استخراج خطوط در اسناد دستنویس فارسی مبتنی بر خوشه-بندی سلسله مراتبی

مجید ایرانپور مبارکه[۱]، علیرضا احمدی‌فرد[۲]

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

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

چکیده

استخراج خط از اسناد دستنویس یکی از مهم‌ترین مراحل پیش‌پردازش درآنالیز 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

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 متغییر است و همچنین معمولا 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 به‌هم چسبیدگی اجزاء 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 یاغلط 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[۵]، روشهای مبتنی بر لکه‌دارکردن[۶]،

---

[1] offline

[2] Word Spotting

[3] Word Image Retrieval

[4] Optical Character Recognition (OCR)

[5] Projection-based Methods

[6] Smearing Methods

روشـــهای مبتنی بر تبـدیل هاف[7]، روش مبتنی بر نازک‌ســـازی[8]و روشهای احتمالی[9].

روشـهای مبتنی بر افکنش معمولا درمورد اسناد تصویری چاپی مورد استفاده قرار می‌گیرد و برای اسناد دستنویسی که خطوط همپوشانی ندارند نیز توسـعه یافته اسـت[۲, ۳]. اما این روش بعلت مشـکلات فراوان، مانند درنظر نگرفتن شـیب خط، در متون اسـناد دسـتنویس تصویری، نمی‌تواند روی را اسناد مستقیما اعمال شود. به همین دلیل جهـت بهبود این روش راهکارهایی مانند اعمال روشـــهای مبتنی بر افکنش به برشـهای عمودی در سـند و سـپس ترکیب نتایج، توسـعه داده شده است[۴-۷]. در این روش سند تصویری به صورت موازی به نوارهـای عمودی تقســیم‌بندی می‌شـــود . ســـپس مرزهای بالقوه جداکننده خطوط در هر بخش بدســـت می‌آید و در مرحله بعد این بخشهای کاندید با استفاده از هیوریستیک یا مدلهای گرافیکی آماری مانند مدل مخفی مارکوف[10] به همدیگر متصل شده و نواحی خطوط را اســـتخراج می‌کننـد[۶, ۷].نهـایتا برای فائق آمدن بر مواردی که خطوط همپوشـــانی دارند از تکنیک‌های تقاطع[11] ، الگوریتم تعقیب کـانتور و چگـالی گاوســـی دو متغییره برای بازنمایی خطوط [۴, ۵] استفاده شده است.

روش مذکور دارایی نقایصـی اسـت که در زیر به آنها اشاره شده است [۸] و شـامل موارد ذیل است. این روش خطوط بالقوه زیادی را تولید می‌کند، پارامتر عرض نوارهای برش باید از پیش تعرف شـــده باشـــد، خطوط نباید انحنای زیاد داشـــته باشـــند و اگر نوار ابتدایی و انتهایی شامل خطوط بالقوه نباشند، استخراج خط بطور کامل ناممکن است.

برخی شـیب خط را نیز برای جداسـازی خطوط مورد اسـتفاده قرار داده‌اند[۴] که این شـیب معمولاً بسـادگی قابل محاسـبه نیست و در یک سـند بین خطوط متفاوت اسـت. بنابراین این متدها نیز در این شرایط نتایج خوبی را ارائه نمی‌کنند.

در روشـــهـای مبتنی بر لکـه‌دار کردن روشـــهـایی مـانـنـد الگوریتم هموارسازی طول اجرا[12] [۹] یا الگوریتم آنالیز طول اجرای جهت دار بهبود یافته[13] [۱۰] بکار گرفته شده است. استفاده از این الگوریتم‌ها باعث خواهد شـد که بلوک‌های پیکسلی در جهت افقی لکه‌دار شوند؛ یعنی اگر فاصله سفید(پس زمینه) بین آنها از یک حد آستانه‌ای کمتر بود با پیکسـل سـیاه(پیش‌زمینه) پر شود. این روش ابتداروی تصاویر خاکسـتری اعمال شـده و سـپس تصـویر باینری می‌شـود و عمل مورفولوژی فرسـایش[14] روی پیش‌زمینه و پس‌زمینه برای بدسـت 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 متصـل بدسـت 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 می‌شـود. سپس این نتایج حاصـل شـده با اسـتفاده از حذف هم‌ترازی‌های نادرسـت با استفاده از هم‌ترازی بین اجزاء متصل با استفاده از اطلاعات محتوایی، تصـویر خروجی نهایی را ارائه می‌کند[۱۱]. در کار دیگری روشـی را مبتنی بر تبـدیـل هـاف بر روی ســـند بلوک‌بندی شـــده ارائه کرده است[۱۲].

