Vol. 2, No. 5, Jan-Mar 2023,(63-79)

Journal of New Achievements in Electrical, Computer and Technology (JNAECT)

مجله دستاوردهای نوین در برق،کامپیوتر و فناوری

دوره ۲، شماره ۵، زمستان ۱۴۰۱ (۷۹-۶۳)

journal homepage: https://jnaect.ir

تقسیم‌بندی و تشخیص زیرکلمات اسناد قدیمی دست‌نویس فارسی

سمیه صبوری[۱] ، حمیدرضا غفاری [۲]

[۱] دانشجوی دکترا ، دانشگاه آزاد اسلامی واحد بیرجند، Sabouri.somaye@gmail.com

[۲] استادیار دانشگاه آزاد اسلامی واحد فردوس، hamidghaffary53@yahoo.com

چکیده

اسناد تاریخی همواره مورد توجه مورخان و زبان شناسان است. اسناد مهم معمولاً توسط روش‌های تقسیم‌بندی و شناسایی به صورت دیجیتال تبدیل می‌شود. دیجیتالی کردن اسناد برای تحقیق بر روی این اسناد و حفاظت از آنها اهمیت فراوانی دارد. این مقاله یک چهارچوب تقسیم‌بندی و تشخیص کلی برای تصاویر اسناد تاریخی فارسی پیشنهاد شده است. ابتدا با حذف نویزها، رفع کجی، حذف مهرها و ... پیش پردازش روی اسناد انجام شده و تصویر سند به یک تصویر دو سطحی تبدیل می‌شود. در مرحله دوم یک روش تقسیم بندی متن سند به خطوط پیشنهاد شده است. در مرحله سوم یک روش تقسیم‌بندی خطوط به زیرکلمات رسم الخط فارسی را ارائه کرده و زیرکلمات این اسناد را استخراج نموده سپس از شبکه‌های عمیق برای آموزش زیرکلمات پرتکرار و تشخیص آنها استفاده شده و نتایج بر مبنای معیارهای مختلف گزارش شده است.

واژه‌های کلیدی: اسناد دست‌نویس فارسی، تقسیم‌بندی سند، تقسیم‌بندی خطوط، تقسیم‌بندی زیرکلمات ، شبکه عصبی

.

۱. مقدمه

متون و اسناد خطی نه تنها میراث فرهنگی یک کشور محسوب می‌شوند، بلکه حتی بعد از گذشت قرن‌ها، حاوی مطالب مفید علمی ، حقوقی و آموزشی برای محققان و پژوهشگران علوم انسانی، مذهبی، تاریخی، پزشکی، قضایی و ... هستند. غالباً نسخه اصلی این اسناد در بخش خصوصی از کتابخانه‌های بزرگ با تدابیر خاصی نگهداری می‌شود و به متقاضیان این مجموعه‌ها تنها تصاویر اسکن شده صفحات در قالب DVD، CD و یا میکروفیلم ارائه می‌گردد؛ و یا اینکه این اسناد ممکن است سندهای ملکی و حقوقی بوده که مالک سند تمایلی ندارد که اصل آن را در اختیار دیگران قرار دهد. از آنجایی که نسخه تایپی بسیاری از این اسناد تهیه نشده است لذا تحقیق روی این اسناد و استفاده از آنها تا حد زیادی محدود شده است. در ایران، گنجینه‌های بسیار گران بهایی از کتب، متون و اسناد بسیار مهم تاریخی و مذهبی دست‌نویس وجود دارد که در کتابخانه‌های بزرگ مانند کتابخانه ملی، کتابخانه آستان قدس و ... بصورت آرشیو در مکان‌های مخصوص نگهداری و محافظت می‌شوند. مجموع آمار و کتب و نسخ خطی موجود در ایران چیزی بالغ بر ۵۰۰ هزار جلد برآورد شده و از این میان، بیشترین نسخ مربوط به کتابخانه آستان قدس رضوی است که کل نسخ آن در طول سنوات تاریخ ۵۰ هزار نسخه خطی است. در عین حال، علاوه بر کتب و اسناد خطی که در کتابخانه‌ها وجود دارد، از آمار نسخ خطی که توسط اشخاص در منازل نگهداری می‌شود اطلاعاتی در دست نیست[۱]. متاسفانه این اسناد بدون هیچگونه حفاظت و شزایط نگهداری خاص در گوشه‌ای رها و در معرض تخریب قرار دارند. یک راه موثر برای حفظ این اسناد استفاده از فناوری تشخیص اسناد با کارایی بالا برای تبدیل تصاویر این اسناد به متون الکترونیکی است.

در این مقاله ما سعی داریم یک چهارچوب منسجم را برای بخش‌بندی طرح، تقسیم‌بندی خطوط متن، بخش‌بندی رشته کاراکتر و تشخیص زیرکلمات را ارائه نماییم.

۲. کارهای مرتبط

مدل‌های متفاوتی برای تشخیص سندهای دست‌نویس وجود دارد که هر کدام تلاش کرده‌اند از ترکیب روش‌های مختلف برای بهبود کار استفاده نمایند. در [۲] تشخیص سند تاریخی شامل چهار مرحله اصلی است: پیش‌پردازش، تقسیم‌بندی طرح، خط متن، بخش‌بندی رشته کاراکتر و در نهایت تشخیص کاراکتر. تقسیم‌بندی طرح گام مهمی در دیجیتالی‌سازی خودکار اسناد تاریخی است که روش‌های زیادی برای این کار در حوزه اسناد چاپی یا دست‌نویس ارائه شده است در [۱] این روش‌ها به سه دسته تقسیم شده‌اند: دسته اول تقسیم‌بندی بر اساس یک طرح خاص و از پیش تعیین شده انجام می‌شود [۳] ، [۴]. دسته دوم سعی می کند با تغییرات طرح سازگار شود تا بتواند طرح‌بندی را با الگوهای یکسان انجام دهد [۵]، [۶]، [۷]. در [۸] یک روش فیلتر بر اساس تشخیص نقطه گوشه هریس پیشنهاد شده است. در [۹] از یک تکنیک ساده برای استخراج متون استفاده شده است بدین صورت که تصویر به بلوک‌هایی تقسیم می‌شود و چگالی نقطه هر بلوک محاسبه می‌شود سپس بلوک‌های شامل نقاط بیشتر به عنوان بلوک‌های متنی طبقه‌بندی می شوند. دسته سوم با ترکیب بلوک‌های متنی [۱۰]، [۱۱]، [۱۲] یا با استفاده از فناوری هوش مصنوعی[۱۳]، [۱۴] بر محدودیت‌های دو روش قبل غلبه کند. روش‌های تقسیم‌بندی خط متن بر اساس تشخیص خط پایه [۱۵] و ردیابی منحنی خطوط [۱۶] برای تقسیم‌بندی مناطق متن به خطوط متنی مورد استفاده قرار می‌گیرند. برای به دست آوردن نرخ تشخیص خوب، خطوط متن باید با دقت بالایی تقسیم شوند. در [۱۷] یک رویکرد جدید برای تقسیم‌بندی ارقامی که همپوشانی دارند ارائه شده است. در [۱۸] یک روش پیشنهادی برای جداسازی متن از تصویر سند را با استفاده از الگوریتم تقسیم‌بندی Otsu و روش تبدیل Hough ارائه شده است. در [۱۹] از یک شبکه عصبی

