خوشه‌بندی تصاویر زیر- کلمات چاپی فارسی با استفاده از ویژگیهای مکان مشخصه و الگوریتم k- میانگین

افشین ابراهیمی استادیار دانشکده مهندسی برق، دانشگاه صنعتی سهند
احسان‌اله کبیر دانشیار گروه مهندسی الکترونیک، دانشگاه تربیت مدرس

چکیده

در این مقاله از ویژگیهای مکان مشخصه برای توصیف شکل کلی زیر- کلمات چاپی فارسی استفاده شده است. در محاسبه این ویژگیها تعداد برخوردها با بدنه زیر- کلمه به ۲ محدود شده است. از این ویژگیها، با روش PCA، ۱۲ ویژگی ناهمبسته انتخاب شده‌اند. از روش k- میانگین با معیار فاصله اقلیدسی برای خوشه‌بندی تصاویر زیر- کلمات استفاده شده است. تصاویر ۹۴۴۵ زیر- کلمه با قلم لوتوس ۱۲ و درجه تفکیک ۴۰۰ نقطه در اینچ، به ۱۵۰ و ۳۰۰ خوشه تقسیم شدند. مقادیر کمترین و بیشترین تعداد نمونه‌های خوشه‌ها در خوشه‌بندی به ۱۵۰ خوشه به ترتیب ۱۱ و ۹۱ زیر- کلمه و در خوشه‌بندی به ۳۰۰ خوشه به ترتیب ۲ و ۵۸ زیر- کلمه به‌دست آمد. در یک آزمایش برای ارزیابی خوشه‌بندی، تصاویر ۲۰۰ زیر- کلمه که دوباره روبش شدند، به ۳۰۰ خوشه طبقه‌بندی شدند. در این طبقه‌بندی از معیار فاصله اقلیدسی از میانگین خوشه‌ها استفاده شد. در انتخاب اول، پنج انتخاب اول و ده انتخاب اول به ترتیب ۸۰/۶۹٪، ۹۷/۵۲٪ و ۱۰۰٪ از این زیر- کلمات به درستی طبقه‌بندی شدند.

کلمات کلیدی: زیر- کلمه، ویژگی‌های مکان مشخصه، خوشه‌بندی، k- میانگین، PCA، طبقه‌بندی، فاصله اقلیدسی.

Clustering of printed Farsi subwords using characteristic loci features and k-means algorithm

A. Ebrahimi Faculty of Electrical Eng., Sahand University of Technology
E. Kabir Dept. of Electrical Eng., Tarbiat Modares University

Abstract

In this paper the characteristic loci features are used for the description of printed Farsi subwords. In extraction of these features, the number of crossings with subword bodies are restricted to 2. Using PCA, 12 uncorrelated features are selected. The images of subwords are clusterd using k-means algorithm with Euclidian distance. 9445 subwords of Lotus 12 font, with 400dpi resolution, are clustered to 150 and 300 clusters. Minimum and maximum cluster sizes, are 11 and 91 subwords, respectively, for 150 clusters and 2 and 58 subwords, for 300 clusters. In a test, for clustering verification, images of 200 subwords, were rescanned and classified to 300 clusters. In this classification, the Euclidian distance from cluster means is used. In first, first five and first ten choices, 80.69%, 97.52% and 100% of these subwords were correctly classified.

Key words: Subword, Characteristic loci features, Clustering, k-means, PCA, Classification, Euclidian distance.

۱-مقدمه

اولین تحقیقات در زمینه بازشناسی متون مربوط به سالهای ۱۹۲۹ و ۱۹۳۳ میلادی است [۱۵]. این سیستمها حروف چاپی را با روش تطبیق کلیشه‌ای شناسایی می‌کردند و به دلیل استفاده از تکنولوژی اپتومکانیکی کاربردی نبودند. تصور دسترسی به دستگاهی برای بازشناسی حروف تا دهه ۱۹۵۰ میلادی و ظهور کامپیوترهای رقمی به‌صورت یک رؤیا باقی ماند. از اواسط این دهه OCR به‌صورت یک زمینه فعال برای تحقیق درآمد. امروزه نرم افزارهای نسبتا خوبی را در این زمینه می‌توان با قیمتی کمتر از ۱۰۰ دلار خرید. البته این نرم افزارها تنها قادر به بازشناسی مستندات چاپی با کیفیت مناسب و حروف مجزای دستنویس هستند. تحقیقات فعلی در زمینه OCR مربوط به متونی است که با سیستمهای موجود قابل بازشناسی نیستند. نمونه‌هایی از این متون عبارتند از: متون چاپی با قلمهای گوناگون[۱] یا با کیفیت تصویری پایین و متون دستنویسی که بدون محدودیت خاصی[۲] نوشته شده باشند. چون در مقایسه با سیستمهای موجود هنوز یک تایپیست ماهر خطای کمتری انجام می‌دهد، برای بازشناسی متون چاپی با کیفیت تصویری بالا نیز تحقیقاتی برای کاهش میزان خطا و وازدگی انجام شده است [۲۲].

تحقیقات در زمینه بازشناسی متون چاپی فارسی و عربی نیز از سال ۱۹۸۰ شروع شده است [۱ و ۲]. روشهای بازشناسی متون از دو رویکرد مبتنی بر جداسازی کلمات به حروف و زیر-حروف و رویکرد مبتنی بر شکل کلی کلمات، استفاده می‌کنند. در بازشناسی متون فارسی و عربی مبتنی بر جداسازی، علاوه بر مشکلاتی مانند وجود نقاط و علائم و تنوع قلمها، مشکل جداسازی حروف نیز وجود دارد. مناسب نبودن کیفیت سند، پایین بودن درجه تفکیک روبش آن یا کجی تصویر سند نیز مشکلاتی هستند که به سختی کار می‌افزایند. در رویکرد مبتنی بر شکل کلی کلمات یا زیر- کلمات می‌توان از روشهای مختلف توصیف شکل استفاده کرد. در زمینه بازشناسی متون چاپی مبتنی بر بازشناسی بدون جداسازی کارهای مختلفی انجام شده است.