برخی محققین برای اســـتخراج خطوط هنـدی و چینی از روش نـازک‌ســـازی اســـتفـاده کرده‌انـد[۱۰, ۱۳, ۱۴]، که در[۱۴] ابتدا نازک‌سازی صورت گرفته و در ادامه عملیات پس‌پردازشی روی تمامی نواحی پس‌زمینه برای تشـخیص مرز جداکننده خطوط اعمال شـده است.

اخیراً برخی از روشـهای مجموعه سـطوح[15]، کانتور فعال[16] و بیز تغییرات[17] برای اســـتخراج خطوط بهره برده‌اند[۱۳, ۱۵, ۱۶]. در [۱۵]از روش تخمین چگالی و مجموعه سطوح استفاده شده است. در این کار یک نگاشـت احتمالی از سـند تصـویری ورودی تخمین زده می‌شـود که هر المان احتمال تعلق هر پیکسـل به یک خط را نشـان می‌دهد. روش مجموعه ســـطوح برای تعیین ارزیابی مرزی با خطوط همسـایه مورد اسـتفاده قرار گرفته اسـت. در[۱۶] در ابتدا یک بانک فیلتر تطبیق یافته برای هموارسازی تصویر ورودی مورد استفاده قرار گرفته اسـت. مراکز خطوط اجزاء تشـکیل دهنده خط با اسـتفاده از مرزبندی تصـویر هموار شـده محاسـبه می‌شود. در انتها برای بدست آوردن نتیجه نهایی، کانتورهای فعال روی مرزها وفق داده می‌شـوند. در کـار [۱۳] روش بیز تغییرات برای اســـتخراج خطوط در متون دســـتنویس چینی مورد بهره برداری قرار گرفته اســـت. یک ســـند به‌صـــورت یک مدل ترکیب گاوســـی[18] در نظر گرفته می‌شـــود،

---

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)

13 Improved Directional Run-Length Analysis

14 Erosion

15 Level Set

16 Active Contour

17 Variational Bayes

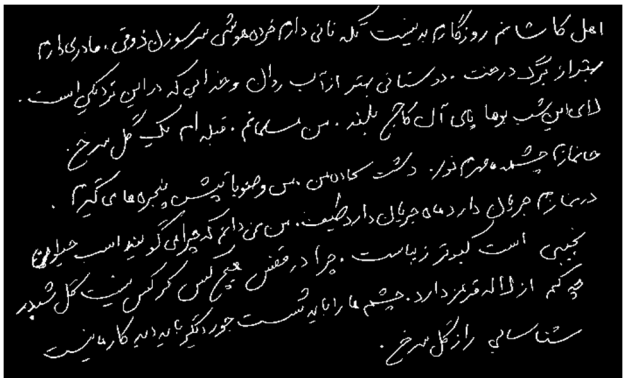
18 Gaussian Mixture Model

طوری‌که هر جزء متعلق به یک خط باشـــد. برای تخمین پارامترهای چگـالی و تعیین تعـداد خطوط چارچوب کاری بیز تغییرات بکار برده شده است.

