
بسم الله الرحمن الرحيم

دانشگاه صنعتی شاهرود

دانشکده مهندسی برق و رباتیک

گروه الکترونیک

بازشناسی متن تایپی نوشته شده با قلم Iranian sans

دانشجو: زینب باقری

استاد راهنما:

جناب آقای دکتر خسروی

پایان نامه ارشد جهت اخذ درجه کارشناسی ارشد

زمستان ۱۳۹۲

تقدیم به:

همسر دلسوز و فداکارم

و

پدر و مادر مهربانم

تشکر و قدردانی

با نثار تشکر و قدردانی بی‌انتها از زحمات و راهنمایی‌های صبورانه استاد ارجمند، جناب آقای دکتر حسین خسروی، دوره کارشناسی ارشد برای من فرصتی بود که در محضر این بزرگوار و سایر اساتید گران‌قدر، درس انسانیت، درستی و تحقیق را فراگیرم.

تعهد نامه

اینجانب زینب باقری دانشجوی دوره کارشناسی ارشد رشته برق دانشکده برق و رباتیک دانشگاه صنعتی شاهرود نویسنده پایان نامه بازشناسی متن تایپی نوشته شده با قلم Iranian sans تحت راهنمائی جناب آقای دکتر حسین خسروی متعهد می‌شوم.

- تحقیقات در این پایان نامه توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است.
- در استفاده از نتایج پژوهش‌های محققان دیگر به مرجع مورد استفاده استناد شده است.
- مطالب مندرج در پایان نامه تاکنون توسط خود یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ جا ارائه نشده است.
- کلیه حقوق معنوی این اثر متعلق به دانشگاه صنعتی شاهرود می‌باشد و مقالات مستخرج با نام « دانشگاه صنعتی شاهرود » و یا « Shahrood University of Technology » به چاپ خواهد رسید.
- حقوق معنوی تمام افرادی که در به دست آمدن نتایج اصلی پایان نامه تأثیرگذار بوده‌اند در مقالات مستخرج از پایان نامه رعایت می‌گردد.
- در کلیه مراحل انجام این پایان نامه ، در مواردی که از موجود زنده (یا بافت‌های آن‌ها) استفاده شده است ضوابط و اصول اخلاقی رعایت شده است.
- در کلیه مراحل انجام این پایان نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته یا استفاده شده است اصل رازداری ، ضوابط و اصول اخلاق انسانی رعایت شده است.

.

تاریخ

امضای دانشجو

مالکیت نتایج و حق نشر

- کلیه حقوق معنوی این اثر و محصولات آن (مقالات مستخرج، کتاب، برنامه های رایانه ای، نرم افزار ها و تجهیزات ساخته شده است) متعلق به دانشگاه صنعتی شاهرود می باشد. این مطلب باید به نحو مقتضی در تولیدات علمی مربوطه ذکر شود.
- استفاده از اطلاعات و نتایج موجود در پایان نامه بدون ذکر مرجع مجاز نمی باشد.

چکیده

در دهه‌های اخیر تحقیقات گسترده‌ای در زمینه‌ی بازشناسی الگوهای نوشتاری شامل حروف، ارقام و سایر نمادهای متداول در اسناد مکتوب، به زبان‌های مختلف انجام شده است. با توجه به پیشرفت‌های حاصل شده در این زمینه، فناوری بازشناسی خودکار متون تحت عنوان بازشناسی نوری حروف یا اُ.سی.آر شکل گرفته است. بازشناسی متن یکی از مهم‌ترین بخش‌های دولت الکترونیک به شمار می‌رود و در سال‌های اخیر در کشور ما نیز تقاضا برای یک سیستم بازشناسی متن فارسی، به شدت افزایش یافته است. با توجه به آنکه حجم زیادی از اسناد کاغذی موجود، توسط اسکنرها یا دوربین‌ها به اسناد تصویری دیجیتالی تبدیل می‌شوند؛ ذخیره‌سازی، بازیابی و مدیریت کارآمد این اسناد تصویری، در بسیاری از برنامه‌ها نظیر اتوماسیون اداری و کتابخانه‌های دیجیتالی اهمیت دارند.

به طورکلی سامانه بازشناسی متن شامل بخش‌های مختلفی از قبیل دریافت تصویر، پیش‌پردازش، آنالیز پیکربندی، تشخیص زبان، تشخیص قلم و در نهایت تشخیص متن می‌باشد. تحقیقات انجام شده در بعضی از این زمینه‌ها از قبیل پیش‌پردازش مستقل از زبان متن بوده و برای هر زبانی قابل استفاده است. لیکن برخی از قسمت‌های دیگر مانند تشخیص قلم و تشخیص متن به زبان متن وابسته بوده و نمی‌توان به طور مستقیم نتایج تحقیقات انجام شده برای سایر زبان‌ها را برای فارسی اعمال کرد. بیشتر تحقیقات انجام شده در زمینه‌ی بازشناسی متون فارسی روی تصاویری با درجه تفکیک زیاد، تصاویر متنی تمیز و غیرواقعی و شناسایی متن با چند قلم معروف بوده است. در تحقیقات انجام شده برای بازشناسی متون فارسی سه رویکرد عمده مبتنی بر جداسازی حروف، مبتنی بر بازشناسی شکل کلی زیرکلمات و روش ترکیبی وجود دارد.

در این پایان‌نامه هدف، بازشناسی متن تایپی نوشته شده با قلم Iranian sans، با حداقل اندازه ۹ و درجه تفکیک ۳۰۰ نقطه بر اینچ است. این قلم با توجه به دو خصوصیت زیبایی و خوانایی، بسیار مورد توجه واقع شده و روز به روز بر حجم استفاده از آن در محیط رایانه و اینترنت افزوده می‌شود. این قلم

قابلیت جایگزینی با قلم تاهما که قلم پیش‌فرض سیستم عامل ویندوز است، را دارد. علیرغم خوانایی، فاصله استاندارد بین سطرها، زیبایی و سازگاری با لاتین، این قلم دارای پیچیدگی ساختاری خاصی بوده که این خود عمل بازشناسی آن را پیچیده می‌نماید.

در این پایان‌نامه ابتدا با تولید پایگاه داده مناسب، به آموزش دو طبقه‌بند برای حروف گسسته و پیوسته پرداخته شده و سپس با رفع مشکل همپوشانی زیرکلمات، از رویکرد مبتنی بر جداسازی برای جداسازی حروف استفاده می‌شود. طبقه‌بندهای مورد استفاده از نوع شبکه عصبی چند لایه می‌باشند. در نهایت، نتایج کارایی سامانه‌ی مذکور برای پردازش چند تصویر با متن چاپی، ارائه می‌شود که در آن در بخش جداسازی دقت ۹۶٪ و در بخش شناسایی دقت ۸۵٪ حاصل شد.

کلمات کلیدی: بازشناسی متن، قلم Iranian sans، رویکرد مبتنی بر جداسازی، طبقه‌بند شبکه عصبی

فهرست مطالب

عنوان صفحه

فهرست شکل‌ها

فهرست جدول‌ها

عنوان **صفحه**

فصل اول

مقدمه

۱-۱-مقدمه

در گذشته تنها وسیله ارتباطی بین مردم برای ثبت وقایع یا اطلاع‌رسانی، کاغذ بوده است. در نتیجه نسخ کاغذی بسیاری وجود داشت که مدیریت، بایگانی و جستجو در میان آن‌ها کار دشواری بود. علاوه بر این بسیاری از نسخ مهم و ارزشمند به دلایل مختلفی از جمله حوادث طبیعی از بین می‌رفتند و امکان بازیابی مجدد آن‌ها وجود نداشت. کم‌کم با پدید آمدن رایانه و آشنایی با قابلیت‌های این محیط الکترونیکی لزوم استفاده از سیستمی که بتواند اطلاعات کاغذی را به رایانه منتقل کرده و در آن‌ها امکان ویرایش به وجود بیاورد روز به روز بیشتر شد.

امروزه نیز مستندات کاغذی به دو صورت چاپی (شامل روزنامه‌ها، کتاب‌ها، مقالات و...) و دست‌نویس وجود دارد که تنوع فراوان، موارد استفاده، لزوم بایگانی و تهیه‌ی آرشیوها باعث انجام تحقیقات گسترده‌ای در زمینه بازشناسی متون شده است.

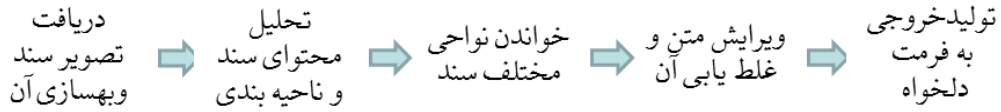
در این فصل ابتدا به معرفی سیستم بازشناسی نوری حروف پرداخته می‌شود و سپس ضمن بیان کاربردهای آن، تقسیم‌بندی این سیستم‌ها از نظر نوع ورودی و نحوه‌ی ورود اطلاعات مطرح شده و در آخر ویژگی‌های زبان فارسی و محدودیت‌های آن بیان می‌گردد.

۱-۱-۱- معرفی بازشناسی نوری حروف

اُ.سی.آر مخفف بازشناسی نوری نویسه‌ها و یا نویسه‌خوان نوری است. یک روبشگر، تصویر را برای سامانه‌ی اُ.سی.آر فراهم می‌کند، سپس این سامانه نواحی مختلف تصویر را شناسایی کرده و به فرمت مناسب ذخیره می‌نماید. در نتیجه یک فایل تصویری، به متن قابل ویرایش و جستجو تبدیل شده و حجم آن نیز کاهش می‌یابد. در فرآیند اُ.سی.آر ورودی، تصویر سند مورد نظر و خروجی آن فایل دیجیتالی قابل ویرایش از اطلاعات سند مثلاً به فرمت وورد[1] می‌باشد. شکل ۱-۱ فرآیند اُ.سی.آر را با

[1] Word

جزئیات بیشتری نشان می‌دهد. در این شکل مراحل مختلف اُ.سی.آر تشریح شده است.



شکل ۱-۱: مراحل مختلف اُ.سی.آر

در این فرایند پس از دریافت تصویر سند، قسمت‌های مختلف تصویر از نظر معنایی و فیزیکی ناحیه‌بندی شده، بعد از شناسایی نواحی مختلف، ویرایش و غلط‌یابی صورت گرفته و سپس خروجی به فرمت دلخواه تولید می‌شود. در زمینه‌ی متون چاپی در زبان لاتین تحقیقات زیادی انجام شده ولی به علت تفاوت‌های عمده‌ای که در ساختار نوشتاری فارسی با لاتین وجود دارد (مثلاً در مراحلی مانند تشخیص قلم[1] و جداسازی) نمی‌توان آن نتایج را مستقیماً برای زبان فارسی به کار برد.

۱-۲-کاربردهای اُ.سی.آر

در ادامه برخی از کاربردهای سیستم‌های بازشناسی متن مطرح می‌شود.

۱-۲-۱- دستگاه‌های چک‌خوان

از جمله کاربردهای سیستم‌های بازشناسی متن، دستگاه‌های چک‌خوان هستند که در بانک‌ها برای ورود داده استفاده می‌شوند. این سیستم‌ها برای خواندن بخش‌های کوچکی از تصویر چک شامل نویسه های چاپی و دست‌نویس مانند شماره حساب، امضای مشتری، شماره برگ چک و مبلغ چک طراحی می‌شوند. ساختار کاغذ چک معمولاً ثابت است و نواحی مشخصی از آن توسط ماشین خوانده می‌شود. دستگاه‌های چک‌خوان در حدود ۱۵۰۰۰ چک را در یک ساعت می‌خوانند و نرخ خطای آن‌ها کم است. این سیستم‌ها در کیفیت‌های پایین چاپ نیز به خوبی عمل می‌کنند.

[1] Font

۱-۲-۲- ورود متن در سیستم‌های اتوماسیون اداری

از جمله نیازهای اساسی در سیستم‌های اتوماسیون اداری دسترسی به امکان ویرایش، اصلاح و جستجو کردن در سندهای اسکن شده می‌باشد که سیستم بازشناسی متن نقش مهمی در این امر دارد. این سیستم‌ها اسناد با قلم‌های محدود و کیفیت چاپ مشخصی را بازشناسی می‌کنند. این سیستم‌ها برای ورود حجم زیادی از متن به شکل قابل پردازش با نرم‌افزارهای پردازشگر متن استفاده می‌شوند. قابلیت این سیستم‌ها در ورود متن، رقیب قدرتمندی برای وارد کردن دستی اطلاعات است. کارایی این سیستم‌ها به کیفیت چاپ اسناد وابستگی زیادی دارد.

۱-۲-۳- خودکارسازی فرایند

در این کاربرد، کنترل یک فرایند مشخص، به بازشناسی دقیق متن نوشته شده اولویت دارد. به عنوان مثال در خودکارسازی فرایند جداسازی نامه در مراکز پستی، بازشناسی نام شهر مقصد، از بازشناسی دقیق کل متن نوشته شده بر روی پاکت نامه اولویت بالاتری دارد. نرخ بازشناسی این سیستم‌ها کاملاً به کیفیت تصویر پاکت نامه وابسته است. سرعت جداسازی نامه‌ها بر حسب آدرس، معمولاً در حدود چند ده هزار نامه در ساعت است.

۱-۲-۴- بازیابی اسناد[1]

هزینه‌ی بالای کار با پایگاه‌های داده بزرگ از تصاویر اسناد، نیاز به روش‌های خودکار و کارآمد برای دستیابی به اطلاعات داخل این تصاویر را به وجود آورده است. در تلاش برای حرکت به سوی ادارات بدون کاغذ، حجم بزرگی از اسناد کاغذی روبش شده و به صورت تصویری ذخیره می‌شوند، بدون اینکه برچسب اطلاعاتی مناسبی به آن‌ها زده شود یک راه برای برچسب‌زنی و بازیابی پایگاه داده، تبدیل کامل اسناد به شکل الکترونیکی است که می‌توانند به طور خودکار برچسب زده شوند. موانعی مثل

1 Document Retrieval

کیفیت پایین تصویر و هزینه بالای تبدیل در مقابل این روش وجود دارد. از طرفی چون بسیاری از اجزاء غیر متن را نمی‌توان به طور کامل در شکل تبدیل شده نمایش داد، داشتن یک کپی از سند در شکل تصویری لازم به نظر می‌رسد. نگهداری و دسترسی به یک پایگاه داده متنی مسئله‌ی مهمی است. در چنین پایگاهی، مستند جدیدی که وارد می‌شود لزوماً ساختاری مشابه مستندات قبلی ندارد، بنابراین تعریف ساختارهای مختلف از قبل، غیر ممکن است. توقعی که کاربران از یک سیستم پایگاه داده مستندات دارند، بازیابی دقیق مستند مورد نظر نیست، اما انتظار می‌رود که مکانیزمی برای بازیابی مستندات، به صورت مرتب شده بر اساس شباهت به مستند مورد پرس و جو، وجود داشته باشد. برای بازیابی اسناد مبتنی بر محتوا، می‌توان بلوک‌های متن تصویر سند را بازشناسی کرده و از آن‌ها برای بازیابی سند استفاده کرد [۱].

۱-۲-۵- کتابخانه دیجیتال

در یک کتابخانه‌ی دیجیتال برای دسترسی به کتاب مورد نظر از کلید واژه‌ها استفاده می‌شود. کلید واژه‌ها به صورت متن وارد می‌شوند و نرم افزار کتابخانه، کتاب‌های مرتبط با آن‌ها را بازیابی می‌کند. پیدا کردن کلمات کلیدی از تصاویر روبش شده‌ی کتاب‌ها، هسته‌ی اصلی نرم افزار جستجوی کتابخانه است.

۱-۲-۶- کتاب‌های الکترونیکی

با فناوری اُ.سی.آر تصاویر صفحات کتاب به متن رایانه‌ای تبدیل شده و امکان ویرایش متن، جست‌وجوی متن، چاپ مجدد متن با کیفیت بالا و انتقال الکترونیکی سریع آن فراهم می‌شود. اُ.سی.آر به ویژه برای دیجیتالی نمودن منابع و نسخ خطی بسیار مفید است، چرا که برای نسخ خطی، باید اصالت منبع با حفظ شکل و قالب اصلی متون و نه به صورت تایپ شده حفظ شود.

۱-۲-۷- کنترل ترافیک

برای تشخیص پلاک خودروها می‌توان از سیستم بازشناسی متن استفاده کرد. صحنه‌ی ترافیک

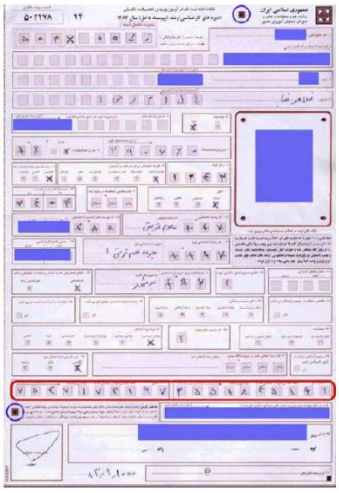
معمولاً با دوربین‌های سریع تصویربرداری می‌شود. از محدود بودن حروف استفاده شده در پلاک‌ها، می‌توان برای بهبود نتایج استفاده کرد. پس از جدا شدن تک تک عناصر موجود در پلاک، عناصر به صورت مجزا به سیستم بازشناسی متن جهت خواندن ارجاع داده شده و خوانده می‌شود.

۱-۲-۸- تبدیل متن به صحبت

از یک سیستم بازشناسی متن در کنار یک سیستم پردازش صوت می‌توان در کاربردهای مختلف مانند کمک به نابینایان به منظور درک متون چاپی و خودکارسازی پاسخ به ارباب رجوع در ادارات استفاده کرد.

۱-۲-۹- فرم خوان‌ها

این سیستم‌ها همان سیستم شناسایی متون دست‌نویس گسسته است که برای خواندن فرم‌های خاص طراحی می‌شوند. این سیستم‌ها معمولاً از روبشگرهایی که رنگ خاصی را حذف می‌کنند استفاده کرده و به هنگام اسکن رنگ زمینه را حذف می‌کنند. حروف و ارقام به صورت مجزا در خانه‌هایی با رنگ مورد نظر نوشته می‌شوند و روبشگر هنگام روبش فقط این نواحی را اسکن کرده و به صورت خودکار رنگ زمینه‌ی این خانه‌ها را حذف می‌کند. بنابراین این سیستم‌ها نرخ بازشناسی بالایی دارند. درشکل ۱-۲ نمونه ای از یک فرم ثبت نام کنکور کارشناسی ارشد ارائه شده است.



شکل ۱-۲: نمونه ای از یک فرم ثبت نام کنکور کارشناسی ارشد[۲]

۱-۲-۱۰- نقشه خوانی خودکار

بازشناسی نویسه‌های روی نقشه‌ها با مسائلی مانند ادغام حروف با خطوط نقشه، چاپ شدن متن با زوایای مختلف، متون با قلم‌های متعدد، متون چاپ شده بر زمینه رنگی و وجود متن دست‌نویس روبرو است.

۱-۲-۱۱- نت خوان‌ها

بازشناسی نت‌های موسیقی یکی دیگر از کاربردهای سیستم بازشناسی متن است. نت‌های موسیقی به دو صورت چاپی و دست‌نویس وجود دارند.

۱-۳-انواع اُ.سی.آر

سیستم‌های بازشناسی عمدتاً از دو منظر نوع ورودی و نحوه‌ی ورود اطلاعات مورد بررسی و طبقه‌بندی قرار می‌گیرند.

۱-۳-۱- از نظر نوع ورودی

در یک تقسیم‌بندی کلی می‌توان سیستم‌های اُ.سی.آر را از لحاظ نوع الگوی ورودی به دو گروه اصلی تقسیم کرد:

الف -سیستم‌های بازشناسی متون چاپی

ب -سیستم‌های بازشناسی متون دست‌نویس

در سیستم‌های بازشناسی متون چاپی، ورودی به صورت تصویری است که از دوربین دیجیتال یا اسکنر به دست آمده است. ولی سیستم شناسایی متون دست‌نویس شامل تصویر متون نوشته شده است که این متون به دو بخش گسسته و پیوسته تقسیم می‌شود. متون دست‌نویس گسسته شامل متونی است که در آن حروف جدا از هم نوشته می‌شوند مانند حروفی که در فرم‌های آزمون‌ها در کادرهایی به صورت مجزا نوشته می‌شود، ولی متون دست‌نویس پیوسته در قالب نامه یا متن یا گزارش

توسط شخص روی کاغذ نوشته می‌شود. متون دست‌نویس پیوسته به علت مشکلاتی از قبیل کشیده یا شکسته نوشته شدن بعضی از حروف یا به طور کلی نوشتاری که مطابق قواعد نگارش رسم‌الخط فارسی نیست، هنوز به صورت یک سیستم نرم افزاری جامع درنیامده است و همچنان تلاش در ایجاد سیستم شناسایی متون دست‌نویس پیوسته با دقت قابل قبول وجود دارد. ولی برای شناسایی متون دست‌نویس گسسته به علت گسسته بودن حروف و همچنین متون تایپی به علت اینکه عموماً از قواعد خاصی پیروی می‌کنند سیستم‌های نرم افزاری زیادی وجود دارند، که بسیاری از آن‌ها با دقت بالایی عملیات شناسایی را انجام می‌دهند.

۱-۳-۲- تقسیم‌بندی از لحاظ نحوه ورود اطلاعات

از جنبه‌ی نحوه‌ی ورود اطلاعات، سیستم‌های اُ.سی.آر به دو دسته‌ی سیستم‌های برخط و سیستم‌های برون‌خط تقسیم‌بندی می‌شوند.

۱-۳-۲-۱ سیستم‌های برخط[1]

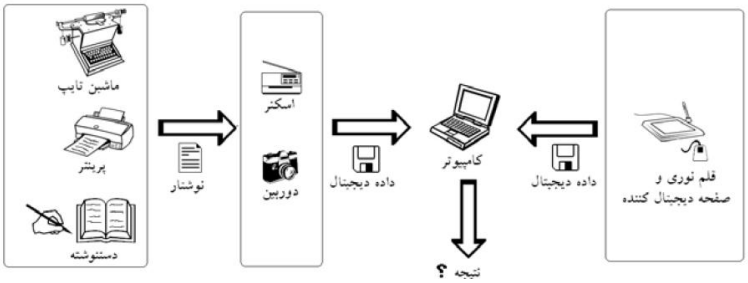
ورودی سیستم بازشناسی برخط، به صورت پیوسته همزمان با نوشتن نویسنده بر روی یک صفحه رقمی‌کننده به سیستم وارد می‌شود. بازشناسی برخط فقط در بازشناسی دست‌نوشته کاربرد دارد. بازشناسی برخط نوشتار به دلیل راحت‌تر بودن نوشتن از تایپ کردن، عدم امکان تایپ در بعضی مکان‌ها، عدم وجود یک صفحه کلید کامل روی رایانه‌های کوچک و سخت بودن تایپ نویسه‌ها در بعضی زبان‌ها به دلیل تعداد زیاد آن‌ها، مورد توجه خاصی قرارگرفته است. در روش‌های برخط از داده های یک بعدی استفاده می‌شود. در این سیستم‌ها می‌توان قسمت‌های مختلف متن را ذخیره کرده و سپس عملیات شناسایی را توسط روش‌های برون خطی انجام داد. برای بازشناسی دست‌نوشته‌ی برخط

1 Online

از ویژگی‌هایی مانند مختصات مکانی حرکت قلم روی صفحه، سرعت نوشتن و میزان فشار قلم بر صفحه استفاده می‌شود.

۱-۳-۲-۲سیستم‌های برون‌خط[1]

در سیستم بازشناسی برون‌خط، تصویر اسکن شده به عنوان ورودی در نظر گرفته می‌شود. بازشناسی برون‌خط هم شامل بازشناسی نوشتار چاپی و هم نوشتار دست‌نویس می‌شود. روش‌های شناسایی برون‌خط شامل کاربرد شبکه عصبی، منطق فازی و غیره می‌باشد که بر روی تصاویر دوبعدی اعمال می‌شوند. سیستم برون‌خط دقت کمتری از سیستم‌های برخط دارند. نرخ بازشناسی برای سیستم‌های برخط بالاتر از این نرخ در سیستم برون‌خط گزارش شده است.



شکل ۱-۳ :سیستم‌های شناسایی برون‌خط و برخط

۱-۴- موانع موجود بر سر تکامل اُ.سی.آر فارسی

در دهه‌های اخیر تحقیقات زیادی در زمینه‌ی بازشناسی ارقام و حروف به زبان‌های مختلف، و از جمله زبان فارسی انجام شده است. بیشتر تحقیقات انجام شده در زمینه‌ی بازشناسی متون چاپی فارسی با محدودیت‌هایی از جمله درجه تفکیک زیاد (۴۰۰ و بیشتر)، استفاده از تصاویر متنی تمیز و غیر واقعی و شناسایی متن با چند قلم خاص مشابه همراه بوده است، هر چند در سال‌های اخیر تا حد امکان بسیاری از این محدودیت‌ها برطرف شده است. پیچیدگی‌های زبان فارسی، عدم سرمایه‌گذاری

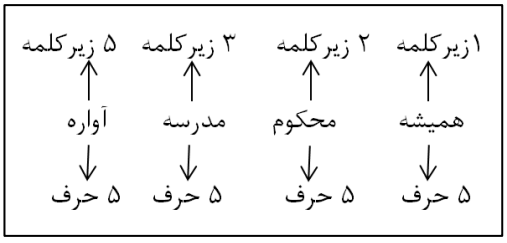
1 Offline

کافی، عدم آشنایی شرکت‌ها با اُ.سی.آر، تنوع قلم‌ها، فقدان لغت نامه و پایگاه داده‌ی جامع و استاندارد و وجود جمعیت کم کاربر زبان فارسی از جمله عواملی هستند که مانع ایجاد سیستم جامع بازشناسی متون فارسی می‌شود.

۱-۵- معرفی ویژگی‌های خط فارسی

آشنایی با قواعد نگارش فارسی در انتخاب روش‌های مناسب برای بازشناسی متون نقش اساسی دارد. در زیر قواعد مهم نگارش فارسی بیان می‌شود.

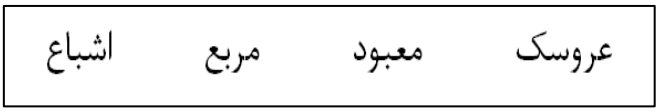
- خط فارسی بر خلاف لاتین از راست به چپ نوشته می‌شود. البته اعداد فارسی از چپ به راست نوشته می‌شوند.
- هر کلمه از یک یا چند زیرکلمه و هر زیرکلمه از یک یا چند حرف متصل به هم تشکیل می‌شود. مطابق قواعد نگارش زبان فارسی، باید بین کلمات فاصله وجود داشته باشد و بین زیرکلمات فاصله اضافه وجود نداشته باشد، که در بسیاری از موارد این قواعد رعایت نمی‌شود.



شکل ۱-۴ : آشنایی با مفهوم حرف و زیرکلمه

- در کلمات فارسی، بیشتر حروف از دو طرف به حروف مجاور خود می‌چسبند. با این حال حروف {ا، د، ذ، ر، ز، ژ، و} به حرف بعدی خود نمی‌چسبند و فقط می‌توانند از سمت راست به حرف قبلی خود وصل شوند.
- حروف فارسی بسته به موقعیت آن‌ها در زیرکلمه می‌توانند تا چهار شکل متفاوت داشته

باشند: حروف ابتدایی، میانی، انتهایی و مجزا. برای مثال شکل‌های مختلف حرف "ع" درشکل ۱-۵ نشان داده شده است.

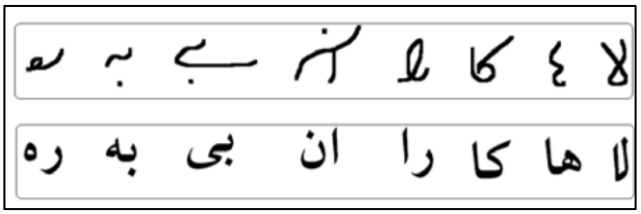


شکل ۱-۵ : شکل‌های مختلف یک حرف در کلمه

- بعضی از حروف فارسی نقطه دارند که تعداد نقاط این حروف بین یک تا سه متغیر است. نقاط ممکن است در بالا "ز"، پایین "ب" یا داخل بدنه‌ی "ج" حرف قرار گیرند. حروف "ک" و "گ" سرکش دارند. روی حرف الف علامت مد"آ" نیز قرار می‌گیرد. حروف "ط" و "ظ" نیز دسته دارند.
- قسمتی از حرف یا زیرکلمه که نقاط آن حذف شده باشد را بدنه‌ی حرف یا زیرکلمه می‌گویند. برخی از حروف مانند"ب پ ت ث" دارای بدنه‌ی یکسان هستند و تفاوت آن‌ها در تعداد نقاط و محل قرار گرفتن آن می‌باشد.
- اندازه‌ی تمام حروف فارسی یکسان نیستند. مثلاً حروف (ب) و (س) در حالت جدا، اندازه‌ی بزرگ‌تری نسبت به حروف (د) و (ه) دارند. این تنوع در اندازه‌ی حروف، کار قطعه‌بندی حروف را مشکل می‌کند.
- در برخی از شیوه‌های نوشتاری زبان فارسی، ترکیب شدن دو یا چند حرف کنار هم، شکلی ایجاد می‌کند که شباهتی به حروف تشکیل دهنده‌ی آن ندارد. به این ترکیب، حروف ادغام شده[1] می‌گویند. چنین مواردی نه‌تنها در نوشتار دست‌نویس، بلکه در متون تایپی در بعضی از قلم‌های خاص نیز وجود دارد. متداول‌ترین ترکیب در متون تایپی، ادغام دو حرف "ل" و"ا" به صورت "لا" است. در نوشته‌های دست‌نویس فارسی نیز با

1 Ligature

توجه به سلیقه‌ی نویسنده شکل بعضی از حروف کنار هم، به کلی تغییر می‌کند (شکل ۱-۶).

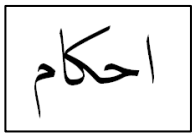


شکل ۱-۶: نمونه‌هایی از ادغام حروف مجاور در متون دست‌نویس

- حروف واقع در یک کلمه ممکن است همپوشانی داشته باشند، در این صورت نمی‌توان با رسم خطوط عمودی، حروف را به صورت کامل از یکدیگر مجزا نمود.

شکل ۱-۷ : همپوشانی حروف

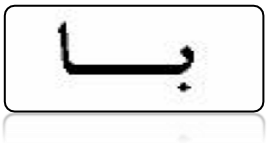
- در برخی از قلم‌ها بعضی از حروف، در دو محل (از یک سمت) به یکدیگر اتصال دارند.



شکل ۱-۸: اتصال حروف از دو محل

- حروف فارسی می‌توانند در بالا یا پایین بدنه‌ی خود دارای اعراب باشند. سه اعراب-َ-ِ-ُ در زبان فارسی اعراب‌های اصلی بوده و اعراب -ً در برخی کلمات عربی رایج در زبان فارسی دیده می‌شود. کلمات عربی دارای اعراب-ٌ و-ٍ در زبان فارسی عمومیت نیافته‌اند. هر چند کاربرد اعراب در زبان فارسی نسبت به زبان عربی بسیار محدودتر است، اما در مواردی که کلمه‌ای نامتداول باشد و یا به دلیل تشابه نگارشی آن با کلمه‌ی دیگر، تأکید بر تلفظ صحیح آن باشد، مورد استفاده قرار می‌گیرند.

- در بالای بدنه‌ی یک حرف ممکن است علامت تشدید وجود داشته باشد.
- برخی از حروف شامل همزه هستند (" ئا"،" أ"،" ؤ").
- حروفی که از طرف چپ قابلیت اتصال به حرف مجاور خود را دارند، می‌توانند به صورت کشیده نوشته شوند.



شکل ۱-۹ : اتصال حروف از سمت چپ

موارد مذکور نمایانگر پاره‌ای از مشکلات است که عمل بازشناسی را با محدودیت‌های زیادی مواجه می‌کند.

۱-۶- نتیجه‌گیری

در این فصل لزوم استفاده از سیستم‌های بازشناسی متن مطرح گردید و پس از تعریف مختصری درباره‌ی سیستم‌های بازشناسی متن، کاربردهای آن بیان شد. همچنین تقسیم‌بندی سیستم‌های بازشناسی متن از نظر نوع ورودی و نحوه‌ی ورود اطلاعات معرفی گردید. از طرفی با توجه به این که این پایان‌نامه در رابطه با بازشناسی متون فارسی است، به بررسی ویژگی‌های متون فارسی پرداخته شد و عوامل پیچیدگی بازشناسی زبان فارسی نیز ذکر گردید.

در فصل آتی به بررسی روند پیدایش سیستم‌های بازشناسی متن و تحقیقات انجام شده در زمینه‌ی اُ.سی.آر پرداخته خواهد شد.

فصل دوم

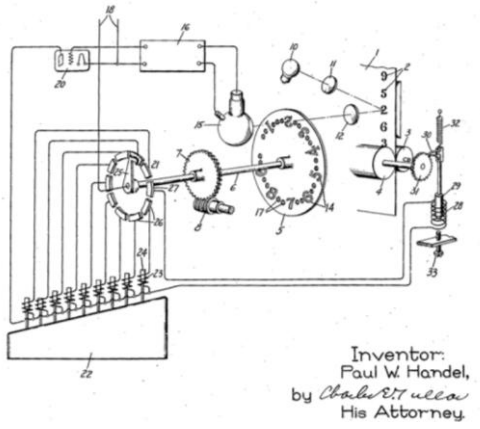
کارهای پیشین

۲-۱-مقدمه

در این فصل با آگاهی از مقدمات اولیه راجع به اُ.سی.آر و بیان ضرورت استفاده از آن در بسیاری از مواقع، به مروری بر سیر تحولی اُ.سی.آر پرداخته می‌شود که در آن نسل‌های مختلف اُ.سی.آر معرفی شده و روند تحقیقاتی آن معرفی می‌گردد. از آن جا که موضوع مورد بحث در مورد سیستم‌های بازشناسی متن به زبان فارسی است، به بررسی تحقیقاتی در زمینه‌ی اُ.سی.آر فارسی پرداخته می‌شود.