کارهای انجام شده در زمینه بازشناسی و بازیابی متون انگلیسی به دو گروه متون چاپی و دستنویس تقسیم می‌شوند.

در زمینه متون چاپی، کلمات با کد کردن شکل کلی آنها توصیف می‌شوند. بدین ترتیب که شکل یک کلمه را با ویژگیهای ناحیه‌ای مانند جای نوار زمینه، داشتن یا نداشتن بالا رونده و پایین رونده، تعداد آنها، جای تقریبی آنها در کلمه، داشتن یا نداشتن حفره و نقطه مدل می‌کنند [۱۰، ۱۱، ۱۲، ۱۳، ۱۴، ۱۶، ۱۹ و ۲۱].

در زمینه متون دستنویس، این ویژگیهای ناحیه‌ای با استفاده از روش ناحیه‌بندی[۳] یا هیستوگرامهای افقی و عمودی تعداد نقاط سیاه استخراج می‌شوند [۱۸ و ۲۳].

در زمینه بازشناسی متون فارسی و عربی نیز کارهای مختلفی انجام شده است. در یک تحقیق الگوریتمی برای خوشه‌بندی زیر- کلمات چاپی و ساختن سه نوع دیکشنری تصویری برای کمک به بازشناسی آنها ارائه شده است [۲۵]. در مرحله ایجاد دیکشنری اول، کانتور بالایی تصویر هر زیر-کلمه استخراج می‌شود و به کمک مجموعه‌ای از قواعد به پاره مسیرهایی برچسب خورده تبدیل می‌شود. با استفاده از این برچسبها به هر زیر- کلمه اندیسی نسبت داده می‌شود که جایگاه آن را در دیکشنری تصویری مشخص می‌کند.

دیکشنری دوم با استفاده از توصیفگرهای فوریه کانتور زیر- کلمات ایجاد شده است. از فاز توصیفگرهای فوریه کانتور زیر- کلمات برای به‌دست آوردن اندیس آنها استفاده شده است.

دیکشنری سوم نیز با استفاده از ویژگیهای مکان مشخصه زیر- کلمات ساخته شده است. پس از محاسبه ویژگیهای مکان مشخصه یک زیر- کلمه، ۱۰ مولفه بزرگتر به‌عنوان اندیس آن در دیکشنری در نظر گرفته شده است.

زیر- کلمات مختلفی که اندیس یکسانی دارند، همسایگی خاص خود را تشکیل می‌دهند. در مرحله طبقه‌بندی، مجموعه زیر کلمات موجود در دیکشنری که اندیس آنها با اندیس کلمه ورودی یکسان است، مشخص می‌شود.

در مرحله بعدی از حروف شاخص زیر- کلمات، مانند "ا"، "ک" و "ل"، که بازشناسی آنها آسان است، استفاده می‌شود. با یافتن حروف شاخص کلمه ورودی، محدوده جستجو در بین کلمات همسایه در دیکشنری تصویری کاهش می‌یابد.

1- Omnifont
2- Unconstrained Handwritten

3- Zoning

برای ایجاد دیکشنری از ۲۷۷۲ تصویر بدنه زیر- کلمات ۲، ۳ و ۴ حرفی فارسی شامل ۶۹۳۰۰ نمونه از پنج نوع قلم در پنج اندازه مختلف استفاده شده است. متوسط اندازه همسایگی‌های تصویری ۷۴/۳۷ زیر- کلمه و متوسط اندازه دسته‌های ایجاد شده با توجه به حروف شاخص ۴/۳ زیر- کلمه بوده است. میزان دسته‌بندی درست نمونه‌ها، ۹۸/۶۱٪ گزارش شده است. ذکر این نکته لازم است که نمونه‌های زیر- کلمات در محیط کامپیوتری ساخته و در همان محیط بازشناسی شده‌اند. بنابراین با اجتناب از مراحل چاپ و روبش نمونه‌ها، هیچگونه نویز یا اعوجاجی در تصاویر وجود ندارد.

در تحقیق دیگری از ویژگیهای شکل کلمات چاپی در بازشناسی متون عربی، نوشته شده با سه قلم متداول، استفاده شده است [۵]. ویژگیهایی مانند نقاط، همزه، پاره‌خطهای جهتی، نقاط انتهایی و اتصالها، منحنی‌های جهتی، حفره‌ها، پایین روندها و فواصل درون کلمه‌ای، از تصویر کلمات چاپی عربی استخراج و در یک دیکشنری ذخیره می‌شوند. در این تحقیق از یک دیکشنری عربی شامل ۴۸۲۰۰ کلمه استفاده شده است. برای تصویر کلمه ورودی بردارهای ویژگی استخراج می‌شوند و با لغات دیکشنری مقایسه می‌شوند. رتبه‌بندی این لغات بر حسب میزان شباهت به کلمه ورودی انجام می‌شود. تصاویر متون استفاده شده در این تحقیق با درجه تفکیک ۳۰۰ نقطه در اینچ روبش شده‌اند. با آزمایش این روش بر روی تصویر ۸۴۳۶ کلمه چاپی عربی، نرخ بازشناسی صحیح ۶۵٪ گزارش شده است.