در[۱۷, ۱۸] یـک روش مبتنی بر بلاک‌بنـدی معرفی شـــده کـه پارامترهای آن به‌صـورت وفقی براساس سند ورودی تعیین می‌گردد. در هر بلاک برای تشـخیص جهت از فیلتر کردن تصویر با هسته‌های گاوسی دوبعدی ناهمسانگرد[19] در زوایای بین ۳۰- درجه تا ۳۰ درجه استفاده شده است. پس از تشخیص جهت خطوط در هر بلاک از هم جدا می‌شوند.

اسـناد دسـتنویس فارسـی و عربی دارای تعداد زیاد و متنوعی نقطه، سـرکش و نشـانه هستند (شکل ۱) و همچنین در این نوع اسناد یک جزء متصـل می‌تواند یک کلمه، زیرکلمه، کاراکتر، نقطه یا بخشـی از نقاط پیوسـته یا یک سـرکش باشـد. با توجه به دلایل مذکور مسـئله استخراج خطوط دستنویس برای اکثر روش‌ها چالش‌برانگیزتر خواهد بود و برخی از روش‌هـا مـانند تبدیل هاف، لکه‌دار کردن طول اجرای جهـت دار بهبود یـافتـه و کانتور فعال در مواجهه با این نوع اســـناد شـکسـت خواهند خورد. علت این شکست‌ها را می‌توان به‌صورت زیر تشـریح کرد: ۱-روشـــهای مبتنی بر تبدیل هاف اهمیت بدنه اصـلی کلمه و نقاط و سـرکش‌های کوچک را در فضـای هاف یکسان در نظر می‌گیرد. ۲- در روشـهایی که از آنالیز طول اجرای اسـتفاده می‌شـود ممکن است اجزاء کوچک برخی خطوط مانند نقطه و سرکش به خط دیگری متصـل شوند. ۳-در برخی روشها با استفاده از فیلتری نقاط و سـرکش‌ها حذف می‌شـوند در صـورتی‌که این اجزاء در زبان فارسی و عربی نقش تعیین‌کننده‌ای دارند.

در[۸] روشـی مبتنی بر تکنیک نقاشـی[20] ارائه شـده اسـت. در این روش یک سند به نوارهای عمودی تقسیم‌بندی می‌شود وبا 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. سـپس از عمل مورفولوژی انبسـاط[21] جهت برقراری اتصـال بین لکه‌های ایجاد شده درمرحله قبل و تشکیل خط استفاده می‌شود. ســـپس روی پس‌زمینـه نتیجه مرحله قبل عمل نازک‌ســـازی انجام می‌گیرد سـپس با استفاده از برخی قواعد هیوریستیک خطوط نهایی اســتخراج می‌شـوند. اما این روش باتوجه به اینکه از روشـــهای قبلی بسـیار بهتر نسبت به فارسی و عربی کار میکند اما باز هم درمورد این دو زبان درصـد خطاهایی بسـیار بیشـتری را گزارش می‌کند. اما در روش ارائه شـده در این مقاله سعی شده تا این مشکلات را با استفاده از خوشـه‌بندی سـلسـله مراتبی با استفاده از یک معیار فاصله معرفی شده مرتفع سازد.



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

در این مقاله روشـی ارائه شـده است که از خوشه‌بندی سلسله‌مراتبی پیکسـل‌های قلم (سیاه یا پیش‌زمینه) موجود در سند جهت استخراج خطوط اسـتفاده می‌کند. این روش در ادامه مقاله در بخش دوم بطور کامل تشـریح شـده است. در بخش سوم نیز نتایج آزمایشات روی دو مجموعه اســتاندارد زبان فارســی گزارش شـــده و ســـپس بحث و نتیجه‌گیری در بخش چهارم آورده شده است.