۲-۲-روند تحقیقاتی اُ.سی.آر

در ۱۹۰۰ میلادی، تیورین[1]، دانشمند روسی، وسیله‌ای برای خواندن متن برای نابینایان به ثبت رساند [۳]. در سال ۱۹۲۹ میلادی تاشک[2] در آلمان و در سال ۱۹۳۲ هندل[3] و همکارانش در آمریکا سامانه‌ی اُ.سی.آر خود را در زمینه‌ی سیستم بازشناسی حروف ثبت کردند (شکل ۲-۱) [۴]. روش آن‌ها، تطبیق الگو نام دارد و به این صورت کار می‌کند که به هر حرف، نور تابیده می‌شود و نور باز تابیده شده از حروف، از قالب‌های مکانیکی عبور داده می‌شود. هرگاه نوری از قالب عبور نکرد، حرف تشخیص داده می‌شود. این اختراع به دلیل تکنولوژی اپتومکانیکی مورد استفاده در آن کاربردی نبود.



شکل ۲-۱: سامانه اُ.سی.آرهندل [۵]

1 Tyurin

2 Tausheck

3 Handel

در دهه‌ی ۱۹۴۰ میلادی، اولین سامانه‌های کامپیوتری بازشناسی نویسه معرفی شدند. این سامانه‌ها به بازشناسی متون چاپی یا ارقام و حروف دست‌نویس مقید[1] می‌پرداختند.

در دهه‌ی ۱۹۵۰، سیستم‌های تجاری اُ.سی.آر برای خواندن ارقامی که با قلم خاص چاپ می‌شدند، به بازار آمدند. همچنین در همین دهه، اولین سامانه‌های تجاری برای بازشناسی نویسه‌های برخط، که با قلم و صفحه الکترونیکی خاص نوشته می‌شدند، مطرح شدند [۳].

در سال ۱۹۵۹ میلادی اولین اُ.سی.آر برای خواندن یک صفحه متن چاپی به بازار آمد که می‌توانست متون نوشته شده با یک قلم خاص در یک اندازه‌ی خاص را بخواند [۶].

در دهه‌ی ۱۹۷۰[۴] تمرکز بیشتر مقالات روی روش‌های استخراج ویژگی از نظر یافتن ویژگی‌های مناسب و استخراج آن‌ها به منظور تمایز بین کلاس‌ها بود که به کلاس‌های ارقام، حروف و علائم محدود می‌شدند. در آن زمان به علت محدود بودن امکانات پیاده‌سازی، استفاده از تعداد زیادی ویژگی میسر نبود و امکانات محدود پردازشی، تأثیری تعیین‌کننده بر روش‌های استخراج ویژگی داشت.

در دهه‌ی ۱۹۸۰ تحقیقات اُ.سی.آر به بازشناسی حروف و ارقام دست‌نویس غیرمقید و حتی متون دست‌نویس گسترش یافت و نتایج خوبی بدست آمد. تا چند دهه، تحقیقات تنها بر زبان‌های لاتین متمرکز بود و از اوایل دهه‌ی ۱۹۸۰ پژوهش‌ها در زمینه‌ی عربی و فارسی نیز آغاز شد. همچنین روش‌های طبقه‌بندی جدیدی مثل شبکه های عصبی و روش‌های فازی بکار گرفته شدند.

در دهه‌ی ۱۹۹۰ موضوعات تازه‌ای در پژوهش‌های مربوط به اُ.سی.آر مطرح شد. از جمله‌ی آن‌ها تحلیل سند است، به گونه‌ای که سیستم بازشناسی متن می‌توانست ساختار فیزیکی صفحه را بیابد و صفحاتی با ساختارهای پیچیده که شامل عکس، گرافیک و جدول هستند، به عنوان ورودی بپذیرد. از دیگر موضوعات، ایده‌ی استفاده از شکل کلی کلمات و بازشناسی با استفاده از آن‌ها در تصاویری بود که در مرحله‌ی جداسازی و شکستن کلمات به حروف مشکل جدی داشتند. موضوع دیگر، بازشناسی

1 Constrained

متون چاپی بدون توجه به نوع قلم آن‌ها بود که تا حد زیادی برای قلم‌های گوناگون میسر شد. در این دهه کارایی سامانه‌های اُ.سی.آر به حد قابل استفاده در کاربردهای واقعی رسید. روش‌های طبقه‌بندی توانمندتری به کار گرفته شدند و همچنین موضوع ترکیب نتایج حاصل از چند طبقه‌بند ساده به جای استفاده از یک طبقه‌بند قوی مطرح شد.

در دهه‌ی ۲۰۰۰ میلادی پیشرفت‌های زیادی حاصل شد. ازجمله ترکیب طبقه‌بندها که در دهه‌ی گذشته مطرح شده بود، با توجه به نتایج خوبی که از آن به دست آمد، مورد توجه بیشتری قرار گرفت. از دیگر موضوعات مطرح شده در این دهه می‌توان بهسازی و بازشناسی مستندات تاریخی و شناسایی افراد از روی دست‌نوشته‌های آنان را نام برد که پژوهش‌های زیادی در این زمینه انجام شده است. در زمینه‌ی بازشناسی متون، توجه بیشتری به بازشناسی گرافیک و همچنین نمادهای ریاضی شده است. در این دوره نتایج قابل توجهی درباره تشخیص ساختار یک صفحه از متن چاپی حاصل شد.

با گسترش کاربرد دوربین‌های دیجیتال و افزایش کیفیت تصاویر آن‌ها، استفاده از این دوربین‌ها به جای روبشگرها در بازشناسی مستندات مورد توجه قرار گرفت. تفاوت بین تصاویر حاصل از این دو دستگاه، موضوعات پژوهشی تازه‌ای را مطرح کرد. بازشناسی متن در تصاویر ویدیویی نیز مورد توجه قرار گرفت. با رواج استفاده از تصاویر رنگی، موضوع به‌کارگیری اطلاعات رنگ در بازشناسی متون نیز مطرح شد. بازیابی تصویری مستندات[1] را می‌توان یکی از مهم‌ترین موضوعات مطرح در سال‌های اخیر دانست. در یک سامانه‌ی بازیابی تصویر مستندات، می‌توان با روش‌های مختلف پرس و جو، مستند مورد نظر کاربر را در یک بانک تصویری بزرگ پیدا کرد. سابقه این موضوع به بازیابی اطلاعات و بازیابی تصویر برمی‌گردد. بازیابی تصویر از مهم‌ترین موضوعات پردازش تصویر در سال‌های اخیر است. امروزه سیستم‌های اُ.سی.آر قادر به تشخیص دقیق کاراکترهای تایپی لاتین با انواع قلم‌ها و در اندازه‌های متفاوت هستند و شناسایی متون دست‌نویس لاتین و یا قلم‌هایی که در آن‌ها از خطوط خمیده استفاده

1 Document Image Retrieval

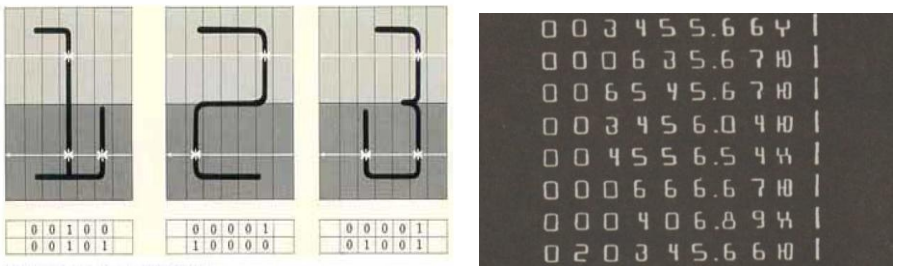
می‌شود (مثل فارسی و عربی)، رو به پیشرفت است.

۲-۲-۱- نسل‌های مختلف نویسه خوان‌های نوری

سیستم‌های تجاری عرضه شده در زمینه‌ی اُ.سی.آر را می‌توان به چهار نسل تقسیم‌بندی نمود.

۲-۲-۱-۱سیستم‌های نسل اول

این نویسه خوان‌ها در اوایل دهه‌ی ۱۹۶۰ میلادی به بازار آمدند. این سیستم‌ها فقط قابلیت تشخیص کاراکترهای خاص با اندازه و قلم مشخصی را داشتند. در آن‌ها از روش‌هایی استفاده شده بود که نسبت به تغییر جای کاراکترها، اندازه و دوران آن‌ها فوق‌العاده حساس بودند. بارزترین این نویسه خوان‌ها NCR420 بود که می‌توانست اعداد و پنج نماد دیگر را که با قلم خاص این نویسه خوان نوشته می‌شدند، بخواند. این قلم NOF یا Bicodes نامیده می‌شد.



شکل ۲-۲: قلم **NOF** یا **Bicodes** [۷]

۲-۲-۱-۲ سیستم‌های نسل دوم

این سیستم‌ها در اواسط دهه‌ی ۱۹۶۰ به بازار آمدند و فقط قادر به شناسایی دست‌نویس‌های مجزا مانند ارقام بودند. یک نویسه خوان شاخص از نسل دوم رتینا[1] است. این سامانه می‌توانست مجموعه‌ی اعداد، پنج حرف بزرگ و دو علامت دست‌نویس را بشناسد. در مورد نویسه های چاپی، این سامانه

[1] RETINA

می‌توانست ۴۰ نویسه را با قلم‌های مختلف بشناسد. البته لازم بود که کاربر نوع قلم را مشخص کند. سرعت خواندن این ماشین ۲۴۰۰ نویسه در ثانیه بود و علیرغم قیمت بالایش، کاربرد زیادی پیدا کرد.

۲-۲-۱-۳سیستم‌های نسل سوم

مهم‌ترین مشخصه‌های ماشین‌های نسل سوم که در اواخر دهه‌ی ۱۹۶۰ میلادی مطرح شدند، پرداختن آن‌ها به نویسه‌های تایپی با کیفیت پایین و همچنین مجموعه‌های کامل نویسه‌های دست‌نویس لاتین بود. محصولات موفق تجاری این نسل، از سال‌های ۱۹۷۵ تا ۱۹۸۵ به بازار آمدند. سامانه‌های این نسل شامل سه بخش خواندن اسناد چاپی با کیفیت تصویری نامطلوب، خواندن اسناد حاوی نویسه‌های مجزای دست‌نویس و بسته‌های نرم افزاری بود.

۲-۲-۱-۴ سیستم‌های نسل چهارم

نویسه خوان‌های این نسل می‌توانند اسناد پیچیده شامل متن، جدول، تصویر، گرافیک و روابط ریاضی را پردازش کرده و ساختار اصلی آن‌ها را در خروجی حفظ کنند و همچنین اسناد رنگی و نویزی با کیفیت تصویری پایین را بازشناسی می‌کنند. این نسل تا به امروز هم ادامه دارد و فعالیت‌های گسترده‌ای در این زمینه در حال انجام است.

۲-۲-۲- سیر تحول اُ.سی.آر فارسی

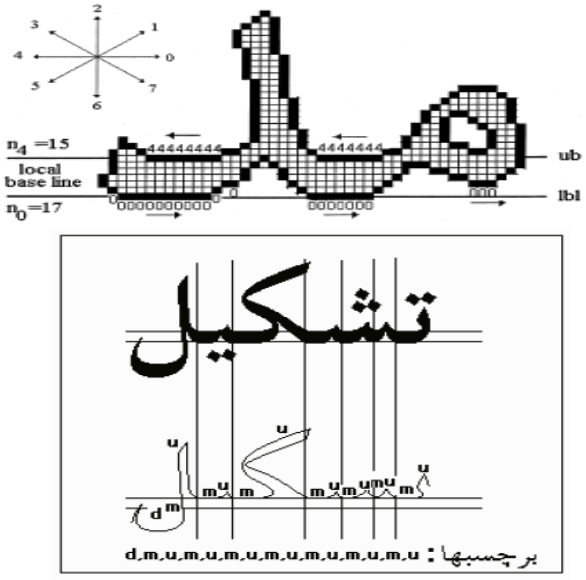
۲-۲-۲-۱ بررسی اجمالی برخی از تحقیقات

در سال‌های اخیر تلاش‌های قابل توجهی از سوی برخی شرکت‌های فعال در زمینه‌ی پردازش تصویر انجام شده که برخی از آن‌ها منجر به محصولات قابل قبولی شده است. برخی از تحقیقاتی که در زمینه‌ی اُ.سی.آر فارسی انجام شده به شرح زیر می‌باشد.

در مورد بازشناسی متون چاپی فارسی مقاله‌ی [۸] روشی برای بازشناسی متون فارسی با قلم‌های درشت مانند تیترهای روزنامه‌ها ارائه کرده است.

در [۹] از ویژگی‌های شبه زرنیکی برای بازشناسی متون تایپی و دست‌نویس با شبکه عصبی احتمالاتی استفاده شده است. این روش مبتنی بر شکل کلی زیرکلمات بوده و تعداد زیرکلمات ۱۰۰۰ فرض شده است. برای بازشناسی از شبکه عصبی استفاده شده و از ۲۸ ویژگی گشتاورهای شبه زرنیکی مرتبه ۴ برای توصیف شکل زیرکلمات استفاده شده است. با آزمایش این روش بر روی مجموعه‌ای شامل ۵۰۰ زیرکلمه دست‌نویس نوشته شده با خط نسخ، و یک صفحه متن چاپی در هر دو مورد، نرخ بازشناسی صحیح ۹۶ ٪ گزارش شده است.

عزمی یک الگوریتم جدید برای جداسازی زیرکلمات فارسی به حروف بر اساس برچسب‌زنی مقید به هر زیرکلمه ارائه کرده است. وی همچنین روشی برای میزان کردن خط پایه برای کانتور بالایی زیرکلمات ارائه کرده است [۱۰]. در این مقاله دقت جداسازی ۹۸/۹٪ گزارش شده است. برای بازشناسی از کدهای فریمن به عنوان ویژگی و از یک طبقه‌بند آماری برای طبقه‌بندی استفاده شده است. روش دومی که به کار برده شده استفاده از تبدیل هاف جهت استخراج ویژگی و شبکه عصبی برای طبقه‌بندی است.

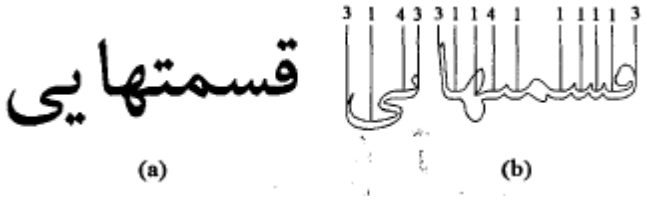


شکل ۲-۳ : جداسازی بر اساس کانتور بالایی [۱۰]

عزمی همچنین از روش مبتنی بر شکل کلی کلمه برای بازشناسی استفاده کرده است. در این کار

تعداد ۵۷۳۷ زیر کلمه تهیه شده است که با حذف نقاط به ۲۷۹۰ زیرکلمه کاهش می‌یابند. در این مقاله سه ویژگی مکان مشخصه، توصیفگر فوریه و کانتور بالایی آزمایش شده است [۱۱].

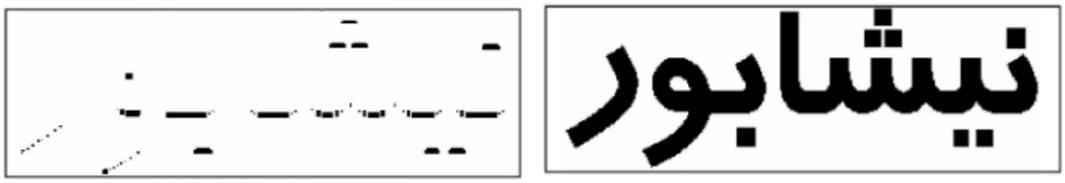
منهاج یک روش جداسازی و بازشناسی توأم برای متون چاپی فارسی ارائه کرده است. وی توصیفگرهای فوریه را به عنوان ویژگی و شبکه‌ی عصبی چند لایه را به عنوان طبقه‌بند استفاده کرده است [۱۲]. برای جداسازی از برچسب‌زنی به کانتور زیرکلمه استفاده شده است (شکل ۲-۴). این روش روی سه قلم تیتر، زر و تایمز نیو رومن آزمایش شده است. برای تصاویر تولیدی در کامپیوتر (کاملاً خالی از نویز) دقت ۱۰۰ ٪ گزارش شده است. برای متونی که هر سه قلم را داشته‌اند این دقت به ۹۳ ٪ کاهش یافته است. برای تصاویر اسکن شده با درجه تفکیک ۴۰۰، دقت ۹۰ ٪ گزارش شده است.



شکل ۲-۴: جداسازی بر مبنای برچسب زنی به کانتور زیرکلمه [۱۲]

در [۱۳] یک سیستم اُ.سی.آر مبتنی بر جداسازی ارائه شده است. این سیستم شامل ماژول ناحیه بندی، جداسازی و بازشناسی است. برای جداسازی از اطلاعات کانتور بالای زیرکلمات استفاده شده است. برای بازشناسی از شبکه‌ی عصبی و SVM استفاده شده و ویژگی‌های مورد استفاده نیز یک سری ویژگی‌های آماری مثل افکنش عمودی، مشتق افکنش عمودی و فاصله‌ی حرف تا خط پایه است. برای بهبود نتایج بازشناسی از NLP استفاده شده است که در آن دقت ۹۸/۳ ٪در سطح حرف و ۹۰/۱۷٪ در سطح کلمه گزارش شده است که این دقت با استفاده از NLP به ۹۴/۱۹٪ افزایش یافته است.

در [۱۴] از تبدیل موجک برای جداسازی زیرکلمات به حروف استفاده شده است. این تبدیل برای آشکارسازی لبه‌های افقی پایین کلمات و یافتن خط پایه به کار رفته است (شکل ۲-۵).



شکل ۲-۵ : یافتن لبه های افقی پایین کلمات [۱۴]

پس از یافتن لبه‌های پایین، از افکنش افقی روی آن‌ها استفاده شده تا موقعیت نقاط جداسازی مشخص شود. برای بازشناسی حروف پس از جداسازی، از شبکه‌ی عصبی استفاده شده است. این سیستم روی ۱۰۰۰ کلمه‌ی چاپی با قلم ترافیک آزمایش شده و دقت ۹۷/۸۳٪ گزارش شده است. سرعت این روش ۳۹ حرف در ثانیه است.

در [۱۵] یک روش سریع و دقیق برای بازشناسی متن چاپی فارسی با درجهٔ تفکیک ۳۰۰ نقطه بر اینچ معرفی شده است. این روش مبتنی بر جداسازی زیرکلمات به حروف و زیر حروف سازنده آن‌ها بوده و فرایند بازشناسی در چندین مرحله، با استفاده از طبقه‌بندهای شبکه‌ی عصبی تقویت شده انجام می‌گیرد. با توجه به این که جداسازی زیرکلمات، همواره یکی از مشکل‌ترین بخش‌های بازشناسی متون فارسی و عربی بوده است، کمترین اشتباه در فرایند جداسازی، موجب گسترش خطا در فرایند کلی بازشناسی می‌شود. در مقاله‌ی مذکور علاوه بر ارائه‌ی روش ساده و سریع برای جداسازی، با استفاده از نتایج مرحله‌ی بازشناسی، خطاهای مرحله‌ی جداسازی تصحیح می‌شود. به عبارتی، سیستم دارای یک حلقه‌ی بازخورد است که باعث افزایش قابلیت اعتماد آن می‌شود. داده‌های هدف در این تحقیق، متون فارسی با قلم‌های لوتوس، نازنین و میترا بوده است. البته الگوریتم به گونه‌ایست که برای سایر قلم‌ها قابل توسعه است. این روش روی ۸ صفحه متن فارسی با درجه‌ی تفکیک ۳۰۰ نقطه بر اینچ آزمایش شده و دقت بازشناسی ۹۹٪ حاصل شده است.

در [۱۶] یک سیستم پیشنهادی جهت بازشناسی کاراکترهای چاپی فارسی با رویکردی بر مراحل بازنمایی و بازشناسی ارائه شده است. سیستم به صورت برون‌خط بوده و در ابتدا سیستم تصویر متنی را پیش‌پردازش می‌کند که شامل عملیات باینری‌سازی، کاهش نویز، اصلاح انحراف و نازک‌سازی است.

در مرحله‌ی قطعه‌بندی به ترتیب خطوط متن، نوار زمینه، زیرکلمات، کاراکترها و نقاط جداسازی می‌شوند. در مرحله‌ی بازنمایی برای اولین بار در یک سیستم بازشناسی، کاراکترهای چاپی چند فونتی فارسی از هفت گشتاور ثابت هندسی سازگار شده، با موفقیت استفاده شده که به طور ابتکاری شماره فرم کاراکتر (جداگانه، ابتدایی، میانی، انتهایی) نیز به بردار ویژگی‌ها اضافه می‌شود.

در [۱۷] یک روش جدید برای جداسازی خطوط متن، در تصاویر متون دست‌نوشته و چاپی ارائه شده است. در این روش ابتدا تکه‌های بهم پیوسته‌ی تصویر مشخص شده و برچسب‌گذاری می‌شود. سپس با در نظر گرفتن یک شش ضلعی پیرامون هر کدام از تکه‌های پیوسته، یک محدوده برای آن‌ها به دست می‌آید. با همپوشانی این محدوده‌ها با همدیگر، یک محدوده‌ی بزرگ‌تر برای هرکدام از خطوط متن حاصل می‌شود که به این ترتیب، هر کدام از خطوط متن در یک محدوده‌ی جدا از دیگر خطوط قرار می‌گیرد. در نهایت با جدا کردن این محدوده‌ها از یکدیگر، خطوط متن جداسازی می‌شوند. در آزمایش‌های انجام شده، درصد موفقیت این روش، برای متون دست‌نوشته‌ی فارسی و انگلیسی به ترتیب برابر با ۸۷/۲٪ و ۹۱/۳٪ و برای متون چاپی ۱۰۰٪ به دست آمده است.

در [۱۸] برای غلبه بر مشکلات جداسازی حروف فارسی به علت اتصال حروف، روشی ترکیبی برای تشخیص حروف فارسی ارائه شده است که در آن جداسازی حروف (که منبع اصلی خطا در سیستم‌های تشخیص حروف فارسی کنونی است) به صورت کامل انجام نشده و فقط حرف اول هر زیرکلمه جدا شده و تشخیص داده می‌شود. سپس این حرف اول، معیاری برای دسته‌بندی زیرکلمات قرار می‌گیرد و هر زیرکلمه فقط در دسته‌ی مربوط تشخیص داده نمی‌شود. این روش روشی انعطاف‌پذیر و قدرتمند بوده که می‌تواند به آسانی توسعه داده شده و برای قلم‌های مختلف به کار برده شود. همچنین با استفاده از روش‌های پیش‌پردازش و پس‌پردازش مناسب، چارچوب ساده و کاملی پیشنهاد می‌دهد.

مقاله‌هایی درباره اُ.سی.آر فارسی در مجله های معتبر خارجی منتشر شده است که از آن جمله می‌توان به موارد زیر اشاره کرد.

موضوع مقاله‌های [۱۹] و [۲۰] بازشناسی مجموعه‌ی از کلمات دست‌نویس است. در این مقالات سامانه‌ی جامع بازشناسی متن فارسی ارائه شده است که در آن از مدل مخفی مارکوف و طبقه‌بند فازی استفاده شده است.

در [۲۱] با استفاده از روش‌های نوین، عملیات پیش‌پردازش که شامل دوسطحی کردن تصویر، حذف حاشیه‌های نامطلوب می‌باشد به همراه تشخیص تصاویر به صورت خودکار انجام شده است.

در [۲۲] با توجه به پیچیدگی‌های خط فارسی از جمله اشکال مختلف یک حرف در قسمت‌های مختلف کلمه، همپوشانی حروف و ...، از یک طبقه‌بند فازی برای شناسایی استفاده شده است.

روش‌های قبلی پیشنهاد شده برای تشخیص قلم اغلب با فیلترهای گابور و بر اساس تشخیص نوع قلم در یک بلوک از متن به جای یک خط و یا یک عبارت استفاده می‌شد، اما در [۲۳] خسروی با استفاده از فیلتر سوبل و رابرتز در هر خط و یا عبارت به تشخیص قلم پرداخته است.

در [۲۴] با توجه به شبیه بودن نقاط حروف فارسی به نویزهای موجود در تصویر، با تخمین اندازه‌ی نقاط در هر ناحیه، به حذف نویز در آن‌ها می‌پردازد.

در [۲۵] شیوه‌ای نوین برای تشخیص و پردازش معادلات ریاضی فارسی ارائه شده است که در بخش تشخیص کاراکتر، از یک سیستم فازی برای کلاسه‌بندی استفاده شده است.

در [۲۶] جهت افزایش سرعت آموزش و تست شبکه عصبی، از کارت‌های پردازش گرافیکی استفاده شده است.

۲-۳-نتیجه‌گیری

در این فصل به مروری بر سیر تحولی اُ.سی.آر پرداخته شد که در آن نسل‌های مختلف اُ.سی.آر معرفی شده و روند تحقیقاتی آن معرفی گردید. همچنین در پایان به معرفی تحقیقاتی در زمینه‌ی اُ.سی.آر فارسی پرداخته شد و موانع موجود بر سر راه تکامل آن بیان گردید. در فصل آتی نیز به معرفی مراحل مختلف پیاده‌سازی اُ.سی.آر پرداخته خواهد شد.

فصل سوم

مراحل پیاده‌سازی اُ.سی.آر

۳-۱-مقدمه

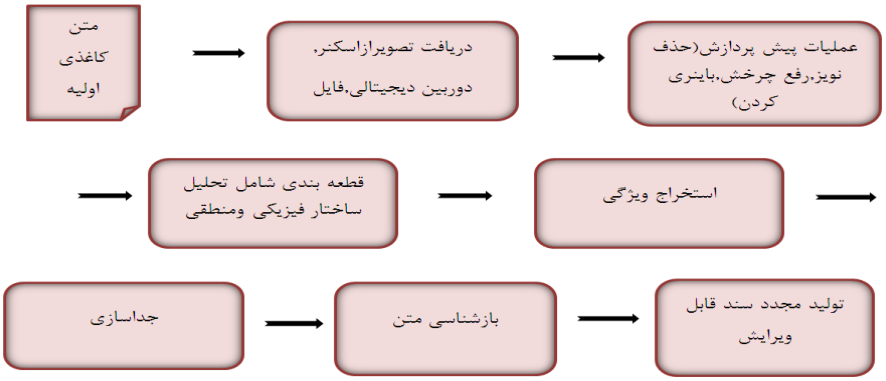
در فصل گذشته به معرفی سیر تکاملی اُ.سی.آر و روش‌های گوناگون استفاده شده در طراحی سیستم‌های بازشناسی متن پرداخته شد. در این فصل تعاریف، مفاهیم، اصطلاحات بازشناسی متن و تمام مراحل مختلفی که برای یک سیستم بازشناسی متن لازم است، به صورت کلی شرح داده می‌شود. این مراحل شامل عملیات پیش‌پردازش، قطعه‌بندی، استخراج ویژگی و بازشناسی متن و روش‌های کلی برای پیاده سازی هرکدام از این مراحل است. پیاده‌سازی بعضی از این بخش‌ها ساده و بخش‌هایی از قبیل جداسازی و بازشناسی، کار پیچیده‌ای است.

۳-۲-سیستم بازشناسی متن

به طورکلی برای تبدیل هر گونه تصویر به فایل قابل ویرایش و قابل فهم برای رایانه، نیاز به یک سیستم بازشناسی متن می‌باشد.

اجزای اصلی یک سیستم بازشناسی متن عبارتند از:

- تصویربرداری از متن ورودی با سطوح خاکستری و درجه تفکیک مناسب
- پردازش‌های اولیه
- قطعه‌بندی
- استخراج ویژگی‌ها
- بازشناسی با یک یا چند طبقه‌بند
- به‌کارگیری اطلاعات جانبی مانند مجموعه لغات معتبر، اطلاعات آماری مربوط به رخداد حروف، اطلاعات دستوری و معنایی



شکل ۳-۱: مراحل سیستم بازشناسی متن

۳-۲-۱- پیش‌پردازش

بعد از اینکه تصویر متن از روبشگر به محیط الکترونیکی انتقال یافت به علت‌هایی از قبیل کثیف بودن صفحه کاغذ یا سطح روبشگر، استفاده از تصاویر کاربنی، کج گذاشتن صفحه، رنگی بودن آن و غیره باید عملیات اولیه‌ای روی آن انجام پذیرد تا برای مراحل بعدی بازشناسی آماده شود. در غیر این صورت به علت نواقص موجود در بازشناسی نتیجه خوبی گرفته نشده و باعث پایین آمدن دقت سیستم بازشناسی می‌شود. عملیات پیش‌پردازش می‌تواند شامل باینری کردن تصویر، حذف نویز و رفع چرخش[1] باشد.

۳-۲-۱-۱ باینری کردن تصویر

بسیاری از الگوریتم‌های بازشناسی حروف از شکل دوسطحی[2] تصاویر ورودی استفاده می‌نمایند. در نتیجه اگر تصویر ورودی خاکستری یا رنگی باشد باید ابتدا به تصویر دوسطحی تبدیل شود. برای دو سطحی کردن تصویر ورودی می‌توان از یک سطح آستانه‌ی ثابت یا یک سطح آستانه‌ی وفقی استفاده نمود. الگوریتم مورد نظر نباید باعث ایجاد نویز در مرز و داخل حروف، چسبیدگی، خوردگی یا

[1] Skew correction

2 Binary

جداسازی آن‌ها شود.

۳-۲-۱-۲ نویز

وقتی تصویر از یک منبع غیر دیجیتالی مانند کاغذ به دست می‌آید به علت‌های مختلفی از جمله جنس کاغذ، رنگ کاغذ، استفاده از کاغذ به همراه کاربن، استفاده از نسخه‌ی کاغذی پس از چند بار کپی گرفتن، وجود پارگی‌ها یا سوراخ دستگاه پانچ بر روی کاغذ و همچنین کثیف بودن سطح روبشگر و غیره نویز در تصویر ایجاد می‌شود. برای حذف نویز می‌توان از فیلتر نمک و فلفل، حذف بافت پس زمینه و ... استفاده کرد. حذف نویز باعث بهبود بازشناسی می‌گردد و از این رو حائز اهمیت است.

۳-۲-۱-۳ رفع چرخش

کاربر هنگام گذاشتن صفحات کاغذی بر روی سطح اسکنر ممکن است دقت کافی نداشته باشد و با کج گذاشتن صفحات در تصویر ایجاد شده چرخشی پدید می‌آید. همچنین ممکن است هنگام عکس‌برداری با کج نگه داشتن دوربین یا درست قرار نگرفتن تصویر در جلوی دوربین، تصویر حاصل نسبت به مبنا اندکی چرخش داشته باشد. در این صورت حاشیه‌های سیاه ناخواسته‌ای نیز اطراف تصویر ایجاد می‌شود. در نتیجه برای اینکه در مرحله بازشناسی با دقت بیشتری شناسایی صورت بگیرد، باید زاویه‌ی چرخش مربوطه را یافته و با روش‌های مناسب حذف کرد. همچنین حاشیه‌های سیاه اطراف تصویر نیز باید حذف شوند. برای رفع چرخش روش‌های زیادی متداول است که از آن جمله می‌توان افکنش افقی، استفاده از تبدیل هاف و استفاده از عملگرهای مورفولوژی را نام برد.

۳-۲-۱-۳-۱ افکنش افقی

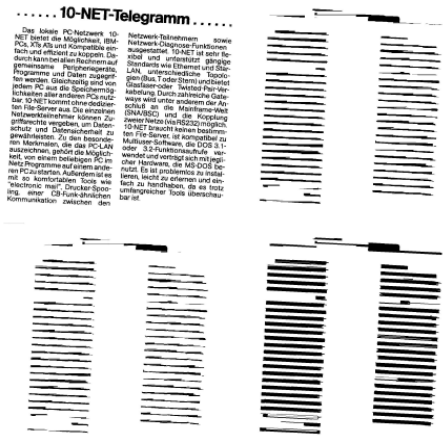
در [۲۷] ساده‌ترین روش‌ها با استفاده از افکنش افقی تصویر در جهات مختلف و یافتن بیشینه‌ی افکنش، زاویه‌ی چرخش یافت می‌شود. مزیت اصلی این روش، سادگی آن است ولی مشکل عمده‌ی این روش این است که برای محاسبه‌ی زاویه‌ی کلی تصویر اندکی کند است.

۳-۲-۱-۲-۳ تبدیل هاف

این تبدیل که اساساً برای یافتن خطوط در جهات مختلف استفاده می‌شود، برای یافتن زاویه‌ی چرخش سند، دو عیب عمده دارد : اول اینکه حجم محاسباتی بسیار بالا و در نتیجه سرعت بسیار کمی دارد. دوم اینکه در اسناد متنی معمولاً خط ممتد گرافیکی نداریم و این تبدیل تلاش می‌کند، خطوط پایه‌ی بلوک‌های متنی را پیدا کند که اگر فاصله‌ی بین خطوط کم باشد یا قلم متن خیلی بزرگ باشد، با مشکل روبرو می‌شود.