کانولوشن به عنوان یک طبقه بند بر روی اسناد خطی سریانی استفاده نموده است. در [۲۰] یک روش استخراج اطلاعات از جداول را پیشنهاد می‌دهد که برمبنای آن مرزهای جداول تشخیص داده شده و تصویر سند به همان نواحی تقسیم می‌شود. با توجه به ویژگی‌های خط فارسی و متفاوت بودن سبک‌های نوشتاری این زبان، محققین با چالش‌های زیادی در این حوزه رو به رو هستند وکارهای زیادی نیز در این حوزه انجام نشده است. در [۲۱] یک روش جدید برای بهبود سیستم تشخیص کاراکتر بر اساس پایداری و انعطاف‌پذیری پیشنهاد شده است. در [۲۲] لازم است کاربر با الگوبرداری از سبک نوشتاری نویسنده کتاب، کلمه مورد جست‌وجوی خود را با یک قلم طراحی و تا حد ممکن شبیه دست خط نویسنده ترسیم نموده و تصویر طراحی شده را به سیستم ارائه دهد.

۳. چهارچوب تقسیم‌بندی و شناخت کلی اسناد قدیمی دست‌نویس فارسی

۳-۱ مقدمه

در این بخش یک فرآیند کامل برای طراحی الگوریتم و روشی برای جست‌وجو در تصاویر اسکن شده‌ی اسناد دست‌نویس تاریخی پیشنهاد می‌شود. از آنجا که رسم الخط فارسی با سایر زبان‌ها خصوصاً رسم الخط لاتین تفاوت‌های زیادی دارد لذا قبل از هر کاری این تفاوت‌ها را که تأثیر بسزایی در طراحی سیستم خواهند داشت بیان می‌شود. مهمترین این تفاوت‌ها عبارتند از:

الف) رسم الخط فارسی بر خلاف رسم الخط لاتین از راست به چپ نوشته می‌شود.

ب) کلمات فارسی از چند بخش مجزا تشکیل شده که به آنها زیر کلمه می‌گویند و هر زیر کلمه شامل یک یا چند حرف است. به عنوان مثال کلمه‌ی "دست‌نویس" از ۴ زیر کلمه‌ی {د، ست، نو، یس} تشکیل شده است. زیر کلمات بدلیل نچسبیدن برخی از حروف فارسی مثل {ا، د، ذ، ر، ز، و} به حرف بعدی بوجود می‌آیند.

ج) حروف فارسی برخلاف حروف لاتین دارای نقطه می‌باشند. از ۳۲ حرف فارسی، ۱۷ حرف دارای نقطه هستند که تعداد این نقاط از یک تا سه متغیرند که ممکن است در بالا یا پایین حروف قرار بگیرند. به بخشی از حرف یا زیرکلمه که نقاط آن حذف شده باشد بدنه‌ی حرف یا زیر کلمه گفته می‌شود.

د) برخی از حروف دارای بخش مجزایی هستند مثلاً حرف " آ " دارای یک علامت مد است که همواره از بدنه‌ی اصلی جداست. علاوه بر این حروف "ک" و "گ" دارای سرکش هستند که گاهی اوقات سرکش‌ها مجزا از بدنه‌ی حرف نوشته می‌شود.

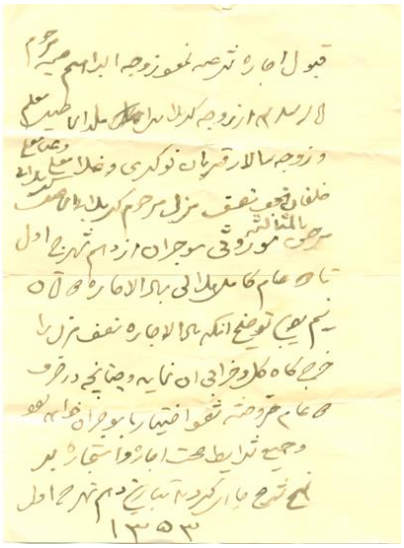
ه) در بالای بدنه‌ی بعضی حروف ممکن است تشدید وجود داشته باشد مثل "الله". بعضی حروف مانند {أ، إ، ؤ} نیز دارای همزه هستند که بالا یا پایین حروف نوشته می‌شود.

برخی از حروف بدنه‌ی مشابی دارند و تفاوت آن‌ها تنها در تعداد یا محل قرار گیری نقطه‌هاست. این موضوع باعث می‌شود که زیرکلماتی بوجود بیایند که زیرکلمات متفاوت با بدنه‌ی اصلی یکسان داشته باشیم مانند کلمات {رحیم – رجیم، منیره – منیژه، نقی – تقی}.

۳-۲ ایجاد پایگاه داده

برای طراحی و آزمایش هر سیستم مکان‌یابی کلمات، به یک پایگاه داده نیاز است. برای این منظور از تصاویر اسکن شده‌ی اسناد قدیمی دست‌نویس فارسی که در منازل افرادی در بافت‌های مسکونی شهرستان گناباد نگه‌داری می‌شد، استفاده شده

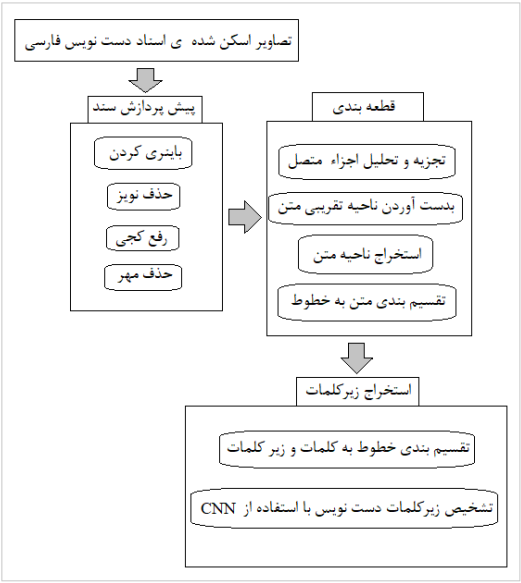
است. این پایگاه داده شامل ۳۵۰سند مختلف بوده که تمامی آنها با وضوح 300 dpi و با فرمت TIF اسکن شده‌اند. متن این اسناد با قلم نی، با جوهرهای رنگی و دستخط‌های متفاوت و توسط افراد مختلفی نوشته شده است. این اسناد شامل اسناد ملکی (زمین، منزل، باغ و ...) و حقوقی می‌باشند. در شکل ۱ یک نمونه از این اسناد (سند DOC_0004) را ملاحظه می‌کنید.