از تبدیل فوریه دوبعدی شکل کلمات نیز برای بازشناسی متون چاپی عربی با چهار قلم متداول استفاده شده است [۹]. تصویر کلمه به یک تصویر قطبی نرمالیزه شده تبدیل می‌شود، سپس به این تصویر تبدیل فوریه دوبعدی اعمال می‌شود. طیف حاصل نسبت به تغییرات اندازه، چرخش و جابه‌جایی مقاوم است. از مجموعه‌ای از ضرایب فوریه، برای نمایش تصویر هر کلمه استفاده شده است. بازشناسی با محاسبه کمترین فاصله اقلیدسی نرمالیزه از هر یک از لغات دیکشنری انجام می‌شود.

در بازشناسی متون چاپی مبتنی بر شکل کلی کلمات می‌توان از روشهای متداول در بازشناسی متون دستنویس ایده گرفت. از مدل مخفی مارکف برای بازشناسی اسامی دستنویس شهرهای ایران استفاده شده است [۳]. ویژگیهای استفاده شده در این تحقیق از اطلاعات کانتور کلمات استخراج شده‌اند. کدهای زنجیره‌ای جهتی کانتور کلمه دستنویس محاسبه می‌شود. این تصویر در راستای افقی به پنج قسمت مساوی و در جهت عمودی به قسمتهایی که ۵۰٪ با هم همپوشانی دارند تقسیم می‌شود. عرض تقسیمات عمودی دو برابر میانگین تکه‌های سیاه عمودی تصویر انتخاب می‌شود. برداری شامل هیستوگرام‌های جهتی کدهای زنجیره‌ای در هر یک از این پنجره‌ها به‌عنوان مدل کلمه انتخاب می‌شود.

برای کم کردن تعداد مشاهده‌هایی که به مدل مخفی مارکف گسسته اعمال می‌شود، فضای ویژگی با استفاده از یک شبکه عصبی خود سامانده کوهنن (SOM)[1] ، کوآنتیزه شده است. برای هر اسم شهر یک HMM[2] گسسته با الگوریتم Baum Welch به‌طور مجزا آموزش داده می‌شود. برای هر تصویر ورودی اسامی شهرها بر حسب میزان شباهت مدل آنها به کلمه ورودی مرتب می‌شوند.

با آزمایش این روش بر روی یک مجموعه کلمات دستنویس شامل اسم ۱۹۸ شهر، درصد بازشناسی صحیح ۶۵٪ در انتخاب اول و ۹۵٪ در ۲۰ انتخاب اول گزارش شده است. مجموعه تمرین شامل ۱۷۰۰۰ نام شهر است که به‌وسیله افراد مختلف نوشته شده‌اند.

در تحقیق دیگری از خوشه بندی FCM[3] و مدل مخفی مارکف برای بازشناسی اسامی دستنویس شهرهای فارسی استفاده شده است [۴]. مراحل پیش پردازش و استخراج ویژگی و مدل مخفی مارکف مانند مرجع [۳] است. خوشه‌بندی اولیه با استفاده از روش FCM انجام شده است. با آزمایش این روش بر روی مجموعه‌ای شامل ۱۹۸ اسم دستنویس شهرها درصد بازشناسی صحیح ۶۷/۲٪ در انتخاب اول و ۹۶/۵٪ در ۲۰ انتخاب اول گزارش شده است. مجموعه تمرین شامل ۱۷۰۰۰ نام شهر است که به‌وسیله افراد مختلف نوشته شده‌اند.

از الگوریتم DTW[4] دوبعدی نیز برای بازشناسی اسامی دستنویس ۵۰۰ شهر ایران استفاده شده است [۲۶]. مجموعه تمرین شامل ۲۰ نمونه برای هر شهر است که افراد مختلفی آنها را نوشته‌اند. از اطلاعات کانتور کلمات برای بازشناسی آنها استفاده شده است. کانتور زیر- کلمه با منحنی خطی تکه‌ای تقریب زده می‌شود. طول و زاویه هر یک از این پاره‌خطها نسبت به راستای افقی در بردار ویژگی زیر- کلمه ذخیره می‌شود. برای بازشناسی تصویر ورودی، میزان شباهت زیر- کلمات آن با زیر- کلمات

1- Self organizing map
2- Hidden Markov model
3- Fuzzy c-means
4- Dynamic time warping

دیکشنری با روش DTW دوبعدی محاسبه می‌شود و با روش فازی K همسایه نزدیکتر، طبقه‌بندی می‌شود. برای هر زیر- کلمه ۵ انتخاب اول در نظر گرفته می‌شود. برای بازشناسی کلمه ورودی، اولین تطابق ترکیبهای مختلف این انتخابها با کلمات معتبر به‌عنوان کلمه بازشناسی شده معرفی می‌شود. در آزمایش این روش برای مجموعه اسامی ۱۰۰، ۳۰۰ و ۵۰۰ شهر ایران، نرخهای بازشناسی صحیح ۸۸، ۸۴ و ۷۵ درصد گزارش شده است.

در ادامه در بخش ۲، روش تهیه مجموعه تصاویر زیر- کلمات چاپی فارسی آمده است. در بخش ۳، نحوه استخراج ویژگی از زیر- کلمات و خوشه‌بندی آنها آمده است. در بخش ۴، روش طبقه‌بندی زیر- کلمات و در بخش ۵، بررسی نتایج خوشه‌بندی زیر- کلمات ارائه شده است. در بخش ۶، نتیجه‌گیری آمده است.

۲- تهیه مجموعه تصاویر زیر- کلمات چاپی فارسی

برای این کار از پایگاههای داده الکترونیکی دو روزنامه کیهان و همشهری، استفاده شده است [۲۴]. از مستندات این پایگاهها، متداول‌ترین کلمات، با تعداد تکرار بیشتر از ۳۰، و زیر- کلمات مربوط به آنها استخراج شده‌اند. برای ۲۹۷۳۹ کلمه متداول، تعداد زیر- کلمات ۱۲۷۰۰ است.