۲- روش پیشنهادی

مقالات در سـاختار کلی روش پیشـنهادی، که در شـکل ۲ نشان داده شده است، شیوه نگارش متن فارسی به‌صورت دستنویس مد نظر قرار داده شـود. از آنجایی که معمولاً دسـت‌نوشـته‌ها ممکن اسـت شـامل خطوط نزدیک به همدیگر باشـند یا نقاط و سرکش‌ها باعث اشتباه در خوشـه‌بندی 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 فارسـی و عربی شیب خط مبنای استاندارد مستقیم است و در متون دستنویس این شیب به مقدار کمی می‌تواند خط مبنا را به سـمت بالا یا پایین متمایل کند یا در آن انحنای کمی را ایجاد نماید معیار فاصـله‌ای معرفی شـده است که این مسئله را در نظر بگیرد. نتایج نهایی خوشه‌بندی با استفاده از تخمین شیب خط و هیوریسـتیک‌هایی مبتنی بر شـیوه نگارش فارسـی اصـلاح و تکمیل می‌گردد. در ادامه تمامی مراحل کار به تفصیل آورده شده است.

۲-۱- پیش‌پردازش

[19] Anisotropic 2D Gaussian Kernels

[20] Painting Technique

[21] Dilation

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 وقفی.

- **نازک‌سازی**

با توجه با اینکه در متون دستنویس فارسی بالارونده‌هایی مانند حروف الف ولام 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 وقفی پارامترهای مورد نیاز**

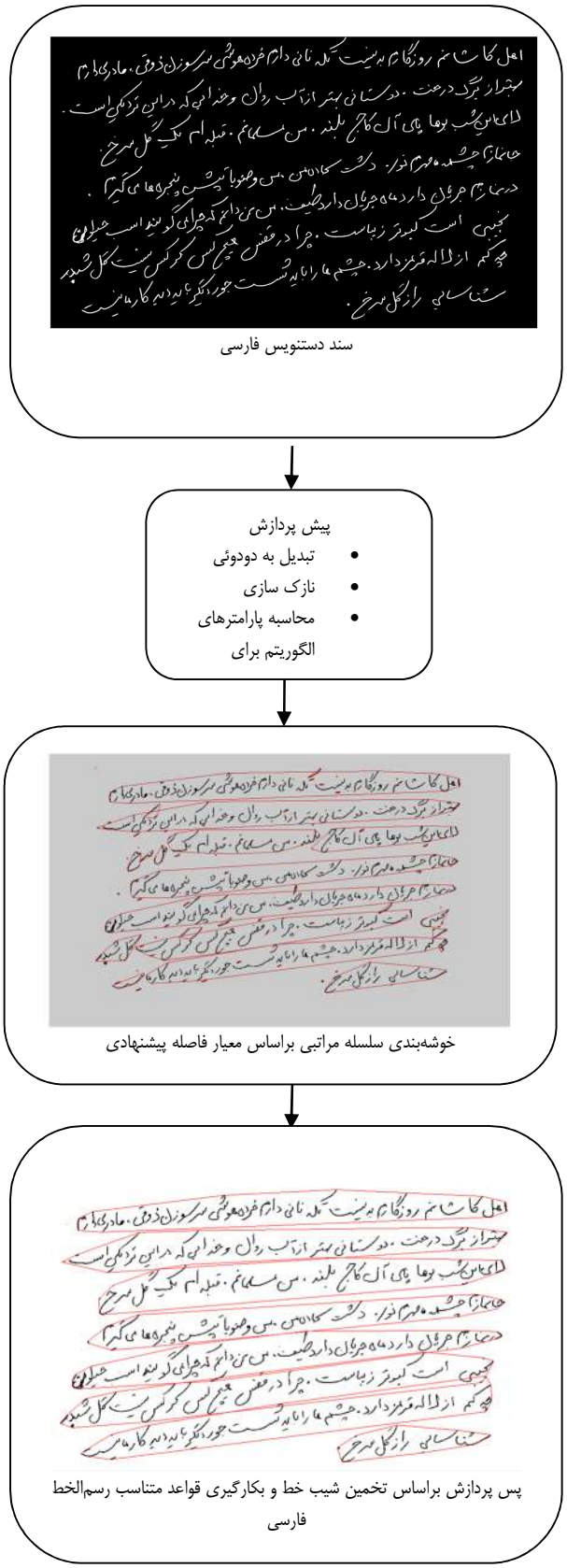
با توجه به اینکه در اسناد ورودی تعداد خطوط، اندازه تصویر و نوع دستخط از نظر اندازه متفاوت است نیاز به تعیین برخی پارامترها برای هر سند وجود دارد. در ادامه این پارامترها و رابطه بدست آمده جهت استخراج آنها را تشریح می‌کنیم.