شکل ۱ : نمونه‌ای از اسناد پایگاه داده تهیه شده

۳ – ۳ ساختار کلی سیستم در اسناد تاریخی دست‌نویس فارسی

ساختار کلی این سیستم برای متون دست‌نویس فارسی را می‌توان به صورت شکل ۲ طراحی نمود. در ادامه، بخش‌های مختلف سیستم طراحی شده که بر همین اساس پیاده‌سازی شده است مورد بررسی قرار می‌گیرد.



شکل ۲ : چهارچوب کلی سیستم تقسیم‌بندی و نشخیص زیرکلمات اسناد دست نویس فارسی

۳ – ۳ – ۱ پیش‌پردازش

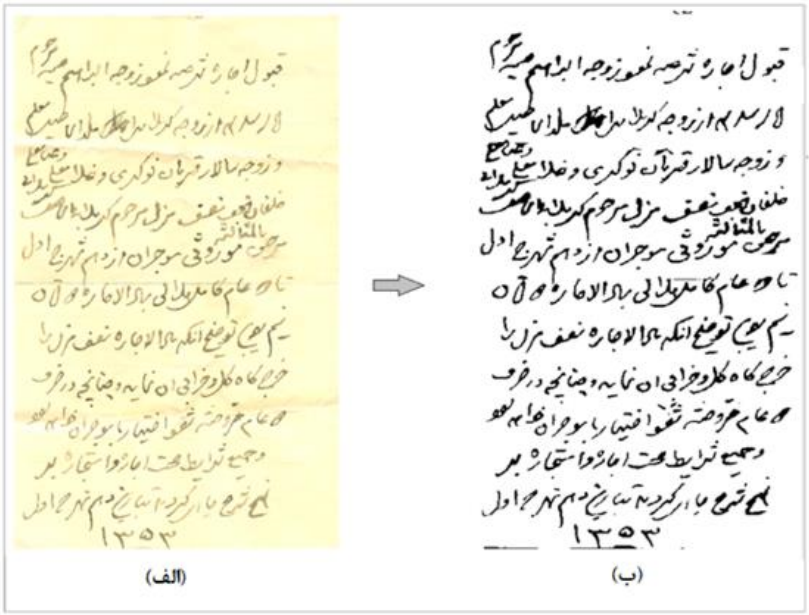
از آنجایی که شرایط نگهداری، قدمت و نوع کاغذ استفاده شده در اسناد و کتب تاریخی یکسان نیستند بنابراین نویزهای بوجود آمده در هریک از آنها ممکن است متفاوت باشند؛ از این رو انجام تمام مراحل پیش پردازش که در بخش دوم به آنها اشاره شده است ضرورتی ندارد. کیفیت تصاویر موجود در پایگاه داده، دقت مورد نیاز، نوع ویژگی‌های لازم برای استخراج از

تصاویر و نوع الگوریتم تطبیقی همگی ارتباط مستقیم با این قضیه دارند. با توجه به به اسناد موجود در پایگاه داده، پیش پردازش‌های زیر برای طراحی این سیستم در نظر گرفته شده است:

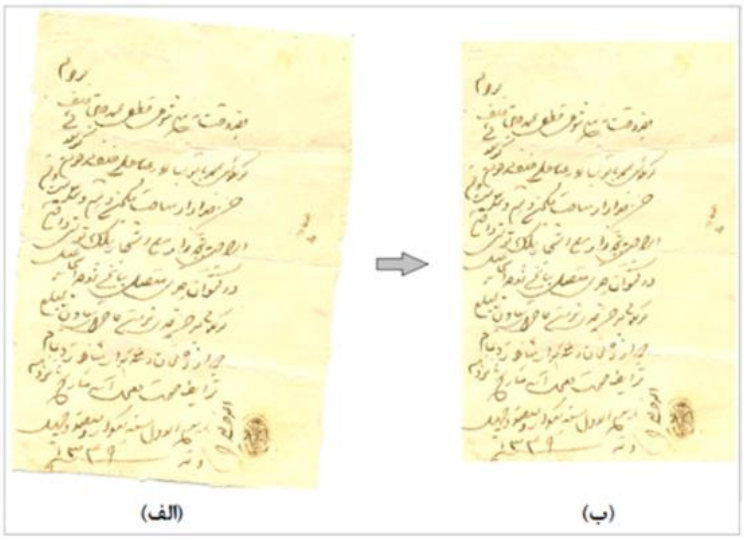
– دو سطحی سازی : در اولین گام از عملیات پیش پردازش ابتدا تصاویر پایگاه داده را که به صورت رنگی اسکن شده‌اند به تصاویر خاکستری رنگ تبدیل نموده و سپس با مقدار آستانه‌گیری به تصاویر دو سطحی(سیاه – سفید) تبدیل می‌شوند. در [۲۴] و [۲۵] برای آستانه‌گیری از روش otsu استفاده نموده‌اند در این اسناد از مقدار آستانه‌ی ۰.۸۵ استفاده شده است که نتایج بهتری را نسبت به روش otsu دارد.

– حذف نویز: از آنجا که بعد از دو سطحی‌سازی تصاویر، یک تصویر با نقطه‌های ریز سیاه رنگ شبیه به نویزهای فلفل - نمکی بوجود می‌آید، لذا برای حذف آنها از عملیات فیلترینگ نویز فلفل- نمکی استفاده شده است . در انتها نیز با عملیات مورفولوژیکی حفره‌های موجود را پر شده است.

– رفع کجی: در گام بعد باید کجی تصاویر که می‌تواند ناشی از مرحله‌ی اسکن کردن و یا عوامل دیگری باشد برطرف شود. برای این منظور، از تصویر یک پروفایل‌گیری افقی انجام شده و با روشی که در[۲۵] بیان شده است زاویه‌ی کجی سند را پیدا نموده و با چرخش تصویر در جهت معکوس این مشکل برطرف شده است. در شکل‌های ۳ و ۴ نتیجه‌ی این عملیات نشان داده شده است.



