدل‌های گرافیکی آماری مانند مدل مخفی مارکوف^{۱۰} به همدیگر متصل شده و نواحی خطوط را استخراج می‌کنند [۶، ۷]. نهایتاً برای فائق آمدن بر مواردی که خطوط همپوشانی دارند از تکنیک‌های تقاطع^{۱۱}، الگوریتم تعقیب کانتور و چگالی گاوسی دو متغیره برای بازنمایی خطوط [۴، ۵] استفاده شده است.

روش مذکور دارای نقایصی است که در زیر به آنها اشاره شده است [۸] و شامل موارد ذیل است. این روش خطوط بالقوه زیادی را تولید می‌کند، پارامتر عرض نوارهای برش باید از پیش تعرف شده باشد، خطوط نباید انحنای زیاد داشته باشند و اگر نوار ابتدایی و انتهایی شامل خطوط بالقوه نباشند، استخراج خط بطور کامل ناممکن است. برخی شیب خط را نیز برای جداسازی خطوط مورد استفاده قرار داده‌اند [۴] که این شیب معمولاً بسادگی قابل محاسبه نیست و در یک سند بین خطوط متفاوت است. بنابراین این متدها نیز در این شرایط نتایج خوبی را ارائه نمی‌کنند.

در روش‌های مبتنی بر لکه‌دار کردن روش‌هایی مانند الگوریتم هموارسازی طول اجرا^{۱۲} [۹] یا الگوریتم آنالیز طول اجرای جهت دار بهبود یافته^{۱۳} [۱۰] بکار گرفته شده است. استفاده از این الگوریتم‌ها باعث خواهد شد که بلوک‌های پیکسلی در جهت افقی لکه‌دار شوند؛ یعنی اگر فاصله سفید (پس زمینه) بین آنها از یک حد آستانه‌ای کمتر بود با پیکسل سیاه (پیش‌زمینه) پر شود. این روش ابتداری تصاویر

خاکستری اعمال شده و سپس تصویر باینری می‌شود و عمل مورفولوژی فرسایش^{۱۴} روی پیش‌زمینه و پس‌زمینه برای بدست آوردن محدوده خطوط انجام شده است.

روش‌های مبتنی بر رهیافت تبدیل هاف برای استخراج خطوط مستقیم در تصاویر بسیار پرکاربرد است. شیب خطوط دستنویس با استفاده از اعمال این نوع تبدیل به مرکز ثقل هر جزء متصل بدست می‌آید. اگر اکثر همسایه‌های نزدیک در یک تراز متعلق به گروه اجزاء تشکیل دهنده هم‌ترازی باشند، این هم‌ترازی ویژگی‌های پیوستگی و نزدیکی را دارا خواهد بود، و این بعنوان یک خط پذیرفته می‌شود. سپس این نتایج حاصل شده با استفاده از حذف هم‌ترازی‌های نادرست با استفاده از هم‌ترازی بین اجزاء متصل با استفاده از اطلاعات محتوایی، تصویر خروجی نهایی را ارائه می‌کند [۱۱]. در کار دیگری روشی را مبتنی بر تبدیل هاف بر روی سند بلوک‌بندی شده ارائه کرده است [۱۲].

برخی محققین برای استخراج خطوط هندی و چینی از روش نازک‌سازی استفاده کرده‌اند [۱۰، ۱۳، ۱۴]، که در [۱۴] ابتدا نازک‌سازی صورت گرفته و در ادامه عملیات پس‌پردازشی روی تمامی نواحی پس‌زمینه برای تشخیص مرز جداکننده خطوط اعمال شده است.

اخیراً برخی از روش‌های مجموعه سطوح^{۱۵}، کانتور فعال^{۱۶} و بیز تغییرات^{۱۷} برای استخراج خطوط بهره برده‌اند [۱۳، ۱۵، ۱۶]. در [۱۵] از روش تخمین چگالی و مجموعه سطوح استفاده شده است. در این کار یک نگاشت احتمالی از سند تصویری ورودی تخمین زده می‌شود که هر المان احتمال تعلق هر پیکسل به یک خط را نشان می‌دهد. روش مجموعه سطوح برای تعیین ارزیابی مرزی با خطوط همسایه مورد استفاده قرار گرفته است. در [۱۶] در ابتدا یک بانک فیلتر تطبیق یافته برای هموارسازی تصویر ورودی مورد استفاده قرار گرفته است. مراکز خطوط اجزاء تشکیل دهنده خط با استفاده از مرزبندی تصویر هموار شده محاسبه می‌شود. در انتها برای بدست آوردن نتیجه نهایی، کانتورهای فعال روی مرزها وفق داده می‌شوند. در کار [۱۳] روش بیز تغییرات برای استخراج خطوط در متون دستنویس چینی مورد بهره برداری قرار گرفته است. یک سند به صورت یک مدل ترکیب گاوسی^{۱۸} در نظر گرفته می‌شود،

