بسم الله الرحمن الرحيم

دانشکده مهندسی کامپیوتر و فناوری اطلاعات

رساله دکتری هوش مصنوعی

جستجوی کلمه در اسناد تصویری فارسی مبتنی بر خصیصه‌ها

نگارنده

مجید ایرانپور مبارکه

استاد راهنما

دکتر علیرضا احمدی‌فرد

استاد مشاور

دکتر حمید حسن‌پور

بهمن ۱۳۹۵

دانشکده : مهندسی کامپیوتر و فناوری اطلاعات

گروه : هوش مصنوعی

پایان‌نامه رساله دکتری آقای مجید ایرانپور مبارکه

تحت عنوان: جستجوی کلمه در اسناد تصویری فارسی مبتنی بر خصیصه‌ها

در تاریخ **۱۳۹۵/۱۱/۱۸** توسط کمیته تخصصی زیر جهت اخذ مدرک دکتری ارزیابی گردید و با درجه **بسیار خوب** مورد پذیرش قرار گرفت.

اساتید راهنما	**امضاء**	**اساتید مشاور**	**امضاء**
نام و نام خانوادگی: دکتر علیرضا احمدی‌فرد		نام و نام خانوادگی: دکتر حمید حسن‌پور	

اساتید داور	**امضاء**	**نماینده تحصیلات تکمیلی**	**امضاء**
نام و نام خانوادگی:		نام و نام خانوادگی:	
نام و نام خانوادگی:			
نام و نام خانوادگی:			

سپاس و ستایش مر خدای را جل و جلاله که آثار قدرت او بر چهره روز روشن، تابان است و انوار حکمت او در دل شب تار، درفشان.

آفریدگاری که خویشتن را به ما شناساند و درهای علم را بر ما گشود و عمری و فرصتی عطا فرمود تا بدان، بنده ضعیف خویش را در طریق علم و معرفت بیازماید.

تقدیم به پدرم به استواری کوه،

مادرم به زلالی چشمه

و

همسرم به صمیمیت باران.

سپاس خدای را که سخنوران، در ستودن او بمانند و شمارندگان، شمردن نعمت های او ندانند و کوشندگان، حق او را گزاردن نتوانند. و سلام و دورد بر محمّد و خاندان پاک او، طاهران معصوم، هم آنان که وجودمان وامدار وجودشان است؛ و نفرین پیوسته بر دشمنان ایشان تا روز رستاخیز. بدون شک جایگاه و منزلت معلم، اجّل از آن است که در مقام قدردانی از زحمات بی شائبه ی او، با زبان قاصر و دست ناتوان، چیزی بنگاریم.

اما از آنجایی که تجلیل از معلم، سپاس از انسانی است که هدف و غایت آفرینش را تأمین می کند و سلامت امانت هایی را که به دستش سپرده اند، تضمین؛

بر حسب وظیفه و از باب " من لم یشکر المنعم من المخلوقین لم یشکر اللهَ عزّ و جلّ: "

از پدر و مادر عزیزم، این دو معلم بزرگوارم، که همواره بر کوتاهی و درشتی من، قلم عفو کشیده و کریمانه از کنار غفلت هایم گذشته اند و در تمام عرصه های زندگی یار و یاوری بی چشم داشت برای من بوده اند؛

از استاد با کمالات و شایسته؛ جناب آقای دکتر علیرضا احمدی فرد که در کمال سعه صدر، با حسن خلق و فروتنی، از هیچ کمکی در این عرصه بر من دریغ ننمودند و زحمت راهنمایی این رساله را بر عهده گرفتند؛

از استاد صبور و با تقوا، جناب آقای دکتر حمید حسن پور، مدیریت محترم دانشکده، که زحمت مشاوره این رساله را در حالی متقبل شدند که بدون مساعدت ایشان، این پروژه به نتیجه مطلوب نمی رسید؛

تعهـــد نامه

این‌جانب **مجید ایرانپور مبارکه** دانشجوی دوره دکتری رشته مهندسی کامپیوتر/هوش مصنوعی دانشکده مهندسی کامپیوتر و فناوری اطلاعات دانشگاه صنعتی شاهرود نویسنده‌ی پایان‌نامه **جستجوی کلمه در اسناد تصویری فارسی مبتنی بر خصیصه‌ها** تحت راهنمایی دکتر علیرضا احمدی‌فرد متعهد می‌شوم :

- تحقیقات در این پایان‌نامه توسط این‌جانب انجام شده است و از صحت و اصالت برخوردار است.
- در استفاده از نتایج پژوهش‌های محققان دیگر به مرجع مورداستفاده استناد شده است.
- مطالب مندرج در پایان‌نامه 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 Technologhy » به چاپ خواهد رسید.
- حقوق معنوی تمام افرادی که در به دست آمدن نتایج اصلی پایان‌نامه تأثیرگذار بوده‌اند در مقالات مستخرج از پایان‌نامه رعایت می‌گردد.
- در کلیه مراحل انجام این پایان‌نامه ، در مواردی که از موجود زنده (یا بافتهای آن‌ها) استفاده شده است ضوابط و اصول اخلاقی رعایت شده است.
- در کلیه مراحل انجام این پایان‌نامه، 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

جستجوی کلمه، روشی برای اندیس‌گذاری، جستجو و بازیابی کلمات در اسناد تصویری است. در این مسئله هدف مشخص کردن موقعیت رخدادهای کلمه پرسشی موردنظر کاربر در اسناد تصویری است. این زمینه تحقیقاتی در زمره چالش‌های مهم در حوزه آنالیز اسناد تصویری است که در سال‌های اخیر مورد توجه محققان قرار گرفته است. کارهای متنوعی در این حوزه بر روی اسناد لاتین انجام شده، اما در مورد زبان‌های فارسی و عربی کارهای محدودی انجام شده است که اکثراً در زمینه اسناد چاپی می‌باشد. هدف از این رساله ارائه یک سیستم کامل جهت جستجوی کلمات پرسشی دلخواه از اسناد تصویری فارسی است که وابسته به شیوه نگارش نویسنده نبوده و در اسناد دارای چند نویسنده نیز به خوبی کار می‌کند. علاوه بر این سیستم مذکور توانایی دریافت کلمه پرسشی به‌صورت متنی و تصویری را نیز دارا می‌باشد.

ازآنجایی‌که تعداد کلاس‌ها (کلمات زبان) در این مسئله بسیار زیاد است و از طرف دیگر تنوع درون کلاسی ناشی از شیوه‌های نگارش متفاوت نیز وجود دارد، نیاز به استفاده از رده‌بندی مبتنی بر خصیصه احساس می‌شود. رده‌بندی مبتنی بر خصیصه با تبادل بین کلاسی، توانایی شناسایی کلاس‌های دیده نشده را دارد. در این روش، متن (برچسب کلاس کلمات) و تصویر در یک زیر فضای برداری مشترک تعبیه می‌شوند. این زیر فضا بر اساس چارچوب کاری که از رده‌بندی مبتنی بر خصیصه، تعبیه برچسب و رگرسیون زیرفضای مشترک ایجاد شده است، کار می‌کند. در این زیرفضای ایجادشده، متن و تصویر یک کلمه یکسان، در زیرفضای مشترک آموزش داده شده، نزدیک به هم قرار می‌گیرند و مسئله جستجوی کلمات تبدیل به یک مسئله ساده نزدیک‌ترین همسایه می‌گردد.

در گام اول، این روش جهت جستجو و بازشناسی کلمات دست‌نویس مجزا و اسناد قطعه‌بندی شده به‌کاررفته است. سپس بر اساس ساختار نوشتار زبان فارسی خصیصه‌های متناسب پیشنهاد شده است. دو روش بازنمایی جهت خصیصه‌ها زبان فارسی معرفی‌شده که یکی بر اساس شکل متفاوت حروف در موقعیت‌های مختلف و دیگری بر اساس دسته‌بندی حروف مبتنی بر اشتراک بدنه‌ها، نقاط و سرکش‌ها در حروف فارسی پیشنهاد شده است. به‌منظور ارزیابی روش‌های پیشنهادی از مجموعه داده‌های فارسی که شامل کلمات مجزای دست‌نویس هستند، فارسا و ایرانشهر، استفاده شده است. میانگین دقت متوسط خصیصه‌های پیشنهادی در جستجوی کلمه برای این مجموعه داده‌ها به ترتیب ۹۵/۶۷ و ۹۶/۲۵ درصد و دقت بازشناسی کلمات دست‌نویس به ترتیب ۹۶/۱۵ و ۹۷/۵۵ درصد به‌دست‌آمده است.

در گام دوم، جستجوی سند در سطح خط انجام شد. با استفاده از روشی پیشنهادی که مبتنی بر خوشه‌بندی سلسله مراتبی می‌باشد، خطوط هر سند استخراج می‌شود و سپس پنجره‌گذاری روی خطوط استخراج‌شده صورت

می‌پذیرد. برای فائق آمدن بر مشکلات پنجره گذاری، یک روش پنجره گذاری مبتنی بر اجزاء متصل پیشنهاد شده است. در این روش اندازه پنجره اهمیتی ندارد بلکه تعداد جزء متصلی که پنجره آن را در برمی‌گیرد مهم است. این پنجره‌ها به‌عنوان کلمات مستقل در نظر گرفته شده و بر اساس روش ارائه‌شده مبتنی بر زیرفضای مشترک برای کلمات قطعه‌بندی شده، بازیابی می‌شوند. به‌منظور ارزیابی روش پیشنهادی از مجموعه داده متون دست‌نویس فارسی (FHT)، که مجموعه‌ای از اسناد دست‌نویس فارسی است، استفاده شده است و در مورد کلمه پرسشی متنی و تصویری، میانگین دقت متوسط به‌دست‌آمده به ترتیب برابر ۸۱/۰۲ و ۷۸/۱۴ درصد می‌باشد.

ازآنجایی‌که همیشه قطعه‌بندی کار پرخطا و زمان‌بری است، روش‌های مستقل از قطعه‌بندی از جایگاه ویژه‌ای برخوردارند. ازاین‌رو در ادامه روشی جهت جستجوی کلمات مستقل از قطعه‌بندی پیشنهاد شده است که به‌صورت دومرحله‌ای کار می‌کند. در مرحله اول اجزاء متصل در کلمه پرسشی در سند جستجو می‌شود. سپس بر اساس اولویت اجزاء متصل بازیابی شده از سند و فاصله بین آن‌ها، نواحی کاندید برای کلمه مورد جستجو تعیین می‌شوند. جهت اطمینان نواحی کاندید گسترش داده می‌شوند. در مرحله بعد، رتبه‌بندی این نواحی کاندید بر اساس کلمه پرسشی ورودی انجام می‌پذیرد. این رتبه‌بندی با استفاده از سیستم پنجره‌گذاری استفاده شده در سطح خط صورت می‌پذیرد. جهت ارزیابی روش پیشنهادی در این مرحله نیز از مجموعه داده FHT استفاده شده است و نتایج تجربی به‌دست‌آمده برای کلمات پرسشی تصویری و متنی به ترتیب ۷۵/۰۷ و ۸۰/۲۸ درصد می‌باشد.

نکته حائز اهمیت در ارزیابی روش‌های پیشنهادی مبتنی بر قطعه‌بندی خط و مستقل از قطعه‌بندی این است که آموزش مدل با استفاده از مجموعه داده فارسا و ایرانشهر (کلمات مجزا) انجام شده است و مرحله آزمون با مجموعه دادهFHT انجام پذیرفته است. این کار توانایی روش پیشنهادی در مقاوم بودن نسبت به تغییرات شیوه نگارش را نشان می‌دهد. علاوه بر این هیچ‌کدام از کلاس‌های کلمات مورد جستجو در مرحله آموزش مورد استفاده قرار نگرفته است، بنابراین روش پیشنهادی توانایی جستجوی هر کلمه دلخواه را به کاربر می‌دهد.

کلمات کلیدی: جستجوی کلمه در اسناد تصویری، اسناد دست‌نویس فارسی، تعبیه برچسب، رده‌بندی مبتنی بر خصیصه، اجزاء متصل

مقالات مستخرج از رساله

مقالات ژورنالی

- ✓ Isolated Persian/Arabic word spotting by label embedding, **Journal of Engineering Research** (JER), Decumber 2016, Accepted.
- ✓ Segmentation-free Word Spotting on Persian Multi-writer Handwritten image documents. Ready to submit.
- ✓ Line-based Persian word spotting by label embedding. Ready to submit.

- ✓ ارائه خصیصه‌های خاص زبان فارسی جهت بازیابی و بازشناسی کلمات تصویری فارسی با استفاده از تعبیه برچسب. **مجله ماشین بینایی و پردازش تصویر**، چاپ‌شده.
- ✓ مروری بر روش‌های جستجوی کلمه در اسناد تصویری، **مجله محاسبات نرم**، تحت داوری.

مقالات کنفرانسی

- ✓ ارائه و مقایسه چندین روش تعبیه برچسب جهت بازیابی و بازشناسی کلمات در اسناد تصویری دست‌نویس فارسی (**مقاله برتر کنفرانس**) نهمین کنفرانس ماشین بینایی و پردازش تصویر ایران، ۱۳۹۴
- ✓ جداسازی خطوط متن و استخراج برون‌خط دنباله‌ای از مؤلفه‌های متصل در دست‌نوشته فارسی، نهمین کنفرانس ماشین بینایی و پردازش تصویر ایران، ۱۳۹۴
- ✓ مکان‌یابی متون فارسی در تصاویر صحنه طبیعی با استفاده امضای رنگ لبه‌های تصویر، هفتمین کنفرانس بین‌المللی فناوری اطلاعات و دانش، ۱۳۹۴
- ✓ جستجوی کلمات در یک تصویر دست‌نوشته فارسی با استفاده از روش تعبیه مشخصه‌ها، سومین کنفرانس ملی پژوهش‌های کاربردی در مهندسی کامپیوتر، ۱۳۹۴
- ✓ استخراج خط در اسناد دست‌نویس فارسی مبتنی بر خوشه‌بندی سلسله مراتبی، هشتمین کنفرانس بین‌المللی فناوری اطلاعات و دانش، ۱۳۹۵

فهرست مطالب

فهرست شکل‌ها

فهرست جدول‌ها

واژه‌نامه

اندیس گذاری معنایی پنهان	Latent semantic indexing (LSI)
آنالیز تفکیکی منظم	Regularized discriminant analysis (RDA)
بازنمایی مشترک	Joint representation
بازیابی محتوایی تصویر	Content-based image retrieval (CBIR)
بخشی کلمه عربی	Part of arabic word (PAW)
بردار فیشر	Fisher vector (FV)
پیچش زمان پویا	Dymanic time wraping (DTW)
تابع سازگاری	Compatibility function
تحلیل همبستگی کانونی	Canonical correlation analysis (CCA)
تحلیل همبستگی کانونی دارای هسته	Kernelized canonical correlation analysis (KCCA)
تخمین حداکثر درستنمایی	Maximum likelihood estimation (MLE)
تشخیص کاراکتر نوری	Optical character recognition (OCR)
تشخیص و تحلیل اسناد تصویری	Document image analysis and recognition (DIAR)
تطبیق انعطاف‌پذیر توالی	Flexible sequence matching (FSM)
تعبیه برچسب	Label embedding
تعبیه ورودی	Input embedding
جستجوی کلمه	Word spotting
خارج از فرهنگ لغت	Out of vocabulary (OOV)
خصیصه	Attribute
دقت	Precision
رده‌بندی مبتنی بر خصیصه	Attribute-based classification
رگرسیون زیرفضای مشترک	Common subspace regression (CSR)
سبد بصری کلمات	Bag-of-visual-word (BoVW)
سیفت	Scale invariant feature transform (SIFT)
شبکه‌های عصبی بازگشتی	Recurrent neural network (RNN)
شبکه عصبی با حافظه کوتاه‌مدت بلند دوجهته	Bidirectional long short-term memory (BLSTM)
فراخوانی	recall

کتابچه کد	Codebook
کدگذاری شکل کاراکتر	Character shape coding(CSC)
کدگذاری شکل کلمه	Word shape coding (WSC)
کلمه پرسشی تصویری	Query by example (QBE)
کلمه پرسشی متنی	Query by string (QBS)
کمینه متوسط فاصله افقی	Min average horizontal distance (MAHD)
کمینه متوسط فاصله عمودی	Min average vertical distance (MAVD)
ماشین بردار پشتیبان مبتنی بر یک نمونه	Exemplar SVM
مجموعه درستی	Ground Truth
مدل تمایزی	Discriminative model
مدل زایشی	Generative model
مدل مخفی مارکوف	Hidden markov model (HMM)
مدل مخفی مارکوف شبه گسسته	Semi-continuous hidden markov model (SC-HMM)
مدل مخلوط گاوسی	Gussian mixture model (GMM)
مقیاس‌بندی پلاتس	Platts Scaling
میانگین دقت متوسط	Mean average precision (Map)
میزان کاهش تعداد اجزاء متصل	Connected components reduction rate (CCRR)
نرخ تشخیص	Detection rate (DR)
نرخ تطابق پیکسلی	Pixel level hit rate (PLHR)
هرم مکانی کاراکترها	Spatial pyramid of characters (SPOC)
هیستوگرام گرادیان	Histogram of gradient (HOG)
هیستوگرام هرمی کاراکترهای فارسی	Persian pyramid histogram of characters (PPHOC)
هیستوگرام هرمی کاراکترهای موقعیتی فارسی	Persian pyramid histogram of positional characters (PPHOPC)
هیستوگرام هرمی نشانه‌های موقعیتی فارسی	Persian pyramid histogram of positional signs (PPHOPS)
وصله	Patch
یادگیری شات صفر	Zero-shot learning

فصل

۱- مقدمه

۱-۱- تعریف مسئله

از گذشته‌های بسیار دور متن به‌عنوان یکی از منابع مهم اطلاعاتی مورد استفاده قرار گرفته است و امروزه بسیاری از این نوع داده در کتابخانه‌های دیجیتال، اسناد، تصاویر و ویدئوها وجود دارد. پیشرفت فناوری دیجیتال باعث شده است که اسناد و کتاب‌ها اسکن شده و به‌صورت تصویر ذخیره شوند. اسناد وقتی‌که به‌صورت تصویری درمی‌آیند، ذخیره‌سازی و انتقال ساده‌تری خواهند داشت. بنابراین نگهداری اسناد و کتاب‌ها به‌صورت اسناد تصویری موردتوجه بسیاری قرار گرفته است. جهت پردازش و استفاده از محتوای متنی اسناد تصویری، روش‌های تشخیص و تحلیل اسناد تصویری[۱] مورد توجه قرار گرفته و توسعه‌یافته‌اند. جستجو و بازیابی اطلاعات همواره به‌عنوان یکی از مهم‌ترین شاخه‌های تحلیل داده، مورد توجه محققین بوده است. ازاین‌رو جستجو در اسناد تصویری نیز اهمیت خاصی را در حوزه تحلیل اسناد دارد. این مسئله در حوزه بازیابی محتوایی[۲] تصاویر قرار می‌گیرد.

رهیافت اولیه جهت جستجو و بازیابی اطلاعات متنی از اسناد تصویری، استفاده از روش‌های تشخیص کاراکتر نوری[۳] جهت تبدیل اسناد تصویری به متن و سپس بکار بردن روش‌های بازیابی متن است [۱]. روش‌های تشخیص کاراکتر نوری فعلی در مورد فونت‌های ساده با پس‌زمینه تمیز عملکرد قابل قبولی دارد، اما کارایی این روش با کاهش کیفیت اسناد یا در مورد اسناد دست‌نویس به شدت کاهش می‌یابد[۲]. علاوه بر این، روش مذکور در مورد پایگاه داده‌ها با تعداد زیادی سند تصویری، بسیار زمان‌بر است [۳].

بنابراین، محققین به دنبال روش‌های بازیابی بوده‌اند که توانایی جستجوی اطلاعات متنی بدون نیاز آنالیز طرح‌بندی[۴] و تبدیل کل سند تصویری به متن را داشته باشد. ازاین‌رو روش‌هایی که قادر به بازیابی تصاویر حاوی یک واژه مشخص باشند یا بتوانند لغات را در یک تصویر تشخیص دهند مورد توجه قرار گرفته‌اند. ایده جستجوی کلمات[۵]، که در برخی از مقالات جستجوی کلمات کلیدی[۶] نیز خوانده می‌شود، در زمینه پردازش گفتار ارائه شده است[۴]. سپس این روش برای اندیس‌گذاری اسناد چاپی و دست‌نویس به کار گرفته شده است. این رهیافت موقعیت کلمه انتخاب شده توسط کاربر را در یک سند تصویری مشخص می‌کند. به‌عبارت‌دیگر، سیستم جستجو کلمه، کلمه پرسشی موردنظر کاربر را در سند تصویری جستجو می‌کند. جستجوی کلمه، روشی برای اندیس‌گذاری، جستجو و بازیابی کلمات در اسناد تصویری است. در این مسئله هدف مشخص کردن موقعیت رخدادهای کلمه پرسشی موردنظر کاربر در اسناد

---

[1] Document image Analysis and Recognition (DIAR)
[۲] Content-Based Image Retrieval (CBIR)
[3] Optical Character Recognition (OCR)
[۴] Layout analysis
[5] Word spotting
[6] Keyword spotting

تصویری است. در این سیستم کاربر کلمه پرسشی[1] را از سند انتخاب کرده و به سیستم ارائه می‌کند. وظیفه سیستم یافتن کلمات مشابه در سند مورد جستجو می‌باشد. یکی از مسائل مهمی که برای سیستم‌های تشخیص کاراکتر نوری چالش مهمی محسوب می‌شود، مشکلات تبدیل طرح‌بندی صفحات سند است. تصاویر، گرافیک و تشخیص فونت و اندازه آن نیز به‌عنوان چالشی برای تبدیل تصویر به متن مشابه وجود دارد. استفاده از این روش باعث می‌شود که طرح‌بندی صفحات حفظ شود و بنابراین چالش‌های ذکرشده وجود نداشته باشند[۵].

یکی از کاربردهای مطرح این روش، جستجوی کلمه در اسناد تاریخی تصویری می‌باشد؛ به طور مثال جستجوی شهر یا ناحیه خاصی در نقشه‌های اسکن شده قدیمی، یا جستجوی تصاویر با کلمه پرسشی "Figure" می‌تواند از کاربردهای مهم این روش ذکر شوند. علاوه بر این، جستجوی کلمه ابزار ارزشمندی برای کاربران جهت جستجو روی محتوای کتب قدیمی و دست‌نویس و بازیابی نامه‌های دست‌نویس خواهد بود.

۱-۲- اهداف و چالش‌ها

زبان فارسی یا پارسی، از دسته زبان‌های هندواروپایی محسوب می‌شود که در کشورهای ایران، افغانستان، تاجیکستان و ازبکستان به آن سخن گفته می‌شود. بیش از ۱۵۰ میلیون نفر از این زبان برای نوشتن استفاده می‌کنند. یکی از مهم‌ترین مشخصه‌های زبان فارسی ماهیت پیوستگی آن هم در متون تایپی و هم در متون دست‌نویس است، که باعث شده دقت تشخیص در متون تایپی فارسی کمتر از متون تایپی لاتین باشد. الفبای فارسی دارای ۳۲ حرف است که هرکدام از آن‌ها بسته به موقعیتش در کلمه ۲ تا ۴ شکل دارد. تمام متون فارسی از راست به چپ نوشته می‌شوند.

جدول ۱-۱ شکل‌های مختلف حروف الفبای زبان فارسی را نشان می‌داد که تفاوت ظاهری حروف بر اساس موقعیت مکانی در کلمه (ابتدا، وسط، انتها و منفرد) را نشان می‌دهد. مثلاً حرف - غ- که غین نامیده می‌شود دارای چهار شکل است. شکل آن در ابتدای کلمه - غ - و در وسط کلمه به‌صورت - غ - و در انتها به‌صورت- غ - می‌باشد. بعضی از حروف تنها دارای دو شکل متفاوت می‌باشند. مانند حرف - د - که دال نامیده می‌شود. در ابتدای کلمه به‌صورت - د – و در وسط و انتهای کلمه به‌صورت - ـد- می‌باشند.

---

[1] Query

جدول ۱-۱. شکل‌های مختلف حروف الفبای فارسی با توجه به محل قرار گرفتن در کلمه

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

تعدادی حروف در الفبای فارسی وجود دارند که از سمت چپ قابل اتصال نیست و اگر در ابتدا و وسط کلمه قرار بگیرند باعث جدا شدن کلمه و تشکیل زیر کلمات می‌شود. این مشخصه زبان فارسی باعث شده که فاز جداسازی در سیستم‌های مبتنی بر قطعه‌بندی با مشکل روبه‌رو شود، به این دلیل که فاصله‌گذاری مناسب بین کلمات و زیر کلمات در بیشتر مواقع رعایت نمی‌شود. این حروف شامل (ر، ز، ژ، د، ذ، و، ا) می‌باشد. مثلاً در کلمه "مدرسه" وجود دو حرف "د" و "ر" در کلمه باعث شده که کلمه "مدرسه" شامل سه زیر کلمه شود.

در زبان عربی و فارسی کلماتی وجود دارد که شکل یکسانی دارند اما تلفظ و معنی متفاوتی دارند، برای نشان دادن تلفظ و معنی موردنظر از یک سری علائم استفاده می‌شود. مثلاً کلمه فارسی "مرد" دارای دو تلفظ است که معنی‌های متفاوتی دارند: "مُرد" به معنی از دنیا رفتن و " مَرد" که به معنی آقا می‌باشد. البته در فارسی، در بیشتر مواقعی که در متن آورده می‌شود معنی آن با توجه به متن مشخص می‌شود. از دیدگاه تشخیص کلمه، وجود این علائم دقت تشخیص را کاهش می‌دهد.

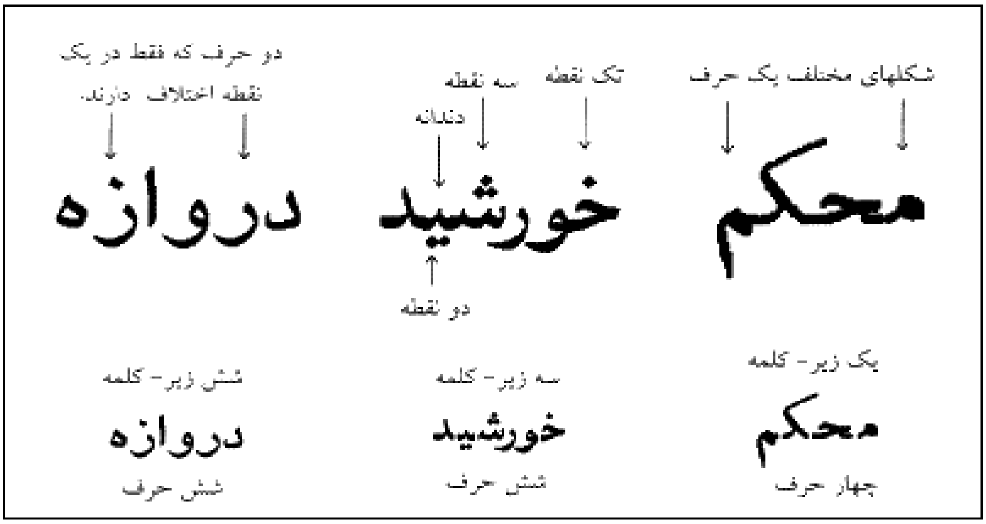
یکی دیگر از مشخصات زبان فارسی وجود نقطه در بعضی حروف است. ۱۸ حرف از ۳۲ حرف زبان فارسی شامل یک یا دو یا سه نقطه در بالا یا پایین خط اصلی می‌باشند:

- ده حرف فارسی دارای یک نقطه می‌باشند (ب، ج، خ، ذ، ز، ض، ظ، غ، ف، ن)
- سه حرف شامل دو نقطه می‌باشند (ی ق، ت)
- پنج حرف شامل سه نقطه می‌باشند (پ، چ ، ژ، ش، ث)

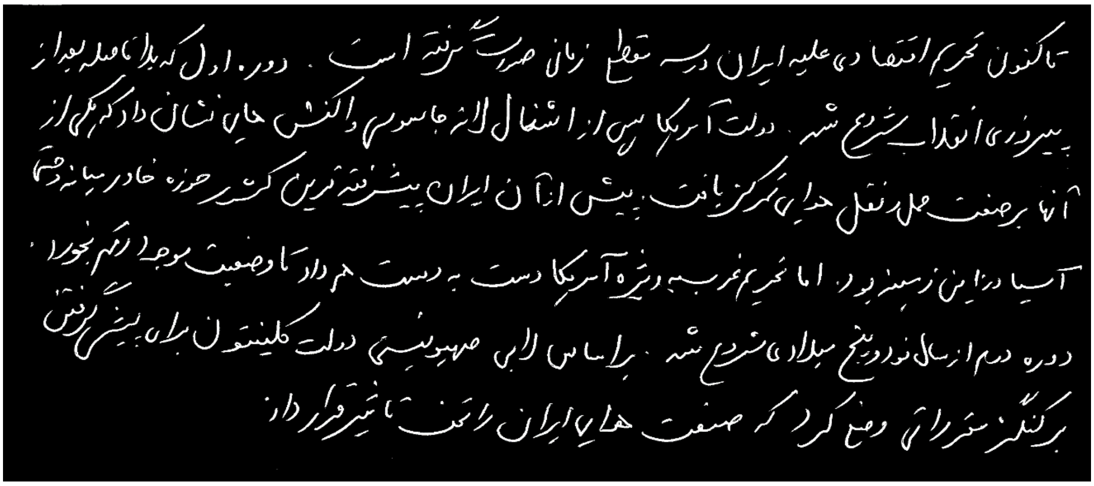
علاوه بر این در فارسی برخی از حروف دارای مکمل‌هایی به غیر از نقطه مانند سرکش حروف (ک وگ)یا آکلاد حرف (آ) می‌باشند.

وجود نقاط و علائم در کلمات، خصوصاً در اسناد دست‌نوشته، باعث می‌شود عمل تشخیص مشکل‌تر شود. دلیل آن را می‌توان این‌گونه بیان کرد که در متون دست‌نوشته اندازه و نحوه نوشتن نقاط، مکمل‌ها و علائم بسیار متغیر است و ممکن است بعضی از آن‌ها آن‌قدر کوچک باشد که در فاز پیش‌پردازش حذف شوند یا آن‌قدر بزرگ باشند که با بعضی حروف اشتباه گرفته شوند. مثال‌هایی که در شکل ۱-۱ و ۱-۲ آورده شده‌اند، برخی ویژگی‌های زبان فارسی را به ترتیب در سطح کلمه و سند نشان می‌دهند.

بیشتر تحقیقات انجام شده در این زمینه بر روی متون زبان لاتین ارائه شده است و تعدادی از آن‌ها در مورد زبان‌های دیگری مثل چینی، هندی و کره‌ای می‌باشند. تحقیقات انجام شده در زبان فارسی و عربی بسیار کم هستند که بیشتر آن‌ها نیز بر روی اسناد چاپی انجام شده است. به همین دلیل با توجه به اهمیت و کاربرد مسئله و اینکه طبق بررسی‌های صورت گرفته کار قابل‌توجهی در زبان فارسی انجام نشده است، نیاز به انجام تحقیق در این زمینه احساس می‌شود.



شکل ۱-۱. برخی ویژگی‌های زبان فارسی[۶]



شکل ۱-۲. نمونه‌ای از اسناد دست‌نویس فارسی

ازآنجایی‌که نوشتارهای زبان‌های مختلف خواص متفاوتی دارند، در نظر گرفتن ساختار نوشتاری زبان در روش‌های جستجوی کلمه به‌عنوان یک نکته کلیدی می‌تواند در نظر گرفته شود. ساختار نوشتاری زبان فارسی پیچیدگی‌های بیشتری نسبت به انگلیسی در بسیاری از کارهای آنالیز اسناد مانند جستجوی کلمه ایجاد می‌کند. برخی از مهم‌ترین این ویژگی را می‌توان به‌صورت زیر ذکر کرد:

۱. ذات نوشتار فارسی پیوسته است.
۲. کلمه‌ها از مجموعه‌ای از یک یا چند زیر کلمه[۱] تشکیل می‌شوند.
۳. هر زیر کلمه می‌تواند یک حرف یا مجموعه‌ای از حروف متصل باشد.
۴. هر زیر کلمه می‌تواند از یک یا چندین جزء متصل ساخته شود.

---

[۱] Subword در برخی از مقالات به جای "زیر کلمه" از "بخشی از کلمه فارسی" (Part of Persian Word (PPW استفاده شده است.

۵. تعداد زیادی از حروف از چندین جزء متصل، شامل بدنه حرف و نقاط یا سرکش‌ها، ساخته شده‌اند.
۶. شکل اکثر حروف بر اساس موقعیت آن در کلمه تغییر می‌کند.
۷. حدود کلمات در نوشتار فارسی، خصوصاً در مورد اسناد دست‌نویس، به راحتی قابل‌تشخیص نیست.

بر این اساس چالش‌های زیاد و متفاوتی در حوزه آنالیز اسناد در زبان فارسی وجود دارد که آن را متمایز با زبان‌های دیگری چون انگلیسی می‌کند. بنابراین با توجه به اینکه بر اساس تحقیقات صورت گرفته، پژوهش چشمگیری در زمینه جستجوی کلمه در اسناد دست‌نویس فارسی انجام نشده است، نیاز به انجام تحقیق در این زمینه احساس می‌شود.

به‌طورکلی سیستم‌هایی که روی دست‌نوشته کار می‌کنند حساسیت زیادی روی تنوع شیوه‌های نگارش (دستخط) دارند. ازاین‌رو بسیاری از کارهای اولیه انجام شده در زمینه جستجوی کلمه تک نویسنده است. روش‌هایی که اخیراً معرفی شده‌اند، سعی در فائق آمدن بر مسئله پیچیدگی و تنوع شیوه‌های نگارش داشته‌اند.

روش‌هایی مانند مدل مخفی مارکوف[۱] و پیچش زمان پویا[۲] در جستجوی کلمه در اسناد تصویری بسیار مورداستفاده قرار گرفته‌اند. اما این روش‌ها در سیستم‌هایی استفاده شده است که کلمه پرسشی به‌صورت تصویری به سیستم داده می‌شود. یکی از چالش‌های دیگر مربوط به کلمه پرسشی تصویری است. در کاربرد واقعی یافتن و ارائه کلمه پرسشی تصویری موردنظر کاربر کاری مشکل و زمان‌بر خواهد بود. اما اگر کلمه پرسشی به‌صورت متنی به سیستم جستجوی کلمه ارائه شود، مشکلات مربوط به آماده‌سازی کلمه پرسشی وجود نخواهد داشت و ارائه کلمه پرسشی به سیستم ساده خواهد شد، اما برای این منظور سیستم پیچیده‌تری باید طراحی گردد.

با در نظر گرفتن تمامی موارد ذکرشده در بالا، در این رساله در گام اول یک سیستم مبتنی بر یادگیری استفاده شده است که بر اساس رده‌بندی مبتنی بر خصیصه[3]، تعبیه برچسب[۴] و رگرسیون زیرفضای مشترک[5] کار می‌کند [۷]. در این رساله خصیصه‌هایی[۶] مبتنی بر شیوه نگارش در زبان فارسی پیشنهاد شده است که بهتر کار می‌کند و امکان به اشتراک‌گذاری اطلاعات بین کلاسی را افزایش می‌دهد. علاوه

---

[۱] Hidden Markov Model (HMM)

[۲] Dymanic Time Wraping (DTW)

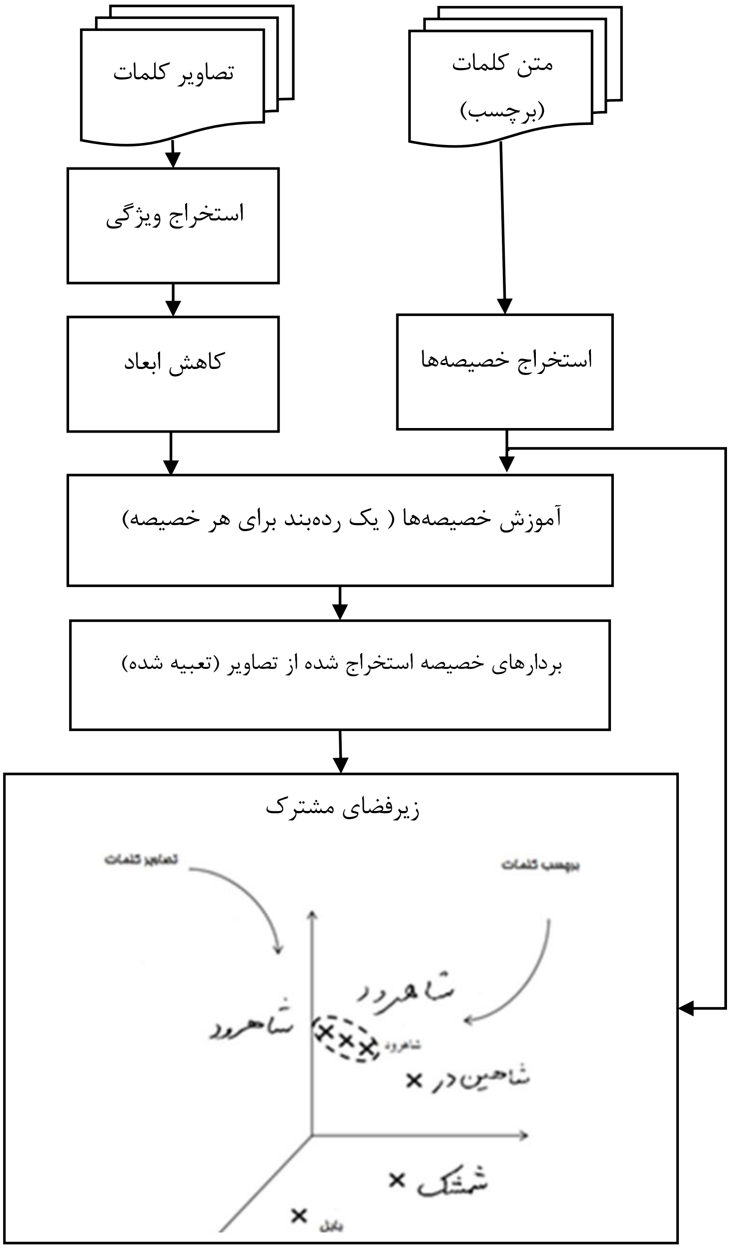
[3] Attribute-based classification

[4] Label embedding

[5] Common Subspace Regression (CSR)

[6] Attributes

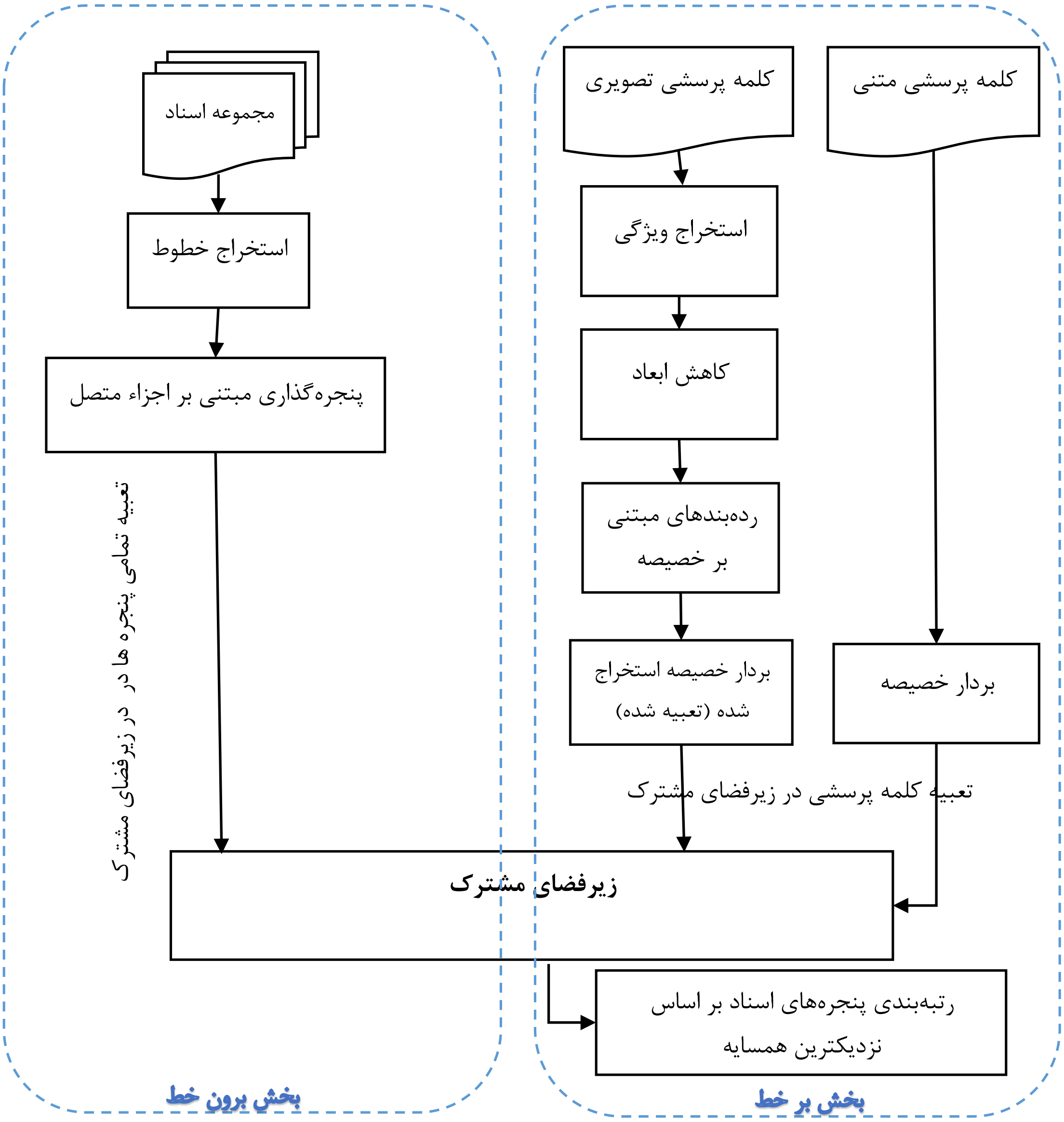
بر این خصیصه‌های پیشنهادی امکان جستجوی زیرکلمات و اجزاء متصل مربوط به کلمات را نیز فراهم می‌کند. خصیصه‌های پیشنهادی و زیرفضای مشترک آموزش داده شده و برای جستجوی کلمات مبتنی بر قطعه‌بندی مورد استفاده قرار گرفته است. شکل ۱-۲ ساختار کلی روش مذکور را در فاز آموزش نشان می‌دهد.