شکل ۳ : الف) تصویر اولیه‌ی سند ب) سند دو سطحی شده با حداقل نویز



شکل ۴: الف)تصویر سند کج اسکن شده ب) تصویر سند پس از رفع کجی

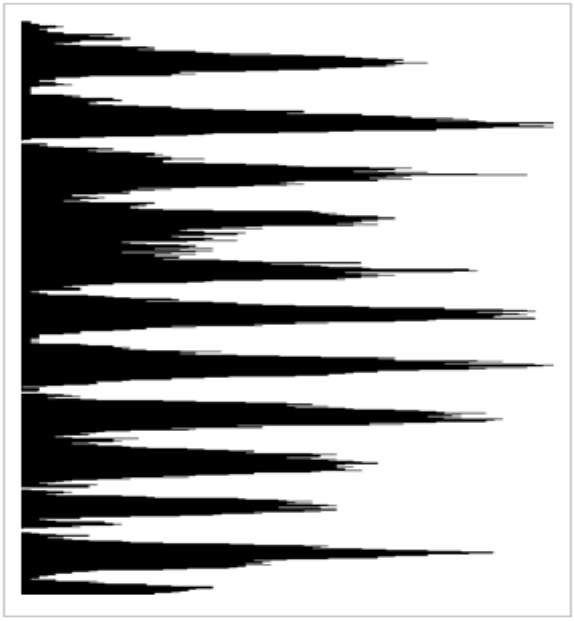
– حذف مهر: از آنجایی که اسناد جمع آوری شده اسناد ملکی بوده که جنبه‌ی حقوقی دارند، لذا شامل مهر و اثر انگشت بوده که نشان دهنده‌ی هویت سند می‌باشند. یکی از کارهایی که در مرحله‌ی پیش‌پردازش مطرح شد، حذف مهرها و اثر انگشت‌های درج شده در اسناد بود، که حذف آنها چالش‌های زیادی را به همراه داشت. از جمله‌ی آنها می‌توان به همرنگ بودن مهر یا اثر انگشت با جوهر متن نوشته شده و یا پخش‌شدگی جوهر مهر روی کاغذ اشاره نمود. برای حذف مهرها و اثر انگشت‌ها از عملیات مورفولوژی استفاده شد[۲۷] که نتایج قابل قبولی داشت. نمونه‌ای از این اسناد در شکل ۵ نشان داده شده است.

شکل ۵ : مهرهای درج شده بر روی سند با رنگ‌ها و طرح‌های مختلف

۳ - ۳ - ۲- قطعه‌بندی

در این مرحله باید کار جداسازی زیر کلمات را از درون تصاویر پیش پردازش شده سند مورد آزمایش، انجام شود. این کار در دو مرحله انجام خواهیم شد؛ یعنی ابتدا از تصویر سند خطوط را جدا نموده و سپس از درون این خطوط زیر کلمات را جداسازی خواهند شد. علت دو مرحله‌ای نمودن این عملیات به تفاوت میان نوشتار لاتین و فارسی مربوط است. همانطور که در ابتدا بیان شد زبان فارسی برخلاف زبان لاتین دارای نقاط، سرکش و مد است، اگر بخواهیم مستقیماً جداسازی زیرکلمات را انجام دهیم، در جاهایی که فاصله‌ی خطوط از هم کم است، کار اختصاص این نقاط و سرکش‌ها دشوارتر و با خطای زیادی همراه خواهد بود.

الف) جداسازی خطوط: برای جداسازی خطوط سند، از نمودار هموارسازی شده‌ی پروفایل افقی استفاده شده است. همانطور که در ۶ مشاهده می‌شود ابتدا یک پروفایل‌گیری افقی از سند انجام شده و سپس خطوط از تصویر سند جدا شده‌اند. همچنین با توجه به متنوع بودن اسناد سعی شده از فیلترهای دو بعدی جهت‌دار [۲۸] نیز برای جداسازی خطوط استفاده شود. در شکل ۷ خط اول یکی از اسناد نشان داده شده است.



شکل ۶ : پروفایل‌گیری افقی از تصویر سند شکل ۱

شکل ۷ : جداسازی خطوط سند شکل ۳۱ با پروفایل‌گیری افقی

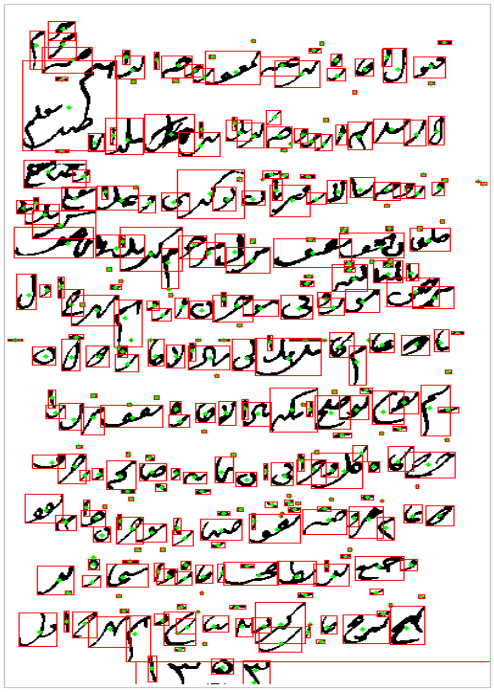
ب) جداسازی زیر کلمات از تصویر سند: به دلیل نوع و سبک نوشتاری متون دست‌نویس فارسی، این مرحله از کار یکی از سخت‌ترین مراحل می‌باشد. زیرا اولاً در بسیاری از موارد زیر کلمات بخاطر عدم دقت نویسنده و یا پخش شدن جوهر قلم هنگام نوشتن به هم چسبیده هستند و ثانیاً در بسیاری از این اسناد بدلیل سبک نگارش نویسنده و یا برای زیبایی نوشتار، زیر کلمات سوار بر هم نوشته شده‌اند که این کار جداسازی زیر کلمات و اختصاص نقاط به بدنه‌ی اصلی زیر کلمه‌ی مربوط را

دشوار و با خطای بیشتری مواجه می‌کند . برای جداسازی زیر کلمات در تصویر یک سند، ابتدا باید تمامی اجزاء متصل در سند را شناسایی کرده سپس به کمک ویژگی‌هایی مانند مساحت، نسبت ظاهری و موقعیت نسبت به خط زمینه، نوع هر جزء را که مربوط به یکی از گروه‌های زیر خواهند بود تعیین شوند.

- بدنه‌ی اصلی زیر کلمات (به جز حرف "ا")
- نقطه‌ها و تشدیدها
- سرکش‌های جدا از بدنه مثل "ک" ، "گ" و مد روی حرف "آ"
- زیر کلمه "ا"

در ادامه می‌بایست نقاط، تشدیدها و سرکش‌ها به بدنه‌ی اصلی که به آن تعلق دارند اختصاص داده شود. همانطور که قبلاً هم گفته شد، این کار در متون تاریخی که همپوشانی بین کلمات زیاد است و حتی گاهی زیر کلمات بر روی هم نوشته شده‌اند، امری دشوار و با خطای نسبتاً بالایی همراه خواهد بود. برای انجام این کار ابتدا یک قطعه‌بندی اولیه از کل سند انجام شده که نتیجه‌ی آن در شکل ۸ نشان داده شده است.

