
بسم الله الرحمن الرحيم

دانشگاه سمنان

دانشکده مهندسی برق و کامپیوتر

پایان‌نامه کارشناسی ارشد مهندسی برق - الکترونیک

عنوان:

جداسازی خطوط در متون دست‌نویس فارسی

نگارش:

عاطفه سهرابی

استاد راهنما:

دکتر سعید مظفری

اسفند ۱۳۹۱

بسمه تعالی

اینجانب *عاطفه سهرابی* متعهد می‌شوم که محتوای علمی این نوشتار با عنوان *جداسازی خطوط در متون دست‌نویس فارسی* که به عنوان پایان‌نامه کارشناسی ارشد رشته *مهندسی برق گرایش الکترونیک* به گروه مهندسی *برق* دانشکده مهندسی برق و کامپیوتر دانشگاه سمنان ارائه شده، دارای اصالت پژوهشی بوده و حاصل فعالیت‌های علمی این جانب می‌باشد.

در صورتی که خلاف ادعای فوق هر زمانی محرز شود، کلیه حقوق معنوی متعلق به این پایان‌نامه از اینجانب سلب شده و موارد قانونی مترتب به آن نیز از طرف مراجع ذیربط قابل پیگیری است.

نام و نام خانوادگی

شماره دانشجویی امضا

تقدیم به مهربان فرشتگانی که:

لحظات ناب باور بودن، لذت و غرور دانستن، جسارت خواستن، عظمت رسیدن، و

تمام تجربه های زیبای زندگیم، مدیون حضور سبز آنهاست.

تقدیم به پدر و مادر عزیز

و

همسر مهربانم

چکیده

هدف این رساله، تقسیم‌بندی یک سند دست‌نویس به خطوط مجزا از یکدیگر است. جداسازی و استخراج خط، اولین و مهمترین مرحله پیش‌پردازش برای آنالیز و بازیابی تصویر یک سند است. هنگامی‌که یک خط جداسازی شد، (خط ایزوله شده) برای مراحل جداسازی کلمه، شناسایی آن، بازیابی زیرکلمات و حروف و سایر مراحلی که برای آنالیز یک سند مورد نیاز است، تحت بررسی قرار می‌گیرد. بنابراین استخراج صحیح خطوط، بمنظور درست انجام شدن سایر مراحل، دارای اهمیت فراوانی است. از طرفی جداسازی خطوط در اسناد دست‌نویس بدون محدودیت، به دلیل ویژگی‌های خاص این اسناد، یک مشکل اساسی است. تغییر در اندازه حروف، عدم وجود یک امتداد برای خط کرسی در سراسر یک خط، فاصله کم و متغیر بین خطوط و تماس و در هم فرو رفتگی خطوط مجاور و ... از جمله عواملی است که سبب می‌شود جداسازی خطوط با مشکلات زیادی مواجه گردد. در این رساله، ابتدا خط کرسی از یک خط با روش ورژن هموار شده نمودار منحنی طرح، تشخیص داده شده و سپس برای جداسازی کامل خط، به بررسی و نسبت دادن المان‌های موجود بین دو خط مجاور، به یکی از خطوط فوقانی یا پایینی پرداخته می‌شود، تا بدین ترتیب خط، جداسازی و استخراج گردد. یکی از مراحل اساسی که در این رساله برای استخراج خط کرسی و تخصیص مؤلفه‌های بین خطوط، مورد استفاده است، بکارگیری مدل مارکوف مخفی است. نتایج بدست آمده برای جداسازی خطوط به روش گفته شده، در زبان فارسی ۹۷.۷۵ درصد و برای سایر زبان‌ها ۹۸.۲۵ بدست آمده است.

کلید واژگان: اسناد دست‌نویس[۱]، جداسازی خط[۲]، خط کرسی[۳]، بین خطوط[۴]، مؤلفه‌های پیوسته[۵]، مدل مارکوف مخفی[۶]

[1] Handweitten Documents
[2] Line Segmentation
[3] Base Line
[4] Between lines
[5] Connected Component
[6] Hidden markov model

