



کاهش فضای جستجو در بازشناسی زیرواژگان تایپی فارسی با استفاده از موقعیت نقاط و علائم

اسماعیل میری، سید محمد رضوی* و ناصر مهرشاد

گروه الکترونیک، دانشکده مهندسی برق و کامپیوتر، دانشگاه بیرجند، بیرجند، ایران

چکیده

در این مقاله با استفاده از روشی ساده، اما کارا سعی شده دامنه جستجوی زیرواژگان به شدت کاهش یابد. در گام آموزش، داده‌های آموزشی براساس موقعیت علائم گروه‌بندی می‌شوند، در گروه‌هایی که تعداد عناصر بیش از ده زیرواژه است، برای کاهش فضای جستجو با توجه به تعداد عناصر گروه، با استخراج ویژگی‌های ساده‌ای از پروفایل‌های افقی و عمودی خوشه‌بندی صورت می‌گیرد. در مرحله بازشناسی در نخستین مرحله با تعیین نسبت پهنا به ارتفاع زیرواژه (با علائم و بی‌علائم) و کد موقعیت نقاط و علائم، دامنه جستجو به زیرواژگانی با این کد موقعیت که در محدوده‌ای از نسبت‌های یادشده باشند، محدود می‌شود؛ در صورتی که تعداد زیرواژگان محدود شده در این مرحله کمتر از ده باشد، این محدوده پذیرفته و در غیر این صورت در مرحله بعد با استخراج ویژگی‌های ساده‌ای از پروفایل‌های افقی و عمودی فضای جستجو به تعدادی از نزدیکترین خوشه‌ها به این زیرواژه که شرط نسبت پهنا به ارتفاع را نیز ارضا کنند محدود می‌شود. با اعمال روش پیشنهادی این مقاله فضای جستجو تا حد قابل قبولی کاهش یافته است.

واژگان کلیدی: بازشناسی، زیرواژگان تایپی فارسی، کاهش فضای جستجو، موقعیت نقاط و علائم

Search Space Reduction for Farsi Printed Subwords Recognition by Position of the Points and Signs

Esmail Miri, Seyyed Mohammad Razavi* & Nasser Mehrshad

Faculty of Electrical and computer Engineering, University of Birjand, Birjand, Iran

Abstract

In the field of the words recognition, three approaches of words isolation, the overall shape and combination of them are used. Most optical recognition methods recognize the word based on break the word into its letters and then recognize them. This approach is faced some problems because of the letters isolation difficulties and its recognition accuracy in texts with a low image quality. Therefore, an approach based on none separating recognition could be useful in such cases.

In methods based on the overall shapes for subword recognition after extraction of subword features usually these features are searched in the image dictionary created in the training phase. Therefore, by considering that we are faced with massive amounts of classes, proposing ways to limit the scope of the search are the main challenges in the overall shape methods. Thus, the information of the overall shape usually is used to reduce the scope search in a hierarchical form.

In this paper, it is tried to reduce the search space of the subwords severely by using a simple and efficient method. In training phase, training data is grouped based on the location of the points and signs, in the groups where have more than 10 subwords, to reduce the search space, according to the number of elements in the group, by extracting the simple features of horizontal and vertical profiles clustering takes place. In

* Corresponding author

* نویسنده عهده‌دار مکاتبات

recognition phase, in the first step, by determining the width to height ratio of the subword (with signs and without signs) and the position code of the points and signs, the search scope is limited to subwords with this position code that are within the range of the ratios mentioned. This range would be accepted if the number of subwords in this phase is less than ten. Otherwise, in the next step, by extracting the simple features of the horizontal and vertical profiles of the subwords, the search space will be limited to a number of the closest clusters to this subword that also satisfies the width-to-height ratio. By using the proposed method of this paper, the search space has fallen to an acceptable level.

In this study, a database of 12700 subwords with five Lotus, Zar, Nazanin, Mitra and Yaghut fonts scanned 400 dpi was used. The four Lotus, Zar, Nazanin and Mitra fonts were used in the training phase and in the test phase, Yaghut font is used.

Keywords: Recognition, Farsi Typed Subwords, Search Space Reduction, Position of the Points and Symbols.

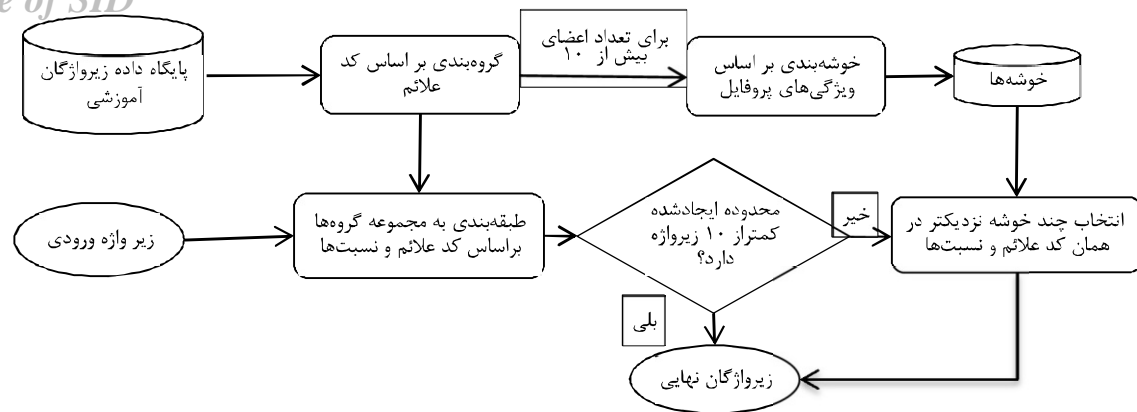
مواجه هستیم، ارائه روش‌هایی برای محدود کردن دامنه این جستجو از چالش‌های اصلی موجود در روش‌های مبتنی بر شکل کلی است؛ از این رو اطلاعات شکل کلی به‌طور معمول برای کاهش دامنه جستجو در یک سامانه سلسله‌مراتبی به‌کار می‌رود [4-6]. علاوه بر سامانه‌های بازشناسی، استفاده از اطلاعات شکل کلی واژه در بازیابی واژه میان مجموعه محدود واژگان، حجم پردازش را به شکل قابل ملاحظه‌ای کاهش می‌دهد. همچنین، توصیف شکل کلی واژه، روشی کارآمد برای نشان کردن واژگان پرس و جو در تصاویر اسناد است [7-13]. کاهش فضای جستجو، علاوه بر کاهش حجم محاسبات مورد نیاز در مراحل بعد، دقت نهایی سامانه بازشناسی را نیز افزایش خواهد داد.

هدف از انجام این پژوهش ارائه یک روش ساده و در عین حال کارا جهت کاهش فضای جستجوی زیرواژگان فارسی است. در نخستین مرحله در گام آموزش زیرواژگان بر اساس کد موقعیت نقاط و علائم آنها گروه‌بندی می‌شوند؛ سپس در گروه‌هایی که تعداد اعضای آن بیش از ده زیرواژه است با استخراج ویژگی‌های ساده‌ای از پروفایل‌های افقی و عمودی با توجه به تعداد اعضا، خوشه‌بندی با این ویژگی‌ها صورت می‌گیرد. در گام بازشناسی در نخستین مرحله با تعیین نسبت پهنا به ارتفاع زیرواژه (با علائم و بی‌علائم) و کد موقعیت نقاط و علائم، دامنه جستجو به زیرواژگانی با این کد موقعیت که در محدوده‌ای از نسبت‌های یادشده باشند، محدود می‌شود. در صورتی که تعداد زیرواژگان محدود شده در این گام کمتر از ده باشد این محدوده پذیرفته می‌شود، در غیر این صورت در مرحله بعد با استخراج ویژگی‌های ساده‌ای از پروفایل‌های افقی و عمودی فضای جستجو به تعدادی از نزدیک‌ترین خوشه‌ها (مربوط به همان کد موقعیت نقاط) به این زیرواژه که شرط نسبت پهنا به ارتفاع را نیز ارضا کنند محدود می‌شود. شکل (۱) اجزای اصلی روش پیشنهادی را نشان می‌دهد.

۱- مقدمه

امروزه به‌منظور ایجاد امکان ذخیره اسناد کتابخانه‌ها، موزه‌ها و بایگانی‌ها به‌صورت دیجیتالی و با قابلیت جستجو، روش‌های مختلف بازشناسی واژگان مورد توجه قرار گرفته‌اند. در زمینه بازشناسی واژگان سه رویکرد جداسازی واژگان، شکل کلی و ترکیبی از این دو استفاده شده است. واژگان در زبان فارسی از کنار هم قرار گرفتن حروف شکل می‌یابند که برخی از این حروف به هم می‌چسبند و برخی به‌صورت مجزا در ترکیب واژه حضور دارند. به قسمت‌های به هم چسبیده یک واژه زیرواژه گفته می‌شود. اکثر روش‌های بازشناسی نوری نویسه با شکستن واژه به حروف تشکیل‌دهنده آن و بازشناسی این حروف، واژه را بازشناسی می‌کنند. این رویکرد به دلیل مشکلات جداسازی حروف و بازشناسی دقیق آنها در متون با کیفیت تصویری پایین با مشکلاتی مواجه است؛ بنابراین رویکرد مبتنی بر بازشناسی بدون جداسازی در چنین مواردی کارساز خواهد بود. از طرفی پژوهش‌های روان‌شناسی زیادی در مورد نحوه قرائت انسان انجام شده است که یکی از نتایج این پژوهش‌ها این است که چشم انسان در زمان مطالعه یک خط متن، به‌طور پیوسته از راست به چپ حرکت نمی‌کند؛ بلکه به‌صورت گسسته از یک نقطه تمرکز به نقطه تمرکز دیگر جهش می‌کند. این پژوهش‌ها نشان داده زمان لازم برای بازشناسی یک کلمه چهارحرفی برابر با زمان لازم برای بازشناسی یک حرف مجزا است. پژوهش‌های انجام‌شده بر اهمیت شکل کلی واژه در فرآیند بازشناسی تأکید کرده و با توجه به این موضوع برای به‌کارگیری ویژگی‌های تصویری در سطح واژه، روش‌هایی پیشنهاد شده است [1-3].