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:

$$Line\_D_e = \frac{\sum_{i \in Corr\_D} i * Hist(i)}{\sum_{i \in Corr\_D} Hist(i)} \quad (۱)$$

$$Corr\_D = \{j | Hist(j) > 0.25 * \max(Hist)\} \quad (۲)$$

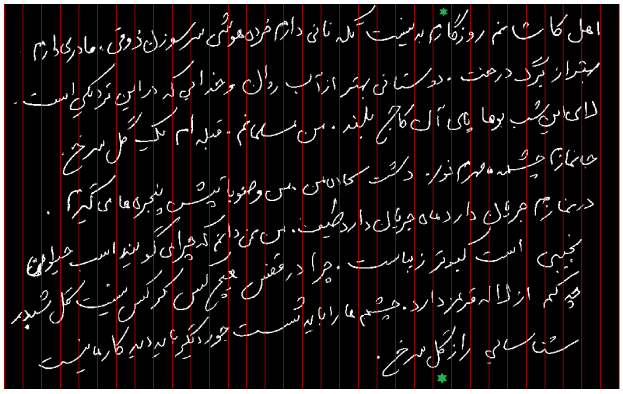
که در این روابط Corr_D شامل اندیس فاصله‌هایی می‌شود که تعدد آنها حداقل از ۲۵ درصد حداکثر بیشتر باشد. Hist هیستوگرام فاصله بین خطوط بدست آمده را نشان می‌دهد. Line_De نشان دهنده فاصله موثر بین خطوط است. از نصف این فاصله بعنوان پارامتر معیار توقف خوشه‌بندی استفاده شده است.



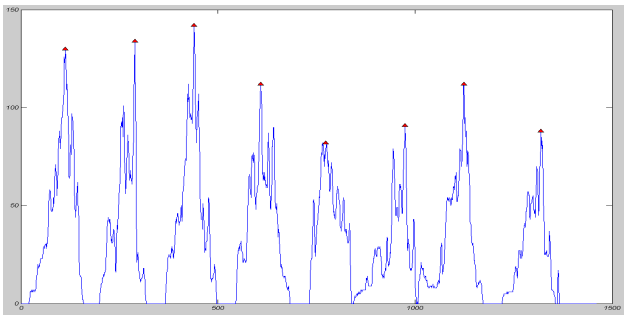
شکل ۲: شمای کلی روش پیشنهادی.

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.

$$Dist_{Area}(i) = \frac{sortedArea(i+1) - sortedArea(i)}{sortedArea(i+1) + sortedArea(i)} \quad (۳)$$



(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

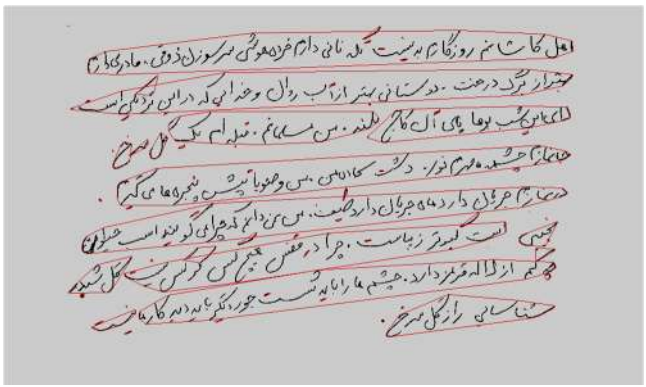
جهت استخراج خطوط از روش خوشه بندی سلسله مراتبی کمترین فاصله[۲۲] استفاده شده است. برای این‌کار موقعیت افقی و عمودی تمامی نقاط قلم (پیکسل‌های سیاه) بعنوان ویژگی‌های نمونه‌ها مورد استفاده قرار گرفته است. برای این‌که عملیات خوشه‌بندی سریع‌تر انجام شود هر جزء متصل بعنوان یک خوشه اولیه درنظر گرفته شده و عمل 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.