شکل ۱-۳. فلوچارت کلی آموزش روش پیشنهادی مبتنی بر قطعه‌بندی

سپس برای فائق آمدن بر خطای زیاد مربوط به قطعه‌بندی کلمات، که در فارسی به دلیل مشخص نبودن محدوده کلمات در کارهای آنالیز اسناد متنی مشکلات زیادی را ایجاد می‌کند، یک روش مبتنی بر

قطعه‌بندی در سطح خط پیشنهاد شده است که از اجزاء متصل موجود در هر خط به‌عنوان معیاری جهت پنجره‌گذاری استفاده شده است (شکل ۱-۴).



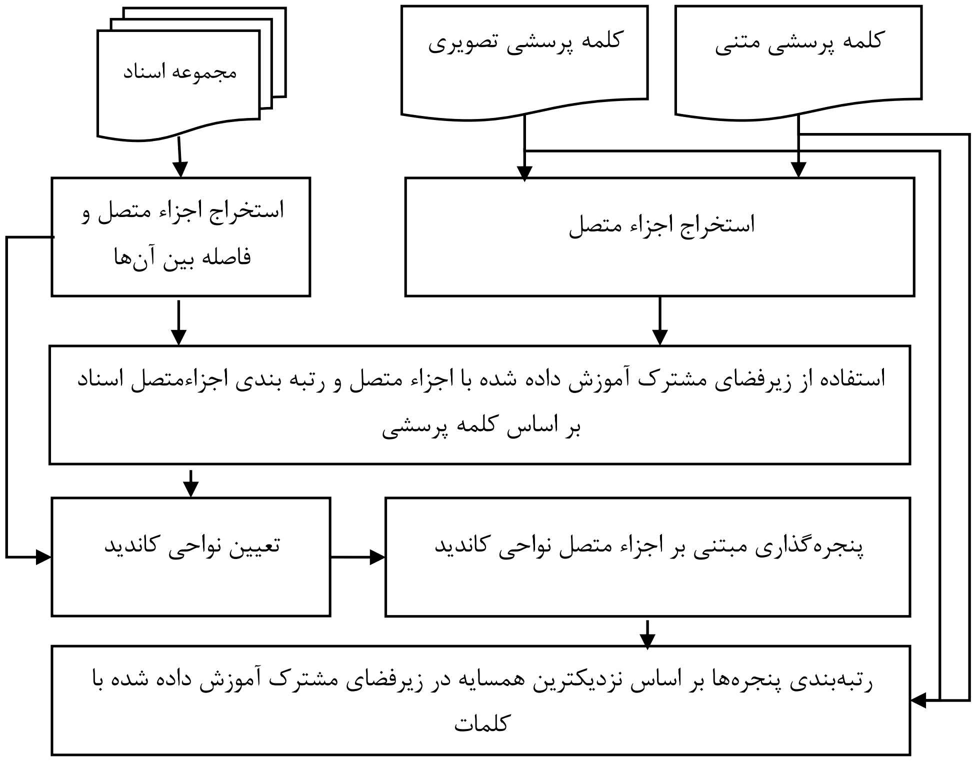
شکل ۱-۴. فلوچارت کلی روش پیشنهادی مبتنی بر قطعه‌بندی در سطح خطوط

در گام بعدی نیز یک چهارچوب کاری سلسله مراتبی با استفاده از قابلیت‌های خصیصه‌های پیشنهادی ارائه شده است که بر اساس جستجوی اجزاء متصل سازنده کلمه پرسشی ورودی نواحی کاندید را تعیین می‌کند. در این روش نواحی کاندید با استفاده از سیستم جستجوی کلمات مبتنی بر قطعه‌بندی رتبه‌بندی می‌شوند. ازآنجایی‌که در روش مبتنی بر خط تعداد مثبت اشتباه[1]زیاد است و کاهش تعداد آن به‌طور معناداری باعث افزایش کارایی سیستم جستجوی کلمه می‌گردد، این روش جدید که تعداد

---

[1] False positive

کاندیدها را کاهش می‌دهد پیشنهاد شده است. علاوه بر این، روش مذکور به قطعه‌بندی سند نیز نیاز ندارد (شکل ۱-۵).

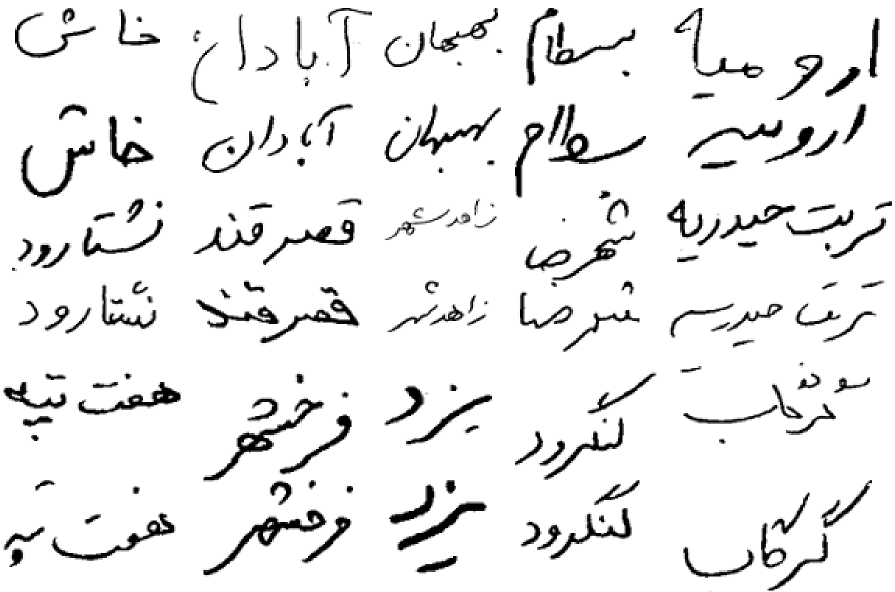


شکل ۱-۵. فلوچارت کلی روش پیشنهادی مستقل از قطعه‌بندی

۱-۳- مجموعه داده‌ها

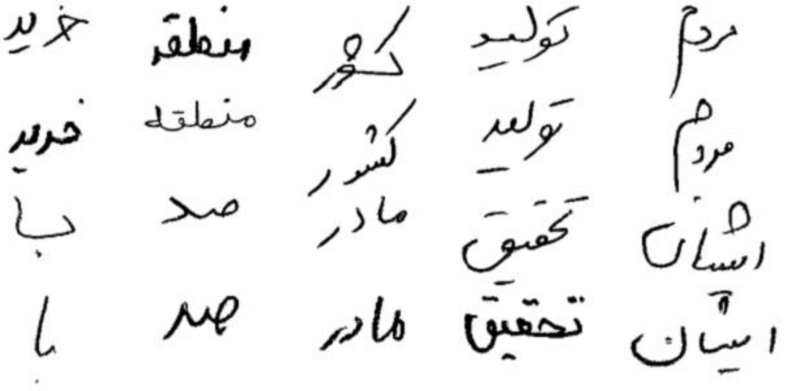
بررسی روش پیشنهادی بر اساس نتایج حاصل‌شده بر روی مجموعه داده‌های استاندارد همواره به‌عنوان یکی از روش‌های ارزیابی مهم به‌حساب می‌آید. در روش‌هایی که در این رساله پیشنهاد شده است از ۳ مجموعه داده استاندارد زبان فارسی استفاده شده است که در زیر معرفی می‌شود.

مجموعه داده ایرانشهر: اسامی ۵۰۳ شهر ایران در یک مجموعه‌ای بیش از ۱۷۰۰۰نمونه‌ای را فراهم آورده است. در هر کلاس از کلمات تصویری شامل نام شهرهای ایران حداقل ۲۵ نمونه وجود دارد. شکل ۱-۶ نمونه‌هایی از کلمات موجود در این مجموعه داده را نشان می‌دهد.



شکل ۱-۶. چند نمونه از تصویر کلمات مجموعه داده ایرانشهر

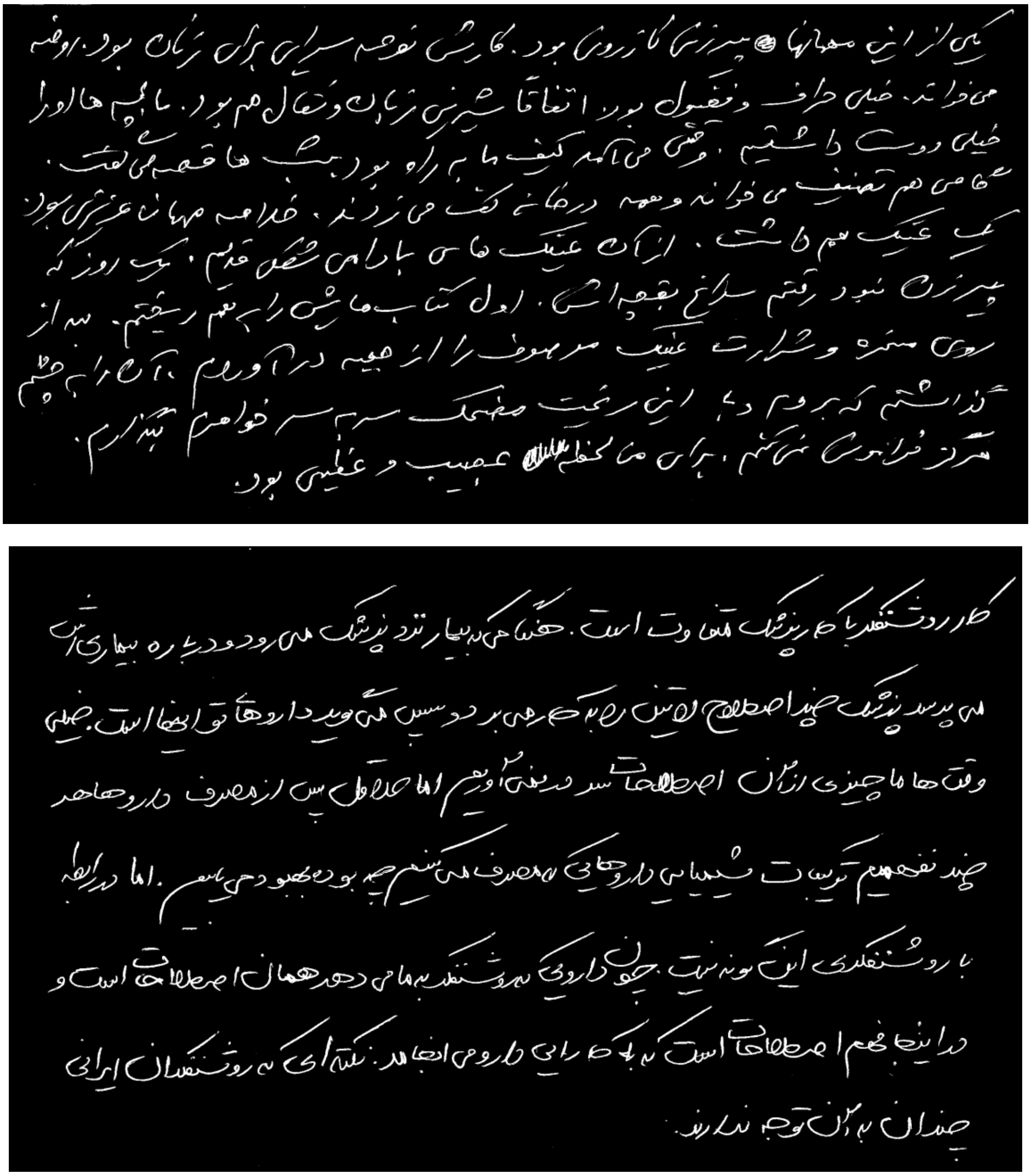
مجموعه داده فارسا[۸]: مجموعه داده فارسا که در دانشگاه صنعتی شاهرود جمع آوری شده است مجموعه‌ای از ۳۰۰ کلمه پرکاربرد فارسی است که توسط ۱۰۰ نویسنده مختلف به‌صورت دست‌نویس نوشته شده است که مجموعه‌ای مشتمل بر ۳۰۰۰۰ تصویر کلمه دست‌نوشته می‌باشد. نمونه‌هایی از برخی از کلمات موجود در این مجموعه داده در شکل ۱-۷ نمایش داده شده است.



شکل ۱-۷. چند نمونه از تصویر کلمات مجموعه داده فارسا

مجموعه داده اسناد دست‌نویس فارسی (FHT) [۹]: این پایگاه داده شامل ۱۰۰۰ سند می‌باشد که دارای ۴۰ دسته متن متفاوت است که توسط ۲۵ نویسنده با سطح سواد متفاوت نوشته شده‌اند. درواقع از هر فرد تقاضا شده است ۴ متن متفاوت را بنویسد که درمجموع این پایگاه داده، ۲۵۰ نویسنده دارد. هر متن به‌طور میانگین دارای ۶/۴۵ خط، ۱۰۶/۶ کلمه، ۲۳۰/۱۷۵ زیر کلمه، ۱۴۰۶ حرف و ۱۳۲/۱

نقطه است. یکی از مزیت‌های این پایگاه داده وجود خطوط صحیح[1] برای هر خط می‌باشد. این اسناد هم شامل دستخط خوانا و خوش‌خط و هم غیر خوانا و بد خط است. شکل ۱-۸ دو متن نمونه از این مجموعه داده را نمایش می‌دهد.



شکل ۱-۸. چند نمونه از تصویر کلمات مجموعه داده FHT

---

[1] Ground Truth

۱-۴- ساختار پایان‌نامه

در فصل بعدی کارهای انجام شده در زمینه جستجوی کلمات مورد بررسی قرار گرفته است. در این فصل انواع دسته‌بندی‌های ارائه شده از منظرهای مختلف آورده شده است. فصل سوم روش پیشنهادی جهت جستجوی کلمات فارسی به‌صورت قطعه‌بندی شده، در سطح کلمه ارائه شده است. این فصل شامل پیشنهاد مدل اصلی استفاده‌شده در این رساله و خصیصه‌های پیشنهادی می‌باشد. ضمناً در انتهای فصل ارزیابی و نتایج تجربی آورده شده است. ازآنجایی‌که از این روش جهت بازشناسی دست‌نوشته‌های فارسی نیز می‌توان بهره برد، مقایسه‌ای نیز با بهترین کارهای انجام شده در این حوزه آورده شده است. جستجوی کلمه در روشی مبتنی بر قطعه‌بندی در سطح خطوط در فصل چهارم ارائه شده است. این بخش شامل روش پیشنهادی جهت استخراج خطوط و پنجره‌گذاری مبتنی بر اجزاء متصل می‌باشد و مدل یادگیری آن مدل پایه‌ی پیشنهادی است. در فصل پنجم یک روش مستقل از قطعه‌بندی پیشنهاد می‌شود. این روش چهارچوب کاری متفاوتی را بر اساس سیستم پایه معرفی می‌کند که با تعیین نواحی کاندید فضای جستجو را به‌طور چشم‌گیری کاهش می‌دهد. فصول چهارم و پنجم نیز شامل ارزیابی و نتایج تجربی سیستم معرفی شده در آن فصل هستند. نهایتاً فصل ششم به نتیجه‌گیری و ارائه پیشنهادات اختصاص داده شده است.

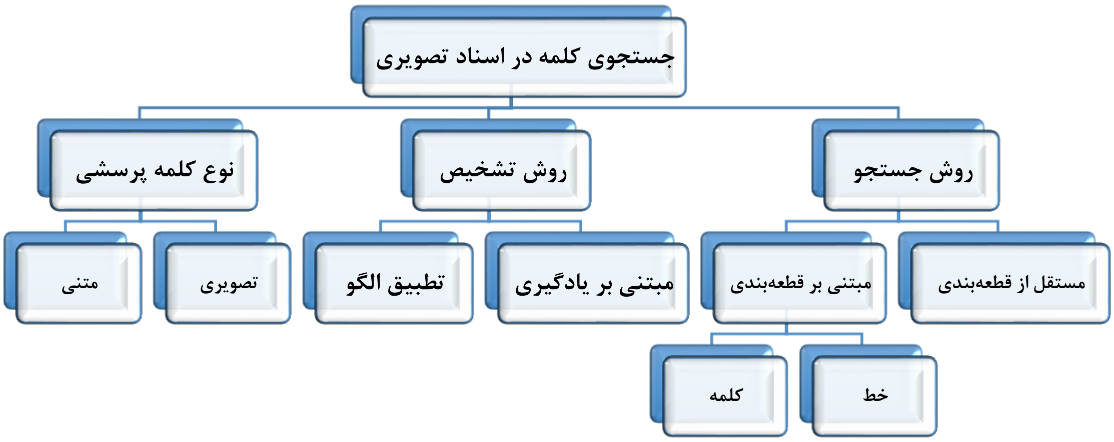
.

فصل

۲- مروری بر کارهای پیشین

۲-۱- مقدمه

روش‌های جستجوی کلمه یک الگوی ساختاری یکسان را دنبال می‌کنند؛ در این الگو ابتدا اسناد در یک فاز برون‌خط اندیس‌گذاری و ذخیره می‌شوند، سپس کاربر یک پرسش را مطرح می‌کند و سیستم شباهت آن را با کلمات اسناد ارزیابی می‌کند. به‌طورکلی این سیستم‌ها، کلمه موردنظر کاربر را در سند تصویری جستجو می‌کنند. با توجه به تنوع کارهای انجام شده روی اسناد، روش‌های موجود در جستجو کلمه را می‌توان از چند دیدگاه تقسیم‌بندی نمود (شکل ۲-۱).



شکل ۲-۱: دسته‌بندی روش‌های موجود جستجوی کلمه

در این دسته‌بندی روش‌های موجود از سه منظر متفاوت دسته‌بندی شده‌اند: اولین مورد روش‌ها را بر اساس نوع کلمه پرسشی دسته‌بندی می‌کند. روش‌های موجود کلمه پرسشی را به‌صورت تصویری[1]، متنی[2] یا به هر دو صورت می‌پذیرند. در اکثر روش‌هایی که کلمه پرسشی صرفاً تصویری وارد می‌شود، کلمه موردنظر از سند بریده می‌شود و به‌عنوان کلمه پرسشی مورداستفاده قرار می‌گیرد، اما روش‌های جدید از رشته متنی ورودی به‌عنوان کلمه پرسشی بهره می‌برند. دسته‌بندی دیگر بر اساس روش جستجو می‌باشد. روش‌هایی وجود دارند که ابتدا سند را به خطوط و سپس به کلمات سازنده قطعه‌بندی کرده و سپس عمل جستجو را انجام می‌دهند، این دسته روش‌ها مبتنی بر قطعه‌بندی در سطح کلمه نامیده می‌شوند. برخی دیگر قطعه‌بندی را فقط در سطح خطوط انجام می‌دهند. برای پیشگیری از وقوع خطای قطعه‌بندی، روش‌های مستقل از قطعه‌بندی نیز ارائه شده است. آخرین دسته‌بندی بر اساس روش تشخیص صورت گرفته است. عمل تشخیص می‌تواند مبتنی بر یادگیری یا بر اساس روش‌های تطبیق الگو باشد. روش‌های تطبیق الگو معمولاً یا عمل تطبیق را روی داده‌های خام و در سطح پیکسل و یا بر اساس ویژگی‌های استخراج‌شده از تصاویر انجام می‌دهند.

---

[1] Query By Example (QBE)

[2] Query By String (QBS)

۲-۲- دسته‌بندی بر اساس روش جستجو

در زمینه جستجوی کلمات، روش‌های موجود ممکن است سند مورد جستجو را در سطح خط یا کلمه قطعه‌بندی کنند یا بدون قطعه‌بندی کل سند را مورد جستجو قرار دهند. درصورتی‌که نیاز به قطعه‌بندی کلمه وجود داشته باشد، ابتدا باید خطوط و سپس کلمات هر خط از سند جدا شود و در نوع دیگر فقط خطوط استخراج می‌شوند. به‌طورکلی قطعه‌بندی از دسته مراحل پیش‌پردازشی است که میزان خطای بالایی دارد و این خطا تأثیر زیادی بر روی عملکرد مراحل بعدی خواهد داشت. در آخرین مورد هیچ عمل قطعه‌بندی انجام نمی‌شود و از بروز این خطا پیشگیری می‌کند، اما چون نیاز به بکار بردن روش‌هایی چون پنجره لغزان وجود دارد پردازش پرهزینه خواهد بود.

۲-۲-۱- روش‌های مبتنی بر قطعه‌بندی

یکی از پیش‌پردازش‌هایی که در بسیاری از روش‌های پردازش اسناد تصویری انجام می‌شود عمل قطعه‌بندی است. این عمل می‌تواند در سطوح مختلف از قبیل خط، کلمه یا حرف انجام شود. نکته قابل‌ذکر در مورد قطعه‌بندی این است که، این مسئله تا حدود زیادی وابسته به زبان است. در برخی از زبان‌ها با توجه به حروف و ساختار نگارش آن‌ها قطعه‌بندی ممکن است ساده‌تر باشد، ازاین‌رو در بسیاری از کارهایی که روی زبان‌هایی مانند چینی و برخی کارهای انگلیسی صورت گرفته است، قطعه‌بندی در سطح حروف دیده می‌شود. در این روش‌های پایه اصلی روش مبتنی بر کدگذاری شکل کاراکتر[1] است. به این صورت که پس از استخراج کاراکترها، با مجموعه کدهای از پیش تعریف‌شده آن‌ها را تفسیر می‌کنند. محدودیت اصلی این روش‌ها حساسیت به خطای قطعه‌بندی است. به علت مشکلات دستگاه‌های تصویربرداری و گسسته سازی و یا در اثر کیفیت پایین اسناد و یا در اسناد دست‌نویس، امکان به هم چسبیدن یا شکسته شدن حروف وجود دارد. این به هم چسبیدگی یا شکستگی باعث می‌شود که کدگذاری اشتباه انجام شود. به دلیل خطای بالای قطعه‌بندی در سطح حروف در زبان‌هایی که ساختار نگارشی پیوسته‌ای دارند، مانند زبان فارسی و عربی، از این دسته روش‌ها استفاده نشده است.

کدگذاری شکل کلمه[2]، به جای قطعه‌بندی کلمه به حروف و سپس کد کردن حروف، به‌صورت کل‌نگر کدگذاری را روی کلمه اجرا می‌کند. در این دسته روش‌ها تصویر کلمه به‌عنوان کوچک‌ترین جزء در نظر گرفته می‌شود. کدگذاری‌هایی که معمولاً در سطح کلمه در نظر گرفته می‌شود از ساختار و شکل حروف

---

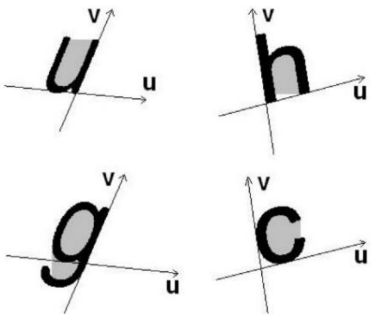
[1] Character Shape Coding(CSC)

[2] Word Shape Coding (WSC)

استفاده می‌کنند که معمولاً ویژگی‌هایی مانند حفره‌ها، پایین‌رونده و بالارونده‌ها، نقاط و سایر ویژگی‌های ساختاری وابسته به شکل نگارش زبان بهره می‌برند. معمولاً این روش‌ها از تطبیق الگو یا معیارهای شباهت جهت سنجش میزان شباهت کلمه پرسشی تصویری و کلمات کاندید در سند استفاده می‌کنند.

روش کدگذاری شکل کاراکتر در تحقیق [۱۰] استفاده شده تا موقعیت کلمات در تصاویری که با استفاده از دوربین دیجیتال اخذ شده است را مشخص نماید. در مرحله پیش‌پردازش استخراج خطوط و کلمات با استفاده از آنالیز اجزاء متصل انجام شده است. حروف i و l به‌عنوان حروف مبنا در نظر گرفته شده‌اند که جهت و میزان کجی خطوط با استفاده از آن‌ها سنجیده می‌شود و با استفاده از آن خط مبنا شناسایی می‌گردد. ویژگی‌های ساختاری حفره، مخازن آب‌ها[1] و بالارونده و پایین‌رونده برای کد کردن کلمات استفاده می‌شوند. شکل ۲-۲ نمونه‌هایی از وقوع این ویژگی‌ها را در برخی از حروف انگلیسی نشان می‌دهد. تعداد حفره‌های درون کاراکتر و محل قرار گرفتن مخازن آب در چهار جهت چپ، راست، پایین و بالا ۵ عدد اول دنباله ۶ رقمی کد کلمه را می‌سازند. بخش آخر از کد، بالارونده و پایین‌رونده بودن را با اختصاص یکی از ارقام ۱، ۱- و ۰ مشخص می‌کند. جدول ۲-۱ کدهای حروف مختلف انگلیسی را با استفاده از این روش کدگذاری ۶ رقمی نشان می‌دهد.

برای تعیین میزان مشابهت، فاصله همینگ[2] کد کلمه پرسشی با کد کلمات با طول یکسان در سند محاسبه شده است. متوسط فراخوانی و دقت در ارزیابی تجربی به ترتیب ۷۵/۷٪ و ۶۹/۰۸٪ گزارش شده است.



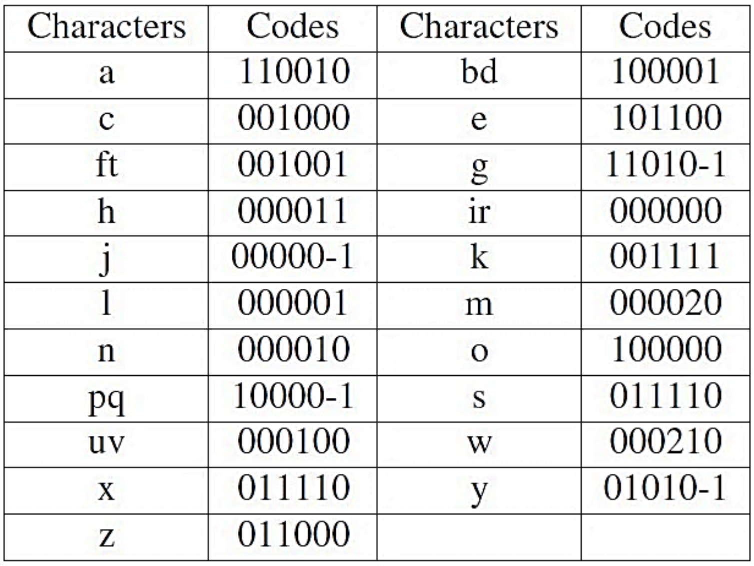
شکل ۲-۲: نمایش حفره و مخازن آب در برخی از حروف نمونه انگلیسی [۱۰]

---

[1] Water reservoirs

[2] Hamming distance

جدول ۲-۱: کدگذاری شکل حروف انگلیسی ارائه‌شده در [۱۰]

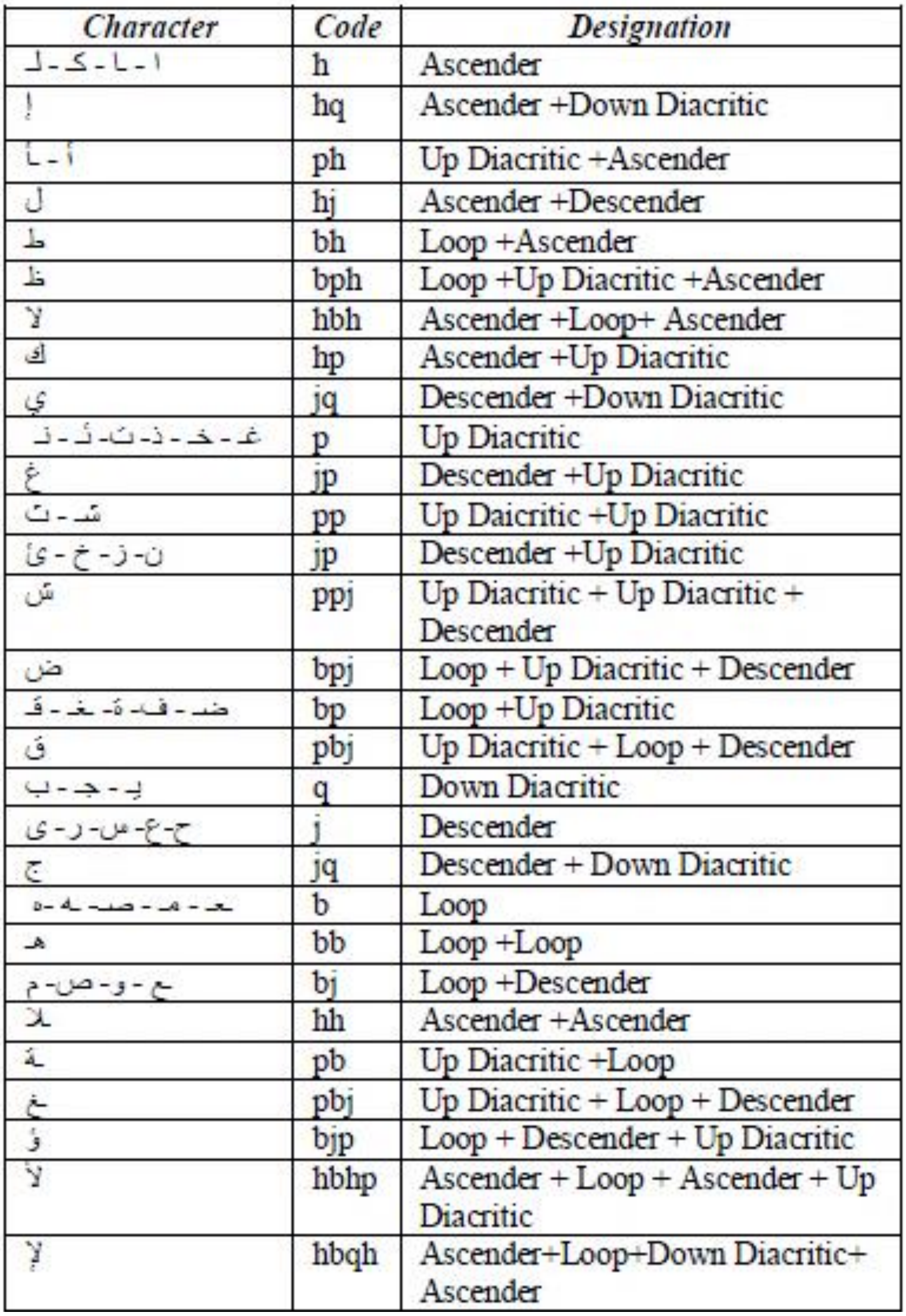


Characters	Codes	Characters	Codes
a	110010	bd	100001
c	001000	e	101100
ft	001001	g	11010-1
h	000011	ir	000000
j	00000-1	k	001111
l	000001	m	000020
n	000010	o	100000
pq	10000-1	s	011110
uv	000100	w	000210
x	011110	y	01010-1
z	011000		

روش کدگذاری بسیار وابسته به روش نگارش در زبان است. در زبان فارسی و عربی که ذات ساختار نگارشی پیوسته است استفاده از کدگذاری شکل حروف جایگاهی ندارد اما به جای حروف می‌توان از کدگذاری شکل کلمات بهره برد. این عمل در تحقیق [۱۱]، با استفاده از ویژگی‌های ساختاری روی زبان عربی انجام شده است. جدول ۲-۲ کدهای استفاده شده برای حروف عربی سازنده کلمات در این تحقیق را نشان می‌دهد.

کدگذاری شکل کلمات در مورد اسناد چاپی فارسی نیز مورداستفاده قرار گرفته شده است [۱۲-۱۴]. عملکرد این روش بر اساس تشخیص فونت سند مورد جستجو و سنتز کلمه پرسشی مطابق با فونت و سایز تشخیص داده شده استوار است. بعد از ایجاد کلمه پرسشی مطابق با سند مورد جستجو، خطوط سند با استفاده از افکنش استخراج می‌شوند و جستجو روی خطوط با استفاده از پنجره‌گذاری انجام می‌گیرد. جهت اندازه‌گیری میزان مشابهت نیز از XNOR استفاده شده است. این مرحله در واقع کاندیدها را استخراج می‌کند و جهت بازیابی دقیق به مرحله بعدی می‌فرستد. در مرحله نهایی از ویژگی‌های ساختاری مانند تعداد حفره‌ها، نقاط و بالارونده‌ها و پایین‌رونده‌ها استفاده شده است. شکل ۲-۳ و ۲-۴ ساختار کلی روش و نمونه‌هایی از ویژگی‌های ساختاری استفاده‌شده در مورد چند کلمه نمونه را نشان می‌دهد.

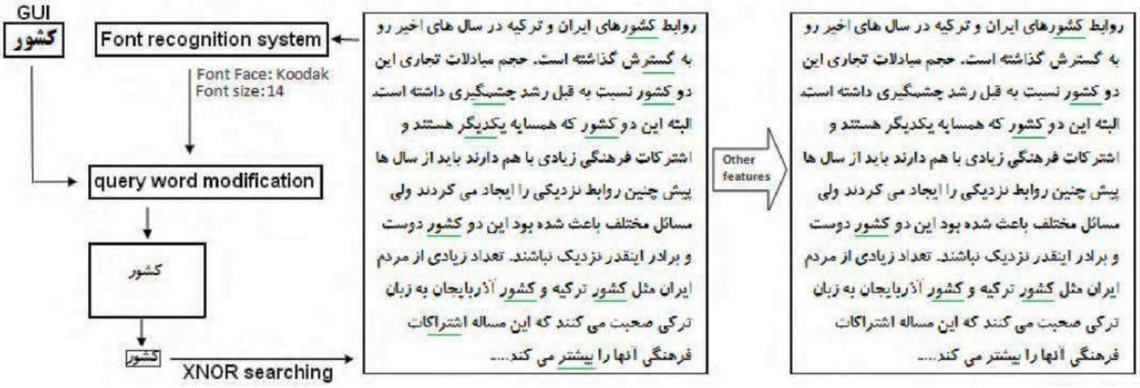
جدول ۲-۲: کدگذاری به‌کاررفته در حروف عربی ارائه‌شده در [۱۱]



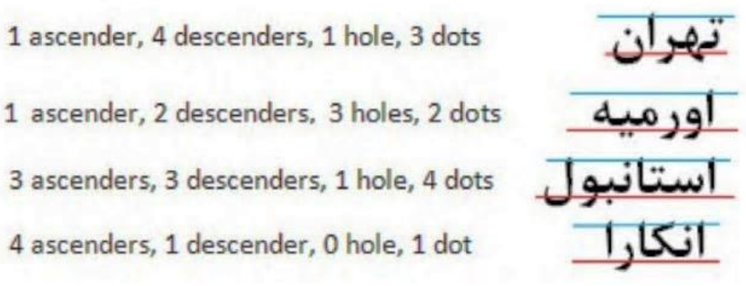
Character	*Code*	*Designation*
ا - ـا - كـ - ـلـ	h	Ascender
إ	hq	Ascender +Down Diacritic
أ - ـأ	ph	Up Diacritic +Ascender
ل	hj	Ascender +Descender
ط	bh	Loop +Ascender
ظ	bph	Loop +Up Diacritic +Ascender
لا	hbh	Ascender +Loop+ Ascender
ك	hp	Ascender +Up Diacritic
ي	jq	Descender +Down Diacritic
غـ - خـ - ذ- نـ- ثـ - نـ	p	Up Diacritic
غ	jp	Descender +Up Diacritic
شـ - ث	pp	Up Daicritic +Up Diacritic
ن- ز- خ - ئ	jp	Descender +Up Diacritic
ش	ppj	Up Diacritic + Up Diacritic + Descender
ض	bpj	Loop + Up Diacritic + Descender
ضـ - فـ - ة- ـغـ - قـ	bp	Loop +Up Diacritic
ق	pbj	Up Diacritic + Loop + Descender
بـ - جـ - ب	q	Down Diacritic
ح- ع- س- ر - ى	j	Descender
ج	jq	Descender + Down Diacritic
ـعـ - مـ - صـ - ـهـ - ه	b	Loop
هـ	bb	Loop +Loop
ـع - و - ص- م	bj	Loop +Descender
ـلا	hh	Ascender +Ascender
ـة	pb	Up Diacritic +Loop
ـغ	pbj	Up Diacritic + Loop + Descender
ؤ	bjp	Loop + Descender + Up Diacritic
لأ	hbhp	Ascender + Loop + Ascender + Up Diacritic
لإ	hbqh	Ascender+Loop+Down Diacritic+ Ascender

جهت ارزیابی روش مذکور، ۶۶۴ سند که بخشی از آن به‌صورت اسکن شده و بخش بزرگ‌تری از آن به‌صورت کامپیوتری تهیه شده بود استفاده گردید. دقت در مورد اسنادی که به‌صورت کامپیوتری تهیه شده بودند به دلیل عدم وجود مشکلات سخت‌افزاری و نویز در زمان تبدیل سند به تصویر بالاتر بوده و

برابر ۹۷/۵٪ است درحالی‌که برای اسناد اسکن شده دقتی برابر ۸۷/۶٪ گزارش شده است. همین مسئله در مورد معیار فراخوانی نیز صادق است و به ترتیب برابر ۹۲/۱٪ و ۷۹/۳٪ می‌باشد.



شکل ۲-۳: ساختار کلی روش ارائه‌شده در [۱۲]



شکل ۲-۴: نمونه‌ای از ویژگی‌های ساختاری چند کلمه فارسی [۱۲]

در [۱۵] و [۱۶] از ویژگی‌های مبتنی بر پروفایل دودویی با نام "گرادیانی، ساختاری و خمیدگی"[1] استفاده شده است. برای استخراج این بردار ویژگی، هر تصویر به پنجره‌های ۴ در ۸ وفقی تقسیم می‌شود (شکل ۲-۵). بردار ویژگی نهایی شامل ۱۹۲ ویژگی گرادیانی، ۱۹۲ ویژگی ساختاری و ۱۲۸ ویژگی تقعر برای هر کلمه است. در این روش از معیار همبستگی[2] برای محاسبه میزان شباهت بین دو بردار باینری استفاده شده است.

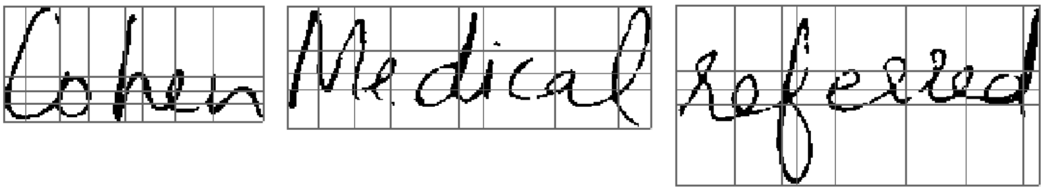
$$D^b(X,Y) = \frac{1}{2} - \frac{S_{11}S_{00} - S_{10}S_{01}}{2((S_{10}+S_{11})(S_{01}+S_{00})(S_{11}+S_{01})(S_{00}+S_{10}))^{1/2}} \qquad (۱-۲)$$

$D^b(X,Y)$ معیار عدم شباهت متناظر با معیار همبستگی است. در این رابطه، S_{ij} تعداد تطبیق i در الگوی اول با j الگوی دوم در موقعیت‌های متناظر است.

---

[1] Gradient, Structural and Convexity (GSC)

[2] Correlation measure



شکل ۲-۵: نمونه‌هایی از قطعه‌بندی وفقی کلمات در [۱۵]

در [۱۷] از توصیفگرهای مبتنی بر کانتور برای بازنمایی کلمات و از تطبیق کانتور کشسان استفاده می‌شود، درحالی‌که استفاده از ویژگی‌های گابور و فاصله اقلیدسی در [۱۸] بررسی شده است. ویژگی‌های کل‌نگر کلمه با یک مدل آماری حاشیه‌نویسی[1] در [۱۹] پیشنهاد داده شده است که می‌تواند کلمات دلخواه را جستجو کند.

برخی از رهیافت‌های محلی رایج، یک کلمه را به‌صورت توالی از ویژگی‌ها که با استفاده از یک پنجره لغزان روی کلمه استخراج شده است، نمایش می‌دهد. جهت مقایسه چنین توالی‌هایی، پیچش زمان پویا رایج‌ترین راهکار استفاده شده در جستجوی کلمات می‌باشد [۲۰, ۲۱] و به‌طور گسترده‌ای مورداستفاده قرار گرفته است[۲۲]. یک مقایسه بین پیچش زمان پویا و روش‌های پیکسلی در [۲۳] ارائه شده و مقایسه روش‌های مبتنی بر "گرادیانی، ساختاری و خمیدگی" با پیچش زمان پویا در [۱۵] و [۲۴] ارائه شده است.

تمامی رهیافت‌هایی که تا اینجا ذکر شد، نیازمند این بود که قبل از شروع عملیات جستجوی کلمه در مرحله پیش‌پردازش اسناد تصویری به کلمات تشکیل دهنده آن، قطعه‌بندی شود. با توجه به خطای بالای این نوع قطعه‌بندی، سناریوی متفاوت این است که سند فقط به خطوط قطعه‌بندی گردد و محل وقوع کلمه پرسشی در خطوط بررسی گردد یا اینکه خطوط بر اساس کلمه پرسشی رتبه‌بندی شوند. در [۲۵] روشی معرفی شده است که با استفاده از ویژگی‌های مبتنی بر پروفایل هر خط را به‌صورت یک سیگنال یک‌بعدی بازنمایی می‌کند و عمل جستجوی کلمه را با استفاده از پیچش زمان پویا انجام می‌دهد.

مدل مخفی مارکوف و تعمیم‌های مختلف آن، تکنیک‌های پرکاربردی در بازشناسی کلمات تصویری هستند[۲۶] که در جستجوی کلمه نیز مورد استفاده قرار گرفته‌اند. در[۲۷] یک مدل مخفی مارکوف شبه دوبعدی جهت جستجوی کلمه در اسناد چاپی دارای کیفیت پایین پیشنهاد شده است و ادعا شده که نسبت به کجی و اندازه فونت مقاوم است. تفاوت مدل مخفی مارکوف تعمیم‌یافته ارائه شده در [۲۸] نسبت به مدل پایه در انتخاب مدل انتقال است. در مدل مخفی مارکوف هر حالت مخفی تنها می‌تواند یک مشاهده داشته باشد درحالی‌که در مدل تعمیم‌یافته برای هر حالت مخفی مجموعه‌ای از مشاهده‌ها

---

[1] Annotation

می‌تواند اتفاق بیافتد. در تحقیق [۲۹] پیشنهادی ارائه شده است که از مزایای مدل مخفی مارکوف شبه گسسته[1] استفاده می‌کند تا مدلی وفق پذیر با شیوه‌های نگارش مختلف بسازد. خاصیت مدل مخفی مارکوف شبه گسسته این است که پارامترهای مدل کلمه را جدا می‌کند. این پارامترها به یک کتابچه کد[2] مشترک از اشکال و مجموعه‌ای از پارامترهای خاص هر کلمه، تجزیه می‌شود. درواقع ادعا شده است که پارامترهای وابسته به کلمه از پارامترهای وابسته به نویسنده با استفاده از این روش جدا می‌شود.