در روش‌های مبتنی بر شکل کلی، برای بازشناسی زیرواژه به‌طور عمومی پس از استخراج ویژگی‌های زیرواژه، این ویژگی‌ها در واژه‌نامه تصویری تشکیل‌شده در مرحله آموزش جستجو می‌شود؛ لذا با توجه به این که با حجم انبوهی از رده‌ها



(شکل-۱): اجزای اصلی روش پیشنهادی

(Figure-1): The main components of the proposed process

قبل استفاده می‌شود. نتایج روی پایگاه داده IFN/ENIT که شامل ۲۶۴۵۹ تصویر زیرواژه است، کاهش ۹۲/۵ درصدی فضای جستجو با دقت ۷۴٪ را نشان می‌دهد.

پژوهش [19] یک روش کاهش فضای جستجو برای اسناد تاریخی عربی را ارائه می‌دهد که تصویر زیرواژه ورودی را با مدخل‌های لغت‌نامه مقایسه کرده و موارد با بیشترین تشابه را انتخاب می‌کند. در مقایسه تصاویر زیرواژگان، اهمیت بیشتری به مناطق شاخص شکل داده می‌شود، مناطق شاخص مناطق محلی از زیرواژه تعریف می‌شود که آن را از سایر زیرواژگان لغت‌نامه متفاوت می‌کند. در این روش ابتدا یک معیار مبتنی بر بازیابی برای محاسبه نمره تمایز برای هر منطقه محلی اعمال می‌شود که نشان می‌دهد که آن ناحیه از شکل چقدر شاخص است. این نمرات در اندازه فاصله تعریف‌شده برای تعدیل وزن‌های ویژگی‌های شکل مربوطه استفاده می‌شود؛ به‌طوری‌که که مناطق با تمایز بیشتر، وزن بیشتری را به خود اختصاص می‌دهند. یک مرحله کاهش فضای ویژگی مبتنی بر ویژگی‌های کلی‌نگر مکان مشخصه به‌منظور تکمیل این توصیف‌گر محلی استفاده شده است. با روش پیشنهادی در پایگاه داده ابن‌سینا، حاوی بیش از دوازده‌هزار زیرواژه استخراج‌شده از یک سند تاریخی عربی، میزان کاهش ۹۸/۱۵٪ با دقت ۹۰/۱۵٪ به‌دست آمده است.

پژوهش [20] یک روش بازشناسی واژگان دست‌نویس فارسی مبتنی بر کاهش فضای جستجو ارائه داده است. در این روش پس از استخراج ویژگی، واژگان در فرهنگ لغت خوشه‌بندی می‌شوند. میانگین هر خوشه در فضای ویژگی به‌عنوان نماینده خوشه و مدخل مشترک اعضای آن خوشه در فرهنگ لغت در نظر گرفته می‌شود. در این روش روی مجموعه داده ایرانشهر در مرحله بازشناسی با انتخاب پنج خوشه نزدیک‌تر به واژه مورد آزمون با دقت ۹۳/۳۷٪ حدود ۷۶/۶۵٪ کاهش فضای جستجو را شاهد هستیم.

پایگاه داده استفاده‌شده در این مقاله مجموعه ۱۲۷۰۰ زیرواژه رایج زبان فارسی [14] است که با پنج قلم لوتوس، زر، نازنین، میترا و یاقوت با اندازه قلم چهارده نگارش و چاپ شده و با درجه تفکیک چهارصد نقطه در اینچ روبش شده‌اند. در ادامه در بخش ۲ به مرور روش‌های موجود پرداخته و در بخش ۳ ساختار کلی سامانه پیشنهادی ارائه می‌شود. ارزیابی روش پیشنهادی و نتایج در بخش ۴ آمده و در بخش ۵ جمع‌بندی و نتیجه‌گیری ارائه شده است.

۲- مروری بر کارهای گذشته

پیوستگی حروف در برخی خط‌ها و در بعضی از شیوه‌های نگارش، قطعه‌بندی شکل واژه را پیچیده‌تر می‌سازد و موجب می‌شود که پژوهش‌گران گرایش بیشتری به سمت روش‌های مبتنی بر شکل کلی داشته باشند. خطوط فارسی، عربی و همچنین دست‌نویس انگلیسی از این دسته هستند. پژوهش‌های متعددی درباره بازشناسی خطوط پیوسته انجام شده است [15].

در مرجع [16] روش‌های مختلف ارائه‌شده برای بازشناسی واژگان دست‌نویس با رویکرد مبتنی بر شکل کلی بررسی شده و مرجع [17] به مرور روش‌های بازشناسی واژگان عربی با هر دو رویکرد مبتنی بر قطعه‌بندی و مبتنی بر شکل کلی پرداخته است.

پژوهش [18] یک استراتژی دومرحله‌ای برای حذف نامزدهای غیر شبیه قبل از بازشناسی واژگان دست‌نویس عربی برای افزایش سرعت بازشناسی ارائه داده است. اصول این روش شامل استخراج نقاط و زیرواژه‌ها از تصویر واژه پیوسته عربی برای توصیف شکل آن است. در نخستین قدم از کاهش فضای جستجو، تعداد زیرواژگان واژه ورودی تخمین زده می‌شود؛ سپس در دومین مرحله از اطلاعات نقاط در نامزدهای مرحله

افقی و عمودی، فضای جستجو به تعدادی از خوشه‌های منتخب محدود شده است. در دومین مرحله با تعیین نسبت پهنا به ارتفاع زیرواژه، دامنه جستجو به زیرواژگانی با محدوده‌ای از این نسبت محدود شده است. در مرحله سوم با توجه به موقعیت علائم تنها زیرواژه‌هایی مورد جستجو قرار می‌گیرند که موقعیت علائم آنها با زیرواژه ورودی یکسان باشند. با اعمال روش پیشنهادی فضای جستجو تا حد قابل قبولی کاهش یافته است.

۳- ساختار کلی روش پیشنهادی

ساختار روش پیشنهادی در شکل (۲) دیده می‌شود. در این پژوهش از پایگاه داده این‌گونه استفاده شده که از چهار قلم لوتوس، زر، نازنین، میترا برای آموزش و قلم یاقوت در مرحله آزمایش استفاده شده است.

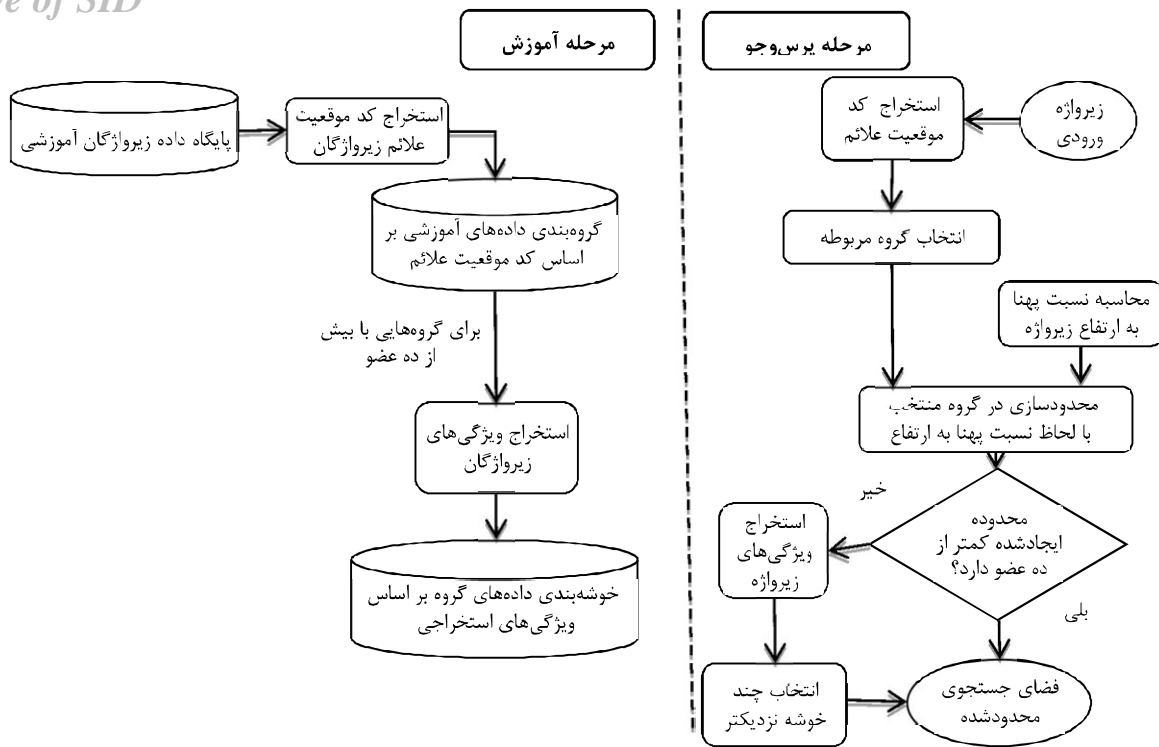
در گام آموزش ابتدا پایگاه داده زیرواژگان آموزشی بر اساس کد موقعیت علائم گروه‌بندی می‌شوند؛ سپس برای گروه‌هایی که بیش از ده عضو داشته باشند، ویژگی‌های ساده‌ای از پروفایل‌های افقی و عمودی استخراج شده و با توجه به تعداد اعضا (به‌ازای هر ده عضو یک خوشه) خوشه‌بندی با استفاده از این ویژگی‌ها صورت می‌گیرد. در گام بازشناسی، کاهش فضای جستجو در چند مرحله صورت می‌گیرد. در مرحله نخست کد موقعیت علائم زیرواژه ورودی و نسبت پهنا به ارتفاع (با علائم و بدون علائم) آن استخراج می‌شود. با توجه به کد موقعیت علائم استخراجی، فضای جستجو به محدوده‌ای از زیرواژگان که کد موقعیت علائم یکسان دارند، محدود می‌شود (گروه مربوطه) در مرحله بعد با توجه به نسبت پهنا به ارتفاع زیرواژه ورودی، فضای جستجو در گروه مربوطه به زیرواژگانی که نسبت پهنا به ارتفاع (با علائم و بدون علائم) در محدوده مشخص داشته باشند، محدود می‌شود. در این مرحله اگر فضای جستجوی ایجاد شده حاوی کمتر از ده زیرواژه باشد، محدودسازی پایان یافته است؛ در غیر آن صورت با استخراج ویژگی‌های پروفایل‌های افقی و عمودی با سنجش فاصله این ویژگی‌ها از مراکز خوشه‌ها، تعدادی از خوشه‌های نزدیک انتخاب و اعضای این خوشه‌ها با لحاظ محدودیت نسبت پهنا به ارتفاع (با علائم و بدون علائم) به‌عنوان محدوده مورد نظر انتخاب می‌شوند. در ادامه هر کدام از مراحل یادشده تشریح شده است.