فهرست مطالب

عنوان **صفحه**

۱

مقدمه

۱-۱) پردازش تصویر

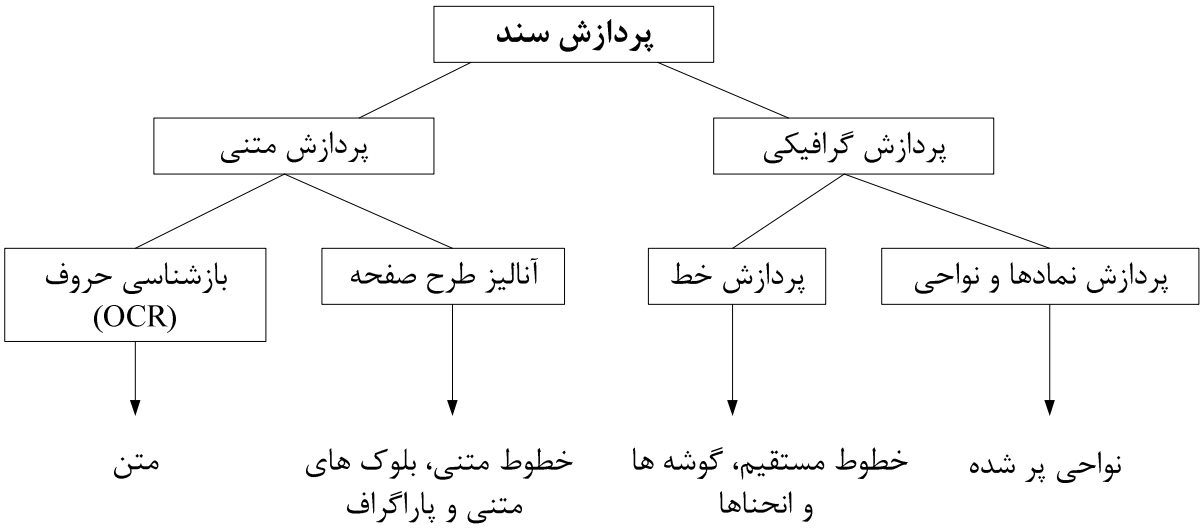
هر نوع پردازش سیگنال که دارای ورودی و خروجی است، پردازش تصویر[۱] نامیده می‌شود. ورودی، یک تصویر مانند عکس است. خروجی نیز می‌تواند یک تصویر یا یک مجموعه از نشانه‌های ویژه و یا متغیرهای مربوط به تصویر باشد. اغلب تکنیک‌های پردازش تصویر، شامل برخورد با تصویر به عنوان یک سیگنال دو بعدی، و به کار بستن تکنیک‌های استاندارد پردازش سیگنال روی آنها است. پردازش تصاویر امروزه بیشتر به موضوع پردازش تصویر دیجیتال گفته می‌شود که شاخه‌ای از دانش رایانه است که با پردازش سیگنال دیجیتال که نماینده تصاویر برداشته شده با دوربین دیجیتال یا پویش شده توسط پویشگر[۲] است سروکار دارد. یکی از انواع تصاویر ورودی، تصویر سند است.

۱-۲) تصویر سند

روش متداول ذخیره‌سازی و ارسال اطلاعات، به وسیله اسناد کاغذی است. اما با توجه به مشکلات و سختی‌های سروکار داشتن با کاغذ، در چند دهه اخیر استفاده از کامپیوتر به منظور ذخیره‌سازی اسناد، با رشد چشم‌گیری همراه بوده است. شروع استفاده از این تکنولوژی از اوایل سال ۱۹۸۰ در آمریکا بود. یکی از مشکلات ذخیره‌سازی سند بصورت دیجیتال، سختی کار در بازیابی[۱] این اسناد است. بنابراین پردازش سند، بعنوان زیرمجموعه‌ای قابل توجه از پردازش تصویر مورد بررسی قرار گرفت. امروزه سیستم‌های پردازش اسناد دارای قابلیت‌های بسیار گسترده

[1] Image processing
[2] Scanner

هستند. به طور مثال این سیستم‌ها هم اکنون نوع سند را تشخیص می‌دهند، می‌توانند قسمت‌های اساسی آن را استخراج کنند، و در صورت نیاز آن را از یک فرمت به فرمت دیگر تبدیل کنند. انواع پردازش تصویر سند در شکل ۱-۱ نشان داده شده است[1].



شکل۱-۱ سلسله بندی پردازش تصویر سند در نواحی مختلف

۱-۳) بازشناسی نوری حروف (OCR)

OCR [۲] بعنوان یک سیستم بازشناسی متن، معرفی می‌شود (بازشناسی متن، تبدیل تصویر به متن قابل ویرایش است). OCR کوته‌نوشت بازشناسی نوری نویسه‌ها[۳] یا نویسه‌خوان نوری[۴] است. OCR در ابتدا تنها در مورد بازشناسی ارقام و حروف چاپی بکار گرفته می‌شد. با گذشت زمان و پیشرفت قابل توجه در این زمینه، روش‌های بازشناسی دست‌نوشته و متون چاپی مطرح شدند که دامنه کار را به کلمات و عبارات رساندند. با وجود عدم تطبیق OCR با این موارد، این نام برای این روش‌ها و تا حدی برای بازشناسی دست‌نوشته‌های برخط هم استفاده شد و رواج پیداکرد. بدین ترتیب، امروزه سیستم OCR برای بازشناسی کلیه مستندات چاپی و مکتوبات

[1] Recognition
[2] Optical Character Recognition
[3] Optical Character Recognition
[4] Optical Character Reader

تایپی[1] مثل صفحات کتاب‌ها، مجله‌ها و نامه‌های چاپی و همچنین اسناد دست‌نویس[2] ، هم در نوشته‌های برخط[3] و هم در نویسه‌های برون‌خط[4] بکار می‌رود.