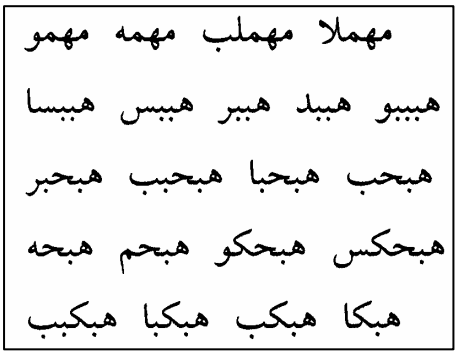
تعدادی از حروف الفبای فارسی مانند "ب"، "ت"، "ث" و "پ" بدنه مشابهی دارند و تفاوت آنها تنها در تعداد و جای قرار گرفتن نقاط است. در استخراج زیر- کلمات، حروف با بدنه یکسان با یکی از این حروف، نماینده گروه، جاگذاری شده‌اند. در جدول (۱)، حروف با بدنه یکسان ونماینده آنها آمده است.

جدول (۱) نمایندگان گروهها برای حروف الفبای فارسی با بدنه یکسان

نماینده گروه	ب	ب	ح	ع	ص	س	ک	د	ر	ط	ف
حروف با بدنه یکسان	ب پ ت ث	ب پ ت ث	ح چ خ	ع غ	ص ض	س ش	ک گ	د ذ	ر ز ژ	ط ظ	ف ق

بدین ترتیب بدون در نظر گرفتن نقاط زیر- کلمات، تعداد آنها به ۹۴۴۵ کاهش می‌یابد.

مجموعه زیر- کلمات بدون توجه به نقاط آنها با چهار قلم لوتوس، میترا، زر و یاقوت و سه اندازه قلم ۱۲، ۱۴ و ۱۶ با یک چاپگر لیزری HP1100 چاپ و سپس با یک روبشگر HP ScanJet 6300c با درجه تفکیک ۴۰۰dpi روبش شدند. در شکل (۱) نمونه‌ای از تصاویر زیر- کلمات آمده است. دلیل استفاده از این چهار قلم، متداول بودن این قلمها در چاپ مقاله، کتاب، روزنامه و نامه‌های اداری است.



شکل ۱- نمونه‌ای از تصاویر زیر- کلمات استخراج شده با قلم لوتوس

در مرحله پیش پردازش تصویر زیر- کلمات، حذف نقاط و تعیین چهارچوب محیطی انجام شده است. برای حذف نقاط به اجزا متصل تصویر زیر- کلمه برچسب زنی شده است. جزء متصل با تعداد نقاط سیاه بیشتر حفظ شده و بقیه اجزا متصل به‌عنوان نقاط یا نویز حذف شده‌اند.

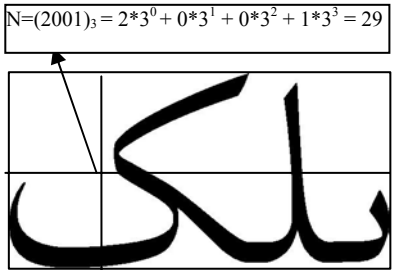
۳-۱- استخراج ویژگی زیر- کلمات و خوشه‌بندی آنها

در این مرحله ویژگیهای مکان مشخصه از تصویر زیر- کلمات چاپی فارسی استخراج شده‌اند. نحوه استخراج این ویژگیها در بخش ۳-۱ آمده است. از روش PCA برای تبدیل این ویژگیها به ویژگیهای ناهمبسته استفاده شده است. سپس تصویر زیر- کلمات با این ویژگیهای ناهمبسته و روش k- میانگین خوشه‌بندی شده‌اند. روش خوشه‌بندی در بخش ۳-۲ آمده است.

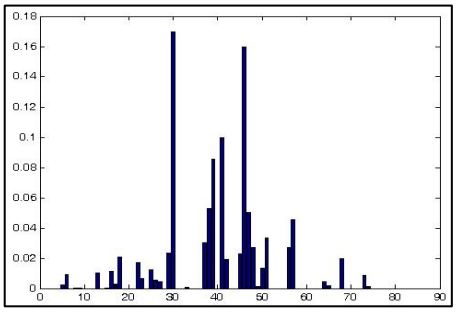
۳-۱- استخراج ویژگی زیر- کلمات

برای خوشه‌بندی زیر- کلمات از ویژگیهای مکان مشخصه آنها استفاده شده است. ویژگیهای مکان مشخصه معمولا در راستاهای عمودی و افقی تعریف می‌شوند [۶]. همانطور که در شکل (۲) آمده است، بردارهای مکانهای مشخصه به این صورت محاسبه می‌شود که به هر نقطه از زمینه تصویر، یک عدد نسبت

می‌دهیم. این عدد با توجه به اینکه خطوط عمودی و افقی رسم شده از آن نقطه در جهتهای چهارگانه بالا، پایین، راست و چپ، بدنه زیر- کلمه را در چند نقطه قطع می‌کنند، محاسبه می‌شود. تعداد قطع بدنه را به ۲ محدود می‌کنیم، بنابراین یک عدد چهار رقمی در مبنای ۳ به‌دست می‌آید. برای نمایش مکانهای مشخصه از معادل مبنای ۱۰ این عدد استفاده می‌شود. بردارهای مکان مشخصه در این حالت ۸۱ عنصر دارند که هر کدام فراوانی عدد مربوط به خود یا به عبارتی سطح مکان مشخصه مربوطه را در زمینه تصویر نشان می‌دهند. برای نرمالیزه کردن این ویژگیها، عناصر بردار به تعداد نقاط سفید زمینه تصویر تقسیم می‌شوند. بدین ترتیب شکل کلی هر زیر- کلمه با بردار فراوانی ویژگیهای مکان مشخصه در تصویر آن نمایش داده می‌شود. در شکل (۳) نمودار فراوانی نرمالیزه شده ویژگیهای مکان مشخصه برای زیر- کلمه شکل (۲) آمده است.