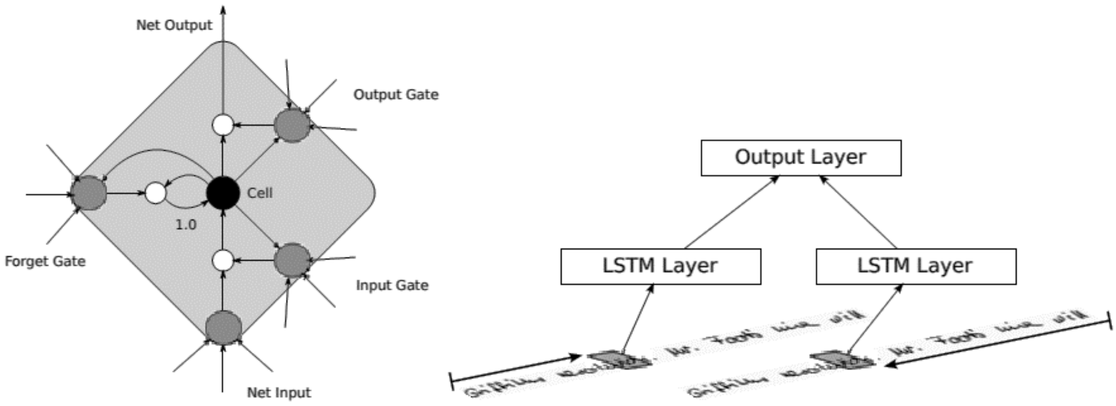
مدل‌های مخفی مارکوف چندین عیب شناخته‌شده دارند. یکی از این عیوب این است که در مدل مخفی مارکوف فرض می‌شود که احتمال هر مشاهده فقط به حالت جاری وابسته است؛ ازاین‌رو مدل کردن اثرات محتوایی مشکل است. علاوه بر این، مدل مخفی مارکوف یک مدل زایشی[3] است، درصورتی‌که مدل‌های تمایزی[4] کارایی بهتری را در رده‌بندی و برچسب‌گذاری دارند.

نوعی از شبکه‌های عصبی بازگشتی[5] نیز برای جستجوی کلمه مورداستفاده قرار گرفته است. در این روش یک شبکه عصبی بازگشتی برای برچسب‌گذاری یک سری آموزش داده می‌شود. قابلیت این نوع شبکه عصبی این است که یک حافظه از ورودی‌های قبلی در حالات درونی شبکه باقی می‌ماند و استفاده از محتوای قبلی را ممکن می‌سازد، بنابراین قابلیت پذیرش سری را دارا می‌باشد. میزان اطلاعات محتوایی که شبکه عصبی بازگشتی پایه می‌تواند حفظ کند بسیار محدود است (مشکل شیب میرا[6]). برای حل این مشکل، شبکه عصبی با حافظه کوتاه‌مدت بلند[7] معرفی شده است که لایه‌های مخفی آن شامل زیر شبکه‌هایی با اتصالات بازگشتی است که بلوک حافظه نامیده می‌شوند. شبکه عصبی با حافظه کوتاه‌مدت بلند دوجهته[8] هم در جستجوی کلمه استفاده‌شده است[۳۰]. این کار در دو بخش انجام شده است که در بخش اول از شبکه عصبی مذکور استفاده شده است. در این مرحله شبکه عصبی هر موقعیت از سری ورودی را به یک بردار نگاشت می‌کند که احتمال وقوع هر کاراکتر را در آن موقعیت خاص نشان می‌دهد. در مرحله دوم الگوریتم گذر نشانه CTC[9] این سیستم توانایی بازیابی و جستجوی کلمات خارج از فرهنگ لغت[10] را نیز داراست. علاوه بر این برای آموزش نیازی به کلمات و حروف مجزا نیز ندارد. شمای کلی عملکرد این سیستم در شکل ۲-۶ نمایش داده شده است. در این روش از شبکه عصبی بازگشتی استفاده

---

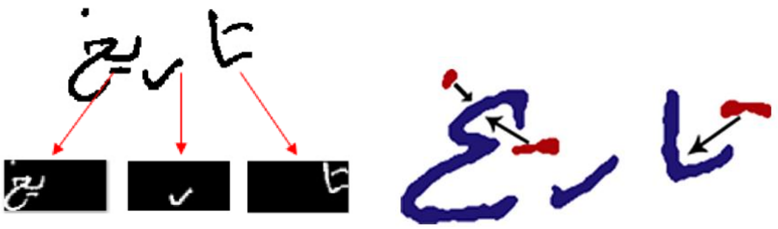
[1] Semi-continuous hidden Markov model (SC-HMM)
[2] Codebook
[3] Generative
[4] Discriminative
[5] Recurrent Neural Network (RNN)
[6] Vanishing gradient problem
[7] Long Short-Term Memory (LSTM)
[8] Bidirectional long short-term Memory (BLSTM)
[9] CTC Token Passing Algorithm
[10] Out of Vocabulary (OOV)

شده است و مانند بسیاری دیگر از روش‌های موجود بزرگ‌ترین چالشی که دارد تفاوت در نوع دست‌خط نویسنده‌ها می‌باشد.



شکل ۲-۶: شمای کلی شبکه عصبی با حافظه کوتاه‌مدت طولانی دوجهته [۳۰]

در رهیافتی جدید روشی در سطح خط و مبتنی بر یادگیری ارائه شده است که از مدل زبانی جهت جستجوی کلمه در اسناد عربی بهره می‌برد[۳۱]. این سیستم در ابتدا "بخشی از کلمه عربی"[1] را جدا می‌کند. بخشی از کلمه عربی که در برخی اسناد زیرکلمه گفته می‌شود شامل یک جزء متصل از بدنه اصلی کلمه به همراه نقاط و نشانه‌های آن می‌باشد (شکل ۲-۷). سپس با استفاده از مدل زبانی عمل بازیابی کلمات دست‌نویس انجام می‌شود. در این روش خطوط جداشده از سند و کلمه پرسشی تصویری به بخشی از کلمات عربی قطعه‌بندی می‌شود. ویژگی‌های گرادیانی و رده‌بند آنالیز تفکیکی منظم[2] جهت بازشناسی بخشی کلمات عربی استفاده شده است. برای هرکدام از بخش‌های استخراج‌شده از کلمه پرسشی یک رده‌بند آموزش داده می‌شود. درنتیجه یک توالی از رده‌بندها برای هر کلمه پرسشی وجود خواهد داشت. بخشی کلمات موجود در هر خط از سند به این رده‌بند سلسله مراتبی داده می‌شود. گراف‌هایی برای بخشی کلمات با ضریب اطمینان بیشتر از یک حد آستانه ساخته می‌شود. سپس مسیرهای گراف جهت پذیرش یا رد کلمه موردنظر ارزیابی می‌شود (شکل ۲-۸).

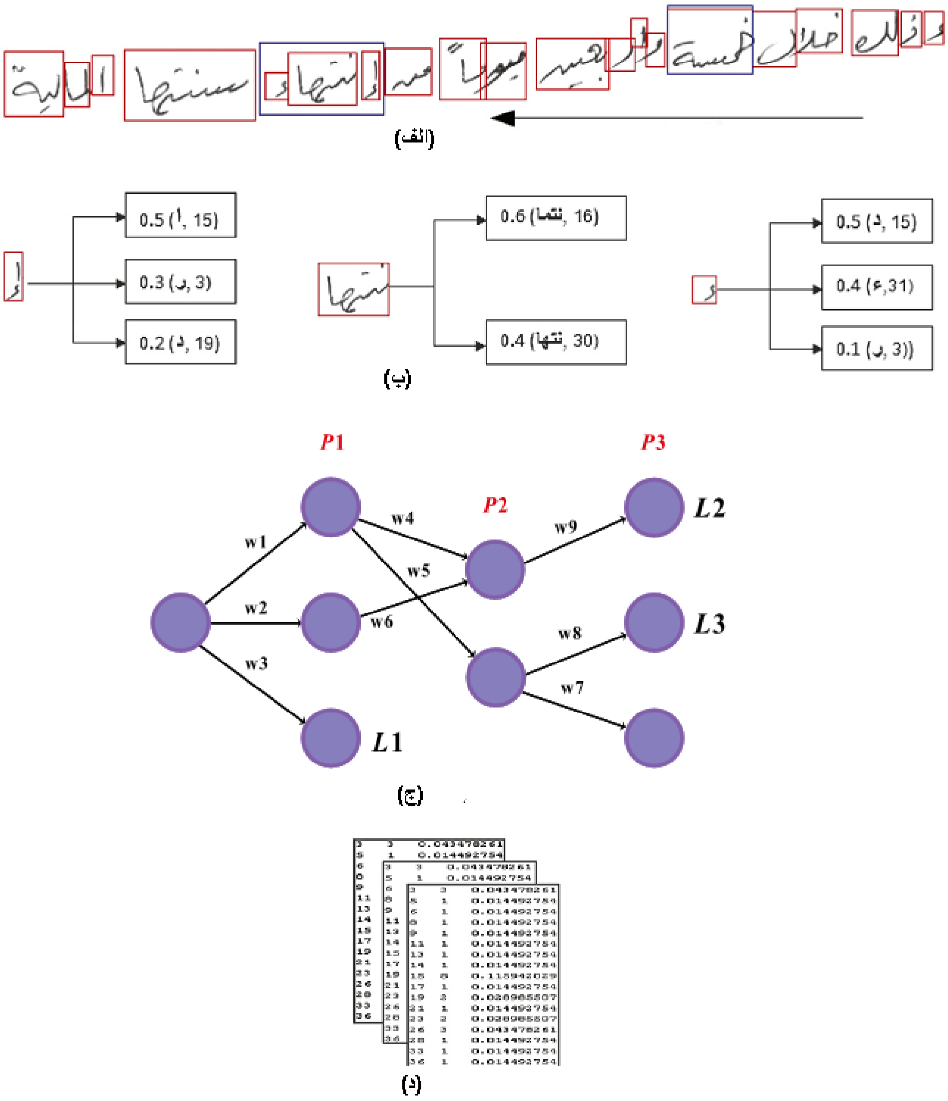


شکل ۲-۷: نمونه‌ای از قطعه‌بندی به بخشی کلمه عربی برای کلمه "تاریخ"

---

[1] Part of Arabic Word (PAW)

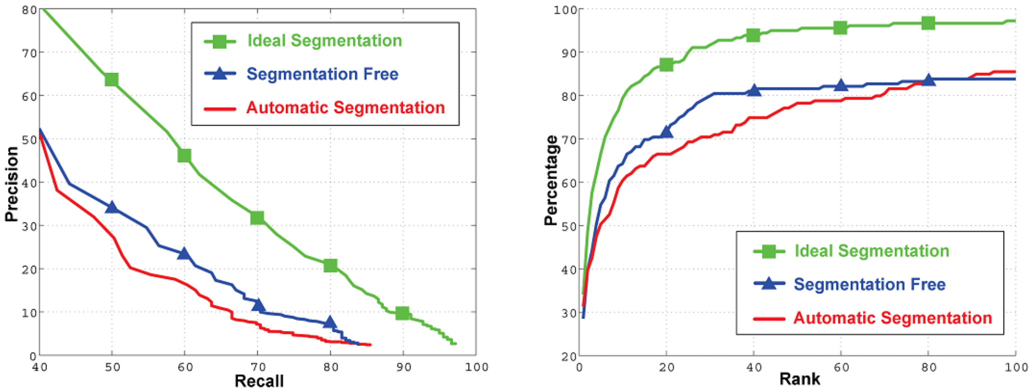
[2] Regularized Discriminant Analysis (RDA)



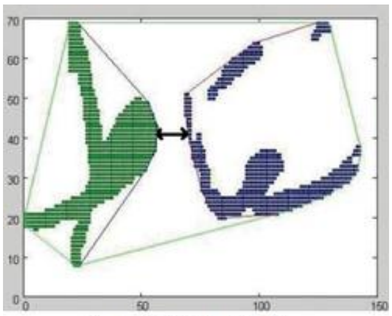
شکل ۲-۸: فرآیند آزمون رده‌بند سلسله مراتبی [۳۱]. الف)قطعه‌بندی. ب) نتایج رده‌بندی. ج) ساخت گراف. د) مدل‌های زبانی

۲-۲-۲- روش‌های مستقل از قطعه‌بندی

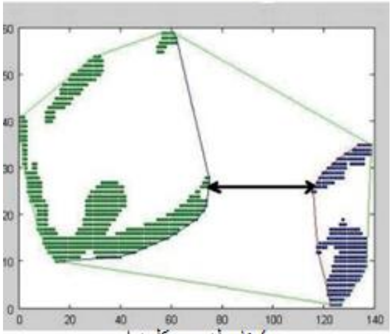
روش‌های جدید جستجوی کلمه سعی کرده‌اند تا به علت پرخطا بودن عمل قطعه‌بندی از قطعه‌بندی کلمه یا به‌طورکلی قطعه‌بندی اجتناب کنند. جهت مقایسه در [۳۲, ۳۳] دو روش مبتنی بر قطعه‌بندی و مستقل از قطعه‌بندی بر روی دست‌نوشته‌های عربی موردبررسی قرار گرفته است و نشان داده شده است که روش‌های مستقل از قطعه‌بندی دقتی در حدود ۵ تا ۱۰ درصد بالاتر از روش‌های مبتنی بر قطعه‌بندی در دست‌نوشته‌های به زبان عربی دارند. شکل ۲-۹ مقایسه روش مبتنی بر قطعه‌بندی اتوماتیک، مستقل از قطعه‌بندی و مبتنی بر قطعه‌بندی ایده آل را نشان می‌دهد.



شکل ۲-۹: مقایسه دقت و فراخوانی در روش مبتنی بر قطعه‌بندی و مستقل از آن در عربی [۳۲]



الف) فاصله بین زیرکلمات



ب) فاصله بین کلمه ای

شکل ۲-۱۰: قطعه‌بندی کلمه عربی "نقاط" [۳۲]

روش مبتنی بر قطعه‌بندی ارائه‌شده ابتدا زیرکلمات عربی را استخراج می‌کند. این کار بر اساس اجزاء متصل صورت می‌پذیرد. به این صورت که اجزاء متصل استخراج‌شده و اجزاء کوچکی مانند نقاط به اجزاء بزرگ‌تر نسبت داده‌شده و یک خوشه را تشکیل می‌دهند. نمونه‌ای از این خوشه‌ها در شکل ۲-۱۰ نشان داده شده است. حرف ابتدای کلمه در بسیاری از کلمات در زبان عربی " الف" است. این مسئله به‌عنوان معیاری جهت تشخیص اول کلمه در این روش استفاده شده است. بعدازاین باید تشخیص داده شود که فواصل بین خوشه‌ها مربوط به فاصله بین کلمات است یا فاصله بین زیرکلمات تشکیل‌دهنده یک کلمه،

ازاین‌رو جهت رده‌بندی فواصل بین خوشه‌ها از شبکه عصبی مصنوعی استفاده شده است . شکل ۲-۱۰ دو خوشه‌ی مجاور از کلمه "نقاط" را نشان می‌دهد. ویژگی‌هایی که جهت تشخیص این فاصله مورداستفاده قرار می‌گیرند عبارت‌اند از: طول خوشه‌ها، وجود یا عدم وجود حرف "الف" در خوشه‌ها، تعداد اجزاء متصل خوشه‌ها، تفاوت کادر محاط کننده[۱] محتوای دو خوشه، حداقل فاصله بین بدنه‌ی محدب[۲] خوشه‌ها و نسبت مجموع بدنه‌ی محدب در هر دو خوشه به بدنه‌ی محدب کلمه‌ای که هر دو خوشه را شامل می‌شود.

در تحقیقی که اخیراً گزارش شده است [۳۴] اسناد با استفاده از شبکه‌ای از توصیفگرهای هیستوگرام گرادیان[۳] بازنمایی شده است. اسناد به پنجره‌های مربعی شکل با طول ثابتی به اندازه ۵ پیکسل تقسیم‌بندی شده و ویژگی‌های هیستوگرام گرادیان از آن استخراج می‌شود. رهیافت پنجره لغزان راهکاری است که جهت جستجوی کلمه استفاده شده است. با استفاده از روش بدون ناظر ماشین بردار پشتیبان مبتنی بر یک نمونه[۴] و بردار فیشر[۵] [۳۵] نواحی بازیابی شده مجدداً مرتب‌سازی می‌شوند. روش فیشر با توجه به فشرده‌سازی زیاد اجازه می‌دهد تا تعداد سند بیشتری به‌طور هم‌زمان در حافظه اصلی قرار گیرد و سرعت جستجو را افزایش می‌دهد. مراحل این سیستم در شکل ۲-۱۲ نشان داده شده است. در این روش اسناد و کلمات پرسشی به پنجره‌های کوچک هم‌اندازه‌ای تقسیم شده و ویژگی هیستوگرام گرادیان از هر پنجره استخراج می‌شود. امتیاز نواحی اسناد از کانولوشن کلمه پرسشی نسبت به نواحی سند به دست می‌آید. این عمل با استفاده از ضرب داخلی توصیفگرهای هیستوگرام گرادیان، به‌عنوان یک معیار مشابهت، صورت می‌پذیرد. علاوه بر این از ماشین بردار پشتیبان مثالی، برای یادگیری بدون ناظر در زمان جستجو با استفاده از کلمه پرسشی استفاده شده است.

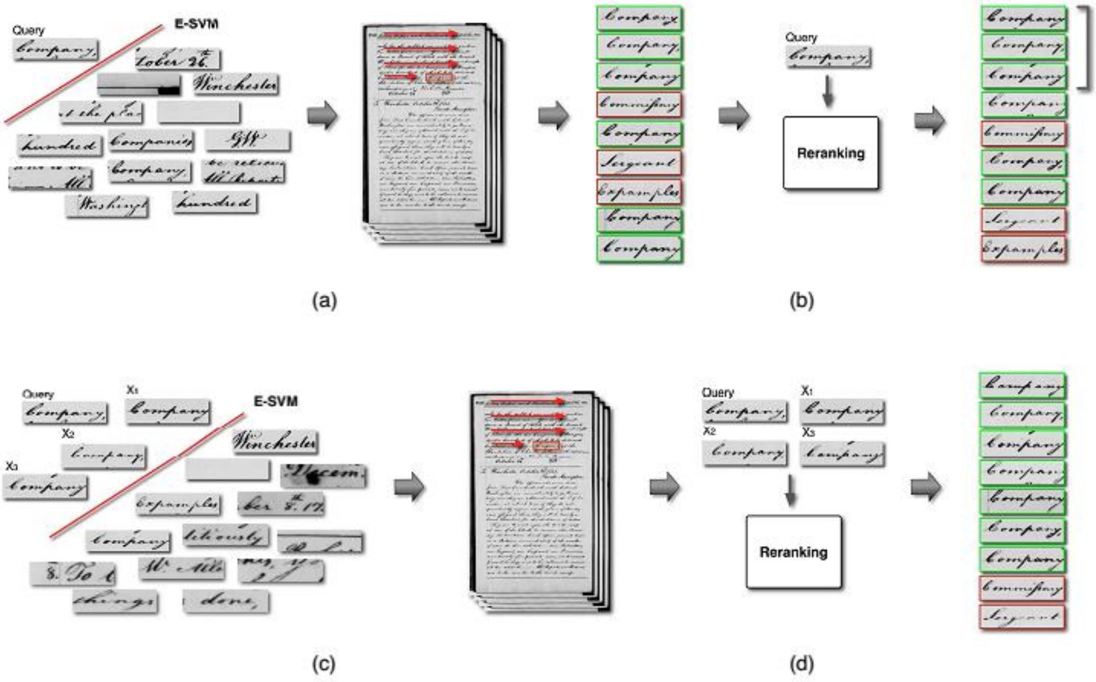
روش مستقل از قطعه‌بندی دیگری با بهره‌گیری از قابلیت توصیفگر سیفت[۶] در یک چارچوب کاری مبتنی بر وصله[۷] تصویر یک مدل سبد بصری کلمات[۸] ایجاد کرده است [۳۶]. توصیفگرهای هر وصله تصویر با اعمال اندیس‌گذاری معنایی پنهان[۹] پالایش شده‌اند.

تأثیر قطعه‌بندی نامناسب در روش‌های جستجوی کلمه نیز موردبررسی قرار گرفته است[۳۷]. در این تحقیق زمانی که برچسب کلمات و یک روش قطعه‌بندی مناسب وجود داشته باشد، استفاده از بردار

---

[1] Bounding Box
[2] Convex Hull
[3] Histogram of Gradient (HOG)
[4] Exemplar SVM
[5] Fisher Vector
[6] Scale invariant feature transform (SIFT)
[7] Patch
[8] Bag-of-Visual-Word (BoVW)
[9] Latent Semantic Indexing (LSI)

فیشر بهترین انتخاب معرفی شده است. سبد بصری کلمات نیز برای روش‌های مستقل از یادگیری بهترین نتایج را گزارش کرده است[۳۷].



شکل ۲-۱۲: شمای کلی روش معرفی‌شده در [۳۴]

۲-۳- دسته‌بندی بر اساس روش تشخیص

روش‌های جستجوی کلمه در برخی از مقالات بر اساس روش تشخیص دسته‌بندی شده‌اند. روش تشخیص می‌تواند بر اساس تطبیق قالب در سطح پیکسل یا بر اساس ویژگی‌های استخراج‌شده انجام شود. عمل تطبیق کلمه پرسشی تصویری با کلمات سند بر اساس یک معیار مشابهت صورت می‌پذیرد. این معیارهای مشابهت تنوع بسیار زیادی در کارهای جستجوی کلمه بر اساس تطبیق دارند. از معیار شباهت ساده‌ای مانند یای انحصاری[۱] [۱۲, ۳۸] تا معیارهای مشابهت معروفی مانند فاصله اقلیدسی و معیار کساین جهت ارزیابی مشابهت استفاده شده است. همچنین معیارهایی برای حل این مسئله خاص ساخته شده‌اند. ویژگی‌های مبتنی بر پروفایل و پیچش زمان پویا از جمله روش‌های قدیمی و پرکاربرد در این زمینه است[۲۳]. اگرچه روش‌های دیگری نیز معرفی شده‌اند، اما روش پیچش زمان پویا و نسخه‌های تغییریافته آن به‌عنوان یکی از معیارهای مشابهت پرکاربرد در این دسته از روش‌ها

---

[1] XOR

مورداستفاده قرار گرفته است. بزرگ‌ترین مشکل این دسته روش‌ها، حساس بودن زیاد به تنوع شیوه نگارش و تبدیلات اَفاین[1] است. همچنین روش‌های بیان‌شده بسیار وابسته به نویسنده است و با تغییر نویسنده یا تنوع آن دقت نتایج به‌طور شگفت‌آوری کاهش می‌یابد.

استفاده از ویژگی‌هایی که با پنجره‌گذاری به دست می‌آیند و طول متغیری وابسته به طول و شیوه نگارش کلمه دارند، بسیار رایج است. ویژگی‌های با طول متغیر به دلیل توانایی بیشتر در وفق‌پذیری با تنوع شیوه نگارش و طول کلمه، انعطاف‌پذیری بسیار بیشتری نسبت به روش‌های با ویژگی‌های با طول ثابت از خود نشان می‌دهند[۳۹, ۴۰]. ازاین‌رو، زمانی که به دنبال استفاده از مدل‌های یادگیری در مسئله جستجوی کلمه بوده‌اند، مدل مخفی مارکوف همواره به‌عنوان یکی از پرکاربردترین روش‌ها مطرح بوده است. مشکلات ویژگی با طول متغیر و استفاده از مدل مخفی مارکوف به شرح زیر است: اولاً، محدودیت در بازیابی و جستجوی کلمه پرسشی وجود دارد. در این حالت فقط توانایی جستجوی مجموعه محدودی از کلمات که در فاز آموزش استفاده شده است امکان‌پذیر خواهد بود و توانایی شناسایی کلمات خارج از مجموعه آموزشی وجود نخواهد داشت. ثانیاً، روش‌هایی که شباهت بین توالی‌ها را ارزیابی می‌کنند معمولاً روش‌های بسیار کندی هستند که این مسئله زمانی که تعداد اسناد افزایش پیدا می‌کند به‌صورت چشم‌گیری خود را نشان می‌دهد. معمولاً در این دسته روش‌ها نیاز به قطعه‌بندی در سطح کلمه وجود دارد. مدل مخفی مارکوف که یکی از پرکاربردترین رده‌بندهایی است که در چنین سیستم‌هایی بکار گرفته شده است، زمانی برای اسناد چند نویسنده نتایج خوبی را به همراه خواهد داشت که مجموعه داده آموزشی بزرگی فراهم شده باشد[۴۱].

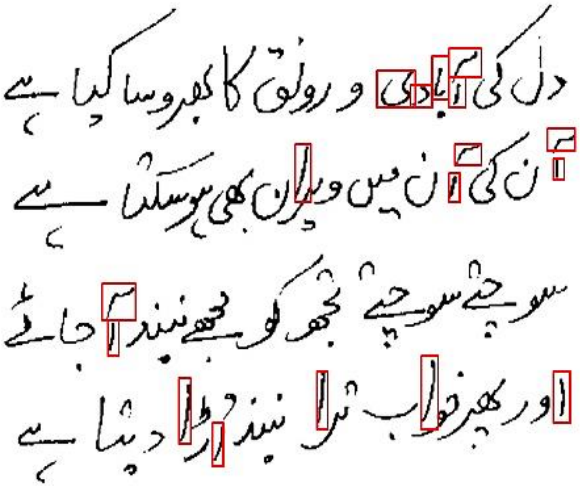
۲-۳-۱- روش‌های مبتنی بر تطبیق الگو

تحقیق [۴۲] نیز روشی را جهت جستجوی کلمه روی اسناد دست‌نویس اردو پیشنهاد داده است. ساختار اصلی و حروف در نوشتار زبان اردو، عربی و فارسی بسیار به هم نزدیک هستند. ازآنجایی‌که ساختار نوشتاری در این زبان‌ها ذات پیوسته‌ای دارد و هر کلمه در این زبان می‌تواند از کنار هم قرار گرفتن زیرکلمات تشکیل شود، روش ارائه‌شده از دو بخش اصلی اندیس‌گذاری و بازیابی تشکیل شده است. در مرحله اندیس‌گذاری قطعه‌بندی در سطح زیرکلمات انجام می‌شود و ویژگی‌های زیرکلمات قطعه‌بندی شده هر سند ذخیره می‌شود. ویژگی‌های استخراج‌شده ساختاری هستند و شامل موارد اسکالر مانند نسبت طول به عرض و مساحت ناحیه محدب، ویژگی‌های برداری مانند پروفایل‌ها، افکنش‌ها و گذر جوهر به پس‌زمینه می‌باشد.

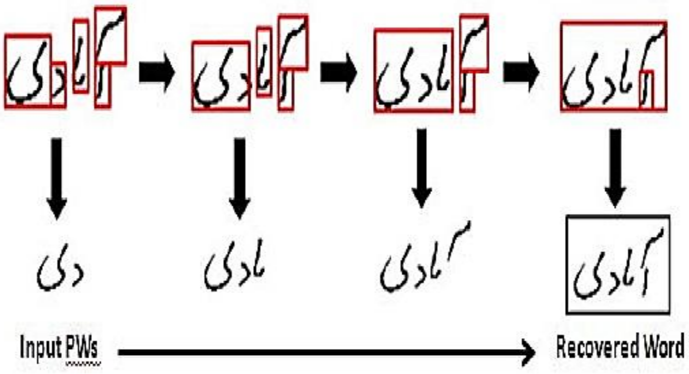
---

[1] Affine Transform

در زمان جستجو یک فرآیند سه مرحله‌ای انجام می‌شود. مرحله اول کاهش فضای جستجو است که برخی از زیرکلمات از فاز جستجو حذف می‌گردند. این مرحله بر اساس ویژگی‌های اسکالر که بسیار کلی هستند و مربوط به اندازه و طول و عرض ناحیه محدب زیرکلمه می‌باشند انجام می‌شود. پیچش زمان پویا و ویژگی‌های ساختاری برداری در مرحله بعد برای تطبیق زیرکلمات سند و کلمه پرسشی استفاده می‌شود. این مرحله روی زیرکلماتی از اسناد انجام می‌گیرد که از مرحله اول گذر کرده و حذف نشده‌اند. مرحله نهایی ترکیب زیرکلمات کنار هم و به دست آوردن کلمه پرسشی مورد جستجوی کاربر است. شکل ۲-۱۳ و ۲-۱۴ مراحل ذکرشده را نشان می‌دهد.



شکل ۲-۱۳: تشخیص زیرکلمات بر اساس ویژگی‌های ساختاری برای کلمه پرسشی "آبادی" [۴۲]

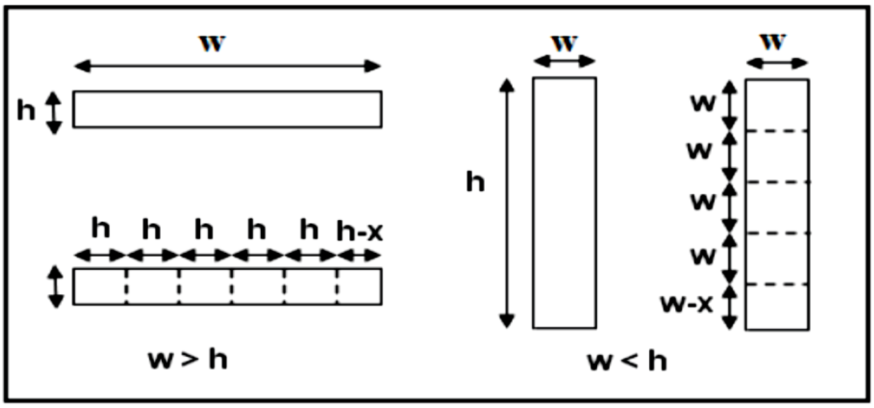


شکل ۲-۱۴: نمونه‌ای از ترکیب نواحی یافت شده برای زیرکلمات کلمه پرسشی "آبادی" [۴۲]

اولین تحقیق در زمینه‌ی جستجوی کلمات فارسی در اسناد چاپی، بر اساس ادعای نویسنده، در [۱۴] گزارش شده است که با بهره‌گیری از شناسایی اندازه و فونت سند مورد جستجو کار می‌کند. ویژگی‌های

استفاده‌شده جهت بازنمایی کلمات مستقل از فونت است و شامل مواردی مانند تعداد زیرکلمات و نسبت طول به عرض و ویژگی‌هایی که با استفاده از ناحیه‌بندی تصویر به دست می‌آیند، می‌باشد. ویژگی‌هایی که با استفاده از ناحیه‌بندی تصویر به دست می‌آیند تصویر را در جهتی که گستردگی بیشتری دارد بلوک‌بندی می‌کنند. اندازه این بلوک‌ها به اندازه جهتی است که اندازه کوچک‌تری دارد. در هر ناحیه تعداد پیکسل‌های قلم محاسبه می‌گردد. شکل ۲-۱۵ نحوه انجام این بلوک‌بندی را نمایش می‌دهد.

در فاز جستجو، اولین مرحله تشخیص فونت و اندازه سند مورد جستجو می‌باشد. پس از شناسایی فونت و اندازه، کلمه پرسشی بر اساس فونت و اندازه تشخیص داده‌شده بازنویسی می‌شود و ویژگی‌های مذکور استخراج می‌شوند.



شکل ۲-۱۵: تشریح روش ناحیه‌بندی جهت استخراج ویژگی [۱۴]

در مرحله جستجو عمل تطبیق به‌صورت چند سطحی انجام شده است. ویژگی نسبت طول به عرض به‌عنوان یک ویژگی ساده که قابلیت کاهش فضای جستجو را دارد، مورد استفاده قرار گرفته است. در مرحله نهایی فقط عمل تطبیق با کلمات سند موجود در فضای جستجوی کاهش‌یافته انجام می‌گیرد. در این مرحله ویژگی‌های پروفایل مورد استفاده قرار می‌گیرد. مشکل اصلی گزارش‌شده در این روش مسئله خطای قطعه‌بندی سند به خطوط و کلمات است.

در تحقیق [۴۳] روشی جهت جستجوی کلمات در اسناد تاریخی فارسی ارائه شده است که بر اساس اندیس‌گذاری زیرکلمات کار می‌کند و از ویژگی‌های ساختاری مبتنی بر پروفایل و پیچش زمان پویا جهت جستجوی کلمه بهره برده است. در این تحقیق ارزیابی روش پیشنهادی روی یکی از کتب تاریخی دانشگاه بیرجند گزارش شده است. برخی نیز سعی کرده‌اند با ارائه روش‌های تطبیق دقیق‌تر و سریع‌تر مشکلات موجود در این حوزه را حل نمایند. روش پیچش زمان پویای دوطرفه با استفاده از ویژگی‌های هیستوگرام گرادیان یکی از روش‌های پیشنهادی در این زمینه بوده است[۴۴]. اخیراً روش تطبیق

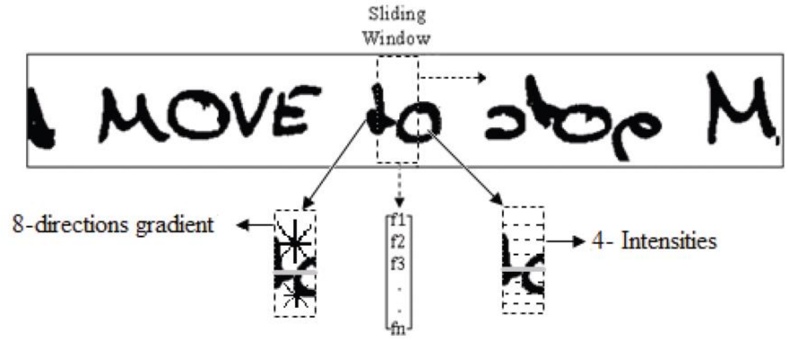
انعطاف‌پذیر توالی[1] ارائه شده است که توانایی بازیابی کلماتی را دارد که بخشی از آن مشابه با کلمه پرسشی موردنظر کاربر است[۴۵].

۲-۳-۲- روش‌های مبتنی بر یادگیری

روش‌های جستجوی کلمه مبتنی بر یادگیری جهت کار با اسناد چند نویسنده پیشنهاد شده و نتایج قابل قبولی را نشان داده‌اند. علاوه بر این، پایگاه داده‌های بزرگ نیاز به سیستم‌های مبتنی بر یادگیری دارند. مدل مخفی مارکوف رایج‌ترین رده‌بندی است که در سیستم‌های جستجوی کلمه مورد استفاده قرار گرفته است[۴۶, ۴۷]. روش دیگری که قبلاً شرح داده شده از شبکه عصبی با حافظه کوتاه‌مدت بلند دوجهته بهره می‌برد[۴۸].

در تحقیق [۴۹] مدل مخفی مارکوف به‌عنوان یک مدل یادگیری جهت جستجوی کلمه در اسناد دست‌نویس مورد استفاده قرار گرفته است. در این تحقیق، مدل مخفی مارکوف روی حروف آموزش داده شده است و ترکیب این مدل‌ها و استفاده از مدل فیلتر، روشی مستقل از قطعه‌بندی کلمه را ارائه کرده است که روی زبان‌های مختلفی ازجمله انگلیسی، عربی و دوانگاری[2] مورد ارزیابی قرار گرفته است.

استخراج ویژگی در سند با استفاده از حرکت پنجره لغزان روی خطوط انجام می‌پذیرد. این پنجره لغزان در جهت نگارش و با هم‌پوشانی ۸۵ درصدی حرکت می‌کند. ویژگی‌هایی که از هر فریم از پنجره استخراج می‌شود، ویژگی‌های سطح روشنایی و گرادیانی است. ویژگی‌های سطح روشنایی با تقسیم‌بندی فریم به چهار ناحیه و محاسبه نسبت پیکسل‌های سیاه و سفید محاسبه می‌گردد. به‌صورت شماتیک روش استخراج ویژگی در شکل ۲-۱۶ نشان داده شده است.



شکل ۲-۱۶: روند استخراج ویژگی از یک فریم از پنجره لغزان [۴۱]

---

[1] Flexible sequence matching (FSM)

[2] Devanagar

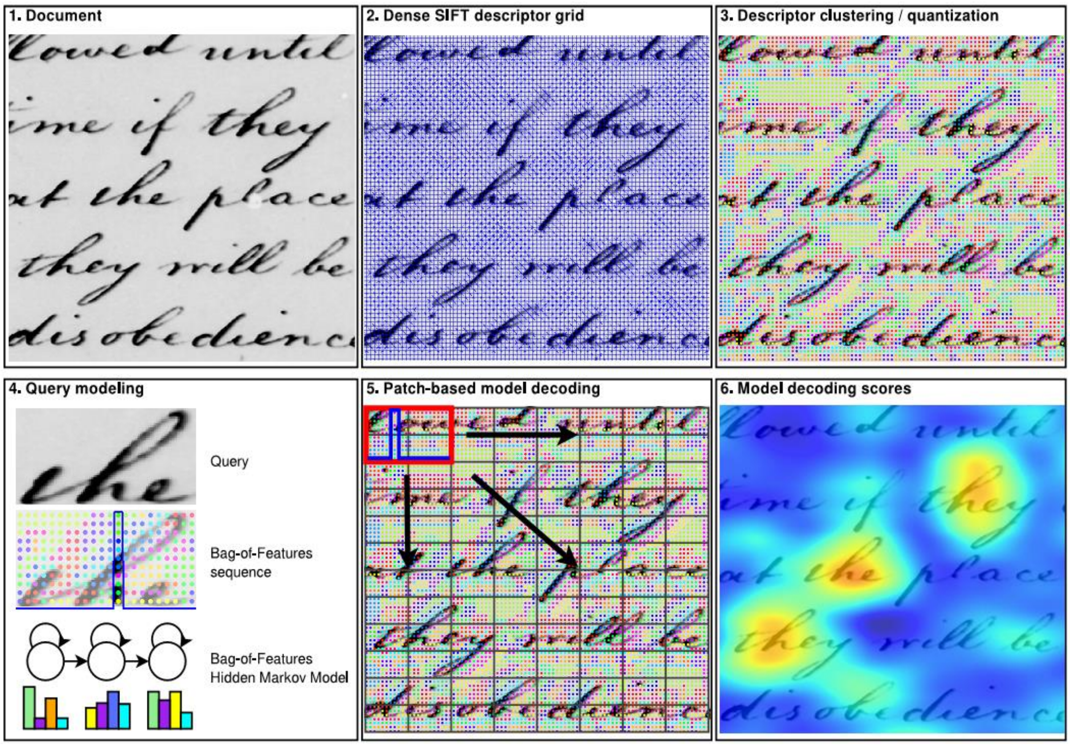
رهیافت سبد بصری کلمات یکی از روش‌های رایج و پرکاربرد در بینایی ماشین است[۵۰, ۵۱]. رهیافت سبد بصری کلمات روشی است که از مفهوم سبد کلمات در پردازش زبان طبیعی الگوبرداری شده است. در تصاویر، مفهوم کلمه همانند پردازش زبان طبیعی وجود ندارد. بنابراین بخش‌هایی از تصویر که شامل اطلاعات مهم‌تری هستند و می‌توانند تصویر را به‌خوبی توصیف کنند به‌عنوان یک کلمه بصری در نظر گرفته می‌شود. معمولاً این بخش‌های حاوی اطلاعات مهم به‌صورت اتوماتیک، استخراج و خوشه‌بندی شده و نماینده هر خوشه کلمات بصری را می‌سازند. به این منظور، نقاط کلیدی استخراج‌شده از تصویر با استفاده از یکی از توصیفگرهای مناسب بازنمایی می‌شوند و بر اساس این ویژگی‌های استخراج‌شده خوشه‌بندی می‌شوند. در این صورت هر خوشه معرف یک کلمه‌ی بصری[1] است و مجموعه خوشه می‌تواند به‌عنوان یک فرهنگ لغت بصری در نظر گرفته شود. هر تصویر با استفاده از بردار هیستوگرامی که رخدادهای هر کلمه را در سند تصویری نشان می‌دهد، توصیف می‌گردد [۵۲]. ازآنجایی‌که کیف بصری کلمات در جوامع آنالیز اسناد تصویری محبوبیت بسیاری پیدا کرده است، اخیراً کارهای زیادی با استفاده از این چهارچوب ارائه شده است[۵۳].

ترکیبی از کیف بصری کلمات و مدل مخفی مارکوف جهت جستجوی کلمه در اسناد دست‌نویس مستقل از قطعه‌بندی در [۵۴] پیشنهاد شده است. این روش برای کلمات با طول کوتاه بهبود خوبی را نشان می‌دهند. علاوه بر این تنها با داشتن یک نمونه می‌تواند مدل کلمه پرسشی را تخمین بزند. ساختار کلی این روش در شکل ۲-۱۷ نشان داده شده است.

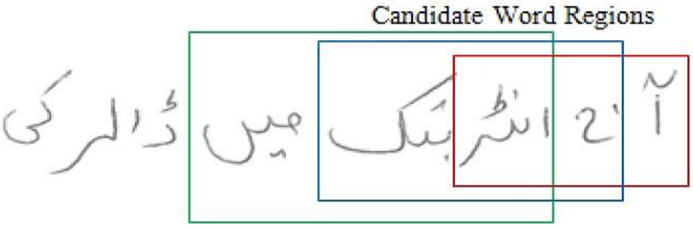
تحقیق [۵۵]، روشی را جهت جستجوی کلمه روی اسناد دست‌نویس اردو پیشنهاد کرده است. این روش مبتنی بر قطعه‌بندی در سطح خط انجام شده است. پنجره لغزان بر اساس اجزاء متصل تشکیل‌دهنده زیرکلمات روی خطوط از راست به چپ حرکت می‌کند و نواحی کاندید را ایجاد می‌نماید. شکل ۲-۱۸ نمونه‌ای از حرکت پنجره لغزان جهت ایجاد نواحی کاندید روی خطوط را نمایش می‌دهد. ویژگی‌های گرادیانی و ساختاری از نواحی کاندید استخراج شده و با استفاده از رده‌بند ماشین بردار پشتیبان، جستجوی کلمه انجام می‌شود. یک مرحله اضافی با استفاده از رتبه خروجی ماشین بردار پشتیبان تعداد پیکسل‌های سیاه و تعداد نقاط، نتیجه خروجی مرحله قبل را وارسی و تائید می‌نماید. دقتی برابر ۵۰/۷۵٪ به همراه فراخوانی ۷۰/۱٪ جهت اسناد دست‌نویس اردو گزارش شده است.