یک سامانه OCR، مثل یک نفر ماشین‌نویس، یک متن را می‌خواند و آن را به قالب مناسب برای ذخیره در کامپیوتر تبدیل می‌کند. معمولاً یک پویشگر، تصویر متن را برای OCR فراهم می‌کند. این تصویر معمولاً از نقاط سیاه و سفید تشکیل می‌شود. سامانه OCR ، اشیاء موجود در این تصویر را که ارقام، حروف، علائم، و کلمات هستند، بازشناسی کرده و نام آن‌ها را در قالب مناسب ذخیره می‌کند. یک فایل تصویری، حجم زیادی دارد و جستجوی متنی در آن ممکن نیست. این در حالی است که فایل خروجی یک سامانه OCR بسیار کم‌حجم و قابل جستجو است[۲] . بنابراین استفاده از سیستم‌های OCR دو مزیت عمده دارد:

۱. افزایش چشمگیر سرعت دسترسی به اطلاعات. زیرا در متن بر خلاف تصویر، امکان جستجو و ویرایش وجود دارد.

۲. کاهش فضای ذخیره سازی. زیرا حجم فایل متنی استخراج شده از یک تصویر، معمولاً بسیار کمتر از حجم خود فایل تصویری است.

۱-۴) معرفی کلی یک سیستم بازشناسی متن

طبق آنچه تاکنون بیان شد برای تبدیل اطلاعات داخل تصاویر متن، به شکل قابل فهم برای ماشین، نیاز به یک سیستم بازشناسی متن (OCR) است. در این بخش، مفاهیم، تعاریف و اصطلاحات بازشناسی متن معرفی می‌شود. اجزای اصلی یک سیستم بازشناسی متن عبارتند از :

جداسازی : جداسازی بلوک‌های تصویر مانند متن، گراف و شکل در این مرحله انجام می‌شود.

استخراج ویژگی : برای معرفی هر یک از اجزای تصویر به شکل قابل فهم برای رایانه، مشخصات هر یک از اجزای تصویر در این مرحله استخراج می شود.

[1] Machine-Printed
[2] Handwritten
[3] Online
[4] Offline

بازشناسی یا طبقه‌بندی متن : بلوک‌های متن استخراج شده در مرحله قبل، در این مرحله بازشناسی می شوند. بازشناسی هر یک از اجزای تصویر متن با طبقه‌بندی آن‌ها در یک مجموعه حروف، زیرکلمات یا کلمات معتبر انجام می‌شود.

اجزای مکمل برای یک سیستم بازشناسی متن عبارتند از:

پیش‌پردازش : پردازش‌های لازم برای بهبود اولیه کیفیت تصویر، مانند رفع کجی و حذف نویز در این مرحله انجام می‌شود.

پس‌پردازش : کلمات بازشناسی شده در مرحله بازشناسی، با یک واژه نامه مقایسه و تأیید می-شود.

بازشناسی فرمول : روش بازشناسی که قادر به تبدیل تصویر فرمول به متن قابل ویرایش باشد.

تشخیص متن داخل جدول : قابلیت سیستم بازشناسی در استخراج متون داخل جدول و بازشناسی آن‌ها، و تبدیل این متون با شکل جدول متنی.

تحلیل ساختار فیزیکی : استخراج بلوک‌های داخل سند به نحوی که موقعیت فیزیکی هر یک در سند ثبت شده و پس از بازشناسی محتوای آن‌ها، برای بازتولید سند متنی، قابل استفاده باشد.

تحلیل ساختار منطقی : تعیین جنس هر یک از بلوک‌های تصویر مانند عنوان، چکیده، مقدمه و متن اصلی برای بلوک‌های متنی.

غلط‌یاب املایی : قابلیت سیستم بازشناسی در تشخیص غلط‌های املایی.

پس‌پردازش نحوی : قابلیت سیستم بازشناسی در تشخیص غلط‌های نحوی.

بازسازی سند[1] : پس از بازشناسی تمام اجزای تصویر، شکل ظاهری سند با استفاده از اجزای بازشناسی شده، مانند تصویر اصلی بازتولید می‌شود.