13 Improved Directional Run-Length Analysis

14 Erosion

15 Level Set

16 Active Contour

17 Variational Bayes

18 Gaussian Mixture Model

7 Hough Transform

8 Methods based on thinning operation

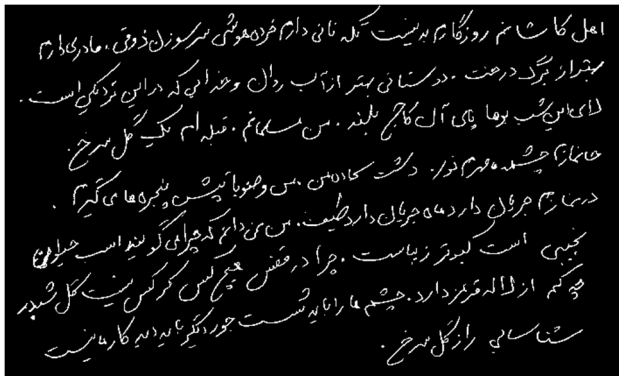
9 Stochastic Methods

10 Hidden Markov Model (HMM)

11 Crossing technique

12 Run-Length Smoothing Algorithm (RLSA)

از خوشه‌بندی سلسله‌مراتبی با استفاده از یک معیار فاصله معرفی شده مرتفع سازد.



شکل ۱: نمونه‌ای از متن دستنویس فارسی موجود در مجموعه داده.

در این مقاله روشی ارائه شده است که از خوشه‌بندی سلسله‌مراتبی پیکسل‌های قلم (سیاه یا پیش‌زمینه) موجود در سند جهت استخراج خطوط استفاده می‌کند. این روش در ادامه مقاله در بخش دوم بطور کامل تشریح شده است. در بخش سوم نیز نتایج آزمایشات روی دو مجموعه استاندارد زبان فارسی گزارش شده و سپس بحث و نتیجه‌گیری در بخش چهارم آورده شده است.

۲- روش پیشنهادی

مقالات در ساختار کلی روش پیشنهادی، که در شکل ۲ نشان داده شده است، شیوه نگارش متن فارسی به صورت دستنویس مد نظر قرار داده شود. از آنجایی که معمولاً دست‌نوشته‌ها ممکن است شامل خطوط نزدیک به همدیگر باشند یا نقاط و سرکش‌ها باعث اشتباه در خوشه‌بندی شوند از یک مرحله پیش‌پردازش استفاده شده است که احتمال خوشه‌بندی اشتباه را در موارد ذکر شده کاهش دهد.

با توجه با اینکه در متون فارسی و عربی شیب خط مبنای استاندارد مستقیم است و در متون دستنویس این شیب به مقدار کمی می‌تواند خط مبنا را به سمت بالا یا پایین متمایل کند یا در آن انحنای کمی را ایجاد نماید معیار فاصله‌ای معرفی شده است که این مسئله را در نظر بگیرد. نتایج نهایی خوشه‌بندی با استفاده از تخمین شیب خط و هیوریستیک‌هایی مبتنی بر شیوه نگارش فارسی اصلاح و تکمیل می‌گردد. در ادامه تمامی مراحل کار به تفصیل آورده شده است.

۲-۱- پیش‌پردازش

طوری که هر جزء متعلق به یک خط باشد. برای تخمین پارامترهای چگالی و تعیین تعداد خطوط چارچوب کاری بیز تغییرات بکار برده شده است.

در [۱۷، ۱۸] یک روش مبتنی بر بلاک‌بندی معرفی شده که پارامترهای آن به صورت وقتی براساس سند ورودی تعیین می‌گردد. در هر بلاک برای تشخیص جهت از فیلتر کردن تصویر با هسته‌های گاوسی دوبعدی ناهمسانگرد^{۱۹} در زوایای بین ۳۰- درجه تا ۳۰+ درجه استفاده شده است. پس از تشخیص جهت خطوط در هر بلاک از هم جدا می‌شوند.