$$D_{p_1,p_2} = |X_{p_2} - X_{p_1}| + |Y_{p_2} - Y_{p_1}|^5 \quad (۴)$$

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 به‌ترتیب مختصات افقی و عمودی هرکدام 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 این‌کار معیار توقف حداکثر فاصله بین دو خوشه درنظر گرفته شده وبرای هر سند به‌صورت 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

[22] Single linkage

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 صورتی‌که 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 درصورتیکه 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 می‌پیوندد.

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، درصورتی‌که مینیممی 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 [] 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 ۱۳می‌باشد.

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 هرخط استخراج شده معیار تطابقی بالای ۹۵٪ را بدست آورد این خط بعنوان یک خط تشخیص داده شده به‌حساب می‌آید. این معیار تطابق به‌صورت زیر محاسبه می‌گردد.

$$MatchScore_{i,j} = T(G_i \cap R_j)/T(G_i \cup R_j) \quad (۵)$$

4-

که در آن تابع $T(a)$ تعداد پیکسل‌های پیش‌زمینه a را بر می‌گرداند. G_i نشان‌دهنده iامین تصویر از مجموعه درستی[۲۵] و R_j نشان دهنده jامین خط استخراج شده توسط رهیافت پیشنهادی است. این معیار، معیار استانداردی 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 تطبیق 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

---

[23] Convex hull

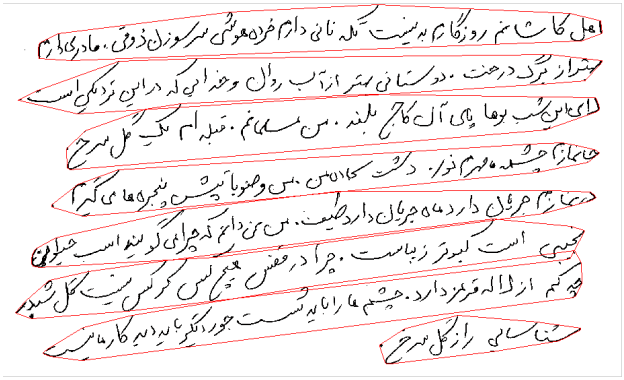
[24] Detection Rate (DR)

[25] Ground Truth

[26] International Conference on Document Analysis and Recognition (ICDAR)

[27] Pixel Level Hit Rate (PLHR)

به‌دست آمده است. در جداول ۱ و۲ مقایسه استخراج خط دو مجموعه داده معرفی شده با روش‌های دیگر نشان داده شده است.



شکل ۵: یک سند نمونه خروجی مرحله خوشه‌بندی.

جدول۱ مقایسه روی مجموعه داده FHT

روش	نرخ تشخیص	نرخ تطابق پیکسلی
روش مقاله [۱۷]	۹۱,۲۲	۹۸,۶۷
روش ارائه شده	**۹۲,۱۷**	**۹۷,۵۳**

جدول۲ مقایسه روی مجموعه داده PHDT

روش	نرخ تشخیص	نرخ تطابق پیکسلی
روش مقاله [۱۷]	۹۰,۳۲	۹۴,۷۳
روش ارائه شده	**۹۱,۱۷**	**۹۳,۰۰**

[۱] L. Likforman-Sulem, A. Zahour, and B. Taconet, "Text line segmentation of historical documents: a survey," *International Journal of Document Analysis and Recognition (IJDAR),* vol. 9, pp. 123-138, 2007.

[۲] R. Manmatha and J. L. Rothfeder, "A scale space approach for automatically segmenting words from historical handwritten documents," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 27, pp. 1212-1225, 2005.