تکامل و توسعه سیستم OCR، با سرعت چشمگیری انجام شد، اما علی‌رغم پیشرفت‌های زیادی که در زمینه پردازش متون تایپی و دست‌نویس بدست آمده است، قدرت ماشین‌ها در خواندن متون، هنوز از توانایی بشر فاصله زیادی دارد. با این وجود، خواندن خودکار متون تایپی و دست‌نویس در بسیاری از زبان‌ها به مراحل بلوغ رسیده است. تاکنون کارهای زیادی بر روی متون به زبان‌های لاتین، چینی و ژاپنی انجام شده است، اما برای متون فارسی و عربی و امثال آن‌ها، کارهای انجام

[1] Document Regeneration

شده نسبتاً کم و پراکنده بوده است. ذکر این نکته ضروری است که بیش از یک‌سوم جمعیت جهان به زبان‌هایی صحبت می‌کنند که برای نوشتن آن‌ها از نویسه‌های عربی یا شبیه به آن استفاده می‌کنند. از زبان‌های غیرعربی می‌توان به فارسی، دری، پشتو، اردو، کردی و جاوی اشاره کرد. با استناد به این مسئله، مشخص است که رشد OCR در زبان‌های عربی و مشابه آن، بسیار کند بوده و افق‌های دست‌یابی به تکامل آن چندان نزدیک به نظر نمی‌رسد. از جمله علل این امر را می‌توان عدم سرمایه‌گذاری کافی، خصلت پیوسته بودن این خطوط، و فقدان پایگاه داده‌ها و لغت-نامه‌های استاندارد و جامع برای متون فارسی و عربی و امثال آن‌ها ذکر کرد [۲] .

۱-۵) جداسازی خطوط

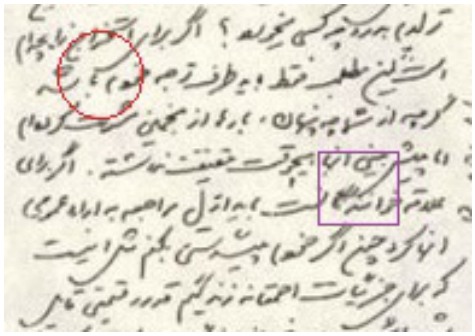
در این بخش با توجه به عنوان و هدف رساله، تمرکز بر روی زمینه‌های محدودتری از سیستم کلی OCR است. جداسازی خطوط متنی، یکی از زیرمجموعه‌های پیچیده و پرکاربرد آنالیز تصاویر اسناد است. این شاخه از آنالیز تصاویر، اطلاعات بسیار قاطع و تعیین‌کننده‌ای، بمنظور تصحیح کجی‌ها و انحناها، تقسیم‌بندی نواحی و بازشناسی حروف را فراهم می‌کند. بنابراین جداسازی خطوط، مرحله‌ای مهم بمنظور آنالیز برون‌خط اسناد است، چرا که جداسازی نادرست خطوط، منجر به بروز خطاهایی در مرحله شناسایی خواهد گشت. اگرچه جداسازی خطوط متنی در اسناد تایپی یا دست‌نویس، معمولاً بعنوان یک مسئله حل شده به نظر می‌رسد [3]، اما خطوط متنی اسناد دست‌نویس آزاد[1]، هنوز بعنوان یک مشکل قابل توجه وجود دارد. در جدول ۱-۱، لیستی از چند مقایسه بین جزئیات اسناد تایپی و دست‌نویس فهرست شده است [4].

جدول ۱-۱ مقایسه اسناد دست‌نویس و تایپی

	خط متن	فاصله بین خطوط همسایه	طرح (حاشیه)
اسناد تایپی	مستقیم	قابل توجه	منظم
اسناد دست‌نویس	منحنی	ناچیز	نامنظم

با توجه به جدول فوق، مرحله جداسازی خطوط برای اسناد تایپی، تا حدود زیادی به تکامل خود دست یافته است، اما در خطوط دست‌نویس موجود در اسناد، که از قالب خاصی تبعیت نمی‌کنند،

هم‌چنان یک مشکل مهم تلقی می‌شود و همچنان بعنوان یکی از مشکلات پیچیده در بهبود OCR موثق، باقی مانده است. علت این امر نیز ؛همان‌طور که در جدول فوق دیده می‌شود؛ طبیعت متون دست‌نویس است که فرآیند جداسازی خطوط را بسیار پرچالش می‌کند، زیرا در اسناد دست‌نویس فاصله داخلی بین خطوط و میزان انحراف خط کرسی متغیر است. مشکلات زمانی حادتر می‌شود که دو خط مجاور در متن سند، در محل‌هایی دچار تماس[2] یا هم‌پوشانی[3] شده باشند. نمونه آن در شکل ۱-۲ نشان داده شده است.