اسناد دستنویس فارسی و عربی دارای تعداد زیاد و متنوعی نقطه، سرکش و نشانه هستند (شکل ۱) و همچنین در این نوع اسناد یک جزء متصل می‌تواند یک کلمه، زیرکلمه، کاراکتر، نقطه یا بخشی از نقاط پیوسته یا یک سرکش باشد. با توجه به دلایل مذکور مسئله استخراج خطوط دستنویس برای اکثر روش‌ها چالش‌برانگیزتر خواهد بود و برخی از روش‌ها مانند تبدیل هاف، لکه‌دار کردن طول اجرای جهت دار بهبود یافته و کانتور فعال در مواجهه با این نوع اسناد شکست خواهند خورد. علت این شکست‌ها را می‌توان به صورت زیر تشریح کرد: ۱- روش‌های مبتنی بر تبدیل هاف اهمیت بدنه اصلی کلمه و نقاط و سرکش‌های کوچک را در فضای هاف یکسان در نظر می‌گیرد. ۲- در روش‌هایی که از آنالیز طول اجرای استفاده می‌شود ممکن است اجزاء کوچک برخی خطوط مانند نقطه و سرکش به خط دیگری متصل شوند. ۳- در برخی روش‌ها با استفاده از فیلتری نقاط و سرکش‌ها حذف می‌شوند در صورتی که این اجزاء در زبان فارسی و عربی نقش تعیین‌کننده‌ای دارند.

در [۸] روشی مبتنی بر تکنیک نقاشی^{۲۰} ارائه شده است. در این روش یک سند به نوارهای عمودی تقسیم‌بندی می‌شود و با استفاده از تکنیک‌های نقاشی اجزاء نزدیک به همدیگر در هر نوار عمودی به هم متصل می‌شوند که در اینصورت ادعا شده است که شیب خط بدست می‌آید. سپس از عمل مورفولوژی انبساط^{۲۱} جهت برقراری اتصال بین لکه‌های ایجاد شده در مرحله قبل و تشکیل خط استفاده می‌شود. سپس روی پس‌زمینه نتیجه مرحله قبل عمل نازک‌سازی انجام می‌گیرد سپس با استفاده از برخی قواعد هیوریستیک خطوط نهایی استخراج می‌شوند. اما این روش با توجه به اینکه از روش‌های قبلی بسیار بهتر نسبت به فارسی و عربی کار میکند اما باز هم در مورد این دو زبان درصد خطاهایی بسیار بیشتری را گزارش می‌کند. اما در روش ارائه شده در این مقاله سعی شده تا این مشکلات را با استفاده

²¹ Dilation

¹⁹ Anisotropic 2D Gaussian Kernels

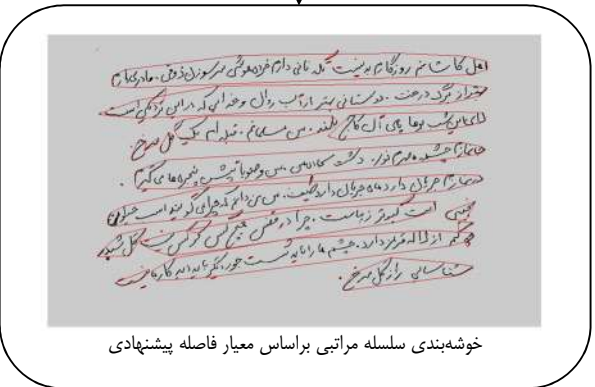
²⁰ Painting Technique

فاصله موثر بین خطوط است. از نصف این فاصله بعنوان پارامتر معیار توقف خوشه‌بندی استفاده شده است.