در این الگوریتم ما نقاط و تشدیدها را به بدنه‌ی زیر کلمه‌ای نسبت می‌دهیم که بیشترین همپوشانی عمودی را با آن دارد و اگر مقدار همپوشانی برای دو زیر کلمه یکسان شد، زیر کلمه‌ای انتخاب خواهد شد که فاصله‌ی افقی مرکز نقطه تا مرکز آن زیر کلمه کمتر است. در مورد سرکش‌ها نیز، به بدنه‌ی اصلی اختصاص خواهد یافت که فاصله‌ی افقی مرکز آن بدنه با ضلع چپ قاب دربردارنده‌ی سرکش مینیمم باشد. شناسایی حرف "ا" نیز برای اختصاص دادن مدهای شناسایی شده به آنها صورت می‌گیرد. سپس سعی شده تا هر یک از گروه‌های بالا را جداسازی شده و با رنگ‌های مختلف نمایش داده شوند. در این جداسازی، گروه اول یعنی بدنه‌ی اصلی زیرکلمات (به جز حرف "ا") با رنگ قرمز، گروه دوم یعنی نقطه‌ها و تشدیدها با رنگ آبی و گروه سوم یعنی زیرکلمه‌ی "ا"با رنگ سبز نشان داده شده است. این جداسازی در شکل ۹ نشان داده شده است.

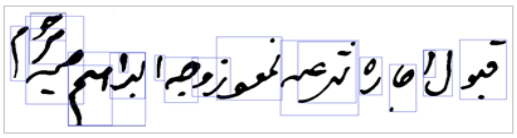


شکل۸: قطعه‌بندی اولیه سند شکل ۱

شکل ۹ : جداسازی اجزاء پیوسته در یک خط از سند و تمایز انواع آن

در ادامه می‌بایست نقاط، تشدیدها و سرکش‌ها را به بدنه‌ی اصلی که به آن تعلق دارند اختصاص داده شود. همانطور که قبلاً هم گفته شد، این کار در متون تاریخی که همپوشانی بین کلمات زیاد است و حتی گاهی زیر کلمات بر روی هم نوشته شده‌اند، امری دشوار و با خطای نسبتاً بالایی همراه خواهد بود.

نتیجه‌ی ادغام نقاط، تشدیدها و سرکش‌های یک زیر کلمه در شکل ۱۰ نشان داده شده است. این قطعه‌بندی با خطاهایی نیز همراه بوده است، مثلاً حرف "ا" مربوط به کلمه‌ی "اجاره" به حرف "ل" مربوط به "قبول" در یک قاب قرار گرفته‌اند؛ این خطاها به دلیل اینکه حروف سوار بر هم نوشته شده‌اند اجتناب ناپذیر است.



شکل ۱۰ : زیرکلمات جداسازی شده نهایی برای یک خط سند شکل ۱

۳ - ۳ - ۳- استخراج ویژگی

در این مرحله می‌بایست از تصاویر زیرکلمات ویژگی‌های مناسبی استخراج شود بطوریکه در فضای این ویژگی‌ها اولاً بین الگوهای مختلف یک زیرکلمه‌ی یکسان تشابه بالا وجود داشته باشد و ثانیاً تمایز کافی بین الگوهای زیرکلمات متفاوت ایجاد کند. برای این منظور از دو دسته از ویژگی‌ها استفاده شده است.

دسته‌ی اول، ویژگی‌های آماری مبتنی بر ستون‌های تصویر هستند. این ویژگی‌ها عبارتند از : پروفایل بالا و پایین، پروفایل افکنش عمودی و تعداد گذار بین پیکسل‌های جوهر و پس زمینه در راستای عمودی. البته با توجه به ساختار نوشتار فارسی پروفایل‌های بالا و پایین را یکبار برای تصویر کلی زیرکلمه و بار دیگر تنها برای بدنه‌ی اصلی زیرکلمه محاسبه خواهد شد.

پروفایل بالا، هیستوگرامی است که از متصل کردن بالاترین پیکسل جوهر در ستون‌های تصویر به یکدیگر ایجاد می‌گردد. در مورد پروفایل پایین، پایین‌ترین پیکسل‌های جوهر مورد نظر خواهند بود. فرض کنید تصویر I با ابعاد h سطر و w ستون داریم و شدت پیکسل سطر r ام، ستون c ام را بصورت I(r,c) نمایش دهیم در اینصورت این دو ویژگی را می‌توان از روابط ۱ و ۲ محاسبه کرد:

$$up(I,c)=\begin{cases} undefined & if \quad \forall r, I(r,c)=1 \\ \{\min(r) \mid I(r,c)=0\} & otherwise \end{cases} \qquad (1)$$

$$lp(I,c)=\begin{cases} undefined & if \quad \forall r, I(r,c)=1 \\ \{\max(r) \mid I(r,c)=0\} & otherwise \end{cases} \qquad (2)$$

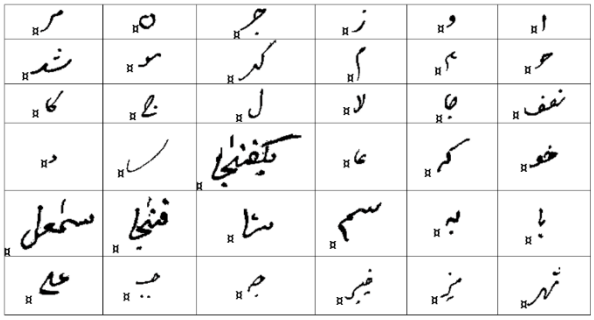
پروفایل افکنش عمودی نموداری است که از شمارش تعداد پیکسل‌های جوهر در هر ستون تصویر ایجاد می‌شود. با در نظر گرفتن مفروضات قسمت قبل، افکنش عمودی را از معادله‌ی ۳ بدست می‌آوریم:

$$pp(I,c)=\sum_{r=1}^{h}\left(1-I(r,c)\right) \quad (3)$$

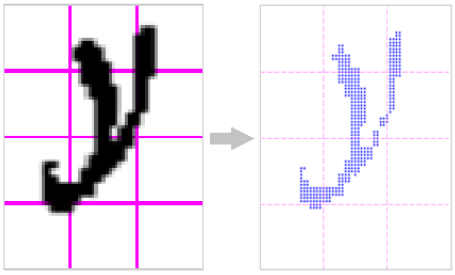
پس از بدست آوردن پروفایل افکنش از رابطه‌ی بالا، همانند پروفایل مرزی آن را هموارسازی کرده و سپس به بازه‌ی [0 1] نرمالیزه می‌کنیم.