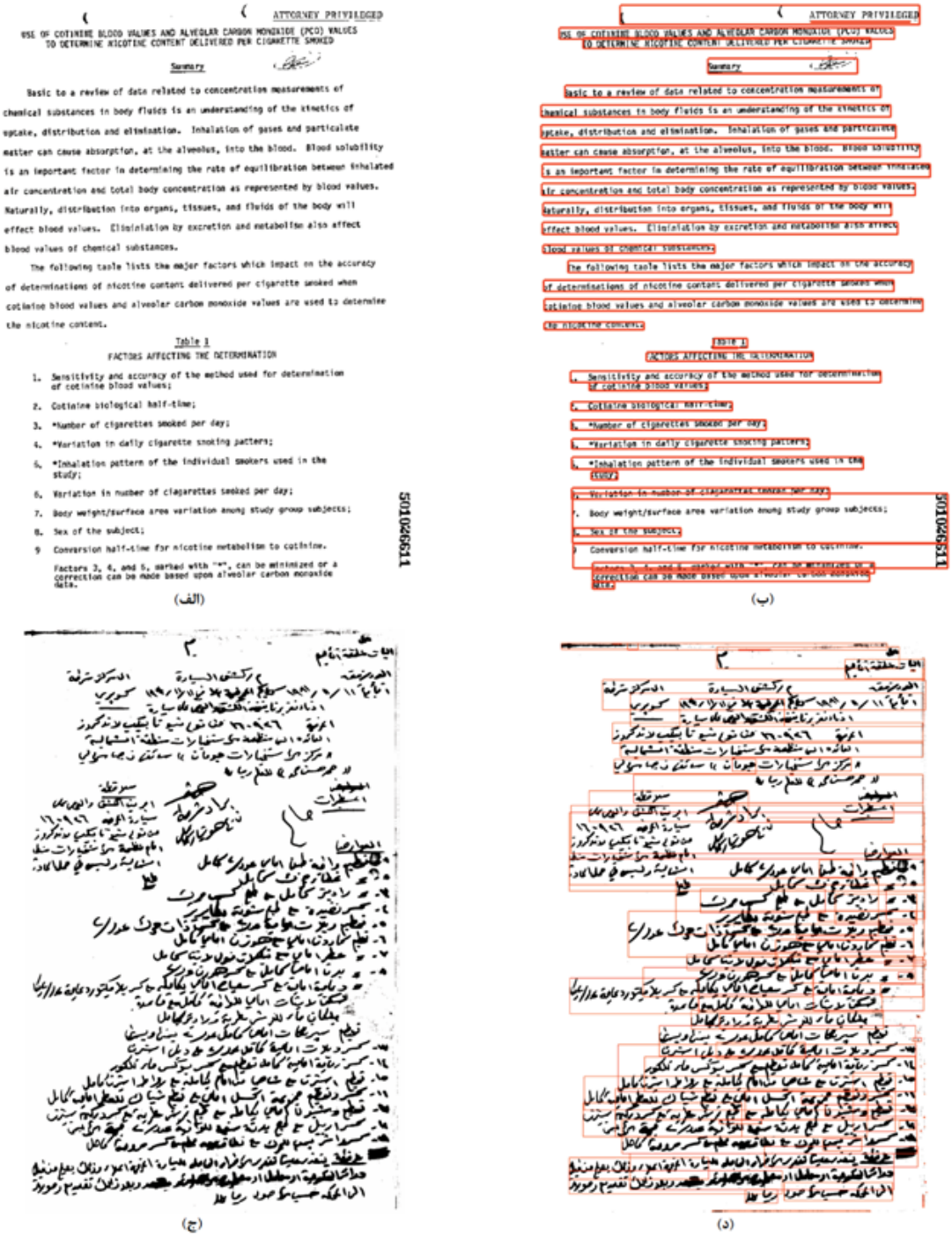
شکل ۱-۲ هم‌پوشانی (مستطیل) و تماس (دایره) دو کلمه

متون تایپی دارای ویژگی‌های مناسبی برای جداسازی هرچه بهتر خطوط هستند. از جمله مهمترین این ویژگی‌ها، یکسان بودن فاصله خطوط از یکدیگر در تمام امتداد دو خط مجاور، و همچنین متصل بودن تمام اجزای یک حرف به هم است. خلأ وجود این ویژگی‌ها، در اکثریت قریب به اتفاق متون دست‌نویس، مشکلاتی را نظیر هم‌پوشانی و یا اتصال اجزا در دو خط مجاور ایجاد می‌کند.

شکل ۱-۳، تأییدی بر توصیفات فوق است. یکی از روش‌هایی که بصورت گسترده در جداسازی *مطلوب* نواحی، خطوط، کلمات و حروف در متون تایپی مورد استفاده قرار می‌گیرد، الگوریتم Docstrum (الگوریتم کادرها و مستطیل‌های محیطی[1] است) [5] و [6]. با توجه به شکل، این روش در جداسازی خطوط موجود در اسناد دست‌نویس، به نتیجه غیرقابل قبول نشان داده شده منجر می‌شود. علت این امر را علاوه بر دست‌نویس بودن اسناد، باید در ویژگی‌های خاص نوشتار زبان فارسی نیز، جستجو کرد که در دست‌نویس شدن این زبان، این ویژگی‌ها، موضوع اصلی را با

[1] Freestyle
[2] Touch
[3] Overlap

چالش‌های زیادی مواجه می‌کنند. این شاخص‌ها، خاص زبان‌های شبه عربی مانند فارسی هستند و در ادامه بطور جامع بیان خواهند شد.