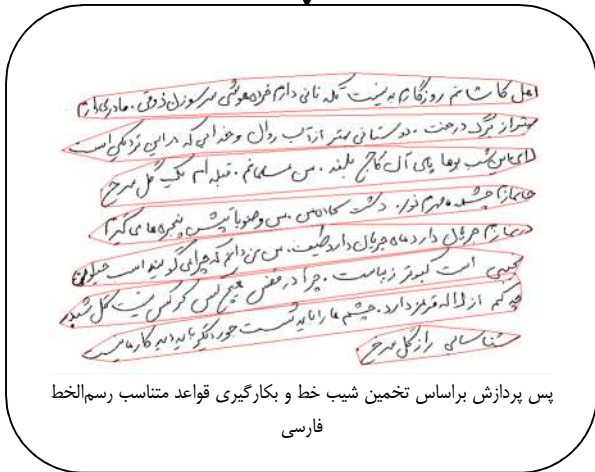


سند دستنویس فارسی

- پیش پردازش
- تبدیل به دودویی
- نازک سازی
- محاسبه پارامترهای الگوریتم برای



خوشه‌بندی سلسله مراتبی براساس معیار فاصله پیشنهادی



پس پردازش براساس تخمین شیب خط و بکارگیری قواعد متناسب رسم الخط فارسی

شکل ۲: شمای کلی روش پیشنهادی.

مرحله پیش‌پردازش یکی از مهم‌ترین مراحل در کارهای شناسایی الگو می‌باشد. در کارهای پردازش اسناد تصویری یکی از مراحل اولیه پیش‌پردازش تبدیل سند تصویری به یک تصویر دودویی است. بعنوان اولین مرحله پیش‌پردازش در این کار با استفاده از الگوریتم Otsu [۱۹] سند ورودی به تصویر دودویی تبدیل می‌گردد و سپس روی تصویر دودویی به‌دست آمده دو کار انجام می‌گیرد: نازک‌سازی و استخراج پارامترهای مورد نیاز برای سند ورودی بصورت وقتی.

• نازک‌سازی

با توجه با اینکه در متون دستنویس فارسی بالارونده‌هایی مانند حروف الف و لام و پایین‌روندهایی مانند ر، ز و واو امکان دارد با همدیگر یا با سرکش‌ها در یک فاصله افقی کم از هم قرار گیرند با استفاده از نازک‌سازی امکان این برخورد را کاهش می‌دهیم. عمل نازک‌سازی بر پایه مورفولوژی انجام شده است. اما در مرحله پایانی خوشه‌بندی اجزاء متصل اصلی در هر خوشه جایگزین می‌شوند.

• تعیین وقتی پارامترهای مورد نیاز

با توجه به اینکه در اسناد ورودی تعداد خطوط، اندازه تصویر و نوع دستخط از نظر اندازه متفاوت است نیاز به تعیین برخی پارامترها برای هر سند وجود دارد. در ادامه این پارامترها و رابطه بدست آمده جهت استخراج آنها را تشریح می‌کنیم.

یکی از مهمترین پارامترهایی که طی عملیات خوشه‌بندی سلسله مراتبی به آن نیاز داریم، حد آستانه‌ای جهت توقف ترکیب خوشه‌ها و استخراج خوشه‌بندی نهایی است. برای شرط توقف اتصال خوشه‌ها از معیار حداکثر فاصله استفاده می‌شود و میزان این حد آستانه از پارامتر فاصله موثر بین دو خط مجاور بدست می‌آید. برای بدست آوردن این پارامتر تصویر به ۲۰ نوار عمودی با عرض برابر تقسیم می‌شود و سپس متوسط فاصله خطوط مجاور برای ۱۴ نوار میانی اندازه گرفته می‌شود. بدین صورت که برای هر کدام از نوارهای مذکور پروفایل افکنش افقی محاسبه می‌گردد و فاصله بین قله‌ها محاسبه می‌گردد شکل ۳ تصویری از این بلوک‌بندی و قله‌های مذکور در یکی از نوارها را نمایش می‌دهد. اگر این فاصله برای تمامی نوارهای مذکور محاسبه شود، فاصله موثر بین دو خط مجاور از رابطه زیر بدست می‌آید:

$$Line_De = \frac{\sum_{i \in Corr_D} i * Hist(i)}{\sum_{i \in Corr_D} Hist(i)} \quad (1)$$

$$Corr_D = \{j | Hist(j) > 0.25 * \max(Hist)\} \quad (2)$$

که در این روابط Corr_D شامل اندیس فاصله‌هایی می‌شود که تعداد آنها حداقل از ۲۵ درصد حداکثر بیشتر باشد. Hist هیستوگرام فاصله بین خطوط بدست آمده را نشان می‌دهد. Line_De نشان دهنده