۳-۲-۱-۳-۳ عملگرهای مورفولوژی

در [۲۸] با استفاده از عملگر بستن[1] که ترکیبی از عملگر گسترش[2] و فرسایش[3] است، هر خط به یک مؤلفه‌ی متصل و تقریباً یکنواخت تبدیل می‌شود. این کار به ازای عناصر ساختاری خاصی که برای زوایای مختلف حساب شده‌اند، انجام می‌شود. زاویه‌ای که در آن ضخامت خطوط مذکور بیشترین مقدار را داشته باشد، جایی است که عملگر مورد نظر در راستای خطوط متن قرار گرفته و از این رو به عنوان زاویه‌ی چرخش سند در نظر گرفته می‌شود (شکل ۳-۲)



شکل ۳-۲ : یافتن زاویه‌ی چرخش با اعمال چند عملگر مورفولوژی در زوایای مختلف [۲۸]

1 Closing

2 Dilation

3 Erosion

این روش نسبت به سایر روش‌ها دقیق‌تر به نظر می‌رسد؛ لیکن از آنجا که برای هر زاویه، چندین عملگر با عناصر ساختاری تقریباً بزرگ بر روی تصویر اعمال می‌شود، مدت زمان اجرای الگوریتم زیاد است. البته روش‌هایی برای پیاده‌سازی بهینه‌ی عملگرهای مورفولوژی، خاصه روی تصاویر باینری، ابداع شده است که سرعت کار را بهبود می‌بخشد [۲۹].

۳-۲-۱-۳-۴ یافتن چرخش بر اساس یافتن زاویه‌ی خطوط متن

در اسناد ورودی به سیستم بازشناسی متن همیشه تعدادی بلوک متنی وجود دارد که هر کدام از تعدادی خط متنی تشکیل شده‌اند. خطوط یک بلوک متنی غالباً در یک راستا هستند، در نتیجه اگر زاویه‌ی چرخش این خطوط را بیابیم، زاویه‌ی چرخش کل سند یافته می‌شود. در این روش [۳۰] ابتدا تصویر ورودی به ابعاد مناسبی تغییر مقیاس داده می‌شود. پس از دوسطحی‌سازی، تصویر کلمات یک خط با استفاده از عملگرهای مورفولوژی به صورت یک مؤلفه پیوسته درمی آید. با استفاده از الگوریتم برچسب‌زنی مؤلفه‌ها سه خطی که خواص خطوط متن را دارند انتخاب شده و زاویه‌ی چرخش هرکدام از این خطوط از روشی مثلاً افکنش افقی محاسبه می‌شود. زاویه‌ی نهایی از متوسط‌گیری سه زاویه‌ی به دست آمده، حاصل می‌شود. این روش به علت انتخاب چرخش از روی خطوط متن به بلوک‌های بزرگ در تصویر حساسیت ندارد و سرعت خوبی دارد. پس از یافتن زاویه‌ی چرخش سند، تصویر آن با همان زاویه و در خلاف جهت چرخانده می‌شود تا چرخش آن حذف شود. نتیجه‌ی این کار در شکل ۳-۳ نشان داده شده است. در این شکل ضمن چرخاندن تصویر، سعی شده است که ابعاد آن تغییر نکند، به عبارتی حاشیه‌ی اضافی که در حین چرخش ایجاد می‌شود، حذف شده است. تصویر نهایی سند در شکل ۳-۳ مشاهده می‌شود.

شکل ۳-۳ :تصویر نهایی پس از رفع چرخش

۳-۲-۲- قطعه‌بندی

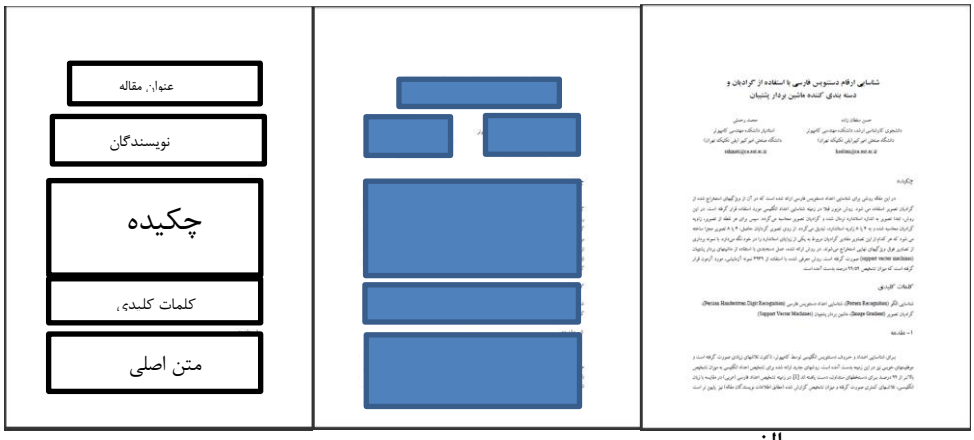
قطعه‌بندی عبارت است از روش‌هایی که بخش‌های مختلفی همچون جدول‌ها، تصاویر، پاراگراف‌ها، جملات یا کلمات و حروف را از تصویر سند استخراج می‌نمایند.

۳-۲-۳-تحلیل ساختار فیزیکی و منطقی

هر تصویر متن به طورکلی می‌تواند شامل متن، عکس، جدول، گراف و غیره باشد. بعد از مرحله‌ی پیش‌پردازش برای شناسایی بهتر می‌توان هر یک از بلوک‌های تصویر را از هم جدا کرده و سپس هر یک به صورت جدا به سیستم بازشناسی سپرده شود. در نتیجه هر یک از بلوک‌ها با توجه به موقعیت آن‌ها در صفحه و همچنین موقعیت هر یک نسبت به همدیگر از هم جدا می‌شوند. همچنین از نظر معنایی می‌توان به هر یک از بلوک‌های متن متناسب با عملکرد معنایی آن‌ها برچسب زد.

پس از تحلیل ساختار اسناد، بلوک‌های مختلف تصویر از هم جدا می‌شوند. جداسازی با در نظر گرفتن ویژگی‌های هر بلوک انجام می‌شود. به عنوان مثال بلوک‌های متن را می‌توان با توجه به خاصیت قرار گرفتن منظم حروف، زیرکلمات و کلمات در کنار هم و تشکیل خطوط موازی تشخیص داد یا بلوک‌های گراف را می‌توان از فضاهای سفید بین اجزای آن و داشتن خطوط راست در ساختار آن از

دیگر بلوک‌ها جدا کرد. پس از تعیین بلوک‌های متن، اجزای متن باید با توجه به روش بازشناسی از یکدیگر جدا و به عنوان ورودی به مرحله بازشناسی فرستاده شوند. در این مرحله تصویر بلوک متن به تصاویر خطوط متن تفکیک می‌شود. از تصویر خط، تصاویر کلمات، زیرکلمات و حروف استخراج می‌شوند.



شکل ۳-۴ : الف) تصویر صفحه اول یک مقاله ب) ساختار فیزیکی ج) ساختار منطقی

۳-۲-۴-استخراج ویژگی[1]

در این مرحله حروف و زیرکلمات به شکل مناسب‌تری که بازشناسی آن‌ها ساده‌تر باشد توصیف می‌شوند. از تصاویر حروف یا زیرکلمات، مشخصاتی که بیان کننده‌ی ویژگی هر یک از آن‌ها باشد و اختلاف بین آن‌ها را بهتر بیان کند، استخراج می‌شود. این مشخصات به صورت ویژگی‌های توصیف‌کننده‌ی حرف یا زیرکلمه یا هر جزء متن می‌باشد. بنابراین از این مرحله به بعد هر جزء متن با یک بردار ویژگی به رایانه معرفی می‌شود. در واقع استخراج ویژگی‌ها مجموعه‌ی کلیه‌ی محاسباتی است که روی الگوهای بدست آمده از مرحله‌ی پیش‌پردازش انجام می‌شود تا بردار ویژگی‌های متناظر با هر الگو تعیین گردد. روش استخراج ویژگی‌ها بر اساس دو نوع از ویژگی‌ها است :۱) آماری ۲) ساختاری

1 Feature Extraction

از ویژگی‌های ساختاری می‌توان به تعداد و نوع پاره خط‌های مختلف (عمودی، افقی، بالا رونده و پایین رونده) اشاره کرد. گشتاورها و مکان‌های مشخصه، ناحیه‌بندی، افکنش‌ها و پروفایل‌ها، نقاط تقاطع و فاصله‌ها نمونه‌هایی از ویژگی‌های آماری هستند.

۳-۲-۵- بازشناسی متن (با یک یا چند طبقه‌بند)

مرحله‌ی بازشناسی شامل روش‌هایی است که در آن بردار ویژگی مناسب از تصویر استخراج شده و در مرحله‌ی اول با بردارهای ویژگی موجود در پایگاه داده مقایسه شده و نهایتاً بر مبنای کم‌ترین فاصله با یکی از بردارهای مرجع متناظر می‌شود. بازشناسی هر یک از اجزای تصویر متن با طبقه‌بندی آن‌ها در یک مجموعه حروف، زیرکلمات یا کلمات معتبر انجام می‌شود. روش‌های طبقه‌بندی به سه شاخه‌ی روش‌های نحوی یا ساختاری، روش‌های آماری و روش‌های مبتنی بر شبکه‌های عصبی تقسیم می‌شوند [۳۱].

۳-۲-۵-۱ روش‌های نحوی یا ساختاری

در این روش‌ها یک الگو[1] بر اساس اجزای تشکیل‌دهنده آن و روابط بین این اجزاء طبقه‌بندی می‌شود. طبقه‌بندی کننده در ابتدا اجزای تشکیل دهنده‌ی نویسه را بازشناسی کرده و با کنار هم قرار دادن آن‌ها بر اساس یک سری روابط نحوی، آن نویسه را بازشناسی می‌کند.

۳-۲-۵-۲ روش‌های آماری

در این روش‌ها، بردار ویژگی متناظر الگو با کلاس‌های آموزش داده شده به سیستم مقایسه می‌شود و نزدیک‌ترین کلاس مشخص می‌شود. روش نزدیک‌ترین همسایه متعلق به این دسته از روش‌های طبقه‌بندی است. در این روش، الگوی ورودی با الگوهای آموزش داده شده به سیستم در مرحله‌ی

1 Pattern

آموزش، مقایسه می‌شود. الگوی با نزدیک‌ترین فاصله به الگوی ورودی به عنوان کلاس الگوی ورودی تعیین می‌شود.

۳-۲-۵-۳ شبکه‌های عصبی

شبکه‌های عصبی از شبکه‌های بهم متصل و ساده با واحدهای پردازش غیرخطی تشکیل می‌شوند. شبکه‌های عصبی مصنوعی بر اساس توپولوژی شبکه، مشخصات نرون‌ها و الگوریتم آموزش آن‌ها از هم متمایز می‌شوند. از شبکه‌ی عصبی برای آموزش الگوهای مختلف به سیستم استفاده می‌شود. آموزش شبکه‌ی عصبی به کمک ویژگی‌های استخراج شده از الگوها انجام می‌شود. پس از آموزش شبکه، با ارائه‌ی ویژگی استخراج شده از یک الگوی ناشناخته به آن، کلاس متناظر آن الگو در خروجی شبکه عصبی مشخص می‌شود.

۳-۲-۶- پس‌پردازش

با بازشناسی تمام تصاویر اجزای متن یک بلوک، متن متناظر با هر یک از آن‌ها مشخص می‌شود. برای بدست آوردن متن نهایی باید پس‌پردازش‌هایی روی این متن انجام شود مثلاً از به هم چسباندن حروف بازشناسی شده در مرحله‌ی قبل، زیرکلمات و کلمات تشکیل می‌شوند. تشکیل این زیرکلمات و کلمات با استفاده از موقعیت آن‌ها در تصویر، یک مرحله پس‌پردازش است. مرحله‌ی بعدی می‌تواند شامل مقایسه و تایید کلمات بازشناسی شده در مرحله‌ی بازشناسی با یک واژه‌نامه باشد.

۳-۳-روش‌های بازشناسی متن

سیستم‌های بازشناسی متن را با توجه به روش بازشناسی می‌توان به سه دسته تقسیم کرد.

۳-۳-۱- روش‌های مبتنی بر جداسازی حروف

در این روش‌ها زیرکلمات و کلمات به حروف شکسته می‌شوند و حروف بازشناسی می‌شوند به

گونه‌ای که از بهم چسباندن این حروف، زیرکلمات و کلمات تشکیل می‌شوند.

روش‌های متداول موجود برای جداسازی حروف فارسی را می‌توان به پنج گروه عمده تقسیم کرد:

۳-۳-۱-۱روش‌های مبتنی بر افکنش عمودی

روش‌های مبتنی بر افکنش از ساده‌ترین روش‌های جداسازی هستند، در این روش‌ها از هیستوگرام عمودی تصویر استفاده می‌شود [۳۲]، [۳۳]. جاهایی که مقادیر هیستوگرام صفر است نقاط جداسازی محسوب می‌شوند. این روش در مواردی که بین حروف همپوشانی وجود دارد (به عنوان مثال حروفی که زیر سرکش کاف هستند) مناسب نیست.

۳-۳-۱-۲ روش‌های مبتنی بر کانتور

روش‌های مبتنی بر آنالیز کانتور، با حرکت روی کانتور کلمات و یافتن مینیمم و ماکزیمم‌های محلی و تحلیل آن‌ها نقاط جداسازی را پیدا می‌کنند [۳۴]، [۱۰]، [۱۲] و [۳۵]. در زبان انگلیسی از این روش‌ها غالباً برای جداسازی حروف کلمات دست‌نویس استفاده می‌شود. در مورد زبان فارسی هم از این روش‌ها استفاده شده که نتایج قابل قبولی داشته است. این روش‌ها نیز غالباً در برخورد با همپوشانی‌های عمودی دچار اشکال می‌شوند.

۳-۳-۱-۳ روش‌های مبتنی بر آنالیز پروفایل

روش‌های مبتنی بر آنالیز پروفایل شبیه به روش‌های مبتنی بر کانتورند با این تفاوت که به جای استفاده از کانتور کلمه، تنها از پروفایل بالایی کلمه استفاده می‌کنند [۳۶]. پروفایل بالایی مجموعه‌ی اولین نقاط بالایی تصویر در هر ستون آن است. در این روش‌ها، مشابه روش‌های مبتنی بر کانتور احتمال پنهان شدن برخی حروف در زیر سایر حروف زیاد است مثلاً در فارسی سرکش کاف یا سرکش الف چنین حالتی را ایجاد می‌کند و یا در انگلیسی ممکن است حروفی مثل i در زیر حروفی مثل T ناپدید شوند.

۳-۳-۱-۴ روش‌های مبتنی بر اسکلت

روش‌های مبتنی بر اسکلت که از دقت مناسب‌تری برخوردارند، ابتدا با نازک‌سازی تصویر زیرکلمه، اسکلت آن را تولید کرده و با یافتن خط پایه‌ی اسکلت حاصله و تحلیل آن، نقاط جداسازی را می‌یابند. اشکال این روش زمان گیر بودن و توزیع خطا در صورت یافتن اسکلت نادرست است.

۳-۳-۱-۵ روش‌های مبتنی بر اسکلت هوشمند یا جداسازی همراه با بازشناسی

در چند سال اخیر، الگوریتم‌های هوشمند به دلیل اینکه در آن‌ها امکان تصحیح نتایج جداسازی وجود دارد از محبوبیت بیشتری برخوردار شده‌اند [۳۷]، [۲۳]. در این الگوریتم‌ها، پس از جداسازی اولیه، یک موتور بازشناسی حروف، نتایج جداسازی را تائید یا رد می‌کند.

۳-۳-۲- روش‌های مبتنی بر شکل کلی زیرکلمات

در این روش کل زیرکلمه به صورت یک الگو در نظر گرفته می‌شود و ویژگی‌های لازم برای طبقه‌بندی از تصویر کل کلمه استخراج می‌شوند. ابتدا باید یک مجموعه پایگاه داده‌ی کامل از لغات پرکاربرد در نظر گرفته شود که تهیه‌ی این پایگاه به دلیل نبود یک مرجع، کار دشواری است. با توجه به اینکه بسیاری از زیرکلمات شباهت زیادی باهم دارند و تنها تفاوت آن‌ها در تعداد نقاط و موقعیت قرار گرفتن آن‌ها می‌باشد، برای بررسی تأثیر نقاط و علائم، دو رویکرد برای بازشناسی زیرکلمات در نظر گرفته می‌شود:

- بازشناسی زیرکلمات بدون حذف نقاط و علائم
- بازشناسی زیرکلمات با حذف نقاط و علائم

برای بازشناسی زیرکلمه‌ی ورودی باید تعیین شود کدام یک از زیرکلمات موجود بیشترین شباهت را به زیرکلمه‌ی ورودی دارد. به دلیل زیاد بودن تعداد کلاس‌های خروجی، نمی‌توان از روش‌های مرسوم بازشناسی الگو استفاده کرد. بنابراین باید از روش‌های خوشه‌یابی کمک گرفت. در نتیجه ابتدا

در فاز آموزش، ویژگی‌های مناسبی از زیرکلمات استخراج شده و با استفاده از الگوریتم‌های خوشه‌یابی، زیرکلمات تهیه شده برای آموزش به چندین خوشه تقسیم می‌شوند. سپس در فرایند بازشناسی زیرکلمه‌ی ورودی، ابتدا بر اساس ویژگی‌های مورد استفاده در خوشه‌یابی، نزدیک‌ترین خوشه‌ها به زیرکلمه‌ی ورودی تعیین می‌شود و پس از تعیین خوشه، خود زیرکلمه شناسایی می‌شود.

در این مرحله، ویژگی‌های دیگری که توصیف بهتری شامل جزئیات زیرکلمه را داشته باشند، استخراج شده و با ویژگی‌های متناظر از زیرکلمات کاندید مقایسه می‌شود. نزدیک‌ترین زیرکلمه به زیرکلمه‌ی ورودی به عنوان برنده انتخاب می‌شود.

در رویکرد بازشناسی با حذف نقاط و علائم، ابتدا با استفاده از روش برچسب‌زنی مؤلفه‌ها نقاط را حذف کرده و بدنه‌ی زیرکلمه شناسایی می‌شود که در این صورت پیکسل‌های نویز هم از بین می‌رود. با حذف نقاط و علائم حجم پایگاه داده به حدود نصف کاهش یافته و با این کار فرایندهای خوشه‌یابی و بازشناسی ساده‌تر و در عین حال دقیق‌تر می‌شود و همچنین می‌توان زیرکلماتی را که در پایگاه داده نبوده‌اند ولی بدنه‌ی آن‌ها مشابه یکی از بدنه‌های موجود باشد را شناسایی کرد. پس از بازشناسی اطلاعات می‌توان نقاط را به بدنه اضافه کرد. برای تولید پایگاه تصویری زیرکلمات بدون نقطه، مشابه حالت با نقطه عمل می‌شود با این تفاوت که یک مرحله‌ی حذف نقاط قبل از افزودن به پایگاه داده قرار می‌گیرد.

مزیت استفاده از شکل کلی در جاهایی است که جداسازی امکان پذیر نیست. مثلاً برای قلم‌های خاص که جداسازی آن‌ها خیلی مشکل است یا در مواردی که نقاط و علائم به بدنه چسبیده‌اند و یا اندازه‌ی قلم خیلی کوچک است.

۳-۳-۳- روش‌های ترکیبی (شکل کلی-جداسازی)

روش‌های ترکیبی بدین ترتیب است که با توجه به نوع قلم و زبان متن در مواردی که روش جداسازی پاسخ مناسب‌تری برای شناسایی بدهد از این روش استفاده کرده و در سایر موارد که شکل

کلی مناسب‌تر باشد و یا جداسازی به راحتی امکان پذیر نباشد، برای بازشناسی از شکل کلی استفاده می‌کند. در این پایان‌نامه، رویکرد پیش‌بینی شده بازشناسی بر اساس جداسازی خواهد بود.

۳-۴- قابلیت‌های سیستم‌های بازشناسی متن

در بررسی یک سیستم بازشناسی متن، قابلیت آن در بازشناسی متون مختلف سنجیده می‌شود که از آن جمله می‌توان به متون چاپ شده با چند قلم محدود و نامحدود، بازشناسی متون داخل جداول، بازشناسی متون چند زبانی، بازشناسی متون چند ستونی، بازشناسی متون دست‌نویس داخل متن چاپی، بازشناسی فرمول، بازشناسی متون رنگی و خاکستری، بازشناسی متون دست‌نویس با محدودیت و بدون محدودیت و بازشناسی دست‌نوشته برخط با صفحات رقمی کننده مختلف اشاره کرد. در ادامه برخی از موارد بالا به اختصار بررسی می‌شوند.

۳-۴-۱- متن‌های تک قلمی

سیستم بازشناسی متن ممکن است برای بازشناسی یک قلم مشخص طراحی شود. در چنین مواردی کارایی این سیستم در بازشناسی متون با قلم‌های دیگر پایین‌تر از قلم آموزش داده شده به آن خواهد بود.

۳-۴-۲- متن‌های چند قلمی[1]

سیستم بازشناسی متن ممکن است برای بازشناسی چند قلم محدود طراحی شده و آموزش داده شود. کارایی این سیستم در بازشناسی متون با قلم‌هایی که به سیستم آموزش داده نشده‌اند پایین‌تر از متون چاپ شده با قلم‌های آموزش دیده خواهد بود. در مورد قلم‌های مشابه با قلم‌های آموزش، ممکن است کارایی سیستم بهتر باشد.

1 Multi-Font

۳-۴-۳-متن‌های دست‌نویس مقید[1]

در یک سیستم بازشناسی متون دست‌نویس (برخط و برون‌خط) ممکن است محدودیت‌هایی برای نوشتن بعضی علائم و نقاط، حروف، زیرکلمات یا کلمات فرض شود. به عنوان مثال، سه نقطه باید به صورت یک نیم دایره نوشته شود و سه نقطه‌ی جدا از هم به عنوان سه تک نقطه بازشناسی می‌شوند. نرخ بازشناسی در این سیستم‌ها بسته به محدودیت‌های اعمال شده می‌تواند بسیار بالا باشد.

۳-۴-۴-متن‌های دست‌نویس بدون قید[2]

تنوع دستخط افراد مختلف و کیفیت معمولاً پایین تصاویر، بازشناسی متون دست‌نوشته را به کاری پیچیده تبدیل کرده است. در حالتی که این متون بدون محدودیت نیز نوشته شوند، دشواری بازشناسی آن‌ها دو چندان می‌شود. به این دلیل نرخ بازشناسی سیستم‌های بازشناسی متون بدون محدودیت نسبت به سیستم‌های با محدودیت پایین‌تر است.

۳-۴-۵- متن‌های داخل جداول و فرم‌ها

بسیاری از اسناد اداری و متون علمی شامل جداول و فرم‌هایی با متون چاپی هستند. در برخی موارد متن داخل جدول به خطوط آن می‌چسبد. توانایی یک سیستم در جداسازی جدول از تصویر و بازشناسی این متون، قابلیت مهمی برای یک سیستم بازشناسی متن است.

۳-۴-۶-متن‌های چند زبانی

در تصویر یک سند ممکن است متون با چند زبان وجود داشته باشد یا هر سند به یک زبان نوشته شده باشد. یک سیستم بازشناسی متن که قابلیت بازشناسی زبان‌های مختلف در تصاویر مجزا را دارد،

1 Constrained

2 Unconstrained

ممکن است قادر به بازشناسی متون زبان‌های مختلف در یک تصویر کنار هم نباشد. بنابراین در انتخاب یک سیستم بازشناسی متن چند زبانی باید به این قابلیت دقت شود.

۳-۴-۷-متن‌های شامل فرمول

اگر سیستم بخواهد در سطح وسیع عملیات بازشناسی را انجام بدهد توجه به نکاتی از قبیل وجود فرمول‌های زیاد در متون علمی نباید نادیده گرفته شود. در نتیجه برای بازشناسی متون علمی، باید سیستمی با قابلیت بازشناسی فرمول انتخاب شود. باید توجه داشت که به دلیل حضور علائم و برخی نویسه‌ها در فرمول‌های ریاضی، قابلیت بازشناسی فرمول با قابلیت بازشناسی متون لاتین متفاوت است.

۳-۴-۸-متن‌های چند ستونی

برخی متون کتاب‌ها، مجلات و روزنامه‌ها به صورت چند ستونی چاپ می‌شوند. بنابراین برای بازشناسی متون این اسناد، سیستم بازشناسی باید قابلیت جداسازی متون چند ستونی را داشته باشد.

۳-۴-۹- تصویرهای خاکستری و رنگی

همه‌ی تصاویر روبش شده‌ی متون، دودویی نبوده و سهم عمده‌ای از آن‌ها به صورت خاکستری و بخشی از آن‌ها نیز به صورت رنگی هستند. بنابراین یکی از ویژگی‌های سیستم بازشناسی متن، قابلیت بازشناسی متون خاکستری و رنگی است.

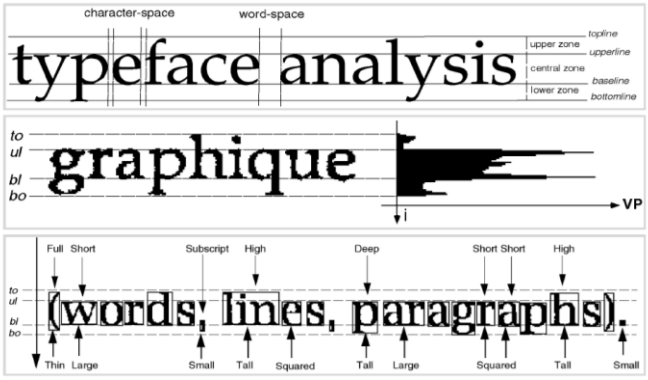
۳-۴-۱۰- بازشناسی قلم

دو رویکرد مرسوم در زمینه‌ی بازشناسی قلم وجود دارد:

- بازشناسی قلم بر مبنای ویژگی‌های تایپوگرافی

روش‌های تایپوگرافی نوع قلم را در سطح یک خط یا عبارت تشخیص می‌دهند [۳۸] و [۳۹]. ویژگی‌های تایپوگرافی، مانند شیب حروف، ضخامت حروف، عرض فاصله‌ی بین حروف و افکنش در نواحی بالا، پایین و وسط خط معمولاً مشخصات یک قلم را به خوبی نمایندگی می‌کنند، ولی استخراج

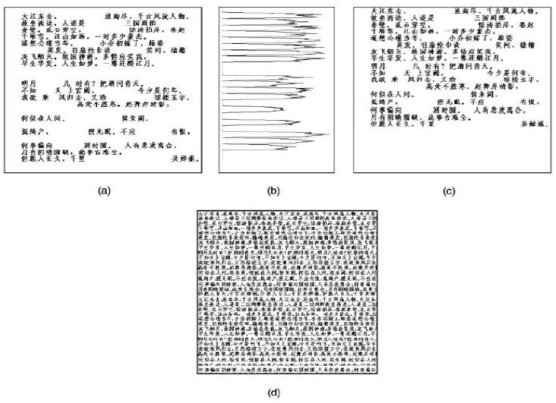
این ویژگی‌ها نیازمند این است که سند ورودی خالی از نویز باشد و با درجه تفکیک بالا، مثلاً ۳۰۰ نقطه بر اینچ، اسکن شده باشد (شکل ۳-۵).



شکل ۳-۵ : ویژگی‌های تایپوگرافی [۴۰]

- با استفاده از ویژگی‌های بافتی

ویژگی‌های بافتی بیش از سایر ویژگی‌ها برای بازشناسی قلم استفاده شده‌اند. این ویژگی‌ها غالباً با استفاده از فیلترهای گابور و تبدیل موجک حاصل می‌شوند. غالب کارهای صورت گرفته با استفاده از این ویژگی‌ها، در سطح یک بلوک متنی انجام شده است نه در سطح یک خط یا کلمه. در این روش‌ها ابتدا یک بلوک متنی پردازش شده و بعد از حذف فضاهای خالی بین خطوط و بین حروف، یک بافت یکپارچه حاصل می‌شود. سپس ویژگی‌های بافتی استخراج شده و قلم متن بر اساس این ویژگی‌ها بازشناسی می‌شود.



شکل ۳-۶ : تشکیل یک بافت یکپارچه از روی بلوک متنی

مسئله‌ی مهم دیگر در بازشناسی قلم، پیچیدگی محاسباتی روش است که مستقیماً سرعت سامانه‌ی نویسه خوان را تحت تأثیر قرار می‌دهد. برخی ویژگی‌های بافتی مانند فیلترهای گابور که به کرات استفاده شده است بسیار زمان‌بر هستند و استفاده از آن‌ها در سیستم اُ.سی.آر سرعت کلی سیستم را به شدت کاهش می‌دهد.

۳-۵-نتیجه‌گیری

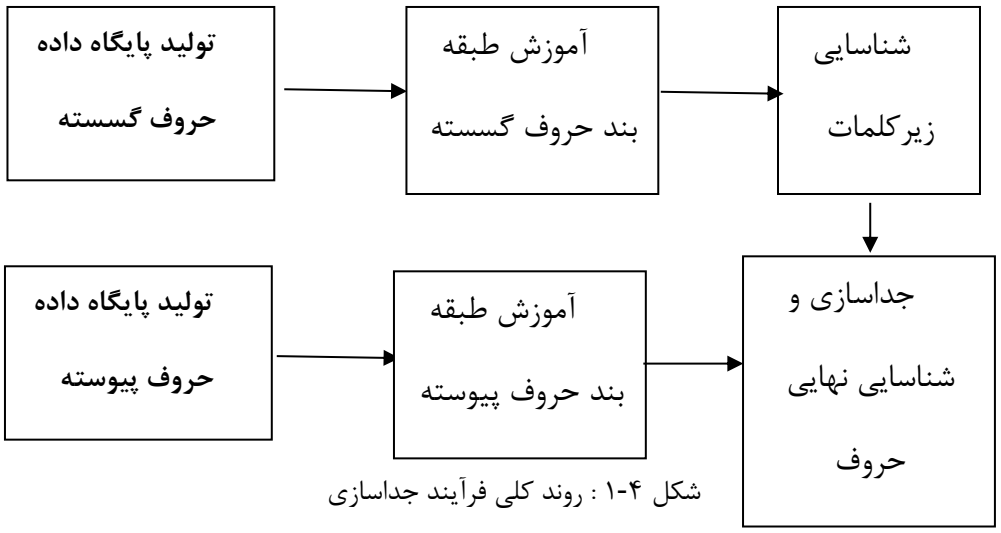
در این فصل مفاهیم و اصطلاحات و مراحل مختلف یک سیستم بازشناسی متن به صورت کلی شرح داده شده و هرکدام به اختصار توضیح داده شد. انواع روش‌های بازشناسی متن و قابلیت‌های یک سیستم بازشناسی متن بیان شد. قابل ذکر است در این پایان‌نامه بعضی از این موارد پیاده‌سازی شده و برخی فقط جهت کامل بودن عملیات یک سیستم بازشناسی متن بیان گردیده است.

فصل چهارم

تولید پایگاه داده

۴-۱- مقدمه

در فصل گذشته راجع به مراحل کلی سیستم بازشناسی متن توضیحاتی داده شد. در این پایان‌نامه هدف، تولید سیستم بازشناسی است که متون نوشته شده با قلم Iranian sans را شناسایی کند. به منظور شناسایی از دو طبقه‌بند حروف گسسته و پیوسته استفاده می‌شود. برای آموزش این طبقه‌بندها نیاز به پایگاه داده می‌باشد. روند کلی کار بدین صورت است که ابتدا با استفاده از طبقه‌بند حروف گسسته زیرکلمات شناسایی شده و برای جداسازی به سمت طبقه‌بند حروف پیوسته فرستاده می‌شود. برای آموزش طبقه‌بندهای فوق نیاز به پایگاه داده می‌باشد. این روند در شکل ۴-۱ نشان داده شده است.



