کاهش فضای جستجو در بازشناسی زیر کلمات تایپی فارسی با استفاده از ویژگی‌های زیست‌سنجه مینوشیا

امین تیمورپور[۱] مهران تقی‌پور گرجی‌کلایی[۲] سیّد محمّد رضوی[۳]

۱- دانش آموخته کارشناسی ارشد- دانشکده مهندسی برق و کامپیوتر- دانشگاه بیرجند- بیرجند- ایران

aminteymourpour@gmail.com

۲- استادیار- دانشکده مهندسی برق و کامپیوتر- دانشگاه بیرجند- بیرجند- ایران

mtaghipour@birjand.acir

۳- دانشیار- دانشکده مهندسی برق و کامپیوتر- دانشگاه بیرجند- بیرجند- ایران

smrazavi@birjand.ac.ir

چکیده: با توجه به گسترده بودن زیرکلمات تایپ شده فارسی، یافتن یک زیرکلمه و به تبع آن یک کلمه در یک متن چاپ شده کار بسیار زمان‌بری خواهد بود. در این مقاله، روشی مبتنی بر نقاط زیست‌سنجه مینوشیا ارائه شده است که فضای جستجوی زیرکلمات تایپ شده فارسی را به صورت قابل توجهی کاهش می‌دهد. لذا تعداد نقاط و مختصات مینوشیای انشعابی و انتهایی که دو ویژگی مطرح در حوزه زیست‌سنجه می‌باشند، بعنوان ویژگی‌هایی برای کاهش فضای جستجو در قالب یک روش دو مرحله‌ای استفاده شده‌اند. در گام نخست نقاط مینوشیا از تصویر زیرکلمه استخراج شده و در چهارخوشه که از لحاظ تعداد نقاط به یکدیگر نزدیک هستند دسته‌بندی می‌شوند، به این ترتیب فضای جستجو تقریباً نصف خواهد شد. در گام دوم با ایجاد یک مخزن از فواصل اولین تا آخرین نقطه انتهایی برای هر زیرکلمه در هر خوشه و تطبیق فاصله مذکور در تصویر آزمایشی با مخزن، فضای جستجو به مقدار قابل توجهی کاهش می‌یابد. نتایج بدست آمده از اعمال روش پیشنهادی بر روی تصاویر زیرکلمه موجود در پایگاه داده نشان می‌دهد، فضای جستجو از ۱۲۷۰۰ زیرکلمه بطور متوسط حدود ۹۸/۸ درصد، با دقت تقریبی بیشتر از ۹۸ درصد کاهش یافته است.

واژه‌های کلیدی: زیرکلمات فارسی، زیست‌سنجه، فضای جستجو، ویژگی مینوشیا

نوع مقاله: پژوهشی

DOI :10.52547/jiaeee.19.2.187

تاریخ ارسال مقاله: ۱۳۹۹/۰۵/۰۲

تاریخ پذیرش مشروط مقاله:۱۴۰۰/۰۵/۱۰

تاریخ پذیرش مقاله: ۱۴۰۰/۰۶/۱۶

نام نویسنده‌ی مسئول: دکتر مهران تقی‌پور گرجی‌کلایی

نشانی نویسنده‌ی مسئول: ایران – خراسان جنوبی – بیرجند –دانشگاه بیرجند– دانشکده مهندسی برق و کامپیوتر – گروه الکترونیک

۱- مقدمه

یکی از اهداف متعالی بشر در آینده‌ای نزدیک حرکت به سوی آرمان‌شهری است که تمام اقدامات در آن بر بستر فضاهای الکترونیک و سایبر شکل بگیرد. اگرچه روی آوردن به زندگی مدرن نیاز به اسناد مکتوب را به حداقل رسانده است، اما رمز موفقیت در آینده حفظ دانش گذشته است. بسیاری از متون ارزشمند فعلی در ادبیات فارسی بصورت مکتوب می‌باشند که متاسفانه نسخه الکترونیکی از آنها موجود نیست. لذا بدیهی است یکی از نیازهای اساسی، حفظ آن آثار و تبدیل آنها به متون قابل ویرایش و الکترونیکی است. اگرچه متون دست‌نویس از پیچیدگی بالایی برخوردار هستند اما در متون تایپ شده نیز به دلیل تنوع فونت‌ها و سایزها چالش‌های فراوانی در بازشناسی آنها وجود دارد. بصورت کلی چالش‌های موجود در این حوزه به دو گروه عمومی و تخصصی (منحصر به زبان فارسی) تقسیم می‌شوند که عباتند از [۱]:

چالش‌های عمومی:

- دشوار بودن پردازش لغوی تمام زبان‌های طبیعی برای رایانه به دلیل عدم دقت کافی کاربر در ترکیب کلمات و زیرکلمات در حین تایپ کردن.
- مشکل ایجاد کلماتی که از منظر تفسیرپذیری برای رایانه دشوار است.
- ایجاد کلمات جدیدی که از ترکیب زیرکلمات بدست می‌آید. اگرچه از دید یک انسان با معنی به نظر می‌رسد اما رایانه آن را در لغت‌نامه خود ندارد.

چالش‌های تخصصی:

- قواعد متفاوت در ساخت کلمات که در بسیاری از موارد بر خلاف زبان‌های رایج بین‌المللی ممکن است یک کلمه از ترکیب چندین زیرکلمه تشکیل شود.
- چالش موجود در زیرکلماتی که گاهی مفهوم ضمیر مفعولی دارند و گاهی نشانه جمع هستند.

و بسیاری از چالش‌های موجود در زبان فارسی که لزوم دسته‌بندی درست و صحیح کلمات و زیرکلمات را نمایان می‌سازند. زبان فارسی، زبان رسمی بیش از ۱۵۰ میلیون نفر در سراسر جهان است. تفاوت‌های عمده‌ی زبان فارسی با زبان‌های لاتین (شامل تفاوت الفباء، پیوسته بودن حروف، تغییر شکل حروف در موقعیت‌های مختلف و شباهت زیاد در بدنه برخی حروف) باعث شده است که بازشناسی متون فارسی پیچیده‌تر از بازشناسی متون لاتین باشد. تحقیقات در زمینه‌ی بازشناسی متون فارسی و عربی از سال ۱۹۸۰ شروع شده است. در بسیاری از روش‌های مطرح شده در منابع موجود بدلیل ساختار منحصربه‌فرد زبان فارسی، ابتدا کلمات به زیرکلمات تقسیم می‌شوند و سپس فرآیند بازشناسی صورت می‌پذیرد. زیرکلمه عبارت است از ترکیبی از حروف به هم پیوسته که تشکیل یک بدنه واحد را بدهند. به‌عنوان مثال کلمه "فارسی" شامل سه زیرکلمه"فا"،"ر" و"سی" است. در این روش هر زیرکلمه به‌عنوان یک شکل واحد دیده شده و مستقل از حروف سازنده آن بازشناسی می‌شود. با توجه به تعداد زیاد زیر-کلمات رایج در زبان فارسی (حدود ۱۳۰۰۰ زیر-کلمه)، بررسی این تعداد کلاس از اصلی‌ترین چالش‌های موجود در بازشناسی متون فارسی است. روش‌های بازشناسی متون تایپ شده فارسی (که عمدتاً تمرکز آنها بر روی زیرکلمات می‌باشد) را می‌توان از جهات مختلف مورد بررسی قرار داد، نگرش جزئی و کلی به زیرکلمات، روش‌هایی که برای استخراج ویژگی مورد استفاده قرار می‌گیرند و تکنیک‌ها و رویکردهایی که برای طبقه‌بندی در نظرگرفته می‌شوند.