در پژوهش‌ها و کارهای انجام‌گرفته قبلی درخصوص زیرواژگان تایی فارسی به طرق مختلف در جهت کاهش دامنه جستجو تلاش شده است. مرجع [21] از تخمین توصیف‌گر سازگار با شکل زیرواژه جهت کاهش فضای جستجو استفاده کرده است. روش پیشنهادی به این ترتیب است که ویژگی‌های مکان مشخصه، ناحیه‌بندی و تبدیل فوریه به‌عنوان سه ویژگی مناسب جهت بازشناسی زیرواژگان فارسی انتخاب شده‌اند. در مرحله آموزش، یک شبکه عصبی جهت تشخیص توصیف‌کننده مناسب از بین سه توصیف‌کننده انتخاب‌شده آموزش می‌بیند. در مرحله آزمایش، ابتدا شبکه عصبی آموزش‌دیده ویژگی توصیف‌گر مناسب جهت زیرواژه ورودی را تخمین می‌زند، سپس بر اساس این ویژگی تعدادی از نزدیک‌ترین زیرواژگان به زیرواژه ورودی به‌عنوان فضای جستجوی کاهش یافته انتخاب می‌شوند.

ابراهیمی برای کاهش دامنه جستجو ابتدا زیرواژگان پایگاه داده را که از قلم‌ها و اندازه‌های متفاوت بودند با ویژگی‌های مکان مشخصه به سیصد خوشه تقسیم کرد [4]. در مرحله طبقه‌بندی، زیرواژه ورودی با مرکز خوشه‌های یادشده مقایسه شده و نزدیک‌ترین خوشه‌ها با استفاده از ویژگی‌های توصیف‌گرهای فوریه مورد جستجو قرار می‌گیرند. در پژوهش دیگری دودی از خوشه‌بندی با ویژگی‌های سراسری و افزایش میزان اطمینان به خوشه انتخابی براساس ویژگی‌های محلی شکل جهت کاهش تعداد خوشه‌های انتخابی و در نتیجه کاهش بیشتر فضای جستجو استفاده کرده است [22].

در پژوهش دیگری فتحی از حروف شاخص برای کاهش دامنه جستجو بهره برده است [23]. در این پژوهش حروف شاخص نخست و آخر زیرواژه بدون جداسازی شناسایی و در مرحله طبقه‌بندی زیرواژه ورودی تنها در زیرواژگانی که دارای این حروف شاخص‌اند جستجو می‌شود. در پژوهش دیگری علی‌بیگی از کد نقاط به این منظور استفاده کرده است [24]. در این پژوهش در مرحله طبقه‌بندی زیرواژه ورودی تنها در بین زیرواژگانی که دارای کد نقاط مشابه هستند جستجو می‌شود؛ لذا دامنه جستجو به‌شدت کاهش می‌یابد.

در پژوهش دیگری میری از خوشه‌بندی، کد موقعیت علائم و نسبت پهنا به ارتفاع زیرواژگان برای محدودسازی فضای جستجو استفاده کرده است [25]. در این پژوهش در نخستین مرحله با استخراج ویژگی‌های ساده‌ای از پروفایل‌های



(شکل-۲): ساختار روش پیشنهادی
(Figure-2): The structure of the proposed method

۶. اگرچه فاصله علائم، تابعی از اندازه قلم است، اما این فاصله‌ها در قلم‌های مختلف متفاوت است.

(جدول ۱): موقعیت علائم در چند زیرواژه نمونه

(Table-1): The position of the syboles in a few sample subwords

قلم نمونه	میترا	نازنین	لوتوس	زر	یاقوت
ثت	ثت	ثت	ثت	ثت	ثت
ثت	ثت	ثت	ثت	ثت	ثت
ثت	ثت	ثت	ثت	ثت	ثت
تنت	تنت	تنت	تنت	تنت	تنت
تنت	تنت	تنت	تنت	تنت	تنت
گج	گج	گج	گج	گج	گج

با لحاظ خواص قیدشده و با روش سعی و خطا درنهایت در جهت ایجاد قواعد یکنواخت برای استخراج کد علائم از پنج قلم مورد نظر مقاله، مراحل زیر طراحی شد: مراحل استخراج کد علائم عبارتند از:

۱. برجسب‌زنی به اجزای زیرواژه و جداسازی بدنه زیرواژه (بزرگ‌ترین جز)
۲. استخراج پهنای قلم (پرتکرارترین ضخامت موجود در بدنه زیرواژه)
۳. ترکیب نقاط مربوط به یک حرف

۳-۱- کد موقعیت علائم زیرواژه

در کارهای انجام‌شده قبلی از جمله در [4] و [23] و [24] از کد علائم زیرواژه در مراحل مختلف کار استفاده شده که در نوشتار با اندازه قلم کوچک و دقت تفکیک پایین منجر به خطای بالایی می‌شود؛ لذا در این مقاله برای حل این مشکل از کد موقعیت علائم جهت محدودسازی فضای جستجو استفاده شده است. جدول (۱) چند نمونه زیرواژه با علائم مختلف را نشان می‌دهد. برخی از خواص علائم که از این جدول قابل استنباط است عبارتند از:

۱. به‌طور تقریبی در تمام قلم‌ها با عنایت به شکل خاص سه نقطه تشخیص و ترکیب آن در نخستین قدم به ذهن متبادر می‌شود.
۲. دو نقطه مربوط به یک حرف در بیش‌تر قلم‌ها به هم چسبیده است؛ اما با توجه به نسبت پهنا به ارتفاع از یک نقطه قابل شناسایی است.
۳. فاصله دو نقطه مربوط به یک حرف، به‌مراتب کمتر از دو نقطه مربوط به دو حرف است.
۴. در دو حرف علامت‌دار کنار هم موقعیت عمودی مرکز علامت در بیش‌تر قلم‌ها متفاوت است.
۵. گاهی زیر یا رو بودن علامت تنها با داشتن خط زمینه قابل تشخیص نیست (گج).

برخورد کردیم؛ موقعیت علامت را پایین در نظر می‌گیریم در غیر آن صورت موقعیت علامت بالا در نظر گرفته می‌شود.

اگر $X_{cmean} > X_{bmax}$ از نقطه‌ای به مختصات (X_{bmax}, Y_{cmean}) به سمت بالا حرکت می‌کنیم، اگر به بدنه برخورد کردیم، موقعیت علامت پایین و در غیر این صورت موقعیت علامت بالا در نظر گرفته می‌شود.

اگر $X_{cmean} < X_{bmin}$ از نقطه‌ای به مختصات (X_{bmin}, Y_{cmean}) به سمت بالا حرکت می‌کنیم، اگر به بدنه برخورد کردیم، موقعیت علامت پایین و در غیر این صورت موقعیت علامت بالا در نظر گرفته می‌شود.

در مرحله بعد به هر زیرواژه یک کد موقعیت علائم اختصاص می‌یابد. گفتنی است که در تشخیص نوع علامت امکان خطای زیادی وجود دارد؛ ولی در تعیین موقعیت علامت با الگوریتم استفاده شده تنها ممکن است، خطاهای نادری ناشی از کیفیت پایین روبش تصاویر اتفاق بیفتد.

۵. اختصاص کد (تعیین گروه)

روش اختصاص کد به زیرواژگان به این ترتیب است که زیرواژه مورد نظر از راست به چپ مورد بررسی قرار می‌گیرد. در هر حرف اگر یکی از علائم (یک نقطه، دو نقطه (ترکیب شده)، سه نقطه (ترکیب شده)، سرکش، مد، تشدید، همزه) در بالای بدنه قرار گرفته باشد به آن حرف کد یک اختصاص می‌یابد و در صورتی که هر یک از علائم در زیر بدنه قرار گرفته باشد به آن کد دو تعلق می‌گیرد. در غیر آن صورت کدی به آن حرف تعلق نمی‌گیرد و در نهایت از کنار هم قرار گرفتن کد حروف از چپ به راست، کد زیرواژه به دست می‌آید. به زیرواژه‌ای که فاقد هرگونه علامت در پایین و بالا باشد کد صفر اختصاص می‌یابد. دوباره یادآوری می‌شود اعداد و نسبت‌های به کار گرفته شده در مراحل استخراج کد موقعیت علائم با سعی و خطا و کوشش در جهت ارائه یک طرح و فرمول فراگیر برای پوشش هر پنج قلم به دست آمده است.