شکل ۲- نحوه محاسبه ویژگیهای مکان مشخصه



شکل ۳- نمودار فراوانی نرمالیزه شده ویژگیهای مکان مشخصه زیر- کلمه شکل ۲

با توجه به اینکه از ویژگیهای مکان مشخصه برای توصیف شکل کلی زیر- کلمات استفاده شده است. محدود کردن تعداد برخوردها با بدنه زیر- کلمات به ۲، منجر به یک توصیف کلی‌تر می‌شود. در کارهای بعدی انجام شده توسط مولفان این مقاله محدودیت تعداد برخوردها را به ۳ افزایش داده‌ایم.

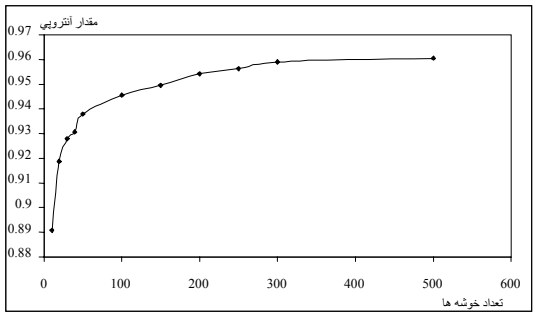
۳-۲- خوشه‌بندی زیر- کلمات

برای خوشه‌بندی از الگوریتم k-میانگین با معیار فاصله اقلیدسی استفاده شده است. به دلیل اینکه تعداد زیادی از عناصر بردارهای مکان مشخصه صفر هستند و بعضی ویژگیها همبسته هستند، ازروش PCA[1] برای تبدیل آنها به ویژگیهای ناهمبسته استفاده شده است. ماتریس تبدیل یک ماتریس ۸۱*۸۱ خواهد بود. برای به‌دست آوردن تعداد مناسب خوشه‌های زیر- کلمات، با تمام ۸۱ ویژگی در فضای ناهمبسته، آنها با روش k- میانگین برای k های مختلف خوشه‌بندی و در هر مرحله معیار آنتروپی برای تمام خوشه‌ها محاسبه شده است [۱۷ و ۲۵]. این معیار در رابطه (۱) آمده است.

$$H = \frac{1}{N\log(NC)}\sum_{k=1}^{N}\sum_{i=1}^{NC}[D^{-1}(k,i)\log(D^{-1}(k,i))] \qquad (۱)$$

که در آن N تعداد کل زیر- کلمات، NC تعداد خوشه‌ها، D(k,i) فاصله زیر- کلمه k ام از خوشه i ام و $D^{-1}(k,i)$ نشان دهنده درجه تعلق آن است.

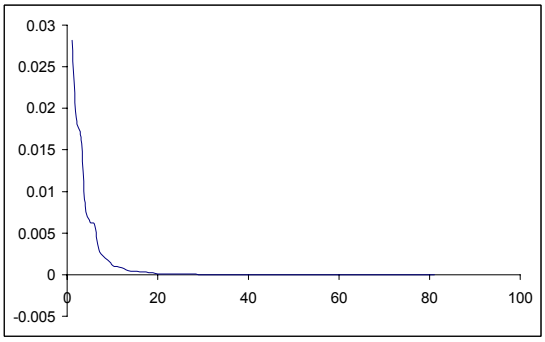
نتایج خوشه‌بندی برای K های مختلف در شکل(۴) آمده است. با توجه به این شکل، با خوشه‌بندی زیر- کلمات به حدود ۳۰۰ خوشه و بیشتر، معیار آنتروپی تغییر قابل ملاحظه‌ای نمی‌کند. پس ۳۰۰ خوشه برای دسته‌بندی نمونه‌های زیر- کلمات مناسب است.



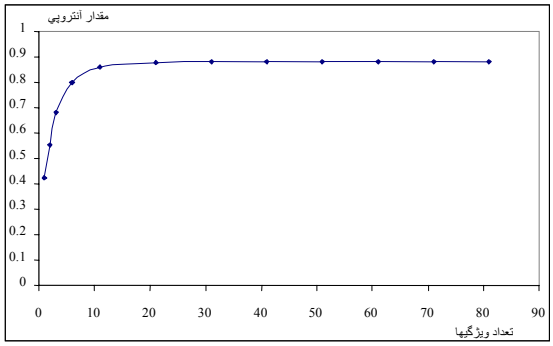
شکل ۴- معیار آنتروپی برای خوشه‌بندی زیر- کلمات به تعداد خوشه‌های مختلف با ۸۱ ویژگی

1- Principle component analysis

نمودار مقادیر ویژه به‌دست آمده از روش PCA در شکل (۵) آمده است. با توجه به شکل تعداد زیادی از مقادیر ویژه برابر صفر هستند. بنابراین در مرحله بعد برای انتخاب ویژگیهای مناسب از این ۸۱ ویژگی، مقادیر ویژه از بزرگ به کوچک مرتب شدند. با انتخاب چند ویژگی اول متناظر با مقادیر ویژه بزرگتر و خوشه‌بندی نمونه‌ها با آنها، دوباره معیار آنتروپی محاسبه شد. نتایج در شکل (۶) ارائه شده است. با توجه به این شکل، ۱۲ ویژگی اول، که دارای مقادیر ویژه متناظر بزرگتر هستند، برای خوشه‌بندی انتخاب شده‌اند.