بررسی مقالات نشان می‌دهد، ویژگی‌های استخراج شده از تصویر زیرکلمات تایپ شده می‌توانند بیشترین تأثیر را در نتیجه نهایی بگذارند. برای مثال در مقاله [۲] سعی شده است با استخراج ویژگی‌هایی همچون زوایای موجود در زیرکلمه، تعداد سوراخ‌ها، تعداد نقاط و موقیت مکانی آنها و تعداد بالارونده‌ها و پایین‌رونده‌ها به بازشناسی زیرکلمات فارسی بپردازند. همچنین در مقاله [۳] استخراج ویژگی با قراردادن زیرکلمات در یک فضای محدود و اعمال خطوط عمودی و افقی بدست می‌آید. بعبارتی تعداد خطوط قطع شده توسط زیرکلمه بیانگر ویژگی استخراج شده است. بدلیل گسترده بودن زیرکلمات و تنوع زیاد آنها در زبان فارسی چنانچه بخواهیم بصورت مستقیم فرآیند بازشناسی را انجام دهیم با یک مسأله چندین هزار کلاسه مواجه خواهیم بود که کمتر طبقه‌بند کننده‌ای قادر به غلبه بر پیچیدگی آن است. برخلاف مقالات مذکور مطالعات نشان می‌دهند به دلیل بالا بودن حجم داده‌ها و تنوع بالای کلمات و زیرکلمات تایپ شده فارسی برای رسیدن به یک سیستم بازشناسی قابل اطمینان و دقیق نیاز است ابتدا فرآیند خوشه‌بندی بر روی داده‌ها صورت بپذیرد. بنابراین استفاده از ویژگی‌هایی که بتواند قدرت تفکیک‌پذیری را افزایش داده و خوشه‌هایی با میزان شباهت برون‌کلاسی کمتر در عین شباهت درون‌کلاسی بالاتر در اختیار کاربر قرار دهد از اهمیت بسزایی برخوردار است. مقالات [۴، ۵ و ۶] توانسته‌اند با بهره‌گیری از ویژگی‌های قابل قبول در اولین گام و یا در یک ساختار سلسله‌مراتبی فضای جستجو را کاهش دهند. در مقاله [۴] از ویژگی‌های مکان مشخصه، هیستوگرام گرادیان و تبدیل فوریه روی کانتور بعنوان ویژگی استفاده شده است. در مقاله [۵] نیز استخراج ویژگی براساس شکل کلی زیر کلمه، تعداد نقاط آن، بالاترین و پایین‌ترین موقعیت مکانی کلمه، تعداد حلقه و غیره می‌باشد که کلمه مورد آزمایش در یک ساختار سلسله مراتبی در یک درخت تصمیم با کاهش فضای جستجو به کلمه مورد نظر می‌رسد. همچنین در مقاله [۶] از شکل کلی زیرکلمه استفاده می‌شود. به این نحو که ابتدا نقاط حذف شده و سپس استخراج ویژگی با استفاده از توصیفگر فوریه صورت می‌پذیرد. در این

مقاله عملاً فضای جستجو از این طریق کاهش می‌یابد. در مقاله [۷] نیز فضای جستجو با استخراج ویژگی‌هایی همچون ارتفاع، طول، ارتفاع نسبی و تعداد پیکسل‌های مشکی تصویر کاهش یافته است. همچنین در مقاله [۸] از ویژگی‌های پروفایل زیرکلمه برای خوشه‌بندی و کاهش فضای جستجو استفاده شده است.

با توجه به مطالب فوق درمی‌یابیم استفاده از یک فرآیند مناسب برای استخراج ویژگی‌هایی که بتواند فضای جستجو را به نحوی کاهش دهد که در کنار افزایش سرعت، دقت و قابلیت اطمینان را نیز افزایش دهد می‌تواند این چالش مهم را در تهیه نسخه الکترونیکی از متون تایپ شده فارسی برطرف سازد. بررسی‌ها بر روی اسکلت و ساختار زیرکلمات فارسی نشان می‌دهد بر خلاف زبان‌ها مطرح دنیا به ویژه انگلیسی نقاط شروع و پایان، نقاط انشعاب متعدد و همچنین تلاقی‌های فراوانی در کلمات و زیرکلمات وجود دارد که می‌توان براساس آنها یک خوشه‌بندی کارآمد برای زیرکلمات فارسی ارائه نمود. مقالات [۹] و [۱۰] نیز همانند مقاله [۸] مبنای کاهش فضای جستجو را براساس تمرکز بر روی ساختار بدنه زیرکلمات قرار داده است، به نحوی که در مقاله [۹]، از پروفایل‌های عمودی و افقی چهل ویژگی از جمله ضرایب فوریه، درون‌یابی، نرمالسازی و غیره مورد استفاده قرار گرفته است. همچنین در مقاله [۱۰]، تصویر زیرکلمه به صورت سلسله‌مراتبی به نحواحی کوچکتر تقسیم شده و با توجهبه حضور پیکسل‌های مشکلی که بیانگر حضور بخشی از زیرکلمه در ناحیه مذکور است نسبت طول به عرض پروفایل‌های عمودی و افقی بعنوان ویژگی استخراج شده است. بصورت کلی می‌توان گفت در هر سه مقاله آخر که مورد بررسی قرار گرفته‌اند از فرآیندهای سلسله مراتبی استفاده شده است. اما توجه به این نکته حائز اهمیت است که سرعت در کنار دقت یک نکته کلیدی برای طراحی چنین سیستم‌هایی است. به منظور کاهش بار محاسباتی در بخش تصمیم‌گیری و اجتناب از بکارگیری طبقه‌بندی کننده‌های مختلف (همانند مقاله [۱۰]) که می‌تواند هزینه محاسباتی به همراه داشته باشد، استفاده از ویژگی‌های با قدرت تفکیک‌پذیری بالا می تواند نیاز به ترکیب طبقه‌بند کننده‌ها را کاهش دهد. ویژگی مینوشیا که در احراز هویت افراد با استفاده از اثر انگشت کاربردی بسیار گسترده دارد می‌تواند راه‌گشای مناسبی برای یافتن ویژگی‌های مدنظر باشد. لذا در این مقاله با استفاده از ویژگی‌های مینوشیا و استخراج ویژگی‌هایی با قدرت تفکیک‌پذیری بالا فضای جستجو بصورت قابل ملاحظه‌ای کاهش داده شد و نتایج بدست آمده حاکی از عملکرد مطلوب و امیدوار کننده روش پیشنهادی دارد. روش پیشنهادی از دو مرحله کاهش فضای جستجو تشکیل شده است. پیش از استخراج نقاط مینوشیا، تصویر ورودی رفع نویز شده و کیفیت آن بهبود می‌یابد، سپس نرمالیزه شده و اسکلت زیرکلمه آن استخراج می‌شود تا ضخامت خطوط به حدود یک پیکسل کاهش یابد. سپس در گام نخست از کاهش فضای جستجو تصاویر ورودی براساس تعداد نقاط مینوشیا خوشه‌بندی می‌شوند، در گام دوم فاصله اولین و آخرین نقطه مینوشیای انتهایی به عنوان معیاری با مخزن فواصل که به همین ترتیب برای تصاویر آموزشی موجود در خوشه ایجاد شده است، مقایسه شده و زیرکلمه درست انتخاب می‌شود. نتایج بدست آمده نشان می‌دهد، با استفاده از روش پیشنهادی می‌توان فضای جستجو را از ۱۲۷۰۰ زیرکلمه به حدود ۵۰۰ زیرکلمه با دقتی بیش از ۹۰ درصد کاهش داد.

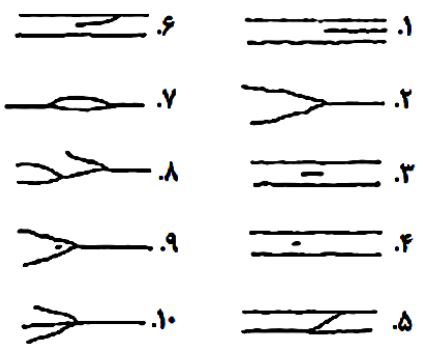
در ادامه مقاله، ابتدا در فصل دوم ویژگی مینوشیا مورد مطالعه قرار خواهد گرفت، سپس روش پیشنهادی در فصل سوم مطرح خواهد شد و در فصل چهارم نیز شبیه‌سازی‌ها و نتایج بدست آمده مورد تحلیل و بررسی قرار خواهند گرفت. در پایان نیز نتیجه‌گیری در فصل پنجم ارائه خواهد شد.

۲- ویژگی زیست‌سنجه مینوشیا

هدف اصلی در این تحقیق، ارائه یک روش متناسب با شکل کلی زیرکلمه است که بتواند فضای جستجو را تا حد امکان کاهش دهد. نقاط مینوشیا[۱] الهام گرفته از ساختار اثرانگشت[۲] می‌باشند و علی‌رغم نمونه‌های مختلف از نقاط مینوشیا که بالغ بر ۱۵۰ نمونه می‌باشند، تنها نقاط انتهایی[۳] و انشعابی[۴] مورد توجه پژوهشگران در این زمینه بوده است و فقط نمونه‌های بسیار نادر از این تحقیقات را می‌توان یافت که خارج از مرز بیومتریک (زیست‌سنجه) کار کرده باشند. از جمله این تحقیقات می‌توان به استفاده از نقاط مینوشیا برای جداسازی[۵] حروف دست‌نویس در مرجع [۱۴] و یا استفاده از این نقاط در تشخیص امضا[۶] [۱۵] اشاره کرد.