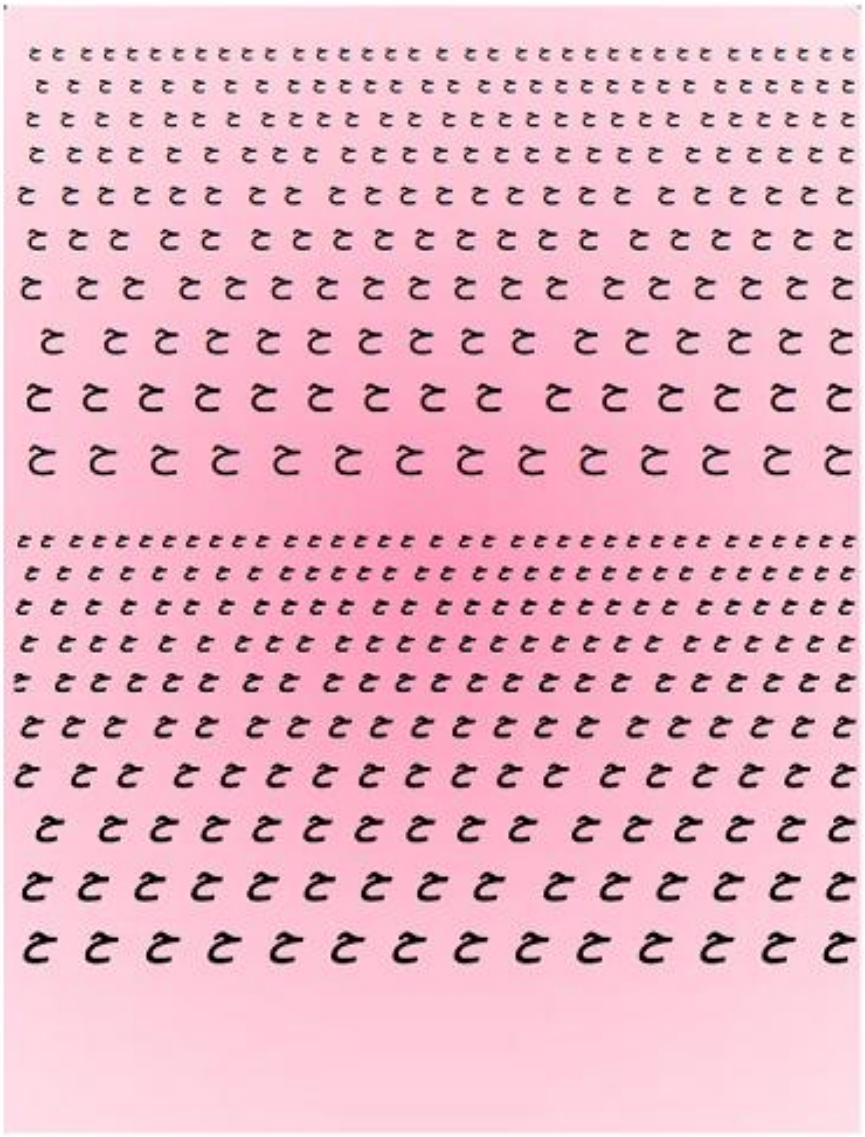
شکل ۴-۱ : روند کلی فرآیند جداسازی

با توجه به اینکه هیچ پایگاه داده اولیه‌ی آماده‌ای در مورد قلم مذکور وجود ندارد، ابتدا به تولید پایگاه داده پرداخته و شبکه آموزش داده می‌شود. پایگاه داده برای حروف پیوسته و گسسته تولید می‌شود. اندازه‌های قلم بین ۹ تا ۲۴ بوده و حروف در حالت‌های ساده، پررنگ و زاویه‌دار نوشته شده‌اند. پس از تولید داده، مجموعه‌ی حروف بدون توجه به نقاط آن‌ها ابتدا با یک چاپگر لیزری مدل HP چاپ و سپس با یک اسکنر مدل HP با درجه تفکیک ۳۰۰ نقطه بر اینچ اسکن شده و به صورت فایل دیجیتالی درآمده‌اند. در حین چاپ و اسکن سعی شده با تغییر سطح روشنایی و استفاده از کاغذهایی با پس‌زمینه‌های مختلف تصاویری با کیفیت متفاوت ایجاد شود. در نهایت پس از اعمال پیش‌پردازش بر روی تصاویر فوق، مؤلفه‌های مختلف پایگاه داده با برچسب‌زنی مؤلفه‌ها تشخیص داده

شده و پس از استخراج ویژگی، نمونه‌ها برای آموزش به شبکه داده می‌شوند.

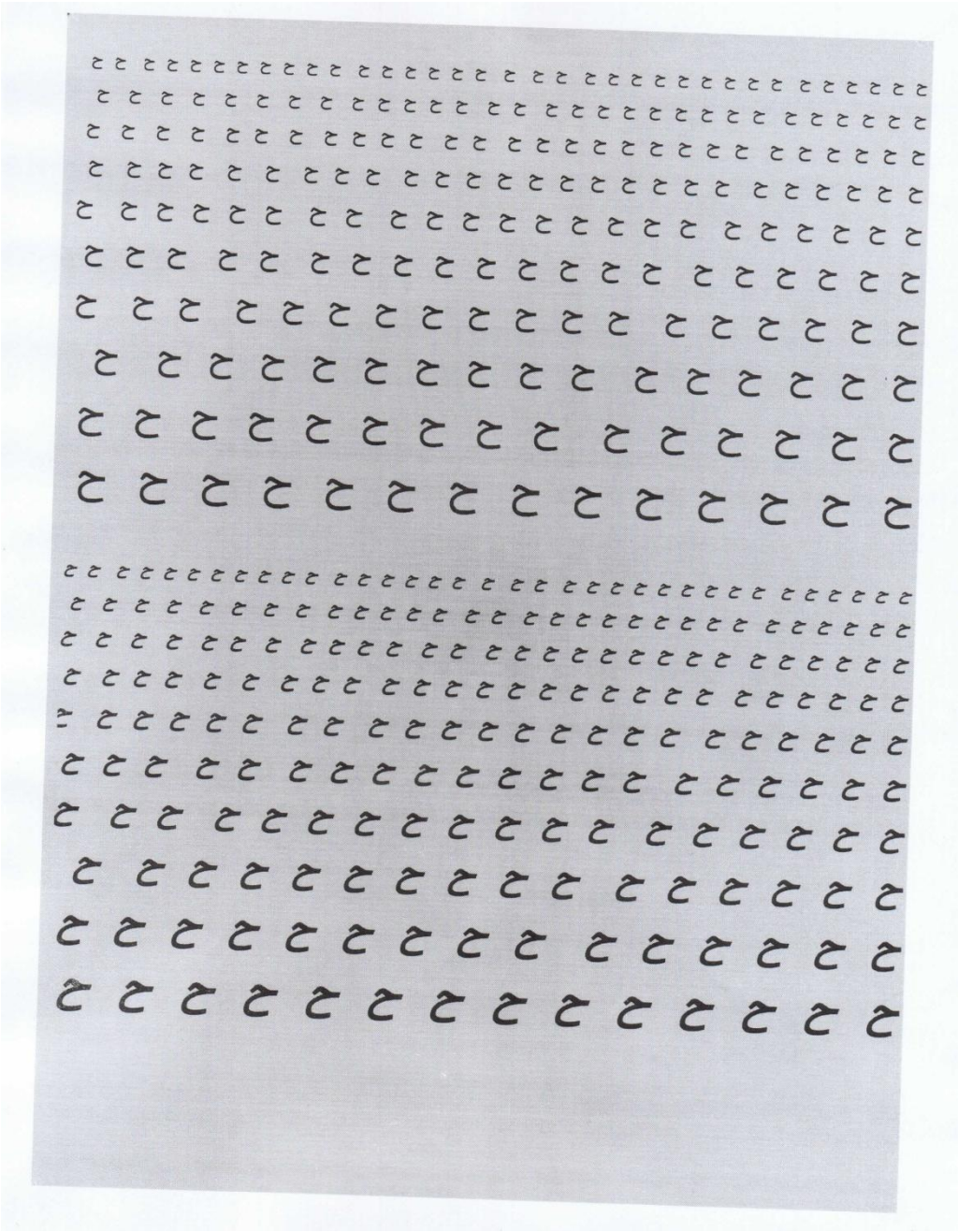
۴-۲- تولید پایگاه داده حروف جدا

در مرحله‌ی نخست با نگارش مناسب هر حرف الفبای فارسی با قلم مذکور مطابق قالب زیر منابع اولیه برای ساخت پایگاه داده فراهم آورده شد که در سطرهای آن از اندازه ۹، ۱۰، ۱۱، ...۲۴ و سپس ضخیم شده و اندکی مورب شده‌ی آن استفاده شده است که نمونه‌ی آن در شکل ۴-۲ آمده است.



شکل ۴-۲ : منابع اولیه تولید شده توسط نرم‌افزار وورد برای ساخت پایگاه داده

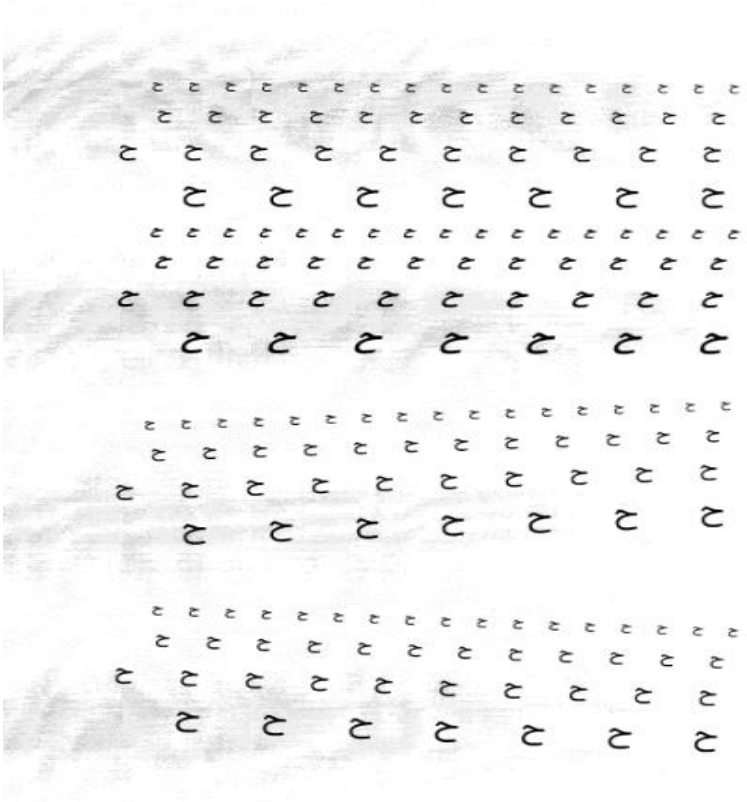
این کار برای حروف دیگر نیز تکرار گردید. سپس در مرحله‌ی بعد نوشته‌های قسمت قبل پرینت گرفته شده و با اندکی چرخش و با عمق ۳۰۰ نقطه بر اینچ اسکن گردید که نتایج آن در شکل ۴-۳ آمده است.



شکل ۴-۳ : تصویر اسکن شده حروف

همان‌طور که مشاهده می‌شود در ابتدا در زمان اسکن چرخش اولیه‌ای معادل ۱۰ درجه در تصویر ایجاد شد و همچنین از رنگ زمینه به منظور ایجاد نویز در تصویر استفاده شد.

جهت شناسایی هرچه بهتر و دقیق‌تر و با توجه به اینکه پایگاه داده قبلی در مرحله شناسایی برای جملاتی که به صورت زاویه‌دار نوشته شده بودند با مشکل مواجه می‌شد، علاوه بر حروف قبل چند سطر به صورت زاویه‌دار مطابق شکل ۴-۴ ایجاد می‌شود.



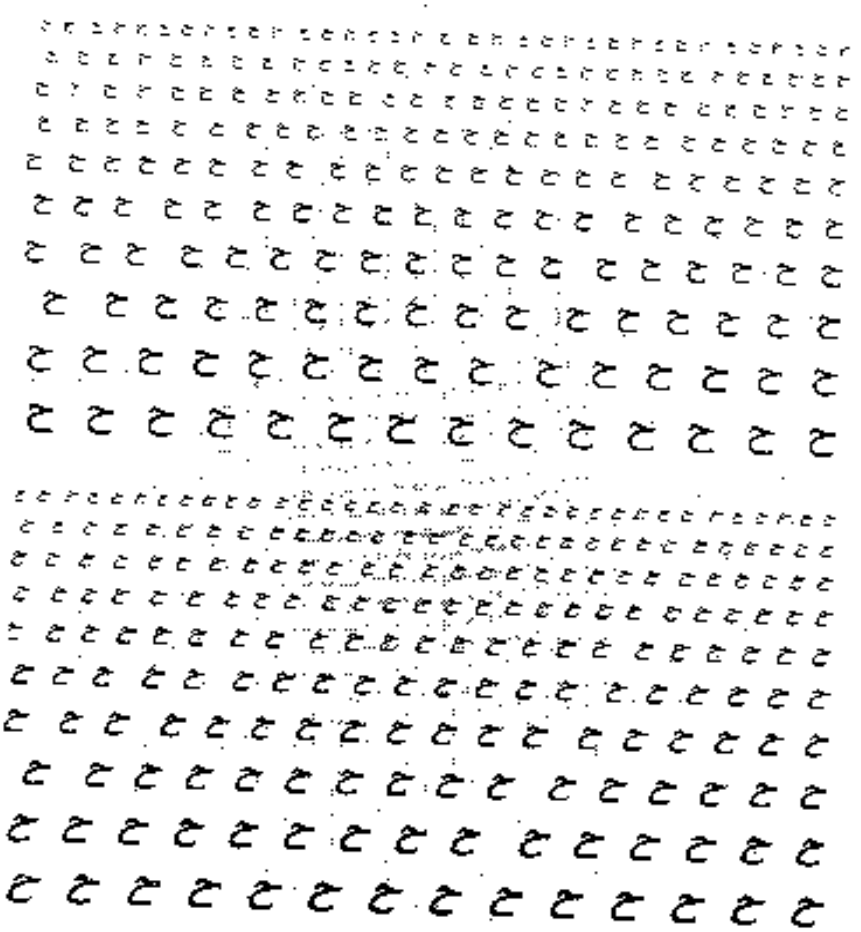
شکل ۴-۴ : تولید پایگاه داده متنوع‌تر

۴-۲-۱- پیش‌پردازش

پیش‌پردازش شامل کلیه‌ی اعمالی است که روی تصویر اولیه انجام می‌شود تا موجب بهتر شدن روند مراحل بعدی شود. ازجمله اعمال پیش‌پردازش می‌توان به باینری کردن تصویر، حذف نویز، رفع چرخش و غیره اشاره کرد.

۴-۲-۱-۱ باینری کردن تصویر

جهت پردازش مناسب بر روی تصویر اسکن شده، باید تصویر حاصل به صورت یک ماتریس با مقادیر ۰ و ۱ تبدیل شود.



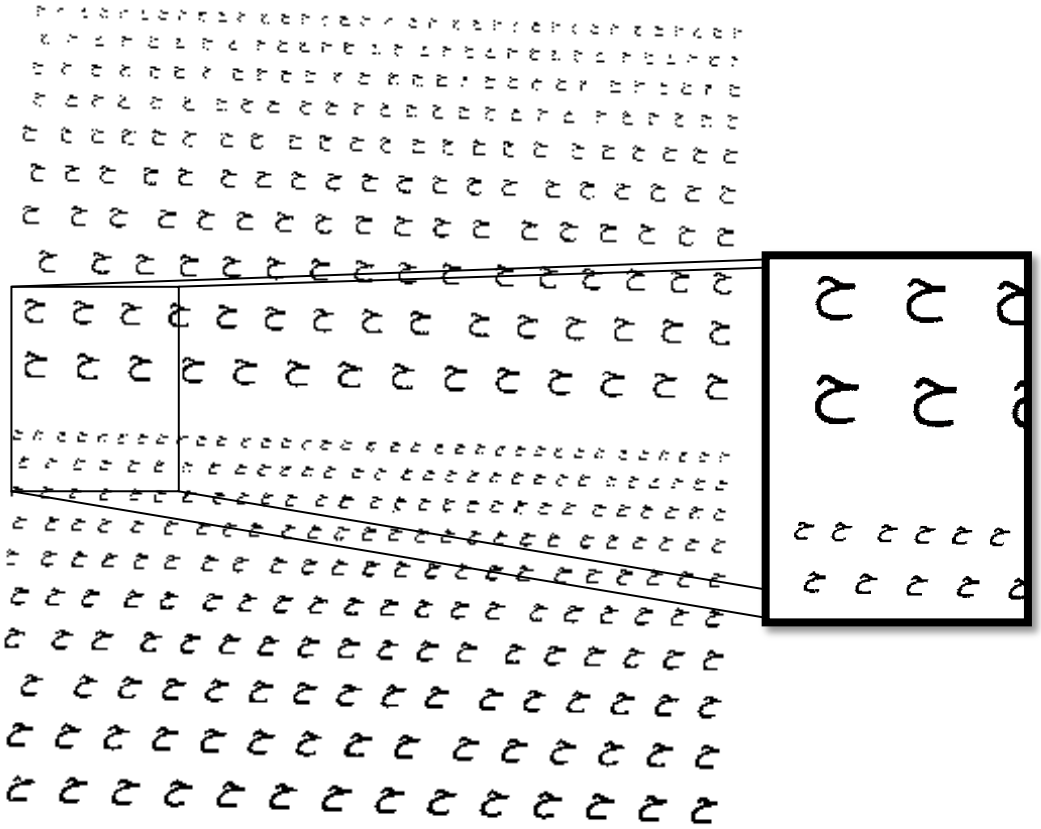
شکل ۴-۵: تصویر باینری شده با نویز فراوان

۴-۲-۱-۲ حذف نویز

با توجه به اینکه زمینه‌ی تصویر رنگی می‌باشد پس از اسکن کردن و باینری کردن آن، نقاط تیره‌ای در تصویر به وجود آمده است؛ لذا برای شناسایی بهتر در مرحله بازشناسی باید به طریق مناسب این نقاط تیره از تصویر حذف شوند. جهت رسیدن به این هدف در این مرحله ابتدا سعی شد با استفاده از روش برچسب‌زنی مؤلفه‌ها و حذف مؤلفه‌های کوچک‌تر از مؤلفه‌های اصلی، نویزها حذف شوند ولی این روش به علت وجود نویز زیاد در تصویر و مدت زمان زیاد پردازش برای برچسب‌زنی همه‌ی مؤلفه‌های نویز استفاده نشد. در نتیجه از دو روش زیر برای حذف اولیه‌ی نویز استفاده شد:

روش اول: ابتدا تقسیم بلوکی تصویر به بلوک‌های k در k انجام شده و سپس با محاسبه‌ی تعداد

پیکسل‌های سیاه در هر بلوک و تعیین سطح آستانه[1] برای آن به صورت مقدار L، پی به وجود نقاط نویز در بلوک‌ها برده شده که اگر تعداد آن از مقدار L کمتر باشد، آن نقاط نویز تلقی شده و از تصویر حذف خواهند شد. حاصل این کار در شکل ۴-۶ آمده است.



شکل ۴-۶ : حذف نویز از تصویر

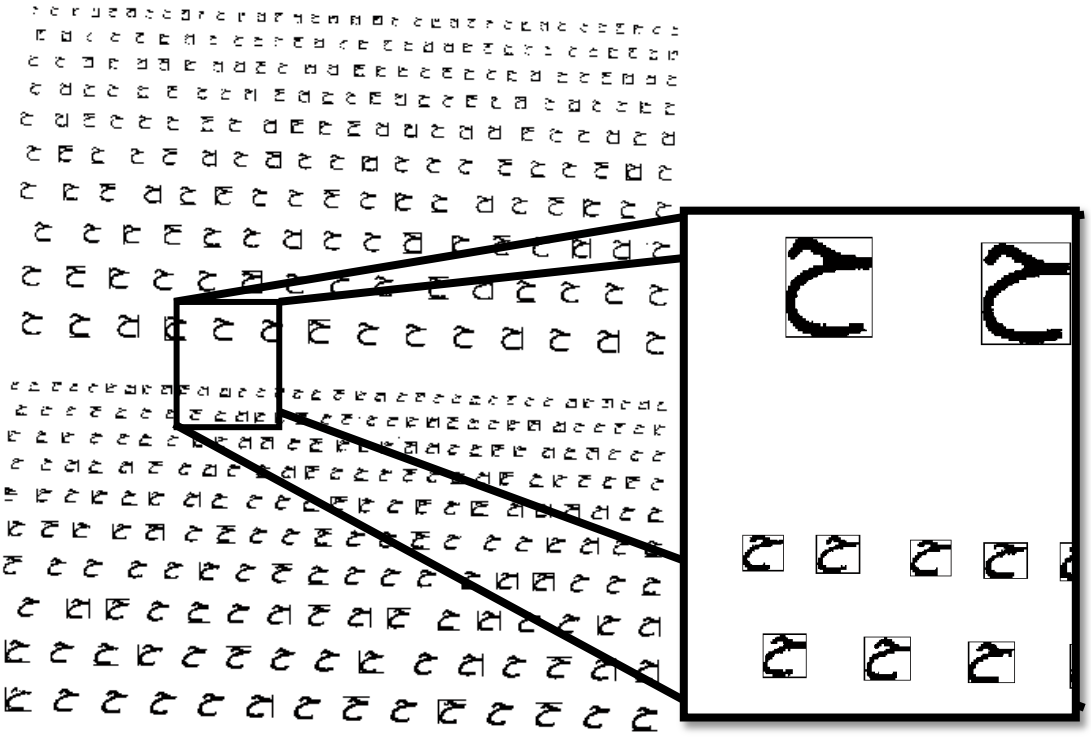
روش دوم : در این روش نیز ابتدا تقسیم بلوکی تصویر به بلوک‌های ۳ در ۳ انجام شده و سپس با محاسبه تفاضل پیکسل‌های موجود در هر بلوک از مقدار پیکسل مرکزی و تعیین سطح آستانه برای آن به صورت مقدار L، پی به وجود نقاط در بلوک‌ها برده شده که اگر مقدار تفاضل از مقدار L کمتر باشد، آن نقاط نویز تلقی شده و از تصویر حذف خواهند شد. که نتایج حاصل از این روش نیز مشابه روش قبل بود با این تفاوت که در آن بجای معیاری برای تعداد نقاط سیاه در هر بلوک از نوعی معیار

[1] Threshhold

گرادیان استفاده شده و سریع‌تر است. پس از انجام این دو روش در نهایت از برچسب‌زنی برای حذف نقاط باقی مانده به عنوان نویز استفاده شد. از این مرحله نمی‌توان قبل از این دو روش استفاده نمود چون قبل از این مراحل با توجه تعداد زیاد مؤلفه‌های نویز مدت زمان زیادی برای حذف آن‌ها در مرحله اول لازم است.

۴-۲-۲- استخراج و جدا کردن حروف

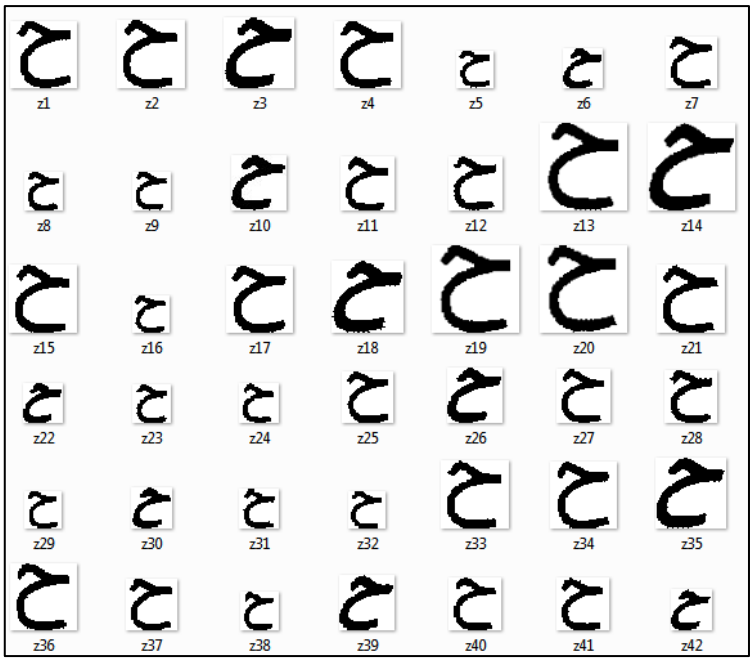
در این مرحله از برچسب‌زنی مؤلفه‌ها برای استخراج حروف استفاده می‌شود، که با استفاده از آن مکان و اندازه‌ی تصویر هر حرف مشخص شده و در متغیری مناسب جهت تولید پایگاه داده ذخیره می‌شود.



شکل ۴-۷ : برچسب‌زنی مؤلفه‌ها برای تولید پایگاه داده

۴-۲-۳- شروع عملیات جداسازی

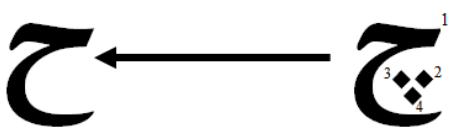
مؤلفه‌های برچسب زده شده در قسمت قبل جهت استانداردسازی، به صورت مربعی ذخیره می‌شوند. به این صورت که با توجه به اندازه هر تصویر به طول یا عرض آن فضای سفید اضافه شده تا مربعی شده و در تصویری مجزا ذخیره می‌شود.



شکل ۴-۸ : مربعی کردن و ذخیره‌سازی حروف

۴-۲-۴- حذف نقطه

با توجه به اینکه شناسایی اولیه بر اساس بدنه‌ی حروف (حروف بی‌نقطه) است، برای حذف نقطه‌ی حروف از برچسب‌زنی مؤلفه‌ها به گونه‌ای استفاده می‌شود که در هر برچسب بزرگ‌ترین اندازه، تصویر اصلی شناخته شده و سایر برچسب‌هایی که اندازه‌ی آن‌ها از ۲ برابر عرض قلم کوچک‌تر باشند حذف می‌گردد.



شکل ۴-۹ : حذف مؤلفه‌های ۲،۳و۴ برای حذف نقاط

۴-۲-۵- حروف گسسته

در این پایان‌نامه تمام حروف گسسته بسته به شکل بدنه‌ی آن‌ها و برخی از مهم‌ترین علائم و اعداد به ۴۰ کلاس تقسیم شده‌اند. تعدادی از حروف الفبای فارسی مانند «ب»، «پ»، «ت»، «ث» بدنه‌ی مشابهی دارند و تفاوت آن‌ها تنها در تعداد و جای قرار گرفتن نقاط آن‌ها است. در این صورت حروف با بدنه یکسان با یکی از این حروف به عنوان نماینده جایگزین می‌شوند و عمل استخراج ویژگی تنها برای بدنه‌ی حروف (بدون نقطه) انجام می‌شود. برخی کلاس‌ها مانند کلاس ۱۲ تنها شامل ۱ حرف و برخی دیگر مانند کلاس ۱۳ شامل ۴ حرف هستند. البته زیرکلماتی چون لا، لله، سی، صی به دلیل اینکه شکل خاصی دارند و در مرحله‌ی بازشناسی به راحتی قابل جداسازی نیستند، در حروف گسسته به عنوان کلاس‌هایی مجزا در نظر گرفته می‌شوند.

جدول ۴-۱ : کلاس‌های مورد استفاده در طبقه‌بند حروف گسسته‌ی فارسی

شماره کلاس	حرف	شماره کلاس	حرف	شماره کلاس	حرف	شماره کلاس	حرف
۱	۱	۱۱	آ	۲۱	ف	۳۱	لله
۲	۲	۱۲	ا	۲۲	ق	۳۲	لا
۳	۳	۱۳	ب پ ت ث	۲۳	گ	۳۳	ء
۴	۴	۱۴	ج ح خ چ	۲۴	ک	۳۴	.
۵	۵	۱۵	د ذ	۲۵	ل	۳۵	صی
۶	۶	۱۶	ر ز ژ	۲۶	م	۳۶	شی سی
۷	۷	۱۷	س ش	۲۷	ن	۳۷	؟
۸	۸	۱۸	ص ض	۲۸	و	۳۸	!
۹	۹	۱۹	ط ظ	۲۹	ه	۳۹	،
۱۰	۰	۲۰	ع غ	۳۰	ی	۴۰	زیرکلمه

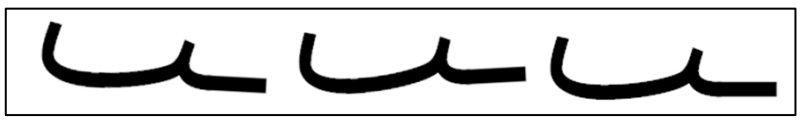
۴-۳- تولید پایگاه داده حروف پیوسته

برخلاف حروف گسسته، حروف پیوسته از لحاظ نحوه‌ی اتصال به حروف دیگر شامل تنوع بیشتری می‌باشد که در جدول زیر این تنوع نمایان شده است. همچنین یک کلاس به عنوان کلاس حروف نامعتبر در نظر گرفته می‌شود که شامل ترکیباتی است که تشکیل یک حرف نمی‌دهند یا شامل بیش از یک حرف هستند. اصولاً حروف پیوسته به سه نوع تقسیم می‌شوند: آغازین، میانه و پایانی که در جدول ۴-۲ نمایش داده شده است.

جدول ۴-۲ : کلاس‌های مورد استفاده در طبقه‌بند حروف پیوسته فارسی

شماره کلاس	حرف	شماره کلاس	حرف	شماره کلاس	حرف	شماره کلاس	حرف	شماره کلاس	حرف
۴۱	ـا	۵۱	ـگ	۶۱	ک	۷۱	ـد	۸۱	ط
۴۲	ـب	۵۲	گ	۶۲	ـک	۷۲	ـر	۸۲	ـط
۴۳	ـح	۵۳	ـگ	۶۳	ـل	۷۳	ـص	۸۳	ـو
۴۴	ـع	۵۴	ـی	۶۴	ـل	۷۴	صـ	۸۴	ی
۴۵	ـع	۵۵	حـ	۶۵	ـل	۷۵	ـصـ	۸۵	ـی
۴۶	عـ	۵۶	ـحـ	۶۶	ـم	۷۶	ـصی	۸۶	س
۴۷	ـد	۵۷	ـه	۶۷	مـ	۷۷	ـسی	۸۷	لا
۴۸	ـف	۵۸	هـ	۶۸	ـمـ	۷۸	سـ	۸۸	ناقص
۴۹	ـف	۵۹	ـهـ	۶۹	ـں	۷۹	ـسـ		
۵۰	ـف	۶۰	ـک	۷۰	ـہ	۸۰	ـط		

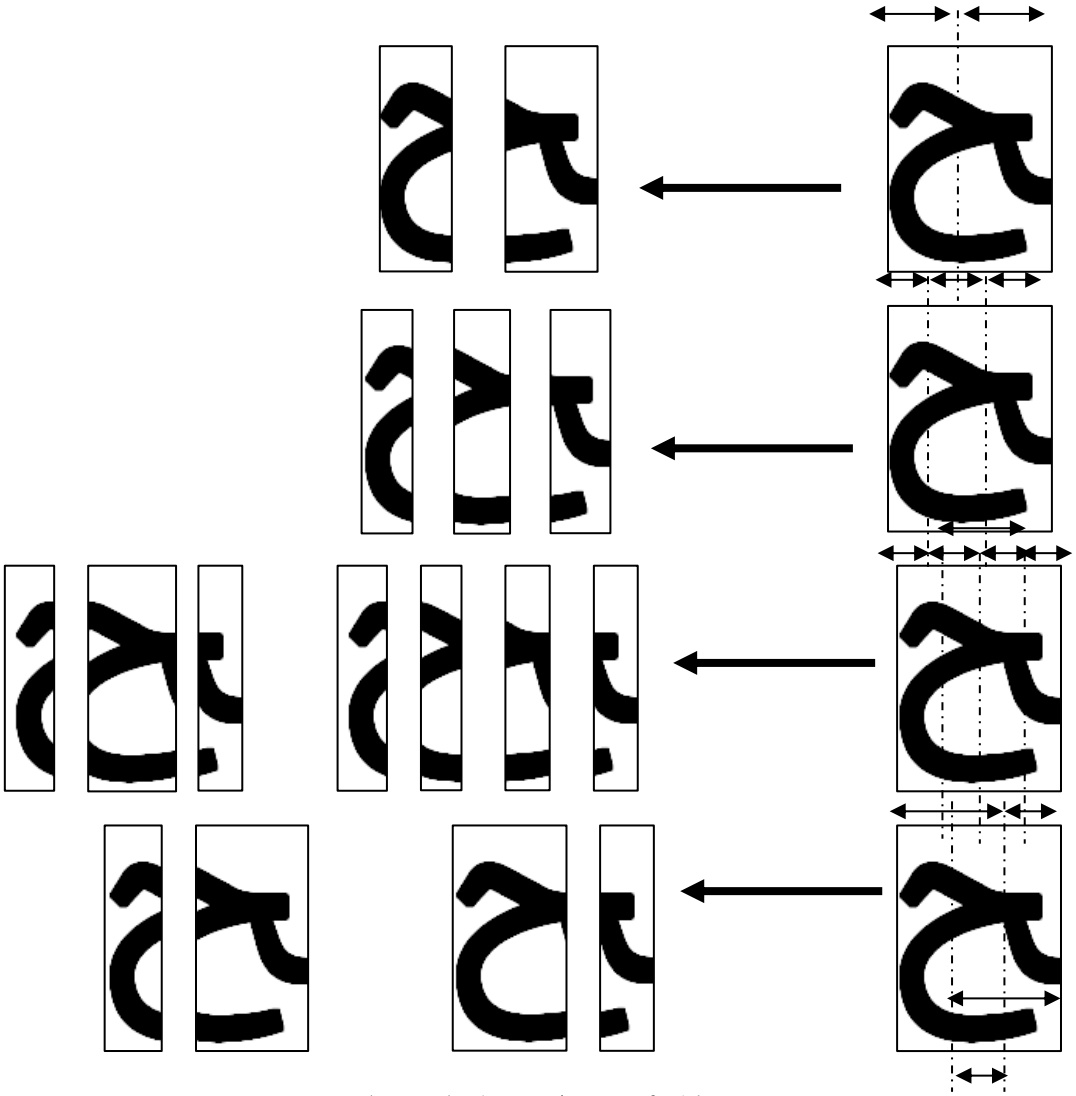
با توجه به اینکه نقاط جداسازی کلمه ممکن است در موقعیت‌های مختلف اتصال دو حرف قرار بگیرد، سعی می‌شود نمونه داده‌های تولید شده با طول‌های مختلفی باشند و همچنین برای تولید نمونه‌های متنوع‌تر از حروف، اندکی چرخش نیز در بعضی نمونه‌ها ایجاد می‌شود.



شکل ۴-۱۰ : چرخش ۱- و ۱ درجه برای تولید داده‌های متنوع‌تر

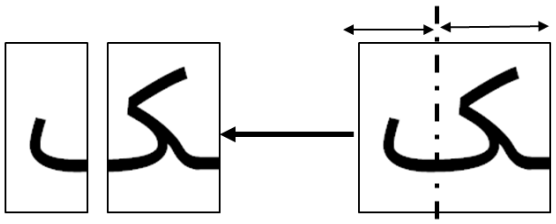
۴-۳-۱- تولید حروف ناقص پیوسته

تولید حروف ناقص بدین علت انجام می‌شود که در زمان جداسازی بتوان صحت جداسازی را رد و یا تایید کرد. برای تشکیل کلاس حروف نامعتبر از ناقص کردن حروف پیوسته‌ی معتبر استفاده می‌شود. کلاس نامعتبر شامل تمام حروف و زیرحروفی است که تشکیل یک حرف نمی‌دهند که نحوه‌ی تولید آن در شکل ۴-۱۱ نمایش داده شده است. مثلاً حرفی مانند "ح" را ابتدا به قسمت‌های مختلف تقسیم کرده و ترکیبی از این قسمت‌ها به عنوان یک زیرکلمه‌ی نامعتبر شناخته شده و در کلاس حروف ناقص قرار داده می‌شود.