---

[1] Visual word



شکل ۲-۱۷: روند کلی روش ارائه‌شده در [۵۴]



شکل ۲-۱۸: حرکت پنجره لغزان روی یک خط‌نمونه و ایجاد نواحی کاندید [۵۵]

۲-۴- تقسیم‌بندی بر اساس نوع کلمه‌ی پرسشی

در اکثر روش‌های جستجوی کلمه در اسناد تصویری، کلمه پرسشی به‌صورت تصویری در نظر گرفته شده که از تصویر سند اصلی بریده شده است؛ بر این اساس آماده کردن کلمه پرسشی کار زمان‌بر و طاقت‌فرسایی است. علاوه بر این، کارایی در سیستم‌هایی که کلمه پرسشی را فقط به‌صورت تصویری می‌پذیرند، نتایج مطلوبی را بر روی اسناد چند نویسنده ارائه نمی‌کنند. علاوه بر این، در روش‌هایی که جستجو در آن‌ها بر اساس یادگیری انجام شده است، مشکل وابستگی به فرهنگ لغت وجود دارد و توانایی جستجو و بازیابی کلمات خارج از فرهنگ لغت وجود ندارد. جستجو با این نوع کلمه پرسشی

می‌تواند به‌صورت یک مسئله کلاسیک بازیابی تصویر[1] فرمول‌بندی شود. تصویر کلمه‌ی پرسشی با مجموعه‌ای از تصاویر کاندید در پایگاه داده مقایسه شده و سیستم شبیه‌ترین تصاویر را از پایگاه داده برمی‌گرداند و یا به‌صورت کلی‌تر تمام کاندیدها را با توجه به میزان شباهت به تصویر کلمه‌ی پرسشی به‌صورت نزولی رتبه‌بندی می‌کند. ازجمله مهم‌ترین معایب این روش می‌توان به این دو مسئله اشاره کرد: اولاً کاربر برای یافتن یک نمونه از کلمه مورد جستجو و مشخص کردن محدوده آن، باید سند را به‌صورت دستی مورد جستجو قرار دهد. دوماً وابستگی شدید به شیوه نگارش وجود دارد. دو مؤلفه مهم در چنین سیستمی، روش بازنمایی تصاویر و معیار شباهت بکار رفته می‌باشد.

در روش‌های جدیدتر، کلمه پرسشی به‌صورت یک رشته متنی در نظر گرفته شده است که به این دسته از روش‌ها مبتنی بر کلمه پرسشی متنی گفته می‌شود. این دسته از سیستم‌ها به‌عنوان کاربردی از بازیابی محتوایی تصویر شناخته می‌شوند. هر یک از سیستم‌ها با کلمه پرسشی تصویری یا متنی مزایای خاص خودشان را دارند. از طرفی سیستم‌هایی که با کلمه پرسشی متنی کار می‌کنند تمام مزایای سیستم‌های شناسایی دستخط مثل انعطاف‌پذیری در جستجوی هر کلمه‌ی کلیدی یا امکان استفاده از حجم زیاد داده‌های برچسب‌دار موجود را دارند. سیستم‌هایی که با کلمه پرسشی تصویری کار می‌کنند به‌طور خاص برای مواردی که داده‌های برچسب‌دار در دسترس نیستند یا با هزینه بالا به دست می‌آیند مناسب هستند. از طرف دیگر تحقیق [۳۸] نشان می‌دهد که رهیافت‌های کلمه پرسشی تصویری منحصراً بر اساس تطبیق تصویر کار می‌کنند؛ یعنی فاز آموزش با نظارت نیاز ندارد و می‌تواند دقت کافی موردنیاز سناریوهای عملی را به دست آورد. با توجه به مقایسه فوق انتظار می‌رود که روش‌های مبتنی بر کلمه پرسشی متنی با توجه به توانایی بازیابی فرم‌ها و نوشتارهای مختلف از کلمه ورودی، فراخوانی بهتری داشته باشد. مقایسه این دو دسته سیستم در جدول ۲-۳ آمده است.

نکته قابل‌ذکر دیگر این است که سیستم‌های مبتنی بر تطبیق الگو با کلمه پرسشی تصویری کار می‌کنند و برخی از سیستم‌های مبتنی بر یادگیری توانایی بازیابی مبتنی بر کلمه پرسشی متنی را دارا هستند. علاوه بر این، برای رهایی از مشکل آماده کردن کلمه پرسشی در این حالت در برخی از کارها از سنتز کردن کلمه پرسشی استفاده کرده‌اند. به‌این‌ترتیب که کلمه پرسشی را به‌صورت متنی دریافت می‌کنند و تصویر آن را ایجاد کرده و به‌عنوان ورودی به سیستم جستجوی کلمه ارائه می‌دهند[۵۶-۵۸]. سنتز کردن کلمه پرسشی به شیوه نگارش بستگی دارد و در صورت لزوم جستجوی اسناد با چندین نویسنده در اسناد دست‌نویس یا فونت‌های متنوع در اسناد چاپی باید چند مدل متفاوت برای سنتز کردن کلمات ایجاد شود. در زبان فارسی نیز روشی برای جستجوی کلمه روی اسناد چاپی ارائه شده که از سنتز کردن

---

[1] Image retrieval

کلمه پرسشی استفاده کرده است و تصویر کلمه را بر اساس فونت و اندازه‌ای که از سند مورد جستجو تشخیص داده است، می‌سازد[۱۲].

جدول ۲-۳: مقایسه روش‌های جستجوی کلمه با کلمه پرسشی متنی و تصویری

	کلمه پرسشی تصویری	کلمه پرسشی متنی
مزایا	عدم نیاز به دانش اولیه مستقل از الفبا و زبان عدم نیاز به تشخیص و آموزش	نیاز به ساخت مدل نوشتار برای هر سند در اکثر روش ها وابسته به الفبا و زبان نیاز به تشخیص و آموزش
معایب	نیاز به پیدا کردن دستی کلمه پرسشی به‌طور دستی عدم یافتن حالات متفاوت گرافیکی کلمه مورد جستجو عدم یافتن حالات متفاوت گرامری کلمه مورد جستجو حساسیت بیشتر به تغییرات دست‌نوشته و فونت هزینه محاسباتی بسیار زیاد تطبیق	وارد کردن کلمه پرسشی به‌طور متنی یافتن حالات متفاوت گرافیکی کلمه مورد جستجو یافتن حالات متفاوت گرامری کلمه مورد جستجو

یک موتور جستجوی کلمات با استفاده از کلمات پرسشی تصویری ارائه شده است[۵۹]. این روش بر اساس انتخاب نواحی مورد توجه[1] و مقایسه این نواحی با استفاده از یک روش تطبیق کشسان منسجم[2] کار می‌کند. این نواحی مورد توجه بر اساس عملگر مورفولوژی روی تصویر خاکستری به دست می‌آید. المان ساختاری که با استفاده از آن مورفولوژی انجام می‌شود، وابسته به زبان انتخاب می‌شود.

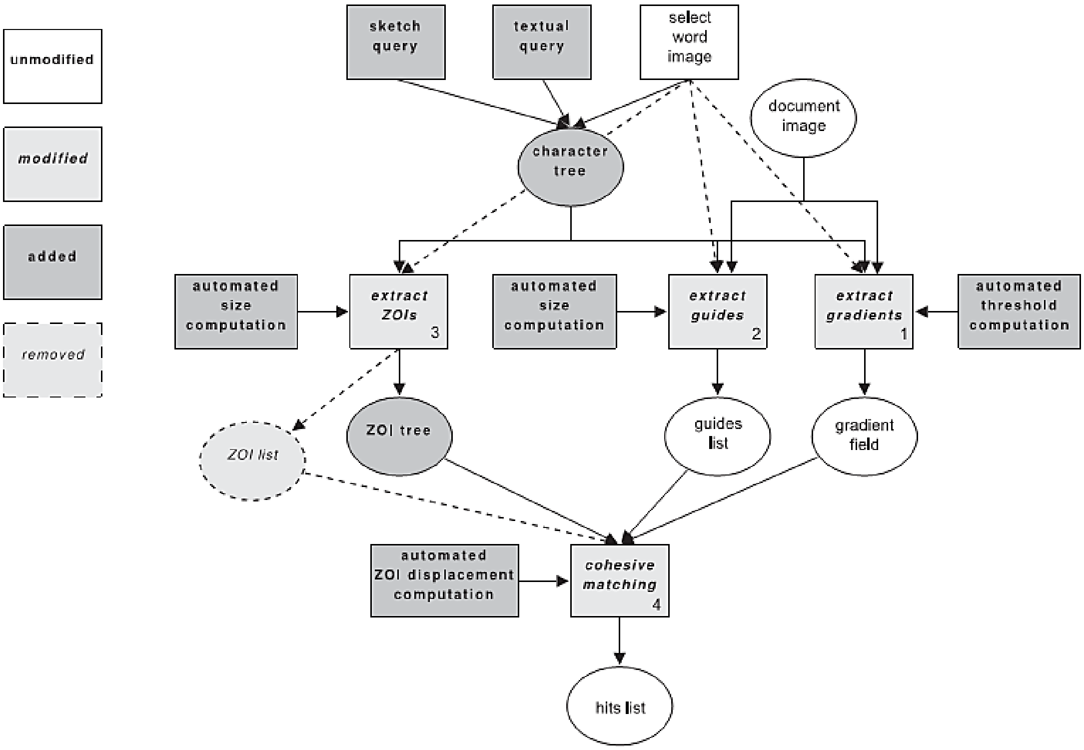
محققان در ادامه کار خود را توسعه داده‌اند و روشی مبتنی بر کلمه پرسشی متنی بر اساس روش ارائه شده قبلی پیشنهاد داده‌اند[۶۰]. در این روش برای هر سند مدلی تعریف شده که ترکیبی از یک الفبا، یک فهرست علائم[3] و یک گرامر می‌باشد. ساختار کلی سیستم در شکل ۲-۱۹ نشان داده شده است. البته در این روش بهبودی در انتخاب نواحی مورد توجه با استفاده از ویژگی‌های تفاضلی نیز داده شده است (شکل ۲-۲۰). نواحی مورد توجه بخش‌ها دارای اطلاعات بیشتر در تصویر هستند. مهم‌ترین نکته ذکرشده در این روش چنین بیان شده است که این تغییر باعث بهبود در جستجوی کلمات انگلیسی، چینی و ... شد اما روی زبان عربی به خاطر وجود نقاط و سرکش‌ها نتایج ضعیف‌تری را ارائه کرد.

---

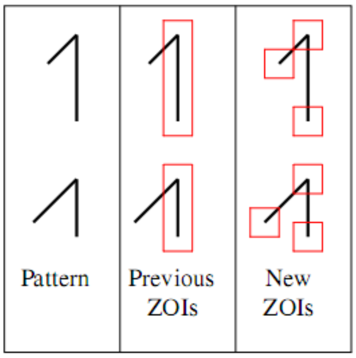
[1] Zone of Interest (ZOI)

[2] Cohesive elastic matching

[3] Glyph book



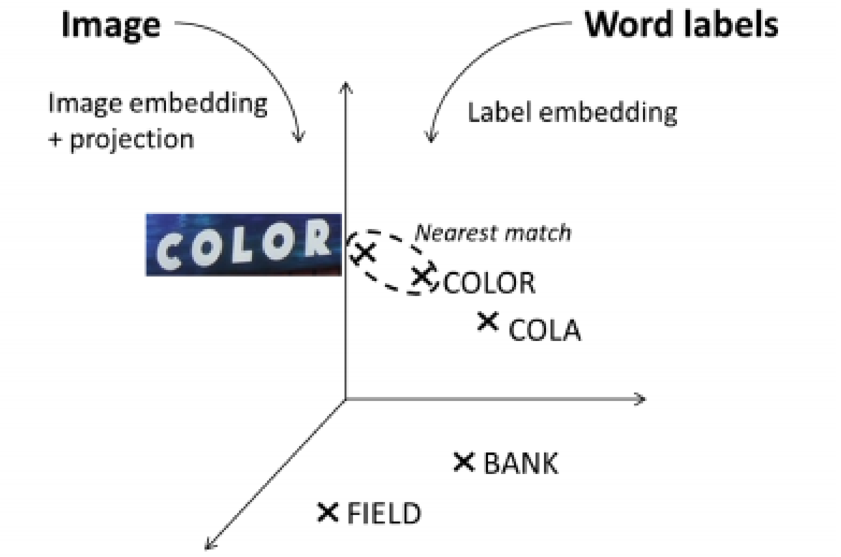
شکل ۲-۱۹: فرآیند کلی روش پیشنهادشده در [۶۰] با نمایش تغییرات نسبت به کار قبلی در [۵۹]



شکل ۲-۲۰: تفاوت انتخاب نواحی مورد توجه در [۵۹] و [۶۰]

اخیراً روشی مبتنی بر قطعه‌بندی جهت جستجوی کلمه ارائه شده که توانایی جستجو با کلمات پرسشی متنی و تصویری را داراست و می‌تواند برای تشخیص دست‌نوشته نیز مورد استفاده قرار گیرد [۶۱]. در این روش، تصویر هر کلمه و کلاس (رشته‌ی متنی) مربوط به آن در یک فضای برداری مشترک تعبیه می‌شود. در این فضا تصاویر و رشته‌ای که نشان‌دهنده‌ی کلمه پرسشی یکسان هستند نزدیک به یکدیگر قرار می‌گیرند که این اجازه را می‌دهد که عملیات بازشناسی به‌صورت مسئله نزدیک‌ترین همسایگی انجام شود. در این روش از ترکیب رده‌بندی مبتنی بر خصیصه، تعبیه برچسب و زیرفضای مشترک استفاده شده است.

در بینایی ماشین از تعبیه تصویر ورودی[1]، یعنی بازنمایی آن به فضای ویژگی، بسیار استفاده شده است. روش‌های متعددی مانند کدگذاری بر اساس پچ‌های تصویر، کاهش ابعاد، فشرده‌سازی و روش‌های مبتنی بر هسته از این نمونه هستند. اما به تعبیه برچسب کمتر پرداخته شده است. تعبیه برچسب یعنی برچسب کلاس در یک فضای جدید با قابلیت تمایز بالاتر نگاشت شوند به‌نحوی‌که کلاس‌های مشابه نزدیک به هم قرار گیرند. اخیراً، روشی بر اساس تعبیه کردن رشته (برچسب کلاس در بازشناسی و بازیابی کلمات تصویری) در زبان انگلیسی در [۶۲] پیشنهاد شده است. به‌این‌ترتیب که هر کلاس در فضای برداری خصیصه تعبیه می‌شود. به‌شرط اینکه تابع تعبیه مناسب انتخاب شود، یعنی کلاس‌های مشابه در فضای تعبیه بر اساس فاصله اقلیدسی نزدیک باشند، تعبیه برچسب روشی مؤثر برای اشتراک پارامترهای بین کلاسی است. هدف این است که یک بازنمایی مشترک[2] برای تصویر و رشته‌ی متنی کلمه وجود داشته باشد که مشخصه‌های استخراج‌شده از رشته و مشخصه‌های پیش‌بینی‌شده از تصویر در یک زیرفضای مشترک قابل‌مقایسه باشند. شمای کلی این سیستم در شکل ۲-۲۱ نمایش داده شده است.



شکل ۲-۲۱: شمای کلی سیستم [۶۲]

با توجه به قدرت و کارایی این روش، اخیراً محققان زیادی سعی در گسترش و بهبود این روش داشته‌اند. این روش با تغییراتی جزئی برای زبان یونانی نیز مورداستفاده قرار گرفته است[۶۳]. علاوه بر این

---

[1] Input embedding

[2] Joint representation

چارچوبی جهت استفاده از این روش با استفاده از پنجره‌گذاری نیز معرفی شده است که قابلیت مستقل بودن از قطعه‌بندی را نیز به آن می‌دهد[۶۴]. استفاده از شبکه عصبی کانولوشنی جهت استخراج هیستوگرام سلسله مراتبی کاراکترها که کلمه پرسشی را به‌صورت متنی و تصویری می‌پذیرد نیز توسعه جدیدی از این کار می‌باشد[۶۵].

فصل

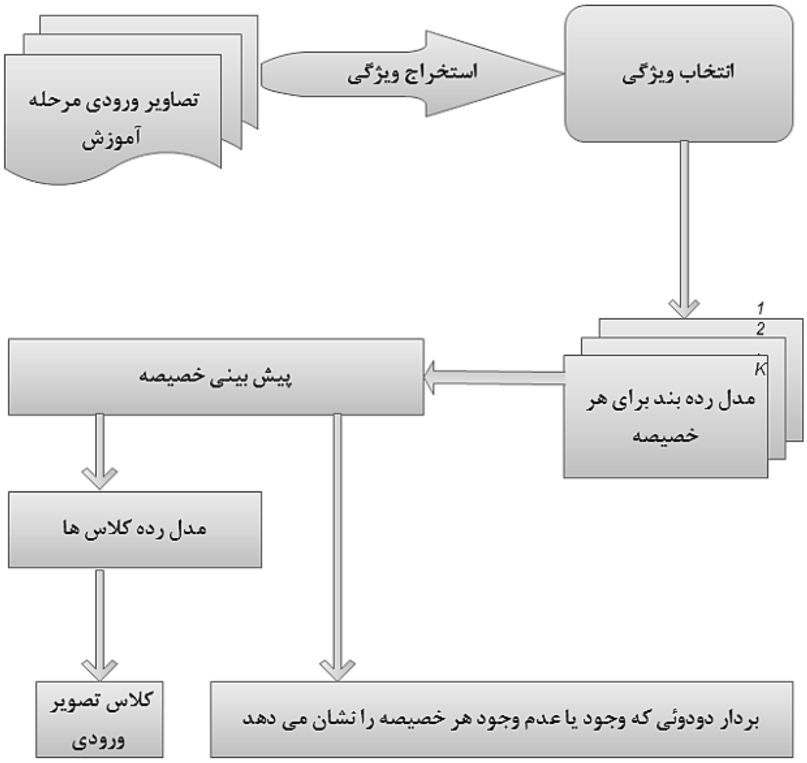
۳- روش پیشنهادی اول (مبتنی بر قطعه‌بندی کلمه)

۳-۱- مقدمه

با توجه به تعداد و تنوع زیاد کلاس‌های اشیاء در دنیای امروز، فراهم آوردن و برچسب‌گذاری تصاویر آموزشی برای تمامی کلاس‌ها در بسیاری از کاربردها غیرممکن به نظر می‌رسد. این تنوع کلاس‌ها در بسیاری از کاربردهای بینایی ماشین دیده می‌شود که شناسایی، جستجو و بازیابی کلمات دست‌نویس یکی از این نمونه‌های کاربردی است. برای حل این مشکل، رده‌بندی مبتنی بر خصیصه معرفی شده است. تفاوت این روش با روش‌های رده‌بندی کلاسیک در این است که به جای نام‌گذاری کلاس‌ها، آن‌ها را توصیف می‌کند. با توجه به اینکه این توصیفات می‌تواند بین کلاس‌های مختلف به اشتراک گذاشته شود، در مرحله آموزش یک هم‌افزایی ایجاد خواهد شد. علاوه بر این توانایی شناسایی و توصیف کلاس‌های دیده نشده هم به وجود خواهد آمد.

در این رهیافت تشخیص شیء با استفاده از توصیفات سطح بالای خاص انسان که خصیصه نامیده می‌شوند، انجام می‌گردد. این خصیصه‌ها می‌توانند ویژگی‌های ساختاری یا معنایی مانند شکل، رنگ و یا حتی اطلاعات جغرافیایی شیء باشند[۶۶, ۶۷]. اشتراک خصیصه‌ها بین چندین کلاس در مجموعه آموزشی علاوه بر این‌که می‌تواند باعث آموزش بهتر شود، برای شناسایی کلاس‌های دیده نشده نیز مورد استفاده قرار می‌گیرد. این روش سعی بر یادگیری به شیوه انسان دارد. انسان درحالی‌که کلاس اشیاء را بر اساس مثال‌های مشاهده‌شده یاد می‌گیرد، توانایی شناسایی کلاس‌های جدید را بر اساس توصیفات مشترک با کلاس‌های قبلی دارد. برای نمونه اگر در شناسایی حیوانات در یک مجموعه آموزشی سه خصیصه "بزرگ بودن"، " دارای رنگ خاکستری" و " دارای خرطوم" یاد گرفته شده باشد، چنانچه حیوانی که این سه خصیصه را به‌طور هم‌زمان داراست "فیل" نامیده شود، برای انسان قابل‌شناسایی خواهد بود، حتی اگر قبلاً نمونه‌ای از آن دیده نشده باشد. این روش به‌جای انتصاب یک نمونه به یک کلاس، آن را بر اساس خصیصه‌هایش توصیف می‌کند. خصیصه‌ها به‌عنوان یک روش بازنمایی میانی عمل می‌کنند که امکان به اشتراک‌گذاری پارامتر بین کلاس‌ها را ممکن می‌سازند. چنین رهیافتی خصوصاً در مواقعی که داده‌های آموزشی کم هستند، یادگیری را بهبود می‌بخشد. ساختار کلی این روش در شکل ۳-۱ آمده است.

اخیراً برای شناسایی و جستجوی کلمات در متون انگلیسی روش‌هایی ارائه شده است که از روش رده‌بندی مبتنی بر خصیصه استفاده می‌کند[۷, ۴۱, ۶۸- ۷۰]. ازآنجایی‌که این روش به‌عنوان روش پایه در این رساله مورداستفاده قرار گرفته است، در این فصل این روش پایه به همراه خصیصه‌هایی که برای زبان فارسی پیشنهاد و ارزیابی‌شده شرح داده می‌شود.



شکل ۳-۱: شمای کلی روش رده‌بندی مبتنی بر خصیصه

در این روش برای شناسایی و جستجو و بازیابی کلمات تصویری، آموزش یک بازنمایی مشترک برای رشته متنی (کلاس کلمه) و کلمه تصویری پیشنهاد شده است. با استفاده از این بازنمایی، مسئله شناسایی، جستجو و بازیابی کلمات تصویری به یک مسئله نزدیک‌ترین همسایه[1] ساده تبدیل می‌شود. در این روش رشته متنی به یک بردار d بعدی در فضای باینری تبدیل می‌شود که درواقع این بازنمایی نشان می‌دهد که آیا کاراکتری خاص در یک ناحیه مکانی خاص از رشته قرار گرفته است یا نه. این بردار به‌عنوان منبعی از خصیصه‌های کاراکتری مورداستفاده قرار می‌گیرد. جهت هماهنگ‌سازی تصویر کلمه با متن تصاویر کلمه نیز به یک فضای تفکیک‌کننده d بعدی نگاشت می‌شود که محتمل بودن وقوع یک کاراکتر خاص در یک ناحیه خاص را کد می‌کند (هیستوگرام هرمی کاراکترها[2]). با آموزش این خصیصه‌های کاراکتری داده‌های آموزشی بهتر مورد استفاده قرار می‌گیرند و بازشناسی و بازیابی کلمات دیده نشده در فرهنگ لغت آموزشی ساده و آسان خواهد بود.

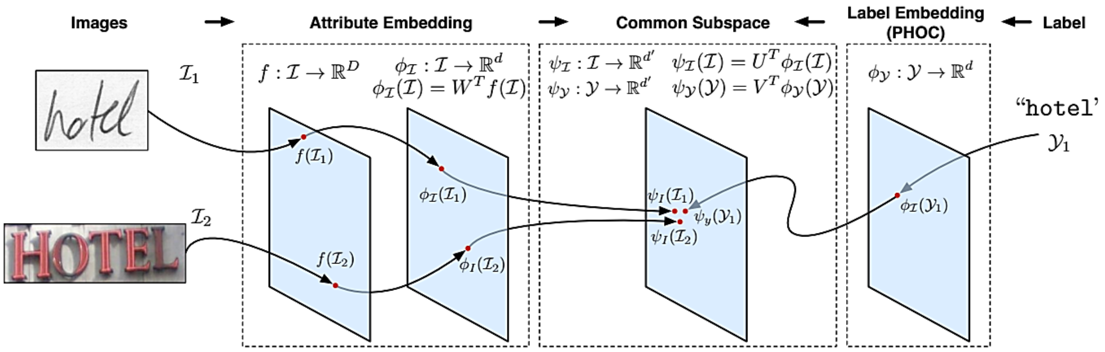
به دلیل این‌که بازنمایی d بعدی برای رشته متنی، باینری است و برای تصویر، امتیازات خصیصه‌ها است که باینری نیستند (درواقع امتیازات خروجی ماشین‌های بردار پشتیبان هستند) ، مقایسه آن‌ها به‌صورت مستقیم بهینه نیست و نیاز به کالیبره کردن وجود دارد. نهایتاً آموزش یک زیرفضای مشترک با ابعاد

---

[1] Nearest neighboar

[2] Pyramid Histogram of characters (PHOC)

کم با یک معیار مشترک بین هر دو بازنمایی پیشنهاد شده است. این کار دو مزیت اساسی دارد: اولاً مقایسه‌ی مستقیم تصویر و رشته متنی را معنی‌دار می‌کند و ثانیاً بازنمایی مثال‌هایی از تصاویر یک کلمه، نزدیک به هم قرار می‌گیرند. ساختار کلی روش معرفی شده در شکل ۳-۲ آمده است.



شکل ۳-۲: شمای کلی روش معرفی‌شده در [۷, ۴۱, ۶۸]

به‌منظور بازیابی و بازشناسی کلماتی که در زمان آموزش دیده نشده‌اند، لازم است بتوان امکان انتقال دانش بین نمونه‌های آموزش و آزمون ایجاد کرد. یکی از معروف‌ترین روش‌ها برای انجام این آموزش شات صفر[1] در بینایی ماشین استفاده از خصیصه‌های بصری[2] است[۶۶, ۶۷]. در این کار از خصیصه‌های کاراکتری برای انتقال اطلاعات بین نمونه‌های آموزش و آزمون استفاده شده است. هرچند ایده قطعه‌بندی کلمه به کاراکترها و آموزش در سطح کاراکتر قبلاً بارها استفاده شده است، اما این‌گونه رهیافت‌ها معمولاً وابسته به مدل‌های HMM با ویژگی‌هایی با طول متغیر هستند و بنابراین کارایی توسط این مدل‌ها محدود شده است. در روش پیشنهادی چارچوبی وسیع‌تر ارائه شده است که هیچ محدودیتی در انتخاب ویژگی‌ها یا روش‌های آموزش خصیصه‌ها وجود ندارد.

روش پیشنهادی وابسته به روش‌های تعبیه برچسب نیز می‌باشد که برچسب‌ها درون یک فضای متفاوت تعبیه می‌شوند و یک تابع سازگاری[۳] بین تصاویر و برچسب‌ها تعریف شده است[۷, ۶۸, ۶۹, ۷۱, ۷۲]. در [۶۹] روشی جهت بازشناسی متن با استفاده از رهیافت تعبیه متن و هرم مکانی کاراکترها[۴] پیشنهاد شده است. تفاوت اصلی این روش با روش ارائه‌شده در [۷, ۶۸] و چارچوب اصلی روش پیشنهادی در این رساله در نحوه تعبیه کردن برچسب‌ها می‌باشد. روش اول سعی دارد تا با استفاده از چهارچوب

---

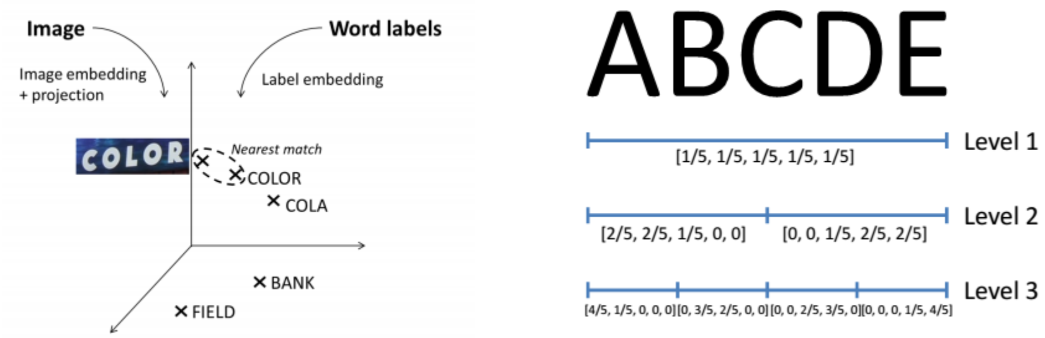
[1] Zero-shot learning

[2] Visual attribute

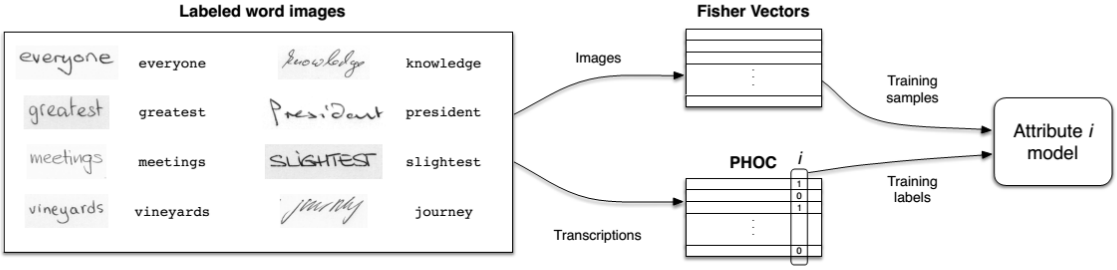
[۳] Compatibility function

[4] Spatial pyramid of characters (SPOC)

ماشین بردار پشتیبان بین بازنمایی تصویر و SPOCهای متناظر یک زیرفضای مشترک پیدا کند، درصورتی‌که در روش دوم PHOCها به‌عنوان خصیصه مورداستفاده قرار می‌گیرند و فقط سعی شده است که یک زیر فضای مشترک بین خصیصه‌ها و PHOCها پیدا شود. شمایی از این دو روش در شکل ۳-۳ و ۳-۴ آمده است.



شکل ۳-۳: روش تعبیه SPOC و مدل تعبیه برچسب در [۶۹].



شکل ۳-۴: فرآیند آموزش مدل خصیصه iام [۷]

۳-۲- بازنمایی کلمه مبتنی بر خصیصه

یکی از رایج‌ترین روش‌ها برای استفاده از آموزش با نظارت در مسئله جستجوی کلمه، آموزش مدل برای هر کلمه به‌صورت مجزا می‌باشد. با استفاده از مجموعه‌ای از نمونه‌های مثبت و منفی، برای هر کلمه کلیدی یک مدل آموزش داده می‌شود. در زمان آزمون احتمال کلمه ورودی بر اساس مدل‌های آموزش داده‌شده محاسبه می‌شود و به‌عنوان امتیاز مورداستفاده قرار می‌گیرد. این دسته از روش‌ها تعداد کلمات کلیدی قابل جستجو را محدود می‌کند. روشی نیز مبتنی بر شبکه عصبی شبه گسسته معرفی شده است [۲۹] که مدل توانایی وفق یافتن با هر کلمه پرسشی ورودی را به‌صورت برخط دارد. بنابراین این روش توانایی جستجو و بازیابی کلمات خارج از فرهنگ لغت را نیز دارا است، اما به‌هرحال برچسب‌ها در فاز آموزش استفاده نمی‌شوند.

از معایب دیگر این روش‌ها آموزش در سطح کلمه است، یعنی اطلاعات بین کلمات مشابه به اشتراک گذاشته نمی‌شود. به‌عنوان‌مثال اگر مدلی برای کلمه "بانک" آموزش داده شد، کلمه "باند" یک نمونه منفی برای آن در نظر گرفته می‌شود. درحالی‌که به اشتراک‌گذاری این اطلاعات مشترک بین کلاس‌های کلمات می‌تواند برای یادگیری یک بازنمایی تفکیک‌کننده خوب، مهم باشد. بنابراین استفاده از خصیصه‌ها یکی از روش‌های رسیدن به این هدف است. خصیصه‌ها، ویژگی‌های معنایی هستند که با استفاده از آن‌ها می‌توان تصاویر و دسته‌های آن‌ها را توصیف کرد. اخیراً، استفاده از رده‌بندی مبتنی بر خصیصه برای مسئله رده‌بندی و بازیابی تصاویر شهرت زیادی پیدا کرده است[۶۷, ۷۲, ۷۳].

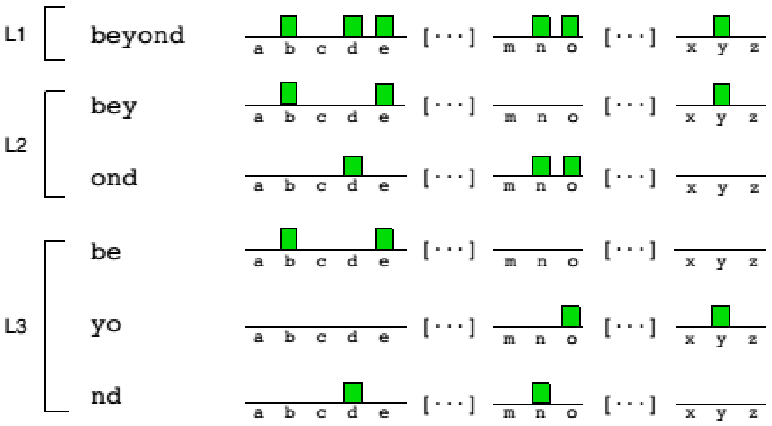
با توجه به ویژگی‌های ذکرشده در مورد خصیصه‌ها، استفاده از آن می‌تواند برای بازنمایی کلمات مناسب باشد. دلیل این ادعا را می‌توان به این صورت ذکر کرد که در این روش کلمات مختلف در فاز آموزش می‌توانند به تبادل اطلاعات بپردازند و این منجر به ایجاد امضاهای فشرده می‌شود. انتخاب این خصیصه‌ها فرآیندی وابسته به کاربرد است، بنابراین برای مسئله جستجوی کلمه باید طوری تعریف شوند که بتوانند به خوبی کلمات مختلف را از همدیگر متمایز کنند.

در زیر بخش بعدی ابتدا روش تعبیه برچسب PHOC که رشته متنی را در یک فضای برداری دودویی تعبیه می‌کند و پایه‌ی روش‌های پیشنهادی این بخش است را معرفی می‌کنیم. سپس روش‌های پیشنهادی مناسب برای زبان فارسی را شرح خواهیم داد. نهایتاً چگونگی استفاده از این بازنمایی جهت کد کردن تصویر کلمه را بیان خواهیم کرد.

۳-۲-۱- تعبیه برچسب متنی

۳-۲-۱-۱- روش PHOC [۷]

ساده‌ترین روش برای تعبیه رشته متنی ساخت یک هیستوگرام از کاراکترهاست که بر اساس تعداد کاراکترهای الفبای زبان ساخته می‌شود. ازآنجایی‌که این روش برای زبان انگلیسی معرفی شده است و ارقام صفر تا نه را نیز در بر می‌گیرد، برداری دارای ۳۶ بعد می‌تواند در نظر گرفته شود که هر بعد مشخص می‌کند که رشته متنی شامل کاراکتر خاصی می‌شود یا خیر. اما این روش بازنمایی متمایز کننده خوبی را برای کلمات ارائه نمی‌کند. به این دلیل که بسیاری از کلماتی که حروف یکسان با جایگشتی متفاوت از کاراکترها را دارند، مانند "پاکت" و "پاتک" و "تکپا"، در این بازنمایی یکسان نمایش داده خواهند شد. بنابراین یک نسخه هرمی از این هیستوگرام کاراکتری معرفی شده است. شکل ۳-۵ نمونه‌ای از این بازنمایی را در سه سطح اول نمایش داده است.



شکل ۳-۵: بازنمای PHOC در سه سطح ۱ و۲ و ۳ برای رشته متنی "beyond" [۷]

در این روش به جای بررسی وجود کاراکتری خاص در کل کلمه، این کار در نواحی متفاوتی از کلمه صورت پذیرفته است. در سطح دوم وجود هر کاراکتر در نیمه اول و دوم مورد بررسی قرار می‌گیرد و در سطح سوم کلمه به سه بخش تقسیم شده و برای هر سه بخش این بررسی انجام می‌شود و به همین ترتیب تا هر سطحی که لازم باشد پروسه می‌تواند ادامه پیدا کند. مثلاً در تحقیق گزارش شده در [۶۹] همین پروسه برای سطوح ۲ تا ۵ اجرا شده که هیستوگرامی با ۵۰۴ = ۳۶× (۲+۳+۴+۵) بعد تولید شده است. برای کد کردن رابطه بین کاراکترهای همسایه در این کار ۵۰ بایگرام[1] رایج انگلیسی نیز در سطح دوم مورد بررسی قرار گرفته است که باعث افزایش ۱۰۰ بعدی هیستوگرام و تبدیل شدن آن به ۶۰۴ بعد می‌شود. در صورت استفاده از این روش مشکل عدم متمایزکنندگی روش قبلی از بین خواهد رفت.

با توجه به ساختار PHOC برای تبدیل رشته متنی به بردار دودویی به تعیین نواحی هرمی که به هر کاراکتر اختصاص داده می‌شود، نیاز داریم. برای این منظور ناحیه اشغال‌شده نرمال شده توسط کاراکتر kام از یک رشته n کاراکتری به‌صورت یک فاصله به شرح زیر تعریف می‌شود:

$$Occ(k,n) = \left[\frac{k}{n},\frac{k+1}{n}\right] \tag{۳-۱}$$

که در این رابطه k از صفر شروع می‌شود. این اطلاعات از برچسب کلمات (رشته متنی) استخراج می‌شوند. از همین رابطه برای به دست آوردن ناحیه اشغال‌شده r در سطح l نیز استفاده می‌شود. اگر اشتراک این نواحی اشغال‌شده بزرگ‌تر یا مساوی ۵۰٪ بود، کاراکتر به این ناحیه اختصاص داده می‌شود.

$$\frac{Occ(k,n) \cap Occ(r,l)}{Occ(k,n)} \geq 0.5 \tag{۳-۲}$$

---

[1] bigram

در اولین مرحله دقیقاً از همین ساختار برای زبان فارسی استفاده شده و ساختار خصیصه پیشنهادی هیستوگرام هرمی کاراکترهای فارسی[۱] نام‌گذاری شده است.

۳-۲-۱-۲- تعبیه برچسب متنی خاص زبان فارسی

ازآنجایی‌که ساختار زبان فارسی ذاتاً پیوسته است و حروف اشکال متفاوتی را بسته به محل قرار گرفتن می‌توانند داشته باشند روش‌های دیگری برای تعبیه خصیصه‌ها در زبان فارسی پیشنهاد شده است که با ساختار نگارشی آن تناسب داشته باشد.

در روش اول پیشنهادی بر اساس تفاوت شکل حروف فارسی در موقعیت‌های مختلف داده شده است. در این روش با توجه به جدول ۱-۱ به جای در نظر گرفتن حروف در هر سطح از هرم، شکل حروف بر اساس موقعیت در نظر گرفته می‌شود. انگیزه اصلی این پیشنهاد را می‌توان شکل بسیار متفاوت برخی حروف در ۴ حالت موقعیتی عنوان کرد. با توجه به این‌که این وضعیت در ۲۴ حرف از حروف فارسی وجود دارد، می‌توان این ذهنیت را ایجاد کرد که مجزا در نظر گرفتن این حالات می‌تواند باعث آموزش بهتر مدل‌های خصیصه‌ها گردد.

در این حالت به جای ۳۳ حالت حروف زبان فارسی ("ا" و "آ" دو حرف متفاوت در نظر گرفته شده است) ۱۱۴ حالت در نظر گرفته می‌شود. این روش نیز برای ۴ سطح ۲ تا ۵ برای بازنمایی رشته‌های متنی فارسی استفاده شده است، بنابراین برداری به طول ۱۶۲۴ = ۱۱۶× (۵+۴+۳+۲) تولید خواهد شد. این روش بازنمایی و ساختار خصیصه ارائه‌شده، هیستوگرام هرمی کاراکترهای موقعیتی فارسی[۲] نامیده شده است.

بیشتر اشکال کاراکتری فارسی دارای دو قسمت اصلی هستند: بدنه اصلی کاراکتر، نقاط و سرکش‌ها. به دلیل اینکه در بسیاری از کاراکترها این بدنه یا نقاط مشترک هستند در روش دوم پیشنهادی سعی شد تا این قسمت‌ها به‌طور مجزا کد شوند. این وضعیت در حالات چهارگانه موقعیتی حروف نیز یکسان باقی می‌ماند. البته در برخی از حروف تفاوت‌هایی وجود دارد مانند حرف "ن" که حالت میانی و ابتدایی شبیه به دسته حروف "ب پ ت ث" دارد اما حالات تنها و آخر آن متفاوت است. جداول ۳-۱ و ۳-۲ به ترتیب دسته‌بندی نقاط و سرکش‌ها و بدنه‌ها را نشان می‌دهد. تعداد بدنه‌ها ۵۵ حالت مختلف و ۹ حالت نقاط و سرکش‌ها را نشان می‌دهد. در این صورت کل حالات برای ۶۱ دسته مختلف خواهد شد. بنابراین در

---

[1] Persian pyramid histogram of characters (PPHOC)

[2] Persian pyramid histogram of positional characters (PPHOPC)

صورت بازنمایی ۴ سطحی از سطوح ۲ تا ۵ تعداد ابعاد برابر ۹۱۰ = ۶۵× (۵+۴+۳+۲) خواهد بود. این روش، هیستوگرام هرمی نشانه‌های موقعیتی فارسی[1] نام‌گذاری شده است.

جدول ۳-۱: حروف دارای نقاط و سرکش مشترک

ردیف		نقاط و سرکش‌های مشترک
۱	آکولاد	آ ، ا
۲	تک نقطه بالا	خ ، خـ ، ـخـ ، ـخ ، ذ ، ـذ ، ز ، ـز ، ض ، ضـ ، ـضـ ، ـض ، ظ ، ـظ ، غ ، غـ ، ـغـ ، ـغ ، ف ، فـ ، ـفـ ، ـف ، نـ ، ـنـ
۳	تک نقطه پایین	ب ، بـ ، ـبـ ، ـب ، ج ، جـ ، ـجـ ، ـج
۴	دونقطه بالا	ت ، تـ ، ـتـ ، ـت ، ق ، قـ ، ـقـ
۵	دونقطه پایین	یـ ، ـیـ
۶	سه نقطه بالا	ث ، ثـ ، ـثـ ، ـث ، ژ ، ـژ ، ش ، شـ ، ـشـ ، ـش
۷	سه نقطه پایین	پ ، پـ ، ـپـ ، ـپ ، چ ، چـ ، ـچـ ، ـچ
۸	سرکش کاف	ک ، کـ ، ـکـ ، ـک
۹	سرکش گاف	گ ، گـ ، ـگـ ، ـگ

۳-۲-۲- آموزش خصیصه‌ها بر اساس برچسب‌های تعبیه‌شده

همان‌طور که قبلاً ذکر شد، هیستوگرام‌های هرمی می‌تواند به‌عنوان برچسب‌های خصیصه‌ها عمل کنند. این خصیصه‌ها به خوبی توانایی متمایز کردن کلمات از هم را دارا هستند. اگر این خصیصه‌ها بر اساس داده‌هایی که از منابع و یا نویسندگان متفاوت جمع آوری شده است آموزش داده شود، مدل‌هایی که به دست می‌آیند نسبت به تغییرات ظاهری و دستخط مقاوم خواهند بود.