۲-۱- ویژگی‌های مینوشیا در اثرانگشت

اثرانگشت در طول زمان عوض نمی‌شود و حتی در صورت بروز جراحات و سوختگی‌های شدید هم پس از بهبود، دوباره به همان شکل قبلی خواهد بود. برجستگی آن‌ها، باعث ایجاد اصطکاک بیشتر و همچنین به وجود آمدن حس لامسه بهتر می‌شود. این خطوط در ماه چهارم بارداری تشکیل می‌شود و تا هفته‌ی هجدهم بارداری برجسته می‌شود. این خطوط علاوه بر ژنتیک افراد به لحظه‌ی تولد جنین و کشش‌ها و حالات روحی مادر در هنگام زایمان نیز بستگی دارد. به همین دلیل، در دوقلوهای همسان نیز اثرانگشت مشابه، دیده نمی‌شود. از آنجایی که بانک‌های اطلاعاتی وسیعی برای اثرانگشت موجود است، روش دستی برای تطبیق اثرانگشت[۷]، امری ناشدنی است. دو ویژگی مهم این الگوها یعنی انتهای برآمدگی و دوشاخه شدن برآمدگی‌ها بیشتر مورد استفاده قرار گرفته، به این ویژگی‌ها اصطلاحاً مینوشیا گفته می‌شود؛ که در شکل (۱) نشان داده شده است.



شکل (۱): انواع مختلف از خطوط مهم در اثرانگشت.

علاوه بر ویژگی‌های بیان شده، در بسیاری از سیستم‌های تشخیص اثرانگشت، برای افزایش صحت تطبیق، از ویژگی‌های سطح بالا نیز استفاده می‌شود که یکی از این ویژگی‌ها، تشخیص کلاس الگوی اثرانگشت است که می‌توان آن را به ۵ کلاس تقسیم نمود: کمان، کمان مایل، حلقه‌ی چپ، حلقه‌ی راست و مارپیچ. شکل (۲) نمونه‌هایی از کلاس‌های مذکور را نشان می‌دهد.

شکل (۲): نمونه‌های مختلف از انواع کلاس‌های اثرانگشت.

در صورتی که باز هم کیفیت تصویر به گونه‌ای باشد که نتوان از کلاس الگو هم به نتیجه رسید، از ویژگی سطح بالاتری به نام چگالی برآمدگی‌ها استفاده می‌شود؛ که بیانگر تعداد برآمدگی‌ها در واحد طول تعریف می‌شود. به منظور مستقل کردن تعداد برآمدگی‌ها از جهت تصویر، تعداد برآمدگی‌های بین دو نقطه منفرد محاسبه می‌شود. این نقاط منفرد، همان هسته[8] و دلتا[9] می‌باشند.

شکل (۳): نمونه‌هایی از نقاط مینوشیا در یک اثرانگشت

اطلاعات مینوشیا در مؤلفه‌های x و y و زاویه آن‌ها نهفته است و در یک تصویر باکیفیت نسبتاً خوب بین ۷۰ تا ۸۰ مشخصه‌ی مینوشیا می‌توان استخراج کرد که البته با کاهش کیفیت تصویر، این تعداد به ۲۰ الی ۳۰ عدد کاهش می‌یابد که با این تعداد هم می‌توان به شناسایی اثر و یا تطبیق آن پرداخت.

۳- روش پیشنهادی

همانطور که ذکر شد، ویژگی مینوشیا می‌تواند گزینه مناسبی برای الگوهایی باشد که از ترکیب خطوط تبعیت می‌کنند. کلمات و زیرکلمات فارسی با توجه به فرم نگارش آنها حاوی اطلاعات مینوشیای بسیاری هستند که می‌توان از آنها برای کاهش فضای جستجو و بالا بردن دقت بازشناسی استفاده نمود. این کاهش فضای جستجو توسط کشف ویژگی و استخراج آن‌ها از زیر کلمات و خوشه‌بندی آن‌ها در خوشه‌هایی با سایزهای به مراتب کوچک‌تر انجام می‌شود. به این ترتیب علاوه بر افزایش سرعت، به دقت بالاتری در بازشناسی کلمات می‌رسیم.

۳-۱- استخراج نقاط مینوشیا

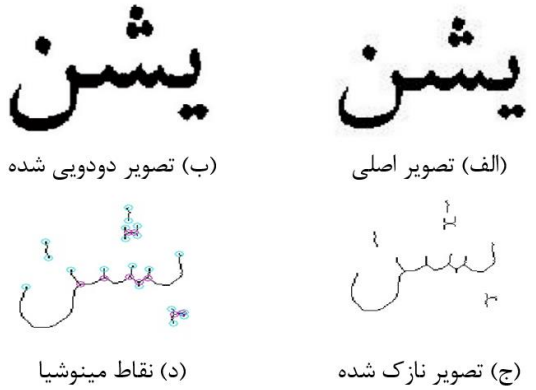
اولین گام در ایجاد خوشه‌های مینوشیا به منظور کاهش فضای جستجو استخراج نقاط مینوشیا از تصویر زیرکلمات است. همانطور که در شکل (۴) مشاهده می‌شود استخراج نقاط مینوشیا در زیرکلمات فارسی شامل ۴ مرحله می‌باشد که عبارتند از:

(الف) اخذ تصویر ورودی

(ب) دودویی(باینری)کردن تصویر

(ج) نازک‌سازی

(د) استخراج نقاط مهم (مینوشیا).



(الف) تصویر اصلی (ب) تصویر دودویی شده

(ج) تصویر نازک شده (د) نقاط مینوشیا

شکل (۴): استخراج نقاط مینوشیا برای زیرکلمات فارسی

دودویی نمودن تصویر ورودی و حذف نویز:

اعمال الگوریتم دودویی‌کردن و باینری‌سازی، یکی از کارهای مرسوم در پیش‌پردازش تصاویر می‌باشد. با این‌کار وضوح بهتری از تصویر ورودی بدست خواهیم آورد و نقاط نویزی که احتمالاً در اثر اسکن‌شدن با وضوح پایین‌تر بوجود آمده، از بین خواهد رفت و همچنین تصویر، برای اعمال نازک‌سازی آمادگی بهتری خواهد داشت. در شکل (۴–ب) زیر نمونه‌ای از تصویر یک کلمه‌ی تایپ‌شده‌ی فارسی و تصویر دودویی شده‌ی آن را ملاحظه می‌کنید.

نازک‌سازی تصویر باینری شده:

مرحله‌ی بعد که جزو مهمترین مراحل در راستای تشخیص و اخذ نقاط مینوشیا می‌باشد، مرحله‌ی نازک‌سازی است. هدف از این امر، نازک

کردن نوشته به ضخامت یک پیکسل است. هرچه قدر این کار بهتر انجام شود، نقاط مینوشیا با سهولت و دقت بهتر استخراج خواهند شد و نقاط مینوشیای جعلی کمتری بوجود خواهد آمد.

استخراج نقاط مینوشیا:

در بهترین حالت ضخامت بدنه کلمه نازک شده باید به یک پیکسل برسد. سپس با اعمال الگوهای انشعابی و نقطه انتهایی که بترتیب در اشکال (۵– الف) و (۵– ب) بصورت یک فیلتر ۳×۳ طراحی شده‌اند، می‌توان نقاط مینوشیا را استخراج نمود.