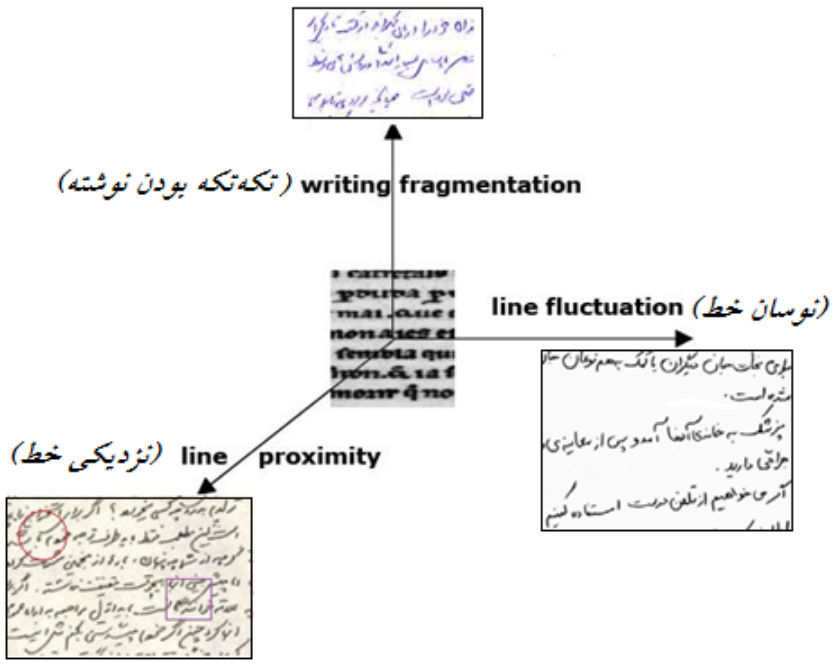


شکل ۱-۳ (الف) تصویر یک سند تایپی. (ب) نتیجه جداسازی خطوط. (ج) تصویر یک سند دست‌نویس. (د) نتیجه جداسازی خطوط.

[1] Bounding Box

۱-۶) چالش‌های جداسازی

مشکلات جداسازی خطوط در متون دست‌نویس را، می‌توان بصورت خلاصه در شکل زیر مشاهده کرد [7].



شکل ۱-۴ پیچیدگی در اسناد دست‌نویس

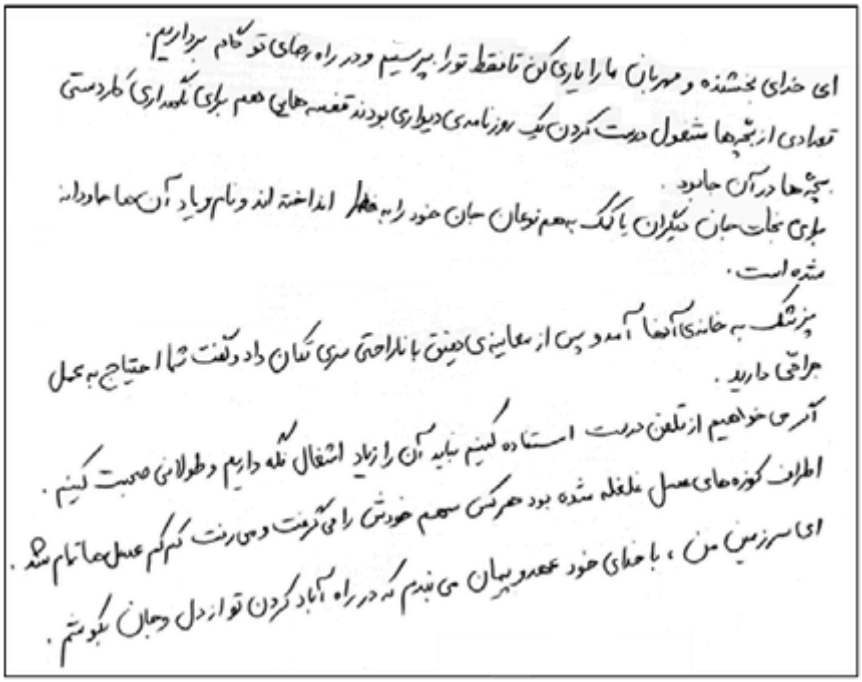
۱-۶-۱) نوسان و انحنای خطوط

خطوط یک متن دست‌نویس عموماً مستقیم نیست. در حالت کلی سه نوع انحنا در متن وجود دارد:

- انحنای سراسری: تمام قطعات صفحه، جهت کجی یکسانی دارند.
- انحنای چندگانه: پاراگراف‌های نامنظم با میزان کجی متفاوت در قطعه‌های مختلف صفحه سند که تنها می‌توان آن‌ها را بصورت غیر خطی نظم بخشید.
- انحنای خطوط (یا تغییر شیب خط): یک خط دارای چندین شیب متفاوت است، مانند خط منحنی‌وار.