جدول (۲) تعدادی از زیرواژه‌ها و کد موقعیت علائم استخراجی آنها را نشان می‌دهد. در جدول (۳) فراوانی کدهای مختلف در پایگاه داده نشان داده شده است. همان‌طور که در این جدول دیده می‌شود، بیشترین فراوانی در مجموعه ۸۶ کد به دست آمده با ۱۶/۸٪ مربوط به کد یک (یک علامت بالا) است. از این ۸۶ کد ۴۶ کد دارای فراوانی کمتر از ده بوده که ۲۲ کد آن تنها یک عضو دارد. از تحلیل مجموعه اطلاعات ارائه شده توسط این جدول، کارایی بالای استفاده از این

در این راستا ابتدا علائم زیرواژه از راست به چپ مرتب می‌شوند؛ سپس نخستین علامت با علامت سمت چپ آن (در صورت وجود) در شروط زیر بررسی می‌شود:

ا. اگر فاصله افقی مرکز دو علامت کمتر از دو برابر پهنای قلم است.

ب. اگر فاصله عمودی لبه‌های دو علامت کمتر از ۱/۷ برابر پهنای قلم است.

در صورت برآورده شدن هم‌زمان شرایط بالا دو علامت با هم ترکیب شده و به عنوان یک علامت لحاظ می‌شوند و بار دیگر مراحل بالا با جزء ترکیبی تکرار و همین مراحل برای سایر علائم به ترتیب اجرا می‌شود. پس از تکمیل این فرایند، اجزای مربوط به سه نقطه و تعدادی از دو نقطه‌ها ترکیب می‌شوند.

در مرحله بعد اجزای ترکیب شده قبلی کنار گذاشته می‌شوند و آنهایی را که مانده‌اند، در شرایط زیر بررسی می‌کنیم؛ بر آورده شدن این شرایط به صورت هم‌زمان باعث خواهد شد دو نقطه‌های مجزای مربوط به یک حرف که در مرحله قبل ترکیب نشده‌اند، با هم ترکیب شوند:

ا. اگر فاصله عمودی مرکز دو نقطه کمتر از یک سوم پهنای قلم است؛

ب. اگر فاصله افقی لبه‌های دو نقطه کمتر از ۱/۵ برابر پهنای قلم است؛

ت. اگر نسبت نسبت‌های پهنای به ارتفاع دوجزئی که شرایط بالا را دارند، کمتر از ۱/۶ برابر است (این شرط باعث خواهد شد که یک نقطه با دو نقطه به هم چسبیده حرف مجاور ترکیب نشود)؛

ث. اگر نسبت مساحت دو جزء بین ۰/۵ تا ۲ است (این شرط از ترکیب همزه و نقطه مجاور جلوگیری می‌کند).

۴. تعیین موقعیت علائم

برای تعیین بالا یا پایین بودن علامت اگر مختصات گوشه چپ بالای چارچوب بدنه را X_{bmin} و Y_{bmin} و مختصات گوشه راست پایینی چارچوب بدنه را X_{bmax} و Y_{bmax} در نظر بگیریم و برای مختصات مربوط به علامت از X_{cmin} ، Y_{cmin} ، X_{cmax} و Y_{cmax} استفاده کنیم، مختصات مرکز چارچوب علامت عبارت خواهد بود از:

$$X_{cmean} = \left(\frac{X_{cmin} + X_{cmax}}{2} \right) \text{ و } Y_{cmean} = \left(\frac{Y_{cmin} + Y_{cmax}}{2} \right)$$

اگر $X_{bmin} \leq X_{cmean} \leq X_{bmax}$ باشد از مرکز چارچوب علامت به سمت بالا حرکت می‌کنیم اگر به بدنه زیرواژه

شکل (۳) نحوه استخراج ویژگی برای پروفایل‌های بالا و سمت راست برای یک زیرواژه نمونه را نشان می‌دهد. نمونه‌های استخراجی و ویژگی نهایی استخراج شده پروفایل بالایی زیرواژه شکل (۳) در جدول (۴) و (۵) آمده است. این ویژگی‌ها در هر چهار جهت به همین ترتیب استخراج و در نهایت جهت نرمال‌سازی در هر بعد بر اندازه همان بعد تقسیم می‌شوند. در سطر دوم جدول (۴) ویژگی‌های نهایی از تقسیم سطر نخست همان جدول بر ۳۵ که ارتفاع زیرواژه شکل (۳) است به دست آمده‌اند.

روش ساده انتخاب ویژگی و تعداد کم ویژگی‌ها (چهل ویژگی) در مقایسه با کارهای قبلی نقطه قوت روش پیشنهادی بوده و باعث افزایش سرعت و سادگی کار شده است. به عنوان نمونه در کارهای [4] و [18] از ویژگی‌های مکان مشخصه با سه تقاطع (۲۵۶ ویژگی) در مرحله خوشه‌بندی استفاده شده است.

(جدول-۲): کد موقعیت علائم چند زیرواژه

(Table-2): The position code of few subwords

زیرواژه	'بمیهن'	'بنگلا'	'بنگلا'	'بنگیا'	'بنگیا'
کد موقعیت علامت	221	211	21	212	2112
زیرواژه	'بنقشه'	'بنقضا'	'بنقطه'	'بنهند'	'بنهیم'
کد موقعیت علامت	2111	211	211	211	212
زیرواژه	'بهجتی'	'بهشتش'	'بهشتی'	'بهشهر'	'بهمکا'
کد موقعیت علامت	221	2111	211	21	2

محدودکننده به روشنی قابل مشاهده است. به عنوان یک نکته اصلاحی با توجه به اینکه امکان اتصال سرکش به بدنه در تعدادی از قلم‌ها وجود دارد، برای جلوگیری از خطا برای حرف "گ" هر دو کد بدون علامت و با یک علامت بالا لحاظ شده که این امر باعث اضافه شدن تعداد عناصر پایگاه داده از ۱۲۷۰۰ به ۱۳۸۸۲ شده است.

۳-۲- استخراج ویژگی

در کارهای انجام شده قبلی برای بازشناسی زیرواژگان از ویژگی‌های ساختاری و آماری متفاوتی استفاده شده است. در این مقاله با توجه به هدف که طراحی روشی ساده و کارا جهت کاهش دامنه جستجو است، ویژگی‌های ساده‌ای از پروفایل‌های افقی و عمودی استخراج شده است. با عنایت به طول متغیر بدنه زیرواژه‌ها و متغیر شدن طول بردار ویژگی، برای تولید بردارهای ویژگی با طول یکسان برای هر زیرواژه روش‌های متفاوتی از جمله انتخاب N ضریب نخست تبدیل فوریه، درون‌یابی، نرمال‌سازی و ... مورد آزمون قرار گرفت و با توجه به نتایج به دست آمده در نهایت با روش نمونه‌برداری که در ادامه تشریح شده برای هر زیرواژه از هر جهت تنها ده ویژگی و در مجموع برای هر زیرواژه چهل ویژگی از پروفایل‌های افقی و عمودی تصویر زیرواژه استخراج شد.

(جدول-۳): کدهای موقعیت علائم و فراوانی آنها در زیرواژگان پایگاه داده

(Table-3): The position codes of the signs and their frequency in the subwords of the database

تعداد	کد	تعداد	کد	تعداد	کد	تعداد	کد	تعداد	کد	تعداد	کد	تعداد	کد	تعداد	کد
2	212121	1	11222	1	22122	1	1222	26	1112	20	1222	482	211	856	0
1	221111	1	12111	3	22211	4	2111	2	1112	88	2111	230	212	213	1
1	221211	1	12112	4	22212	2	2111	18	1121	39	2112	379	221	138	2
1	221212	1	12122	1	11111	13	2112	4	1121	77	2121	120	222	183	11
1	221222	2	12222	1	11121	3	2112	19	1122	17	2122	125	111	116	12
1	222121	2	21111	1	11112	11	2121	16	1211	86	2211	48	111	141	21
1	112111	1	21112	3	11121	2	2121	4	1211	36	2212	172	112	712	22
1	1122211	3	211211	1	111212	4	21221	4	12121	28	221	48	1122	627	111
1	2112111	3	211212	2	111221	8	22111	10	12211	9	2222	118	1211	379	112
		1	211221	8	112111	11	22112	2	12212	16	11111	24	1212	667	121
		1	212112	1	112121	13	22121	4	12221	8	11112	86	1221	212	122

می‌کنیم که نخستین آنها برابر یک و آخرین آنها برابر N باشد. بین یک تا N را به $M-1$ بازه مساوی تقسیم می‌کنیم و $M-2$ عدد حقیقی دیگر را به دست می‌آوریم و این اعداد را گرد می‌کنیم تا شماره عناصری را که باید از N عنصر اولیه انتخاب کنیم، داشته باشیم. در صورتی که N بزرگ‌تر یا مساوی M باشد، شماره‌های عناصر تکراری نخواهد بود؛ ولی اگر N بزرگ‌تر یا مساوی M باشد، شماره‌های عناصر تکراری نیز داریم که نمونه‌ها را چندبار انتخاب می‌کنیم.

در مثال بالا $N=70$ و $M=10$ است. اگر بخواهیم بین یک تا هفتاد را به نه قسمت مساوی تقسیم کنیم، این اعداد را خواهیم داشت:

۱ ۸/۶۷ ۱۶/۳۳ ۲۴ ۳۱/۶۷ ۳۹/۳۳ ۴۷ ۵۴/۶۷ ۶۲/۳۳ ۷۰

گردشده این اعداد عبارتند از:

۱،۹،۱۶،۲۴،۳۲،۳۹،۴۷،۵۵،۶۲،۷۰

بر این اساس عناصری که در جدول (۴) با رنگ قرمز مشخص شده‌اند، انتخاب و در جدول (۵) آمده‌اند؛ ولی برای مثال اگر $N=7$ و $M=10$ باشد، این اعداد را داریم:

۱ ۱/۶۷ ۲/۳۳ ۳ ۳/۶۷ ۴/۳۳ ۵ ۵/۶۷ ۶/۳۳ ۷

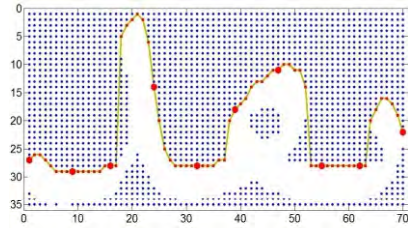
و اگر این اعداد را گرد کنیم اعداد زیر به دست می‌آید:

۱،۲،۳،۴،۵،۶،۷

همان‌طور که مشاهده می‌شود، اعداد ۲،۴،۶ تکرار شده‌اند؛ بنابراین عناصر دوم، چهارم و ششم هر کدام دو بار و بقیه فقط یک‌بار انتخاب خواهند شد.