شکل ۴-۱۱ : تولید حروف ناقص و نامعتبر

در هنگام تولید حروف ناقص با الگوریتم بالا مشاهده شد که برخی از اشکال تولید شده خود عضوی از حروف پیوسته معتبر است، در نتیجه باید از حروف ناقص حذف شوند مانند حرف کاف. همان‌طور که مشاهده می‌شود حرف کاف تولید شده دیگر ناقص نبوده و باید از این گروه حذف شود.



شکل ۴-۱۲ : حروف ناقص تولید شده معتبر

۴-۴-استخراج ویژگی‌ها

پس از تولید داده برای کلاس‌های ۱ تا ۴۰ و ۴۱ تا ۸۸، عملیات استخراج ویژگی برای داده‌های هر کدام از کلاس‌ها به صورت مستقل انجام می‌شود. استخراج ویژگی شامل کلیه محاسباتی است که روی نمونه‌های به دست آمده از مرحله‌ی پیش‌پردازش انجام می‌شود تا بردار ویژگی‌های متناظر با هر نمونه حاصل شود.

در اینجا برای استخراج ویژگی از روش هیستوگرام گرادیان استفاده شده است. این ویژگی که برای بازشناسی ارقام دست‌نویس فارسی معرفی شده است [۴۱]، با انجام تغییراتی برای زیرکلمات فارسی به کار گرفته شد. برای استخراج ویژگی ابتدا تصویر به اندازه‌ی N×N نرمال می‌شود، سپس گرادیان تصویر با استفاده از ماسک‌های سوبل محاسبه می‌شود. در این روش g_x گرادیان در جهت افقی و g_y گرادیان در جهت عمودی با استفاده از دو ماسک مجزا به دست می‌آید. برای هر نقطه از تصویر با استفاده از ماسک‌های فوق بردار گرادیان به صورت $[g_x\ g_y]$ محاسبه می‌گردد. شدت و زاویه‌ی گرادیان برای هر نقطه از تصویر، با به دست آوردن اندازه و جهت بردار فوق به دست می‌آید. پس از به دست آوردن زاویه‌ی گرادیان لازم است مقادیر زاویه به چند زاویه‌ی استاندارد تبدیل گردند. نحوه‌ی استخراج ویژگی برای زیرکلمات فارسی به صورت زیر است:

۱) نرمال کردن تصویر ورودی به اندازه‌ی ۶۰×۶۰ به گونه ای که نسبت ابعاد تصویر حفظ شود.

۲) اعمال فیلتر سوبل که به صورت زیر تعریف می‌شود:

$$g_x(x,y)=f(x+1,y-1)+2f(x+1,y)+f(x+1,y+1)-f(x-1,y-1)$$
$$-2f(x-1,y)-f(x-1,y+1)$$

$$g_y(x,y)=f(x-1,y+1)+2f(x,y+1)+f(x+1,y+1)-f(x-1,y-1)$$
$$-2f(x,y-1)-f(x+1,y-1)$$

رابطه (۴-۱)

از روی این دو مؤلفه‌ی گرادیان، اندازه و جهت گرادیان محاسبه می‌شود. برای ترکیب g_x و g_y ۱۶, حالت اتفاق می‌افتد که شامل زوایای $0,\pm\frac{\pi}{8},\pm\frac{\pi}{4},\pm\frac{3\pi}{8},\pm\frac{\pi}{2},\pm\frac{5\pi}{8},\pm\frac{3\pi}{4},\pm\frac{7\pi}{8},\pm\pi$ شده و یک حالت زمانی رخ می‌دهد که g_x و gy هر دو صفر بوده و زاویه تعریف نشده وجود داشته باشد.

۳) تصویر ورودی به ۱۶ بلوک مساوی ۱۵×۱۵ پیکسلی تقسیم می‌شود.

۴) در هر بلوک مقدار شدت گرادیان به ازای هر ۱۶ زاویه‌ی ذکر شده با هم جمع می‌شوند. به ازای زاویه‌ی تعریف نشده چون شدت گرادیان همواره صفر می‌شود و ویژگی خاصی به دست نمی‌دهد در نظر گرفته نمی‌شود. به این ترتیب در نهایت بردار ویژگی شامل ۴×۴×۱۶ یعنی ۲۵۶ مؤلفه می‌شود.

عمل استخراج ویژگی روی ۳۰۰ تصویر مجزای موجود در هر کلاس انجام شده و برای آموزش طبقه‌بندها استفاده می‌شود.

۴-۵- آموزش طبقه‌بندها

از ویژگی‌های استخراج شده در مرحله‌ی قبل برای آموزش طبقه‌بندها استفاده می‌شود. طبقه‌بندهای مورد استفاده در این پایان‌نامه از نوع شبکه‌های عصبی پرسپترون چند لایه[1] می‌باشد. دو نوع طبقه‌بند، یکی برای طبقه‌بندی حروف گسسته و زیرکلمات و دیگری برای حروف پیوسته آموزش

[1] MLP

داده خواهد شد، که وظیفه‌ی هر یک از آن‌ها مقایسه‌ی کلاس ورودی و نمونه‌ی مورد آزمایش و تشخیص کلاس است. طبقه‌بند گسسته شامل ۴۰ کلاس و طبقه‌بند پیوسته حاوی ۴۸ کلاس است.

۴-۵-۱- طبقه‌بند حروف گسسته و زیرکلمات

همان‌طور که در جدول ۴-۱ نشان داده شد، طبقه‌بند حروف گسسته در ۴۰ کلاس آموزش داده می‌شود، که وظیفه‌ی آن تشخیص کلاس ورودی حروف گسسته و زیرکلمات است. در این طبقه‌بند کلاس‌های ۱ تا ۳۹ برای حروف گسسته بوده و کلاس ۴۰ برای زیر کلمات می‌باشد.

برای استخراج ویژگی کلاس ۴۰ از ۶۶۵۰ زیرکلمه معتبر و موجود در زبان فارسی استفاده شده است. نمونه‌ای از این زیرکلمات در شکل ۴-۱۳ آمده است.

شکل ۴-۱۳ : نمونه‌ای از زیرکلمات معتبر و موجود در زبان فارسی

روند کار بدین ترتیب است که اگر ورودی به صورت زیرکلمه (کلاس ۴۰) شناسایی شود، برای جداسازی به مرحله‌ی بعد فرستاده می‌شود و اگر ورودی حرف یا علامت باشد، یکی از کدهای ۱ تا ۳۹ فعال می‌شوند.

۴-۵-۲- طبقه‌بند حروف پیوسته

طبقه‌بند حروف پیوسته مطابق جدول ۴-۲ شامل ۴۸ کلاس (کلاس‌های ۴۱ تا ۸۸) است. از این طبقه‌بند برای جداسازی زیرکلمات کلاس ۴۰ استفاده می‌شود. روش کار بدین‌صورت است که ابتدا زیرکلمات کلاس ۴۰ توسط طبقه‌بند حروف گسسته شناسایی شده و برای جداسازی به طبقه‌بند حروف پیوسته فرستاده می‌شود. در این مرحله با اعمال روش‌های مناسب، نقاط کاندید اولیه‌ی جداسازی مشخص شده و صحت این نقاط جداسازی با استفاده از طبقه‌بند حروف پیوسته مشخص می‌شود. اگر نقطه‌ی جداسازی معتبر باشد یکی از کلاس‌های ۴۱ تا ۸۷ انتخاب می‌شود و اگر نقطه‌ی جداسازی نامعتبر باشد کلاس ۸۸ انتخاب شده و نقطه‌ی جداسازی جدیدی انتخاب می‌شود.

۴-۶- نتیجه‌گیری

در این فصل با توجه به موجود نبودن پایگاه داده‌ی مناسب برای قلم Iranian sans اقدام به تولید پایگاه داده‌ای مناسب جهت آموزش طبقه‌بندهای حروف پیوسته و گسسته شد. در این فرایند پس از تولید سندهای مناسب و اسکن نمودن آن‌ها عملیات پیش‌پردازش بر روی آن‌ها انجام گرفت و با استفاده از روش هیستوگرام گرادیان ویژگی‌های لازم جهت آموزش طبقه‌بندها استخراج شد. طبقه‌بندهای مورد استفاده در این پایان‌نامه از نوع شبکه‌ی عصبی Mlp می‌باشد.

در فصل آتی از طبقه‌بند حروف گسسته برای طبقه‌بندی حروف جدا و شناسایی زیرکلمات استفاده کرده و از طبقه‌بند حروف پیوسته برای جداسازی زیرکلمات شناسایی شده در مرحله‌ی قبل استفاده خواهد شد.

فصل پنجم

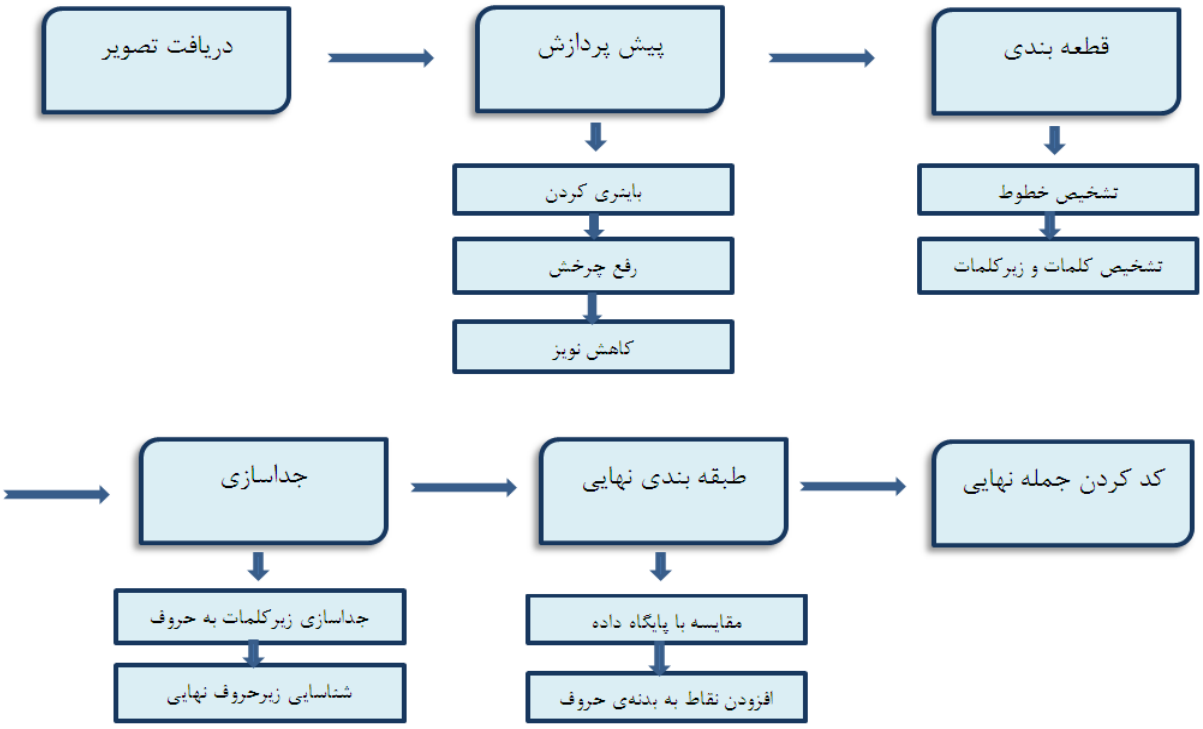
سامانه بازشناسی متون

با قلم Iranian sans

۵-۱- مقدمه

در فصل گذشته به تولید پایگاه داده جهت آموزش طبقه‌بندها پرداخته شد. بدین صورت که یک طبقه‌بند برای حروف گسسته و زیرکلمات و یک طبقه‌بند برای حروف پیوسته آموزش داده شد.

در این فصل به توضیح روند اجرای سیستم بازشناسی متن پرداخته می‌شود. همان‌طور که در شکل زیر نشان داده شده است، روند کار بدین صورت است که در مرحله اول پس از دریافت تصویر پیش‌پردازش‌هایی شامل باینری کردن، اصلاح چرخش و حذف نویز بر روی تصویر اعمال می‌شود و در مرحله جداسازی با شناسایی حدود مکانی جمله و با استفاده از روش‌های مناسب و رفع مشکلات همپوشانی به جداسازی حروف گسسته و زیرکلمات پرداخته می‌شود. با استفاده از طبقه‌بند حروف گسسته، حروف مجزا از زیر کلمات جدا شده و با توجه به شماره‌ی کلاس آن ذخیره می‌شود. زیرکلمات شناسایی شده در این مرحله به قسمت جداسازی حروف رفته و با روش‌های مختلف این جداسازی انجام می‌شود. سپس با استفاده از طبقه‌بند حروف پیوسته به شناسایی حروف پرداخته می‌شود. در نهایت این حروف در کنار یکدیگر قرار گرفته و کد می‌شوند و کلمه و در نهایت جمله بازیابی می‌شود.



شکل ۵-۱ : شمای کلی سیستم بازشناسی متن

۵-۲-بازشناسی متن

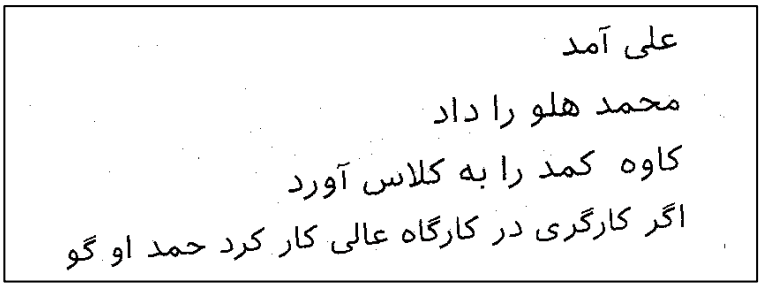
هر یک از اجزای متن با یک بردار ویژگی مشخص می‌شود. هر سیستم بازشناسی متن، این بردارها را در مرحله‌ی آموزش فرا می‌گیرد و به عنوان داده‌های معتبر در حافظه خود نگهداری می‌کند (آموزش طبقه‌بند). برای بازشناسی تصویر یک جزء متن، بردار ویژگی متناظر با آن با همان روش استخراج در مرحله‌ی آموزش، از این تصویر ناشناخته استخراج می‌شود. این بردار ویژگی با داده‌های سیستم مقایسه و بازشناسی می‌شود.

۵-۲-۱- پیش‌پردازش

عملیات پیش‌پردازش فرآیندی برای ارتقای تصویر ورودی است که آن را برای تحلیل و بازشناسی در مراحل بعد مهیا می‌سازد. از مهم‌ترین مراحل پیش‌پردازش می‌توان به دوسطحی سازی، بهبود تصویر با حذف نویز و اصلاح چرخش نام برد.

۵-۲-۱-۱ باینری کردن تصویر

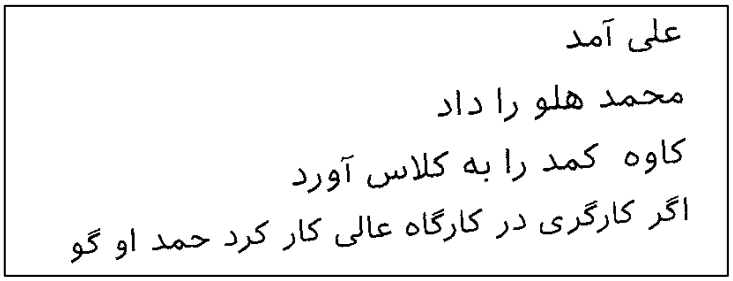
عموماً تصاویر پس از اسکن شدن به صورت خاکستری یا رنگی هستند با توجه به اینکه بسیاری از الگوریتم‌های پردازش تصویر از شکل دوسطحی تصاویر استفاده می‌کنند، نیاز است تصویر خاکستری به دوسطحی تبدیل شود. در اینجا با استفاده از یک سطح آستانه ثابت تصویر ورودی به دوسطحی تبدیل می‌شود.



شکل ۵-۲: باینری کردن تصویر

۵-۲-۱-۲ کاهش نویز

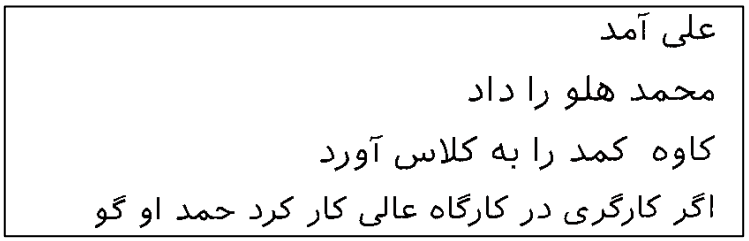
با توجه به اینکه تصاویر استفاده شده از کاغذ و توسط اسکنر پدید آمده است، به دلایل مختلف (که برخی از آن‌ها در قبل توضیح داده شد) دارای نویز می‌باشد. با توجه به اینکه حذف نویز گام موثری در بهبود بازشناسی دارد از روشی مشابه فصل گذشته برای حذف نویز استفاده می‌شود که شامل دو مرحله است. در مرحله‌ی اول با روش میانگین‌گیری بلوکی و تفاضل پیکسل مرکزی از پیکسل‌ها مقداری از نویز کاسته می‌شود و سپس در مرحله‌ی دوم با استفاده از روش برچسب‌زنی مؤلفه‌ها نویز به صورت کامل حذف می‌شود. استفاده از مرحله‌ی دوم به صورت مستقیم به دلیل وجود نویز فراوان نیاز به پردازش زیادی داشته و زمان‌بر است؛ لذا ابتدا در مرحله‌ی اول مقداری از نویز حذف شده و در مرحله‌ی دوم نویز به صورت کامل حذف می‌شود.



شکل ۵-۳ : حذف نویز در دو مرحله

۵-۲-۱-۳ اصلاح چرخش

تصویر اغلب اوقات توسط اسکنرها دریافت می‌شود که ممکن است به دلیل خطای انسانی در گذاشتن صفحات، تصویر سند با محورهای تصویر تراز نباشد یا ممکن است تصویر روبروی دوربین درست قرار نگیرد که باعث ایجاد زاویه‌ای در تصویر می‌شود و می‌تواند در نرخ بازشناسی تأثیر قابل ملاحظه‌ای بگذارد. برای رفع چرخش در تصویر دریافتی از افکنش افقی تصویر استفاده شده است. پس از یافتن زاویه‌ی چرخش، تصویر با همان زاویه و در خلاف جهت چرخانده می‌شود تا چرخش تصویر حذف شود.



شکل ۵-۴ : اصلاح چرخش به روش افکنش افقی

۵-۲-۲- قطعه‌بندی

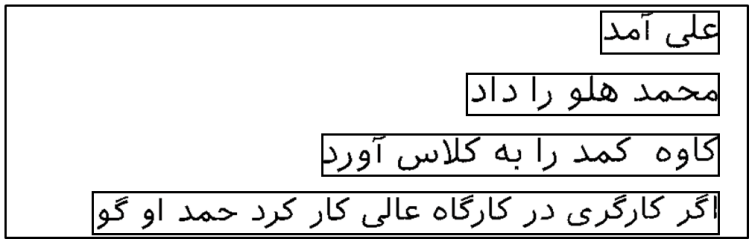
قطعه‌بندی مرحله‌ای مهم است که در آن اجزای تصویر متنی که باید به مرحله شناسایی تحویل داده شوند، از یکدیگر جدا می‌شوند. قطعه‌بندی در سه سطح انجام می‌شود:

- تشخیص خطوط
- تشخیص کلمات یا زیرکلمات
- تشخیص کاراکترها

۵-۲-۲-۱ تشخیص خطوط

متون فارسی اغلب به صورت مجموعه‌ای از خطوط موازی هستند که با فاصله نوشته می‌شوند. الگوریتم جداسازی خطوط، معمولاً از این فاصله‌ی طبیعی استفاده می‌کند. هیستوگرام افقی با شمارش عناصر سیاه تصویر در هر ردیف محاسبه شده و ردیف‌هایی که مقدار هیستوگرام در آن‌ها صفر یا از حد آستانه‌ای کمتر باشد به عنوان ردیف‌های بین خطوط در نظر گرفته می‌شوند. در واقع هر ردیفی از متن که بیشترین فراوانی را داشته باشد نمایانگر خطوط اصلی متن است.

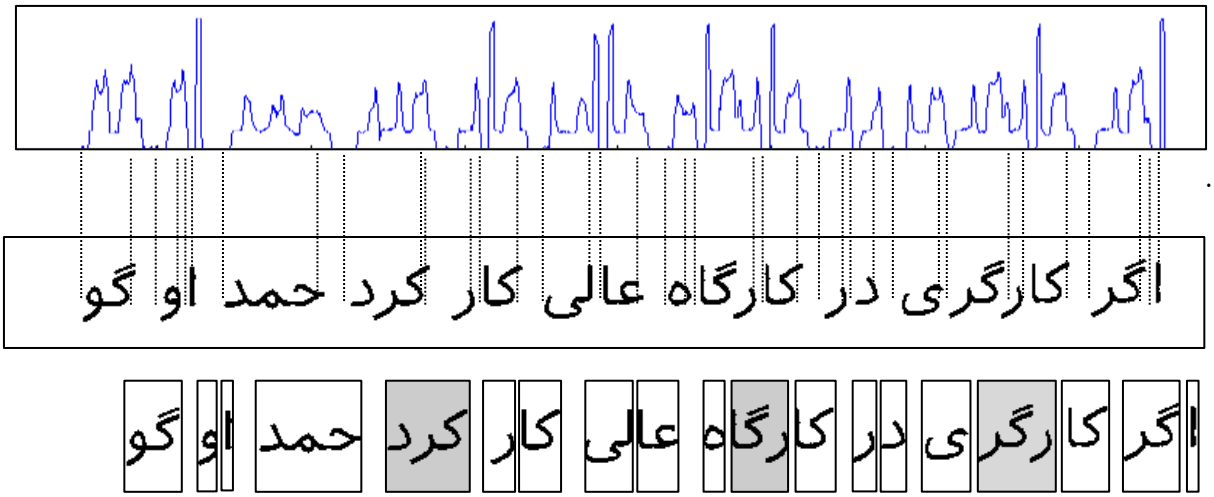
با استفاده از این روش، نقاط و علائمی که بین آن‌ها و خطوط اصلی متن ردیف‌های خالی وجود دارد، به عنوان یک خط جدید شناسایی می‌شوند. این خطا با در نظر گرفتن حد آستانه‌ای برای ارتفاع هیستوگرام افقی خطوط متن برطرف می‌گردد و بخش‌هایی که هیستوگرام افقی آن‌ها از حد آستانه‌ای کمتر باشد به نزدیک‌ترین خط مجاور خود ملحق می‌شوند.



شکل ۵-۵ : قطعه بندی جمله‌ها

۵-۲-۲-۲ تشخیص کلمات و زیرکلمات

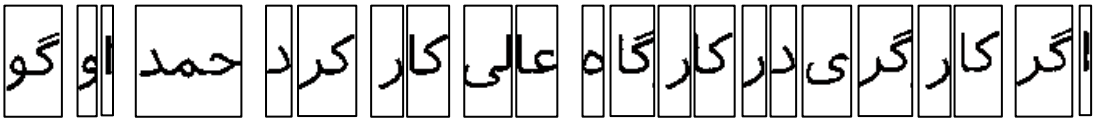
بعد از جداسازی خطوط، سیستم باید هر خط را به زیرکلمات تشکیل دهنده‌ی آن تفکیک نماید. بدین صورت که کلمات و کلمات فرعی درون خط با استفاده از هیستوگرام عمودی تعیین می‌گردد. در این حالت ستون‌هایی که مجموع پیکسل‌های آن صفر هستند محل جداسازی کلمات و زیرکلمات را تعیین می‌کند. اشکال این روش این است که اگر چند زیرکلمه با هم همپوشانی داشته باشند (شکل ۵-۶)، کل آن‌ها یک زیرکلمه در نظر گرفته می‌شوند، چون هیستوگرام صفر ایجاد نمی‌شود.



شکل ۵-۶ : مشکل وجود همپوشانی در جداسازی زیرکلمه‌ها به روش هیستوگرام عمودی

همان‌طور که مشاهده می‌شود به دلیل وجود همپوشانی کلمات، زیرکلمات به درستی تفکیک نمی‌شوند. در این حالت برای برطرف نمودن این مشکل از برچسب‌زنی مؤلفه‌ها استفاده می‌شود. در بدترین حالت این همپوشانی بین سه زیر کلمه متوالی روی می‌دهد. در این حالت ابتدا برچسب‌زنی مؤلفه‌ها انجام شده و سپس درصد همپوشانی مؤلفه‌ها با مؤلفه‌ی اصلی محاسبه می‌شود. این کار برای آن است که ابتدا تشخیص داده شود این مؤلفه جزئی از خود زیرکلمه است (مانند نقطه و یا سرکش گاف و علامت مد در آ) و یا زیر کلمه‌ی مجاور است. در حقیقت با قواعد مناسبی نقاط و علائم به بدنه‌ی زیرکلمه نسبت داده می‌شود. در صورتی که این همپوشانی نزدیک به ۱۰۰٪ باشد، معلوم می‌شود که این مؤلفه جزئی از زیرکلمه است و در غیر این صورت این مؤلفه زیرکلمه‌ی مجاور بوده که دارای همپوشانی شده است. در این حالت ابتدا تصویر مؤلفه‌ی مرکزی و یا به عبارتی بدنه‌ی زیر کلمه را از تصویر جدا کرده تا عامل همپوشانی حذف گردد و با استفاده از هیستوگرام عمودی جداسازی مؤلفه اول و آخر نیز انجام می‌شود.

شکل ۵-۷ : برطرف نمودن مشکل همپوشانی با استفاده از برچسب زنی مناسب مؤلفه‌ها



شکل ۵-۸ : برطرف نمودن مشکل همپوشانی

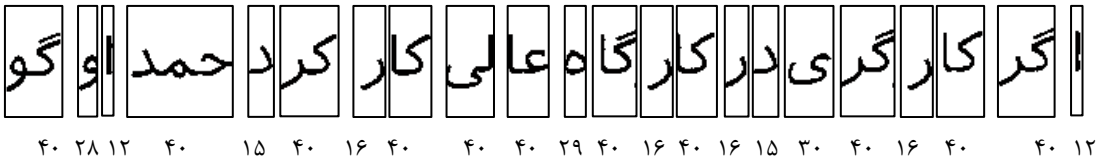
بدین صورت عمل جداسازی به زیرکلمات و حروف گسسته انجام می‌شود. حال قطعات تولید شده‌ی مذکور حذف نقطه می‌شوند. در این فرایند از عرض قلم به عنوان معیار مناسبی برای حذف نقطه استفاده می‌شود.

۵-۲-۲-۲-۱ عرض قلم

برای یافتن عرض قلم کافیست در هر ستون تصویر از برچسب‌زنی مؤلفه‌ها استفاده کرد که در این صورت طول عمودی هر مؤلفه (تعداد پیکسل‌های سیاه) حاصل می‌شود. برداری تشکیل داده می‌شود که هر درایه‌ی آن بیانگر تعداد تکرارهای طول می‌باشد. طولی که تعداد تکرار آن از همه بیشتر باشد به عنوان عرض قلم در نظر گرفته می‌شود. بدین ترتیب با محاسبه‌ی عرض قلم و برچسب‌زنی مؤلفه‌ها، مؤلفه‌هایی که دارای طول و عرضی کمتر از عرض قلم باشد از تصویر حذف می‌شود تا برای شناسایی به قسمت طبقه‌بند حروف گسسته فرستاده شود.

۵-۲-۲-۲-۲ شناسایی حروف گسسته و زیرکلمات

پس از بر طرف نمودن مشکل همپوشانی و حذف نقاط، حروف و زیرکلمات تولید شده با استفاده از طبقه‌بند حروف گسسته شناسایی می‌شود. بدین ترتیب که حروف گسسته با توجه به جدول ۴-۱، از شماره ۱ تا ۳۹ کلاسه‌بندی شده و همچنین تصاویر شناسایی شده که در کلاس ۴۰ (زیرکلمه) قرار می‌گیرند جهت جداسازی به قسمت جداسازی فرستاده می‌شود.



شکل ۵-۹ : شماره کلاس مطابق جدول ۴-۱

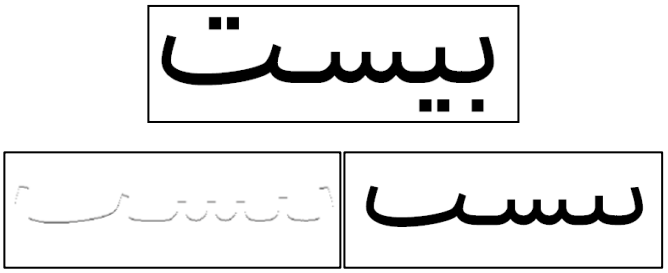
۵-۲-۳-جداسازی

به طور کلی برای بازشناسی متون سه رویکرد عمده وجود دارد: روش‌های مبتنی بر جداسازی، روش‌های مبتنی بر شکل کلی و روش‌های ترکیبی. در این پایان نامه از رویکرد مبتنی بر جداسازی استفاده شده است. جداسازی به این معنی است که یک زیرکلمه به حروف سازنده‌اش تجزیه شود. برای یافتن نقاط جداسازی روش‌های مختلفی وجود دارد که به دلیل پاره‌ای از محدودیت‌ها از جمله همپوشانی حروف، شباهت بسیاری از زیرحروف به حروف و عوامل دیگر، تعیین نقاط دقیق جداسازی کار ساده ای نیست. در این روش برای جداسازی حروف از روش پروفایل بالایی تعمیم یافته استفاده شده است. برای اینکه نقاط جداسازی مطمئن باشد، باید ابتدا نقاط اولیه‌ی جداسازی تعیین شود، سپس به وسیله‌ی موتور بازشناسی (طبقه‌بند حروف پیوسته) که در مرحله‌ی تولید داده آموزش داده شد، نقاط نامعتبر حذف شده و نتایج جداسازی بهبود می‌یابد.

۵-۲-۴- جداسازی حروف به روش پروفایل بالایی تعمیم یافته

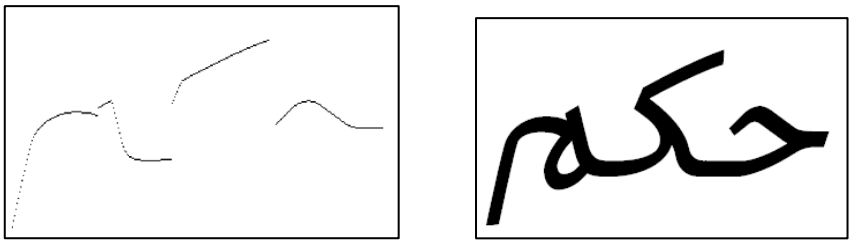
در راستای جداسازی حروف به روش پروفایل بالایی پس از حذف نقاط و علائم ابتدا با استفاده از روش گذر از سفید به سیاه به تشکیل پروفایل بالایی و لبه‌های بالایی تصویر پرداخته می‌شود. روش کار بدین صورت است که پس از تشکیل پروفایل بالایی و محاسبه‌ی مینیمم‌های محلی و استفاده از روش برچسب‌زنی مؤلفه‌ها به بررسی نقاط کاندید برای جداسازی پرداخته می‌شود. به عنوان مثال

برای زیرکلمه‌ی "بیست" در مرحله‌ی اول پس از حذف نقاط به محاسبه‌ی پروفایل بالایی پرداخته می‌شود.



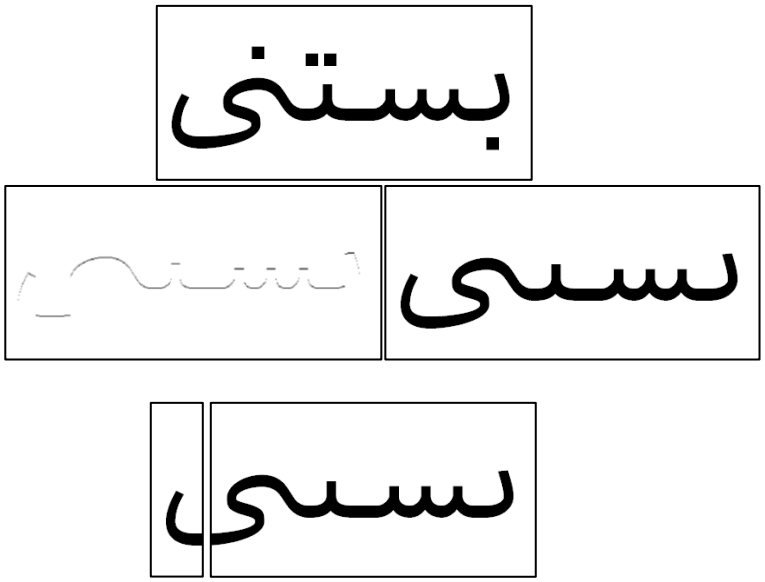
شکل ۵-۱۰ : حذف نقاط، محاسبه لبه بالایی، تعیین مینیمم‌های محلی کاندید برای جداسازی

در استفاده از این روش به تنهایی چند مشکل عمده وجود دارد. اول آنکه با توجه به قرار گرفتن بعضی حروف در زیر حروف دیگر (همپوشانی حروف) پس از اعمال روش پروفایل بالایی نمی‌توان عمل جداسازی را به درستی انجام داد. مثلاً در کلمه‌ای مانند حکم به علت وجود سرکش حرف "ـکـ" که روی حرف "حـ" را پوشانده، محاسبه‌ی مینیمم محلی برای جداسازی، عملی غیر ممکن است و منجر به جداسازی نادرست می‌شود.