دسته‌ی دوم، ویژگی‌های آماری مبتنی بر ناحیه‌بندی تصویر هستند که از چگالی جوهر در نواحی تصویر استفاده خواهیم کرد. با آزمایشات اولیه‌ای که بر روی این ویژگی انجام شده است، استخراج آن برای تنه‌ی اصلی زیر کلمه مناسب‌تر شناخته شد. از این رو در این روش ناحیه‌بندی بر روی تصویر بدنه‌ی اصلی زیرکلمه انجام می‌شود. نمونه هایی از زیر کلمات استخراج شده در جدول ۱ نشان داده شده است.

جدول۱: برخی از زیرکلمات انتخابی برای آزمایش ویژگی‌ها



همانطور که قبلاً نیز بیان شد، در ناحیه‌بندی، قاب در برگیرنده تصویر را به چند ناحیه همپوشان یا غیر همپوشان با مساحت‌های مساوی یا متفاوت تقسیم می‌کنند. در این روش یک ناحیه‌بندی یکنواخت ۳×۴ را بر روی بدنه‌ی اصلی زیر کلمه بکار می‌بریم(برای کلماتی که دارای کشیدگی افقی یا عمودی هستند این ناحیه‌بندی به طور مجزا انجام می‌شود)؛ یعنی ابتدا از تصویر زیرکلمه، بدنه‌ی اصلی را استخراج کرده و سپس قاب دربرگیرنده‌ی آن را در راستای افقی به ۳ قسمت و در راستای عمودی به ۴ قسمت تقسیم می‌کنیم. هدف ما در واقع استخراج چگالی جوهر در این نواحی است. برای این منظور، تعداد پیکسل‌های سیاه در هر ناحیه شمارش شده و بر مساحت ناحیه تقسیم می‌شوند. بنابراین ما یک بردار به طول ۱۲ از مقادیر چگالی بدست می‌آوریم. . نتایج این روند برای زیرکلمه‌ی "لا" در شکل ۱۱ نشان داده شده است.



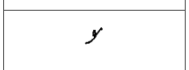
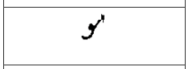
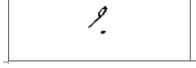
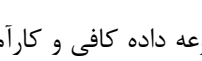
شکل ۱۱: تقسیم زیرکلمه‌ی "لا" به نواحی یکسان

خروجی این مرحله برای پیکسل‌های شمارش شده در یک ماتریس ۳×۴ نشان داده می‌شود. با شمارش پیکسل‌های هر ناحیه باید بردار ویژگی بر اساس آنها استخراج شود؛ برای تولید بردار ویژگی از رابطه‌ی ۴ استفاده می‌شود.

$$A=\begin{bmatrix} s_1 & s_2 & s_3 \\ s_4 & s_5 & s_6 \\ s_7 & s_8 & s_9 \\ s_{10} & s_{11} & s_{12} \end{bmatrix} \quad C=\begin{bmatrix} \frac{s_1}{B}=f_1 \\ \frac{s_2}{B}=f_2 \\ . \\ . \\ \frac{s_{12}}{B}=f_{12} \end{bmatrix} \qquad \frac{width \times height}{N} \Rightarrow B \tag{4}$$

مقدار شباهت بین زیرکلمات بر مبنای معیارهای بیان شده در بخش قبل برای زیرکلمات بدست آمده که مقدار شباهت برای دو زیرکلمه در جدول ۲ نشان داده شده است.

جدول ۲: زیرکلمات مشابه برای دو زیرکلمه‌ی "شهر" و "مر"

زیرکلمه‌ی مورد آزمایش	زیرکلمات با بالاترین شباهت	زیرکلمه‌ی مورد آزمایش	زیرکلمات با بالاترین شباهت
			
			
			
			
			
			

۳ – ۳ – ۴- داده افزایی مجموعه داده پیشنهادی

یکی از چالش‌های مهم شبکه‌های عصبی کانولوشنی، عدم وجود مجموعه داده کافی و کارآمد برای آموزش این شبکه‌ها است. به عبارت دیگر، در اکثر مسائل دنیای واقعی داده‌های میلیونی و برچسب‌دار برای آموزش شبکه عصبی وجود ندارد. در این راستا، برخی محققان تلاش کرده‌اند که مشکل کمبود داده را با تولید داده‌های مصنوعی جبران نمایند [۲۷]. با توجه به اینکه فرآیند جمع‌آوری تصویر اسناد و استخراج زیرکلمات از آنها یک فرآیند زمان‌بر است، در این تحقیق، از روش‌های داده افزایی (مانند ایجاد نویز، چرخش و انتقال تصویر) برای افزایش تصاویر مجموعه داده پیشنهادی استفاده شده است.

۴. نتایج

۴ – ۱ معیارهای ارزیابی.

در این مقاله از تصاویر اسکن شده‌ی اسناد فارسی دست‌نویس قدیمی که با عنوان سند مالکیت، اجاره و ... محسوب می‌شوند استفاده شده، که این اسناد با فرمت Tif و وضوح تصویر 300 dpi اسکن شده و یک پایگاه داده که شامل ۱۰۸ سند مختلف است ایجاد شده است. در سیستم پیشنهادی، ابتدا سعی شده تا عملیات پیش پردازش بطور کامل بر روی اسناد انجام شده تا اسنادی با حداقل نویز برای مراحل بعد تهیه شود. در مرحله‌ی بعد تصاویر هر یک از این اسناد به خطوط جدا از هم تقسیم شده و سپس از درون خطوط جداسازی شده زیرکلمات استخراج می‌شوند. با توجه به اینکه سبک نوشتاری این اسناد به صورتی است که زیرکلمات زیادی برای زیبایی کار به شکل همپوشان و یا سوار بر هم نوشته شده‌اند که در آنها اختصاص نقاط، تشدید و... به زیرکلمه‌ی مربوط کاری سخت بوده و با خطای زیادی همراه است. پس از جداسازی تمام زیرکلمات موجود