باید دقت کرد که در اکثر موارد، خطوط در اسناد دست‌نویس هر دو نوع کجی دسته دوم و سوم را دارند، بعبارتی علاوه بر اینکه خط منحنی‌وار است و عموماً تخمین خطی و یا تکه‌ای- خطی در

مورد آن صحیح نیست، کجی یک فرمت مشخص ندارد و در خطوط متفاوت، جهت متفاوتی نسبت به خط‌های دیگر پیدا می‌کند. شکل ۱-۵ این واقعیت را نشان داده است.



شکل ۱-۵ یک صفحه دست‌نویس فارسی با کجی‌های متفاوت در خطوط مختلف

۱-۶-۲) نزدیکی و مجاورت خطوط

فاصله کم بین خطوط مجاور، سبب هم‌پوشانی یا تماس قسمت‌هایی از کلمات می‌شود که حروفشان نسبت به خط، دارای قسمت‌های بالارونده یا پایین‌آینده هستند.

۱-۶-۳) تکه‌تکه شدن دست‌نوشته

برخی حروف از بیش از یک جزء ساخته شده‌اند. حروف نقطه‌دار در زبان‌های فارسی و عربی (نقطه‌های جداکننده)، مثالی از این حروف هستند.

مشکلات اشاره شده، مشکلاتی هستند که در تمام اسناد دست‌نویس و مستقل از نوع زبان نوشتاری‌شان قابل مشاهده‌اند.

۱-۷) ویژگی‌های خط فارسی

پیچیدگی‌های موجود در اسناد دست‌نویس *فارسی*، حادتر و بارزتر خواهد بود. علت این امر، خصوصیات خاص نگارشی این زبان است. در این بخش برخی ویژگی‌ها و شاخص‌های نوشتاری فارسی بطور مختصر بیان می‌شود. شاید بتوان مهمترین ویژگی نوشتاری زبان فارسی را پیوستگی ذاتی و جهت نگارش آن از راست به چپ دانست که این زبان را از سایر زبان‌ها متمایز کرده است. در ادامه تعدادی از ویژگی‌های مهم نوشتاری خط فارسی بطور خلاصه ذکر می‌گردد [۲] .

الف) اشکال متفاوت یک حرف

شکل حروف در زبان فارسی تابعی از محل قرار گرفتن آن‌ها در کلمه است و هر حرف بر حسب موقعیت‌های مختلف در کلمه (اول، آخر و جدا) می‌تواند اشکال مختلفی داشته باشد (جدول۱-۲). بعنوان مثال، شکل حرف (م) در حالت جدا، باید به شکل (م) و در ابتدا یا وسط کلمه، به شکل (مـ)، نوشته شود. شکل این حرف در انتهای کلمه و در یک متن دست‌نویس ممکن است موجب هم‌پوشانی و یا تماس بین دو خط متوالی گردد، که عمل جداسازی دو خط را با مشکل مواجه می‌کند. بعنوان مثال تداخل حرف (م) در حالت دست‌نویس با سرکج حروف (ک) یا (گ) و یا مد (آ) در خط متوالی بعدی رخ می‌دهد.

ب) نقطه‌دار بودن حروف

بیش از نیمی از حروف فارسی نقطه‌دار هستند. بعبارت دقیق‌تر، ۱۰ حرف دارای یک نقطه، ۳ حرف دارای دو نقطه و ۵ حرف دارای سه نقطه هستند. بحث مهم نقطه‌دار بودن کلمات نیست بلکه محل قرار گرفتن نقاط است که مشکل‌ساز می‌شود. مثلاً با توجه به دو حرف {ب ، ن} یا {چ ، ژ} می‌توان به وجود این مشکل در متون دست‌نویس فارسی پی برد که قرار گرفتن حداقل یک نقطه بین دو خط متوالی، امری کاملاً طبیعی است ولی تشخیص اینکه نقطه مربوط به خط بالایی است یا پایینی، تصمیمی است که روند جداسازی خطوط را سخت می‌کند.

ج) وجود اِعراب

در نوشتار فارسی، هنگامی‌که احتمال تلفظ اشتباه یک کلمه وجود داشته باشد، صدا بصورت اعراب به بالا یا پایین برخی از حروف اضافه می‌شود. قرار گرفتن اعراب بین خطوط نیز، احتمال بروز مشکلی نظیر وجود نقطه‌ها را که در فوق توضیح داده شد، بوجود می‌آورند. علاوه بر اعراب، در

نوشتار فارسی علائمی از قبیل تشدید، تنوین، همزه و مد (در صورت نوشتاری حرف آ) نیز وجود دارند که جدا کردن خطوط را در متون دست‌نویس مشکل می‌سازند. علت بروز این مشکل علاوه بر سخت بودن تشخیص تعلق علائم به کدام خط (خط بالا یا خط پایین)، می‌تواند ناشی از تماس حروف موجود روی خطوط بالا یا پایین با این علائم باشد.