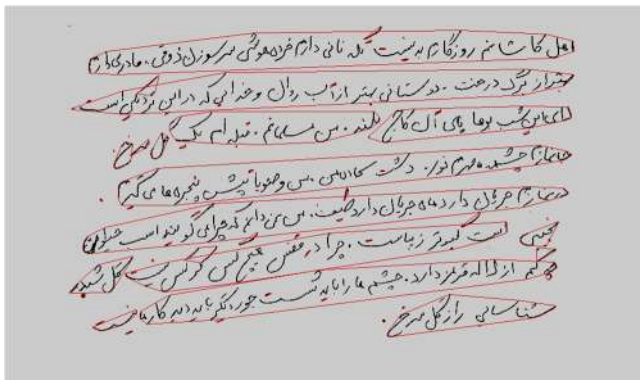
انجام شود هر جزء متصل بعنوان یک خوشه اولیه در نظر گرفته شده و عمل خوشه‌بندی انجام می‌گیرد. در این روش با توجه به ساختار نوشتاری زبان فارسی که نوشتار بصورت افقی نگارش می‌شود فاصله طبق رابطه (۴) طوری تعریف شده است که فاصله در جهت عمودی شامل جریمه باشد.

$$D_{p_1, p_2} = |X_{p_2} - X_{p_1}| + |Y_{p_2} - Y_{p_1}|^5 \quad (4)$$

که در آن p_1 و p_2 دو نمونه از پیکسل‌های مشکی هستند که فاصله بین آنها محاسبه می‌شود و X_p و Y_p به ترتیب مختصات افقی و عمودی هر کدام از دو نقطه مذکور می‌باشند. با استفاده از این معیار فاصله، اجزاء متصلی (خوشه‌هایی) که از نظر افقی در کنار هم قرار گرفته‌اند در یک خوشه قرار خواهند گرفت. با توجه به اینکه در زبان فارسی شبه‌کلمات تشکیل دهنده یک خط در امتداد محور عمودی کشیدگی دارند احتمال قرار گرفتن اجزاء متصل کنار هم حتی در شرایطی که خط بصورت شیب‌دار یا منحنی نوشته شده باشد در کنار هم بسیار زیاد است و نتایج بدست آمده که در بخش سوم به تفصیل آمده است این ادعا را ثابت می‌کند.

اما یکی از مسائل مهم در خوشه‌بندی سلسله‌مراتبی معیار توقف ادغام خوشه‌هاست. برای این کار معیار توقف حداکثر فاصله بین دو خوشه در نظر گرفته شده و برای هر سند به صورت وفقی تعیین می‌گردد که در بخش قبل بطور کامل آورده شده است.

نتیجه این مرحله با شرایط ذکر شده مجموعه‌ای از خوشه‌هاست که معمولاً بیشتر از تعداد خطوط است اما خوشه‌های بزرگی که نمایش‌دهنده و معرف خط اصلی هستند در آن وجود دارد. شکل زیر نمونه‌ای از یک سند تصویری را پس از این مرحله نمایش می‌دهد.

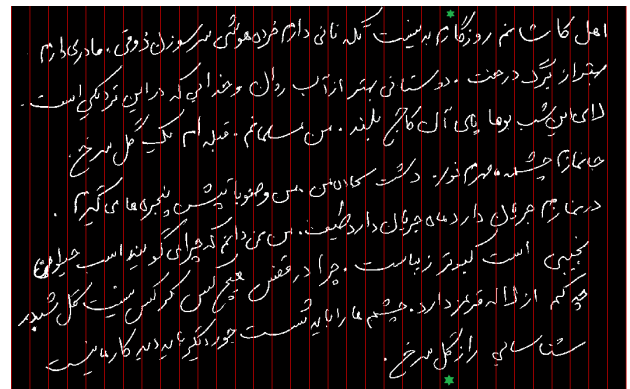


شکل ۴: یک سند نمونه خروجی مرحله خوشه‌بندی.

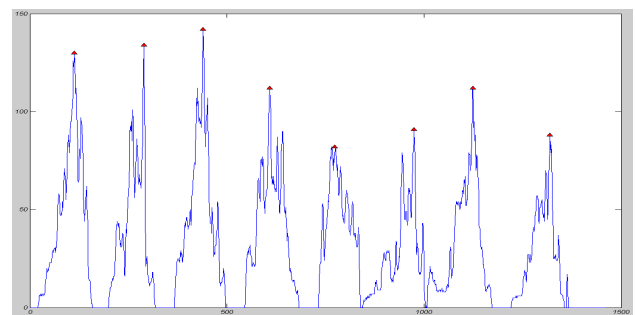
همانطور که در شکل ۴ نیز کاملاً مشخص است معمولاً نقاط و سرکش‌ها با توجه به اینکه در بسیاری موارد فاصله عمودی آن‌ها تا بدنه اصلی زیاد است در یک خوشه مجزا قرار می‌گیرند. یکی دیگر از مواردی که در برخی نمونه‌ها رخ می‌دهد تقسیم‌شدن یک خط واحد