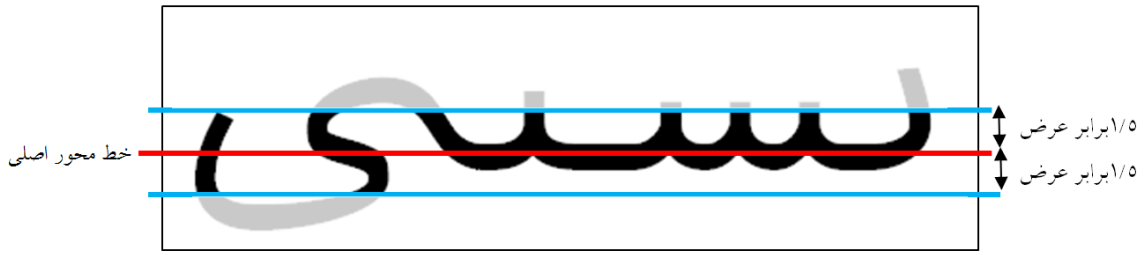
شکل ۵-۱۱ : عدم وجود مینیمم محلی بین حرف "ح" و "ک" به دلیل همپوشانی

از طرفی با توجه به آنکه نقطه‌ی جداسازی باید در محدوده‌ی خط زمینه باشد، برای تعیین نقاط جداسازی مناسب، باید محدوده‌ای مشخص شود. در غیر این صورت تعیین نقاط جداسازی به درستی انجام نمی‌شود. مثلاً مطابق شکل ۵-۱۲ حرف "ی" به صورت نامناسب جدا خواهد شد.



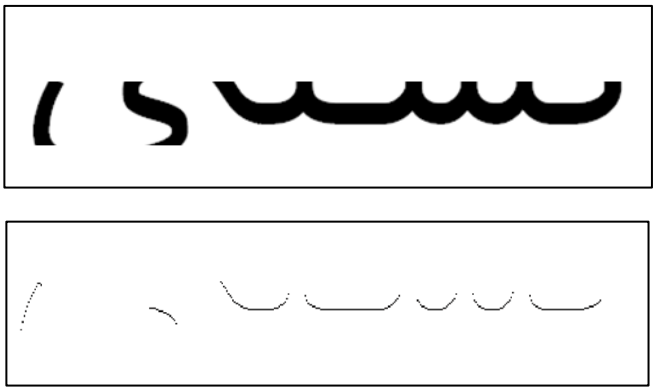
شکل ۵-۱۲ : جداسازی نامناسب حرف ی به دلیل عدم وجود محدوده مناسب

با توجه به مشکلات مذکور باید ابتدا محدوده ای مناسب برای جداسازی مشخص شود. به این ترتیب که ابتدا خط اصلی متن با استفاده از هیستوگرام افقی محاسبه شده و سپس با محاسبه‌ی عرض قلم، محدوده‌ای به اندازه یک و نیم برابر عرض قلم پایین و بالای خط کرسی، به عنوان محدوده‌ی مجاز برای محاسبه‌ی نقاط جداسازی در نظر گرفته شده و پروفایل بالایی فقط برای این محدوده از تصویر در نظر گرفته می‌شود. بدین ترتیب علاوه بر برطرف نمودن مشکل همپوشانی، از جداسازی نامناسب در خارج محدوده‌ی خط مبنا جلوگیری به عمل خواهد آمد.



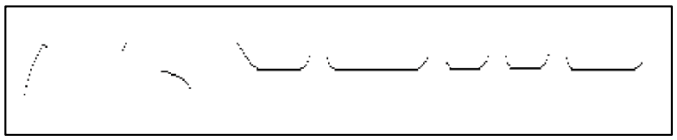
شکل ۵-۱۳ : مشخص نمودن محدوده مجاز برای محاسبه پروفایل بالایی و محل جداسازی

پس از این مرحله محاسبه‌ی پروفایل بالایی برای تصویر حاصل انجام می‌شود که حاصل این کار درشکل ۵-۱۴ آمده است.



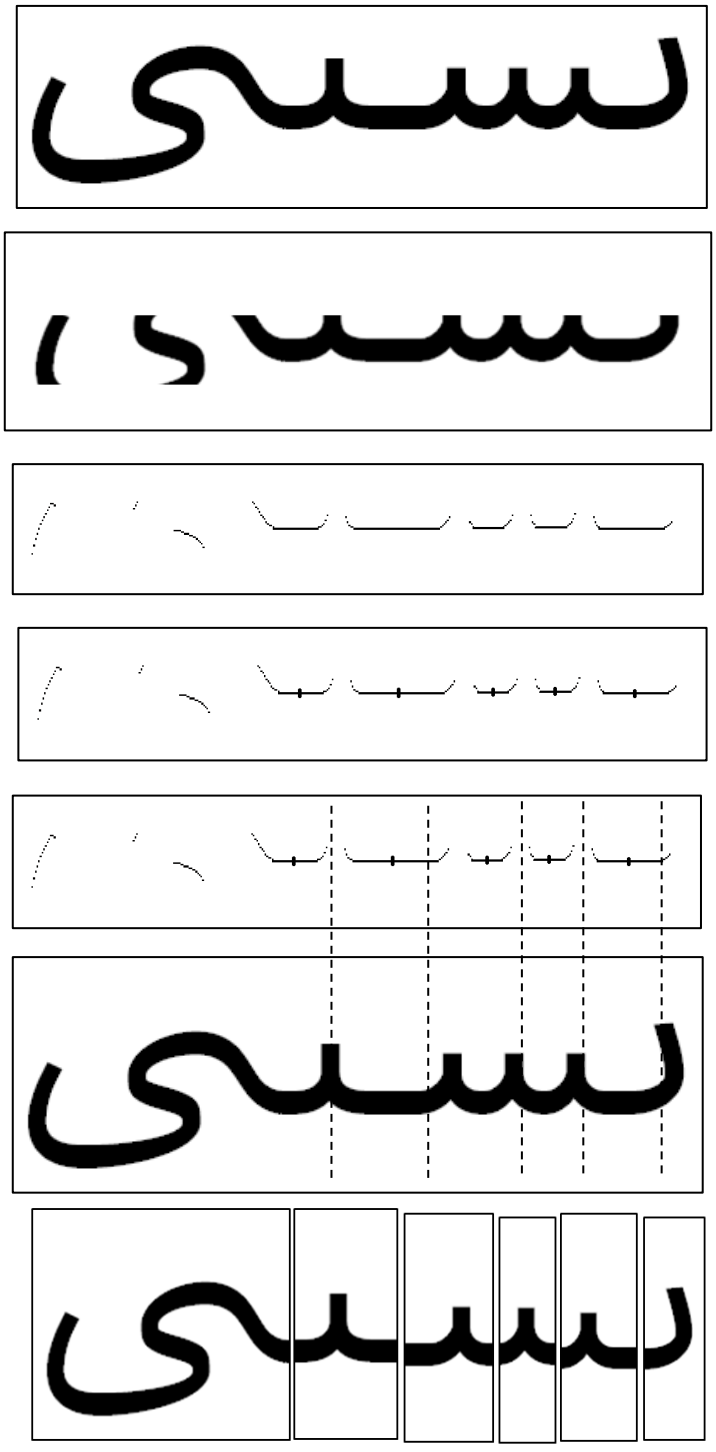
شکل ۵-۱۴ : محاسبه پروفایل بالایی تصویر

پس از این مرحله جهت محاسبه‌ی کاندیدهای مناسب جداسازی در محدوده‌ی مینیمم محلی، مکان‌هایی که به اندازه یک پیکسل با مرکز مینیمم محلی فاصله دارند را در یک راستا قرار داده تا در مرحله‌ی بعد نقاط جداسازی بهینه‌تری حاصل شود.



شکل ۵-۱۵: بهینه نمودن مینیمم‌های محلی

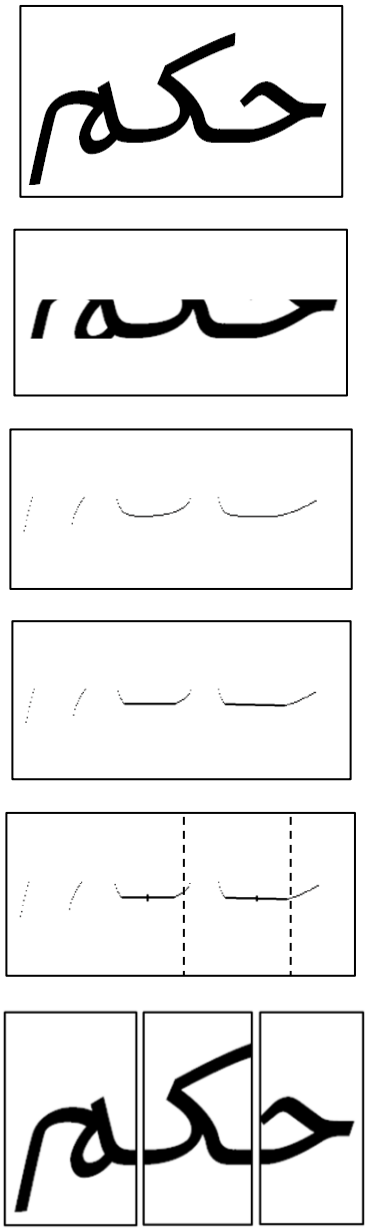
سپس در این مرحله با اعمال هیستوگرام افقی، سه سطری که بیشترین فراوانی‌ها را داشته باشند در نظر گرفته می‌شود. هر سطر را به صورت برداری ذخیره نموده که حاوی تکه خط‌های افقی می‌باشد. سرانجام با استفاده از برچسب‌زنی مؤلفه‌ها وسط تکه خط‌های افقی ایجادشده به عنوان کاندید اولیه‌ی جداسازی در نظر گرفته می‌شوند و با این شرط که فاصله‌ی این نقاط از هم نباید از عرض قلم کمتر باشد، این نقاط به عنوان نقاط جداسازی خواهند بود. بدین ترتیب اگر فاصله‌ی دو نقطه‌ی جداسازی از عرض قلم کمتر بود، آن که امتداد بزرگ‌تری دارد، نقطه‌ی جداسازی در نظر گرفته می‌شود.



شکل ۵-۱۶ : تعیین نقاط جداسازی اولیه

بدین ترتیب نقاط جداسازی اولیه تعیین می‌شود و حروف و زیرحروف ایجادشده به شبکه‌ی بازشناسی حروف پیوسته سپرده می‌شود تا صحت این جداسازی تعیین و در نهایت کلاسه‌بندی شده

و به مرحله‌ی بعد فرستاده شود. در حقیقت با مشخص کردن ناحیه‌ی مجاز برای جداسازی مطابق شکل ۵-۱۷ هم مشکل همپوشانی (شکل ۵-۱۱) و هم مشکل جداسازی حروفی مانند "ی" که نقاط جداسازی غیرمجاز (شکل ۵-۱۲) تولید می‌کند، حل می‌شود.

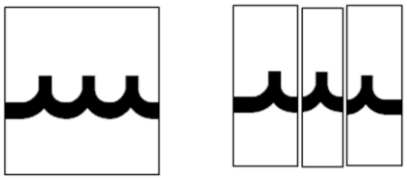


شکل ۵-۱۷ : برطرف کردن مشکل همپوشانی

۵-۲-۴-۱ شناسایی زیرحروف نهایی

در مرحله‌ی قبل نقاط جداسازی اولیه مشخص شده و تصویر به تعدادی زیرکلمه تقسیم می‌شود.

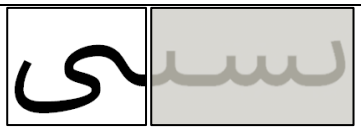
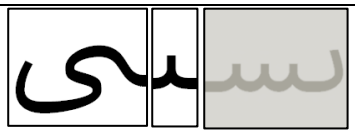
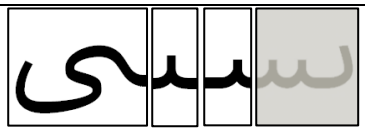
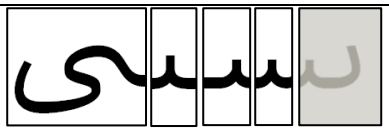
اما هنوز معتبر بودن نقطه‌ی جداشده مشخص نشده است. مثلاً مطابق شکل ۵-۱۸ حرف "س" به تعدادی دندانه تقسیم شده است که این جداسازی معتبر نمی‌باشد.



شکل ۵-۱۸ : جداسازی نادرست حرف سین در کلمه بستنی

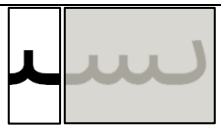
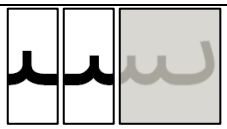
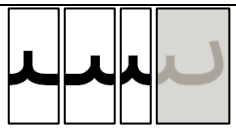
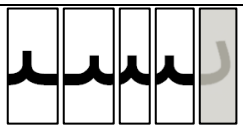
بنابراین لازم است صحت جداسازی مرحله‌ی قبل بررسی شود. در این مرحله ابتدا جداسازی حروف کلمه‌ی **"بستنی"** مطابق شکل ۵-۱۵ و با استفاده از روش پروفایل بالایی انجام می‌شود، سپس از سمت چپ قطعات تولید شده در کنار یکدیگر قرارگرفته و تشکیل یک تصویر جدید می‌دهد. سپس تصویر حاصله برای شناسایی به شبکه بازشناسی حروف پیوسته فرستاده شده و اعتبارسنجی می‌شود. این کار برای چهار تصویر اول انجام می‌شود و اعتبار هر مرحله ثبت می‌شود.

جدول ۵-۱: شناسایی و اعتبارسنجی نقاط جداسازی (مرحله اول)

شماره مرحله	کلاس شناسایی شده مطابق جدول ۴-۲	دقت شناسایی	تصویر حاصل
اول	کلاس ۸۵ (زیر بدنه "یی")	٪۹۹	
دوم	کلاس ۸۸ (حرف ناقص)	٪۹۰	
سوم	کلاس ۸۸ (حرف ناقص)	٪۹۵	
چهارم	کلاس ۸۸ (حرف ناقص)	٪۹۸	
نتیجه	کلاس ۸۵ (زیر بدنه "یی")	٪۹۹	

در مرحله‌ی اول جداسازی مشاهده می‌شود که با دقت ۹۹٪ محل جداسازی درست بوده و مطابق جدول ۴-۲ کلاس ۸۵ شناخته می‌شود. در مرحله‌ی دوم با اضافه نمودن قطعه‌ی بعدی به این تصویر، تصویر جدید با دقت ۹۰٪ نامعتبر و جزو حروف ناقص شناخته می‌شود. این کار تا ۴ مرحله انجام می‌شود وبا دقت ۹۵ و ۹۸ درصد، تصویر ناقص تشخیص داده می‌شود. بنابراین در مجموع این ۴ مرحله، تصویر مرحله اول به عنوان کلاس ۸۵ شناخته می‌شود و از تصویر اصلی جدا شده و این روند برای باقیمانده‌ی تصویر تکرار می‌شود.

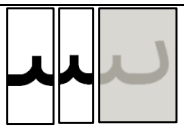
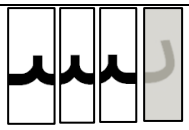
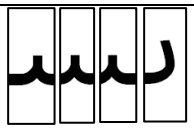
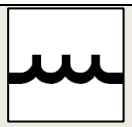
جدول ۵-۲: شناسایی و اعتبارسنجی نقاط جداسازی (مرحله دوم)

شماره مرحله	کلاس شناسایی شده مطابق جدول ۴-۲	دقت شناسایی	تصویر حاصل
اول	کلاس ۷۱ (دندانه)	٪۹۷	
دوم	کلاس ۸۸ (حرف ناقص)	٪۹۰	
سوم	کلاس ۷۹ (ـسـ)	٪۸۰	
چهارم	کلاس ۸۸ (حرف ناقص)	٪۹۸	
نتیجه	کلاس ۷۱ (دندانه)	٪۹۷	

در مرحله‌ی دوم، جداسازی با دقت ۹۷٪ درست بوده و مطابق جدول ۴-۲ تصویر حاصل کلاس ۷۱ شناخته می‌شود. در مرحله دوم با اضافه نمودن قطعه‌ی بعدی به این تصویر، تصویر جدید با دقت ۹۰٪ نامعتبر و جزو حروف ناقص شناخته می‌شود. در مرحله‌ی سوم تصویر حاصله با اعتبار ۸۰٪ حرف (ـسـ) و کلاس ۷۹ تشخیص داده می‌شود (به دلیل تشابه زیاد با حرف س) و همچنین در مرحله

چهارم با اضافه شدن قطعه بعدی، با دقت ۹۸٪ حرف ناقص شناخته می‌شود. بدین ترتیب در مجموع با اعتبار ۹۷٪ کلاس ۷۱ شناخته و ذخیره شده و از تصویر اصلی جدا می‌شود و بقیه‌ی تصویر به مرحله‌ی بعد فرستاده می‌شود.

جدول ۵-۳: شناسایی و اعتبارسنجی نقاط جداسازی (مرحله سوم)

شماره مرحله	کلاس شناسایی شده مطابق جدول ۴-۲	دقت شناسایی	تصویر حاصل
اول	کلاس ۷۱ (دندانه)	٪۹۰	
دوم	کلاس ۸۸ (حرف ناقص)	٪۹۷	
سوم	کلاس ۷۹ (حرف ـسـ)	٪۱۰۰	
چهارم	کلاس ۸۸ (حرف ناقص)	٪۹۸	
نتیجه	کلاس ۷۹ (حرف ـسـ)	٪۱۰۰	

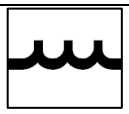
در مرحله‌ی سوم، جداسازی اولیه با دقت ۹۰٪ بوده و مطابق جدول ۴-۲ تصویر حاصل کلاس ۷۱ و یا به عبارتی به صورت یک دندانه شناخته می‌شود. در مرحله‌ی دوم با اضافه نمودن قطعه‌ی بعدی به این تصویر، تصویر جدید با دقت ۹۷٪ نامعتبر و جزو حروف ناقص شناخته می‌شود. در مرحله‌ی سوم تصویر حاصله با اعتبار ۱۰۰٪ حرف (ـسـ) و کلاس ۷۹ تشخیص داده می‌شود و همچنین در مرحله‌ی چهارم با اضافه شدن قطعه‌ی بعدی، با دقت ۹۸٪ حرف ناقص شناخته می‌شود. بدین ترتیب در مجموع با اعتبار ۱۰۰٪ کلاس ۷۹ و یا حرف (ـسـ) شناخته، ذخیره و از تصویر اصلی جداشده و

بقیه‌ی تصویر به مرحله بعد فرستاده می‌شود. از بقیه‌ی تصویر تنها ۱ قطعه باقیمانده که با دقت ٪۹۵ کلاس ۷۰ شناخته می‌شود.

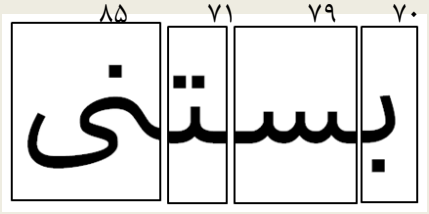
مرحله	کلاس شناسایی شده	دقت شناسایی	
اول	کلاس ۷۰	٪۹۵	

نتیجه نهایی جداسازی و طبقه‌بندی تصویر در جدول ۵-۴ آمده است.

جدول ۵-۴: : نتیجه کلی جداسازی کلمه بستنی

شماره تصویر	کلاس شناسایی شده مطابق جدول ۴-۲	دقت شناسایی	تصویر حاصل
اول	کلاس ۸۵	٪۹۹	
دوم	کلاس ۷۱	٪۹۷	
سوم	کلاس ۷۹ (ـسـ)	٪۱۰۰	
چهارم	کلاس ۷۰	٪۹۵	

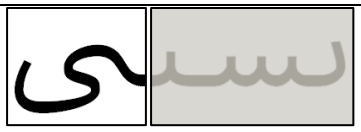
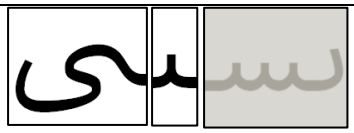
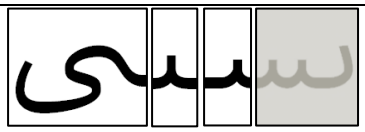
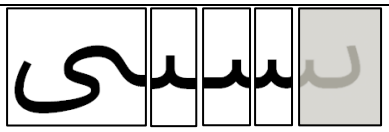
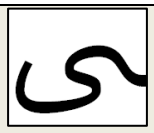
شماره کلاس زیر بدنه: ۷۰ ۷۹ ۷۱ ۸۵



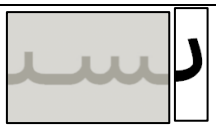
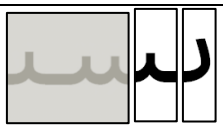
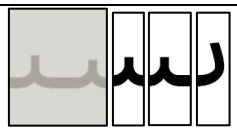
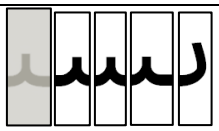
شماره کلاس حروف با نقطه: ؟ ؟ ؟ ؟

برای بهبود عملکرد می‌توان عمل جداسازی را در هر مرحله از طرف دیگر ادامه داد. بدین صورت که ابتدا از سمت چپ شروع کرده و با شناسایی و جداسازی تصویر اول، شناسایی در مرحله‌ی بعد باید از سمت راست ادامه یابد. این کار تا آخرین مرحله‌ی جداسازی حروف انجام می‌شود. مزیت این روش آن است که در صورت بروز اشکال در یکی از مراحل جداسازی، عمل جداسازی متوقف نمی‌شود و از سمت دیگر ادامه می‌یابد.

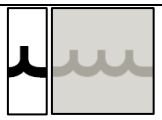
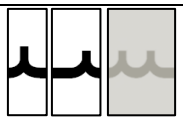
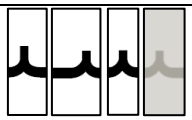
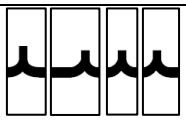
جدول ۵-۵: مرحله اول جداسازی از دو سمت کلمه بستنی

شماره مرحله	کلاس شناسایی شده مطابق جدول ۴-۲	دقت شناسایی	تصویر حاصل
اول	کلاس ۸۵ (زیر بدنه "یی")	٪۹۹	
دوم	کلاس ۸۸ (حرف ناقص)	٪۹۰	
سوم	کلاس ۸۸ (حرف ناقص)	٪۹۵	
چهارم	کلاس ۸۸ (حرف ناقص)	٪۹۸	
نتیجه	کلاس ۸۵ (زیر بدنه "یی")	٪۹۹	

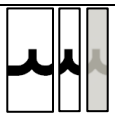
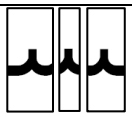
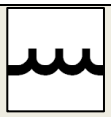
جدول ۵-۶: مرحله دوم جداسازی از دو سمت کلمه بستنی

شماره مرحله	کلاس شناسایی شده مطابق جدول ۴-۲	دقت شناسایی	تصویر حاصل
اول	کلاس ۷۰	٪۹۵	
دوم	کلاس ۸۸ (حرف ناقص)	٪۹۳	
سوم	کلاس ۸۸ (حرف ناقص)	٪۹۵	
چهارم	کلاس ۸۸ (حرف ناقص)	٪۱۰۰	
نتیجه	کلاس ۸۵ (زیر بدنه "یی")	٪۹۹	

جدول ۵-۷: مرحله سوم جداسازی از دو سمت کلمه بستنی

شماره مرحله	کلاس شناسایی شده مطابق جدول ۴-۲	دقت شناسایی	تصویر حاصل
اول	کلاس ۷۱ (دندانه)	٪۹۷	
دوم	کلاس ۸۸ (حرف ناقص)	٪۹۰	
سوم	کلاس ۷۹ (حرف ـسـ)	٪۸۰	
چهارم	کلاس ۸۸ (حرف ناقص)	٪۹۸	
نتیجه	کلاس ۷۱ (دندانه)	٪۹۷	

جدول ۵-۸: مرحله چهارم جداسازی از دو سمت کلمه بستنی

شماره مرحله	کلاس شناسایی شده مطابق جدول ۴-۲	دقت شناسایی	تصویر حاصل
اول	کلاس ۷۱ (دندانه)	٪۹۷	
دوم	کلاس ۸۸ (حرف ناقص)	٪۱۰۰	
سوم	کلاس ۷۹ (حرف ــسـ)	٪۹۸	
نتیجه	کلاس ۸۵ (زیر بدنه "یی")	٪۹۹	

پس از جداسازی حروف پیوسته به روش‌های فوق و شناسایی کلاس بدنه‌ی حروف، باید به شناسایی حروف اصلی که در آن نقاط دیگر حذف نشده‌اند، پرداخت. به عنوان مثال کلاس بدنه‌ی ۷۹ هم می‌تواند مربوط به حرف "س" و هم مربوط به حرف "ش" باشد. که در مرحله‌ی طبقه‌بندی نهایی حروف این تمایز انجام می‌شود.

۵-۲-۵- طبقه‌بندی نهایی

در مراحل گذشته ابتدا با انجام عملیات پیش‌پردازش، به جداسازی زیرکلمه و بر طرف نمودن مشکل همپوشانی پرداخته شد و حروف گسسته در ۳۹ کلاس طبقه‌بندی شدند که در آن زیرکلمه‌ی متناظر با کلاس ۴۰ به بخش جداسازی حروف پیوسته فرستاده شد.

جدول ۵-۹ : کلاس‌های بدنه حروف گسسته

شماره کلاس	حرف	شماره کلاس	حرف	شماره کلاس	حرف
۱	۱	۱۶	ر ز ژ	۳۱	لله
۲	۲	۱۷	س ش	۳۲	لا
۳	۳	۱۸	ص ض	۳۳	ء
۴	۴	۱۹	ط ظ	۳۴	.
۵	۵	۲۰	ع غ	۳۵	صی ضی
۶	۶	۲۱	ف	۳۶	شی سی
۷	۷	۲۲	ق	۳۷	؟
۸	۸	۲۳	گ	۳۸	!
۹	۹	۲۴	ک	۳۹	،
۱۰	۰	۲۵	ل	۴۰	زیرکلمه
۱۱	آ	۲۶	م		
۱۲	ا	۲۷	ن		
۱۳	ب پ ت ث	۲۸	و		
۱۴	ج ح خ چ	۲۹	ه		
۱۵	د ذ	۳۰	ی		

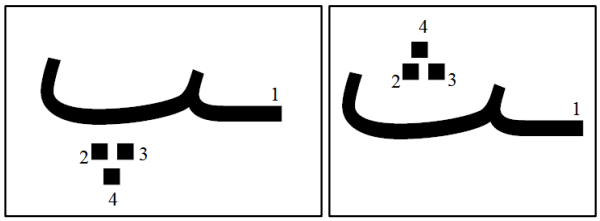
زیرکلمه‌های مرتبط با کلاس ۴۰ نیز با استفاده از روش پروفایل بالایی تعمیم یافته به حروف پیوسته تجزیه و از کلاس ۴۱ تا ۸۷ طبقه‌بندی می‌شوند. اما این زیرحروف نماینده باید در حالتی که در آن نقطه‌ها حذف نشده‌اند نیز شناسایی شوند. به عنوان مثال زیر کلاس ۴۳ می‌تواند مرتبط به حرف ح، ج، چ و یا خ باشد که باید به روش مناسب با استفاده از برچسب‌زنی مؤلفه‌ها این طبقه‌بندی را انجام داد.

جدول ۵-۱۰ : کلاس‌های بدنه حروف پیوسته

شماره کلاس	زیرحرف نماینده	شماره کلاس	زیرحرف نماینده	شماره کلاس	زیرحرف نماینده
۴۱	ـا	۵۷	ـه	۷۳	ـص
۴۲	ـبـ	۵۸	ـهـ	۷۴	صـ
۴۳	ـح	۵۹	ـهـ	۷۵	ـصـ
۴۴	ـعـ	۶۰	ـک	۷۶	ـصی
۴۵	ـع	۶۱	ک	۷۷	ـسی
۴۶	ع	۶۲	ـکـ	۷۸	سـ
۴۷	ـد	۶۳	ـل	۷۹	ـسـ
۴۸	ـف	۶۴	لـ	۸۰	ـط
۴۹	ـو	۶۵	ـلـ	۸۱	طـ
۵۰	ـه	۶۶	ـم	۸۲	ـطـ
۵۱	ـگ	۶۷	مـ	۸۳	ـو
۵۲	گ	۶۸	ـمـ	۸۴	ـی
۵۳	ـگـ	۶۹	ـن	۸۵	ـی
۵۴	ـق	۷۰	نـ	۸۶	ـس
۵۵	حـ	۷۱	ـنـ	۸۷	ـلا
۵۶	ـحـ	۷۲	ـر	۸۸	حروف ناقص

۵-۲-۵-۱ طبقه‌بندی نهایی با استفاده از برچسب‌زنی مؤلفه‌ها

در این قسمت با استفاده از روش برچسب‌زنی مؤلفه‌ها به تشخیص حروف اصلی از بدنه‌ی حروف پرداخته می‌شود. روند کار بدین صورت است که با استفاده از برچسب‌زنی مؤلفه‌ها، تعداد نقاط و محل آن‌ها مشخص می‌شود و متناسب با این معیار حرف اصلی مشخص شده و در کلاس مناسب طبقه‌بندی می‌شود. به عنوان مثال برای شناسایی زیرحرفی که در کلاس ۴۲ قرار دارد ابتدا مؤلفه‌های موجود در تصویر برچسب‌زنی می‌شوند.



شکل ۵-۱۹ : برچسب‌زنی مؤلفه و مشخص نمودن تعداد و محل دقیق نقاط

اگر تعداد مؤلفه‌ها برابر ۲ باشد یعنی ۱ مؤلفه برای زیرحرف نماینده بوده و دیگری نقطه‌ی مربوط به حرف است و مشخص می‌شود که حرف مورد نظر ـبـ است. اگر تعداد مؤلفه‌ها برابر ۳ باشد نشانه‌ی آن است که حرف مورد نظر ـتـ است. اما زمانی که تعداد مؤلفه‌ها برابر ۴ است دو حالت ممکن است رخ داده باشد (ـثـ، ـپـ). حال اگر میانگین عرضی مکان نقطه‌ها پایین‌تر از خط زمینه باشد حرف مورد نظر ـپـ بوده و اگر بالاتر از خط زمینه باشد حرف ـثـ خواهد بود.

شکل ۵-۲۰ : کلاس نهایی حروف

بنابراین زیرحرف کلاس ۴۲ با توجه به تعداد و مکان نقاط، به صورت کلاس ۴۲۰۰، ۴۲۰۱، ۴۲۰۲ و ۴۲۰۳ خواهد بود. برای حروف دیگر نیز کلاسه‌بندی به همین صورت انجام شده که نتیجه‌ی آن در جدول ۵-۷ آمده است.