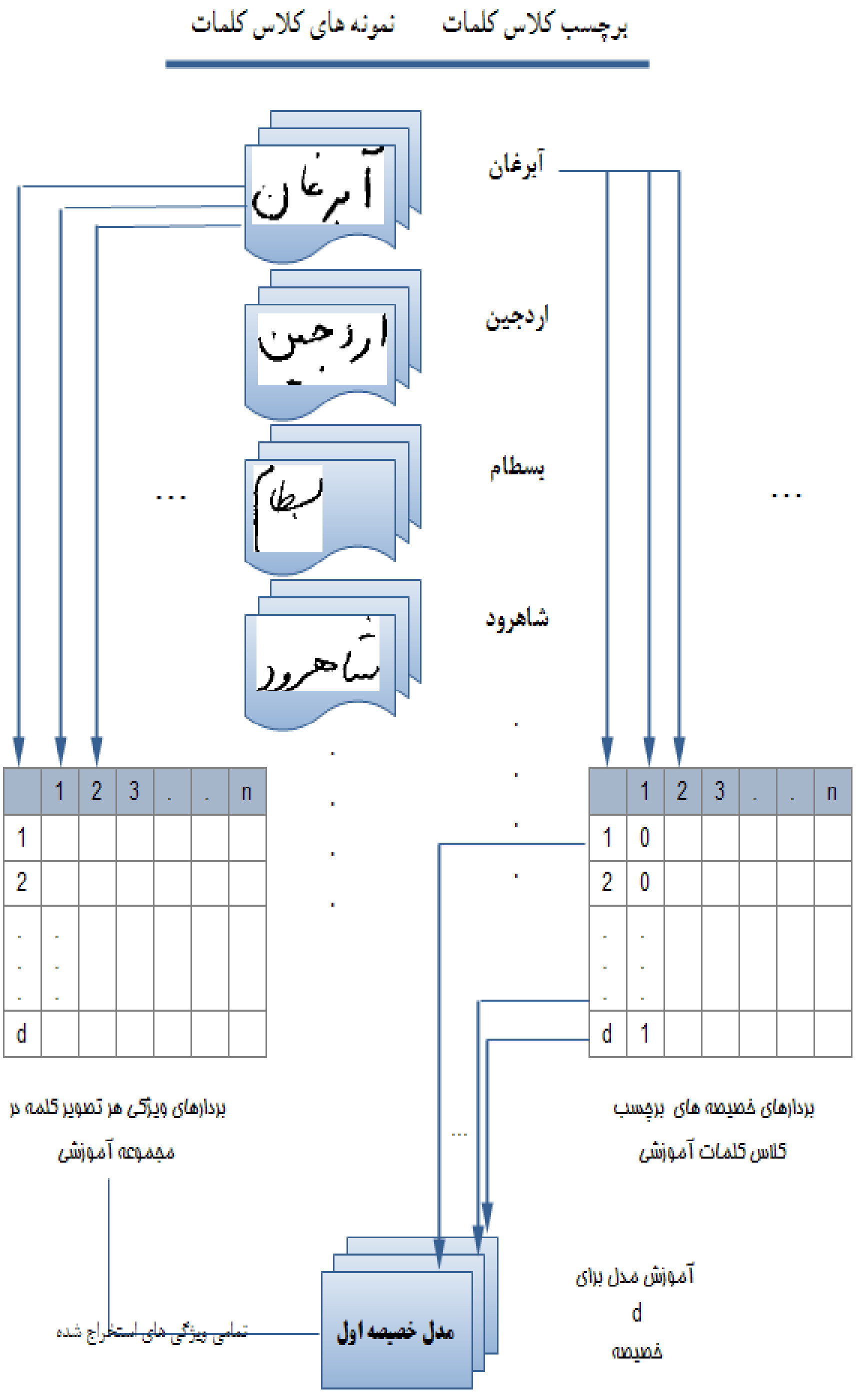
در روش پیشنهادی، برای آموزش این خصیصه‌ها از بردارهای ماشین پشتیبان خطی استفاده شده است. ابتدا بردار ویژگی تصاویر کلمات استخراج می‌شود. این بردارهای ویژگی به همراه برچسب‌های تعبیه‌شده، برای مثال PHOCها، برای آموزش مدل‌های خصیصه بر اساس ماشین بردار پشتیبان استفاده می‌شوند. ساختار کلی روش آموزش در شکل ۳-۶ آمده است.

---

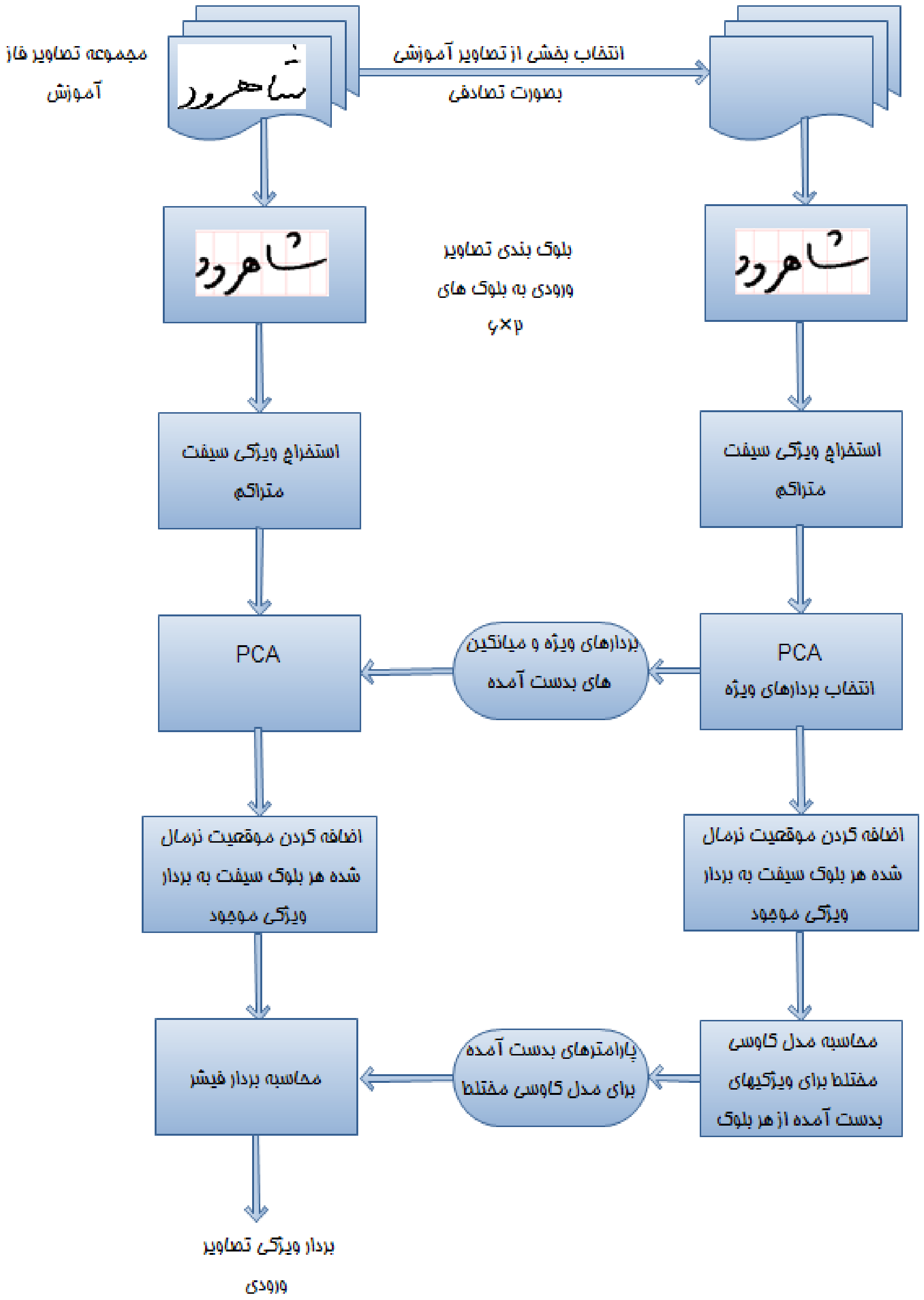
[1] Persian pyramid histogram of positional signs (PPHOPS)

جدول ۳-۲: حروف دارای بدنه مشترک

بدنه‌های مشترک	ردیف	بدنه‌های مشترک	ردیف	بدنه‌های مشترک	ردیف
بـ ، پـ ، تـ ، ثـ ، نـ ، یـ	۳	ب ، پ ، ت ، ث	۲	آ ، ا	۱
ج ، چ ، ح ، خ	۶	ـب ، ـپ ، ـت ، ـث	۵	ـبـ ، ـپـ ، ـتـ ، ـثـ ، ـنـ ، ـیـ	۴
ـج ، ـچ ، ـح ، ـخ	۹	جـ ، چـ ، حـ ، خـ	۸	ـجـ ، ـچـ ، ـحـ ، ـخـ	۷
ی	۱۲	ـد ، ـذ	۱۱	د ، ذ	۱۰
س ، ش	۱۵	ـر ، ـز ، ـژ	۱۴	ر ، ز ، ژ	۱۳
ـس ، ـش	۱۸	ـسـ ، ـشـ	۱۷	سـ ، شـ	۱۶
صـ ، ضـ	۲۱	ـصـ ، ـضـ	۲۰	ص ، ض	۱۹
طـ ، ظـ	۲۴	ط ، ظ	۲۳	ـص ، ـض	۲۲
عـ ، غـ	۲۷	ـعـ ، ـغـ	۲۶	ع ، غ	۲۵
ـف	۳۰	ف	۲۹	ـع ، ـغ	۲۸
فـ ، قـ	۳۳	ـق	۳۲	ق	۳۱
کـ ، گـ	۳۶	ک ، گ	۳۵	ـفـ ، ـقـ	۳۴
ل	۳۹	ـک ، ـگ	۳۸	ـکـ ، ـگـ	۳۷
ـل	۴۲	ـلـ	۴۱	لـ	۴۰
ـمـ	۴۵	مـ	۴۴	م	۴۳
ـن	۴۸	ن	۴۷	ـم	۴۶
ـهـ	۵۱	هـ	۵۰	ه	۴۹
ـو	۵۴	و	۵۳	ـه	۵۲
		ـآ ، ـا	۵۶	ـی	۵۵



شکل ۳-۶: ساختار کلی آموزش مدل خصیصه‌ها

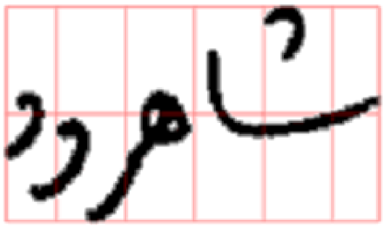


شکل ۳-۷: روند کلی استخراج ویژگی در روش پیشنهادی

برای استخراج ویژگی‌های تصاویر کلمات از بردار فیشر استفاده شده است. بردار فیشر یک روش کدگذاری است که در مسائل بازیابی و بازشناسی تصویر استفاده می‌شود. بردار فیشر می‌تواند به‌صورت سبد بصری کلماتی در نظر گرفته شود که نه‌تنها تعداد کلمات، بلکه آماره‌های سطح بالاتر را نیز کد

می‌کند. به‌طور خلاصه در زمان آموزش توصیفگرهای سطح پایین (در این کار سیفت متراکم استفاده شده است) استخراج می‌شود. این ویژگی‌ها جهت آموزش مدل‌های مخلوط گاوسی[1] با تعداد k هسته مورداستفاده قرار می‌گیرند. روند کلی استخراج ویژگی در شکل ۳-۷ آمده است.

یکی از مسائل مهم در رابطه با روش‌های کدگذاری تصویر مانند بردار فیشر این است که نمی‌توانند مکان ویژگی‌ها که برای توصیف کلمات بسیار مهم هستند، را حفظ کنند. یکی از رهیافت‌های مناسب جهت اضافه کردن اطلاعات مکانی، استفاده از هرم‌های مکانی[2] است[۷۴]. به جای استفاده از ویژگی‌های کل تصویر، ابتدا تصویر به k ناحیه مختلف تقسیم و ویژگی‌های این نواحی در کنار هم قرار داده می‌شود تا یک بردار ویژگی کلی برای تصویر کلمه تشکیل شود. مشکلی که در این حالت وجود خواهد داشت، طول متفاوت کلمات است (شکل ۳-۸ را ببینید). راهکار دیگری که این مشکل را حل می‌کند و با بردار فیشر هم به‌خوبی کار می‌کند در مرجع [۷۵] پیشنهاد شده است.



شکل ۳-۸: یک نمونه از قطعه‌بندی تصویر کلمه (کلمه "شاهرود")

با اضافه کردن مختصات طول و عرض نرمال شده، توصیفگرهای سیفت تصاویر غنی‌تر می‌شوند. مدل مخلوط گاوسی فقط روی ویژگی‌های سیفت آموزش داده نمی‌شود بلکه از نسخه غنی‌شده آن استفاده می‌کند. به همراه توصیفگر بردار فیشر تصویر مکان ویژگی‌های درون کلمه نیز کد می‌شود. مختصات عمودی و افقی بین ۵,۰- تا ۵,۰+ نرمال می‌شوند. تصویر کلمه در کادر محیطی قرار می‌گیرد (شکل ۳-۸) که مرکز این کادر مبدأ مختصات و حداقل و حداکثر مقادیر طول و عرض ۵,۰- و ۵,۰+ خواهد بود.

برای این‌که نتایج این بازنمایی بهتر شود به جای آموزش مدل مخلوط گاوسی با استفاده از ویژگی‌های غنی‌شده کل تصویر، این کار به‌صورت ناحیه‌ای انجام شده است. یعنی برای هر ناحیه از تصویر یک مدل مخلوط گاوسی از بردار ویژگی‌ها به دست می‌آید و مدل‌های به‌دست‌آمده با هم ترکیب شده و وزن‌های آن‌ها مجدداً به یک نرمال‌سازی می‌شوند.

---

[1] Gussian Mixture Model (GMM)

[2] Spatial Pyramids

۳-۲-۳- زیرفضای مشترک بین برچسب‌ها و خصیصه‌ها

در بخش قبلی یک بازنمایی مبتنی بر خصیصه برای تصاویر کلمات معرفی شد. هرچند که این بازنمایی نسبت به تغییرات ظاهری مقاوم است، اما در هنگام مقایسه کلمات مختلف نیاز به مراقبت زیادی دارد. دلیل این مسئله را می‌توان در امتیازات خصیصه‌ها برای کلمات مختلف جستجو کرد. هرچند به خاطر مشابهت موجود، مقایسه مستقیم خصیصه‌های تعبیه‌شده تصاویر و برچسب‌های متنی تعبیه‌شده نتایج قابل قبولی را نشان می‌دهد اما خوش‌تعریف نیست. بااینکه هر دو در یک فضا با ابعاد یکسان قرار می‌گیرند اما برچسب‌های متنی تعبیه‌شده دودویی هستند درحالی‌که امتیازات خصیصه‌های به این صورت نبوده و محدوده‌های متفاوتی دارند. بنابراین کالیبره کردن این امتیازات و برچسب‌های تعبیه‌شده ضروری به نظر می‌رسد.

یکی از رایج‌ترین روش‌ها جهت کالیبره کردن امتیازات ماشین بردار پشتیبان، مقیاس‌بندی پلاتس[1] است که برای به دست آوردن احتمالات کالیبره شده یک تابع سیگموئید روی امتیازات خروجی استفاده می‌شود:

$$P(y = 1|s) = \frac{1}{1 + e^{\alpha s + \beta}} \quad (۳\text{-}۳)$$

که α و β می‌تواند با استفاده از تخمین حداکثر درستنمایی[2] به دست آید.

یکی از مشکلات عمده این‌گونه روش‌ها در نظر نگرفتن همبستگی بین خصیصه‌های مختلف می‌باشد. ازآنجایی‌که به دلیل شیوه کدگذاری چند سطحی، همبستگی زیادی بین خصیصه‌های پیشنهادی وجود دارد، حذف همبستگی بین خصیصه‌ها حائز اهمیت بسیاری است. به این دلیل از کالیبراسیون توأم که همبستگی بین خصیصه‌ها را بهتر می‌تواند استخراج کند، استفاده شده است. چون رگرسیون عادی همبستگی بین مقادیر خود بردارهای خصیصه را در نظر نمی‌گیرد، رگرسیون زیرفضای مشترک که منجر به فرمول‌بندی معادل تحلیل همبستگی کانونی[3] می‌شود پیشنهاد شده است.

فرض کنید $I = \{I_n, n = 1, \dots, N\}$ مجموعه N تصویر در مجموعه آموزشی و $y = \{y_n, n = 1, \dots, N\}$ برچسب‌های مربوط به آن‌ها باشند. در این صورت $A = \phi_I(I) \in \Re^{d \times N}$ نشان دهنده N تصویر تعبیه‌شده در فضای d بعدی خصیصه‌ها است و N برچسب تعبیه‌شده در فضای d بعدی به‌صورت

---

[1] Platts Scaling

[2] Maximum Likelihood Estimation (MLE)

[3] Canonical Correlation Analysis (CCA)

$B = \phi_y(y) \in \{0,1\}^{d\times N}$ می‌باشد. بنابراین روش مستقیم برای ارتباط دادن A و B تعریف یک تابع فاصله به‌صورت زیر است:

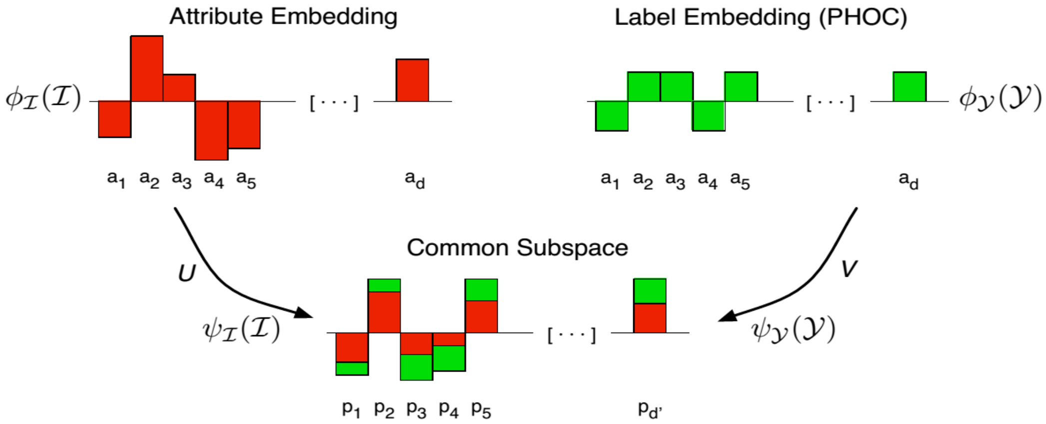
$$F(I_i, y_i; P) = ||P^T \phi_I(I_i) - \phi_y(y_i)||_2^2 \ , \qquad P \in \Re^{d\times d} \tag{۴-۳}$$

برای کمینه کردن فاصله بین تمام نمونه‌ها و برچسب‌های آن‌ها می‌توان از رابطه زیر استفاده نمود:

$$\mathrm{argmin}_P \sum_i^N \frac{1}{2} F(I_i, y_i; P) + \frac{1}{2}\Omega(P) = \mathrm{argmin}_P \frac{1}{2}||P^T A - B||_F^2 + \frac{1}{2}\Omega(P) \tag{۵-۳}$$

که $\Omega(P) = \alpha||P||_F^2$ عبارت منظم سازی و α وزن آن می‌باشد. در این مورد برابر یک مسئله رگرسیون معمولی است و $P = (AA^T + \alpha I)^{-1} AB^T$ که I ماتریس همانی است. چون d تعداد خصیصه‌ها می‌باشد، که زیاد هم نیستند و این مسئله فقط یک بار در زمان آموزش باید حل شود، حل آن بسیار سریع است.

همان‌طور که قبلاً هم ذکر شد، این فرمولاسیون فقط همبستگی بین امتیازات خصیصه‌ها را در نظر می‌گیرد لذا همبستگی بین خود خصیصه لحاظ نمی‌شود. برای این‌که هر دو مسئله در نظر گرفته شود تغییراتی در آن انجام شده است[۷] که هر دو را در یک زیرفضای مشترک با ابعاد d' نگاشت می‌کند (شکل ۳-۹).



شکل ۳-۹: نگاشت امتیازات پیش‌بینی‌شده خصیصه‌ها و برچسب‌های تعبیه‌شده در یک زیرفضای دارای همبستگی بیشتر[۷]

یک تابع فاصله جدید به‌صورت زیر تعریف شده است:

$$\hat{F}(I_i, y_i; U, V) = ||\psi_I(I_i) - \psi_y(y_i)||_2^2 \tag{۶-۳}$$

که در آن $\psi_I(I) = U^T\phi_I(I)$ و $\psi_y(y) = V^T\phi_y(y)$ دو تابع تعبیه خطی هستند و از ماتریس‌های نگاشت $U, V \in \Re^{d\times d'}$ برای تعبیه $\phi_I(I)$ و $\phi_y(y)$ در زیرفضای مشترک استفاده می‌کند. سپس، مشابه با مورد قبلی، هدف کمینه کردن فاصله بین همه نمونه‌ها و برچسب‌های آن‌ها است.

$$\mathrm{argmin}_{U,V}\sum_{i}^{N}\frac{1}{2}\hat{F}(I_i, y_i; U, V) + \frac{1}{2}\Omega(U) + \frac{1}{2}\Omega(V) =$$
$$\mathrm{argmin}_{U,V}\frac{1}{2}||U^TA - V^TB||_F^2 + \frac{1}{2}\Omega(U) + \frac{1}{2}\Omega(V) \qquad (۷\text{-}۳)$$

$$\psi_I(I)\psi_I(I)^T = I$$
$$\psi_y(y)\psi_y(y)^T = I$$

که در آن محدودیت تعامد اطمینان می‌دهد که حل کردن آن آسان نخواهد بود.

با استفاده از ضرایب لاگرانژ نسبت به U و V مشتق گرفته شده و برابر صفر قرار داده می‌شود و به تساوی‌های زیر می‌رسیم.

$$\lambda(AA^T + \alpha I)u_k = AB^Tv_k \qquad (۸\text{-}۳)$$

$$\lambda(BB^T + \alpha I)v_k = BA^Tu_k \qquad (۹\text{-}۳)$$

که v_k و u_k ، k امین ستون ماتریس U و V است و λ به خاطر ضرایب لاگرانژ ظاهر می‌شود. وقتی u^k را حل می‌کنیم به مسئله مدار ویژه تعمیم‌یافته زیر می‌رسیم:

$$AB^T(BB^T + \alpha I)^{-1}BA^Tu_k = \lambda^2(AA^T + \alpha I)u_k \qquad (۱۰\text{-}۳)$$

K بردار ویژه تعمیم‌یافته ابتدایی از K ستون ماتریس نگاشت U می‌باشد که اجازه می‌دهد تا ابعاد نهایی d' انتخاب شود. فرآیندی مشابه برای به دست آوردن ماتریس نگاشت V نیز انجام می‌شود. این مدل با استفاده بیشتر از ارتباط بین داده‌ها مدل بهتری نسبت به رگرسیون در این کار ایجاد می‌کند. علاوه بر این، در این مدل می‌توان ابعاد خروجی را کنترل نمود و کاهش ابعاد را انجام داد.

همچنین این روابط برای حل مسئله تحلیل همبستگی کانونی نیز مورد استفاده قرار می‌گیرند، زیرا تلاش در یافتن نگاشتی است که همبستگی در یک زیرفضای مشترک را حداکثر نماید. تحلیل همبستگی کانونی ابزاری است برای استفاده از اطلاعاتی که از منابع مختلف در دسترس قرار می‌گیرند و برای کاربردهای بازیابی [۷۶] و خوشه‌بندی [۷۷] مورد استفاده قرار می‌گیرند. علاوه بر این تحلیل همبستگی

کانونی برای مرتبط کردن توصیفات تصاویر و برچسب‌های آن نیز مورد استفاده قرار گرفته است که برای کاربردهای بازیابی مزایای زیادی دارد. به این دلیل که ممکن است رابطه بین امتیازات خصیصه‌ها و خصیصه‌های دودویی خطی نباشد، مدل مبتنی بر هسته (تحلیل همبستگی کانونی دارای هسته[1]) می‌تواند منجر به بهینه‌سازی بیشتری شود.

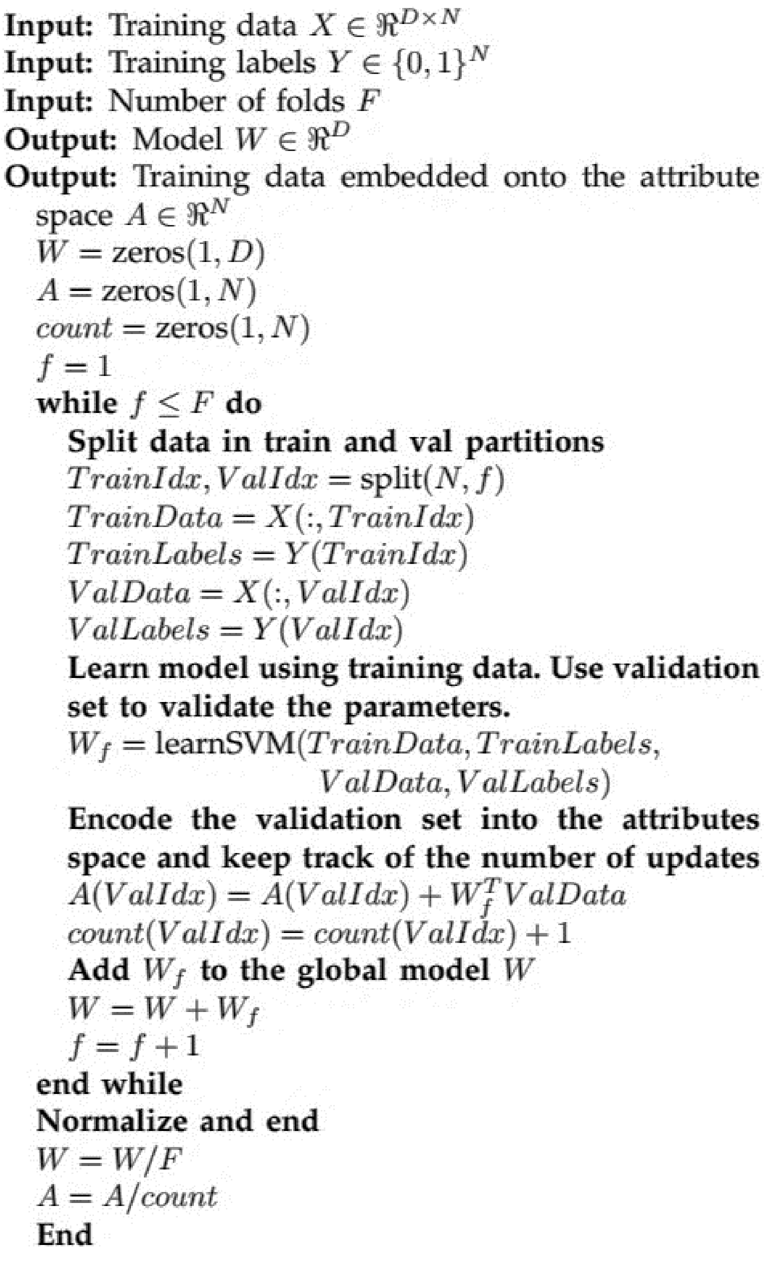
یکی از مشکلات آموزش فضای خصیصه و زیرفضای مشترک در دو گام مجزا نیاز به حجم زیاد داده‌های آموزشی است. به این دلیل که داده‌هایی که برای آموزش زیرفضای مشترک استفاده می‌شوند باید متفاوت با داده‌هایی باشند که برای آموزش فضای خصیصه مورد استفاده قرار می‌گیرند. به خاطر اینکه اگر برای هر دو مرحله داده یکسان استفاده کنیم امتیازات ماشین بردار پشتیبان بیش برازش[2] خواهند شد (اکثراً بسیار نزدیک به ۱ یا ۱- می‌شوند) و در نتیجه زیرفضای آموزش‌دیده با این شرایط به‌شدت بایاس می‌شوند و منجر به نتایج ضعیف خواهند شد. اگر مجموعه داده بزرگی داشته باشیم می‌توان آن را به دو بخش تقسیم کرد و بخشی را برای آموزش زیرفضای مشترک و بخشی را جهت آموزش فضای خصیصه مورد استفاده قرار داد. اما با این روش داده‌های آموزشی به‌خوبی مورد استفاده قرار نمی‌گیرند.

جهت غلبه بر مشکل ذکرشده، گونه‌ای از بَگینگ[3] استفاده شده است. برای این کار داده‌های آموزشی به چندین پوشه‌ی[4] آموزش و ارزیابی تقسیم‌بندی می‌شود، به‌طوری‌که داده‌های ارزیابی و آموزش در هر دسته از همدیگر مجزا هستند اما ممکن است در دسته‌های متفاوت اشتراک داشته باشند. برای هر دسته یک مدل با استفاده از داده‌های مربوط به آن آموزش داده می‌شود و این مدل برای محاسبه امتیاز روی داده‌های ارزیابی استفاده می‌شود. بنابراین امتیازات روی داده‌های ارزیابی تقریباً بایاس شده نیستند. در دسته‌های مختلف امتیازات ارزیابی جمع می‌شوند و برای هر نمونه تعداد دفعاتی که استفاده شده است ثبت می‌گردد. در انتهای فرآیند مدل کلی با میانگین‌گیری از همه این مدل‌های محلی آماده می‌شود. با نرمال‌سازی امتیازات هر نمونه بر اساس تعداد باری که استفاده شده است مجموعه آموزشی از امتیازاتی که بایاس نیستند فراهم می‌شود که می‌تواند برای آموزش بدون مشکل زیرفضای مشترک مورد استفاده قرار گیرد(الگوریتم ۳-۱) .

الگوریتم ۳-۱: آموزش مدل خصیصه‌ها با استفاده از بگینگ

---

[1] Kernelized Canonical Correlation Analysis (KCCA)
[2] Overfit
[3] bagging
[4] fold



```
Input: Training data $X \in \Re^{D \times N}$
Input: Training labels $Y \in \{0,1\}^N$
Input: Number of folds $F$
Output: Model $W \in \Re^D$
Output: Training data embedded onto the attribute space $A \in \Re^N$
  $W = \text{zeros}(1, D)$
  $A = \text{zeros}(1, N)$
  $count = \text{zeros}(1, N)$
  $f = 1$
  while $f \leq F$ do
    Split data in train and val partitions
    $TrainIdx, ValIdx = \text{split}(N, f)$
    $TrainData = X(:, TrainIdx)$
    $TrainLabels = Y(TrainIdx)$
    $ValData = X(:, ValIdx)$
    $ValLabels = Y(ValIdx)$
    Learn model using training data. Use validation set to validate the parameters.
    $W_f = \text{learnSVM}(TrainData, TrainLabels, ValData, ValLabels)$
    Encode the validation set into the attributes space and keep track of the number of updates
    $A(ValIdx) = A(ValIdx) + W_f^T ValData$
    $count(ValIdx) = count(ValIdx) + 1$
    Add $W_f$ to the global model $W$
    $W = W + W_f$
    $f = f + 1$
  end while
  Normalize and end
  $W = W/F$
  $A = A/count$
  End
```

۳-۳- نتایج تجربی و ارزیابی

با توجه به روند کلی مطرح‌شده که در شکل ۳-۷ نیز آمده است، برای آموزش مدل مخلوط گاوسی، ۲۵۰ تصویر به‌صورت تصادفی و غیرتکراری از تصاویر پایگاه داده انتخاب می‌شوند. سپس تصاویر کلمات طوری چیده می‌شوند که هر کلمه توسط کادری محاط شود. این کادر به ۲ ناحیه افقی و ۶ ناحیه عمودی برابر تقسیم می‌شود (شکل ۳-۸) و توصیفگرهای سیفت متراکم از هرکدام از ۱۲ بلوک تصویر استخراج می‌شود. بردار ویژگی با ابعاد ۱۲۸×n برای هر بلوک از تصویر ایجاد می‌شود. درواقع n برابر با مجموع تمام نقاطی است که در آن ویژگی استخراج شده است. با استفاده از آنالیز اجزاء اصلی تعداد

ویژگی‌ها به ۶۲ ویژگی کاهش داده می‌شود و سپس با اضافه کردن مختصات طول و عرض نرمال شده، غنی می‌شوند و برداری به طول ۶۴ را ایجاد می‌کنند. برای هر کدام از نواحی ۱۲ گانه مدل مخلوط گاوسی با استفاده از این بردارهای ویژگی غنی‌شده آموزش داده می‌شود. در هر مدل مخلوط گاوسی از ۱۶ مدل گاوسی استفاده شده است، در نتیجه ۱۹۲ = ۱۲ × ۱۶ دسته پارامتر خواهیم داشت. سپس بردار فیشر استخراج می‌شود که این بردار طولی برابر ۲۴۵۷۶ = ۲ × ۶۴ × ۱۲ × ۱۶ خواهد داشت.

با استفاده از روش تعبیه برچسب، برچسب‌های مجموعه تصاویر آموزشی در فضای هیستوگرام هرمی تعبیه می‌شود. در هر تکرار ۸۰ درصد از داده‌ها برای آموزش و ۲۰ درصد باقیمانده برای اعتبارسنجی مورد استفاده قرار می‌گیرد. آموزش توسط ماشین بردار پشتیبان توسط روش بگینگی که قبلاً شرح داده شد، انجام می‌گیرد. مقادیر اوزان نهایی از میانگین‌گیری نتایج محلی به دست می‌آید و در نهایت برای امتیاز مشخصه‌ها برای داده‌های اعتبار سنجی و نمونه‌های آزمون از رابطه زیر محاسبه می‌گردد.

$$Score = W^T \times FV \qquad (۱۱\text{-}۳)$$

در این رابطه w وزن‌های نهایی به‌دست‌آمده و FV بردار ویژگی تصویر (بردار فیشر) است. این آموزش برای خصیصه‌هایی که تعداد کمتر از ۲ نمونه دارند انجام نمی‌شود.

برای حل این مشکل از روش تحلیل همبستگی کانونی(CCA) استفاده شده است. با استفاده از این روش رتبه مشخصه‌ها و برچسب‌های PHOC در یک فضای برداری مشترک تعبیه می‌شوند که در این فضا همبستگی بین آن‌ها ماکزیمم است. در این فضا رتبه مشخصه و برچسب PHOC که مربوط به یک کلمه هستند، نزدیک به یکدیگر قرار دارند. به‌عبارت‌دیگر در این فضا تصاویر و رشته‌ی متنی که نشان‌دهنده‌ی کلمه یکسان هستند نزدیک به یکدیگرند و این اجازه را می‌دهد که عملیات بازشناسی به‌صورت مسئله نزدیک‌ترین همسایگی انجام شود.

جهت آموزش زیرفضای مشترک از تجزیه‌وتحلیل همبستگی کانونی استفاده شده است، زیرا ابزاری است که توصیفگرهای تصویر و برچسب‌های متناظرشان را به هم مرتبط می‌کند. در این روش با تجزیه‌وتحلیل هم‌زمان مجموعه‌ها و با شناسایی و مشخص کردن عناصری از یک مجموعه متغیر با بیشترین وابستگی و ارتباط با عناصر مجموعه متغیر دیگر، روابط آماری مستقل موجود بین دو مجموعه متغیر را مورد آزمون قرار می‌دهد. در این کار هدف از این مرحله دستیابی به دو ماتریس نگاشت Ux و Vy است که توسط این دو ماتریس، داده‌ها از فضای مشخصه و برچسب به فضای برداری مشترک نگاشت می‌شوند.

اگر X مجموعه‌ی N تصویری باشد که در فضای خصیصه تعبیه شده‌اند و Y مجموعه‌ی N برچسب تعبیه‌شده در فضای برچسب باشد و میانگین آن‌ها μ_x و μ_y باشد، ماتریس‌های C_{xx}، C_{yy} ، C_{yx} و C_{xy} توسط روابط زیر به دست می‌آیند.

$$C_{xx} = \frac{1}{N}(X-\mu_x)(X-\mu_x)^T + \rho I \tag{۱۲-۳}$$

$$C_{yy} = \frac{1}{N}(Y-\mu_y)(Y-\mu_y)^T + \rho I \tag{۱۳-۳}$$

$$C_{xy} = \frac{1}{N}(X-\mu_x)(Y-\mu_y)^T \tag{۱۴-۳}$$

$$C_{yx} = (C_{xy})^T \tag{۱۵-۳}$$

در این روابط ρ ضریب تنظیم است و I ماتریس همانی است. هدف یافتن ماتریس‌های U_x و V_y طبق رابطه‌ی زیر است.

$$argmax_{U_x,V_y}\frac{U_x^T C_{xy} V_y}{\sqrt{U_x^T C_{xx} U_x}\,\sqrt{V_y^T C_{yy} V_y}} \tag{۱۶-۳}$$

با استفاده از روابط زیر ماتریس U_x که معادل با بردارهای ویژه‌ی ماتریس Z است به دست می‌آید. ماتریس V_y نیز به روش مشابهی به دست می‌آید.

$$ZU_{xk} = \lambda_k^2 U_{xk} \;\;,\; Z = C_{xx}^{-1} C_{xy} C_{yy}^{-1} C_{yx} \tag{۱۷-۳}$$

$$U_x = \{U_{x1}, U_{x2}, \ldots\ldots U_{xk}\} \;\;, \qquad U_{xi} \in R^d$$

توسط این دو ماتریس، رتبه مشخصه‌ها و برچسب‌ها از فضای d بُعدی در فضای k بُعدی تعبیه می‌شوند. درواقع این روش اجازه می‌دهد که ابعاد خروجی را کنترل نموده و کاهش ابعاد را نیز انجام دهیم.

حل معادلات بالا و به دست آوردن دو ماتریس U_x و V_y، یک‌بار به‌صورت برون‌خط در مرحله‌ی آموزش انجام می‌شود. در زمان آزمون یک نمونه ($S \in R^d$) با توجه به اینکه در فضای مشخصه قرار دارد یا فضای برچسب، توسط $U_x^T(S-\mu_x)$ یا $V_y^T(S-\mu_y)$ در فضای برداری مشترک تعبیه می‌شود.

با استفاده از روند ذکرشده، آزمایش‌ها روی دو مجموعه داده از کلمات مجزای دست‌نویس فارسی، فارسا و ایرانشهر، انجام شده است. عمل بازیابی و بازشناسی برای هرکدام از دو مجموعه مورد بررسی قرار گرفته است. سپس جهت ارزیابی توانایی روش ذکرشده در بازیابی و بازشناسی کلمات خارج از فرهنگ

لغت، برای آموزش از یکی از این مجموعه‌ها و جهت آزمون از مجموعه دیگری استفاده شده است که هیچ کلاس کلمه و نویسنده مشترکی ندارند.

۳-۳-۱- معیارهای ارزیابی

معیارهای ارزیابی مختلفی جهت بررسی و مقایسه نتایج به‌دست‌آمده مورد استفاده قرار گرفته است. روش جستجوی کلمه در دسته روش‌های بازیابی قرار می‌گیرد، بنابراین از معیارهای ارزیابی روش‌های بازیابی اطلاعات جهت ارزیابی روش جستجوی کلمه استفاده می‌شود. معیارهای مختلفی جهت ارزیابی روش‌های بازیابی اطلاعات معرفی شده است که مهم‌ترین معیار، معیار میانگین دقت متوسط[1] است. میانگین دقت متوسط برای مجموعه‌ای از کلمات پرسشی، میانگین میزان دقت متوسط برای هر کلمه پرسشی است که به‌صورت زیر محاسبه می‌شود:

$$Map = \frac{\sum_{q=1}^{Q} AveP(q)}{Q} \qquad (۱۸\text{-}۳)$$

که Q تعداد کلمات پرسشی است. $AveP$ دقت متوسط را به ازای هر کلمه پرسشی محاسبه می‌کند. ازآنجایی‌که دقت[2] و فراخوانی[3] معیارهایی تک مقداری بر اساس لیست تمامی کلمات بازیابی شده است، برای سیستمی که یک سری مرتب‌شده از کلمات بازیابی شده را بازمی‌گرداند، باید ترتیب بازیابی نیز مدنظر قرار داده شود. با محاسبه دقت و فراخوانی در هر موقعیت لیست مرتب بازیابی شده فقط می‌توان منحنی دقت-فراخوانی[4] را رسم نمود. در این منحنی دقت $P(r)$ به‌عنوان تابعی از فراخوانی رسم می‌شود. دقت متوسط میانگین مقادیر $P(r)$ را روی فاصله ۰ تا ۱ محاسبه می‌کند.

$$AveP = \int_0^1 P(r)dr \qquad (۱۹\text{-}۳)$$

این رابطه مساحت زیر نمودار دقت- فراخوانی را محاسبه می‌کند. در عمل انتگرال با محاسبه مجموع در هر موقعیت از لیست مرتب جایگزین می‌شود:

$$AveP = \sum_{k=1}^{n} P(k)\Delta r(k) \qquad (۲۰\text{-}۳)$$

---

[1] Mean average precision (Map)

[2] Precision

[3] recall

[4] Precision-recall curve

که k موقعیت در لیست بازیابی شده و n تعداد اسناد بازیابی شده را نشان می‌دهد. $P(k)$ دقت در k عنصر اول لیست بازیابی شده و$\Delta r(k)$ تغییر در فراخوانی از $k\text{-}1$ به k است. این مجموع برابراست با:

$$AveP = \frac{\sum_{k=1}^{n}\left(P(k) \times rel(k)\right)}{number\ of\ relevent\ retrived\ sample} \qquad (۳-۲۱)$$

تابع $rel(k)$ درصورتی‌که در موقعیت k ام از لیست بازیابی شده نمونه صحیحی وجود داشته باشد برابر ۱ و در غیر این صورت برابر ۰ خواهد بود.

معیار ارزیابی دیگری که در جستجوی کلمه مورد استفاده قرار می‌گیرد معیار دقت در اولین بازیابی[1] است. این معیار نرخ صحت اولین مورد بازیابی شده را مشخص می‌کند.

معیار ارزیابی به‌کاررفته در سیستم بازشناسی کلمات دست‌نویس، معیار دقت است که نسبت تعداد کلمه دست‌نویسی که درست شناسایی شده است را به کل نمونه‌های صحیح آن کلمه محاسبه می‌کند.

۳-۳-۲- نتایج تجربی

نتایج ارزیابی با استفاده از معیارهای معرفی‌شده در جداول زیر آمده است. دقت بازشناسی کلمات دست‌نویس در جدول ۳-۳ آورده شده است. در این جدول روش‌های پیشنهادی با آخرین روش‌های معرفی‌شده مقایسه شده است. انجام مقایسه با ۴ حالت مختلف انجام شده است. در حالت اول ۷۰٪ از مجموعه فارسا جهت آموزش مدل و ۳۰٪ جهت آزمون اختصاص داده‌شده و نتایج با استفاده از خصیصه‌های پیشنهادی نشان داده شده است. در حالت دوم بر همین اساس روی داده‌های مجموعه داده ایرانشهر دقت محاسبه شده است. نتایج تجربی نشان می‌دهد که روش‌های پیشنهادی که ساختار نوشتاری زبان فارسی را در نظر می‌گیرد دقت بهتری نسبت به روش پایه گزارش می‌کند. با توجه به این‌که در این دو روش ساختار نگارشی زبان مدنظر قرار داده شده است، نتایج حاصله نشان می‌دهد که خصیصه‌های پیشنهادی توانایی اشتراک‌گذاری اطلاعات بین کلاسی را بهتر از روش پایه انجام می‌دهند.