پارامتر جدا کردن خوشه‌های کوچک و بزرگ به صورت زیر تعریف می‌شود. تعداد پیکسل‌های قلم (مساحت) برای هر خوشه محاسبه شده و سپس مرتب می‌شوند. فاصله بین مساحت‌های مرتب شده خوشه‌ها بدست آمده و براساس رابطه ۳ نرمال می‌شود. مساحت شروع بزرگترین فاصله مقدار پارامتر حد آستانه جداسازی خوشه‌ها را تعیین می‌کند.

$$\text{DistArea}(i) = \frac{\text{sortedArea}(i+1) - \text{sortedArea}(i)}{\text{sortedArea}(i+1) + \text{sortedArea}(i)} \quad (3)$$



(الف)



(ب)

شکل ۳: الف) سند قطعه‌بندی شده به نوارهای عمودی (ب) پروفایل افکنش افقی نوار مشخص شده با علامت ستاره سبز در تصویر الف و نمایش قله‌ها.

۲-۲- خوشه‌بندی

جهت استخراج خطوط از روش خوشه بندی سلسله مراتبی کمترین فاصله ۲۲ استفاده شده است. برای این کار موقعیت افقی و عمودی تمامی نقاط قلم (پیکسل‌های سیاه) بعنوان ویژگی‌های نمونه‌ها مورد استفاده قرار گرفته است. برای این که عملیات خوشه‌بندی سریع‌تر

به چندین خوشه مجزا می‌باشد؛ این مشکل بخاطر تغییر ناگهانی شیب خط، فاصله زیاد بین اجزاء متصل یا عدم وجود معیار توقف مناسب و دقیق برای هر سند، رخ می‌دهد. برای غلبه بر این مشکلات از پس‌پردازش استفاده شده است که در قسمت بعدی شرح داده شده است.

۳-۲- پس‌پردازش

در این مرحله از تعدادی قاعده برای ترکیب خوشه‌ها و استخراج دقیق خطوط استفاده شده است که در زیر آورده شده است.

در مرحله اول اندازه هر خوشه بدست می‌آید و خوشه‌ها براساس یک حد آستانه، که در بخش ۲-۱-۲ آورده شده است، به دو دسته خوشه‌های کوچک و بزرگ دسته‌بندی می‌شوند و در صورتی که ناحیه بدنه محدب^{۲۳} یک خوشه کوچک فقط با یک خوشه بزرگ اشتراک بیش از ۵۰ درصد داشته باشد دو خوشه با هم ترکیب می‌شوند. با استفاده از قاعده بیان شده خوشه‌های کوچکی که شامل نقاط و سرکش‌های هستند به خوشه‌های اصلی می‌پیوندند.

در مرحله بعد برای تصمیم‌گیری در مورد سرکش‌ها زوایای خوشه‌های کوچک محاسبه می‌شود و در صورتیکه این زاویه بین ۳۰ تا ۶۰ درجه باشد سرکش تشخیص داده شده و با خوشه بزرگی که زیر آن قرار دارد ترکیب می‌شود.

در ادامه خط تقریبی مبنا برای خوشه‌های بزرگ محاسبه می‌شود به این‌صورت که به ازای مختصات تمام نقاط قلم موجود در یک خوشه بزرگ یک خط برازش می‌شود. این خطوط استخراج شده شیب احتمالی خطوط اصلی سند را نشان می‌دهد. با استفاده از این خطوط به‌دست آمده نیز قوانینی آورده شده است. اگر فاصله دو خط از خطوط برازش شده بسیار کم باشد نشان می‌دهد که دو خوشه بزرگ روی یک خط قرار گرفته‌اند و این دو خوشه با هم ترکیب می‌شوند.

در نهایت فاصله اقلیدسی مراکز خوشه‌های کوچکی که هنوز مورد تصمیم‌گیری قرار نگرفته‌اند با خطوط برازش شده به خوشه‌های بزرگ محاسبه شده و به خوشه مربوط به نزدیکترین خط می‌پیوندند.