در اسناد، بر اساس تقسیم‌بندی صورت گرفته بردار چگالی نواحی برای هر یک از زیرکلمات استخراج شده است. معیار انتخابی برای استخراج ویژگی هر یک از زیرکلمات معیار فاصله‌ی اقلیدسی بوده است. در این مطالعه نیز با اعمال مرحله‌ی قطعه‌بندی بر روی اسناد مورد آزمایش، تعداد ۲۸۴۰ زیرکلمه که توسط سیستم به عنوان زیرکلمه شناسایی شده بودند استخراج شدند. از این میان، کلمات با تکرار حداقل ۵ بار به عنوان زیرکلمات مورد استفاده در این مجموعه انتخاب شدند. سپس بردار چگالی تمام نواحی زیرکلمات مطابق توضیحات داده شده در مراحل قبل استخراج شده و به عنوان بردار ویژگی برای آن زیرکلمه در نظر گرفته شده است. در ادامه با معیار شباهت پیشنهادی، شباهت بین دو زیرکلمه‌ای که شرط اولیه‌ی تطبیق را داشته‌اند محاسبه شده است. شرط اولیه بدین صورت بیان شده که نسبت مساحت دو تصویر و نسبت ابعاد ظاهری آن خیلی بزرگ یا خیلی کوچک نباشد. زیرا به احتمال زیاد چنین تطبیقی نمی‌تواند مربوط به دو زیرکلمه‌ی یکسان باشد. در مورد معیار شباهت بکار رفته هم باید توجه نمود که معیار تطبیقی که برای تصاویر دارای پهنای کم مثل زیرکلمه‌ی"ا" در مقایسه با زیرکلمه‌ی عریضی مثل "فنجا" استفاده شده، متفاوت بوده است.

از روی اسناد پیش‌پردازش شده و قطعه‌بندی شدی ۲۸۴۰ زیر کلمه استخراج شده که این زیر کلمات در ۲۵ کلاس دسته‌بندی شده‌اند. نمونه‌ای از این زیرکلمات در جدول ۲ قابل مشاهده است. در ابتدا از یک طبقه بند KNN با سه همسایگی برای طبقه‌بندی این زیرکلمات استفاده شده است. سپس از معیارهای رابطه ۵ برای محاسبه دقت این طبقه‌بند استفاده شده که به صورت زیر محاسبه می‌شود:

$$\text{Accuracy}=\frac{TP+TN}{TP+\text{TN}+FP+FN} \quad (5)$$

میزان بهینه برای دقت، ۹۱.۶۲٪ بدست آمده است.

۴ – ۲ شبکه کانولوشنی

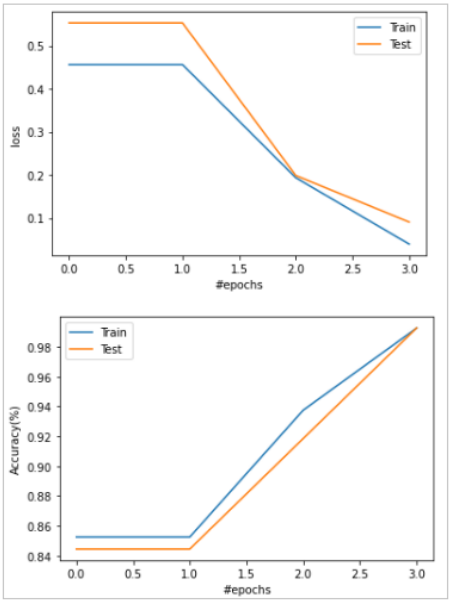
در این تحقیق برای تعیین میزان کارایی مجموعه داده پیشنهادی در تشخیص زیرکلمات، از دو شبکه کانولوشنی VGG-19 و ResNet50 [۲۸] استفاده شده است. در این شبکه‌ها۳۰٪ از داده‌های استخراج شده به عنوان داده‌های تست و ۷۰٪ داده‌های باقی مانده به عنوان داده‌های آموزشی مورد استفاده قرار گرفته است. سپس از سه معیار صحت، فراخوان و معیار F، مطابق روابط ۸-۶ برای ارزیابی میزان کارایی این شبکه‌ها استفاده شده که نتایج آن در جدول ۳ نشان داده شده است.

$$\text{precision}=\frac{TP}{TP+FP} \quad (6)$$

$$\text{Recall}=\frac{TP}{TP+FN} \quad (7)$$

$$\text{F-score}=\frac{2*Precision*Recall}{Precision+Recall} \quad (8)$$

بهترین مقدار برای فاکتور دقت در بازه ۹۶٪ تا ۹۸٪ قرار می‌گیرد که نمودار آن در شکل ۱۲ مشاهده می‌شود.



شکل ۱۲: نمودار خطا و دقت محاسبه شده بر روی پایگاه داده

جدول ۳: مقایسه نتایج بدست آمده از مدل‌های کانولوشنی

F-score	Recall	Accuracy	مدل
81%	82%	98%	VGG-19
73%	75%	91%	Res-Net50

با توجه با اینکه از چالش‌های مهم اسناد دست نویس به دلیل تنوع در سبک نوشتاری آنها بحث جداسازی زیرکلمات می‌باشد؛ در اسناد مختلف انگلیسی، عربی، چینی، هندی و ... از روش‌های مختلفی استفاده شده است که هر یک از این روش‌ها متناسب با سبک نوشتاری اسناد و چالش‌های جداسازی زیر کلمات آنها طراحی شده است. اما در تشخیص کلمات از یادگیری عمیق استفاده شده است که نتایج قابل قبولی داشته‌اند که نمونه‌هایی از آنهادر جدول ۴ نشان داده شده است.

جدول ۴: استفاده از شبکه‌های عمیق در دست نوشته‌های متفاوت

دست نوشته	Accuracy
Handwritten Syriac[19]	97.58%
MNIST[17]	98.86%
English typescripts[18]	91%
Historical Tibetan Document[3]	86.60%

۵. نتیجه‌گیری

در این مطالعه یک چهارچوب کلی برای تقسیم‌بندی و تشخیص زیرکلمات از تصاویر دست‌نویس قدیمی فارسی پیشنهاد شد که بر مبنای آن ابتدا پیش پردازش‌هایی روی سند انجام شده سپس تقسیم‌بندی سند به خطوط و سپس تقسیم‌بندی به کلمات و زیرکلمات انجام می‌شود. برای تقسیم‌بندی متن به خطوط از پروفایل‌گیری افقی و برای تقسیم‌بندی خطوط به

زیرکلمات از ساختار بلوک‌بندی استفاده شده است. درنهایت از یک شبکه عصبی عمیق برای تشخیص زیرکلمات استفاده شده که. نتایج تجربی نشان می‌دهد که روش پیشنهادی عملکرد مطلوب و رضایت بخشی دارد.

۵. منابع و مراجع

۱. علی‌آبادی محمد. یک روش جدید برای مکان‌یابی کلمات در متون تاریخی دست‌نویس فارسی، پایان‌نامه کارشناسی ارشد، دانشگاه بیرجند، ۱۳۹۱

2. Sébastien, P. Gomez-Krämer, J. M. Ogier.(2017). A comprehensive survey of mostly textual document segmentation algorithms since 2008, Pattern Recognition, vol. 64, pp. 1-14.