(الف) فیلتر استخراج مینوشیا انشعابی

(ب) فیلتر استخراج مینوشیا نقطه انتهایی

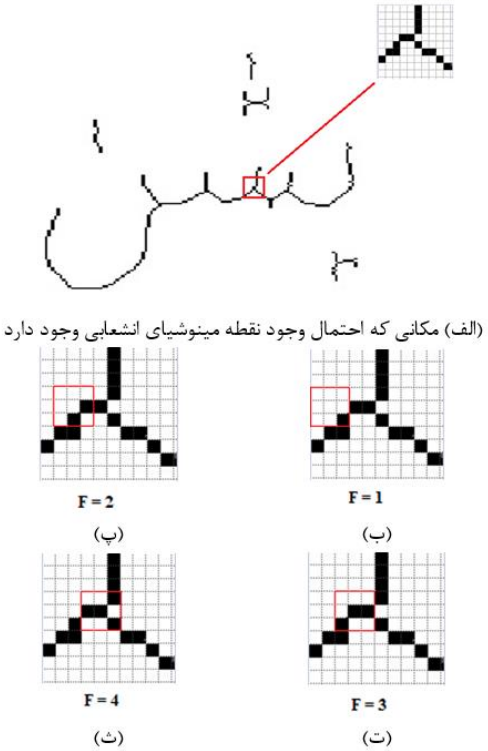
شکل (۵): الگوی استخراج نقاط مینوشیا برای تصاویر نازک شده زیر کلمات فارسی به ضخامت یک پیکسل

چنانچه فرض کنیم تصویر زیر کلمه ورودی یک تصویر $m \times n$ پیکسل باشد که $m > 3$ و $n > 3$، آنگاه فیلتر استخراج نقاط مینوشیا با گام‌های یک پیکسل شروع به روبش تصویر ورودی در پنجره‌های ۳×۳ می‌کند. چنانچه مطابق رابطه (۱) در هر گام حاصلضرب نظیر به نظیر پیکسل‌های فیلتر مینوشیای انشعابی و پنجره تصویر برابر ۴ باشد به شرط آنکه N_9 برابر یک (یعنی پیکسل وسط پنجره برابر یک باشد) آن پیکسل نقطه انشعابی است و چنانچه حاصلضرب فیلتر و پنجره برابر ۲ باشد آن پیکسل نقطه انتهایی است.

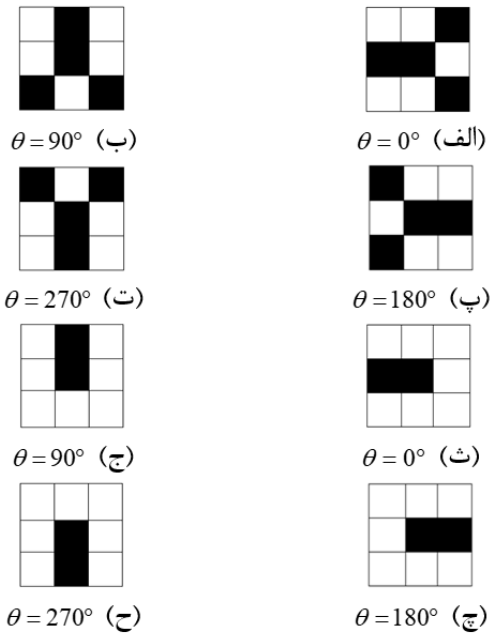
$$F = \begin{cases} \sum_{i=1}^{9} N_i = 2 & End\ \text{Point} \\ \sum_{i=1}^{9} N_i = 4 & Bifurcation\ \text{Point} \end{cases} \qquad (۱)$$

که N_i حاصلضرب نظیر به نظیر هر پیکسل از تصویر و پیکسل i ام الگوی نقاط مینوشیا است که می‌تواند یک یا صفر باشد. شکل (۶) نمونه‌ای از یافتن نقطه مینوشیای انشعابی را نشان می‌دهد. همانطور که ملاحظه می‌شود، فیلتر شکل (۵-الف) در حال روبش تصویر است و مکانی که $F = 4$ می‌شود را بعنوان نقطه انشعابی تشخیص می‌دهد (به شکل (۶ -ث) توجه کنید).

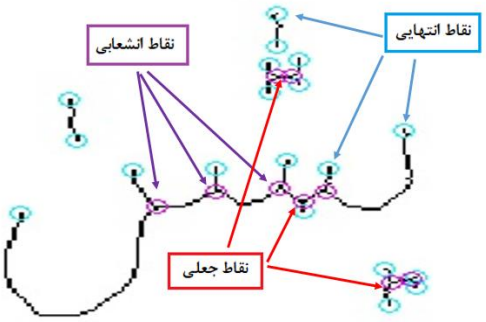
توجه به این نکته بسیار مهم است که همواره نمی‌توان انتظار داشت، دقیقا الگوی پنجره با الگوی فیلتر مطابقت داشته باشد. لذا برای در نظر گرفتن تمام حالات ممکن در هر گام فیلتر همانند شکل (۷) به اندازه ۹۰ درجه می‌چرخد. به این ترتیب با یافتن بیشترین انطباق و با لحاظ نمودن شروط مذکور نقاط مینوشیا همانند شکل (۸) بدست می‌آیند. همانطور که در شکل (۸) ملاحظه می‌شود با توجه به چالشی که در نازک‌سازی وجود دارد در برخی نقاط روش پیشنهادی در تشخیص نقاط مینوشیا دچار خطا می‌شود که نیاز است نقاط جعلی حذف گردند.



شکل (۶): اعمال فیلتر استخراج مینوشیا انشعابی بر روی یک زیرکلمه نازک شده. زمانی پیکسلی بعنوان نقطه انشعابی انتخاب می‌شود که دو شرط را برقرار کند: ابتدا پیکسل وسط پنجره ۳×۳ یک باشد، دوم اینکه حاصلضرب نظیر به نظیر فیلتر انشعابی و آن پنجره همانند (ث) برابر ۴ باشد



شکل (۷): اعمال چرخش بر روی فیلترهای استخراج نقاط مینوشیا به منظور در نظر گرفتن تمام حالات ممکن. (الف) الی (ت) چرخش ۹۰ درجه برای فیلتر استخراج نقطه انشعابی و (ث) الی (ح) نیز چرخش ۹۰ درجه برای فیلتر استخراج نقطه آخر را نشان می‌دهد



شکل (۸): استخراج اولیه نقاط مینوشیا بر روی یک نمونه از زیر کلمه فارسی

۳-۲- حذف نقاط مینوشیا جعلی و استخراج ویژگی‌های نهایی

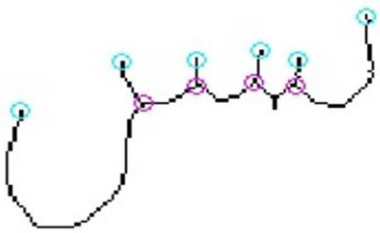
همانطور که در شکل (۸) دیده می‌شود بعضی از نقاط به اشتباه به عنوان نقاط مینوشیا شناسایی شده‌اند و با دقت بیشتر مشاهده می‌شود که این اتفاق، اکثراً در مکان «نقاط» یک کلمه رخ می‌دهد. بنابراین می‌توان از همان ابتدا، نقاط را حذف کرد و روش پیشنهادی تنها بر روی بدنه‌ی کلمه‌ی مورد نظر اعمال گردد و استخراج نقاط مینوشیا صورت بپذیرد. امّا کماکان این تکنیک مشکل نقاط مینوشیای جعلی در بدنه را برطرف نمی‌کند به همین دلیل این نقاط، توسط سه رویکرد حذف خواهند شد:

۱- فاصله‌ی بین نقطه‌ی مینوشیای مربوط به انتهایی و انشعابی، کم‌تر از حد آستانه‌ی D باشد.

۲- فاصله‌ی بین نقاط مینوشیای انشعابی کمتر از D باشد.

۳- فاصله‌ی بین نقاط مینوشیای انتهایی کمتر از D باشد.

شایان ذکر است، مقدار D بر حسب تجربه بدست می‌آید که در اینجا مقدار ۶ قرار داده شده است. شکل (۹) نمونه‌ای از نقاط مینوشیای استخراج شده برای زیرکلمه مورد بررسی را نشان می‌دهد.



شکل (۹): نقاط مینوشیا نهایی بعد از حذف نقاط جعلی

۳-۳- کاهش فضای جستجو

تعداد نقاط مینوشیا برای هر زیر کلمه می‌تواند گزینه مناسبی برای کاهش فضای جستجو در زیر کلمات فارسی باشد. بصورت کلی کاهش فضای جستجو با استفاده از خصایص نقاط مینوشیا را می‌توان در دو گام به شرح زیر انجام داد.

- **کاهش فضای جستجو براساس تعداد نقاط مینوشیا:**

در این گام همانند آنچه در بخش قبل توضیح داده شد نقاط مینوشیا استخراج شده و فضای جستجو به خوشه‌های کوچک‌تر تقسیم می‌شود. اما از آنجاییکه طبق تجربیات به دست آمده با وجود نویزهای ناخواسته احتمال خطا در تولید نقاط مینوشیا و بعبارتی وجود مینوشیاهای جعلی وجود دارد نیاز است خوشه‌های انتخاب شده به نحوی باشند که خطا را به حداقل برساند، لذا باید با سعی و خطا بهترین خوشه‌ها انتخاب شوند.