جدول ۵-۱۱:کلاس‌های نهایی حروف گسسته

شماره کلاس نماینده حرف	زیرحرف	شماره کلاس نهایی	شماره کلاس نماینده حرف	زیرحرف	شماره کلاس نهایی
۱	۱	۱۰۰	۱۵	د	۱۵۰۰
۲	۲	۲۰۰		ذ	۱۵۰۱
۳	۳	۳۰۰	۱۶	ر	۱۶۰۰
۴	۴	۴۰۰		ز	۱۶۰۱
۵	۵	۵۰۰		ژ	۱۶۰۲
۶	۶	۶۰۰	۱۷	س	۱۷۰۰
۷	۷	۷۰۰		ش	۱۷۰۱
۸	۸	۸۰۰	۱۸	ص	۱۸۰۰
۹	۹	۹۰۰		ض	۱۸۰۱
۱۰	٥	۱۰۰۰	۱۹	ط	۱۹۰۰
۱۱	آ	۱۱۰۰		ظ	۱۹۰۱
۱۲	ا	۱۲۰۰	۲۰	ع	۲۰۰۰
۱۳	ب	۱۳۰۰		غ	۲۰۰۱

	ت	۱۳۰۱	۲۱	ف	۲۱۰۰
	پ	۱۳۰۲	۲۲	ق	۲۲۰۰
	ث	۱۳۰۳	۲۳	ک	۲۳۰۰
۱۴	ح	۱۴۰۰	۲۴	گ	۲۴۰۰
	خ	۱۴۰۱	۲۵	ل	۲۵۰۰
	ج	۱۴۰۲	۲۶	م	۲۶۰۰
	چ	۱۴۰۳	۲۷	ن	۲۷۰۰

شماره کلاس نماینده حرف	زیرحرف	شماره کلاس نهایی	شماره کلاس نماینده حرف	زیرحرف	شماره کلاس نهایی
۲۸	و	۲۸۰۰	۳۵	ص	۳۵۰۰
۲۹	ه	۲۹۰۰		ض	۳۵۰۱
۳۰	ی	۳۰۰۰	۳۶	س	۳۶۰۰
۳۱	لله	۳۱۰۰		ش	۳۶۰۱
۳۲	لا	۳۲۰۰	۳۷	؟	۳۷۰۰
۳۳	ء	۳۳۰۰	۳۸	!	۳۸۰۰
۳۴	.	۳۴۰۰	۳۹	،	۳۹۰۰

جدول ۵-۱۲ : کلاس‌های نهایی حروف پیوسته

شماره کلاس نماینده حرف	زیرحرف	شماره کلاس نهایی	شماره کلاس نماینده حرف	زیرحرف	شماره کلاس نهایی
۴۱	ـا	۴۱۰۰	۴۸	ـف	۴۸۰۰
۴۲	ـ ب	۴۲۰۰	۴۹	فـ	۴۹۰۰
	ـ ت	۴۲۰۱		قـ	۴۹۰۱
	ـ پ	۴۲۰۲	۵۰	ـفـ	۵۰۰۰
	ـ ث	۴۲۰۳		ـقـ	۵۰۰۱
۴۳	ـح	۴۳۰۰	۵۱	ـگ	۵۱۰۰
	ـخ	۴۳۰۱	۵۲	گـ	۵۲۰۰
	ـج	۴۳۰۲	۵۳	ـگـ	۵۳۰۰
	ـچ	۴۳۰۳	۵۴	ـق	۵۴۰۰
۴۴	ـع	۴۴۰۰	۵۵	حـ	۵۵۰۰
	ـغ	۴۴۰۱		خـ	۵۵۰۱
۴۵	ـع	۴۵۰۰		جـ	۵۵۰۲
	ـغ	۴۵۰۱		چـ	۵۵۰۳

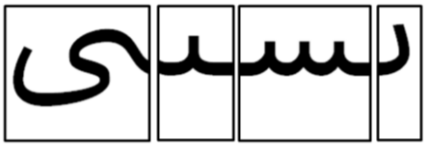
۴۶	ـعـ	۴۶۰۰	۵۶	ـحـ	۵۶۰۰
	ـغـ	۴۶۰۱		ـخـ	۵۶۰۱
۴۷	ـد	۴۷۰۰		ـجـ	۵۶۰۲
	ـذ	۴۷۰۱		ـچـ	۵۶۰۳

شماره کلاس اولیه	زیرحرف	شماره کلاس نهایی	زیرحرف	شماره کلاس	شماره کلاس نهایی
۵۷	ـه	۵۷۰۰	۷۱	ـبـ	۷۱۰۰
۵۸	ـهـ	۵۸۰۰		ـنـ	۷۱۰۱
۵۹	ـهـ	۵۹۰۲		ـیـ	۷۱۰۲
۶۰	ـکـ	۶۰۰۰		ـتـ	۷۱۰۳
۶۱	ـک	۶۱۰۰		ـپـ	۷۱۰۴
۶۲	ـکـ	۶۲۰۰		ـثـ	۷۱۰۵
۶۳	ـلـ	۶۳۰۰	۷۲	ـر	۷۲۰۰
۶۴	ـل	۶۴۰۰		ـز	۷۲۰۱
۶۵	ـلـ	۶۵۰۰		ـژ	۷۲۰۲
۶۶	ـمـ	۶۶۰۰	۷۳	ـصـ	۷۳۰۰
۶۷	ـم	۶۷۰۰		ـضـ	۷۳۰۱
۶۸	ـمـ	۶۸۰۰	۷۴	ـصـ	۷۴۰۰

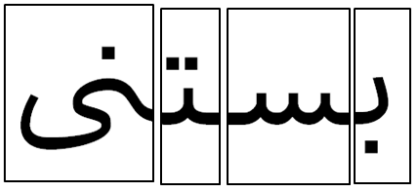
شماره کلاس اولیه	زیرحرف	شماره کلاس نهایی
۶۹	ـن	۶۹۰۰
۷۰	بـ	۷۰۰۰
	نـ	۷۰۰۱
	یـ	۷۰۰۲
	تـ	۷۰۰۳
	پـ	۷۰۰۴
	ثـ	۷۰۰۵
۷۸	سـ	۷۸۰۰
	شـ	۷۸۰۱
۷۹	ـسـ	۷۹۰۰
	ـشـ	۷۹۰۱
۸۰	ـط	۸۰۰۰
	ـظ	۸۰۰۱
۸۱	طـ	۸۱۰۰
	ظـ	۸۱۰۱
۸۲	ـطـ	۸۲۰۰
	ـظـ	۸۲۰۱
۸۳	ـو	۸۳۰۰

زیرحرف	شماره کلاس	شماره کلاس نهایی
	ضـ	۷۴۰۱
۷۵	ـصـ	۷۵۰۰
	ـضـ	۷۵۰۱
۷۶	ـصی	۷۶۰۰
	ـضی	۷۶۰۱
۷۷	ـسی	۷۷۰۰
	ـشی	۷۷۰۱
۸۴	ـی	۸۴۰۰
۸۵	ـبی	۸۵۰۰
	ـنی	۸۵۰۱
	ـیی	۸۵۰۲
	ـتی	۸۵۰۳
	ـثی	۸۵۰۴
	ـچی	۸۵۰۵
۸۶	ـس	۸۶۰۰
	ـش	۸۶۰۱
۸۷	ـلا	۸۷۰۰
۸۸	ناقص	۸۸۰۰

مطابق با جدول بالا شناسایی نهایی کلمه بستنی مطابق شکل ۵-۲۱ خواهد بود.



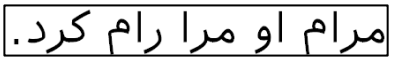
شماره کلاس زیر بدنه: ۷۰ ۷۹ ۷۱ ۸۵



شماره کلاس حروف با نقطه: ۷۰۰۰ ۷۹۰۰ ۷۱۰۳ ۸۵۰۱

شکل ۵-۲۱ : شماره کلاس نهایی حروف با وجود نقاط

مراحل کامل کار برای جمله‌ی "مرام او مرا رام کرد." به صورت زیر خواهد بود.



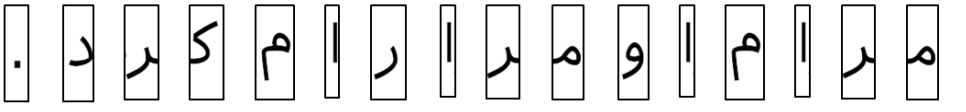
شکل ۵-۲۲ : انجام عملیات پیش پردازش

شکل ۵-۲۳ : جداسازی اولیه زیرکلمات با وجود مشکل همپوشانی

شکل ۵-۲۴ : برطرف نمودن مشکل همپوشانی

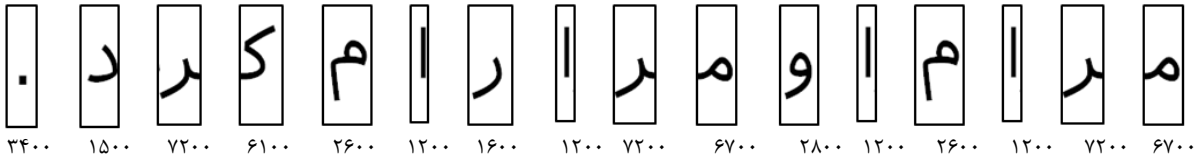
شماره کلاس: ۴۰ ۱۲ ۲۶ ۱۲ ۲۸ ۱۲ ۴۰ ۱۲ ۱۶ ۱۲ ۲۶ ۱۲ ۴۰ ۱۵ ۳۴

شکل ۵-۲۵ : شناسایی اولیه و ارسال زیرکلمات کلاس ۴۰ برای جداسازی



۶۷ ۷۲ ۱۲ ۲۶ ۱۲ ۲۸ ۶۷ ۷۲ ۱۲ ۱۶ ۱۲ ۲۶ ۶۱ ۷۲ ۱۵ ۳۴

شکل ۵-۲۶ : جداسازی زیرکلمات و کلاسه بندی بدنه حروف



شکل ۵-۲۷: کلاسه بندی نهایی حروف

مرام او مرا رام کرد.

شکل ۵-۲۸ : سند تولید شده نهایی با استفاده از کلاسه‌بندی نهایی

بنابراین با توجه به شماره کلاس نهایی تولید شده برای هر حرف جمله مذکور، مطابق جداول طبقه‌بندی نهایی، سند دیجیتالی به صورت متن تکست[1] ذخیره خواهد شد.

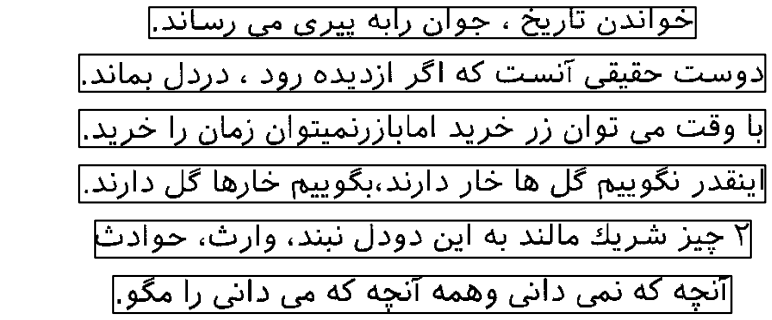
۵-۳-ارزیابی الگوریتم برای یک تصویر کامل

این ارزیابی بر روی مجموعه‌ای از تصاویر اسکن شده انجام شده است که در آن تصاویر به گونه‌ای در نظر گرفته می‌شود که هم دارای چرخش اولیه بوده و هم دارای اندکی نویز است. نمونه‌ای از این تصاویر مطابق شکل ۵-۲۹ است.

خواندن تاریخ ، جوان رابه پیری می رساند.
دوست حقیقی آنست که اگر ازدیده رود ، دردل بماند.
با وقت می توان زر خرید امابازرنمیتوان زمان را خرید.
اینقدر نگوییم گل ها خار دارند،بگوییم خارها گل دارند.
۲ چیز شریك مالند به این دودل نبند، وارث، حوادث
آنچه که نمی دانی وهمه آنچه که می دانی را مگو.

شکل ۵-۲۹ : یک تصویر نمونه برای ارزیابی الگوریتم

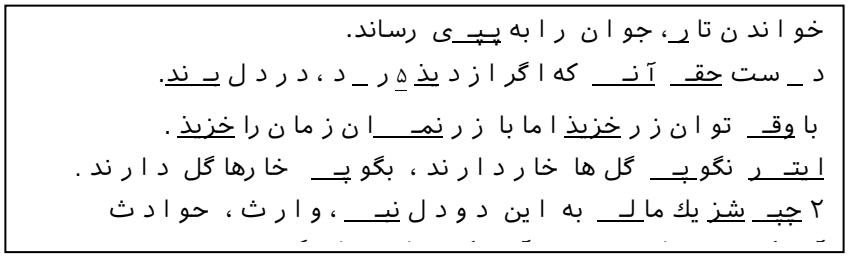
[1] Text



شکل ۵-۳۰: شناسایی جملات و جداسازی آن‌ها

در این قسمت ابتدا با تکنیک‌های ذکر شده در فصول قبل، چرخش‌های لازم انجام شده و نویزهای موجود حذف می‌شود. سپس خطوط جملات موجود در متن تشخیص داده شده و تفکیک می‌شود. پس از جداسازی، جملات به بخش استخراج زیرکلمات فرستاده شده تا برای هر جمله، زیرکلمات آن استخراج شود. پس از شناسایی، زیرکلمات به بخش جداسازی زیرحروف فرستاده شده و با استفاده از روش پروفایل بالایی کاندیدهای اولیه‌ی جداسازی انتخاب شده و با استفاده از طبقه‌بند حروف پیوسته نقاط جداسازی بدنه اصلی زیرکلمه مشخص می‌شود.

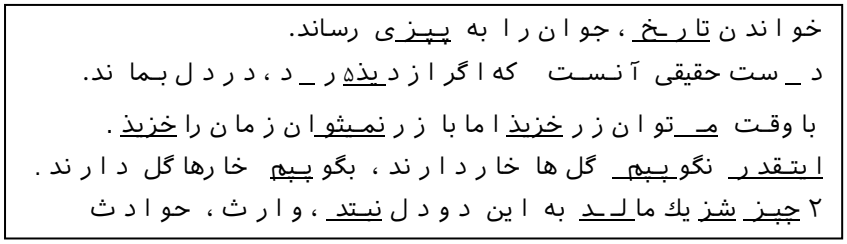
در مرحله‌ی اول، جداسازی زیرکلمه فقط از یک سمت انجام می‌گرفت. مشکل این روش آن است که اگر روند جداسازی از یک سمت انجام شود در صورت عدم شناسایی حروف اولیه یک زیرکلمه، شناسایی حروف بعدی آن‌ها نیز انجام نمی‌شود. نتیجه‌ی جداسازی برای یک جمله درشکل ۵-۳۱ آمده است.



شکل ۵-۳۱ : نتیجه شناسایی برای حالتی که جداسازی از یک سمت انجام می‌شود

در این روند جداسازی و شناسایی با خطای بالایی همراه است. دو مشکل اساسی در این روند وجود دارد. اول آنکه جداسازی و شناسایی فقط از یک سمت انجام شده و در صورت بروز مشکلی در

شناسایی حروف ابتدایی، عمل جداسازی زیرکلمه متوقف می‌شود. بنابراین برای غلبه بر این مشکل باید جداسازی از هر دو سمت انجام شود. نتیجه‌ی بهبود این جداسازی در شکل ۵-۳۲ آمده است.

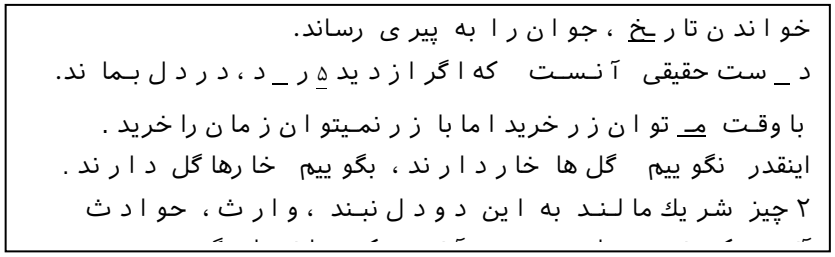


شکل ۵-۳۲ : نتیجه شناسایی برای حالتی که جداسازی از هردو سمت انجام می‌شود

مشکل دیگری نیز وجود دارد که در شناسایی نهایی موجب غلط‌های املایی می‌شود. این مسئله از آنجا ناشی می‌شود که ممکن است در زمان جداسازی زیرکلمات، نقطه‌های تصویر نیز تقسیم شده و در دو تصویر مجاور محاسبه می‌شوند. برای غلبه بر این مشکل با استفاده از برچسب‌زنی مؤلفه‌ها و در نظر گرفتن آستانه‌ی مناسب، نقطه‌های اضافی تولید شده حذف می‌شود. روند تصحیح و حذف نقاط در جدول ۵-۹ آمده است. با اعمال روش مذکور به جملات فوق، غلط‌های املایی بوجود آمده تا حد قابل قبولی برطرف می‌شود. نتیجه‌ی نهایی در شکل ۵-۳۳ آمده است.

جدول ۵-۱۳ : حذف نقاط اضافی در زمان جداسازی

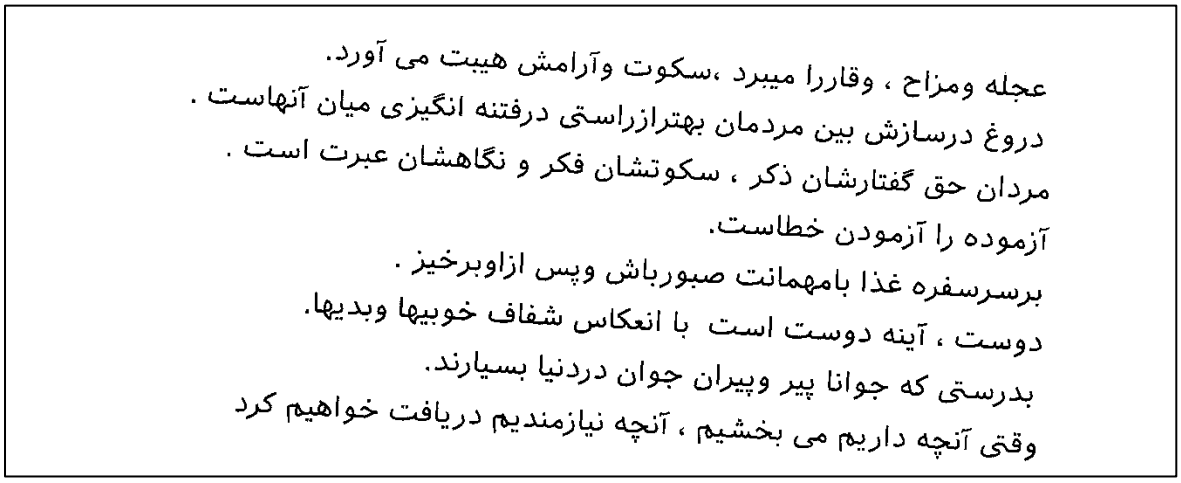
ایجاد نقطه اضافی در زمان جداسازی	شناسایی نادرست	حذف نقاط اضافی	شناسایی درست
پیپری	پپری	پیپری	پیری
دیدده	دیده	دیده	دیده
نمیتتوان	نمیتوان	نمیتوان	نمیتوان
خریید	خزید	خرید	خرید



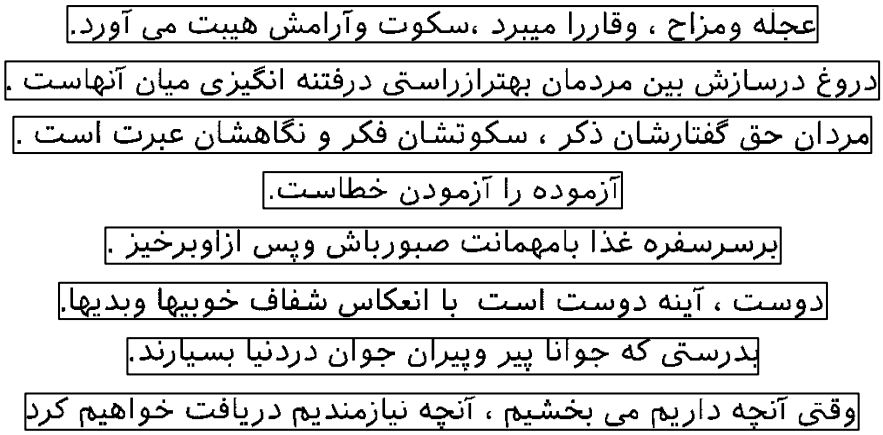
شکل ۵-۳۳ : نتیجه نهایی شناسایی پس از حذف نقاط اضافی تولید شده

۵-۴-ارزیابی الگوریتم برای تصویرهای دیگر

در ادامه الگوریتم پیشنهادی برای چند تصویر دیگر نیز مورد آزمایش قرار گرفت و نتایج آن ثبت گردید. این نتایج روند بهبود شناسایی تصویر را در سه مرحله نشان می‌دهد.



شکل ۵-۳۴ : تصویر دوم، دارای چرخش اولیه و نویز



شکل ۵-۳۵ : حذف نویز و چرخش و استخراج جملات تصویر دوم

عجل و مزاح ، _ قار را میبرد ، سکت و آرامش هیبـ مـ آورد .
دروغ درسازش بین مردمان بهـ از راستی در فتنه انگیـ ی میان آنهاست .
مردان حق گفتار شان ذکر ، سکوتشان فکر و نگاهشان عبرت است .
آزمودـ را آزمودن خطا ست .
بر سر سفرـ غذا با مـــ نث صبور باش _ پس از او بر خیز .
دوست ، آینه دوست است با انعکاس شفاف خوبـ _ بدیـ .
بدرسـ که جوانا پیر و پیر ان جوان در دنیا بسیـ رند .
وقتی آنچه داریم مـ بخشیـ ، آنچه نیاز منذ یـم دریافت خواهیـ کرد

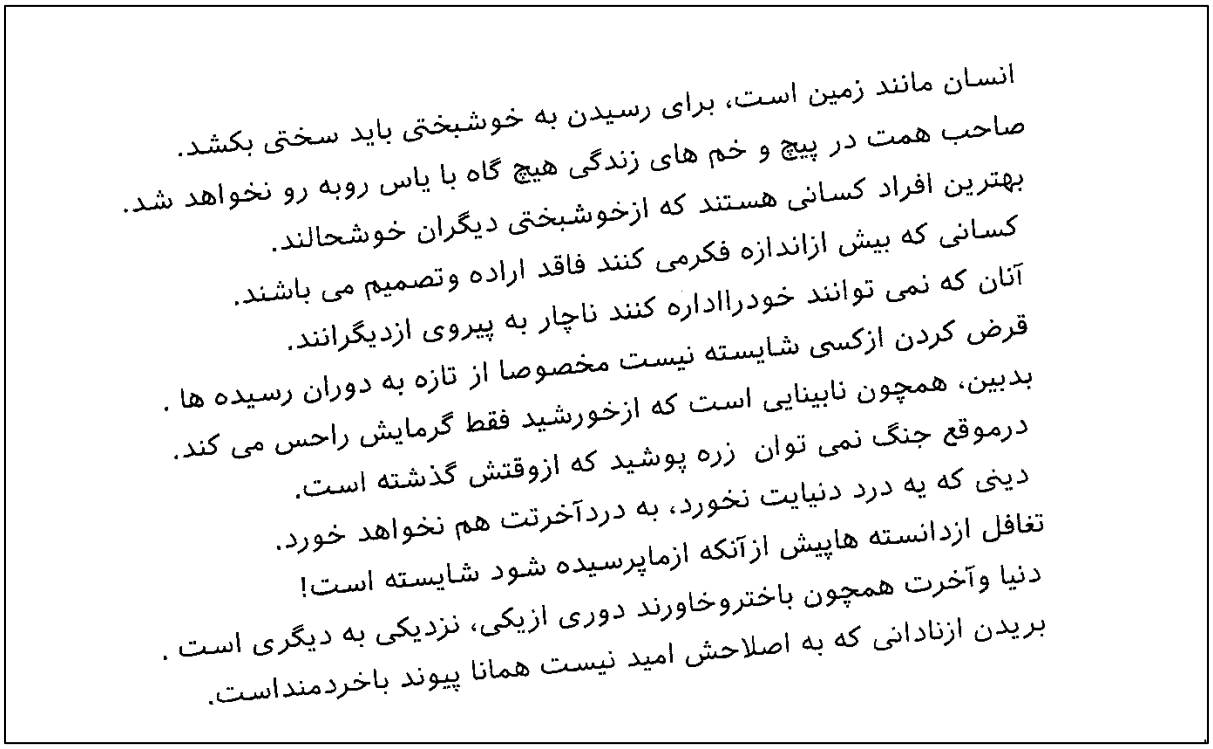
شکل ۵-۳۶ : نتیجه شناسایی تصویر دوم، جداسازی از یک سمت

عجله و مزاح ، _ قار را میبرد ، سکوت و آرامش هیبت مـ آورد .
دروغ درسازش بین مردمان بهتر از راستی در فتنه انگیز ی میان آنها ست .
مردان حق گفتار شان ذکر ، سکوتشان فکر و نگاهشان عبرت است .
آزمودـ را آزمودن خطا ست .
بر سر سفرـ غذا با مهمانث صبور باش _ پس از او بر خیز .
دوست ، آینه دوست است با انعکاس شفاف خوبیها _ بدیها .
بدرستی که جوانا پیر و پیر ان جوان در دنیا بسیار ند .
وقتی آنچه داریم مـ بخشیـ ، آنچه نیاز منذ یـم دریافت خواهیـ کرد

شکل ۵-۳۷ : نتیجه شناسایی تصویر دوم، جداسازی از هر دو سمت

عجله و مزاح ، _ قار را میبرد ، سکوت و آرامش هیبت مـ آورد .
دروغ درسازش بین مردمان بهتر از راستی در فتنه انگیز ی میان آنهاست .
مردان حق گفتار شان ذکر ، سکوتشان فکر و نگاهشان عبرت است .
آزمودـ را آزمودن خطاست .
بر سر سفرـ غذا با مهمانت صبور باش _ پس از او بر خیز .
دوست ، آینه دوست است با انعکاس شفاف خوبیها _ بدیها .
بدرستی که جوانا پیر و پیران جوان در دنیا بسیار ند .
وقتی آنچه داریم مـ بخشیـ ، آنچه نیاز مند یـم دریافت خواهیـ کرد

شکل ۵-۳۸ : نتیجه شناسایی تصویر دوم، جداسازی از هر دو سمت با تصحیح نقاط اضافی



شکل ۵-۳۹ : نتیجه شناسایی تصویر سوم، جداسازی از هر دو سمت

انسان مانند زمین است، برای رسیدن به خوشبختی باید سختی بکشد.

صاحب همت در پیچ و خم های زندگی هیچ گاه با یاس روبه رو نخواهد شد.

بهترین افراد کسانی هستند که ازخوشبختی دیگران خوشحالند.

کسانی که بیش ازاندازه فکرمی کنند فاقد اراده وتصمیم می باشند.

آنان که نمی توانند خودرااداره کنند ناچار به پیروی ازدیگرانند.

قرض کردن ازکسی شایسته نیست مخصوصا از تازه به دوران رسیده ها .

بدبین، همچون نابینایی است که ازخورشید فقط گرمایش راحس می کند.

درموقع جنگ نمی توان زره پوشید که ازوقتش گذشته است.

دینی که یه درد دنیایت نخورد، به دردآخرتت هم نخواهد خورد.

تغافل ازدانسته هاپیش ازآنکه ازماپرسیده شود شایسته است!

دنیا وآخرت همچون باختروخاورند دوری ازیکی، نزدیکی به دیگری است .

بریدن ازنادانی که به اصلاحش امید نیست همانا پیوند باخردمنداست.

شکل ۵-۴۰ : حذف نویز و چرخش و استخراج جملات تصویر سوم

ا نسا ن ما نند ز مـــ ا ست ، بر ا ی ر سیذ ن به خو شبـــ با ید سخثی بکشد .
صا حب همت د ر پیچ و خم ها ی ز ند گی هـ گا ۵ با یا س ر و به ر و نخو ا هد شد .
بهتر ین ا فر ا د کسا نی هسـ که ا ز خو شیخثی د یگر ا ن خو شحا لند .
کسا نی که بیش ا ز ا ند ا ز _ فکر مـ کنند فا قد ا ر ا د ه و تصـ مـ با شند .
آ نا ن که نمی تو ا نند خو د ر ا ا د ا ر ۵ کننـ نا چا ر به پیر و ی ا ز د یگر ا نند .
قر ض کر د ن ا ز کسی شا یسـ نیست مخضو صا ا ز تا ز _ به د و ر ا ن ر سید ه ها .
بد بین ، همچو ن نا بینا یـ ا ست که ا ز خو ر شید فقظ گر ما یـ ر ا حس مـ کند .
د ر مو قع جنگ نمـ تو ا ن ز ر ۵ پو شپد که ا ز و قتـ گذشته است .
د یثی که یه د ر د د نیا یث نخو ر د ، به د ر دآ خر تت هـ نخو ا هد خو ر د .
تغا فل ا ز د ا نستـ ها پیش ا زآ نکه ا ز ماپر سید_ شو د شا یستـ است !
د نیا _ آ خر ت همچون با خثر وخا _ ر ند د و ر ی ا ز یکـ ، نز د یکـ به د یگر ی ا ست .
بر ید ن ا ز نا د ا ني که به ا صلاحش ا مید نیست هما نا پیو ند با خر د مند ا ست .

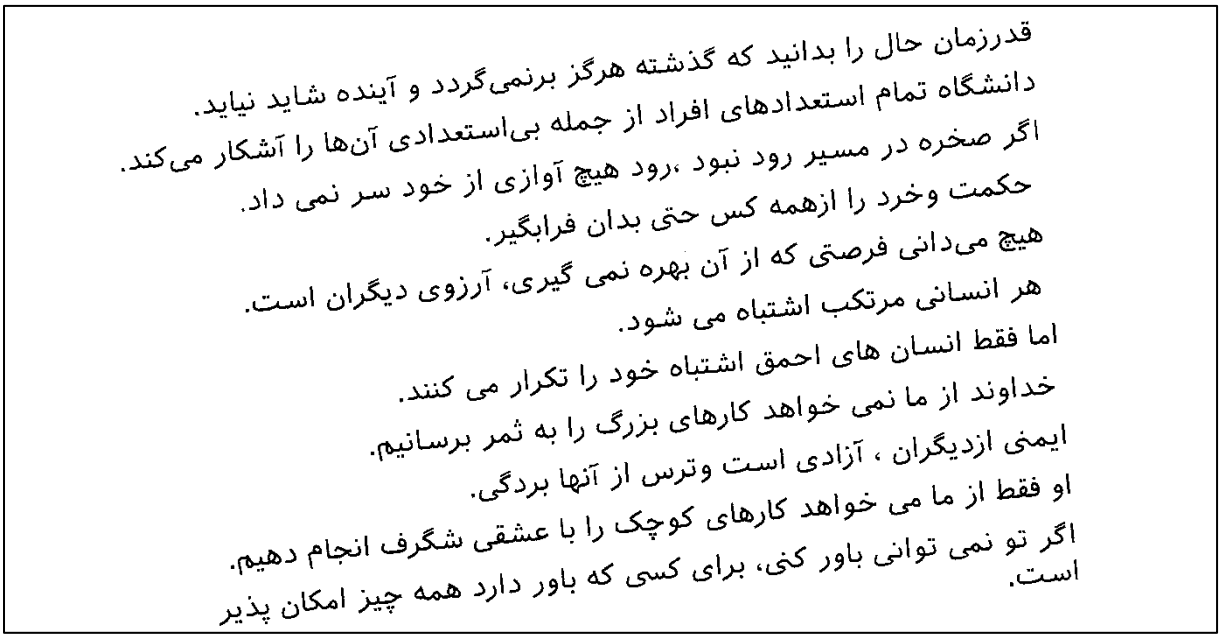
شکل ۵-۴۱ : نتیجه شناسایی تصویر سوم، جداسازی از یک سمت

ا نسا ن ما نند ز مین ا ست ، بر ا ی ر سیذ ن به خو شبخثی با ید سخثی بکشد .
صا حب همت د ر پیچ و خم ها ی ز ند گی هیچ گا ۵ با یا س ر و به ر و نخو ا هد شد .
بهتر ین ا فر ا د کسا نی هستند که ا ز خو شیخثی د یگر ا ن خو شحا لند .
کسا نی که بیش ا ز ا ند ا ز _ فکر مـ کنند فا قد ا ر ا د ه و تصمیم مـ با شند .
آ نا ن که نمی تو ا نند خو د ر ا ا د ا ر ۵ کنند نا چا ر به پیر و ی ا ز د یگر ا نند .
قر ض کر د ن ا ز کسی شا یسته نیست مخضو صا ا ز تا ز _ به د و ر ا ن ر سید ه ها .
بد بین ، همچو ن نا بینا یـ ا ست که ا ز خو ر شید فقظ گر ما یش ر ا حس مـ کند .
د ر مو قع جنگ نمـ تو ا ن ز ر ۵ پو شپد که ا ز و قتش گذشته است .
د یثی که یه د ر د د نیا یث نخو ر د ، به د ر دآ خر تت هـ نخو ا هد خو ر د .
تغا فل ا ز د ا نسته ها پیش ا زآ نکه ا ز ماپر سید_ شو د شا یسته است !
د نیا _ آ خر ت همچون با خثر وخا _ ر ند د و ر ی ا ز یکـ ، نز د یکـ به د یگر ی ا ست .
بر ید ن ا ز نا د ا ني که به ا صلاحش ا مید نیست هما نا پیو ند با خر د مند ا ست .

شکل ۵-۴۲ : نتیجه شناسایی تصویر سوم، جداسازی از هر دو سمت

ا نسا ن ما نند ز مین ا ست ، بر ا ی ر سید ن به خو شبختی با ید سختی بکشد .
صا حب همت د ر پیچ و خم ها ی ز ند گی هیچ گا ۵ با یا س ر و به ر و نخو ا هد شد .
بهتر ین ا فر ا د کسا نی هستند که ا ز خو شبختی د یگر ا ن خو شحا لند .
کسا نی که بیش ا ز ا ند ا ز _ فکر مـ کنند فا قد ا ر ا د ه و تصمیم مـ با شند .
آ نا ن که نمی تو ا نند خو د ر ا ا د ا ر ۵ کنند نا چا ر به پیر و ی ا ز د یگر ا نند .
قر ض کر د ن ا ز کسی شا یسته نیست مخصو صا ا ز تا ز _ به د و ر ا ن ر سید ه ها .
بد بین ، همچو ن نا بینا یـ ا ست که ا ز خو ر شید فقط گر ما یش ر ا حس مـ کند .
د ر مو قع جنگ نمـ تو ا ن ز ر ۵ پو شید که ا ز و قتش گذشته است .
د ینی که یه د ر د د نیا یت نخو ر د ، به د ر دآ خر تت هـ نخو ا هد خو ر د .
تغا فل ا ز د ا نسته ها پیش ا زآ نکه ا ز ماپر سید_ شو د شا یسته است !
د نیا _ آ خر ت همچون با ختر وخا _ ر ند د و ر ی ا ز یکـ ، نز د یکـ به د یگر ی ا ست .
بر ید ن ا ز نا د ا ني که به ا صلاحش ا مید نیست هما نا پیو ند با خر د مند ا ست .

شکل ۵-۴۳ : نتیجه شناسایی تصویر سوم، جداسازی از هر دو سمت با تصحیح نقاط اضافی



شکل ۵-۴۴ : نتیجه شناسایی تصویر چهارم، جداسازی از هر دو سمت

شکل ۵-۴۵ : حذف نویز و چرخش و استخراج جملات تصویر چهارم

قد ر ز م ا ن ح ا ل ر ا ب د ا تید که گذ شـــ هر گز بر نمـ گر د د _ آ یتد ۵ شاید نیا ید .
د ا نشـ _ تما م ا ستغد ا د ها ی ا فر ا د ا ز جمله یـ ا ستغد ا د ی آ ن ها ر ا آ شـــ ر مـ کند .
ا گر صخر ه د ر مسـ ر و د نبو د ، ر و د هیـ آ و ا ز ی ا ز خو د سر نمـ د ا د .
حکمـ و خر د ر ا ا ز همه کس حتی بد ا ن فر ا بگیر .
هیچ می د ا نـ فر صتـ که ا ز آ ن بهر _ نمی گیـ ی ، آ ر ز _ ی د یگر ا ن ا ست .
هر ا نسانی مر ثکب ا شتـه می شو د ،
ا ما فقظ ا نسا ن ها ی ا حمق ا شتـ ۵ خو د ر ا ثکر ا ر می کنند .
خد ا _ ند ا ز ما نمی خو ا هد کا ر ها ی بز ر گ ر ا به ثمر بر سا نیـ .
ا یمنی ا ز د یگر ا ن ، آ ز ا د ی ا سـ _ تز س ا ز آ نها بر د گی .
ا _ فقط ا ز _ می خو ا هد کا ر ها ی کو چک ر ا با عشـ شگـ ا نجا م د هیـ .
ا گر تو نمی تو ا نـ با _ ر کنی ، بر ا ی کسی که با _ ر د ا ر د همه چیز ا مکان پذ یز
ا ست.