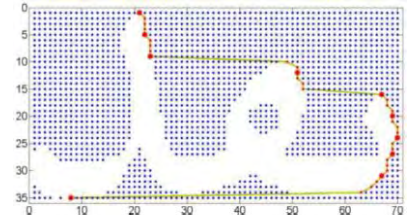
۳-۳- لحاظ نسبت پهنا به ارتفاع واژه

از محدودیت‌های ساده و مفیدی که برای کاهش فضای جستجو استفاده شده، به‌کارگیری نسبت پهنا به ارتفاع زیرواژه ورودی است. ویژگی عمده این محدودکننده این است که به‌علت ساختار نسبی آن مستقل از اندازه قلم است. در این مقاله برای کاهش فضای جستجو و در عین حال پوشش کل زیرواژه‌ها در قلم‌های پنج‌گانه، با انجام آزمایش‌های متعدد، در نهایت فضای جستجو در خوشه‌های مربوطه به زیرواژه‌هایی که نسبت پهنا به ارتفاع بین $0/6$ تا $1/4$ نسبت پهنا به ارتفاع زیرواژه ورودی دارند (با علائم و بدون علائم) محدود شده است. شکل (۴) نحوه محاسبه نسبت پهنا به ارتفاع یک زیرواژه نمونه با علائم و بدون علائم و جدول (۶) همین نسبت‌ها را برای چند زیرواژه نمونه نشان می‌دهند.



(الف)

(A)



(ب)

(B)

(شکل-۳): استخراج ویژگی از بدنه زیرواژه، شکل سفیدرنگ:

زیرواژه پس از حذف علائم، نمودار سبز: پروفایل زیرواژه،

نقطه‌های قرمز روی نمودار سبزرنگ: ویژگی‌های پروفایل،

دایره‌های روی منحنی سبز: ویژگی‌های نمونه برداری شده از

پروفایل (شکل الف ویژگی‌های نمایه بالا و شکل ب ویژگی‌های

پروفایل سمت راست).

(Figure-3): Extracting the feature from the body of subword, White Shape: after the removal of symbols, Green Curve: The profile of the subword, Red dots on the green Curve: Profile features, Circles on the Green Curve: Sample Profiles from the Profile (A: Top Profile features, and B: right Profile features).

(جدول-۴): ویژگی‌های استخراج شده از پروفایل بالایی

زیرواژه نمونه

(Table-4): upper profile Extracted Features of the sample subwords

27	26	26	27	28	29	29	29	29	29	29	29	29	29	29
28	28	28	5	3	2	1	2	6	14	20	25	27	28	28
28	28	28	28	28	28	27	27	20	18	17	16	14	14	14
13	13	12	11	11	10	10	11	11	14	28	28	28	28	28
28	28	28	28	28	28	28	20	18	16	16	17	19	22	22

(جدول-۵): ویژگی‌های نمونه برداری شده پروفایل بالایی

(Table-5): Upper profile sampled Features

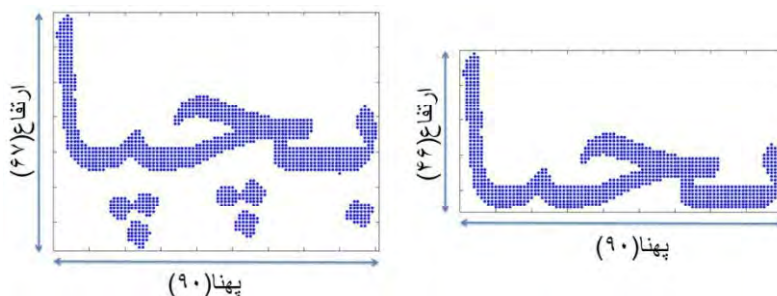
نمونه‌های استخراجی	27	29	28	14	28	18	11	28	28	22
ویژگی‌های نهایی	0.77	0.83	0.8	0.4	0.8	0.51	0.31	0.8	0.8	0.63

همان‌طور که ملاحظه می‌شود، همواره پهنای تصویر به‌اندازه تعداد ویژگی‌های مورد نظر نیست؛ در صورتی که پهنای تصویر N پیکسل باشد و بخواهیم M ویژگی داشته باشیم، بدین صورت عمل می‌کنیم که ابتدا M عدد حقیقی ای تولید

(جدول-۶): نسبت پهنا به ارتفاع با علائم و بدون علائم چند زیرواژه

(Table-6): Width-to-Height ratio with and without symbols of few sample subwords

زیرواژه	'پ'	'چه'	'فقی'	'بتیس'	'صمغی'	'بخشیم'	'اعتصوم'	'نهفتن'	'گلسنگها'
نسبت پهنا به ارتفاع	1.15	1.46	1.25	2.25	2.53	2.12	3.45	2.02	2.73
بدون علائم	2.25	2.48	1.68	2.88	3.38	2.87	3.67	2.7	3.02



(شکل-۴): نسبت پهنا به ارتفاع با علائم و بدون علائم یک زیرواژه

(Figure-4): Width-to-Height ratio with and without symbols of a subword

بیفتند) به عنوان ورودی در نظر گرفته و مراحل کار تشریح شده است.

الف- زیرواژه قشنگ: با عنایت به کد موقعیت علائم این زیرواژه (۱۱۱۱) و جدول (۳)، در مرحله گروه‌بندی فضای جستجو برای این زیرواژه به ۱۲۵ زیرواژه محدود می‌شود. با اعمال محدودیت پهنا به ارتفاع فضای جستجو به ۱۰۷ زیرواژه کاهش می‌یابد؛ لذا با توجه به روال تعریف‌شده خوشه‌بندی روی داده‌ها صورت گرفته (سیزده خوشه) و تعدادی از خوشه‌های نزدیک‌تر (یک ششم) به عنوان فضای منتخب نهایی انتخاب می‌شوند در نتیجه این مراحل فضای جستجوی نهایی برای این نمونه به سیزده زیرواژه محدود شده که در شکل (۵-الف) نشان داده شده است.

ب- زیرواژه تعمیق: با عنایت به کد موقعیت علائم این زیرواژه (۱۲۱) و جدول (۳) در مرحله گروه‌بندی فضای جستجو برای این زیرواژه به ۶۶۷ زیرواژه محدود می‌شود. با اعمال محدودیت پهنا به ارتفاع فضای جستجو به ۴۱۵ زیرواژه کاهش می‌یابد؛ لذا با توجه به روال تعریف‌شده خوشه‌بندی روی داده‌ها صورت گرفته (۶۷ خوشه) و تعدادی از خوشه‌های نزدیک‌تر (یک ششم) به عنوان فضای منتخب نهایی انتخاب می‌شوند در نتیجه این مراحل فضای جستجوی نهایی برای این نمونه به ۷۸ زیرواژه محدود شده که در شکل (۵-ب) نشان داده شده است.

ج- زیرواژه شصتمین: با عنایت به کد موقعیت علائم این زیرواژه (۱۱۲۱) و جدول (۳) در مرحله گروه‌بندی فضای جستجو به ۱۷۲ زیرواژه محدود می‌شود.

۳-۴- گروه‌بندی و خوشه‌بندی

یکی از مراحل که در این پژوهش به منظور کاهش فضای جستجو به کار گرفته شده گروه‌بندی است. در مرحله گروه‌بندی داده‌های آموزشی بر اساس کد موقعیت علائم آنها گروه‌بندی می‌شوند (زیرواژگانی که کد موقعیت علائم آنها یکسان است در یک گروه قرار می‌گیرند). در جدول (۳) کدهای موقعیت علائم و فراوانی آنها در زیرواژگان پایگاه داده ارائه شد. همان‌گونه که در این جدول دیده می‌شود در کل زیرواژگان پایگاه داده را از نظر کد موقعیت علائم می‌توان به ۸۶ گروه تقسیم کرد که تعداد زیادی از این گروه‌ها تنها یک عضو دارند.

گروه‌هایی که بیش از ده عضو دارند، با استفاده از الگوریتم خوشه‌بندی k میانگین با معیار فاصله اقلیدسی و ویژگی‌های استخراج‌شده در مرحله قبل به خوشه‌هایی تقسیم می‌شوند. تعداد خوشه‌ها با توجه به تعداد اعضای هر گروه اولیه (مجموعه‌ای که کد موقعیت علائم یکسان دارند)، یک خوشه به‌ازای هر ده عضو لحاظ شده و از متوسط ویژگی‌های چهار قلم زر، لوتوس، نازنین و میترا به عنوان ویژگی‌های ورودی (مرحله آموزش) و ویژگی‌های ۱۲۷۰۰ زیرواژه با قلم یا قوت به عنوان ویژگی‌های آزمایش استفاده شده است.

به منظور حصول نتیجه مورد نظر در خصوص کاهش فضای جستجو و در عین حال خطای کمتر در بازشناسی (وجود زیرواژه مورد نظر در جمع زیرواژه‌های پیشنهادی) تعدادی از نزدیک‌ترین خوشه‌ها به زیرواژه ورودی به عنوان فضای جستجو انتخاب می‌شوند. در ادامه در جهت بررسی شهودی روال کار، چهار زیرواژه (قشنگ، تعمیق، شصتمین،

د- زیرواژه بیفتند: با عنایت به کد موقعیت علائم این زیرواژه (۲۲۱۲۲) و جدول (۳) در مرحله گروه‌بندی فضای جستجو به هشت زیرواژه محدود می‌شود. با اعمال محدودیت پهنا به ارتفاع فضای جستجو به هفت زیرواژه کاهش می‌یابد؛ لذا با توجه به روال تعریف‌شده نیاز به خوشه‌بندی وجود ندارد و فضای جستجوی محدودشده نهایی در شکل (۵-د) نشان داده شده است.