- **کاهش فضای جستجو براساس مختصات نقاط مینوشیا:**

پس از آنکه خوشه‌های مناسب انتخاب شدند، در گام دوم می‌توان از مینوشیاهای مطمئن در زیرکلمات استفاده نمود که اولاً خطا در استخراج آنها بسیار کم است و ثانیاً امکان تولید جعلی آنها بسیار پایین است. لذا در گام دوم می‌توان از مختصات اولین و آخرین نقاط انتهایی که در ابتدا و انتهای زیرکلمه می‌باشند، بعنوان معیاری برای کاهش دوم فضای جستجو استفاده نمود.

با توجه موارد فوق گام‌های کاهش فضای جستجو را به ترتیب می‌توان بصورت زیر بیان نمود:

۱- ابتدا تصویر ورودی را نرمالیزه کرده تا اثر احتمالی متغیر بودن فونت و ابعاد تصاویر زیرکلمات بر روند کلی الگوریتم را کاهش دهیم. به این ترتیب تصاویر در هر ابعادی باشند بصورت یک تصویر ۱×۱ در می‌آیند.

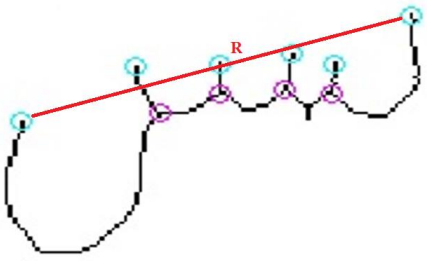
۲- در گام نخست کاهش فضای جستجو، استخراج نقاط مینوشیای نقطه آخر و انشعابی انجام می‌پذیرد و با کمک آنها خوشه‌بندی تصاویر زیرکلمات براساس تعداد نقاط بدست آمده.

۳- در گام دوم کاهش فضای جستجو، محاسبه فاصله اولین نقطه انتهایی و آخرین نقطه انتهایی بعنوان معیاری برای کاهش فضای جستجو درنظر گرفته می‌شود. به این ترتیب برای هر یک از اعضای نمونه‌ی آزمایشی، چهار عدد بین صفر و یک که همان مختصات نرمالیزه شده از نقاط مینوشیا می‌باشند، موجود است. با استفاده از فرمول تعیین فاصله اقلیدسی که در رابطه (۲) بیان شده است عددی برای هر نمونه به دست می‌آید.

$$r = d(p,q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \qquad (۲)$$

همانند شکل (۱۰) این فاصله را برای هر یک از اعضای نمونه‌های آموزشی نیز به دست می‌آوریم و برای هر یک از نمونه‌های آزمایشی، عدد r به دست آمده را با عدد r به دست آمده از نمونه‌های آموزشی مقایسه کرده و به ترتیب از کمترین به بیشترین فاصله مرتب می‌کنیم. در دسته مرتب شده از مقادیر فاصله‌ها، به دنبال برچسب زیر کلمه

موردنظر می‌گردیم. به‌محض رسیدن به برچسب موردنظر، عملیات جستجو را متوقف کرده و طول دسته را یادداشت می‌کنیم. بدیهی است که طول دسته‌ها هر چه کمتر باشند فضای جستجو کاهش بیشتری داشته است.

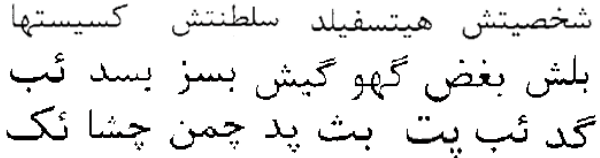


شکل (۱۰): نمونه‌ای از فاصله اولین و آخرین نقطه انتهایی در یک زیرکلمه

۴- شبیه‌سازی‌ها و نتایج

۴-۱- پایگاه داده مورد مطالعه

پایگاه داده مورد مطالعه در این مقاله شامل ۱۲۷۰۰ زیرکلمه است که از دو روزنامه کیهان و همشهری استخراج شده‌اند. به این ترتیب متداولترین کلمات که بیشتر از ۳۰ تکرار در متن داشته‌اند انتخاب و زیرکلمات مربوط به آنها استخراج شده‌اند. به این ترتیب برای ۲۹۷۳۹ کلمه متداول، تعداد زیرکلمات بدست آمده ۱۲۷۰۰ شده است. نمونه‌ای از زیرکلمات موجود در این پایگاه داده در شکل (۱۱) نمایش داده شده است.



شکل (۱۱): نمونه‌هایی از زیرکلمات در پایگاه داده مورد مطالعه

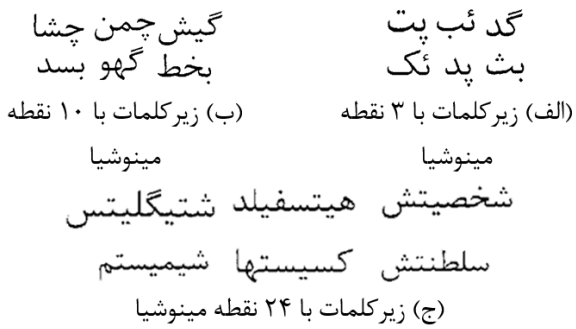
۴-۲- شبیه‌سازی و نتایج

تصاویر موجود در پایگاه داده با فونت نازنین با قلم ۱۴ تهیه و در دسترس می‌باشد. لذا به منظور ارزیابی صحیح روش پیشنهادی ۱۰۰۰ تصویر بصورت تصادفی بعنوان داده‌های آزمایشی از بین ۱۲۷۰۰ تصویر انتخاب شده است و در پنج قلم متفاوت با کیفیت‌های متفاوت اسکن شده‌اند. بنابراین ۵۰۰۰ (۵ دسته ۱۰۰۰ تایی) زیرکلمه فارسی با اندازه قلم‌های متفاوت و کیفیت‌های متفاوت از فونت نازنین بعنوان داده آزمایشی خواهیم داشت. از سویی برای ارزیابی بیشتر روش پیشنهادی ۱۰۰۰ تصویر انتخابی را با فونت لوتوس بازنویسی نموده و همانند فونت نازنین آنها را در کیفیت‌های مختلف و با پنج قلم متفاوت اسکن نمودیم تا ۵۰۰۰ تصویر آزمایشی با فونت لوتوس نیز در اختیار داشته باشیم. شکل (۱۲) نمونه‌ای از تصاویر تولید شده با فونت لوتوس برای یک نمونه از تصویر زیرکلمه در پایگاه داده را نشان می‌دهد.