جدول ۱-۲ شکل‌های مختلف حروف فارسی با توجه به محل قرار گرفتنشان در زیرکلمه

	مجزا	ابتدا	میان	انتها		مجزا	ابتدا	میان	انتها
۱	ا یا آ	ا یا آ	---	ـا	۱۷	ص	صـ	ـصـ	ـص
۲	ب	بـ	ـبـ	ـب	۱۸	ض	ضـ	ـضـ	ـض
۳	پ	پـ	ـپـ	ـپ	۱۹	ط	طـ	ـطـ	ـط
۴	ت	تـ	ـتـ	ـت	۲۰	ظ	ظـ	ـظـ	ـظ
۵	ث	ثـ	ـثـ	ـث	۲۱	ع	عـ	ـعـ	ـع
۶	ج	جـ	ـجـ	ـج	۲۲	غ	غـ	ـغـ	ـغ
۷	چ	چـ	ـچـ	ـچ	۲۳	ف	فـ	ـفـ	ـف
۸	ح	حـ	ـحـ	ـح	۲۴	ق	قـ	ـقـ	ـق
۹	خ	خـ	ـخـ	ـخ	۲۵	ک	کـ	ـکـ	ـک
۱۰	د	---	---	ـد	۲۶	گ	گـ	ـگـ	ـگ
	مجزا	ابتدا	میان	انتها		مجزا	ابتدا	میان	انتها
۱۱	ذ	---	---	ـذ	۲۷	ل	لـ	ـلـ	ـل
۱۲	ر	---	---	ـر	۲۸	م	مـ	ـمـ	ـم
۱۳	ز	---	---	ـز	۲۹	ن	نـ	ـنـ	ـن
۱۴	ژ	---	---	ـژ	۳۰	و	---	---	ـو
۱۵	س	سـ	ـسـ	ـس	۳۱	ه	هـ	ـهـ	ـه
۱۶	ش	شـ	ـشـ	ـش	۳۲	ی	یـ	ـیـ	ـی

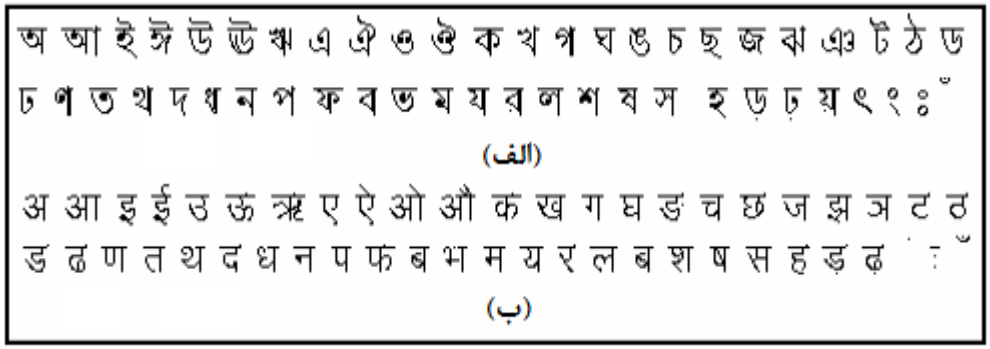
د) تنوع فراوان در شیوه‌های نگارشی

در مقایسه با نوشتار لاتین، فونت‌های فارسی در حالت چاپی دارای واحدهای نوشتاری بسیار زیاد (شامل ارقام، اشکال مختلف حروف، علائم خاص، اجزاء کوچک دارای اهمیت مانند دندانه) و در حالت دست‌نویس دارای سبک‌های متعدد می‌باشد. امری کاملاً واضح است که تعدد سبک نوشتاری اسناد دست‌نویس، فرآیند جداسازی را دشوارتر می‌سازد.

شکل زیر نیز نمونه‌هایی از ویژگی‌های خط نوشتاری فارسی را بخوبی بیان می‌کند.

شکل ۱-۶ ویژگی‌های خط نوشتاری فارسی

علاوه بر آنچه تاکنون ذکر شد، می‌توان به این نکته توجه کرد که در برخی زبان‌ها، وجود ویژگی-های خاص برای حروف الفبای آن زبان، جداسازی خطوط را بسیار آسان می‌کند. مثلاً شکل ۱-۷ دو نمونه از این زبان‌ها را نشان می‌دهد که پیدا کردن خط افقی موجود در قسمت فوقانی حروف، منجر به اجرای الگوریتم موفقی بمنظور جداسازی خطوط، خواهد بود اما با توجه به تنوع اشکال حروف در زبان فارسی، پیداکردن چنین نشانه‌ای که تمام حروف را پوشش دهد امری سخت و بدون اغراق تقریباً ناممکن است.



شکل ۱-۷ وجود یک خط افقی در اکثریت حروف (الف) bangla و (ب) devnagari (از زبان‌های هندی)