شکل ۵- نمودار مقادیر ویژه بدست آمده از روش PCA



شکل ۶- معیار آنتروپی برای انتخاب ویژگیهای مناسب برای ۱۵۰ خوشه

با خوشه‌بندی زیر- کلمات با ۱۲ ویژگی متناظر با مقادیر ویژه بزرگتر در فضای ناهمبسته حاصل از روش PCA، با استفاده از روش خوشه‌بندی k-میانگین به ۳۰۰ خوشه، تعداد نمونه‌های خوشه‌ها بین ۲ تا ۵۸ نمونه متغیر است.

با خوشه‌بندی زیر- کلمات چاپی فارسی با همین ویژگیها و روش دسته‌بندی k-میانگین به ۱۵۰ خوشه، تعداد نمونه‌های خوشه‌ها با میانگین ۴۲/۴۶ زیر- کلمه، بین ۱۱ تا ۹۱ زیر- کلمه متغیر است. با توجه به این نتایج، ملاحظه می‌شود که میانگین تعداد نمونه‌های خوشه‌ها در خوشه‌بندی زیر- کلمات به ۳۰۰ خوشه نسبت به ۱۵۰ خوشه کمتر است. پراکندگی نمونه ها نیز در این حالت کمتر است و تعداد زیر- کلمات بیشتر خوشه‌ها نزدیک میانگین است.

۴- طبقه‌بندی زیر- کلمات

از تصاویر زیر- کلمات چاپی قلم لوتوس اندازه ۱۲ به عنوان مجموعه تمرین و آزمایش برای طبقه‌بندی استفاده شده است. برای طبقه‌بندی زیر- کلمات از معیار فاصله اقلیدسی استفاده شده است. فاصله هر زیر- کلمه از تمام خوشه‌ها با این معیار محاسبه شده و خوشه با کمترین فاصله به عنوان نزدیکترین خوشه انتخاب شده است. در رابطه (۲) نحوه محاسبه فاصله اقلیدسی آمده است.

$$d(i) = \sqrt{(x - m(i)) * (x - m(i))'} \qquad (۲)$$

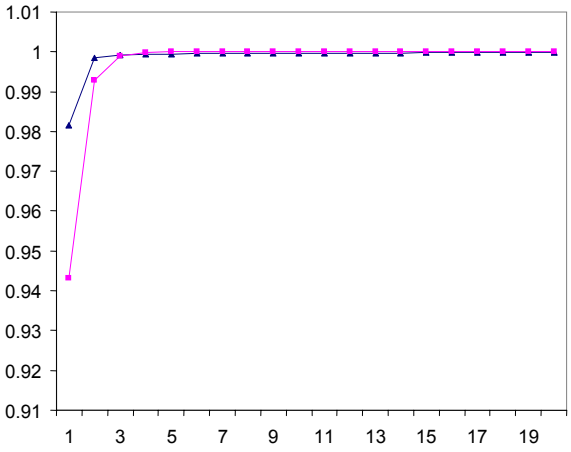
که در آن x زیر- کلمه مورد نظر، m(i) مرکز خوشه i ام، و d(i) فاصله اقلیدسی زیر- کلمه ورودی از خوشه i ام است. فاصله d هر زیر- کلمه ورودی از تمام خوشه‌ها محاسبه شده و به خوشه با کمترین فاصله نسبت داده شده است. نتایج طبقه‌بندی نمونه‌های آموزش، برای زیرکلمات چاپی فارسی با قلم لوتوس و اندازه ۱۲، به ۱۵۰ و ۳۰۰ خوشه با فاصله اقلیدسی در جدول (۲) آمده است. در این آزمایش از مجموعه آموزش، شامل ۹۴۴۵ زیر- کلمه، استفاده شده است.

جدول ۲- میزان طبقه‌بندی درست زیر- کلمات چاپی فارسی برای ۱۵۰ و ۳۰۰ خوشه

تعداد خوشه	طبقه‌بندی درست در انتخاب اول	طبقه‌بندی درست در ۵ انتخاب اول	طبقه‌بندی درست در ۱۰ انتخاب	طبقه‌بندی درست در ۲۰ انتخاب
۱۵۰	٪۹۴/۳۱	٪۱۰۰	٪۱۰۰	٪۱۰۰
۳۰۰	٪۹۸/۱۶	٪۹۹/۹۵	٪۱۰۰	٪۱۰۰

در شکل (۷) نمودارهای میزان طبقه‌بندی درست با فرض ۱۵۰ و ۳۰۰ خوشه، برحسب عمق انتخاب به‌ترتیب با مثلث روی

خط و مربع روی خط نشان داده شده‌اند. با توجه به شکل (۷)، وقتی تعداد انتخابها افزایش می‌یابد، طبقه‌بندی برای ۱۵۰ خوشه نسبت به طبقه‌بندی برای ۳۰۰ خوشه، نتایج بهتری نشان می‌دهد. البته باید به این موضوع توجه داشت که در طبقه‌بندی برای ۱۵۰ خوشه، هر خوشه یا هر انتخاب، به‌طور میانگین شامل ۴۲/۴۶ زیر- کلمه است. ولی در حالت دوم، این تعداد برابر ۲۱/۳۸ زیر- کلمه است. در مرحله بعدی باید زیر- کلمه ورودی را با این تعداد زیر- کلمه مقایسه کرده و شبیه‌ترین زیر- کلمه را انتخاب کرد. با افزایش انتخابها این اختلاف تعداد زیر- کلمات در حالت ۱۵۰ و ۳۰۰ خوشه بیشتر می‌شود. بنابراین طبقه‌بندی به ۳۰۰ خوشه نتایج بهتر و تعداد زیر- کلمات کمتری را تولید می‌کند.