شکل (۱۲): نمونه‌هایی از تصاویر زیرکلمه ایجاد شده با فونت لوتوس در مرحله آزمایش

۴-۲-۱- طراحی خوشه‌ها و تولید مخزن فواصل

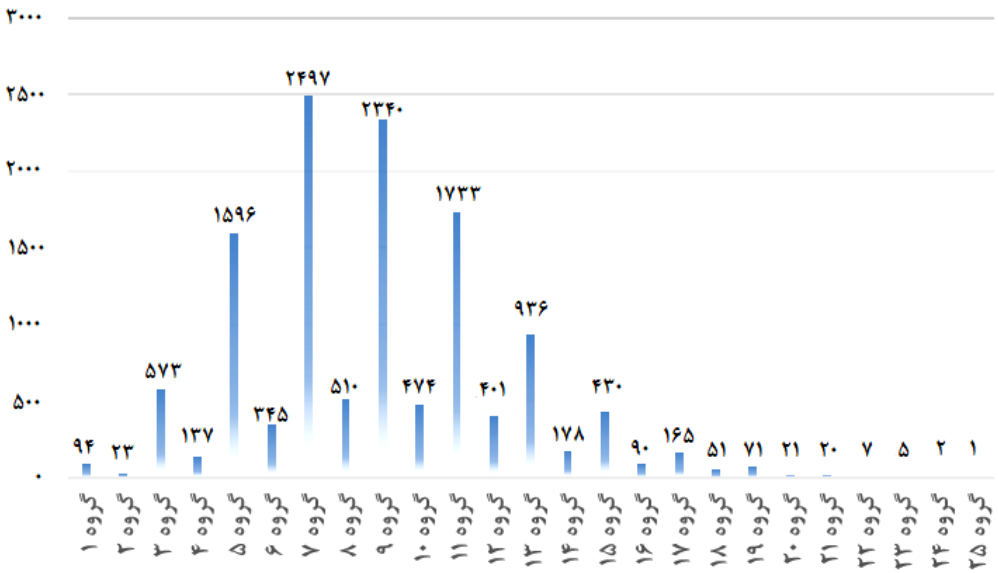
اشکال (۱۳-الف) الی (۱۳-ج) نمونه‌ای از زیرکلمات را نشان می‌دهد که با استفاده از روش پیشنهادی بترتیب شامل ۳، ۱۰ و ۲۴ نقطه مینوشیا می‌باشند. به همین ترتیب شکل (۱۴) بیانگر دسته‌بندی (گروه‌بندی) زیرکلمات پایگاه داده براساس نقاط مینوشیا را نشان می‌دهد. همانطور که ملاحظه می‌شود کمترین تعداد نقاط، ۲ نقطه (که دو نقطه انتهایی در ابتدا و انتهای زیرکلمه است) و بیشترین تعداد نقاط ۲۵ نقطه است. به منظور کاهش خطا در تشخیص زیرکلمه و شباهت ساختاری زیرکلماتی که دارای تعداد نقاط مینوشیای مشابه هستند، زیرکلمات از نقطه‌نظر تعداد نقاط مینوشیا خوشه‌بندی شده‌اند. همانطور که در شکل (۱۵) مشاهده می‌شود، از تصویر آموزشی با فرآیندی که در بخش‌های قبلی توضیح داده شد، نقاط مینوشیا استخراج شده و در چهار خوشه اصلی که بصورت سعی و خطا بدست آمده‌اند دسته‌بندی شده‌اند. در دسته اول تصاویری قرار گرفته‌اند که دارای ۲ الی ۵ نقطه مینوشیا بوده‌اند و به همین ترتیب در دسته چهارم تصاویری قرار گرفته‌اند که بیش از ۱۵ نقطه مینوشیا داشته‌اند.



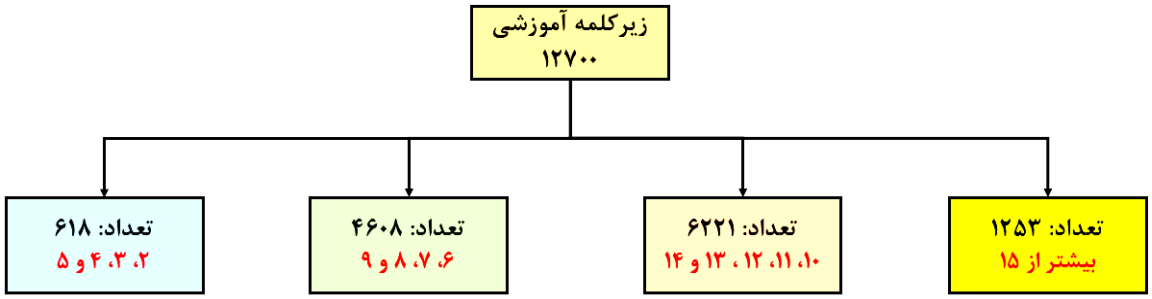
شکل (۱۳): نمونه‌هایی از تصاویر آموزشی خوشه‌بندی شده. (الف) ۳ نقطه مینوشیا، (ب) ۱۰ نقطه مینوشیا و (ج) ۲۴ نقطه مینوشیا

با توجه به خوشه‌های بدست آمده و محاسبه فاصله اولین و آخرین نقطه انتهایی برای هر زیرکلمه در هر خوشه مطابق رابطه (۲) انجام شده و مخزنی از فواصل برای هر خوشه بدست خواهد آمد. بنابراین در

انتها چهار مخزن خواهیم داشت که بترتیب دارای ۶۱۸، ۴۶۰۸، ۶۲۲۱ و ۱۲۵۳ فاصله هستند. بدیهی است که طول فواصل برای خوشه اول کمترین و برای خوشه چهارم بیشترین است.



شکل (۱۴): دسته‌بندی (گروه‌بندی) اولیه زیرکلمات پایگاه



شکل (۱۵): خوشه‌بندی تصاویر زیرکلمات فارسی آموزشی در چهار خوشه که بصورت سعی و خطا بدست آمده است

۴-۲-۲- ارزیابی روش پیشنهادی برای فونت نازنین

همانطور که در جداول (۱) و (۲) دیده می‌شود، در گام نخست از کاهش فضای جستجو، خطا بسیار ناچیز بوده و داده‌های آزمایشی در این گام با بیش از ۹۸ درصد دقت خوشه‌بندی شده‌اند. بطور مثال در خوشه اول که ۶۱۸ تصویر آموزشی قرار گرفته است، چنانچه تصویری از جنس این خوشه باشد با خطای حدود ۱ درصد بازشناسی می‌شود. در خوشه دوم که ۴۶۰۸ تصویر آموزشی قرار گرفته است، خطا حدود ۱/۵ درصد بوده و این مقدار برای زمانی که تصویر آزمایشی ورودی متعلق به خوشه‌های سوم و چهارم باشد بترتیب برابر ۲/۳ درصد و صفر درصد خواهد بود. در گام دوم زیرکلمه آزمایشی با مخزن خوشه انتخاب شده مقایسه می‌شود و به محض رسیدن به زیرکلمه صحیح متوقف می‌شود. به این ترتیب برای مثال همانطور که در جداول (۱) و (۲) دیده می‌شود حدود ۴۳ درصد از تصاویر تخصیص داده شده به خوشه اول در همان گام نخست بازشناسایی می‌شوند. براساس نتایج بدست آمده بیش از ۸۵ درصد زیرکلمات آزمایشی در فضای کمتر از ۱۰۰ زیرکلمه شناسایی می‌شوند که کاهش فضای جستجو قابل قبولی را نشان می‌دهد و بصورت متوسط می‌توان گفت در هر خوشه در ۲۵ درصد اول فروانی تصویر آزمایشی به درستی بازشناسی می‌شود.

جدول (۱): میانگین نتایج حاصل از خوشه‌بندی در خوشه‌های اول و دوم و کاهش فضای جستجو برای تصویر آزمایشی با فونت نازنین

خوشه اول (تعداد: ۶۱۸)		خوشه دوم (تعداد: ۴۶۰۸)	
فراوانی	بازشناسی (٪)	فراوانی	بازشناسی (٪)
۱	۴۳/۰۲	۱	۵۰/۰۸
کم‌تر از ۱۰	۵۹/۰۷	کم‌تر از ۱۰	۶۲/۲۸
کم‌تر از ۵۰	۷۸/۴۶	کم‌تر از ۵۰	۷۱/۲۷
کم‌تر از ۱۰۰	۹۵/۳۸	کم‌تر از ۱۰۰	۸۴/۳۹
کم‌تر از ۲۵۰	۹۸/۲۸	کم‌تر از ۲۵۰	۹۱/۱۹
کم‌تر از ۶۱۸	۹۹/۰۶.	کم‌تر از ۵۰۰	۹۳/۴۷
-	-	کم‌تر از ۱۰۰۰	۹۵/۶۱
-	-	کم‌تر از ۲۰۰۰	۹۷/۹۲
-	-	کم‌تر از ۴۶۰۸	۹۸/۳۴
-	-	-	-
خطا	۰/۹۴	خطا	۱/۶۶

جدول (۲): میانگین نتایج حاصل از خوشه‌بندی در خوشه‌های سوم و چهارم و کاهش فضای جستجو برای تصویر آزمایشی با فونت نازنین

خوشه سوم (تعداد: ۶۲۲۱)		خوشه چهارم (تعداد: ۱۲۵۳)	
فراوانی	بازشناسی (٪)	فراوانی	بازشناسی (٪)
۱	۴۹/۳۶	۱	۴۶/۰۷
کم‌تر از ۱۰	۵۹/۳۴	کم‌تر از ۱۰	۶۹/۸۱
کم‌تر از ۵۰	۷۲/۱۹	کم‌تر از ۵۰	۹۰/۰۷
کم‌تر از ۱۰۰	۸۵/۱۱	کم‌تر از ۱۰۰	۹۲/۸۲
کم‌تر از ۲۵۰	۸۷/۳۹	کم‌تر از ۲۵۰	۹۷/۲۸
کم‌تر از ۵۰۰	۹۰/۲۹	کم‌تر از ۵۰۰	۹۸/۹۳
کم‌تر از ۱۰۰۰	۹۲/۶۴	کم‌تر از ۱۲۵۳	۱۰۰
کم‌تر از ۲۰۰۰	۹۳/۳۴	-	-
کم‌تر از ۴۰۰۰	۹۶/۶۴	-	-
کم‌تر از ۶۲۲۱	۹۷/۷۲	-	-
خطا	۲/۲۸	خطا	۰