بجنگند	تنسيق	نخستين	تجز	خيز	خيش	شيق	شيق	ضيق	نچش	تميز	تميز	تفنگ
بيگمنو	نشيم	تخميني	شير	تيمز	حشيش	حضيض	حنين	حنين	خيشو	خيني	فقنا	مننگ
بيختن	مستقيم	سنخي	منيز	منيز	سنجش	شيشو	شيشو	شيمز	شينو	شبخو	قشنگ	فتنشا
بيفتند	منخصيصي	ششمين	تحسين	تحسين	شينو	صفين	ضعين	قيخو	نچشو	نچشو	تفنگد	منفقد
چيفتن	منحصنين	پژوهشي	حشيشي	حشيشي	حقيقتي	نرضي	نچشو	نچفي	نچفي	نچفي	منفقد	فتنشا
پيشگفتا	شستمين	تضميني	سرخي	سرخي	شستو	ضعيفي	تحميق	تعميق	تعميق	تعميق	متفقد	فشنگ
بيفکنند	منخصيصين	محققين	شستو	شستو	نميگو	نميشو	نيستر	خميتر	خميتر	خميتر	فتنقله	خفتند
	منخصيصي	تضمن	نيستو	نيستو	نميو	مفيسو	منچستر	قيمتم	قيمتم	نصحتي		قشگشا
	منخصيصي	تصنيفي	نچفي	نچفي	نچفي	سنيو	حسنين	متجق	متجق	شيخي		

(شکل-۵): فضای محدود شده برای چند زیرواژه منتخب
(Figure-5): Limited space for selected subwords

پایان یافته تلقی می‌شود در غیر این صورت در جهت انتخاب نزدیک‌ترین زیرواژه‌ها به زیرواژه ورودی در محدوده کد مورد نظر پس از استخراج ویژگی با مراکز خوشه‌های استخراج‌شده مقایسه شده و فاصله آن با مراکز خوشه‌ها مشخص شده و نزدیک‌ترین خوشه‌ها به زیرواژه ورودی به‌عنوان خوشه‌های هدف انتخاب می‌شوند. با توجه به تعداد متفاوت خوشه‌ها در کدهای مختلف برای سنجش تعداد مناسب خوشه‌ها نسبت خوشه‌های انتخابی به کل خوشه‌ها مورد استفاده قرار گرفته است.

جداول (۸) و (۹) تغییرات دقت و متوسط تعداد زیرواژه با تغییر نسبت خوشه‌های انتخابی به کل خوشه‌ها برای داده‌های آموزش و آزمایش را بدون محدودیت نسبت پهنا به ارتفاع نشان می‌دهند. همان‌طور که در جدول (۸) دیده می‌شود، بدون اعمال خوشه‌بندی یا به‌عبارت دیگر با لحاظ یک خوشه برای هر گروه متوسط تعداد عناصر گروه ۱۱۵۰ خواهد بود که به معنی کاهش ۹۱٪ فضای جستجو با حفظ دقت ۱۰۰٪ است که با انتخاب یک‌ششم از خوشه‌ها تعداد متوسط زیرواژگان برای داده‌های آزمایش به ۲۰۱ زیرواژه است که به مفهوم کاهش ۹۸/۴٪ فضای جستجو با دقت ۹۹٪ است. این تعداد برای داده‌های آموزش به چهارده زیرواژه (کاهش ۹۹/۹٪ فضای جستجو) با دقت صددرصد کاهش یافته است.

با اعمال محدودیت پهنا به ارتفاع فضای جستجو برای این نمونه به ۴۵ زیرواژه کاهش می‌یابد؛ لذا با توجه به روال تعریف‌شده خوشه‌بندی روی داده‌ها صورت گرفته (۱۷ خوشه) و تعدادی از خوشه‌های نزدیک‌تر (یک‌ششم) به‌عنوان فضای منتخب نهایی انتخاب می‌شوند در نتیجه این مراحل فضای جستجوی نهایی برای این نمونه به هفده زیرواژه محدود شده که در شکل (۵-ج) نشان داده شده است.

۴- آزمایش‌ها و تحلیل نتایج

در این مقاله در یک راه‌یافت بازنمایی زیرواژگان فارسی براساس شکل کلی زیرواژه، ارائه یک روش ساده و در عین حال کارآمد در جهت محدود کردن فضای جستجوی زیرواژگان مد نظر قرار گرفته است. در نخستین مرحله با توصیف قیدشده در مراحل قبل کد موقعیت نقاط و علائم داده‌های آموزشی استخراج و بر اساس این کدها داده‌های آموزشی به ۸۶ گروه تقسیم شده‌اند. با توجه به تعداد زیرواژگان هرگروه برای گروه‌هایی که تعداد زیرواژگان بیشتر از ده دارند خوشه‌بندی براساس ویژگی‌های استخراج‌شده توصیف‌شده مبتنی بر متوسط ویژگی‌ها برای چهار قلم زر، لوتوس، نازنین و میتر صورت گرفته است. تعداد خوشه‌ها در این مرحله یک خوشه به‌زای هر ده زیرواژه لحاظ شده که با توصیف بالا در مجموع برای ۳۹ کد خوشه‌بندی صورت گرفته است که اطلاعات مربوطه از جمله کدهایی که خوشه‌بندی شده‌اند و تعداد خوشه‌ها در هر کدام در جدول (۷) دیده می‌شود.

در گام آزمایش ابتدا کد موقعیت علائم زیرواژه ورودی استخراج گردیده و با توجه به آن گروه زیرواژگان مرتبط با آن انتخاب می‌شود؛ سپس با توجه به نسبت پهنا به ارتفاع محدوده زیرواژگان مورد نظر محدودتر می‌شود. در این حالت اگر تعداد زیرواژگان محدودشده کمتر از ده زیرواژه باشد، کار

(جدول-۷): تعداد خوشه‌ها و تعداد اعضا در کدهای مختلف

(Table-7): The number of clusters and the number of members in different codes

1211	1122	1121	1112	1111	222	221	212	211	122	121	112	111	22	21	12	11	2	1	0	کد موقعیت
118	48	172	48	125	120	379	230	482	212	667	379	627	712	1416	1166	1831	1387	2139	856	تعداد
12	5	18	5	13	12	38	23	49	22	67	38	63	72	142	117	184	139	214	86	تعداد خوشه
22112	21211	21121	22121	12111	11221	11211	11121	11111	2221	2212	2211	2122	1212	2121	2112	2111	1222	1221	1212	کد موقعیت
11	11	13	13	16	19	18	26	16	28	36	86	17	24	77	39	88	20	86	24	تعداد
2	2	2	2	2	2	2	3	2	3	4	9	2	3	8	4	9	2	9	3	تعداد خوشه

(جدول-۸): تغییرات متوسط تعداد زیرواژه با تغییر نسبت خوشه‌های انتخابی به کل خوشه‌ها برای داده‌های آموزش بدون محدودیت

نسبت پهنا به ارتفاع

(Table-8): Mean variation of number of subword by changing the ratio of selected clusters to whole clusters for training data without the width-to-height ratio limiter

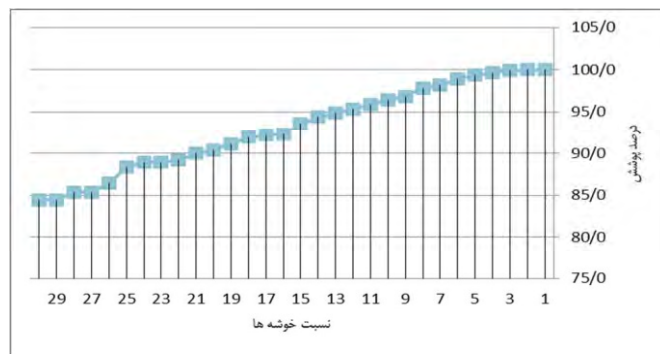
15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	نسبت تعداد خوشه‌ها (۱ به ...)
86	92	99	107	116	125	138	156	175	203	239	295	385	583	1150	تعداد زیرواژه

(جدول-۹): تغییرات دقت و متوسط تعداد زیرواژه با تغییر نسبت خوشه‌های انتخابی به کل خوشه‌ها برای داده‌های آزمایش بدون محدودیت

نسبت پهنا به ارتفاع

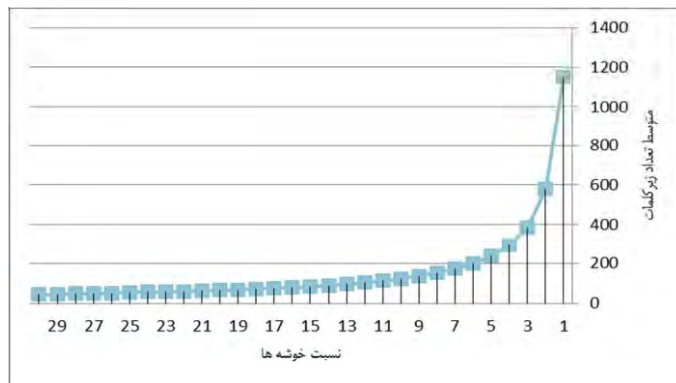
(Table-9): Coverage percentage and mean subword number variation with change in the ratio of selected clusters to total clusters for test data without limitation of width to height ratio

10	9	8	7	6	5	4	3	2	1	نسبت تعداد خوشه‌ها (۱ به ...)
96.4	96.8	97.8	98.2	99	99.4	99.7	99.9	100	100	دقت
123	135	153	173	201	236	294	384	579	1150	تعداد زیرواژه



(شکل-۶): چگونگی تغییرات دقت با تغییرات نسبت خوشه‌ها در داده‌ها آزمایش (بدون لحاظ محدودیت نسبت)

(Figure-6): Coverage percentage changes with cluster ratio changes in test data (without regard to the limitation of the ratio)



(شکل-۷): چگونگی تغییرات متوسط تعداد زیرواژگان محدود شده با تغییرات نسبت خوشه‌ها در داده‌های آزمایش

(بدون لحاظ محدودیت نسبت)

(Figure-7): Moderate changes in the number of subwords limited by changes in the ratio of clusters in the test data (without regard to the limitation of the ratio)

تعداد متوسط زیرواژگان برای داده‌های آزمایش به ۱۲۹ زیرواژه (۹۹٪ کاهش فضای جستجو) با دقت ۹۸/۹٪ کاهش یافته که نشان‌دهنده کاهش ۳۶٪ فضای جستجو نسبت به حالت قبل است. این تعداد برای داده‌های آموزش به ده زیرواژه (۹۹/۹٪ کاهش فضای جستجو) با دقت صد درصد کاهش یافته است. شکل‌های (۸) و (۹) چگونگی تغییرات دقت و متوسط تعداد زیرواژگان را با تغییر نسبت خوشه‌ها نشان می‌دهند.