3. D. T. Ha, N. Duc-Dung, D. H. Le.(2016). An adaptive over-split and merge algorithm for page segmentation, Pattern Recognition Letters, vol. 80, pp. 137-143.

4. Ma, Longlong, Congjun Long, Lijuan Duan, Xiqun Zhang, Yanxing Li, and Quanchao Zhao. (2020). Segmentation and recognition for historical Tibetan document images. IEEE Access 52641-52651.

5. M. Yu, Q. Guo, D. Wang, Y. Yu.(2013). Improved connectivity-based layout segmentation method, Computer Engineering and Applications, vol. 49, no. 17, pp. 195-198.

6. R. Xiao.(1993) .Research on the Method of Extracting Ancient Yi Text from Complex Background, M.S. thesis, South-Center University for Nationalities, Osaka, China.

7. S.S. Bukhari, T. M. Breuel,(2012) Layout Analysis for Arabic Historical Document Images Using Machine Learning, in Proc. International Conference on Frontiers in Handwriting Recognition, pp. 639-644.

8. A.S. Kavitha, P. Shivakumara, G.H. Kumar.(2016). Text segmentation in degraded historical document images, Egyptian Informatics Journal, vol. 17, no. 2, pp. 189-197.

9. F. Zeng, G. Zhang, J. Jiang(2013). Text Image with Complex Background Filtering Method Based on Harris Corner-point Detection., Journal of Software, vol. 8, no. 8, pp. 1827-1834.

10. Y. Vikas, N. Ragot.(2016). Text Extraction in Document Images: Highlight on Using Corner Points, in Proc. IAPR International Workshop on Document Analysis Systems , pp. 281-286.

11. J.Y. Ramel, S. Leriche, M. L. Demonet.(2007). User-driven page layout analysis of historical printed books, International Journal on Document Analysis and Recognition, vol. 9, no. 2, pp. 243-261.

12. A. Winter, T. Andersen, E. H. B. Smith.(2011). Extending Page Segmentation Algorithms for Mixed-Layout Document Processing, in Proc. International Conference on Document Analysis and Recognition , pp. 1245-1249.

13. V. Singh, B. Kumar.(2014). Document layout analysis for Indian newspapers using the contour-based symbiotic approach, in Proc. International Conference on Computer Communication and Informatics, pp. 1-4.

14. K. Chen, M. Seuret, M. Liwicki, J. Hennebert, R. Ingold.(2015). Page segmentation of historical document images with convolutional autoencoders, in Proc. International Conference on Document Analysis and Recognition, pp. 1101-1105.

15. K. Chen, C. Liu, M. Seuret, M. Liwicki, J. Hennebert, R. Ingold. (2016). Page Segmentation for Historical Document Images Based on Superpixel Classification with Unsupervised Feature Learning, in Proc. IAPR Workshop on Document Analysis Systems, pp. 299-304.

16. Z. Li, W. Wang, Y. Chen, Y. Hao.(2019). A novel method of text line segmentation for historical document image of the urchin Tibetan. J. Vis. Commun. Image R., vol. 61, pp. 23-32.

17. F. Zhou, W. Wang, Q. Lin. (2018). A novel text line segmentation method based on contour curve tracking for Tibetan historical documents. International Journal of Pattern Recognition and Artificial Intelligence, vol. 21.

18. Demır, Ali Alper, and Ufuk Özkaya. (2020). A Semantic Segmentation Based Approach for Segmentation and Recognition of Touching and Overlapping Digits 28th Signal Processing and Communications Applications Conference (SIU).

19. Desai, Sujata S., Darshana Rajput, and Kiran Patil. (2020). An approach for Text Recognition from Document Images. IEEE Bangalore Humanitarian Technology Conference (B-HTC).

20. Fermanian, Rita, et al. (2020). Deep Recognition-based Character Segmentation in Handwritten Syriac Manuscripts. Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA).

21. Yin, Z. C., et al. (2020). Recognition and Extraction of Information from Image-based Tables for Electric Power System Operation and Maintenance International Conference on Smart Grids and Energy Systems (SGES).

22. Sadri, J., Suen, C. Y., and Bui, T.D..(2006). A New Clustering Method for Improving Plasticity and Stability in Handwriting Character Recognition System" In 18th Int,1 conf. on Pattern Recognition, pp. 1130-1133.

23. Xiaobo Chen, Jian Yang, Qiaolin Ye, Jun Liang.(2011). Recursive Projection Twin Support Vector Machine via Within-Class Variance Minimization, in Pattern Recognition 44 (10–11), pp. 2643–2655.

24. Farrahi-Moghaddam R., Cheriet M.(2009). Application on Multi-level Classifier and Clustering for automatic Word spotting in Historical Document Image, In 10th Int`1 conf. on Document Analysis and Recognition, Vol.2, pp. 511-515.

25. Otsu, N.(1979). A Threshold Selection Method from Gray-Level Histograms, IEEE Transactions on Systems, Man, and Cybernetics, Vol. 9, No. 1, pp. 62-66.

26. M. Ziaratban .(2020).Script-Independent Handwritten Text line Segmentation Using Directional 2D Filters. Journal of Soft Computing and Information Technology 9.1 :46-60.

۲۷. یغمایی فرزین. (۱۳۹۶). تشخیص عادت‌های نوشتاری و استفاده از آن در سنتز کلمات در دستخط‌های فارسی، مجله ماشین بینایی و پردازش تصویر سال چهارم، شماره اول.

28. Bhattacharyya, A., Chakraborty, R., Saha, S., Sen, S., Sarkar, R., & Roy, K. (2022). A two-stage deep feature selection method for online handwritten bangla and devanagari basic character recognition. SN Computer Science, 3(4), 260.

Classification and recognition of sub-words in old Persian manuscript documents

Somaye Sabouri[1], Hamidreza Ghaffari[2]
[1] .Ph.D student, Islamic Azad University, Birjand branch, sabouri.somaye@gmail.com
[2] Assistant Professor, Islamic Azad University, Firdous branch,
hamidghaffary53@yahoo.com

Abstract— Historical documents are always of interest to historians and linguists. Important documents are usually digitized by segmentation and identification methods. Digitization of documents is very important for research on these documents and their protection. This article proposes a general classification and recognition framework for the images of Persian historical documents. First, pre-processing is done on documents by removing noises, removing skew, removing stamps, etc., and the document image becomes a two-level image. In the second step, a method of dividing the text of the document into lines is proposed. In the third stage, a method of dividing lines into sub-words of Persian script is presented and the sub-words of these documents are extracted, then deep networks are used to train frequent sub-words and recognize them, and the results are reported based on different criteria.

Keywords: Persian handwritten documents, document segmentation, line segmentation, sub-word segmentation, neural network