۴-۲-۳- ارزیابی روش پیشنهادی برای فونت لوتوس

به منظور بررسی روش پیشنهادی در صورت تغییر فونت متن تایپ شده در مرحله دوم ارزیابی، تصاویر آزمایشی با فونت لوتوس مشابه حالت قبل ایجاد شده‌اند. ۵۰۰۰ تصویر آزمایشی در سایزهای مختلف ایجاد شده است که نتایج حاصله بیانگر عملکرد مطلوب روش پیشنهادی برای این فونت نیز می‌باشد. همانطور که در جداول (۳) و (۴) دیده می‌شود، میزان خطا در گام نخست کاهش فضای جستجو به‌ترتیب برای خوشه‌های اول، دوم، سوم و چهارم برابر است با ۵/۱۲ درصد، ۸/۳۷ درصد، ۹/۱۰ درصد و ۳/۲۲ درصد. همچنین در صورت تشخیص درست خوشه در گام نخست برای تصویر آزمایشی بیش از ۸۵ درصد از آنها در ۲۵۰ جستجوی اول بازشناسایی می‌شوند. توجه به این نکته حائز اهمیت است که تصاویر آزمایشی با فونت نازنین تهیه شده‌اند و ضعف نتایج به نسبت بخش قبلی قابل پیش‌بینی بوده است. بدیهی است که نقاط مینوشیا در این دو فونت بدلیل وجود مرحله نرمالیزه کردن کمی متفاوت باشد، لذا با استخراج نقاط مینوشیا از سطح خاکستری تصاویر نتایج بسیار مطلوب‌تر خواهد بود. اما با این وجود نتایج تاحدود زیادی قابل قبول بوده و بیانگر کارآمدی روش پیشنهادی است.

جدول (۳): میانگین نتایج حاصل از خوشه‌بندی در خوشه‌های اول و دوم و کاهش فضای جستجو برای تصویر آزمایشی با فونت لوتوس

خوشه اول (تعداد: ۶۱۸)		خوشه دوم (تعداد: ۴۶۰۸)	
فراوانی	بازشناسی (٪)	فراوانی	بازشناسی (٪)
۱	۳۱/۲۹	۱	۳۷/۱۵
کم‌تر از ۱۰	۴۲/۴۴	کم‌تر از ۱۰	۵۲/۰۳
کم‌تر از ۵۰	۶۱/۲۲	کم‌تر از ۵۰	۶۸/۸۱
کم‌تر از ۱۰۰	۷۶/۳۴	کم‌تر از ۱۰۰	۷۳/۷۱
کم‌تر از ۲۵۰	۸۵/۲۹	کم‌تر از ۲۵۰	۸۲/۴۶
کم‌تر از ۶۱۸	۹۴/۸۸	کم‌تر از ۵۰۰	۸۴/۳۳
-	-	کم‌تر از ۱۰۰۰	۸۶/۴۹
-	-	کم‌تر از ۲۰۰۰	۹۰/۰۶
-	-	کم‌تر از ۴۶۰۸	۹۱/۶۳
-	-	-	-
خطا	۵/۱۲	خطا	۸/۳۷

جدول (۴): میانگین نتایج حاصل از خوشه‌بندی در خوشه‌های سوم و چهارم و کاهش فضای جستجو برای تصویر آزمایشی با فونت لوتوس

خوشه سوم (تعداد: ۶۲۲۱)		خوشه چهارم (تعداد: ۱۲۵۳)	
فراوانی	بازشناسی (٪)	فراوانی	بازشناسی (٪)
۱	۴۰/۰۷	۱	۳۸/۱۲
کم‌تر از ۱۰	۵۴/۳۶	کم‌تر از ۱۰	۵۱/۱۴
کم‌تر از ۵۰	۶۵/۶۱	کم‌تر از ۵۰	۷۰/۰۱
کم‌تر از ۱۰۰	۷۱/۰۰	کم‌تر از ۱۰۰	۸۰/۲۰
کم‌تر از ۲۵۰	۸۳/۸۶	کم‌تر از ۲۵۰	۸۴/۹۳
کم‌تر از ۵۰۰	۸۵/۳۸	کم‌تر از ۵۰۰	۹۲/۶۶
کم‌تر از ۱۰۰۰	۸۸/۲۰	کم‌تر از ۱۲۵۳	۹۶/۷۸
کم‌تر از ۲۰۰۰	۸۹/۰۰	-	-
کم‌تر از ۴۰۰۰	۹۰/۸۲	-	-
کم‌تر از ۶۲۲۱	۹۰/۰۹	-	-
خطا	۹/۱۰	خطا	۳/۲۲

با مقایسه نتایج بدست آمده برای تصاویر فونت نازنین و لوتوس می‌توان نتیجه گرفت، بصورت کلی تشخیص و استخراج صحیح نقاط مینوشیای انشعابی و نقطه آخر می‌تواند گزینه مناسبی برای کاهش فضای جستجو باشد. در این حالت خطا کمتر از ۱۰ درصد بوده و دقت مناسبی برای کاهش فضای جستجو در گام‌های بعدی را فراهم می‌آورد. در گام دوم نیز مطابق نتایج بدست آمده فضای جستجو می‌تواند تا کمتر از ۵۰۰ کلمه کاهش یابد و نرخ بازشناسی بصورت میانگین حدود ۹۰ درصد را در اختیار کاربر قرار دهد که این امر

کاهش چشمگیر فضای جستجو از ۱۲۷۰۰ کلمه به کمتر از ۵۰۰ کلمه را نشان می‌دهد.

۴-۲-۴- مقایسه با روش‌های موجود

در این بخش روش پیشنهادی با روش‌های مطرح شده برای کاهش فضای جستجو که بر روی پایگاه داده مورد مطالعه کرده‌اند، براساس متوسط درصد کاهش فضای جستجو مقایسه شده است. بدیهی است تعداد تصاویر خوشه‌بندی شده از پایگاه داده در خوشه‌های مختلف برای روش‌های مطرح شده یکسان نیست، به همین جهت به منظور رعایت عدالت متوسط کاهش فضای جستجو در گام آخر روش‌ها مورد ارزیابی قرار گرفته است. برای مثال در مقاله [۹] فضای جستجو از ۱۲۷۰۰ زیرکلمه بطور متوسط به حدود ۱۳۰ زیرکلمه رسیده است که معنی کاهش حدود ۹۹/۱ درصدی فضای جستجو است. در روش پیشنهادی نیز بطور متوسط با دقت حدود ۹۰ درصد، فضای جستجو بصورت متوسط به حدود ۹۹ درصد کاهش می‌یابد و همچنین برای دقت بالای ۹۸ درصد، فضای جستجو به حدود ۹۸ درصد کاهش می‌یابد. باید توجه داشت روش پیشنهادی از لحاظ عملکرد در کاهش فضای جستجو با حفظ دقت بازشناسی در قیاس با روش‌های دیگر عملکرد مطلوبی دارد اگرچه نمی‌توان قاطعانه گفت از همه آنها بهتر است. اما توجه به دو نکته حائز اهمیت است؛ اولاً روش پیشنهادی از گام‌های با هزینه محاسباتی پایین استفاده می‌کند و تمرکز آن بر روی مرحله استخراج ویژگی است نه مرحله طبقه‌بندی. دوم اینکه بیشترین زمان صرف شده برای این روش در گام باینری کردن است که در آینده می‌توان با استخراج نقاط مینوشیا از سطح خاکستری تصویر زیر کلمه سرعت را نیز بیشتر نمود. برای مثال برای یک زیرکلمه نمونه با استفاده از یک سیستم مشابه که دارای پردازنده هفت هسته‌ای و هشت گیگابایت حافظه دسترسی تصادفی است، زمان صرف شده برای روش‌های ارائه شده در مقالات [۹] و [۱۰] بترتیب برابر است با ۱۵ و ۱۰/۵ میلی‌ثانیه که این مقدار برای روش پیشنهادی حدود ۷ میلی‌ثانیه است که به وضوح بیانگر سرعت بالاتر روش پیشنهادی است. باید توجه شود هدف از کاهش فضای جستجو در گام نخست کاهش هزینه‌های محاسباتی و در گام دوم بالا بردن دقت است، لذا روش پیشنهادی که از یک ویژگی ساده همانند مینوشیا با تنها دو گام ساده در کاهش جستجو استفاده می‌کند می‌تواند گزینه مناسبی برای طراحی سیستم‌های بازشناسی در این حوزه باشد.