جدول (۱۰) و (۱۱) تغییرات دقت و متوسط تعداد زیرواژه را با تغییر نسبت خوشه‌های انتخابی به کل خوشه‌ها برای داده‌های آموزش و آزمایش با اعمال محدودیت نسبت به ارتفاع نشان می‌دهند. همان‌طور که در جدول دیده می‌شود، بدون اعمال خوشه‌بندی یا به عبارت دیگر با لحاظ یک خوشه برای هر کد متوسط، تعداد عناصر خوشه در این شرایط به ۶۱۶ زیرواژه (۹۵٪ کاهش فضای جستجو) کاهش یافته است که با انتخاب یک‌ششم از خوشه‌های مرحله دوم

(جدول-۱۰): تغییرات دقت و متوسط تعداد زیرواژه با تغییر نسبت خوشه‌های انتخابی به کل خوشه‌ها برای داده‌های آموزش

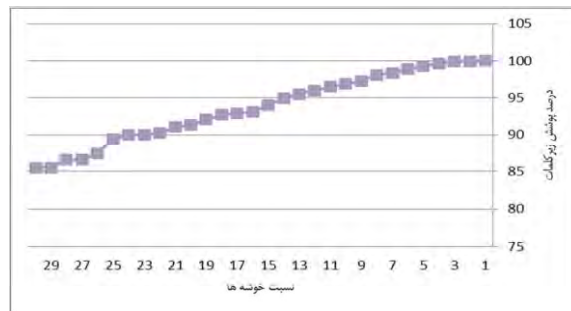
(Table-10): Coverage percentage and mean number of subword variation by changing the ratio of selected clusters to total clusters for training data

نسبت تعداد خوشه ها (۱ به ...)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
دقت	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
تعداد زیرواژه	616	329	230	182	151	130	114	102	91	84	77	72	67	63	59

(جدول-۱۱): تغییرات دقت و متوسط تعداد زیرواژه با تغییر نسبت خوشه‌های انتخابی به کل خوشه‌ها برای داده‌های آزمایش

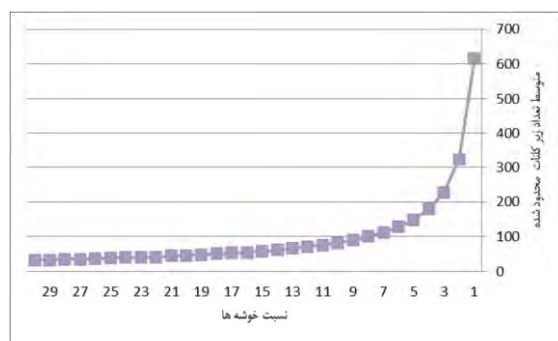
(Table-11): Coverage percentage and mean number of subword variation by changing the ratio of selected clusters to total clusters for test data

نسبت تعداد خوشه ها (۱ به ...)	1	2	3	4	5	6	7	8	9	10
دقت	100	100	99.9	99.6	99.3	98.9	98.4	98.1	97.2	96.8
تعداد زیرواژه	615	324	228	180	149	129	112	100	90	82



(شکل-۸): چگونگی تغییرات دقت با تغییرات نسبت خوشه‌ها در داده‌ها آزمایش

(Figure-8): Coverage percentage variation by changing the ratio of selected clusters for test data



(شکل-۹): چگونگی تغییرات متوسط تعداد زیرواژگان محدود شده با تغییرات نسبت خوشه‌ها در داده‌های آزمایش

(Figure-9): mean number of subwords variation by changing the ratio of selected clusters to total clusters for test data

۵- نتیجه گیری

با توجه به ساختار پیوسته نوشته‌های فارسی و ایرادهای مطرح شده برای روش‌های بازشناسی مبتنی بر شکستن واژه به حروف، کارایی روش‌های مبتنی بر شکل کلی قابل توجه است. درخصوص ایرادها و مشکلات این روش، تعداد زیاد رده‌ها که معادل تعداد زیرواژگان است از گلوگاه‌های اصلی به‌شمار می‌آید. نخستین مرحله برای کاهش فضای جستجو در این رویکرد، تقسیم فضا به بخش‌های مختلف به‌کمک خوشه‌بندی است.

در این مقاله با توجه به هدف که طراحی روشی ساده و کارا جهت کاهش دامنه جستجو است از گروه‌بندی و خوشه‌بندی به‌صورت هم‌زمان استفاده شده است. در نخستین مرحله پایگاه داده به ۸۶ گروه بر اساس کد موقعیت علائم زیرواژگان تقسیم می‌شود. با انتخاب گروه مرتبط با زیرواژه ورودی فضای جستجو به اعضای آن گروه محدود می‌شود. برای ایجاد محدودیت بیشتر در گروه انتخاب شده، تنها اعضای که از نظر نسبت پهنا به ارتفاع در محدوده‌ای از نسبت پهنا به ارتفاع زیرواژه ورودی باشند، انتخاب می‌شوند. در صورتی که محدودیت‌های ایجاد شده به فضایی بیش از ده زیرواژه منتهی شود (برای خوشه‌هایی که تعداد اعضای بیشتری دارند) یک مرحله خوشه‌بندی براساس ویژگی‌های ساده پروفایل‌های افقی و عمودی صورت گرفته و تنها تعدادی از خوشه‌های نزدیک به زیرواژه ورودی با رعایت حدود نسبت پهنا به ارتفاع، محدوده جدید را مشخص می‌کنند. برای مرحله خوشه‌بندی ویژگی‌های ساده‌ای از پروفایل‌های افقی و عمودی استخراج شده است. برای استخراج ویژگی نزدیک‌ترین فاصله پیکسل‌های تصویر بدنه زیرواژه از لبه‌های چهارگانه کادر به‌عنوان ویژگی انتخاب شده و با عنایت به طول متغیر بدنه زیرواژه‌ها و متغیر شدن طول ویژگی، برای هر زیرواژه از هر جهت تنها ده ویژگی به روش درون‌یابی انتخاب شده است و در مجموع برای هر زیرواژه چهل ویژگی خواهیم داشت. از متوسط ویژگی‌های ۱۲۷۰۰ زیرواژه در چهار قلم زر، لوتوس، نازنین و میترا به‌عنوان ویژگی‌ها ورودی (مرحله آموزش) و ویژگی‌های ۱۲۷۰۰ زیرواژه با قلم یاقوت به‌عنوان ویژگی‌های آزمایش استفاده شده است. کاهش نهایی فضای جستجو از ۱۲۷۰۰ به ۱۲۹ (۹۹/۱٪ کاهش) گویای عملکرد قابل توجه روش پیشنهادی است.

۴-۱- مقایسه نتایج با کارهای گذشته

در روش ارائه شده در این مقاله در مقایسه با روش‌های قبلی از جمله روش ارائه شده در [4] و [22] و [25] علاوه بر سادگی و نتایج به‌نسبه بهتر، سرعت بهتری نیز حاصل شده است که این نتایج در جدول (۱۲) ارائه شده است. همان‌طور که در قبل نیز بیان شد، در [4] پس از خوشه‌بندی به سبب خوشه ده خوشه نخست به‌عنوان فضای کاهش یافته معرفی شده و در روش ارائه شده در [22] داوودی موفق شده است با یک مرحله بهبود فضای جستجو را با حفظ دقت ۹۹/۱۷٪ به ۴/۸ خوشه از سبب خوشه کاهش دهد و در روش معرفی شده در [25] کاهش فضای ویژگی در چند مرحله به ۹۹/۲٪ رسیده است. روش ارائه شده در این پژوهش در مقایسه با سه روش یادشده، در عین سادگی با همان دقت، فضای جستجو را تا ۹۹/۱٪ کاهش داده است که از دو روش نخست نتیجه به‌مراتب بهتری حاصل شده و در مقایسه با روش سوم اگرچه نتیجه بهبودی را نشان نمی‌دهد، اما با توجه به دومرحله‌ای شدن گروه‌بندی و خوشه‌بندی از سرعت بیشتری برخوردار است که مزیت این روش نسبت به روش سوم است.

(جدول-۱۲): مقایسه نتایج

(Table-12): Comparing results

درصد کاهش دامنه جستجو	روش
۹۶/۶٪	خوشه‌بندی با ویژگی‌های سراسری (ابراهیمی [۴])
۹۸/۴٪	خوشه‌بندی با ویژگی‌های سراسری و نواحی شاخص (داوودی [۲۲])
۹۹/۲٪	خوشه‌بندی با متوسط ۴ قلم و نسب پهنا به ارتفاع و موقعیت علائم [۲۵]
۹۹/۱٪	روش ارائه شده در این مقاله

جهت مقایسه سرعت بازشناسی روش ارائه شده در [25] و روش پیشنهادی این پژوهش که به‌عنوان نقطه قوت این روش بیان شده است، در شرایط یکسان با استفاده از رایانه‌ای با مشخصات (CPU: Core i7-3520M و RAM: 8GB) و با استفاده از نرم‌افزار متلب، متوسط زمان صرف شده برای محدودسازی فضای جستجو برای یک زیرواژه اندازه‌گیری شده که این زمان برای روش نخست زمان یازده میلی‌ثانیه و در روش دوم ۱۰/۵ میلی‌ثانیه گزارش شده است. درضمن این زمان برای زیرواژگانی که در روش پیشنهادی در مرحله نخست کاهش فضای جستجو گیر می‌افتند که شامل ۴۷ گروه از ۸۶ گروه هستند، به‌شدت کاهش خواهد یافت.