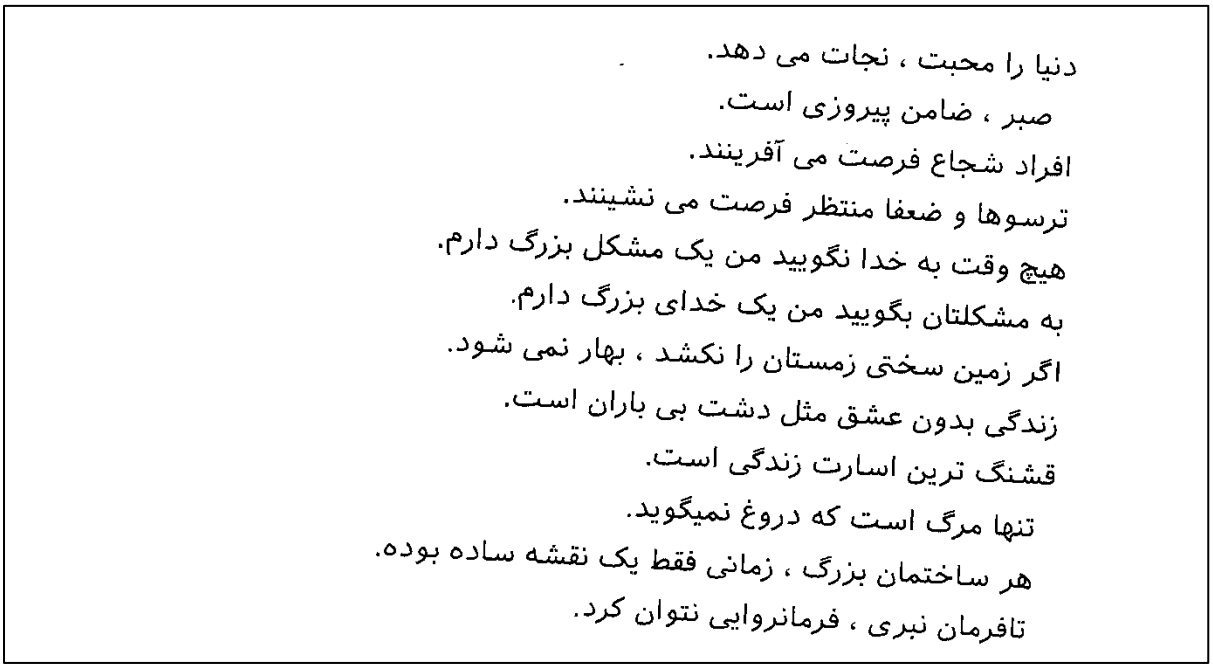
شکل ۵-۴۶ : نتیجه شناسایی تصویر چهارم، جداسازی از یک سمت

قد ر ز م ا ن ح ا ل ر ا ب د ا تید که گذ شته هر گز بر نمـ گر د د _ آ یتد ۵ شاید نیا ید .
د ا نشگا _ تما م ا ستغد ا د ها ی ا فر ا د ا ز جمله یـ ا ستغد ا د ی آ ن ها ر ا آ شکا ر مـ کند .
ا گر صخر ه د ر مسیر ر و د نبو د ، ر و د هیچ آ و ا ز ی ا ز خو د سر نمـ د ا د .
حکمت و خر د ر ا ا ز همه کس حتی بد ا ن فر ا بگیر .
هیچ می د ا نـ فر صتی که ا ز آ ن بهر _ نمی گیـ ی ، آ ر ز _ ی د یگر ا ن ا ست .
هر ا نسانی مر ثکب ا شتبا ه می شو د ،
ا ما فقظ ا نسا ن ها ی ا حمق ا شتبا ۵ خو د ر ا ثکر ا ر می کنند .
خد ا _ ند ا ز _ نمی خو ا هد کا ر ها ی بز ر گ ر ا به ثمر بر سا نیـ .
ا یمنی ا ز د یگر ا ن ، آ ز ا د ی ا ست _ تز س ا ز آ نها بر د گی .
ا _ فقط ا ز ما می خو ا هد کا ر ها ی کو چک ر ا با عشقی شگرف ا نجا م د هیـ .
ا گر تو نمی تو ا نـ با _ ر کنی ، بر ا ی کسی که با _ ر د ا ر د همه چیز ا مکان پذ یز
ا ست.

شکل ۵-۴۷ : نتیجه شناسایی تصویر چهارم، جداسازی از هر دو سمت

قد ر ز م ا ن ح ا ل ر ا ب د ا نید که گذ شته هر گز بر نمـ گر د د _ آ یند ۵ شاید نیا ید .
د ا نشگا _ تما م ا ستعد ا د ها ی ا فر ا د ا ز جمله یـ ا ستعد ا د ی آ ن ها ر ا آ شکا ر مـ کند .
ا گر صخر ه د ر مسیر ر و د نبو د ، ر و د هیچ آ و ا ز ی ا ز خو د سر نمـ د ا د .
حکمت و خر د ر ا ا ز همه کس حتی بد ا ن فر ا بگیر .
هیچ می د ا نـ فر صتی که ا ز آ ن بهر _ نمی گیـ ی ، آ ر ز _ ی د یگر ا ن ا ست .
هر ا نسانی مر تکب ا شتبا ه می شو د ،
ا ما فقط ا نسا ن ها ی ا حمق ا شتبا ۵ خو د ر ا تکر ا ر می کنند .
خد ا _ ند ا ز ما نمی خو ا هد کا ر ها ی بز ر گ ر ا به ثمر بر سا نیـ .
ا یمنی ا ز د یگر ا ن ، آ ز ا د ی ا ست _ تر س ا ز آ نها بر د گی .
ا _ فقط ا ز _ می خو ا هد کا ر ها ی کو چک ر ا با عشقی شگرف ا نجا م د هیـ .
ا گر تو نمی تو ا نـ با _ ر کنی ، بر ا ی کسی که با _ ر د ا ر د همه چیز ا مکان پذ یر
ا ست.

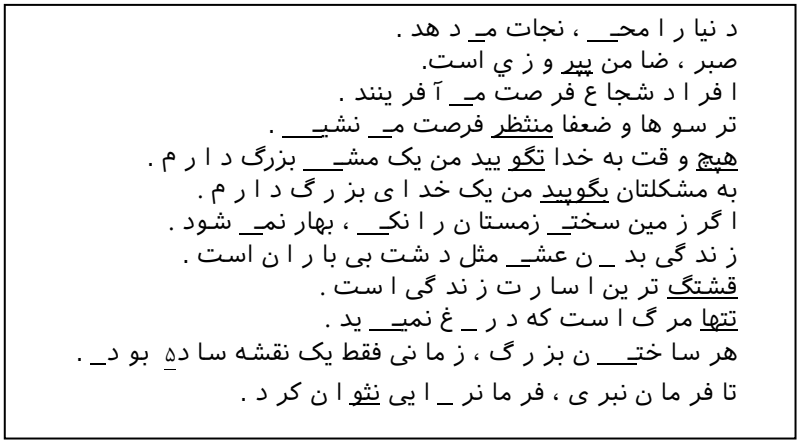
شکل ۵-۴۸ : نتیجه شناسایی تصویر چهارم، جداسازی از هر دو سمت با تصحیح نقاط اضافی


دنیا را محبت ، نجات می دهد.
صبر ، ضامن پیروزی است.
افراد شجاع فرصت می آفرینند.
ترسوها و ضعفا منتظر فرصت می نشینند.
هیچ وقت به خدا نگویید من یک مشکل بزرگ دارم.
به مشکلتان بگویید من یک خدای بزرگ دارم.
اگر زمین سختی زمستان را نکشد ، بهار نمی شود.
زندگی بدون عشق مثل دشت بی باران است.
قشنگ ترین اسارت زندگی است.
تنها مرگ است که دروغ نمیگوید.
هر ساختمان بزرگ ، زمانی فقط یک نقشه ساده بوده.
تافرمان نبری ، فرمانروایی نتوان کرد.

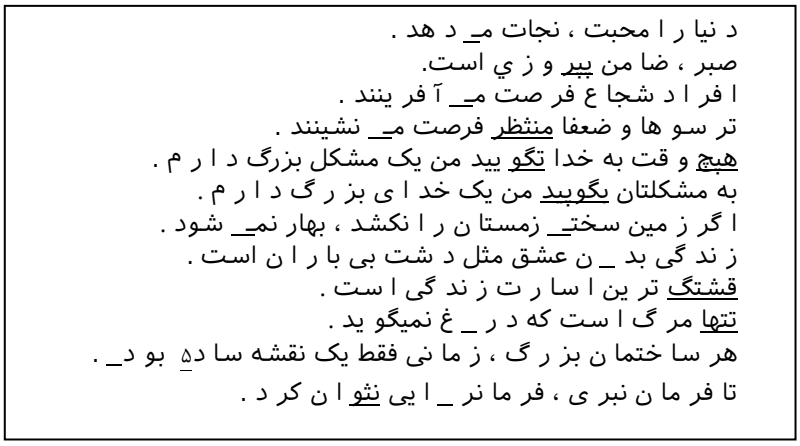
شکل ۵-۴۹ : نتیجه شناسایی تصویر پنجم، جداسازی از هر دو سمت


دنیا را محبت ، نجات می دهد.
صبر ، ضامن پیروزی است.
افراد شجاع فرصت می آفرینند.
ترسوها و ضعفا منتظر فرصت می نشینند.
هیچ وقت به خدا نگویید من یک مشکل بزرگ دارم.
به مشکلتان بگویید من یک خدای بزرگ دارم.
اگر زمین سختی زمستان را نکشد ، بهار نمی شود.
زندگی بدون عشق مثل دشت بی باران است.
قشنگ ترین اسارت زندگی است.
تنها مرگ است که دروغ نمیگوید.
هر ساختمان بزرگ ، زمانی فقط یک نقشه ساده بوده.
تافرمان نبری ، فرمانروایی نتوان کرد.

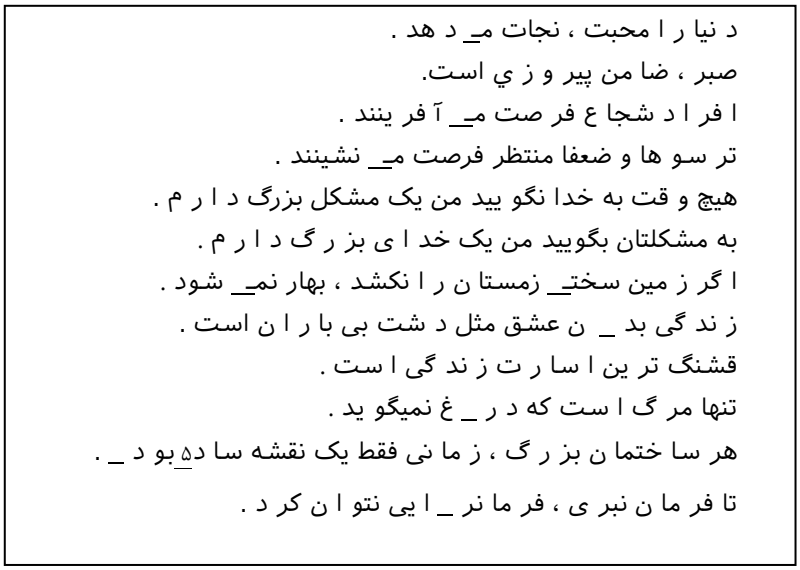
شکل ۵-۵۰ : حذف نویز و چرخش و استخراج جملات تصویر پنجم



شکل ۵-۵۱ : نتیجه شناسایی تصویر پنجم، جداسازی از یک سمت



شکل ۵-۵۲ : نتیجه شناسایی تصویر پنجم، جداسازی از هر دو سمت



شکل ۵-۵۳ : نتیجه شناسایی تصویر پنجم، جداسازی از هر دو سمت با تصحیح نقاط اضافی

۵-۴-۱- نتیجه ارزیابی الگوریتم

در قسمت قبل جهت اعتبارسنجی، روند بازشناسی بر روی ۵ صفحه‌ی متن فارسی با درجه تفکیک ۳۰۰ نقطه بر اینچ آزمایش شد و نتیجه بازشناسی در سه مرحله بهبود یافت.

در مرحله اول شناسایی از یک سمت انجام می‌شود. در این مرحله اگر در زیرکلمه‌ای جداسازی برای حروف اولیه زیرکلمه به درستی انجام نشود عمل جداسازی برای بقیه حروف آن زیر کلمه ادامه نمی‌یابد.

در مرحله‌ی دوم جهت بر طرف نمودن مشکل مرحله‌ی قبل و شناسایی و جداسازی هرچه بهتر، جداسازی از دو سمت انجام می‌شود. بدین صورت که در صورت بروز مشکل در روند شناسایی در یک زیرکلمه عملیات شناسایی و جداسازی از سمت دیگر انجام می‌شود که این خود باعث بهبود قابل ملاحظه‌ی درصد جداسازی و شناسایی می‌شود. در مرحله‌ی سوم با تصحیح نقاط اضافی تولید شده غلط‌های املایی بوجود آمده تصحیح می‌شوند.

همانطور که گفته شد عمل بازشناسی متن در چند فرایند انجام می‌شود که در هر مرحله ممکن است خطایی رخ داده و درصد شناسایی نهایی کاهش یابد. در مرحله‌ی اول پس از عملیات پیش پردازش، فرایند جداسازی زیرکلمات و حروف انجام شده که جهت رفع مشکل همپوشانی ممکن است خطایی رخ دهد. پس از برطرف نمودن مشکل همپوشانی، حروف و زیرکلمات به طبقه‌بند حروف گسسته فرستاده می‌شود. که در این مرحله نیز ممکن است خطایی رخ دهد. مثلاً برخی زیرکلمات بصورت حرف تشخیص داده شده و برای جداسازی به بخش جداسازی فرستاده نمی‌شود و یا برخی حروف به درستی شناسایی نمی‌شوند مانند شناسایی حرف" ۵ "بجای حرف "ه" .

زیرکلمات شناسایی شده به بخش جداسازی فرستاده می‌شود. در این بخش ابتدا نقاط اولیه‌ی کاندید برای جداسازی انتخاب شده و سپس توسط طبقه بند حروف پیوسته نقاط جداسازی نهایی مشخص می‌شود که در هر دو بخش (بخش تعیین نقاط جداسازی اولیه و بخش جداسازی با طبقه‌بند حروف پیوسته) ممکن است خطایی رخ دهد.

با توجه به این که شناسایی هم شامل شناسایی زیرکلمات و حروف گسسته بوده و هم شناسایی برای حروف پیوسته در زمان جداسازی، لذا خطای رخ داده در مرحله شناسایی، مجموع خطای رخ داده در مرحله جداسازی بعلاوه خطای شناسایی حروف گسسته می‌باشد. که در نهایت و در مرحله سوم دقت در سطح جداسازی، ۹۶٪ و در سطح شناسایی نهایی ۸۵٪ حاصل شد.

۵-۵-نتیجه‌گیری

در این فصل به توضیح روند کلی سیستم بازشناسی متن پرداخته شد. در مرحله‌ی اول پیش‌پردازش‌هایی بر روی تصویر ورودی اعمال شد و در مرحله‌ی جداسازی با شناسایی حدود مکانی جمله و با استفاده از روش‌های مناسب و رفع مشکلات همپوشانی به جداسازی حروف گسسته و زیرکلمات پرداخته شد. با استفاده از طبقه‌بند حروف گسسته، حروف مجزا از زیرکلمات جدا شده و با توجه به شماره کلاس آن ذخیره گردید و زیرکلمات شناسایی شده در این مرحله به قسمت جداسازی حروف رفته و با روش‌های مختلف این جداسازی انجام شد. سپس با استفاده از طبقه‌بند حروف پیوسته به شناسایی بدنه حروف پرداخته شد و سپس بدنه‌ی این حروف به همراه نقطه‌های آن تحلیل و در کلاس‌های نهایی طبقه‌بندی گردید. در نهایت این حروف در کنار یکدیگر قرار گرفت و جمله بازیابی شد. جهت اعتبارسنجی، روند مذکور بر روی ۵ صفحه‌ی متن فارسی با درجه تفکیک ۳۰۰ نقطه بر اینچ آزمایش شده است که در آن دقت در سطح جداسازی، ۹۶٪ و در سطح شناسایی نهایی ۸۵٪ می‌باشد. این درصد در مقایسه با [۱۵] که برای قلم‌های میترا، نازنین و لوتوس انجام شده دارای خطای بیشتری بوده که این خطا ناشی از پیچیدگی‌های ساختاری و نوشتاری قلم Iranian sans می‌باشد.

فصل ششم

نتیجه‌گیری و پیشنهادات

۶-۱- نتیجه گیری

امروزه با افزایش تقاضا در همه‌ی زبان‌ها روز به روز نیاز به یک سامانه‌ی بازشناسی متن قوی افزایش می‌یابد. در زبان فارسی با توجه به تنوع بسیار زیاد قلم‌ها برای متون تایپی و همچنین وجود دست‌نوشته‌هایی که افراد با توجه به سلیقه‌ی شخصی و به اشکال متنوع می‌نویسند، برای دست‌یابی به یک سیستم بازشناسی قوی راه طولانی در پیش است. در این پایان‌نامه در فصل اول تعاریف اولیه‌ای از سیستم بازشناسی متن و کاربردهای آن مطرح شد، سپس بعضی از ویژگی‌های زبان فارسی که عملیات بازشناسی را با مشکل روبه رو می‌کند و بعضی از علت‌های عدم تکامل سیستم بازشناسی زبان فارسی مطرح شد. در فصل دوم به معرفی تحقیقاتی که در زمینه‌ی پیدایش اُ.سی.آر انجام شده، پرداخته شد و همچنین روند تحقیقات اُ.سی.آر بیان شد. سپس به مرور بعضی از مقالات در زمینه‌ی اُ.سی.آر فارسی و همچنین روند تحقیقات فارسی پرداخته شد. در فصل سوم در مورد مراحل یک سیستم بازشناسی متن و روش‌های هر یک از آن‌ها به طور کلی توضیحاتی داده شد. در فصل چهارم به منظور فقدان پایگاه داده برای قلم Iranian sans, یک پایگاه داده‌ی نسبتاً کامل از حروف گسسته و پیوسته‌ی این قلم خاص به همراه اعداد و بعضی از علائم ایجاد شد تا به عنوان نمونه‌ی آموزش برای طبقه‌بندها به کار گرفته شوند. در فصل پنجم مراحل مختلف سیستم بازشناسی متن به کار رفته در این پایان‌نامه به طور کامل توضیح داده شد. در این فصل ابتدا با اعمال عملیات پیش‌پردازش، سند ورودی شبکه آماده می‌شود، سپس با قطعه‌بندی جمله‌ها، کلمات و زیرکلمات عملیات جداسازی آغاز می‌شود. جداسازی با استفاده از روش پروفایل بالایی تعمیم یافته و سپس جاروب از طرفین زیرکلمه انجام می‌شود. سپس با افزودن نقاط به بدنه‌ی حروف و اختصاص شماره کلاس‌های مربوطه طبقه‌بندی نهایی انجام شده و در نهایت با توجه به شماره‌ی کلاس‌ها، زیرکلمه‌ها، کلمه‌ها و در نهایت جمله بازیابی می‌شود. در نتیجه یک سیستم بازشناسی متن برای قلم Iranian sans ارائه شد.

۶-۲- نوآوری‌ها

- ✓ ارائه‌ی ساختاری مناسب جهت حذف نویزهای موجود در تصویر
- ✓ تولید پایگاه داده برای حروف گسسته و پیوسته قلم Iranian sans
- ✓ پیاده‌سازی و آموزش طبقه‌بند حروف پیوسته و گسسته برای قلم Iranian sans
- ✓ غلبه بر مشکلات پیچیدگی‌های ساختاری قلم مذکور با افزودن زیرکلماتی مانند "سی", "صی"، "لا"، "ـیی" "لله" به کلاس‌های حروف جدا و حروف پیوسته
- ✓ پیاده‌سازی روشی مناسب برای جداسازی مناسب زیرکلمات و غلبه بر مشکلات همپوشانی
- ✓ پیاده‌سازی الگوریتم مناسب جهت جداسازی بهینه‌ی زیرکلمات به حروف سازنده‌ی آن با استفاده از روش پروفایل بالایی تعمیم یافته و جاروب از طرفین زیرکلمه
- ✓ ایجاد کلاس نامعتبر برای طبقه‌بند حروف پیوسته جهت شناسایی بهتر
- ✓ ارائه‌ی الگوریتم مناسب برای حذف نقاط اضافی ایجاد شده در هنگام جداسازی و تصحیح نتایج شناسایی

۶-۳- پیشنهادات

برای ادامه‌ی کار در زمینه‌ی سیستم‌های بازشناسی متن می‌توان موارد زیر را پیشنهاد داد.

- تلاش برای تقویت هر یک از مراحل بازشناسی متن، مثلاً در مرحله پیش‌پردازش برای باینری کردن، حذف نویز، اصلاح چرخش و تقویت سایر مراحل دیگر
- استفاده از چند طبقه بند بصورت تقویت شده با الگوریتم آدابوست برای بالا بردن دقت شناسایی

- تلاش برای تولید سیستم‌های بازشناسی متن که بتواند متونی نوشته شده با بیشترین تعداد قلم و انواع زبان‌ها را شناسایی کند، که این امر نیاز به آشنایی با نوشتار زبان‌های مختلف و قلم‌های آن‌ها دارد.
- تلاش برای ساخت سامانه‌ای که نسبت به درجه تفکیک حساس نباشد و بتواند هر تصویر با هر درجه تفکیک را شناسایی کند. که طراحی چنین سامانه‌ای کار بسیار سختی است.
- طراحی سامانه‌ای که قابلیت شناسایی تصاویر خاکستری و رنگی را (بدون دوسطحی کردن) داشته باشد.
- برخی حروف یا ارقام در یک قلم چاپی شبیه به حروف یا ارقامی از قلم‌های دیگر هستند و نرم افزارهای اُ.سی.آر را به خصوص در متونی که با چند نوع قلم نوشته شدند به اشتباه می‌اندازند که یک راه حل می‌تواند بازشناسی نوع قلم پیش از بازشناسی متن باشد.
- طراحی روشی برای جداسازی زیرکلماتی از جمله "لم", "می", "سی", "صی" برای قلم Iranian sans که به راحتی و با روش ذکر شده قابل جداسازی نیستند.

فهرست مراجع

[1] D. Doermann, “The indexing and retrieval of document images: a survey,” *Computer Vision and Image Understanding*, Vol. 70, No. 3, pp. 287-298, 1998.

[2] H. Khosravi, E. Kabir, “Introducing a very large dataset of handwritten Farsi digits and a study on their varieties,” *Pattern Recognition Letters*,Vol. 28,No. 10, pp. 1133-1141,Jul 2007.

[3] N. Arica , F. T. Yarman-Vural, “An overview ofcharacter recognition focused on off-line handwriting,” *IEEE Trans. Systems, Man, and Cybernetics*, part C, Vol.31, No. 2, pp. 216-233, 2001.

[4] S. Mori, C. Y. Suen, and K. Yamamoto, “ Historical Review of OCR Research and Development,” *Proc. of IEEE*, Vol.80, No. 7 ,pp.1029-1058, July 1992.

[5] http: //www.uspto.gov

[6] M. Bokser, “Omnidocument Technologies,” *Proceedings of the IEEE*, Vol. 80, No. 7, pp. 1066-1078, 1992.

[7] http:\\www.ncr.co.jp\library.html

[8] B. Parhami, and M. Taraghi, “ Automatic recognition of printed Farsi texts,” *Pattern Recognition letters*, Vol. 14,No. 1-6, pp. 395-403, 1981.

[۹] م.شیرعلی شهرضا و ک .فائز، "تشخیص کلمات و ارقام دست‌نویس فارسی به وسیله شبکه های عصبی (خط نسخ) "، رساله دکترای مهندسی برق - کامپیوتر، دانشگاه صنعتی امیر کبیر،۱۳۷۴.

[10] R. Azmi, E. Kabir, “A new segmentation technique for ominifont Farsi text,” *Pattern Recognition Letters*, Vol. 22 ,No.2 , pp. 97-104,Feb. 2001.

[۱۱] ر.عزمی، "بازشناسی متون چاپی فارسی، "رساله دکتری، بخش مهندسی برق، دانشگاه تربیت مدرس، ۱۳۷۸.

[12] M. B. Menhaj, and M. Adab, “Simultaneous segmentation and recognition of Farsi/Latin printed texts with MLP,” *International Joint Conference on Neural Networks*, Vol. 2, pp. 1534-1539, May 2002.

[13] R. Mehran, H. Pirsiavash, F. Razzazi,“A Front-end OCR for Omni-font Persian/Arabic Cursive Printed Documents,” *Digital Image Computing: Techniques and Applications* ,pp. 385-392, Dec 2005.

[14] A. Broumandnia, J. Shanbehzadeh, et al, “Segmentation of Printed Farsi/Arabic Words,” *IEEE/ACS International Conference on Computer Systems and Applications*, pp.761-766,2007.

[۱۵] ح. خسروی، و ا. کبیر، " بازشناسی متن چاپی فارسی بر مبنای جداسازی هوشمند، " سومین کنفرانس بین‌المللی فناوری اطلاعات و دانش، مشهد، دانشگاه فردوسی مشهد،۱۳۸۶.

[۱۶] ح. بویری، م. عباسی دزفولی، و م.یکتایی، "بازشناسی نوری کاراکترهای چاپی فارسی با استفاده از شبکه عصبی فازی، " اولین کنفرانس دانشجویی فناوری اطلاعات ایران، سنندج، دانشگاه کردستان،۱۳۸۹.

[۱۷] ا. کیومرثی، ح. نظام آبادی پور، و ی. نوروززاده، "یک روش جدید برای جداسازی خطوط متن دست نوشته"، هفتمین کنفرانس ماشین بینایی و پردازش تصویر، تهران، دانشگاه علم و صنعت،۱۳۹۰.

[۱۸] م. زند، " تشخیص حروف چاپی فارسی با استفاده از روش ترکیبی، " همایش منطقه ای علوم کامپیوتر، مهندسی کامپیوتر و فناوری اطلاعات، دورود، دانشگاه آزاد اسلامی واحد دورود،۱۳۹۱.

[19] M. Dehghan, K. Faez, M. Ahmadi, M. Shridhar, “Unconstrained Farsi handwritten word recognition using fuzzy vector quantization and hidden markov models,” *Pattern Recognition Letters*, Vol. 22 , No. 2, pp. 209-214, Feb 2001.

[20] M. Dehghan , K. Faez, M. Ahmadi, M. Shridhar, “Handwritten Farsi (Arabic) word recognition: A holistic approach using discrete HMM,” *Pattern Recognition* ,Vol. 34, No. 5 , pp. 1057-1065, May 2001.

[21] S.Shirali-Shahreza , M.T.Manzuri-Shalmani, M.H. Shirali-Shahreza, “Preparing Persian/Arabic Scanned Images for OCR,”*Information and Communication Technologies,*Vol.1,pp.1332-1336, 2006.

[22] R. Halavati, S.B. Shouraki, "Recognition of Persian online handwriting using elastic fuzzy pattern recognition," *International Journal of Pattern Recognition and Artificial Intelligence* ,Vol. 21 ,No. 3 , pp. 491-513,2007.

[23] H. Khosravi, E. Kabir, "Farsi font recognition based on Sobel–Roberts features," *Pattern Recognition Letters*,Vol. 31 ,No.1 , pp. 75-82 ,January 2010.

[32] M.H. Shirali-Shahreza, S. Shirali-Shahreza, "Removing Noises Similar to Dots from Persian Scanned Documents," *International Colloquium on IEEE*, Vol. 2, pp.313-317,2008.

[25] S. Khalighi,P. Tirdad, H.R. Rabiee, M. Parviz, "A Novel OCR System for Calculating Handwritten Persian Arithmetic Expressions," *International Conference on IEEE*, pp.755-758,2009.

[26] E. Arianyan, S.A. Motamedi, I. Arianyan, " Efficient Optical Character Recognition on Graphics Processing Unit, " *Sixth International Symposium on IEEE*, pp.789-793, 2012.

[27] A. and J. Kanai, "Projection profile based skew estimation algorithm for JBIG," *compressed images, ICDAR*,pp. 401- 405,1997.

[28] L. Najman, "Using mathematical morphology for document skew estimation," *SPIE Document Recognition and Retrieval XI*, pp. 182-191, 2004 .

[29] R. v. d. Boomgaard, and R. v. Balen, "Methods for fast morphological image transform using bitmapped binary images", *Graphical Models and Image Processing*, pp.252-258 , 1992.

[۳۰] ح.خسروی، و ا. کبیر، "یافتن زاویه چرخش سند مبتنی بر زوایای خطوط متن با استفاده از عملگر مورفولوژی"، *سومین کنفرانس بین‌المللی فناوری اطلاعات و دانش*، مشهد، دانشگاه فردوسی مشهد،۱۳۸۶.

[31] B. Al-Badr, S. A. Mahmoud, "Survey and bibliography of Arabic optical text recognition," *Signal Processing*, Vol. 41, pp. 49-77,1995.

[32] H. Nashida, and S. Mori, “An Algebraic Approach to Automatic Construction of Structured Models,” *Pattern Analysis and Machine Intelligence* ,Vol. 15, No.12 , pp. 1298-1311,1993.

[33] H. Y. Abdelazim, and M. A. Hashish, “Arabic Reading Machine”, *Proc. of the 10th National Computer Conference*, Jeddah, pp. 733-744 ,1988.

[34] Houle, G. and M. Shridhar, “Handwritten Word Recognition With OCR-based Segmenter,” *Proc.of the Workshop on Document Image Analysis* ,pp. 51-58 ,1997.

[۳۵] ح .نظام آبادی پور، ا .کبیر و ر.عزمی ، "الگوریتم اصلاح شده جداسازی حروف در متون چاپی با برچسب زدن به کانتور بالایی کلمات" استقلال ۱۳۸۰.

[36] F. Kimura, M. Shridhar, et al., “Improvements of a Lexicon Directed Algorithm for Recognition of Unconstrained Handwritten Words,” *Proc. of 2nd ICDAR Conf*, pp.18-22, 1993.

[37] A. Cheung.,M. Bennamoun, et al., “An Arabic optical character recognition system using recognition-based segmentation,” *Pattern Recognition* ,Vol. 34, pp. 215-233,2001.

[38] B.B. Chaudhuri, and U. Garain, “Automatic Detection of Italic, Bold and All-Capital words inDocument Images”. *14th International Conference on Pattern Recognition*, pp. 610-612 , 1998.

[39] C.B. Jeong, H. K. Kwag, et al., “Identification of Font Styles and Typefaces in Printed Korean,” *6th International Conference on Asian Digital Libraries*, Vol.2911 , pp.666-669 , 2003.

[40] A. Zramdini, and R. Ingold, “Optical Font Recognition Using Typographical Features,” *Pattern Analysis and Machine Intelligence* ,Vol. 20, No. 8 ,PP. 877-882,1998.

[۴۱] ح.خسروی و ا. کبیر، معرفی دو ویژگی سریع و کارآمد برای بازشناسی ارقام دست‌نویس فارسی، *چهارمین کنفرانس ماشین بینایی و پردازش تصویر*، مشهد، دانشگاه فردوسی مشهد،۱۳۸۵.

Abstract

In recent decades, extensive research, writing recognition patterns include letters, numbers and other symbols commonly used in written documents in different languages is done. Given the progress made in the field of automatic text recognition technology called optical character recognition or OCR is formed. Text recognition is considered as an important part of e-government in our country and in recent years as the demand for a Persian text recognition system greatly enhanced. Due to the large amount of paper documents, digital image documents are converted by the scanner or camera, storage, efficient management and retrieval of these documents, files, in many applications, including office automation, and digital libraries are important.

In general, text recognition system includes several parts , such as receiving image preprocessing, configuration analysis, diagnostics language, font and finally text recognition. Research conducted in some topics , such as preprocessing is independent of the text language and can be used with any language . But some other topics , such as font recognition depends on the context and results of research conducted for other languages can not directly be applied to Persian . Most researches in the field of Persian literature on the recognition of images with high resolution images and text clean and false and document identification are with some known font. In research conducted for the recognition of Persian texts there are three approaches that based on separation words based on the overall shape recognition and mix of them.

This thesis aims at recognizing typed text written with the Iranian sans font, with a minimum size of 9 and the resolution is 300 dpi. according to The font style and readability, it is more addressed and every day the volume of computing and Internet environment is enhanced. This font style is replacement for the default Windows operating system font likeTahma, Despite readability, standard spacing between rows, beauty and consistency of Latin, this font has a structural complexity that complicates the process of recognition.

In this thesis after the production of database suitable, discrete and continuous characters classification were trained. then with solving the problem of overlapping of the words, separation approach is used to separate letters. Finally, the performance result of the system for processing some images of printed text, is provided where the separation accuracy of 96% and an accuracy of 85% was achieved in the identification.

Keywords: text recognition pen Iranian sans, separation-based approach, neural network classification