جدول (۵): مقایسه روش‌های مطرح با روش پیشنهادی از لحاظ متوسط میزان کاهش فضای جستجو برای دقت بالای ۹۸ درصد

روش	متوسط درصد کاهش دامنه جستجو
ویژگی‌های نسبت پهنا به ارتفاع در یک روش سلسله مراتبی [۸]	٪ ۹۴/۲
ویژگی‌های ضرایب فوریه، نرمالسازی و غیره در یک روش سلسله مراتبی [۹]	٪ ۹۹/۱
ویژگی‌های نسبت پهنا به ارتفاع و ناحیه بندی گام به گام [۱۰]	٪ ۹۹/۶
ویژگی‌های سراسری [۱۱]	٪ ۹۶/۶
ویژگی‌های سراسری و نواحی شاخص [۱۲]	٪ ۹۸/۴
گراف منفیلد [۱۳]	٪ ۸۹/۹۳
ویژگی‌های مینوشیا	٪ ۹۸/۸

۵- نتیجه‌گیری و کارهای آتی

انتخاب ویژگی مناسب می‌تواند نقش مهمی در افزایش نرخ بازشناسی زیرکلمات و متعاقب آن کاهش فضای جستجو داشته باشد. بررسی‌های نویسندگان نشان می‌دهد نوشتار زبان فارسی از لحاظ ساختاری می‌تواند شباهات زیادی به خطوط اثرانگشت و یا اثر کف دست داشته باشد، بنابراین امکان استفاده از ویژگی‌های مینوشیا که برگرفته از ساختارهای منحصربفرد در این نوع خطوط است می‌تواند کزینه‌ای مناسب برای استخراج ویژگی از زیرکلمات فراسی باشد. در این مقاله از دو ویژگی مینوشیای انشعابی و انتهایی برای استخراج ویژگی در زیرکلمات تایپ شده فارسی استفاده شده است. با استفاده از تعداد و مختصات ویژگی‌های مذکور یک روش دو مرحله‌ای برای کاهش فضای جستجو مطرح شده است. در ایتدا اسکلت تصاویر زیرکلمات پس از بهبود کیفیت تصویر و نرمالیزه شدن استخراج شده و نقاط مینوشیای آن استخراج می‌شوند، تصویر برای اساس تعداد نقاط خوشه‌بندی می‌شود و براساس فاصله اولین و آخرین نقطه انتهایی با مخزن فواصل خوشه مقایسه می‌شود. براساس نتایج بدست آمده روش پیشنهادی توانسته است فضای جستجو را از ۱۲۷۰۰ زیرکلمه به حدود ۵۰۰ زیرکلمه با دقت بالای ۹۰ درصد کاهش دهد. نتایج بدست آمده از روش پیشنهادی بسیار امیدوارکننده بوده است، اما مشکل عمده در روش پیشنهادی در خطای استخراج مینوشیا است که ناشی از فرآیند باینری کردن و استخراج اسکلت زیرکلمه است. لذا در ادامه این پژوهش استخراج نقاط مینوشیا از تصویر سطح خاکستری در دستور کار قرار خواهد گرفت. بدیهی است با توجه شباهت بالای ساختاری و شکلی (از منظر استخراج نقاط مینوشیا) اکثر فونت‌های فارسی (مانند نازنین، لوتوس، تیتر، میترا و غیره) نتایج به مراتب بهتری استخراج شود.

مراجع

[۱] تدوین نواقص دستورالعمل املایی مصوب فرهنگستان به منظور ایجاد خطایاب املایی صرفی و نحوی زبان فارسی، نسخه ۱، دانشگاه علم و صنعت ایران، سال ۱۳۸۸. شماره مستند ۱۹۰/۲۵۳۷/۱/۹. قابل بازیابی از http://arooz.net/attachments/057_09-WritingRule.pdf

[2] Pourasad, Y., Hassibi, H., And Ghorbani, A., A Word Spotting Method For Farsi Machine-Printed Document Images, Turkish Journal Of Electrical Engineering & Computer Sciences, Vol. 21, pp. 734-746, 2013.

[3] Ergün, C., And Norozpour, S., Farsi Document Image Recognition System Using Word Layout Signature, Turkish Journal Of Electrical Engineering & Computer Sciences, Vol. 27, Pp. 1477-1488, 2019.

[۴] خسروی حسین و کبیر احسان اله، "ارزیابی روش‌های بازشناسی متون فارسی بر مبنای شکل کلی زیرکلمات"، نشریه مهندسی برق و کامپیوتر، شماره ۷، دوره ۴، ۲۶۷ – ۲۸۰، ۱۳۸۸.

[5] Keyvanpour, M., Tavoli, R., and Mozaffari, S., HWS: A Hierarchical Word Spotting Method for Farsi Printed Words Through Word Shape Coding, International Journal of Information & Communication Technology Research, Vol. 7, No. 2, pp. 59-70, 2015.

[6] Bahar, P., and Mozafari, S., Farsi Machine-printed Subwords Recognition Using Contour-based Fourier Descriptors, The First International Conference on Persian Language Processing, University of Semnan, 1391.

[7] Soheili, M. R., Kabir, E., and Stricker, D., Sub-word Image Clustering in Farsi Printed Books, Seventh International Conference on Machine Vision (ICMV), 2014.

[8] Miri, E., Razavi S. M., and Mehrshad, N., Search Space Reduction in Printed Persian Sub Word Recognition by a Heretical Method, Indian Journal of Science and Technology, Vol 10, pp. 1-10, 2017.

[۹] اسماعیل میری، سیّد محمد رضوی و ناصر مهرشاد، "کاهش فضای جستجو در بازشناسی زیرواژگان تایپی فارسی با استفاده از موقعیت نقاط و علائم" مجله پردازش علائم و داده‌ها، دوره ۱۶، شماره ۳، ۱۰۱ – ۱۱۵، ۱۳۹۸.

[۱۰] اسماعیل میری، سیّد محمد رضوی و ناصر مهرشاد، "کاهش فضای جستجو برای بازشناسی زیرکلمات تایپی فارسی با استفاده از ویژگی‌های ساده، کوانتیزاسیون ویژگی و ترکیب طبقه‌بندها" مجله رایانش نرم و فناوری اطلاعات، دوره ۹، شماره ۲، ۶۱ – ۷۳، ۱۳۹۶.

[۱۱] افشین ابراهیمی، احسان الله کبیر "یک روش دو مرحله‌ای برای بازشناسی زیرکلمات چاپی"، نشریه مهندسی برق و مهندسی کامپیوتر ایران، دوره ۲، شماره ۲، ۱۳۸۳.

[۱۲] هما داودی، احسان الله کبیر "استفاده از مناطق شاخص زیرکلمات چاپی فارسی برای کاهش فضای جستجو در بازشناسی آنها"، نشریه مهندسی برق و مهندسی کامپیوتر ایران، دوره ۱۲ ، شماره ۱ ، ۱۳۹۳.

[۱۳] رسول حاجی زاده، علی آقاگل زاده و مهدی ازوجی، "آموزش منفیلد با استفاده از تشکیل گراف منفیلد مبتنی بر بازنمایی تنک" نشریه مهندسی برق و الکترونیک ایران، دوره ۱۵، شماره ۲، ۸۱-۹۵، ۱۳۹۷.

[14] Golati, S. S., and Malik, L., Handwritten Marathi Compound Character Segmentation using Minutiae Detection Algorithm, Procedia Computer Science, Vol. 87, pp. 18-24, 2016.

[15] Kholmatov, A., and Yanikoglu, B., Biometric Cryptosystem Using Online Signatures, The 21st International Symposium on Computer and Information Sciences, Istanbul, Turkey, November 1-3, 2006.

زیرنویس‌ها

[1] minutiae

[2] fingerprint

[3] Ending points

[4] bifurcation

[5] segmentation

[6] signature

[7] Fingerprint matching

[8] core

[9] delta