---

[1] Precision at 1 (P@1)

جدول ۳-۳: نتایج دقت روش پیشنهادی جهت بازشناسی کلمه

روش	نرخ تشخیص			
	فارسا	ایرانشهر	آموزش با فارسا آزمون با ایرانشهر	آموزش با ایرانشهر آزمون با فارسا
PPHOC	۹۵/۴۴	۹۵/۸۱	۲۷/۵۳	۴۰/۹۲
PPHOPC	۹۶/۰۷	**۹۷/۵۵**	۲۸/۴۲	۳۹/۳۷
PPHOPS	**۹۶/۱۵**	۹۷/۰۸	**۲۸/۵۰**	**۴۲/۶۸**
Gradient based[78]	_	۸۹/۱۲	–	–
Decision Templates[79]	_	۹۱/۵۵	–	–
Mixture of Experts[80]	۹۱/۱۱	–	–	–
SOM+DHMM[81]	۹۵/۶۳	۹۶/۹۴	–	–
Divide and conquer[82]	۹۰/۵۰	–	–	–

در بخش سوم و چهارم ارزیابی توانایی شناسایی کلاس‌های دیده نشده مورد ارزیابی قرار گرفته است. یکی از توانایی‌های روش‌هایی که از رده‌بندی مبتنی بر خصیصه استفاده می‌کنند، توانایی شناسایی کلاس‌های دیده نشده است. به همین دلیل در این دو بخش آموزش و آزمون مدل با استفاده از دو مجموعه داده مجزا انجام گرفت. لازم به ذکر است که این دو مجموعه داده تعداد کلاس و نویسنده‌های مختلفی دارند. نتایج تجربی نشان می‌دهد که توانایی بازیابی کلاس‌های دیده نشده در سیستم پیشنهادی وجود دارد. دقت‌ها در موردی که مدل با استفاده از مجموعه داده ایرانشهر آموزش داده‌شده بالاتر است. دلیل این امر را می‌توان در تعداد بیشتر کلاس‌های این مجموعه و درنتیجه آموزش بهتر مدل دانست.

جدول ۳-۴ ارزیابی را جهت سیستم جستجوی کلمه مبتنی بر معیار میانگین دقت متوسط نشان می‌دهد. در این جدول هم مشابه جدول ۳-۳ با ۴ روش مختلف مقایسه انجام شده است. در این ارزیابی‌ها نیز

بهترین نتایج در مورد خصیصه‌هایی گزارش شده است که ساختار نگارشی زبان را در نظر گرفته‌اند. علاوه بر این، نتایج تجربی توانایی روش در جستجوی کلمات دیده نشده را نشان می‌دهد.

جدول ۳-۴: نتایج میانگین دقت متوسط (MaP) روش پیشنهادی جهت جستجوی کلمه

مجموعه داده	نوع خصیصه	کلمه پرسشی متنی				کلمه پرسشی تصویری			
		مقایسه مستقیم	*KCCA*	*CCA*	*Platt Scaling*	مقایسه مستقیم	*KCCA*	*CCA*	*Platt Scaling*
فارسا	*PPHOC*	۸۱/۸۰	۹۴/۹۸	۹۳/۹۰	۹۴/۳۲	۷۸/۴۰	۸۶/۸۶	۸۲/۲۹	۸۵/۱۵
	PPHOPC	۸۴/۲۱	۹۵/۵۷	۹۴/۶۸	۹۵/۸۶	۷۶/۸۵	۸۸/۳۵	۸۳/۰۴	۸۸/۰۵
	PPHOPS	۸۵/۲۲	**۹۵/۶۷**	۹۴/۵۶	۹۴/۶۳	۷۷/۹۵	**۸۸/۹۹**	۸۴/۱۹	۸۶/۸۰
ایرانشهر	*PPHOC*	۸۲/۸۱	۹۶/۲۱	۹۵/۴۹	۹۶/۸۳	۸۸/۰۲	۹۱/۵۶	۸۹/۵۱	۹۱/۸۲
	PPHOPC	۷۸/۲۸	۹۶/۲۰	۹۵/۷۰	۹۶/۹۰	۸۸/۱۰	**۹۲/۰۸**	۹۰/۱۲	۹۱/۸۱
	PPHOPS	۸۵/۶۴	**۹۶/۲۵**	۹۵/۵۵	۹۶/۰۱	۸۷/۴۱	۹۲/۰۱	۸۹/۹۹	۹۱/۲۹
آموزش با فارسا آزمون با ایرانشهر	*PPHOC*	۳۹/۵۶	۴۷/۳۴	۴۷/۴۵	۴۲/۲۵	۷۰/۱۱	۸۱/۹۲	۸۲/۸۸	۴۴/۶۵
	PPHOPC	۴۱/۵۶	۴۷/۷۸	**۴۸/۴۳**	۴۲/۶۹	۷۴/۴۲	**۸۳/۸۸**	۸۴/۱۰	۴۳/۶۰
	PPHOPS	۳۹/۵۰	۴۷/۳۲	۴۷/۱۶	۴۲/۲۸	۷۱/۰۹	۸۳/۱۳	۸۳/۷۷	۴۷/۳۷
آموزش با ایرانشهر آزمون با فارسا	*PPHOC*	۶۸/۲۲	۸۳/۷۵	۸۲/۱۶	۶۸/۲۲	۸۰/۸۰	۸۵/۰۱	۴۸/۷۱	۳۷/۱۰
	PPHOPC	۶۸/۰۸	۸۱/۰۸	۸۲/۲۵	۷۰/۴۶	۸۱/۴۰	۸۵/۸۱	۸۴/۹۹	۵۹/۹۱
	PPHOPS	۷۱/۷۶	**۸۸/۵۳**	۸۷/۸۸	۷۷/۷۲	۸۲/۵۰	۸۵/۶۵	**۸۶/۶۳**	۶۸/۹۲

جدول ۳-۵: نتایج دقت در اولین بازیابی (P@1) روش پیشنهادی جهت جستجوی کلمه

کلمه پرسشی متنی				کلمه پرسشی تصویری				نوع خصیصه	مجموعه داده
مقایسه مستقیم	*KCCA*	*CCA*	*Platt Scaling*	مقایسه مستقیم	*KCCA*	*CCA*	*Platt Scaling*		
۹۸/۶۷	**۱۰۰**	**۱۰۰**	**۱۰۰**	۹۸/۷۰	۹۸/۸۸	۹۸/۹۳	۹۴/۷۴	*PPHOC*	فارسا
۹۹/۳۳	**۱۰۰**	**۱۰۰**	**۱۰۰**	۹۸/۵۸	**۹۸/۹۴**	۹۸/۹۳	۹۴/۲۷	*PPHOPC*	
۹۸/۶۷	**۱۰۰**	**۱۰۰**	**۱۰۰**	۹۸/۸۴	۹۸/۸۲	۹۸/۵۸	۹۶	*PPHOPS*	
۹۰/۲۴	۹۹	۹۹	۹۸/۸۱	۸۹/۶۶	۹۳/۸۴	۸۹/۴۰	۹۲/۷۱	*PPHOC*	ایرانشهر
۹۱/۴۳	۹۸/۲۱	۹۹/۲۰	۹۸/۸۱	۸۹/۳۳	۹۴/۳۰	۹۱/۸۵	۹۳/۹۷	*PPHOPC*	
۸۸/۸۴	۹۸/۴۱	**۹۹/۴۰**	۹۸/۸۱	۸۹/۴۶	**۹۴/۴۳**	۹۲/۳۱	۹۴/۳۷	*PPHOPS*	
۴۸/۷۵	۴۸/۷۵	۴۸/۷۵	۴۸/۷۵	۹۸/۳۲	۹۸/۸۸	۹۹/۱۰	۸۰/۷۸	*PPHOC*	آموزش با فارسا آزمون با ایرانشهر
۴۸/۵۰	۴۸/۵۰	۵۰	۵۰	۹۸/۷۸	**۹۹/۱۸**	۹۹/۱۵	۷۶/۰۲	*PPHOPC*	
۴۷/۵۰	۴۷/۵۰	**۵۰**	**۵۰**	۹۸/۵۰	۹۸/۹۸	۹۸/۹۸	۸۳/۳۸	*PPHOPS*	
۹۳/۷۵	۹۲/۵۰	۷۰/۰۷	۸۹/۰۱	۹۹/۲۵	۹۹/۳۲	۹۹/۳۳	۹۸/۸۷	*PPHOC*	آموزش با ایرانشهر آزمون با فارسا
۹۰	۹۱/۲۵	۷۱/۲۵	۸۸/۷۵	۹۹/۳۲	۹۹/۳۳	**۹۹/۳۸**	۸۵/۲۰	*PPHOPC*	
۸۲/۵۰	۹۶/۲۵	**۹۷/۵۰**	۹۲/۵۰	۹۹/۲۲	۹۹/۲۵	۹۹/۲۸	۹۲/۴۵	*PPHOPS*	

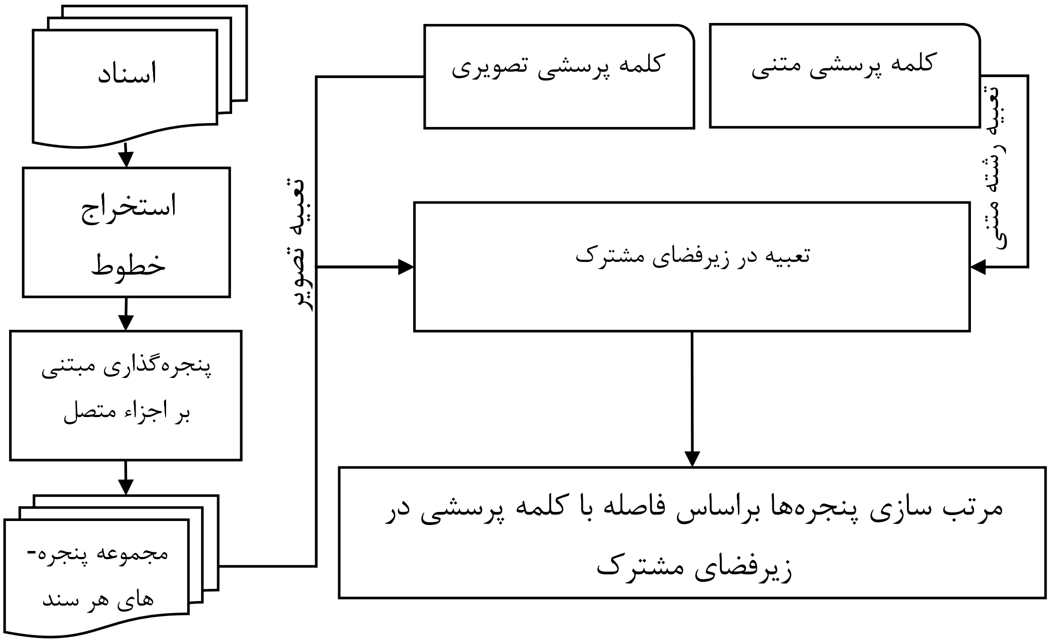
با توجه به نتایج تجربی گزارش‌شده در جداول بالا، روش‌های پیشنهادی توانایی بهتری در بازنمایی خصیصه‌های اسناد زبان فارسی دارد. علاوه بر این نشان داده شده که پیچیدگی و تنوع در شیوه نگارش و کلمات استفاده‌شده در پایگاه داده ایرانشهر بسیار بیشتر از پایگاه داده فارسا است. با توجه به این تنوع در مواردی که آموزش با استفاده از پایگاه داده ایرانشهر انجام شده و آزمون روی مجموعه داده فارسا است، دقت روش بسیار بالاتر است. دلیل این امر مدل کردن بهتر شیوه‌های نگارش متفاوت توسط نمونه‌های آموزشی این مجموعه می‌باشد.

فصل

۴- روش پیشنهادی دوم (مبتنی بر خط)

۴-۱- مقدمه

با توجه به مطالب ذکرشده در فصل‌های قبل، ازآنجایی‌که قطعه‌بندی کلمات و استخراج دقیق محدوده کلمات در نوشته‌های فارسی خصوصاً در مورد اسناد دست‌نویس، معمولاً با خطای بالای قطعه‌بندی روبرو است، محققان به دنبال روش‌هایی بوده‌اند که از این خطا رهایی یابند. یکی از این راهکارها استفاده از قطعه‌بندی در سطح خط است. در این فصل روشی مبتنی بر قطعه‌بندی در سطح خط جهت جستجوی کلمه معرفی شده است. در این روش عمل قطعه‌بندی خطوط با استفاده از روش پیشنهادی انجام می‌شود و بعدازآن بر اساس یک روش پنجره‌گذاری پیشنهادی که مبتنی بر اجزاء متصل است عمل پنجره گذاری انجام می‌شود. مجموعه‌ای از پنجره‌ها که شامل تعداد مختلفی جزء متصل می‌باشند آماده می‌شود. این مجموعه بر اساس روش تعبیه برچسب در فضای آموزش دیده شده توسط مجموعه داده‌های کلمات مجزا تعبیه‌شده و به ازای کلمه‌ی پرسشی ورودی مرتب می‌شوند(شکل ۴-۱).



شکل ۴-۱: ساختار کلی روش پیشنهادی مبتنی بر قطعه‌بندی خط

روند انجام تعبیه برچسب و تعبیه تصاویر بر اساس روش شرح داده شده در فصل سوم انجام می‌گیرد. بنابراین در ادامه این بخش روش پیشنهادی جهت قطعه‌بندی خط و پنجره‌گذاری و نحوه استفاده از پنجره‌ها جهت کاهش تعداد مقایسه شرح داده شده است.

۴-۲- قطعه‌بندی خطوط

استخراج یا قطعه‌بندی خط یکی از مراحل مهم پیش‌پردازش در بسیاری از کاربردهای آنالیز اسناد تصویری مانند شناسایی متون چاپی و دست‌نویس برون‌خط، جستجوی کلمات، بازیابی کلمات تصویری و شناسایی کاراکتر نوری می‌باشد. با توجه به اینکه در اسناد دست‌نویس فاصله خطوط و کلمات در یک سند متغیر است و همچنین معمولاً راستای مستقیم برای خط مبنا رعایت نشده و خطوط یک سند دارای زوایای متفاوتی نسبت به خط مبنا هستند، استخراج خطوط در اسناد دست‌نویس بسیار چالش‌برانگیز است. یکی از مشکلات بزرگ دیگر در این زمینه امکان به هم چسبیدگی اجزاء متصل از دو خط متوالی است؛ که با توجه به وجود نقاط و سرکش‌ها در متون فارسی و عربی این مسئله بیشتر مشکل ایجاد می‌نماید. استخراج صحیح یا غلط خطوط تأثیر مستقیم زیادی بر دقت قطعه‌بندی کلمات و حروف دارد.

روش‌های متفاوتی برای استخراج و قطعه‌بندی خطوط از اسناد دست‌نویس ارائه شده است. روش‌های موجود در استخراج خط در مقالات به روش‌های مختلفی دسته‌بندی شده است که به‌صورت زیر می‌باشند: روش‌های مبتنی بر افکنش، روش‌های مبتنی بر لکه‌دار کردن[1]، روش‌های مبتنی بر تبدیل هاف[2]، روش مبتنی بر نازک‌سازی[3] و روش‌های احتمالی[4].

روش‌های مبتنی بر افکنش معمولاً در مورد اسناد تصویری چاپی مورد استفاده قرار می‌گیرد و برای اسناد دست‌نویسی که خطوط در آن همپوشانی ندارند نیز توسعه یافته است[۸۳, ۸۴]. اما این روش به علت مشکلات فراوان مانند در نظر نگرفتن شیب خط، در متون اسناد دست‌نویس تصویری، نمی‌تواند بر روی اسناد مستقیماً اعمال شود. به همین دلیل جهت بهبود این روش راهکارهایی مانند اعمال روش‌های مبتنی بر افکنش به برش‌های عمودی در سند و سپس ترکیب نتایج، پیشنهاد شده است[۸۵-۸۸]. در این روش سند تصویری به‌صورت موازی به نوارهای عمودی تقسیم‌بندی می‌شود . سپس مرزهای بالقوه جداکننده خطوط در هر بخش به دست می‌آید و در مرحله بعد این بخش‌های کاندید با استفاده از هیوریستیک یا مدل‌های گرافیکی آماری مانند مدل مخفی مارکوف به یکدیگر متصل شده و نواحی خطوط را تشکیل می‌دهند. نهایتاً برای فائق آمدن بر مواردی که خطوط همپوشانی دارند از تکنیک‌های تقاطع[5]، الگوریتم تعقیب کانتور و چگالی گاوسی دومتغیره برای بازنمایی خطوط استفاده شده است. روش مذکور دارای ضعف‌های زیادی می‌باشد که به چند مورد اصلی آن اشاره شده است: این روش خطوط بالقوه زیادی را تولید می‌کند، پارامتر عرض نوارهای برش باید از پیش تعریف شده باشد، خطوط

---

[1] Smearing Methods
[2] Hough Transform
[3] Methods based on thinning operation
[4] Stochastic Methods
[5] Crossing technique

نباید انحنای زیاد داشته باشند و اگر نوار ابتدایی و انتهایی شامل خطوط بالقوه نباشند، استخراج خط به‌طور کامل ناممکن است.

برخی شیب خط را نیز برای جداسازی خطوط مورد استفاده قرار داده‌اند [۸۵] که این شیب معمولاً به سادگی قابل‌محاسبه نیست و در یک سند بین خطوط متفاوت است. بنابراین این روش‌ها نیز در این شرایط نتایج خوبی را ارائه نمی‌کنند.

در روش‌های مبتنی بر لکه‌دار کردن، روش‌هایی مانند الگوریتم هموارسازی طول اجرا[1] [۸۹] یا الگوریتم آنالیز طول اجرای جهت‌دار بهبودیافته[2] [۹۰] بکار گرفته شده است. استفاده از این الگوریتم‌ها باعث خواهد شد که بلوک‌های پیکسلی در جهت افقی لکه‌دار شوند؛ یعنی اگر فاصله سفید(پس‌زمینه) بین آن‌ها از یک حد آستانه‌ای کمتر بود با پیکسل سیاه(پیش‌زمینه) پر شود. این روش ابتدا روی تصاویر خاکستری اعمال‌شده و سپس تصویر باینری می‌شود و عمل مورفولوژی فرسایش روی پیش‌زمینه و پس‌زمینه برای به دست آوردن محدوده خطوط انجام شده است.

روش‌های مبتنی بر رهیافت تبدیل هاف برای استخراج خطوط مستقیم در تصاویر بسیار پرکاربرد است. شیب خطوط دست‌نویس با استفاده از اعمال این نوع تبدیل به مرکز ثقل هر جزء متصل به دست می‌آید. اگر اکثر همسایه‌های نزدیک در یک تراز متعلق به گروه اجزاء تشکیل‌دهنده هم‌ترازی داشته باشند، این هم‌ترازی ویژگی‌های پیوستگی و نزدیکی را دارا خواهد بود و به‌عنوان یک خط پذیرفته می‌شود. سپس نتایج حاصل‌شده با حذف هم‌ترازی‌های نادرست از هم‌ترازی بین اجزاء متصل با استفاده از اطلاعات محتوایی، تصویر خروجی نهایی را ارائه می‌کند[۹۱]. در [۹۲] روشی مبتنی بر تبدیل هاف بر روی سند بلوک‌بندی شده ارائه شده است.

برخی محققین برای استخراج خطوط هندی و چینی از روش نازک‌سازی استفاده کرده‌اند، که در[۹۳] ابتدا نازک‌سازی صورت گرفته و در ادامه عملیات پس‌پردازشی روی تمامی نواحی پس‌زمینه برای تشخیص مرز جداکننده خطوط اعمال شده است.

اخیراً برخی از روش‌های مجموعه سطوح[3]، کانتور فعال[4] و بیز تغییرات[5] برای استخراج خطوط بهره برده‌اند. در [۹۴] از روش تخمین چگالی و مجموعه سطوح استفاده شده است. در این روش یک نگاشت احتمالی از سند تصویری ورودی تخمین زده می‌شود که هر المان احتمال تعلق هر پیکسل به یک خط را نشان می‌دهد. روش مجموعه سطوح برای تعیین ارزیابی مرزی با خطوط همسایه مورد استفاده قرار

---

[1] Run-Length Smoothing Algorithm (RLSA)
[2] Improved Directional Run-Length Analysis
[3] Level Set
[4] Active Contour
[5] Variational Bayes

گرفته است. در[۹۵] ابتدا یک بانک فیلتر تطبیق یافته برای هموارسازی تصویر ورودی مورد استفاده قرار گرفته است. مراکز خطوط اجزاء تشکیل‌دهنده خط با استفاده از مرزبندی تصویر هموارشده محاسبه می‌شود. در انتها برای به دست آوردن نتیجه نهایی، کانتورهای فعال روی مرزها وفق داده می‌شوند. در کار[۹۶] روش بیز تغییرات برای استخراج خطوط در متون دست‌نویس چینی مورد بهره‌برداری قرار گرفته است. یک سند به‌صورت یک مدل مخلوط گاوسی در نظر گرفته می‌شود، طوری که هر جزء متعلق به یک خط باشد. برای تخمین پارامترهای چگالی و تعیین تعداد خطوط چارچوب کاری بیز تغییراتی بکار برده شده است.

در[۹۷, ۹۸] یک روش مبتنی بر بلوک‌بندی معرفی شده که پارامترهای آن به‌صورت وفقی بر اساس سند ورودی تعیین می‌گردد. در هر بلاک برای تشخیص جهت از فیلتر کردن تصویر با هسته‌های گاوسی دوبعدی ناهمسانگرد[1] در زوایای بین ۳۰- درجه تا ۳۰ درجه استفاده شده است. پس از تشخیص جهت، خطوط در هر بلاک از هم جدا می‌شوند.

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

در[۹۹] روشی مبتنی بر تکنیک نقاشی[2] ارائه شده است. در این روش یک سند به نوارهای عمودی تقسیم‌بندی می‌شود و با استفاده از تکنیک‌های نقاشی اجزاء نزدیک به همدیگر در هر نوار عمودی به هم متصل می‌شوند که در این صورت ادعا شده که شیب خط به دست می‌آید. از عمل مورفولوژی انبساط جهت برقراری اتصال بین لکه‌های ایجادشده در مرحله قبل و تشکیل خط استفاده می‌شود. سپس روی پس‌زمینه نتیجه مرحله قبل عمل نازک‌سازی انجام می‌گیرد و با استفاده از برخی قواعد هیوریستیک خطوط نهایی استخراج می‌شوند. این روش با توجه به اینکه از روش‌های قبلی برای زبان‌های

---

[1] Anisotropic 2D Gaussian Kernels

[2] Painting Technique

فارسی و عربی بهتر کار می‌کند، اما باز هم در مورد این دو زبان درصد خطاهای بسیار بیشتری نسبت به لاتین دارد. لذا در روش ارائه‌شده سعی شده تا این مشکلات را با استفاده از خوشه‌بندی سلسله مراتبی با استفاده از یک معیار فاصله معرفی‌شده مرتفع سازیم.

روش پیشنهادشده یک روش جزء به کل می‌باشد که از خوشه‌بندی سلسله‌مراتبی پیکسل‌های قلم (سیاه یا پیش‌زمینه) موجود در سند جهت استخراج خطوط استفاده می‌کند. ساختار کلی روش پیشنهادی، که در شکل ۴-۲ نشان داده شده است، شیوه نگارش متن فارسی به‌صورت دست‌نویس را مدنظر قرار داده است. ازآنجایی‌که معمولاً دست‌نوشته‌ها ممکن است شامل خطوط نزدیک به همدیگر باشند یا نقاط و سرکش‌ها باعث اشتباه در خوشه‌بندی شوند، از یک مرحله پیش‌پردازش استفاده‌شده که احتمال خوشه‌بندی اشتباه را در موارد ذکرشده کاهش دهد.

با توجه به اینکه در متون فارسی و عربی شیب خط مبنای استاندارد مستقیم است و در متون دست‌نویس این شیب می‌تواند به مقدار کمی خط مبنا را به سمت بالا یا پایین متمایل کند یا در آن انحنای کمی را ایجاد نماید، معیار فاصله‌ای معرفی‌شده که این مسئله را در نظر بگیرد. نتایج نهایی خوشه‌بندی با استفاده از تخمین شیب خط و هیوریستیک‌هایی مبتنی بر شیوه نگارش فارسی اصلاح و تکمیل می‌گردد. این روش در ادامه به‌طور کامل تشریح و نتایج آزمایش‌ها روی دو مجموعه استاندارد زبان فارسی گزارش شده است.

۴-۲-۱- پیش‌پردازش

مرحله پیش‌پردازش یکی از مهم‌ترین مراحل در پردازش تصاویر دیجیتال می‌باشد. در سیستم‌های پردازش اسناد تصویری در اکثر روش‌ها اولین گام پیش‌پردازش تبدیل سند تصویری به یک تصویر دودویی است. به‌عنوان اولین مرحله پیش‌پردازش، در این روش با استفاده از الگوریتم اوتسو[1] [۱۰۰] سند ورودی به تصویر دودویی تبدیل می‌گردد. با توجه به اینکه در متون دست‌نویس فارسی، بالارونده‌هایی مانند حروف الف و لام و پایین‌رونده‌هایی مانند ر، ز و واو و سرکش‌ها ممکن است در یک فاصله افقی کم از همدیگر قرار گیرند، استفاده از نازک‌سازی می‌تواند امکان تداخل این موارد را کاهش دهد. بنابراین در این روش از عمل نازک‌سازی بر پایه مورفولوژی استفاده شده است. با توجه به اینکه در اسناد ورودی تعداد خطوط، اندازه تصویر و نوع دستخط از نظر اندازه متفاوت است نیاز به تعیین برخی پارامترها برای هر سند وجود دارد تا روش بتواند خود را بر اساس شرایط سند ورودی تطبیق دهد.

---

[1] Otsu

شکل ۴-۲: ساختار کلی روش پیشنهادی جهت قطعه‌بندی خطوط

یکی از مهم‌ترین پارامترهایی که طی عملیات خوشه‌بندی سلسله مراتبی پیشنهادی به آن نیاز داریم، حد آستانه‌ای جهت توقف ترکیب خوشه‌ها و استخراج خوشه‌بندی نهایی می‌باشد. برای شرط توقف

اتصال خوشه‌ها از معیار حداکثر فاصله استفاده می‌شود و میزان این حد آستانه از پارامتر فاصله مؤثر بین دو خط مجاور به دست می‌آید. برای به دست آوردن این پارامتر، تصویر سند به ۲۰ نوار عمودی با عرض برابر تقسیم می‌شود و سپس متوسط فاصله خطوط مجاور برای ۱۴ نوار میانی اندازه گرفته می‌شود. بدین‌صورت که برای هرکدام از نوارهای مذکور پروفایل افکنش افقی محاسبه‌شده و از آن فاصله بین قله‌های پروفایل استخراج می‌گردد. شکل ۴-۳ تصویری از این بلوک‌بندی و قله‌های مذکور در یکی از نوارها را نمایش می‌دهد. اگر این فاصله برای تمامی نوارهای مذکور محاسبه شود، فاصله مؤثر بین دو خط مجاور از رابطه زیر به دست می‌آید:

$$Line_{D_e} = \frac{\sum_{i \in Corr_D} i * Hist(i)}{\sum_{i \in Corr_D} Hist(i)}$$

(۱-۴)

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

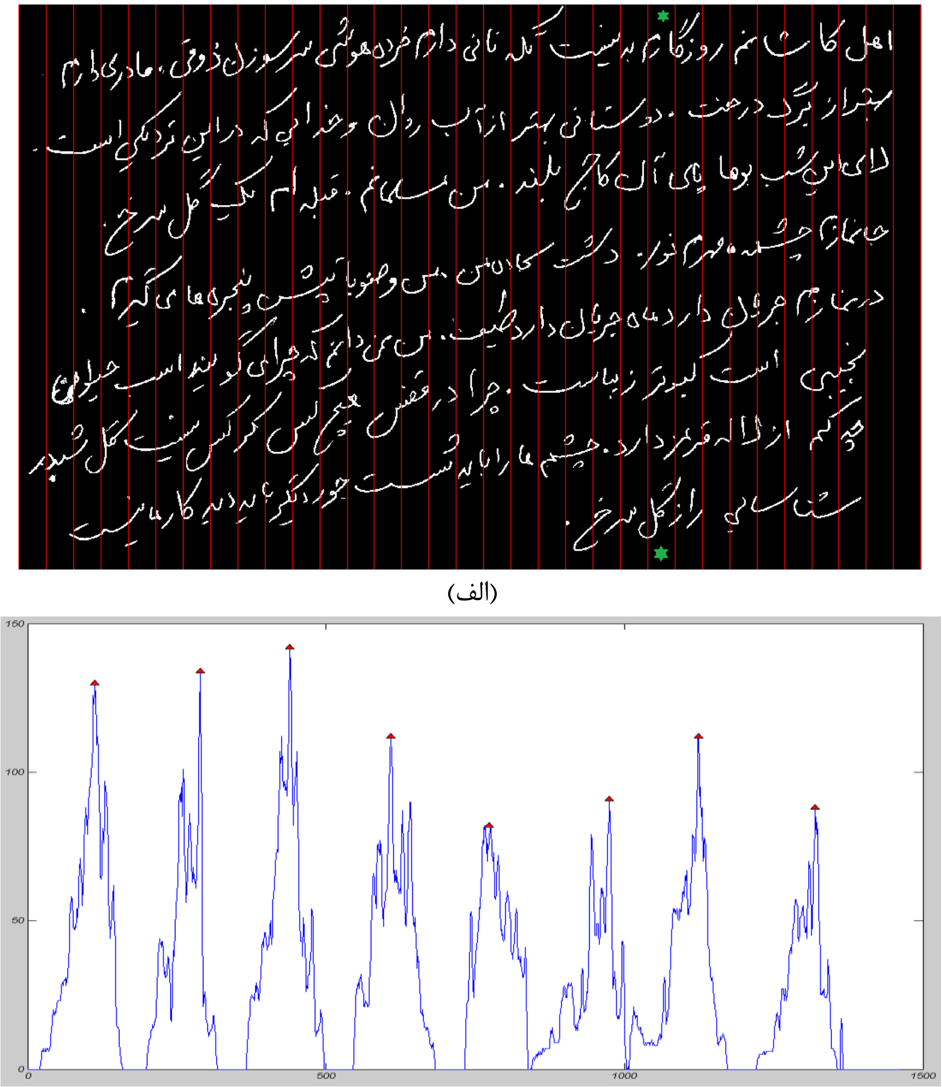
(۲-۴)

که در این روابط Corr_D شامل اندیس فاصله‌هایی می‌شود که تعداد آن‌ها حداقل از ۲۵ درصد حداکثر بیشتر باشد. Hist هیستوگرام فاصله بین خطوط به‌دست‌آمده را نشان می‌دهد. Line_De نشان‌دهنده فاصله مؤثر بین خطوط است. نصف این فاصله به‌عنوان پارامتر معیار توقف خوشه‌بندی استفاده شده است.

پارامتر دسته‌بندی کردن خوشه‌ها به دو دسته‌ی کوچک و بزرگ به‌صورت زیر تعریف می‌شود. تعداد پیکسل‌های قلم (مساحت) برای هر خوشه محاسبه شده و سپس مرتب می‌شوند. فاصله بین مساحت‌های مرتب‌شده خوشه‌ها به‌دست‌آمده و بر اساس رابطه زیر نرمال می‌شود. مساحت شروع بزرگ‌ترین فاصله مقدار پارامتر حد آستانه جداسازی خوشه‌ها را تعیین می‌کند.

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

(۳-۴)



اهل کاشانم روزگارم بد نیست تکه نانی دارم خرده هوشی سر سوزن ذوقی، مادری دارم
بهتر از برگ درخت. دوستانی بهتر از آب روان و خدایی که در این نزدیکی است
لای این شب بوها پای آن کاج بلند. من مسلمانم. قبله ام یک گل سرخ
جانمازم چشمه، مهرم نور. دشت سجاده من. من وضو با تپش پنجره ها می گیرم.
در نمازم جریان دارد ماه جریان دارد طیف. من نمی دانم که چرا می گویند اسب حیوان
نجیبی است کبوتر زیباست. چرا در قفس هیچ کسی کرکس نیست گل شبدر
چه کم از لاله قرمز دارد. چشم ها را باید شست جور دیگر باید دید کار ما نیست
شناسایی راز گل سرخ.

(الف)

(ب)

شکل ۴-۳: الف) سند قطعه‌بندی شده به نوارهای عمودی ب) پروفایل افکنش افقی نوار مشخص‌شده با علامت ستاره سبز در تصویر الف به همراه نمایش قله‌ها.

۴-۲-۲- خوشه‌بندی

جهت استخراج خطوط از روش خوشه‌بندی سلسله مراتبی کمترین فاصله[1] استفاده شده است. برای این کار موقعیت افقی و عمودی تمامی نقاط قلم (پیکسل‌های سیاه) به‌عنوان ویژگی‌های نمونه‌ها مورد استفاده قرار گرفته است. برای این‌که عملیات خوشه‌بندی سریع‌تر انجام شود، هر جزء متصل به‌عنوان یک خوشه اولیه در نظر گرفته شده و عمل خوشه‌بندی انجام می‌گیرد. در این روش با توجه به ساختار

---

[1] Single linkage

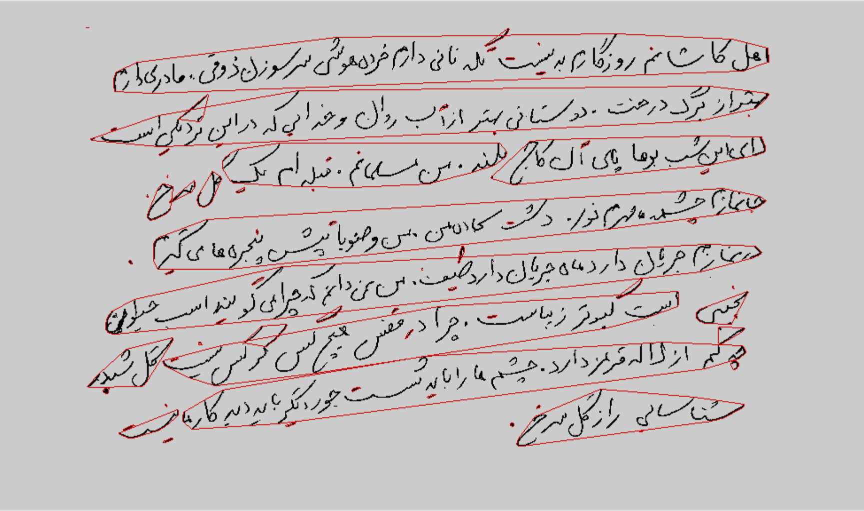
نوشتاری زبان فارسی که به‌صورت افقی نوشته می‌شود، فاصله طبق رابطه زیر طوری تعریف شده است که فاصله در جهت عمودی شامل جریمه باشد.

$$D(p_1, p_2) = \left|X_{p_1} - X_{p_2}\right| + \left|Y_{p_1} - Y_{p_2}\right|^5 \tag{۴-۴}$$

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

اما یکی از مسائل مهم در خوشه‌بندی سلسله‌مراتبی معیار توقف ادغام خوشه‌هاست. برای این کار معیار توقف حداکثر فاصله بین دو خوشه در نظر گرفته شده و برای هر سند به‌صورت وفقی تعیین می‌گردد که قبلاً شرح داده شده است. نتیجه این مرحله با شرایط ذکر شده مجموعه‌ای از خوشه‌هاست که معمولاً بیشتر از تعداد خطوط است و شامل خوشه‌های بزرگی که نمایش‌دهنده و معرف خط اصلی هستند می‌شود. شکل ۴-۴ نمونه‌ای از یک سند تصویری را پس از این مرحله نمایش می‌دهد.

همان‌طور که در شکل ۴-۴ نیز کاملاً مشخص است، معمولاً نقاط و سرکش‌ها که در بسیاری موارد فاصله عمودی آن‌ها تا بدنه اصلی زیاد است، در یک خوشه مجزا قرار می‌گیرند. یکی دیگر از مواردی که در برخی نمونه‌ها رخ می‌دهد تقسیم شدن یک خط واحد به چندین خوشه مجزا می‌باشد؛ این مشکل به خاطر تغییر ناگهانی شیب خط، فاصله زیاد بین اجزاء متصل یا عدم وجود معیار توقف مناسب و دقیق برای هر سند، رخ می‌دهد. برای غلبه بر این مشکلات از پس‌پردازش استفاده شده است که در قسمت بعدی شرح داده شده است.



شکل ۴-۴: یک سند نمونه خروجی مرحله خوشه‌بندی.

۴-۲-۳- پس پردازش

در این مرحله از تعدادی قاعده هیوریستیک مبتنی بر ساختار نگارشی فارسی و بر اساس نتایج اولیه مرحله خوشه‌بندی برای ترکیب خوشه‌ها و استخراج دقیق خطوط استفاده شده است که در زیر آورده شده است.

- در مرحله اول اندازه هر خوشه به دست می‌آید و خوشه‌ها بر اساس یک حد آستانه، که قبلاً ذکر شده است، به دو دسته‌ی خوشه‌های کوچک و بزرگ دسته‌بندی می‌شوند. درصورتی‌که ناحیه بدنه محدب یک خوشه کوچک فقط با یک خوشه بزرگ اشتراک بیش از ۵۰ درصد داشته باشد دو خوشه با هم ترکیب می‌شوند. با استفاده از قاعده بیان شده، اکثر خوشه‌های کوچکی که شامل نقاط و سرکش‌ها هستند به خوشه‌های اصلی می‌پیوندند.
- در مرحله بعد برای تصمیم‌گیری در مورد سرکش‌ها زوایای خوشه‌های کوچک محاسبه می‌شود. درصورتی‌که این زاویه بین ۳۰ تا ۶۰ درجه باشد سرکش تشخیص داده شده و با خوشه بزرگی که زیر آن قرار دارد ترکیب می‌شود.
- در ادامه خط تقریبی مبنا برای خوشه‌های بزرگ محاسبه می‌شود، بدین‌صورت که به ازای مختصات تمام نقاط قلم موجود در یک خوشه بزرگ یک خط برازش می‌شود. این خطوط استخراج شده شیب احتمالی خطوط اصلی سند را نشان می‌دهد. بر مبنای خطوط احتمالی استخراج شده، قواعدی جهت ترکیب خوشه‌ها در نظر گرفته شده است. اگر فاصله دو خط از خطوط برازش شده بسیار کم باشد نشان می‌دهد که دو خوشه بزرگ روی یک خط قرار گرفته‌اند و این دو خوشه با هم ترکیب می‌شوند. درنهایت فاصله اقلیدسی مراکز خوشه‌های کوچکی که

هنوز مورد تصمیم‌گیری قرار نگرفته‌اند با خطوط برازش شده به خوشه‌های بزرگ محاسبه شده و به خوشه مربوط به نزدیک‌ترین خط می‌پیوندد.

۴-۲-۴- نتایج تجربی

به‌منظور ارزیابی روش پیشنهادی از ۱۰۰ سند از مجموعه داده استاندارد استفاده شده است. مجموعه داده، FHT [۹] می‌باشد که شامل ۱۱۲۶ سند دست‌نویس فارسی است که توسط ۲۵ نویسنده مختلف نوشته شده است و متوسط تعداد سطر در هر سند حدود ۷ می‌باشد.

جهت ارزیابی دو معیار استاندارد مورد استفاده قرار گرفته است. اولین معیار نرخ تشخیص[1] است [۹۴] که در استخراج خط بسیار پرکاربرد می‌باشد. این معیار نرخ تشخیص صحیح خط در اسناد ورودی را مشخص می‌کند. اگر هرخط استخراج شده معیار تطابقی بالای ۹۵٪ را به دست آورد این خط به‌عنوان یک خط تشخیص داده شده به‌حساب می‌آید. این معیار تطابق به‌صورت زیر محاسبه می‌گردد.

$$MatchScore_{i,j} = T(G_i \cap R_j)/T(G_i \cup R_j) \qquad (۵-۴)$$

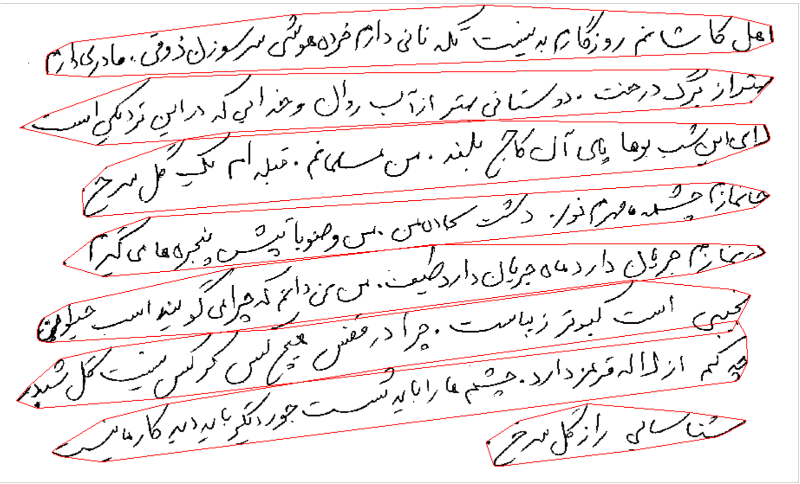
که در آن تابع T(a) تعداد پیکسل‌های پیش‌زمینه a را برمی‌گرداند. G_i نشان‌دهنده iامین تصویر از مجموعه درستی[2] و R_j نشان دهنده jامین خط استخراج شده توسط رهیافت پیشنهادی است. این معیار استاندارد از سال ۲۰۰۷ در مسابقات کنفرانس بین‌المللی بازشناسی و آنالیز اسناد[3] مورد استفاده قرار می‌گیرد. بر اساس این معیار روش پیشنهادی نرخ تشخیصی برابر ۹۲/۷۳٪ دارد.

معیار ارزیابی بعدی که در این مورد ارائه‌شده، معیار نرخ تطابق پیکسلی[4] است. این معیار بر اساس تقسیم تعداد پیکسل‌های قلم مشترک بین بهترین تطبیق در میان تصاویر مجموعه درستی و خط استخراج شده بر تعداد پیکسل‌های قلم این تصویر مجموعه درستی به دست می‌آید. نرخ تطابق پیکسلی برای مجموعه داده مورد ارزیابی قرار گرفته برابر ۹۷/۵۳٪ می‌باشد.

نتایج قابل‌مقایسه‌ای روی این مجموعه داده به دست آمده است که می‌توان به روش مذکور جهت ادامه گام بعدی اعتماد کرد.