شکل ۷- نمودار میزان طبقه‌بندی درست با فرض ۱۵۰ و ۳۰۰ خوشه بر حسب عمق انتخاب

برای آزمایش روش ارائه شده، از مجموعه‌ای جداگانه شامل ۲۰۰ زیر- کلمه استفاده شده است. این زیر- کلمات به‌صورت تصادفی از بین مجموعه زیر- کلمات انتخاب شده‌اند. سپس آنها با یک چاپگر لیزری HP 1100 چاپ و با یک روبشگر HP ScanJet 4200c روبش شدند. پس از پیش‌پردازش‌هایی مانند حذف نویز و نقاط و علائم، ویژگیهای مکان مشخصه آنها استخراج شدند. از معیار فاصله اقلیدسی برای طبقه‌بندی این زیر- کلمات با فرض ۳۰۰ خوشه استفاده شده است. نتایج طبقه‌بندی در جدول (۳) آمده است.

جدول ۳- میزان طبقه‌بندی درست با فرض ۳۰۰ خوشه

تعداد انتخابها	انتخاب اول	پنج انتخاب اول	ده انتخاب اول
میزان طبقه‌بندی درست	٪۸۰/۶۹	٪۹۷/۵۲	٪۱۰۰

همانطور که ملاحظه می‌شود، در ده انتخاب اول میزان طبقه‌بندی به ٪۱۰۰ رسیده است. به عبارت دیگر در مرحله بازشناسی باید تصویر کلمه ورودی را با حدود ۲۲۰ کلمه مقایسه کنیم.

۵- بررسی خوشه‌های حاصل از دسته‌بندی به ۱۵۰ و ۳۰۰ خوشه

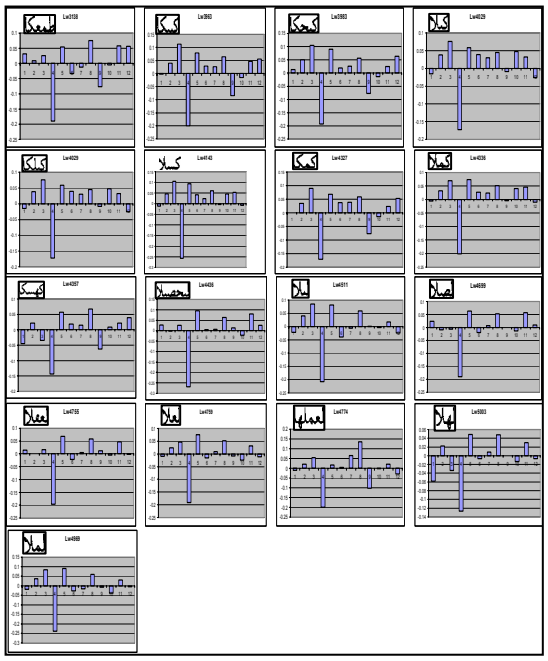
در این مرحله پس از خوشه‌بندی زیر- کلمات چاپی با قلم لوتوس و اندازه ۱۲ به ۱۵۰ و ۳۰۰ خوشه، در هر خوشه واریانس نمونه‌ها برای هر ویژگی محاسبه شده است. سپس خوشه‌هایی که دارای واریانس زیاد در یک یا چند ویژگی هستند به عنوان خوشه‌های نا همگن در نظر گرفته شده‌اند. حد آستانه برای این کار به صورت زیر محاسبه شده است که بزرگترین واریانس منهای کوچکترین واریانس شده و حاصل به چهار تقسیم شده است.

بزرگترین و کوچکترین واریانس از بین تمام ویژگیها در تمام خوشه‌ها محاسبه شده است. در این آزمایش حد آستانه $۱/۲۱۳*۱۰^{-۴}$ در طبقه‌بندی با فرض ۱۵۰ خوشه و حد آستانه $۱/۹۰۹*۱۰^{-۴}$ در طبقه‌بندی با فرض ۳۰۰ خوشه به‌دست آمده است. در دسته‌بندی با فرض ۳۰۰ خوشه، حدود ۳۳ خوشه و در دسته‌بندی با فرض ۱۵۰ خوشه، حدود ۹۶ خوشه واریانسهایی بزرگتر از این حد آستانه داشته‌اند. ملاحظه می‌شود خوشه‌های ناهمگن در حالت ۳۰۰ خوشه به مراتب کمتر از حالت ۱۵۰ خوشه است.

با بررسی خوشه‌هایی که دارای تغییرات زیاد واریانس در حداقل یکی از ویژگیها هستند، خطاهای به وجود آمده در ادامه می‌آیند:

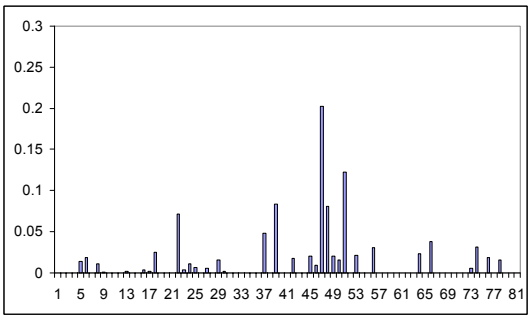
۱- برای مثال یکی از خوشه‌ها که واریانس زیادی در بعد سوم دارد در نظر می‌گیریم. در شکل (۸) تصویر زیر- کلمات این خوشه و نمودار واریانس آنها آمده است. این خوشه شامل ۱۷

زیر- کلمه است، با توجه به شکل (۸)، ملاحظه می‌شود که این زیر- کلمات به صورت نامناسبی در یک خوشه قرار گرفته‌اند.