[۳] M. R. Hashemi, O. Fatemi, and R. Safavi, "Persian cursive script recognition," in *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, 1995, pp. 869-873.

[۴] A. Zahour, B. Taconet, P. Mercy, and S. Ramdane, "Arabic hand-written text-line extraction," in *Document Analysis and Recognition, 2001. Proceedings. Sixth International Conference on*, 2001, pp. 281-285.

[۵] A. Zahour, L. Likforman-Sulem, W. Boussalaa, and B. Taconet, "Text line segmentation of historical arabic documents," in *Ninth .۰۷Document Analysis and Recognition, 2007. ICDAR 20 International Conference on*, 2007, pp. 138-142.

[۶] M. Arivazhagan, H. Srinivasan, and S. Srihari, "A statistical approach to handwritten line segmentation," *Document Recognition and Retrieval XIV, Proceedings of SPIE, San Jose, CA,* ۲۰۰۷, pp. 6500T-1.

[۷] V. Papavassiliou, T. Stafylakis, V. Katsouros, and G. Carayannis, "Handwritten document image segmentation into text lines and words," *Pattern recognition,* vol. 43, pp. 369-377, 2010.

[۸] A. Alaei, U. Pal, and P. Nagabhushan, "A new scheme for unconstrained handwritten text-line segmentation," *Pattern Recognition,* vol. 44, pp. 917-928, 2011.

[۹] Z. Shi and V. Govindaraju, "Line separation for complex document images using fuzzy runlength," in *Document Image Workshop on*, 2004, pp. 306- *Analysis for Libraries, International* .306

[۱۰] P. P. Roy, U. Pal, and J. Lladós, "Morphology based handwritten line segmentation using foreground and background information," in *International Conference on Frontiers in Handwriting Recognition*, 2008, pp. 241-246.

[۱۱] L. Likforman-Sulem, A. Hanimyan, and C. Faure, "A Hough based algorithm for extracting text lines in handwritten documents," in *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, 1995, pp. 774-777.

[۱۲] Louloudis, B. Gatos, I. Pratikakis, and K. Halatsis, "A block- .G based Hough transform mapping for text line detection in handwritten documents," in *Tenth International Workshop on Frontiers in Handwriting Recognition*, 2006.

[۱۳] F. Yin and C.-L. Liu, "A variational bayes method for handwritten text line segmentation," in *Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on*, 2009, pp. 436-440.

[۱۴] Y. A. SHMJITSURUOKA and T. YOSHIKAWA, "THE SEGMENTATION OF A TEXT LINE FOR A HANDWRITTEN UNCONSTRAINED DOCUMENT USING THINNING ".ALGORITHM

[۱۵] Y. Li, Y. Zheng, D. Doermann, and S. Jaeger, "Script-independent text line segmentation in freestyle handwritten documents," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 30, pp. 1313-1329, 2008.

[۱۶] S. S. Bukhari, F. Shafait, and T. M. Breuel, "Script-independent handwritten textlines segmentation using active contours," in *Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on*, 2009, pp ۴۵۰-۴۴۶.

[۱۷] M. Ziaratban and K. Faez, "An Adaptive Script-Independent .Block-Based Text Line Extraction," in *ICPR*, 2010, pp. 249-252

[۱۸] M. Ziaratban and F. Karim, "Adaptive Script-Independent Text Line Extraction," *IEICE transactions on information and systems,* vol. 94, pp. 866-877, 2011.

[۱۹] N. Otsu, "A threshold selection method from gray-level .histograms," *Automatica,* vol. 11, pp. 23-27, 1975

[۲۰] M. Ziaratban, K. Faez, and F. Bagheri, "FHT: An unconstraint Farsi handwritten text database," in *Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on*, 2009, pp. 281-285.

[۲۱] A. Alaei, P. Nagabhushan, and U. Pal, "A new dataset of Persian handwritten documents and its segmentation," in *Machine Vision and Image Processing (MVIP), 2011 7th Iranian*, 2011, pp. 1-5.