- [10] S. Lu and C. L. Tan, "Document image retrieval through word shape coding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 11, pp. 1913-1918, Nov. 2008.
- [11] J. A. Rodriguez-Serrano and F. Perronnin, "Handwritten word spotting using hidden markov models and vocabularies," *Pattern Recognition*, vol. 42, no. 9, pp. 2106-2116, Sep. 2009.
- [12] T. M. Rath and R. Manmatha, "Word spotting for historical documents," *International Journal on Document Analysis and Recognition*, Vol. 9, no. 2-4, pp. 139-152, Apr. 2007.
- [13] Y. Lu and C. L. Tan, "Information retrieval in document image databases," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 16, no. 11, pp. 1398-1410, Nov. 2004.
- [14] A. Ebrahimi and E. Kabir, "A pictorial dictionary for printed farsi sub words," *Pattern Recognition Letters*, Vol. 29, no. 5, pp. 656-663, 2008.
- [15] A. Rehman and T. Saba, "Off-line cursive script recognition: current advances, comparisons and remaining problems," *Artificial Intelligence Review*, vol. 37, no. 4, pp. 261-288, 2012.
- [16] S. G. Madhvanath and V. Govindaraju, "The role of holistic paradigms in handwritten word recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 149-164, Feb. 2001.
- [17] L. M. Lorigo and V. Govindaraju, "Off-line arabic handwriting recognition: a survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 5, pp. 712-724, May 2008.
- [18] S. Mozaffari, K. Facz, V. Märgner and H. Elabed, "Two-stage lexicon reduction for offline Arabic handwritten word recognition," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 22, No. 07: pp. 1323-1341, November 2008.
- [19] H. Davoudi, M. Cheriet and E. Kabir, "Lexicon reduction of handwritten arabic subwords based on the prominent shape regions," *International Journal on Document Analysis and Recognition*, vol 19, Issue 2, pp 139-153, 2016.

[۲۰] برومند، سمیه و ایرانپور مبارکه، مجید، "بازشناسی واژگان دست‌نوشته با ویژگی‌های نوین و کاهش فرهنگ لغت"، مجله پردازش بینایی و تصویر، آماده چاپ، ۱۳۹۶.

6- References

۶- مراجع

- [1] T. Adamek, N. E. Connor, and A. F. Smeaton, "Word matching using single closed contours for indexing Handwritten Historical Documents," *International Journal of Document Analysis and Recognition*, vol. 9, no. 2-4, pp. 153-165, 2007.
- [2] J. R. Pinales, R. J. Rivas, and M. J. C. Blcda, "Holistic Cursive word recognition based on perceptual features," *Pattern Recognition Letters*, vol. 28, no. 13, pp. 1600-1609, 1 Oct. 2007.
- [3] A. Amin, "Recognition of printed arabic text based on global features and decision tree learning techniques," *Pattern Recognition*, vol. 33, no. 8, pp. 1309-1323, 2000.
- [۴] ابراهیمی، افشین، "استفاده از شکل کلی زیرکلمات چایی در بازیابی تصویر مستندات و بازشناسی متون فارسی"، رساله دکتری مهندسی برق- الکترونیک، دانشگاه تربیت مدرس، تهران، ۱۳۸۴.
- [4] A. Ebrahimi, "Using the holistic form of print subwords in retrieving documentary images and recognizing Persian texts", Ph.D. dissertation, Electron. Eng., Tarbiat Modares Univ., Tehran, 1384.
- [۵] خسروی، حسین و کبیر، احسان الله، "ارزیابی روش‌های بازشناسی متون فارسی بر مبنای شکل کلی زیرکلمات"، نشریه مهندسی برق و کامپیوتر ایران، جلد ۷، شماره ۴، صص. ۲۸۰-۲۶۷، ۱۳۸۸.
- [5] H. Khosravi, E. Kabir, " Evaluation of methods for recognizing Persian texts based on the holistic form of subwords," *Iranian Journal of Electrical and Computer Engineering*, vol.7, no.4, pp.267-280, 2005.
- [6] S. Madhvanath, G. Kim, and V. Govindaraju, "Chain code contour processing for handwritten word recognition," *IEEE Transactions on Pattern Recognition and Machine Intelligence*, vol. 21, no. 9, pp. 928-932, Sep. 1999.
- [7] K. Zagoris, K. Ergina, and N. Papamarkos, "A document image retrieval system," *Engineering Application of Artificial Intelligence*, vol. 23, no. 6, pp. 872-879, 2010.
- [8] S. Bai, L. Li, and C. L. Tan, "Keyword spotting in document images through word shape coding," in *Proc. 10th International Conference on Document Analysis and Recognition, ICDAR'09*, pp. 331-335, 26-29 Jul. 2009.
- [9] L. Li, S. Lu, and C. L. Tan, "A fast keyword-spotting technique," in *Proc. 9th Int. Conference*



اسماعیل میری در سال ۱۳۷۲ مدرک کارشناسی مهندسی برق - الکترونیک را از دانشگاه تهران اخذ کرد و سپس در سال‌های ۱۳۹۰ و ۱۳۹۶ در مقاطع کارشناسی ارشد و دکترای مهندسی برق -

الکترونیک از دانشگاه بیرجند دانش‌آموخته شد و هم‌اکنون در شرکت مخابرات ایران مشغول فعالیت است. زمینه‌های پژوهشی مورد علاقه ایشان عبارتند از: بازشناسی الگو و کاربردهای آن، پردازش تصویر.

نشانی رایانامه ایشان عبارت است از:

miri.esmail@birjand.ac.ir



سیدمحمد رضوی در سال ۱۳۷۳ مدرک کارشناسی مهندسی برق - الکترونیک را از دانشگاه صنعتی امیرکبیر (واحد تفرش) اخذ کرد و در سال‌های ۱۳۷۵ و ۱۳۸۵ در مقاطع کارشناسی ارشد و دکترای مهندسی

برق - الکترونیک از دانشگاه تربیت مدرس دانش‌آموخته شد. از سال ۱۳۷۶ به‌عنوان عضو هیأت علمی در دانشگاه بیرجند مشغول انجام وظیفه است. وی هم‌اکنون دانشیار دانشکده مهندسی برق و کامپیوتر دانشگاه بیرجند است. زمینه‌های پژوهشی مورد علاقه ایشان عبارتند از: بازشناسی الگو، پردازش تصویر و سامانه‌های هوشمند.

نشانی رایانامه ایشان عبارت است از:

smrazavi@birjand.ac.ir



ناصر مهرشاد در سال ۱۳۷۳ مدرک کارشناسی خود را از دانشگاه فردوسی مشهد اخذ کرد و در سال‌های ۱۳۷۶ و ۱۳۸۲ در مقاطع کارشناسی ارشد و دکترای

در دانشگاه تربیت مدرس گرایش مهندسی پزشکی دانش‌آموخته شد. وی هم‌اکنون عضو هیأت علمی گروه مهندسی الکترونیک دانشگاه بیرجند بوده و به‌عنوان دانشیار مشغول فعالیت است. زمینه‌های پژوهشی مورد علاقه ایشان شامل ماشین بینایی، پردازش سیگنال دیجیتال و اطلاعات مهندسی پزشکی است.

نشانی رایانامه ایشان عبارت است از:

nmehrshad@birjand.ac.ir

[20] S. Bromand, M. Iranpurmobaraka, "Handwritten words recognition with new features and reducing the dictionary," *Machine Vision And Image Processing*, unpublished.

[21] H. Davoudi, E. Kabir, "Using compatible shape descriptor for lexicon reduction of printed farsi subwords," *International Journal on Document Analysis and Recognition*, vol. 19, Issue 2. pp 139-153, 2016.

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

[22] H. Davoudi, E. Kabir, "Using compatible shape descriptor for lexicon reduction of printed farsi subwords," *Iranian Journal of Electrical and Computer Engineering*, vol. 12, Issue 1., 2014.

[۲۳] فتحی، فائقه، استخراج حروف شاخص از زیرواژگان چاپی فارسی، پایان‌نامه کارشناسی ارشد، دانشگاه صنعتی سهند، تبریز، ایران، ۱۳۸۸.

[23] F. Fathi, "Extraction of index letters from Persian printed subwords", M.S. thesis, Dept. Electron.Eng., Sahand University of Technology, Tabriz, Iran, 2009.

[۲۴] علی‌بیگی، محمد، بازشناسی زیرواژگان تایی فارسی، پایان‌نامه کارشناسی ارشد، دانشگاه بیرجند، بیرجند، ایران، ۱۳۸۹.

[24] M. Alibaigi, "Persian printed subwords recognition", M.Sc. thesis, Department of Electronic Engineering, University of Birjand, Birjand, Iran, 2010.

[۲۵] میری، اسماعیل، رضوی، سید محمد و مهرشاد، ناصر، "روشی ساده برای کاهش فضای جستجو در بازشناسی زیرواژگان تایی فارسی"، نهمین کنفرانس ماشین بینایی و پردازش تصویر ایران، دانشگاه شهید بهشتی، آبان ماه ۱۳۹۴.

[25] E. Miri, S.M. Razavi, N. Mehrshad, "A simple method for search space reduction in Persian typed subwords recognition," *9th Conference on Machine Vision and Image Processing conference*, Shahid Behshti University, Tehran, 2015.