---

[1] Detection Rate (DR)
[2] Ground Truth
[3] International Conference on Document Analysis and Recognition (ICDAR)
[4] Pixel Level Hit Rate (PLHR)



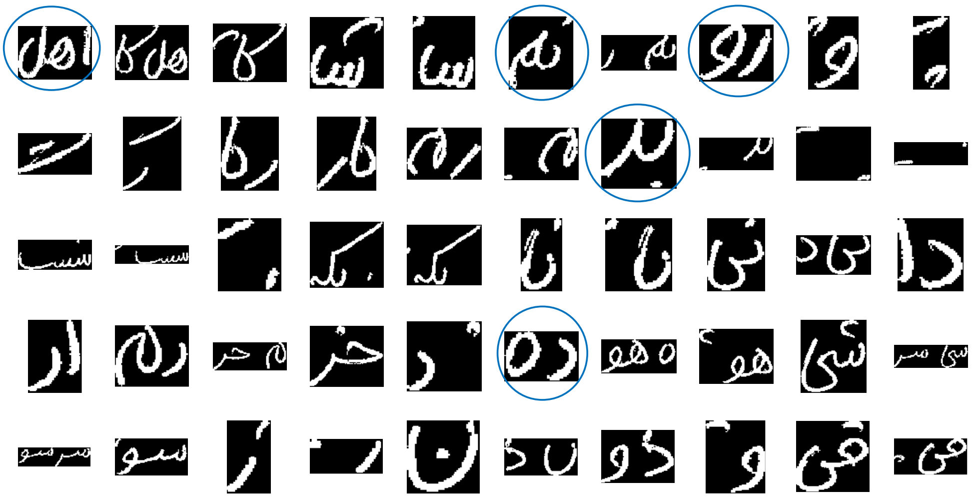
شکل ۴-۵: یک سند نمونه خروجی نهایی

۴-۳- پنجره‌گذاری مبتنی بر اجزاء متصل

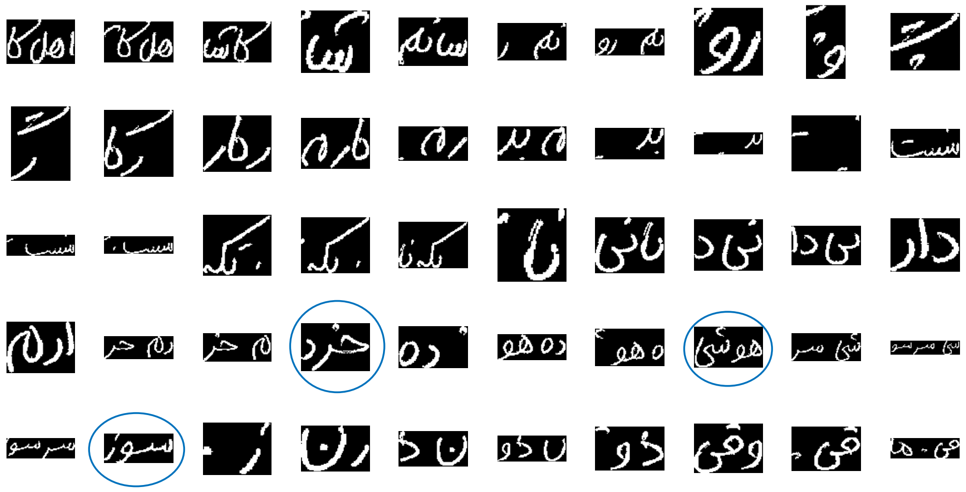
روش رایج پنجره‌گذاری در بینایی ماشین شامل در نظر گرفتن پنجره با سایزهای متفاوت است. این پنجره‌ها با همپوشانی زیاد کل فضای جستجو را پیمایش می‌کنند. این روش پنجره‌گذاری معایب زیادی دارد: تعیین اندازه/اندازه‌های مورد نظر برای پنجره لغزان به‌سادگی قابل‌محاسبه نیست، که این مسئله خصوصاً در مورد اسناد پیچیده‌تر بیشتر نمود می‌کند. حجم بالای پنجره‌هایی که باید پردازش شوند نیز از معایب بزرگ این روش محسوب می‌شود. به‌منظور رفع چالش‌های مربوط به پنجره‌گذاری معمول، روش پنجره‌گذاری خاص مسئله جستجوی کلمه با استفاده از اجزاء متصل پیشنهاد شده است.

این روش پنجره‌گذاری پیشنهادی از اجزاء متصل تشکیل‌دهنده خطوط و چهارضلعی محاطی آن‌ها استفاده می‌کند. نمونه‌ای از یک خط از یک سند و چهارضلعی‌های محاطی آن در شکل ۴-۶ نشان داده شده است. طول پنجره‌ها در این روش بر اساس تعداد اجزاء متصل درون آن تعیین می‌گردد. پنجره‌هایی شامل ۲ تا ۷ جزء متصل مجاور تعیین شده‌اند. از سمت راست به چپ این پنجره شامل توالی‌هایی از اجزاء متصل مجاور خواهد بود و روی این زنجیره اجزاء متصل که خط را تشکیل داده‌اند تا انتها جلو می‌رود.

شکل ۴-۶: یک سطر از یک سند نمونه با چهارضلعی‌های محاطی اجزاء متصل



شکل ۴-۷: نمونه‌های از خروجی پنجره به طول ۲ جزء متصل



شکل ۴-۸: نمونه‌های از خروجی پنجره به طول ۳ جزء متصل

شکل ۴-۹: نمونه‌های از خروجی پنجره به طول ۴ جزء متصل

شکل ۴-۱۰: نمونه‌های از خروجی پنجره به طول ۵ جزء متصل

این نوع پنجره‌گذاری یک مزیت بسیار بزرگی که دارد این است که هر جزء متصل یا به‌صورت کامل در پنجره قرار می‌گیرد و یا درون آن قرار نمی‌گیرد که این مسئله باعث خواهد شد در مرحله بازیابی نتایج بهتری حاصل شود. نمونه‌هایی از خروجی‌های حاصل از اعمال این پنجره‌گذاری در شکل‌های بالا آورده شده است. پنجره‌های به‌دست‌آمده از هر سند با توجه به روش تعبیه برچسب که در فصل سوم به‌طور کامل شرح داده شد، در زیرفضای مشترک تعبیه می‌شوند. این زیرفضا بایستی با استفاده از مجموعه داده‌ای از کلمات مجزای دارای برچسب آموزش داده شود.

در مرحله آزمون برای هر کلمه پرسشی ورودی، کافی است عملیات موردنیاز جهت تعبیه آن را در زیرفضای مشترک انجام دهیم و پنجره‌های استخراج شده در مرحله قبل، که در اینجا هر کدام به‌عنوان یک کلمه مجزا در نظر گرفته می‌شود، بر اساس میزان شباهت با کلمه پرسشی ورودی به‌صورت نزولی مرتب می‌شوند.

۴-۴- نتایج تجربی

برای انجام مرحله‌ی ارزیابی از پایگاه داده‌ی FHT، ۱۰ نوع متن متفاوت انتخاب‌شده که از هر نوع متن، ۱۰دستخط متفاوت وجود دارد. ضمناً این ۱۰ دستخط متفاوت به‌صورت تصادفی از بین حالات مختلف انتخاب‌شده و اسناد مختلف لزوماً یکسان نیستند. به‌طورکلی برای ارزیابی این روش از ۱۰۰ سند تصویری استفاده شده است.

جهت ارزیابی روش پیشنهادی مجموعه‌ای شامل ۶۰ کلمه‌ی متمایز به‌عنوان کلمات پرسشی انتخاب شده‌اند. ۱۰ متن انتخابی دارای موضوعات تقریباً مشابهی هستند تا فراوانی کلمات پرسشی داخل آن‌ها زیاد باشد. به همین دلیل در مجموعه‌ی کلمات پرسشی، ۲۴ کلمه در بیشتر از یک متن ظاهر شده‌اند. درمجموع تعداد این ۲۴ کلمه در داخل ۱۰ متن، ۷۸ می‌باشد. با در نظر گرفتن این ۷۸ کلمه و تعداد کلمات متمایز(۳۱ کلمه)، تعداد کل کلمات پرسشی که برای ۱۰ متن در نظر گرفته شده، ۱۰۹ می‌باشد. برای هرکدام از این ۱۰۹ کلمه، حداقل ۱۰ سند با دستخط متفاوت وجود دارد. بیشترین فراوانی یک کلمه در داخل یک متن ۷ و کمترین آن یک است. جدول ۴-۱ تعداد کلمات پرسشی با فراوانی بین ۱ تا ۷ در هر سند مجزا را نشان می‌دهد. با در نظر گرفتن تعداد فراوانی کلمات، به‌طورکلی ۱۸۱۰ نمونه کلمه در این ۱۰۰ سند وجود دارد که این روش به جستجوی آن‌ها پرداخته است. متوسط تعداد نمونه‌های کلمات پرسشی در متن برابر با ۱/۵۶۵۱می‌باشد. این الگوریتم در روش پرسش با تصویر، ۱۲۵۰ نمونه از کلمات پرسشی موجود در اسناد تصویری را در رتبه اول آشکار کرده است که بر اساس معیار دقت در اولین بازیابی (P@1) معادل با ۷۸,۱۸ ٪ می‌باشد و در روش پرسش با رشته‌ی متنی موفق به آشکارسازی ۱۳۱۲ نمونه از کلمات پرسشی در رتبه اول شده است که بر اساس معیار دقت در اولین بازیابی (P@1) معادل با ۸۱,۳۴ ٪ و ۸۰,۵۵٪ است. مجموعه‌ی کلمات پرسشی از جهت تعداد حروف تشکیل‌دهنده‌ی کلمه متنوع می‌باشند، به‌طوری‌که شامل کلماتی با تعداد حروف کم مثل "ما"، "در"، "شد" و "دو"، کلمات با تعداد حروف متوسط مثل"دارو"،"کشور"، "مردم" و "درآمد" و کلماتی با تعداد حروف زیاد مثل "مدیریت"، "اولویت"،" اقدامات" و "مستقیم" است.

جدول ۴-۱: تعداد کلمات با فراوانی بین ۱ تا ۷ در هر سند مجزا

تعداد نمونه‌های کلمات در متن	۱	۲	۳	۴	۵	۶	۷
تعداد کلمات	۶۷	۲۷	۷	۴	۲	۱	۱

برای هرکدام از اسناد با استفاده از روش جداسازی خطوط پیشنهادی، خطوط سند استخراج شد. سپس با استفاده از روش پنجره‌گذاری پیشنهادی پنجره‌هایی به طول ۲ تا ۱۰ جزء متصل ایجاد و ذخیره شد. زیرفضای مشترکی مبتنی بر مجموعه داده فارسا آموزش داده‌شده و تمامی پنجره‌ها در این فضا نگاشت داده شدند. برای جستجوی هر کلمه پرسشی، تعداد اجزاء متصل کلمه پرسشی ورودی محاسبه می‌شود. سپس این کلمه پرسشی در زیرفضای آموزش داده‌شده تعبیه می‌شود و فاصله اقلیدسی آن با پنجره‌هایی که از ۲ جزء کوچک‌تر تا ۲ جزء بزرگ‌تر از تعداد اجزاء متصل کلمه پرسشی دارند مقایسه می‌گردد و

لیست مرتب شده‌ای از پنجره‌های انتخابی برگردانده می‌شود. جهت ارزیابی نتایج حاصله از معیار میانگین دقت متوسط که در فصل قبل معرفی شد، استفاده شده است و نتایج در جدول ۴-۲ آورده شده است.

جدول ۴-۲: میانگین دقت متوسط با استفاده از روش مبتنی بر قطعه‌بندی خط

نوع کلمه پرسشی / روش	کلمه پرسشی متنی	کلمه پرسشی تصویری
CCA	۸۰/۴۱	۷۷/۷۲
KCCA	۸۱/۰۲	۷۸/۱۴

جدول ۴-۳ لیست مرتب خروجی به ازای چند کلمه پرسشی نمونه را آورده است و بازیابی‌هایی که صحیح انجام شده با کادر قرمز مشخص شده است. این جدول تعداد کلمه پرسشی نمونه در سند و دقت متوسط به ازای هر کلمه پرسشی را نشان می‌دهد. مثلاً برای کلمه پرسشی "در" که در این سند ۶ بار تکرار شده و در رتبه‌های ۱،۳،۴،۵،۶،۷ ظاهر شده است دقت متوسط به‌صورت زیر محاسبه می‌شود:

$$AveP = \frac{1}{1} + \frac{2}{3} + \frac{3}{4} + \frac{4}{5} + \frac{5}{6} + \frac{6}{7} = 0.818$$

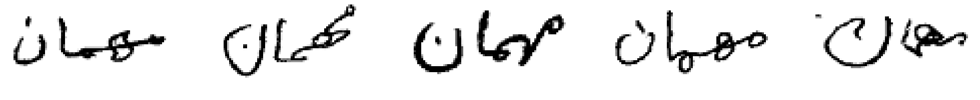
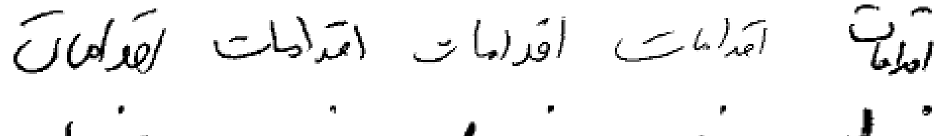
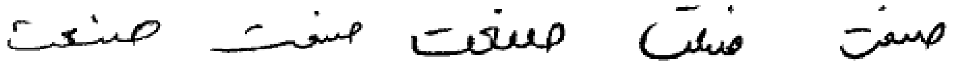
جدول ۴-۳: محاسبه دقت متوسط (AveP) برای چند کلمه پرسشی نمونه

AveP	تعداد نمونه در سند	کلمه پرسشی	Rank7	Rank6	Rank5	Rank4	Rank3	Rank2	Rank1
۰,۹۶۷	۵	دارو							
۰,۷۱۰۲	۵	است							
۱	۷	بود							
۰,۸۱۸	۶	در							
۱	۴	دو							
۰,۸۶۷	۳	ایران							
۰,۸۶۷	۳	تولید							

به ازای تمامی کلمات پرسشی در هر سند معیار دقت متوسط (AveP) محاسبه شده است. سپس میانگین آن و مقدار میانگین دقت متوسط (MaP) به ازای کل سیستم محاسبه می‌شود که در جدول ۴-۲ آورده شده است.

تنوع شیوه نگارش مختلف در مجموعه داده‌های مورداستفاده و توانایی روش پیشنهادی در بازیابی کلمات پرسشی دیده نشده در جدول ۴-۴ نشان داده شده است. لیست موجود در این جدول تنوع نحوه نگارش کلماتی را نشان می‌دهد که الگوریتم پیشنهادی موفق به شناسایی آن‌ها شده است.

جدول ۴-۴: لیست نمونه‌هایی از کلمات بازیابی شده توسط روش پیشنهادی با توجه به تنوع شیوه نگارش

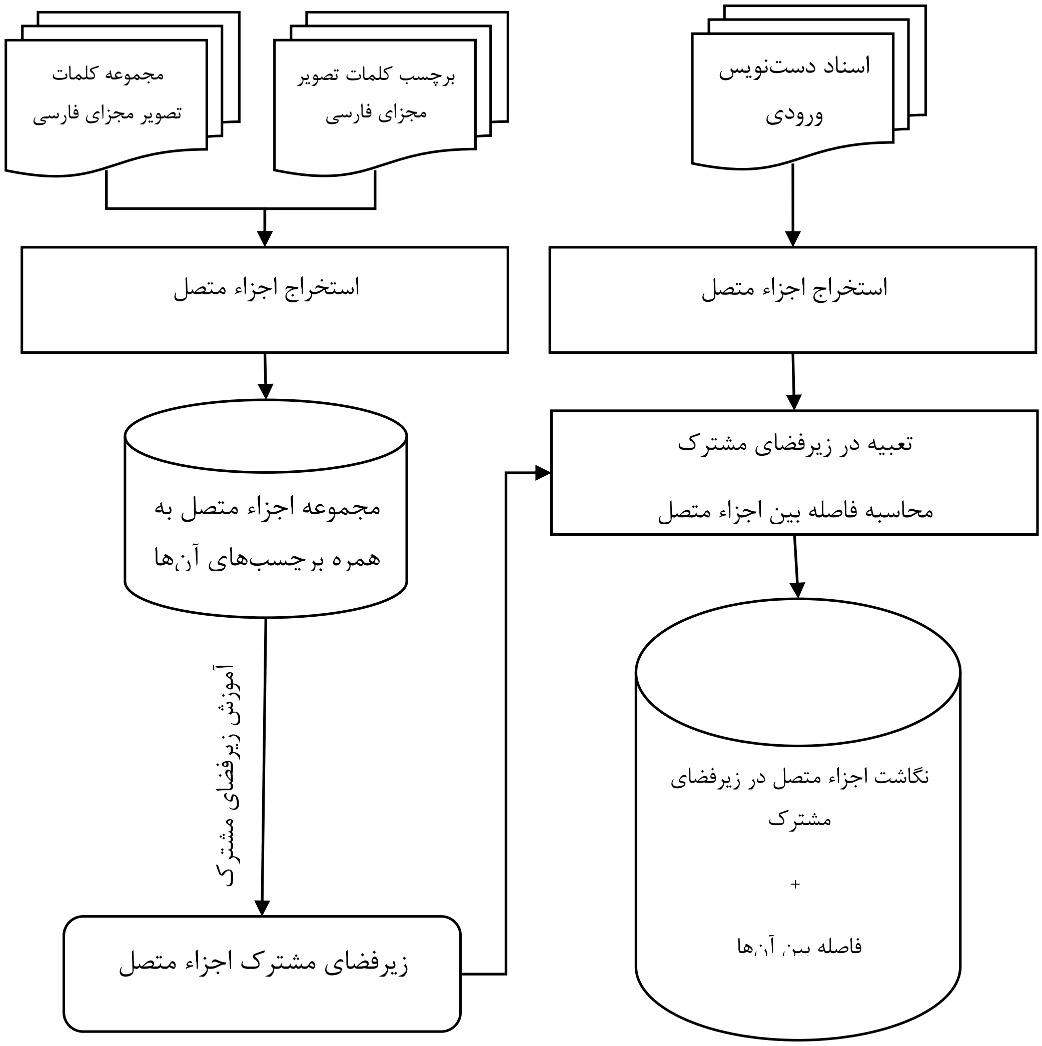
برای	
مهمان	
روز	
اقدامات	
فساد	
مدیریت	
بخش	
صنعت	
کشور	
زیبا	

فصل

۵- روش پیشنهادی سوم (مستقل از قطعه‌بندی)

۵-۱- مقدمه

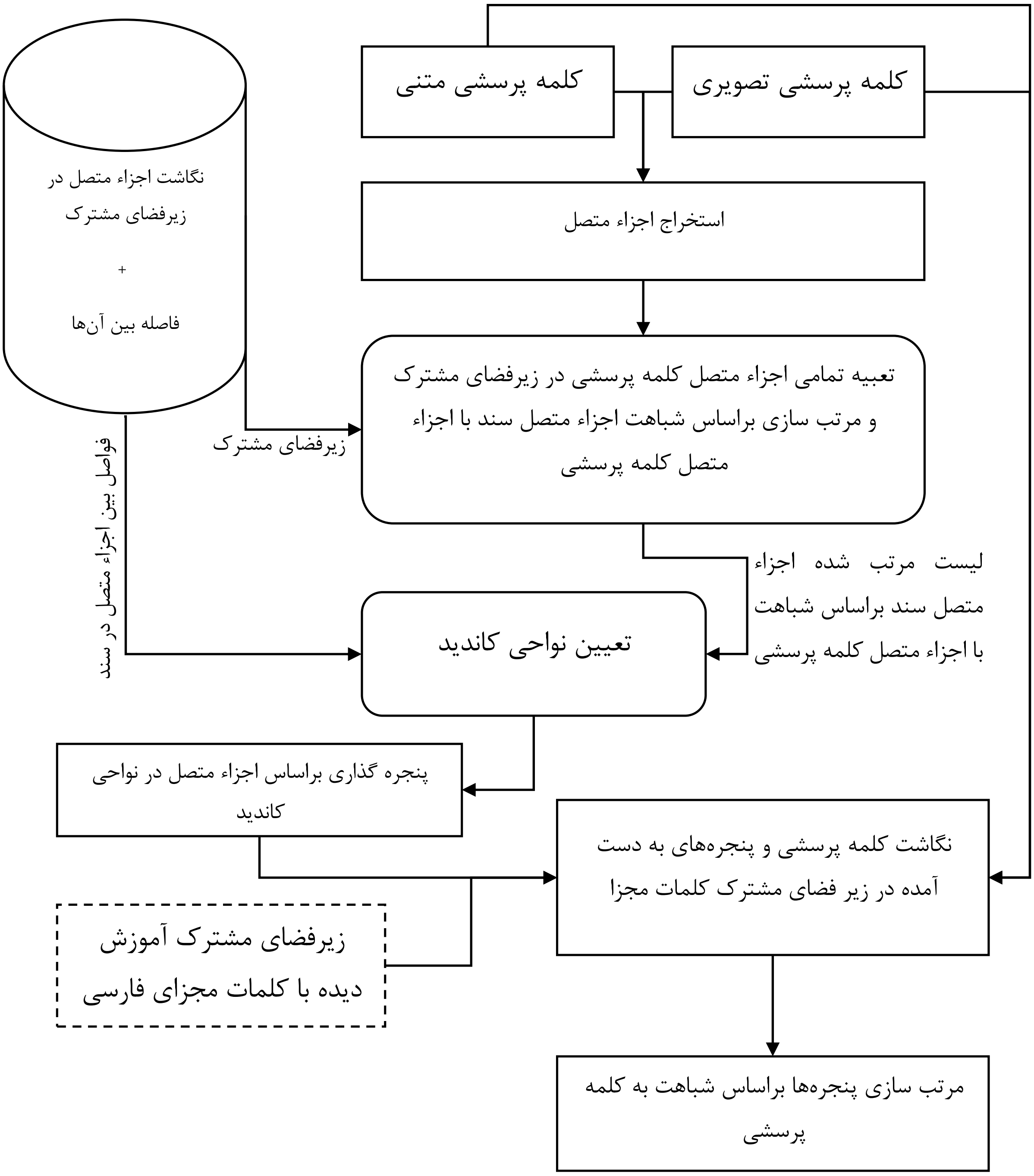
با توجه به مشکلات فراوان ذکرشده در استفاده از روش‌های مبتنی بر قطعه‌بندی، در این فصل روشی مستقل از قطعه‌بندی جهت جستجوی کلمات در اسناد دست‌نویس فارسی پیشنهاد شده است. ساختار کلی این روش بر مبنای تعیین نواحی کاندید بر اساس اجزاء متصل و سپس مرتب‌سازی مجدد[1] نواحی کاندید می‌باشد.



شکل ۵-۱: مرحله برون‌خط روش پیشنهادی

---

[1] reranking



شکل ۵-۲: مرحله برخط روش پیشنهادی

در مرحله برون‌خط روش پیشنهادی، ابتدا یک مجموعه از کلمات تصویری دارای برچسب انتخاب می‌شوند و اجزاء متصل و برچسب‌های تعبیه‌شده آن‌ها بر اساس روش پیشنهادی " هیستوگرام هرمی

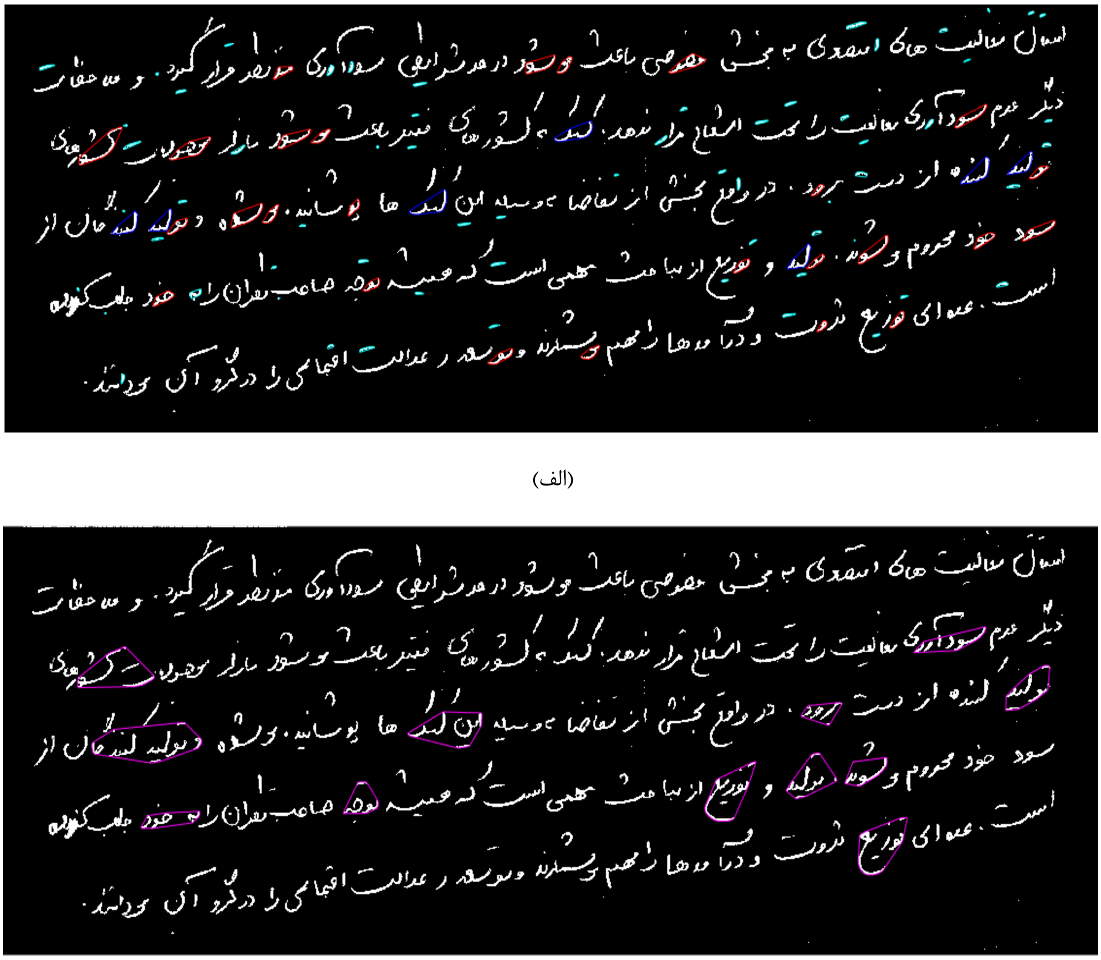
نشانه‌های موقعیتی فارسی" استخراج‌شده و زیرفضای مشترک آموزش داده می‌شود. بعدازاین مرحله، اجزاء متصل اسناد استخراج و در این زیرفضا نگاشت داده می‌شوند. در زیرفضای نگاشت داده شده فاصله کلمات استخراج می‌شود (شکل ۵-۱). از این مدل استخراج شده جهت تعیین نواحی کاندید در مرحله برخط استفاده خواهد شد. علاوه بر این یک مدل مبتنی بر زیرفضای مشترک آموزش دیده روی کلمات مجزا، مانند آنچه در فصل سوم شرح داده شده، نیز برای مرتب‌سازی مجدد نواحی کاندید مورد استفاده قرار می‌گیرد (شکل ۵-۲). در ادامه روش پیشنهادی به تفصیل شرح داده خواهد شد.

۵-۲- تعیین نواحی کاندید

جهت تعیین نواحی کاندید از اجزاء متصل استفاده شده است. ایده اصلی این پیشنهاد بر این اساس استوار است که بازیابی اجزاء متصل نیازی به قطعه‌بندی ندارد، پس می‌تواند به‌عنوان پایه‌ای جهت یک روش مستقل از قطعه‌بندی در نظر گرفته شود. ازاین‌رو درصورتی‌که سیستمی وجود داشته باشد که توانایی بازیابی اجزاء متصل را داشته باشد، می‌توان بر اساس امتیازات اجزاء متصل در سند و نزدیکی اجزاء دارای امتیاز بالا به همدیگر نواحی کاندید را تعیین نمود. بنابراین سیستمی مبتنی بر روش تعبیه برچسب معرفی‌شده در فصل سوم برای این منظور وفق داده شده است که توانایی بازنمایی اجزاء متصل را در دست‌نوشته‌های فارسی داشته باشد. سپس با استفاده از قانونی ساده مبتنی بر نزدیک بودن اجزاء متصل بازنمایی شده با امتیاز بالا نواحی کاندید تعیین شدند.

با استفاده از سیستم معرفی شده در فصل سوم با استفاده از روش تعبیه برچسب و خصیصه‌ها می‌توان سیستمی طراحی کرد که توانایی جستجو و بازیابی اجزاء متصل نوشته‌های دست‌نویس فارسی را داشته باشد. اگر در طراحی سیستم مجموعه خصیصه‌های معرفی شده با نام "هیستوگرام هرمی نشانه‌های موقعیتی فارسی" استفاده شود، ازآنجایی‌که بدنه حروف و نقاط و سرکش‌ها به‌صورت مجزا در نظر گرفته می‌شود توانایی بازیابی و جستجوی اجزاء متصل را خواهد داشت. زیرا بازنمایی اجزاء متصل با توجه به جداول ۳-۱ و ۳-۲ به‌صورت مجزا امکان‌پذیر است.

بنابراین مجموعه‌ای از اجزاء متصل و برچسب‌های آن‌ها جهت آموزش خصیصه‌ها و زیرفضای مشترک موردنیاز است. بعد از انجام این آموزش هر سند می‌تواند بر اساس مجموعه‌ای از اجزاء متصل تشکیل‌دهنده خود بازنمایی شود. این بازنمایی بر اساس نگاشت اجزاء متصل در زیرفضای مشترک و استخراج حداقل فواصل بین اجزاء متصل سازنده سند انجام می‌پذیرد. این مراحل به‌صورت برون‌خط انجام می‌شود و یک مدل از هر سند بر این اساس آماده می‌شود که در زمان جستجو مورداستفاده قرار می‌گیرد.



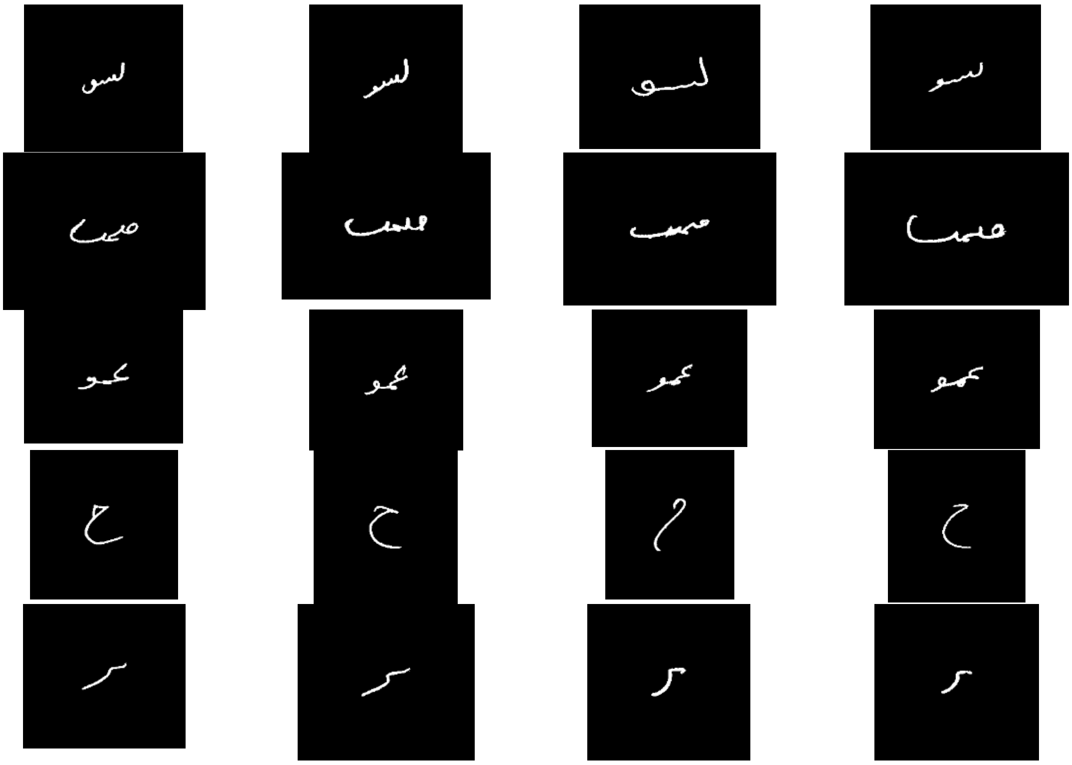
(الف)

(ب)

شکل ۵-۳: الف) نمونه‌ای از بازیابی اجزاء متصل در سند بر اساس کلمه پرسشی "تولید". هر یک از اجزاء متصل کلمه پرسشی در سند نمونه با رنگی متفاوت نشان داده شده است. ب) نواحی کاندید تعیین شده بر اساس اجزاء متصل بازیابی شده.

در زمان جستجو (برخط) اجزاء متصل کلمه پرسشی استخراج می‌شود. در مورد کلمه پرسشی تصویری، این عمل با استفاده از تکنیک‌های پردازش تصویر انجام می‌پذیرد. اما در مورد کلمه پرسشی متنی، مجموعه‌ای از قوانین مبتنی بر زبان فارسی این کار را انجام می‌دهد. این قوانین بر اساس ساختار نوشتاری زبان فارسی تعیین شده است. برای مثال چون همیشه جستجو روی یک کلمه انجام می‌پذیرد، بنابراین اجزاء متصل کلمه پرسشی دو دسته هستند: ۱- نقاط و سرکش‌ها ۲- بدنه‌های اصلی زیر کلمات که از مفهوم بدنه‌های اصلی حروف مشتق شده است. نقاط به‌سادگی از حروف تشکیل‌دهنده قابل‌تشخیص هستند. علت اصلی تشکیل زیرکلمات نیز حروفی مانند " ا د ذ ر ز ژ و " هستند که به حرف بعد از خود

نمی‌چسبند. با قواعدی که برحسب تشخیص این حروف کار می‌کنند می‌توان از روی رشته متنی تعداد اجزاء متصل را تعیین نمود.



شکل ۵-۴: نمونه‌هایی از اجزاء متصل استفاده‌شده در آموزش سیستم جستجو و بازیابی اجزاء متصل

پس از استخراج اجزاء متصل، آن‌ها در زیرفضای مشترک تعبیه می‌شوند و میزان شباهت اجزاء متصل سند با اجزاء متصل کلمه پرسشی محاسبه‌شده و این امتیازات بین ۰ تا ۱ نرمال‌سازی می‌شوند. حال با توجه به تعداد اجزاء متصل کلمه پرسشی، مجموعه‌های مجزا و مرتب‌شده‌ای برای هر جزء متصل کلمه پرسشی داریم که شامل اجزاء متصل سند است. بعدازاین مرحله به‌صورت سلسله مراتبی به دنبال ترکیباتی هستیم که شامل دو نمونه از دو مجموعه متفاوت بوده و فاصله آن‌ها از یک مقدار آستانه‌ای کمتر باشد (کنار هم قرار گرفته باشند). این مقدار آستانه‌ای بر اساس دو پارامتر محاسبه شده است: کمینه متوسط فاصله افقی[1] و کمینه متوسط فاصله عمودی[2].

---

[1] Min average horizontal distance (MAHD)

[2] Min average vertical distance (MAVD)

برای محاسبه کمینه متوسط فاصله افقی برای هر سند کمترین فاصله افقی هر جزء متصل در سند محاسبه می‌شود (HD) و هیستوگرام این فواصل محاسبه می‌گردد:

$$S_{HD} = hist(HD, t) \quad , \quad t = t_0, \dots, \max\{HD\} \qquad (۱\text{-}۵)$$

که t محدوده فواصل هیستوگرام است. سپس MAHD به‌صورت زیر محاسبه می‌گردد:

$$MAHD = (\sum_{k\in K} S_{HD}(k).k)(\sum_{k\in K} S_{HD}(k))^{-1} \qquad K = \{k|S_{HD}(k) > {}^{1}/_{4}\max\{S_{HD}\}\} \qquad (۲\text{-}۵)$$

در این محاسبه فواصل خیلی کم که ممکن است ناشی از اشتباهات یا فواصل اجزاء در خطوط مختلف باشد بر اساس شرایط تعیین شده برای k در نظر گرفته نمی‌شود. مقدار MAVD نیز به همین صورت محاسبه می‌گردد.

با استفاده از این دو پارامتر درصورتی‌که فاصله اجزاء متصل انتخابی از دو مجموعه متفاوت فاصله عمودی و افقی کمتر یا مساوی با مقادیر MAHD و MAVD باشد به‌عنوان یک مجموعه دوتایی انتخاب شده و از مجموعه‌های تکی حذف می‌شوند. بسته به تعداد مجموعه‌ها، این عمل برای مجموعه‌های سه عضوی و بیشتر تکرار می‌شود.

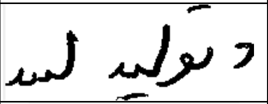
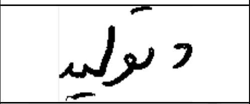
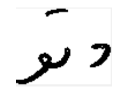
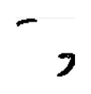
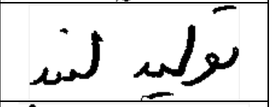
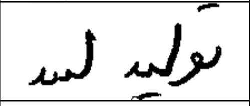
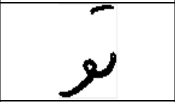
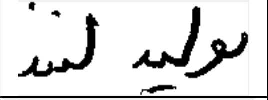
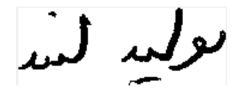
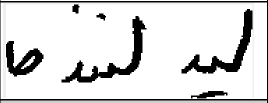
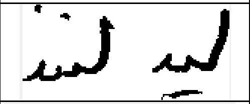
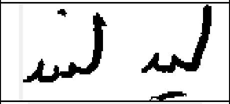
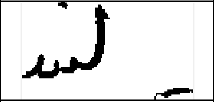
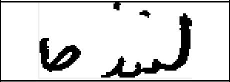
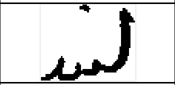
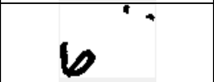
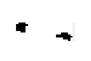
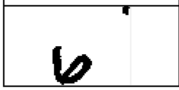
مجموعه‌های بیش از دو عضو به‌عنوان نواحی کاندید در نظر گرفته می‌شود. فقط یک حالت استثناء وجود دارد که آن‌هم در مواردی است که کلمه پرسشی دارای یک جزء متصل باشد که در این صورت این مرحله انجام نمی‌شود. نواحی بیش از دو عضوی انتخاب شده گسترش داده می‌شود، یعنی چند جزء متصل سمت چپ و راست این مجموعه به ناحیه اضافه شده و به‌عنوان ناحیه کاندید معرفی می‌شود. علت این است که با استفاده از این کار درصورتی‌که بخشی از کلمه تشخیص داده شده باشد یا چسبیدگی یا شکستگی اجزاء متصل در سند وجود داشته باشد بتوان تا حدی آن را پوشش داد. پس با استفاده از این روش نواحی کاندید مشخص می‌شوند (شکل ۵-۳ ب).

۵-۳- مرتب‌سازی مجدد

نواحی کاندیدی که در مرحله قبل تعیین شدند در این مرحله به چندین پنجره تقسیم می‌شوند. این پنجره‌ها بر اساس تعداد اجزاء متصل کلمه پرسشی تعیین می‌شوند. ازآنجایی‌که در برخی موارد در

دست‌نوشته‌ها نقاط، سرکش‌ها یا بدنه‌های یک کلمه به هم می‌چسبند برای کاهش میزان خطا پنجره‌هایی شامل تعداد ۲ جزء متصل کمتر تا ۲ جزء بیشتر از اجزاء کلمه پرسشی پنجره ایجاد می‌شود. جدول ۵-۱ نمونه‌ای از این پنجره‌های تولید شده را برای یک ناحیه کاندید برای کلمه پرسشی "تولید" نشان می‌دهد. نهایتاً، لیست پنجره‌های تولید شده به زیرفضای مشترک آموزش داده شده با کلمات مجزا نگاشت داده می‌شود و بر اساس میزان مشابهت با کلمه پرسشی وارد شده مرتب می‌شوند.

جدول ۵-۱: پنجره‌های نمونه برای یک ناحیه کاندید برای کلمه پرسشی "تولید"

ناحیه کاندید موردنظر				
				
تعداد جزء متصل در پنجره				
6	5	4	3	2
				
				
				
				
				
				
				
				

۵-۴- نتایج تجربی

برای انجام مرحله‌ی ارزیابی از پایگاه داده‌ی FHT، ۱۰ نوع متن متفاوت انتخاب شده که از هر نوع متن،۱۰دستخط متفاوت وجود دارد. ضمناً این ۱۰ دستخط متفاوت به‌صورت تصادفی از بین حالات

مختلف انتخاب شده و در مورد اسناد مختلف لزوماً یکسان نیستند. به‌طورکلی برای ارزیابی این روش از ۱۰۰ سند تصویری استفاده شده است.

جهت ارزیابی روش پیشنهادی مجموعه‌ای شامل ۶۰ کلمه‌ی متمایز به‌عنوان کلمات پرسشی انتخاب شده‌اند. ۱۰ متن انتخابی دارای موضوعات تقریباً مشابهی هستند تا فراوانی کلمات پرسشی داخل آن‌ها زیاد باشد. به همین دلیل در مجموعه‌ی کلمات پرسشی، ۲۴ کلمه در بیشتر از یک متن ظاهر شده‌اند. درمجموع تعداد این ۲۴ کلمه در داخل ۱۰ متن، ۷۸ می‌باشد. با در نظر گرفتن این ۷۸ کلمه و تعداد کلمات متمایز(۳۱ کلمه)، تعداد کل کلمات پرسشی که برای ۱۰ متن در نظر گرفته شده، ۱۰۹ می‌باشد. برای هرکدام از این ۱۰۹ کلمه، حداقل ۱۰ سند با دستخط متفاوت وجود دارد. بیشترین فراوانی یک کلمه در داخل یک متن ۷ و کمترین آن یک است. جدول ۴-۵-۲ تعداد کلمات پرسشی با فراوانی بین ۱ تا ۷ در هر سند مجزا را نشان می‌دهد. با در نظر گرفتن تعداد فراوانی کلمات، به‌طورکلی ۱۸۱۰ نمونه کلمه در این ۱۰۰ سند وجود دارد که این روش به جستجوی آن‌ها پرداخته است. متوسط تعداد نمونه‌های کلمات پرسشی در متن برابر با ۱/۵۶۵۱می‌باشد. مجموعه‌ی کلمات پرسشی از جهت تعداد حروف تشکیل‌دهنده‌ی کلمه متنوع می‌باشند، به‌طوری‌که شامل کلماتی با تعداد حروف کم مثل "ما"، "در"، "شد" و "دو"، کلمات با تعداد حروف متوسط مثل"دارو"،"کشور"، "مردم" و "درآمد" و کلماتی با تعداد حروف زیاد مثل "مدیریت"، "اولویت"،" اقدامات" و "مستقیم" است.