با استفاده از قواعد ذکر شده اکثر مشکلات مربوط به استخراج خط حل می‌شود تنها مشکل باقیمانده که باقی می‌ماند مسئله نزدیکی بسیار زیاد یا اتصال دو جزء متصل از دو خط متفاوت است. برای یافتن خطوط با این شرایط ماکزیمم ارتفاع اجزاء متصل هر خوشه نهایی نسبت به خط تقریبی برازش شده برای هر خوشه استخراج می‌شود و اگر خوشه‌ای بیش از ۱,۵ برابر متوسط ارتفاع خوشه‌ها بود

پیکسل‌های خطوط ۲۰٪ میانی افکنش خطوط با شیب خط برازش شده استخراج می‌شود، در صورتی که مینیمی کمتر از ۱۰ وجود داشته باشد از این نقطه خوشه از محلی که کمترین پیکسل قلم را قطع کند با استفاده از قواعد ارائه شده در [۱] به دو خوشه از نقطه جدا شوند تقسیم می‌شود.

۳- نتایج تجربی

به‌منظور بررسی و تست روش پیشنهادی از دو مجموعه داده استاندارد استفاده شده است. مجموعه داده اول، مجموعه داده FHT [20] است که شامل ۱۱۲۶ سند دستنویس فارسی است که توسط ۲۵ نویسنده مختلف نوشته شده است که متوسط تعداد سطر در هر سند حدود ۷ می‌باشد. مجموعه داده دوم، مجموعه داده PHDT [۲۱] است که شامل ۱۴۰ سند دستنویس فارسی می‌باشد که توسط ۴۰ شخص متفاوت نگارش شده است و متوسط تعداد سطر در هر سند در حدود ۱۳ می‌باشد.

جهت ارزیابی دو معیار استاندارد مورد استفاده قرار گرفته است. اولین معیار نرخ تشخیص ۲۴ است [۱۵] که در استخراج خط بسیار پرکاربرد است. این معیار نرخ خطوط تشخیص داده شده است. اگر هر خط استخراج شده معیار تطابقی بالای ۹۵٪ را بدست آورد این خط بعنوان یک خط تشخیص داده شده به حساب می‌آید. این معیار تطابق به‌صورت زیر محاسبه می‌گردد.

$$MatchScore_{i,j} = T(G_i \cap R_j) / T(G_i \cup R_j) \quad (5)$$

4-

که در آن تابع $T(a)$ تعداد پیکسل‌های پیش‌زمینه a را بر می‌گرداند. G_i نشان‌دهنده آلمین تصویر از مجموعه درستی^{۲۵} و R_j نشان‌دهنده آلمین خط استخراج شده توسط رهیافت پیشنهادی است. این معیار، معیار استاندارد است که از سال ۲۰۰۷ در مسابقات کنفرانس بین‌المللی آنالیز و تشخیص سند^{۲۶} مورد استفاده قرار می‌گیرد.

معیار ارزیابی بعدی که در این مورد ارائه شده است معیار نرخ تطابق پیکسلی^{۲۷} است. این معیار براساس تقسیم تعداد پیکسل‌های قلم مشترک بین بهترین تطابق در میان تصاویر مجموعه درستی و خط استخراج شده بر تعداد پیکسل‌های قلم این تصویر مجموعه درستی به‌دست می‌آید.

نتایج قابل مقایسه‌ای روی این مجموعه داده‌های استاندارد و مطرح در اسناد دستنویس فارسی با استفاده از روش ارائه شده در این مقاله

²⁶ International Conference on Document Analysis and Recognition (ICDAR)

²⁷ Pixel Level Hit Rate (PLHR)

²³ Convex hull

²⁴ Detection Rate (DR)

²⁵ Ground Truth

SID



ابزارهای
پژوهش



سرویس ترجمه
تخصصی



کارگاه های
آموزشی



بلاگ
مرکز اطلاعات علمی



سامانه ویراستاری
STES



فیلم های
آموزشی

کارگاه های آموزشی مرکز اطلاعات علمی



تازه های آموزش
آموزش مهارت های کاربردی در تدوین و چاپ مقالات ISI

آموزش مهارت های کاربردی
در تدوین و چاپ مقالات ISI



تازه های آموزش
روش تحقیق کمی

روش تحقیق کمی



تازه های آموزش
آموزش نرم افزار Word برای پژوهشگران

آموزش نرم افزار Word
برای پژوهشگران