جدول ۵-۲: تعداد کلمات با فراوانی بین ۱ تا ۷ در هر سند مجزا

تعداد نمونه‌های کلمات در متن	۱	۲	۳	۴	۵	۶	۷
تعداد کلمات	۶۷	۲۷	۷	۴	۲	۱	۱

جهت ارزیابی استخراج نواحی کاندید به معیاری کارا نیاز است. در مسائلی که نواحی کاندید تعیین می‌شوند، همیشه یک تقابل بین میزان صحت تشخیص این نواحی و میزان کاهش فضای جستجو وجود دارد. ازاین‌رو برای بررسی میزان اثربخشی روش پیشنهادی این دو مسئله باید مورد ارزیابی قرار گیرد. پس در تعیین نواحی کاندید نرخ تشخیص نواحی کاندید صحیح باید مدنظر قرار داده شود و همچنین نرخ کاهش فضای جستجو انجام شده با استفاده از روش تعیین نواحی کاندید بایستی در نظر گرفته شود.

برای مسئله اول معیار نرخ تشخیص پیشنهاد شده است. این معیار نشان دهنده تعداد کل نواحی تشخیص داده شده صحیح بر تعداد کل نواحی صحیح است و به‌صورت زیر تعریف می‌شود:

$$DR = \frac{\sum_{d\epsilon D}\sum_{q\in Q} Positive\_Selected\_Candidate(d,q,o)}{\sum_{d\epsilon D}\sum_{q\in Q} actual\_occurrence(d,q)} \qquad (۵-۳)$$

که D تعداد اسناد موجود در مجموعه داده مورد بررسی و Q تعداد کلمات پرسشی مورد بررسی در هر سند را نشان می‌دهد. اگر کل سند به‌عنوان ناحیه کاندید در نظر گرفته شود، به دلیل اینکه قطعاً شامل ناحیه کاندید مورد جستجو خواهد بود، این معیار ۱۰۰٪ خواهد شد. ازاین‌رو نیاز به معیار دیگری نیز وجود دارد که میزان کاهش فضای جستجو را نمایش دهد. برای این کار سه معیار متفاوت معرفی شده است.

معیار میزان کاهش تعداد اجزاء متصل[1] که نرخ کاهش فضای جستجو را بر اساس تعداد اجزاء متصل در نواحی کاندید نشان می‌دهد و به‌صورت زیر تعریف شده است:

$$CCRR = \frac{\sum_{d\in D} CC\_count(d) - CC\_in\_candidateregion(d))}{\sum_{d\in D} CC\_count(d)} \qquad (۵-۴)$$

معیار میزان کاهش بر اساس تعداد پیکسل نیز به‌صورت زیر تعریف شده است:

$$RR = \frac{\sum_{d\in D}(\sum_{c\in CC} pixel\_count(c,d) - \sum_{cp\in CandidateCC} pixel\_count(cp,d))}{\sum_{d\in D}\sum_{c\in CC} pixel\_count(c,d)} \qquad (۵-۵)$$

در این روابط d اسناد، c تعداد اجزاء متصل، cc تعداد اجزاء متصل در سند و cp تعداد اجزاء متصل در نواحی کاندید را نشان می‌دهد. Pixel_count تعداد پیکسل‌های اجزاء متصل را محاسبه می‌کند.

معیار کاهش بر اساس تعداد پنجره‌ها با روش مطرح شده در فصل ۴ که مبتنی بر خط کار می‌کند نیز مقایسه شده است. این نرخ کاهش تعداد پنجره‌ها را نشان می‌دهد و به‌صورت زیر مطرح شده است:

$$WRR = \frac{\sum_{d\in D}\sum_{q\in Q} Window\_Count(d,q)}{\sum_{d\in D} Query\_count(d) \times Window\_count\_Line(d)} \qquad (۵-۶)$$

اسناد در این رابطه با d نشان داده شده و q نشان دهنده کلمات پرسشی است. Windows_count تعداد پنجره‌های یک سند به ازای یک کلمه پرسشی خاص را نشان می‌دهد.

جدول ۵-۳ نرخ تشخیص را به ازای میزان اشتراک ناحیه کاندید با ناحیه واقعی نشان می‌دهد. بر اساس نتایج تجربی، روش تعیین نواحی کاندید پیشنهادی بین ۹۰ تا ۹۴ درصد از کاندیدهای موجود را به

---

[1] Connected Components Reduction Rate (CCRR)

درستی تشخیص می‌دهد. با توجه به پیچیدگی دست‌نوشته‌های فارسی و ساده بودن مدل پیشنهادی، نتایج قابل‌قبول به نظر می‌رسد. علاوه بر این با توجه به از دست دادن بخشی از نواحی کاندید، فضای جستجو بسیار کاهش می‌یابد که نتایج تجربی این کاهش بر اساس معیارهای پیشنهادی در جدول ۵-۴ نشان داده شده است.

جدول ۵-۳: نرخ تشخیص بر اساس میزان اشتراک ناحیه کاندید با ناحیه واقعی

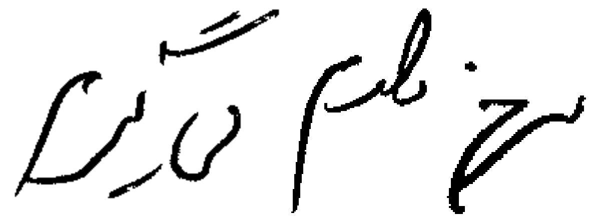
نرخ تشخیص (DR)	نرخ اشتراک
۹۴/۴۶	>0%
۹۳/۰۷	>=50%
۹۲/۴۳	>=60%
۹۱/۷۴	>=70%
۹۰/۷۰	>=80%
۹۰/۰۲	>=90%

جدول ۵-۴: نرخ کاهش بر اساس معیارهای معرفی‌شده

نرخ	معیار
۶۳/۰۹٪	RR
۷۳/۸۰٪	CCRR
۷۶/۶۵٪	WRR

بر اساس نتایج حاصل‌شده، روش معرفی‌شده دقتی در حدود ۹۰ درصد در سخت‌گیرانه‌ترین شرایط توانایی تشخیص نواحی کاندید را داراست و علاوه به راین در حدود ۷۰ درصد از فضای جستجو را کاهش می‌دهد. بر اساس تحلیل خطاهای صورت گرفته، اکثر کاندیدای ازدست‌رفته در اثر مشکلات نوشتاری یا ایراد دستگاه‌های تصویربرداری حاصل می‌شوند. این ایرادات شامل موارد زیر است:

- به هم چسبیدگی اجزاء متصل: در دست‌نوشته‌ها این مسئله مشکل رایجی است که ممکن است بین کلمات کنار هم دریک خط یا بین کلمات دو خط متوالی رخ دهد. برخی نمونه‌ها در مجموعه داده مورد استفاده در شکل ۵-۵ نمایش داده شده است.

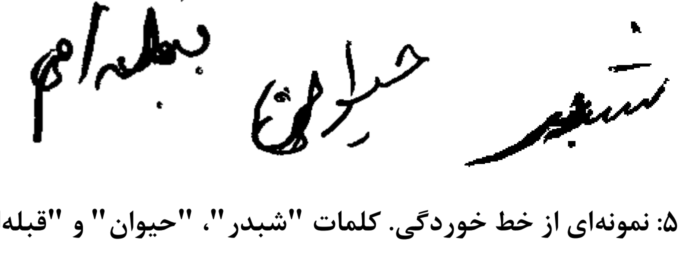


شکل ۵-۵: نمونه‌ای از به هم چسبیدگی اجزاء متصل. کلمات "سرخ"، "دارم" و "می‌گیرم"

- شکستگی یک جزء متصل به دو یا چند بخش: در برخی شرایط یک جزء متصل به دو یا چند جزء متصل شکسته می‌شود. نمونه‌هایی از این حالت درمجموعه داده مورد استفاده در شکل ۶-۵ نشان داده شده است.

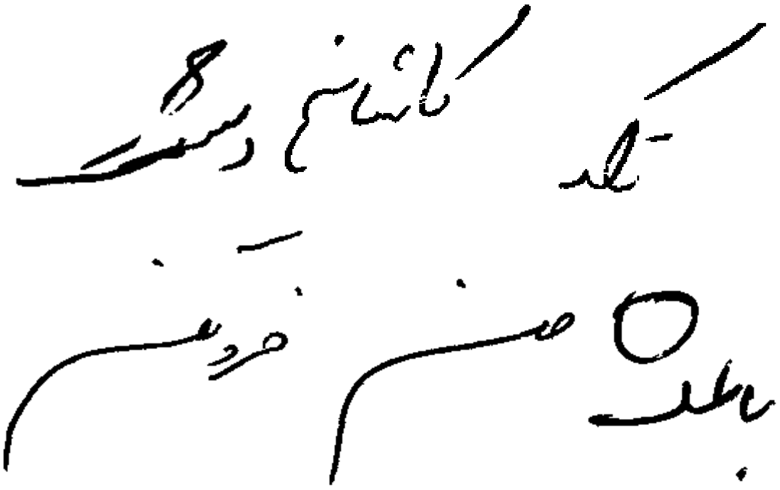
شکل ۵-۶: نمونه‌ای از شکستگی اجزاء متصل. کلمات "بخش" و "دوستانی"

- خط‌خوردگی یا اصلاح نوشتار و بنابراین خارج شدن از فرم استاندارد نگارشی: این حالت خاص نیز در مجموعه داده وجود دارد که باعث کاهش دقت بازیابی چه در سطح اجزاء متصل و چه در سطح کلمه می‌شود.



شکل ۵-۷: نمونه‌ای از خط خوردگی. کلمات "شبدر"، "حیوان" و "قبله‌ام"

- نوشتن با ساختاری متفاوت از استاندارد نگارشی: در برخی موارد دستخط و شیوه نگارش یک کلمه خاص بسیار متفاوت خواهد بود که روش پیشنهادی توانایی شناسایی برخی از حالات که تفاوت زیادی دارند را ندارد (شکل ۵-۸)



شکل۸-۵: نمونه‌ای از ساختار متفاوت با استانداردهای نگارشی. کلمات "تکه"، "کاشانم"، "دشت"، "باعث"، "من" و "فروتن"

- شباهت بخشی از یک کلمه در کلمه‌ای دیگر مانند "در" موجود در کلمه "درآمد"

ارزیابی کل سیستم بر اساس دو معیار معرفی‌شده در فصل سوم انجام شده است: معیار میانگین دقت متوسط (MaP) و معیار دقت در اولین بازیابی (P@1). نتایج این ارزیابی‌ها به ترتیب در جداول ۵-۵ و ۵-۶ نشان داده شده است.

جدول ۵-۵: میانگین دقت متوسط (MaP) برای کلمات پرسشی تصویری و متنی

KCCA	CCA	
۸۰/۲۸	۸۰/۱۴	کلمه پرسشی متنی
۷۵/۰۷	۷۴/۲۵	کلمه پرسشی تصویری

جدول ۵-۶: میانگین دقت اولین بازیابی (P@1) برای کلمات پرسشی تصویری و متنی

KCCA	CCA	
۷۹/۵۳	۷۹/۲۹	کلمه پرسشی متنی
۷۳/۸۶	۷۲/۲۱	کلمه پرسشی تصویری

نتایج تجربی نشان می‌دهد که در روش ارائه‌شده همواره دقت برای کلمات پرسشی تصویری کمتر از کلمات پرسشی متنی است. علت این امر را می‌توان در نگاشت متن و تصویر در فضای خصیصه پیگیری کرد. برای نگاشت متن به فضای خصیصه از مجموعه‌ای از قوانین مرتبط با ساختار زبان فارسی و حروف

تشکیل‌دهنده کلمه پرسشی استفاده شده و یک بردار دودویی به‌عنوان بازنمایی در فضای خصیصه ساخته می‌شود. اما عمل نگاشت در مورد کلمه تصویری از مجموعه‌ای از ماشین‌های بردار پشتیبان بهره می‌برد که توسط مجموعه داده آموزش داده‌شده است. با توجه به اینکه تبدیل در مورد متن بدون خطا و در مورد تصویر در بهترین حالات نیز با خطا همراه است، بنابراین تفاوت دقت دراین‌باره منطقی به نظر می‌رسد.

فصل

۶- نتیجه‌گیری و کارهای آینده

۶-۱- نتیجه‌گیری

در این پایان‌نامه یک رهیافت برای جستجوی کلمه پرسشی در اسناد دست‌نوشته‌ی زبان فارسی پیشنهاد گردید. در این روش کلمه‌ی پرسشی موردنظر کاربر در اسناد تصویری جستجو شده و مکان نمونه‌های آن کلمه در داخل اسناد مشخص می‌شود.کلمه پرسشی می‌تواند به‌صورت تصویر یا رشته‌ی متنی باشد.

یکی از مهم‌ترین مسائل در آنالیز اسناد تصویری، مانند سایر حوزه‌های بینایی ماشین، روش بازنمایی است که در دهه‌های اخیر کارهای بسیار زیادی در جهت حل این مشکل انجام شده است. روش بازنمایی قابل‌قبول است که نسبت به تغییرات مقاوم بوده و قابلیت تفکیک‌پذیری بالایی نیز داشته باشد. بنابراین در حل مسئله مذکور در این رساله، در اولین گام به دنبال یک روش بازنمایی مناسب بوده‌ایم که توانایی مدل کردن مستقل از نوع نگارش را داشته باشد. برای حل مسئله مذکور روشی مبتنی بر تعبیه برچسب، رده‌بندی مبتنی بر خصیصه معرفی شد که با آموزش یک زیرفضای مشترک مسئله را به یک مسئله ساده نزدیک‌ترین همسایگی تبدیل می‌کند. خصیصه‌های مختلفی بر اساس ساختار زبان فارسی معرفی شد که توانایی بازنمایی بهتری را در زبان فارسی داشته باشیم. یکی از این خصیصه‌ها که بدنه حروف را از نقاط و سرکش‌ها به‌صورت مجزا در نظر می‌گیرد (PPHOS) این قابلیت را به سیستم اضافه می‌کند که بتواند علاوه بر کلمات، اجزاء متصل سازنده کلمات را نیز بازیابی کند.

جهت جستجوی کلمه مستقل از قطعه‌بندی سند دو روش پیشنهاد داده شد. اولین روش مبتنی بر قطعه‌بندی خطوط می‌باشد. در این روش، خطوط سند استخراج شده و پنجره‌گذاری مبتنی بر اجزاء متصل با طول ۲ تا ۱۰ جزء ایجاد می‌گردد. این پنجره‌ها در زیر فضای مشترک آموزش داده شده نگاشت می‌شوند. برای هر کلمه پرسشی، متنی یا تصویری، عمل نگاشت به زیر فضای مشترک انجام می‌شود و بر اساس نزدیکی فاصله به این کلمات پنجره‌های سند بازیابی می‌شوند. در این روش مشکلات پنجره-گذاری ساده وجود ندارد و تعداد پنجره به مراتب کمتر است. علاوه بر این دقت مطلوبی گزارش می‌شود.

در روش دوم به‌طور کامل مستقل از هر نوع قطعه‌بندی عمل شده است. بر اساس خصیصه‌ی معرفی شده‌ای که توانایی جستجوی اجزاء متصل سازنده کلمات فارسی را دارد، یک روش دو مرحله ای پیشنهاد داده شد. این روش نگاشت اجزاء متصل هر سند را در زیرفضای مشترک آموزش داده شده به همراه فاصله آن‌ها از یکدیگر ذخیره می‌کند. برای هر کلمه پرسشی ابتدا اجزاء متصل استخراج شده و بر اساس میزان شباهت، اجزاء متصل مشابه در سند استخراج می‌شوند. بر اساس انباشتگی اجزاء در موقعیت‌های مکانی خاصی از سند نواحی کاندید تعیین می‌شود. سپس این نواحی کاندید بر اساس یک زیرفضای مشترک آموزش داده شده با کلمات مجزا نسبت به کلمه پرسشی مرتب می‌شوند.

با توجه به اینکه در مجموعه داده استفاده شده تنوع دستخطها بسیار زیاد بود و مجموعه آموزش از مجموعه آزمون در تمامی مراحل جدا در نظر گرفته شد، نتایج تجربی حاصل‌شده بر توانایی و قدرت روش پیشنهادی صحه گذاشت. علاوه بر این، روش ارائه‌شده مستقل از تمامی محدودیت‌هایی که در مورد روش‌های ارائه‌شده در جستجوی کلمه در نظر گرفته می‌شود، کار می‌کند. یعنی این روش مستقل از فرهنگ لغت کار می‌کند و توانایی شناسایی کلمات خارج از فرهنگ لغت را نیز دارد. همچنین بر روی اسناد چند نویسنده با یک مدل ثابت کار می‌کند و نیاز به مدل‌های متفاوت برای اسناد متفاوت ندارد. توانایی کار کردن با کلمات پرسشی تصویری و متنی را داراست، هر چند در تمامی مراحل دقت کلمه پرسشی متنی بهتر بوده است. علاوه بر این، روش مستقل از قطعه‌بندی پیشنهادی مشکلات و خطاهای قطعه‌بندی را ندارد.

۶-۲- کارهای آینده

در آزمایش‌های اولیه، جهت آموزش بهتر خصیصه‌ها اهمیت وجود یک مجموعه داده آموزشی بزرگ که توانایی بازنمایی خوبی داشته باشد مشخص شد. مسئله وجود یک مجموعه داده آموزشی بزرگ در تمامی کارهای شناسایی الگو همواره به‌عنوان یک مسئله مهم مطرح بوده است. اما در روش پیشنهادی به دلیل ساختار خاص خصیصه‌های معرفی‌شده این مسئله به نحوی متفاوت مطرح می‌شود. به دلیل اینکه در تمامی سطوح مانند کل کلمه، نصف اول و دوم کلمه و یک سوم اول دوم و سوم خصیصه‌ها که شامل تمامی حروف الفبا هستند آموزش داده شوند، باید مجموعه داده طوری آماده شود که امکان چنین حالتی را ممکن سازد. این اهمیت در خصوص زیرسیستم بازیابی اجزاء متصل اهمیت بیشتری دارد. مجموعه داده‌هایی که وجود دارند و از آن‌ها در این رساله استفاده شده است، نمونه‌های مثبت کافی برای آموزش برخی از خصیصه‌ها را فراهم نمی‌کند. حتی در بعضی موارد برخی از خصیصه‌ها هیچ نمونه‌ای برای آموزش در دسترس ندارند. فراهم کردن مجموعه داده بزرگ کاری بسیار زمان‌بر و خسته‌کننده است. علاوه بر این تهیه فهرستی از کلاس کلمات مناسبی که بتواند به‌خوبی تمام کلاس خصیصه‌های موردنظر را آموزش دهد مشکل است. بنابراین به‌عنوان یکی از راهکارهایی که می‌توان در آینده جهت گسترش این کار مطرح نمود، سنتز داده‌های آموزشی بر اساس کلاس کلمات موردنیاز خواهد بود. با این روش می‌توان داده‌ها را طوری تولید کرد که تمامی خصیصه‌ها به‌خوبی و کامل آموزش داده شود، بنابراین مدل‌های خصیصه بهتر تولید خواهد شد و نتایج به‌طور قابل‌توجهی افزایش خواهد یافت.

در نظر گرفتن کلمه پرسشی و ساختار سند به‌صورت گراف به‌عنوان راهکاری است که می‌توان در ادامه کار از آن بهره برد. در این ساختار مبتنی بر گراف هر جزء متصل به‌عنوان یک نود از گراف در نظر گرفته

می‌شود و یال‌های گراف بین اجزاء متصل نزدیک به هم در نظر گرفته می‌شود. این گراف به‌صورت نسبی اجزاء زیر کلمات، توالی زیر کلمات و کلمات را در متن سند نشان می‌دهد. در این روش می‌توان با استفاده از یک معیار شباهت مبتنی بر جهت قرار گرفتن اجزاء متصل و میزان شباهت آن به گراف تشکیل شده از توالی اجزاء متصل در کلمه پرسشی، نواحی کاندید را استخراج نماید. علاوه بر این در صورت تعیین یک معیار شباهت دقیق می‌توان عمل تشخیص را به‌صورت یک مرحله‌ای و بدون نیاز به تعیین نواحی کاندید انجام داد.

یکی از مسائل مهم در روش مستقل از قطعه‌بندی پیشنهادی، مسئله تعیین اجزاء متصل مرتبط با هم در سند است. در روش پیشنهادی صرفاً بر اساس فاصله و جهت قرار گرفتن اجزاء متصل این مسئله حل می‌شود. راهکار دیگری که می‌توان جهت ادامه کار ارائه کرد، استفاده از مدلی شبیه به مدل‌های زبانی در پردازش زبان طبیعی است. در مورد روش پیشنهادی مستقل از قطعه‌بندی که در مرحله اول اجزاء متصل را بازیابی می‌کند، می‌توان از یک مدل شبیه به مدل‌های زبانی جهت بالا بردن دقت تشخیص نواحی کاندید بهره برد. مدل‌های زبانی بر اساس احتمال رخداد توالی کلمات در پردازش زبان طبیعی در نظر گرفته می‌شود. در این مدل خاص، مدل زبانی بر اساس ساختار تصویری و بر اساس اجزاء متصل سازنده کلمات ساخته می‌شود و می‌تواند به‌عنوان راهکاری در جهت تعیین دقیق‌تر نواحی کاندید در نظر گرفته شود. در این روش می‌توان ارزیابی نواحی کاندید مبتنی بر اجزاء متوالی مجاور را بر اساس این مدل و ترکیب آن با میزان شباهت اجزاء متصل کلمه پرسشی با ناحیه کاندید در نظر گرفت. این روش می‌تواند با ساختار گرافی اجزاء متصل اسناد ترکیب شود و دقت روش مذکور را با ساخت گراف دقیق‌تر با استفاده از مدل یاد گرفته شده افزایش دهد.

مراجع

[۱] T. Kameshiro, T. Hirano, Y. Okada, and F. Yoda, "A document image retrieval method tolerating recognition and segmentation errors of OCR using shape-feature and multiple candidates," in *Document Analysis and Recognition, ICDAR'99. Proceedings of the Fifth International Conference on*, 1999, pp. 681-684.

[۲] ح. خسروی, "رویکرد یکپارچه برای بازشناسی متن چاپی فارسی," رساله دکتری, دانشگاه تربیت مدرس، ۱۳۸۷.

[۳] L. Li, S. Lu, and C. L. Tan, "A fast keyword-spotting technique," in *Document Analysis and Recognition. ICDAR 2007. Ninth International Conference on*, 2007, pp. 68-72.

[۴] G. Jones, J. Foote, K. Sparck Jones, and S. Young, "Video mail retrieval: The effect of word spotting accuracy on precision," in *Acoustics, Speech, and Signal Processing .ICASSP-95., 1995 International Conference on*, 1995, pp. 309-312.

[۵] T. M. Rath, V. Lavrenko, and R. Manmatha, "Retrieving historical manuscripts using shape," Massachusset Univ Amherst Center for Intelligent Information Retrieval, 2003.

[۶] ا. ابراهیمی, "استفاده از شکل کلی زیرکلمات چاپی در بازیابی تصویر مستندات و بازشناسی متون فارسی," رساله دکتری, مهندسی برق دانشگاه تربیت مدرس ۱۳۸۴.

[۷] J. Almazán, A. Gordo, A. Fornés, and E. Valveny, "Word Spotting and Recognition with Embedded Attributes," *IEEE transactions on pattern analysis and machine intelligence,* vol. 36, pp. 2552-2566, 2014.

[۸] A. R. A. Z.Imani, A.Zohrevand, "Introduction to Database FARSA digital images of handwritten Farsi words," presented at the Eleventh Iranian Conference on Intelligent Systems (ICIS2013), Kharazmi University. Tehran , Iran, 2013.

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

[۱۰] S. Lu and C.-L. Tan, "Keyword spotting and retrieval of document images captured by a digital camera," in *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, 2007, pp. 994-998.

[۱۱] A. Kefali and C. Chemmam, "A Semi-Automatic Approach of old Arabic Documents Indexing," in *3rd International Conference on Computer Science and its Applications (CIIA 2011)*, Saida , Algeria 2011.

[۱۲] Y. Pourasad, H. Hassibi, and A. Ghorbani, "A word spotting method for Farsi machine-printed document images," *Turkish Journal of Electrical Engineering & Computer Sciences,* vol. 21, pp. 734-746, 2013.

[۱۳] Y. POURASAD, H. HASSIBI, and A. GHORBANI, "A Farsi/Arabic Word Spotting Approach for Printed Document Images," *International Journal of Natural & Engineering Sciences,* vol. 6, 2012.

[۱۴] Y. Pourasad, H. Hassibi, and A. Ghorbani, "Farsi Word Spotting and Font Size Recognition," *Procedia Technology,* vol. 1, pp. 372-377, 2012.

[۱۵] B. Zhang, S. N. Srihari, and C. Huang, "Word image retrieval using binary features," in *Electronic Imaging 2004,* pp. 45-53.

[۱۶] S. Srihari, H. Srinivasan, P. Babu, and C. Bhole, "Spotting words in handwritten Arabic documents," in *Electronic Imaging 2006*.

[۱۷] T. Adamek, N. E. O'Connor, and A. F. Smeaton, "Word matching using single closed contours for indexing handwritten historical documents," *International Journal of Document Analysis and Recognition (IJDAR),* vol. 9, pp. 153-165, 2007.

[۱۸] H. Cao and V. Govindaraju, "Template-free word spotting in low-quality manuscripts," in *Proceedings of the 6th International Conference on Advances in Pattern Recognition*, 2007, pp. 135-139.

[۱۹] T. M. Rath, R. Manmatha, and V. Lavrenko, "A search engine for historical manuscript images," in *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, 2004, pp. 369-376.

[۲۰] T. M. Rath and R. Manmatha, "Word image matching using dynamic time warping," in *Computer Vision and Pattern Recognition. Proceedings. IEEE Computer Society Conference on*, 2003, pp. II-521-II-527 vol. 2.

[۲۱] K. Terasawa and Y. Tanaka, "Slit style HOG feature for document image word spotting," in *Document Analysis and Recognition. ICDAR'09. 10th International Conference on*, 2009, pp.۱۲۰-۱۱۶ .

[۲۲] T. M. Rath and R. Manmatha, "Word spotting for historical documents," *International Journal of Document Analysis and Recognition (IJDAR)*, vol. 9, pp. 139-152, 2007.

[۲۳] R. Manmatha and T. M. Rath, "Indexing of handwritten historical documents-recent progress," presented at the Proceedings 2003 Symposium on Document Image Understanding Technology.(UMD), 2003.

[۲۴] S. N. Srihari, H. Srinivasan, C. Huang, and S. Shetty, "Spotting words in Latin, Devanagari and Arabic scripts," *VIVEK-BOMBAY-*, vol. 16, p. 2, 2006.

[۲۵] A. Kolcz, J. Alspector, M. Augusteijn, R. Carlson, and G. V. Popescu, "A line-oriented approach to word spotting in handwritten documents," *Pattern Analysis & Applications,* vol. 3, pp. 153-168, 2000.

[۲۶] M. Dehghan, K. Faez, M. Ahmadi, and M. Shridhar, "Handwritten Farsi (Arabic) word recognition: a holistic approach using discrete HMM," *Pattern Recognition,* vol. 34, pp. 1057-1065, 2001.

[۲۷] S.-S. Kuo and O. E. Agazzi, "Keyword spotting in poorly printed documents using pseudo 2-D hidden Markov models," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 16, pp. 842-848, 1994.

[۲۸] J. E. Y. W. T. David, F. R. B. M. Maire, and G. Vesom, "Making latin manuscripts searchable using gHMM's," presented at the Advances in Neural Information Processing Systems 17, 2005.

[۲۹] J. A. Rodriguez, F. Perronnin, G. Sánchez, and J. Lladós, "Unsupervised writer style adaptation for handwritten word spotting," in *Pattern Recognition. ICPR. 19th International Conference on*, 2008, pp. 1-4.

[۳۰] V. Frinken, A. Fischer, and H. Bunke, "A novel word spotting algorithm using bidirectional long short-term memory neural networks," in *Artificial Neural Networks in Pattern Recognition*, ed: Springer, 2010, pp. 185-196.

[۳۱] M. Khayyat, L. Lam, and C. Y. Suen, "Learning-based word spotting system for Arabic handwritten documents," *Pattern Recognition,* vol. 47, pp. 1021-1030, 2014.

[۳۲] G. R. Ball, S. N. Srihari, and H. Srinivasan, "Segmentation-based and segmentation-free methods for spotting handwritten arabic words," presented at the Tenth International Workshop on Frontiers in Handwriting Recognition, Suvisoft, 2006.

[۳۳] S. Srihari, H. Srinivasan, P. Babu, and C. Bhole, "Handwritten arabic word spotting using the cedarabic document analysis system," in *Proc. Symposium on Document Image Understanding Technology (SDIUT-05)*, 2005, pp. 123-132.

[۳۴] J. Almazán, A. Gordo, A. Fornés, and E. Valveny, "Segmentation-free word spotting with exemplar svms," *Pattern Recognition,* vol. 47, pp. 3967-3978, 2014.

[۳۵] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the Fisher vector: Theory and practice," *International journal of computer vision,* vol. 105, pp. 222-245, 2013.

[۳۶] R. Shekhar and C. Jawahar, "Word image retrieval using bag of visual words," in *Document Analysis Systems (DAS), 10th IAPR International Workshop on*, 2012, pp. 297-301.

[۳۷] S. Dey, A. Nicolaou, J. Llados, and U. Pal, "Evaluation of the Effect of Improper Segmentation on Word Spotting," *arXiv preprint arXiv:1604.06243,* 2016.

[۳۸] R. Manmatha, C. Han, and E. M. Riseman, "Word spotting: A new approach to indexing handwriting," in *Computer Vision and Pattern Recognition. Proceedings CVPR'96, IEEE Computer Society Conference on*, 1996, pp. 631-637.

[۳۹] A. Fischer, A. Keller, V. Frinken, and H. Bunke, "HMM-based word spotting in handwritten documents using subword models," in *Pattern recognition (icpr), 20th international conference on*, 2010, pp. 3416-3419.

[۴۰] J. Rodriguez-Serrano and F. Perronnin, "A model-based sequence similarity with application to handwritten word spotting," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 34, pp. 2108-2120, 2012.

[۴۱] J. Almazán, "Learning to represent handwritten shapes and words for matching and recognition," PHD, Universitat Autònoma de Barcelona - UAB Barcelona, 2014.

[۴۲] A. Abidi, A. Jamil, I. Siddiqi, and K. Khurshid, "Word spotting based retrieval of urdu handwritten documents," in *Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference on*, 2012, pp. 331-336.

[۴۳] م. علی‌آبادی, ج. صدری, س. م. رضوی, "یک روش ترکیبی برای جست وجوی کلمات در کتب تاریخی دست نویس فارسی," اولین کنفرانس بازشناسی الگو و پردازش تصویر ایران, ۱۳۹۱.

[۴۴] S. Yao, Y. Wen, and Y. Lu, "HoG based two-directional Dynamic Time Warping for handwritten word spotting," in *Document Analysis and Recognition (ICDAR), 13th International Conference on*, 2015, pp. 161-165.

[۴۵] T. Mondal, N. Ragot, J.-Y. Ramel, and U. Pal, "Flexible Sequence Matching Technique: An Effective Learning-free Approach For word-spotting," *Pattern Recognition,* vol. 60, pp. 596-612, 2016.

[۴۶] J. A. Rodríguez-Serrano and F. Perronnin, "Handwritten word-spotting using hidden Markov models and universal vocabularies," *Pattern Recognition,* vol. 42, pp. 2106-2116, 2009.

[۴۷] A. Fischer, A. Keller, V. Frinken, and H. Bunke, "Lexicon-free handwritten word spotting using character HMMs," *Pattern Recognition Letters,* vol. 33 ,pp. 934-942, 2012.

[۴۸] V. Frinken, A. Fischer, R. Manmatha, and H. Bunke, "A novel word spotting method based on recurrent neural networks," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 34, pp. 211-224, 2012.

[۴۹] S. Wshah, G. Kumar, and V. Govindaraju, "Script independent word spotting in offline handwritten documents based on hidden markov models," in *Frontiers in Handwriting Recognition (ICFHR), International Conference on*, 2012, pp. 14-19.

[۵۰] Z. S. Harris, "Distributional structure" *Word,* vol 10, pp 146-162, 1954.

[۵۱] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Workshop on statistical learning in computer vision, ECCV,* 2004, pp. 1-2.

[۵۲] F. Perronnin, "Universal and adapted vocabularies for generic visual categorization," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 30, pp. 1243-1256, 2008.

[۵۳] D. Aldavert, M. Rusiñol, R. Toledo, and J. Lladós, "A study of Bag-of-Visual-Words representations for handwritten keyword spotting," *International Journal on Document Analysis and Recognition (IJDAR),* vol. 18, pp. 223-234, 2015.

[۵۴] L. Rothacker, M. Rusinol, and G. A. Fink, "Bag-of-features HMMs for segmentation-free word spotting in handwritten documents," in *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*, 2013, pp. 1305-1309.

[۵۵] M. W. Sagheer, N. Nobile, C. L. He, and C. Y. Suen, "A novel handwritten urdu word spotting based on connected components analysis," in *Pattern Recognition (ICPR), 20th International Conference on*, 2010, pp. 2013-2016.

[۵۶] M. Kassis and J. El-Sana, "Automatic Synthesis of Historical Arabic Text for Word-Spotting," in *Document Analysis Systems (DAS), 2016 12th IAPR Workshop on*, 2016, pp 239-244.

[۵۷] T. Konidaris, B. Gatos, K. Ntzios, I. Pratikakis, S. Theodoridis, and S. J. Perantonis, "Keyword-guided word spotting in historical printed documents using synthetic data and user feedback," *International Journal of Document Analysis and Recognition (IJDAR),* vol. 9, pp. 167-177, 2007.

[۵۸] J. A. Rodriguez-Serrano and F. Perronnin, "Synthesizing queries for handwritten word image retrieval," *Pattern Recognition,* vol. 45, pp. 3270-3276, 2012.

[۵۹] Y. Leydier, F. Lebourgeois, and H. Emptoz, "Text search for medieval manuscript images," *Pattern Recognition,* vol. 40, pp. 3552-3567, 2007.

[۶۰] Y. Leydier, A. Ouji, F. LeBourgeois, and H. Emptoz, "Towards an omnilingual word retrieval system for ancient manuscripts," *Pattern Recognition,* vol. 42, pp. 2089-2105, 2009.

[۶۱] J. Almazán, A. Gordo, A. Fornés, and E. Valveny, "Word spotting and recognition with embedded attributes," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 36, pp. 2552-2566, 2014.

[۶۲] J. A. Rodriguez-Serrano ,F. Perronnin, and F. Meylan, "Label embedding for text recognition," in *Proc. BMVC*, 2013.

[۶۳] G. Sfikas, A. P. Giotis, G. Louloudis, and B. Gatos, "Using attributes for word spotting and recognition in polytonic greek documents," in *Document Analysis and Recognition (ICDAR), 13th International Conference on*, 2015, pp. 686-690.

[۶۴] S. K. Ghosh and E. Valveny, "A Sliding Window Framework for Word Spotting Based on Word Attributes," in *Pattern Recognition and Image Analysis: 7th Iberian Conference ,IbPRIA, Santiago de Compostela, Spain, June 17-19, Proceedings*, R. Paredes, J. S. Cardoso, and X. M. Pardo, Eds., ed Cham: Springer International Publishing, 2015, pp. 652-661.

[۶۵] S. Sudholt and G. A. Fink, "PHOCNet: A Deep Convolutional Neural Network for Word Spotting in Handwritten Documents," *arXiv preprint arXiv:1604.00187,* 2016.

[۶۶] C. H. Lampert, "Learning to detect unseen object classes by between-class attribute transfer," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 2009, pp. 951-958.

[۶۷] C. H. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorization," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 36, pp. 453-465, 2014.

[۶۸] J. Almazán, A. Gordo, A. Fornés, and E. Valveny, "Handwritten Word Spotting with Corrected Attributes," in *IEEE International Conference on Computer Vision (ICCV)*, 2013.

[۶۹] J. A. Rodriguez-Serrano and F. C. Perronnin, "Label-embedding for text recognition," ed: Google Patents, 2015.

[۷۰] J. A. Rodriguez-Serrano, A. Gordo, and F. Perronnin, "Label embedding: A frugal baseline for text recognition," *International Journal of Computer Vision*, vol. 113, pp. 193-207, 2015.

[۷۱] S. Bengio, J. Weston, and D. Grangier, "Label Embedding Trees for Large Multi-Class Tasks," in *NIPS*, 2010.

[۷۲] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid, "Label-embedding for attribute-based classification," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, 2013, pp. 819-826.

[۷۳] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *Computer Vision and Pattern Recognition,. CVPR. IEEE Conference on*, 2009, pp. 1778-1785.

[۷۴] S. Lazebnik, C. Schmid ,and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 2006, pp. 2169-2178.

[۷۵] J. SáNchez, F. Perronnin, and T. De Campos, "Modeling the spatial layout of images beyond spatial pyramids," *Pattern Recognition Letters,* vol. 33, pp. 2216-2223, 2012.

[۷۶] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural computation,* vol. 16, pp. 2639-2664, 2004.

[۷۷] M. B. Blaschko and C. H. Lampert, "Correlational spectral clustering," in *Computer Vision and Pattern Recognition,. CVPR. IEEE Conference on*, 2008, pp. 1-8.

[۷۸] R. Ebrahimpour, M. Amini, and A. V. Shams, "A New Combination Method Based on Different Representation of Data," *International Journal of Hybrid Information Technology,* vol. 4, pp. 51-60, 2011.

[۷۹] R. Ebrahimpour, R. D. Vahid, and B. M. Nezhad, "Decision Templates with Gradient based Features for Farsi Handwritten Word Recognition," *International Journal of Hybrid Information Technology,* vol. 4, pp. 1-12, 2011.

[۸۰] R. Ebrahimpour, S. Sarhangi, and F. Sharifizadeh, "Mixture of Experts for Persian handwritten word recognition," *Iranian Journal of Electrical & Electronic Engineering,* vol. 7, pp. 217-224, 2011.

[۸۱] Z. Imani, A. Ahmadyfard, A. Zohrevand, and M. Alipour, "Offline handwritten Farsi cursive text recognition using hidden Markov models," in *2013 8th Iranian Conference on Machine Vision and Image Processing (MVIP),* 2013, pp. 75-79.

[۸۲] R. Ebrahimpour, S. Sarhangi, M. Javadi, and F. Sharifizadeh, "Handwritten Word Recognition by Multiple Classifiers: A Divide-and-Conquer Approach," *International Journal of Applied Research on Information Technology and Computing,* vol. 2, pp. 9-21, 2011.

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

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

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

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

[٨٧] M. Arivazhagan, H. Srinivasan, and S. Srihari, "A statistical approach to handwritten line segmentation," *Document Recognition and Retrieval XIV, Proceedings of SPIE, San Jose, CA*, pp. 6500T-1, 2007.

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

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

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

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

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

[٩٣] Y. A. Shmjitsurouka and T. Yoshika WA, "The Segmentation of a text line for a handwritten unconstrained document using thinning algorithm", *the 7th International workshop on frontiers in handwritten recognition,* pp 505-510, 2000.

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

[٩۵] S. S. Bukhari, F. Shafait, and T. M. Breuel, "Script-independent handwritten textlines segmentation using active contours," in *Document Analysis and Recognition, ICDAR'09. 10th International Conference on*, 2009, pp. 446-450.

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

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

[٩٨] M. Ziaratban and F. Karim, "Adaptive Script-Independent Text Line Extraction," *IEICE transactions on information and systems,* vol. 94, pp.٢٠١١ ,٨٧٧-٨۶۶ .

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

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

Abstract

Word spotting is a way to index, search and retrieve words in document images. Word spotting aims to find multiple occurrences of a query word in document images. Word spotting is one of the major challenges in the field of document image analysis which has recently received many attentions. Several word spotting systems have been introduced on Latin script, but a limited number of works exist on Arbic and Persian documents and most of them only proposed for printed documents. The goal of this thesis is to present an end to end multiwriter Persian word spotting system which accepts, independently of lexicon, both word images and text strings.

The number of classes in this problem is very high. Moreover very large intra-class variability due to different writing styles, illumination, typography, etc, can make the same word look very different. For this reason, a classification method using attributes is needed. Attribute-based classification by sharing information between classes makes out-of-vocabulary (OOV) spotting possible. In this approach, both word images and text strings embeds in a common vectorial subspace. This is achieved by a combination of label embedding, attributes learning, and a common subspace regression. In this subspace, images and strings that represent the same word are close together, allowing one to cast recognition and retrieval tasks as a nearest neighbor problem.

In the first step, an attribute-based method is proposed to retrieve isolated handwritten Persian words. In this method we suggest two methods for representing Persion words using attributes. These methods are inspired from two facts, first the appearance of a Persian letter changes depending on its position in a word. Second, Persion letters can be categorized based their shape of main part. To evaluate the performance of proposed methods, two isolated world Persian handwritten datasets, Iranshar and Farsa, are used. Mean average precision (MaP) for presented attributes in word spotting is 95.67%, 96.25% and precision for handwritten word recognition is 96.15% and 97.55% respectively.

Then we introduce a method for segmenting handwritten documents to constitute lines. Using a new approach we slide a window on each line of a document. The proposed sliding window is based on connected components, so it can overcome to the problem of basic methods based on sliding window such as determining window size and movement step. FHT dataset is used to evaluate the proposed line-based method. MaP criterion in this experiment for query by example and query by string is 73.67 and 81.02% respectively.

Finally, a two-pass segmentation-free method is proposed for spotting of words in handwritten documents. In the first pass a number of candidate regions for the query word are selected in the document. These regions are extracted based on the proposed spotting method at connected component level. Then considering neighboring connected components to the candidate regions, the regions which likely match the query word are extracted. Evaluation of this method is also performed using FHT dataset and MaP 79.45% for text query and 70.25% for image query are obtained.

It is noteworthy that we trained the proposed line based and segmentation free methods using Farsa and Iranshahr datasets whereas FHT dataset is used for test. This illustrates that the proposed methods are robust to writing style. Furthermore, any word from class of queries is not used in training phase so the proposed methods can spot and retrieve unseen words.

Keywords: Word spotting, Persian documents, label-embedding, attribute-based classification, connected component.

Shahrood University of Technologhy
Faculty of Computer Engineering and IT

Ph.D. Thesis in Artificial Intelligence

Persian word spotting based on attributes

By:
Majid Iranpour Mobarakeh

Supervisor:
Dr. Alireza Ahmadyfard

Adviser:
Dr. Hamid Hasanpour

February 2017
۱۱۲