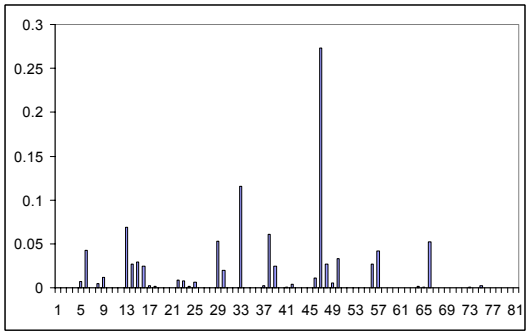


شکل ۸- تصویر زیر- کلمات و بردار ویژگی آنها در یک خوشه

واریانس زیاد ویژگی سوم در شکل (۸) نیز مشهود است. از این ویژگی می‌توان برای خوشه‌بندی ثانوی در داخل خوشه استفاده کرد. با توجه به شکل (۸)، زیر- کلماتی که بیش از ۲ بالارونده دارند، در این خوشه قرار گرفته‌اند. بنابراین زیر- کلماتی مانند "کیلکا" و "لیلا" در یک دسته قرار گرفته‌اند. در شکلهای (۹ و ۱۰) به ترتیب نمودار ویژگیهای مکان مشخصه این دو زیر- کلمه آمده است.



شکل ۹- نمودار فراوانی ویژگیهای مکان مشخصه برای زیر- کلمه "کیلکا"



شکل ۱۰- نمودار فراوانی ویژگیهای مکان مشخصه برای زیر- کلمه "لیلا"

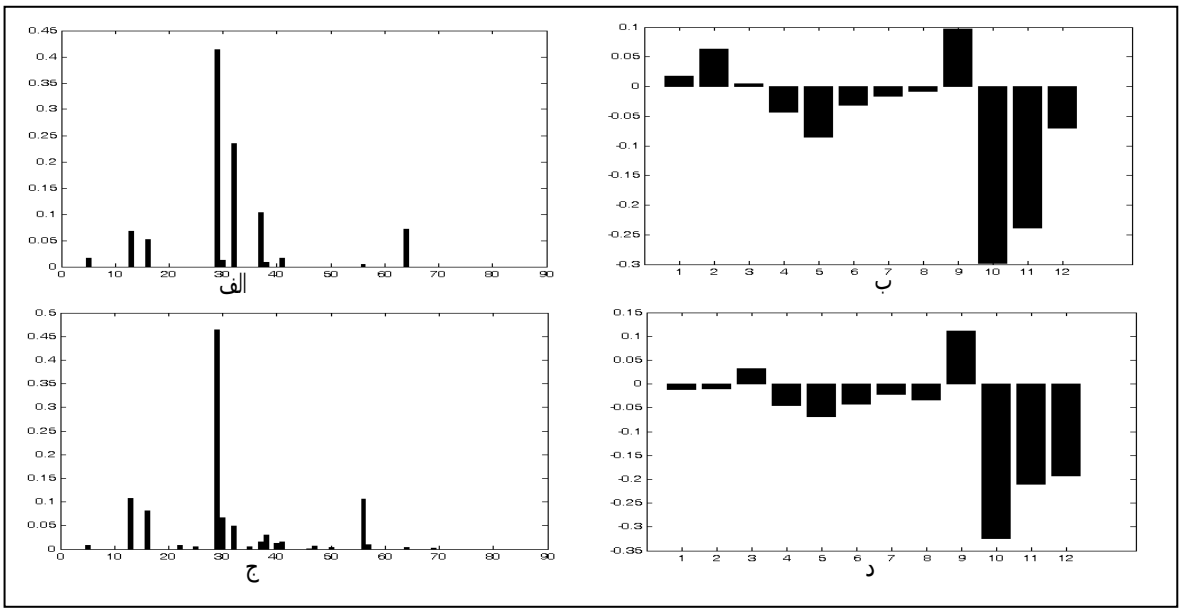
با توجه به این شکلها، ویژگی شماره ۴۷ مکان مشخصه برای هر دو زیر- کلمه اندازه بزرگی دارد. ویژگی شماره ۴۷ مکان مشخصه، به نقاطی از زمینه تصویر زیر- کلمه اختصاص دارد که خطوط افقی و عمودی رسم شده از این نقاط، بدنه زیر- کلمه را به ترتیب یکبار در سمت راست، دو بار در سمت چپ و یکبار در سمت پایین قطع کنند. به دلیل اینکه تعداد برخورد با بدنه را به دو محدود کرده‌ایم، اگر در زیر- کلمه‌ای تعداد برخوردها ۲ یا بیشتر باشد در محاسبات تاثیر نمی‌گذارد. بنابراین برای این دو زیر- کلمه ویژگی شماره ۴۷ آنها معادل می‌شود.

۲- دو زیر- کلمه "د" و "لو" در یک خوشه قرار گرفته‌اند. نمودار ویژگیهای مکان مشخصه و PCA این دو زیر- کلمه در شکل (۱۱) آمده است. با توجه به شکل (۱۱- الف) و (۱۱- ج)، ویژگیهای شماره ۱۳، ۱۷ و ۲۹ هر دو زیر- کلمه اندازه بزرگ و نزدیک به هم دارند. ولی در سایر ویژگیها با هم متفاوتند. اختلاف بین ویژگیهای PCA این دو زیر- کلمه در ویژگیهای ۲ و ۱۲، آنها است، که واریانس این دو ویژگی نیز آن را نشان می‌دهد.

۳- متفاوت بودن شکل بالارونده‌ها مانند " ک " و " ل ":

یکی از خوشه‌ها، شامل زیر- کلمات با دو بالارونده است. از نظر شکل کلی زیر- کلمات این خوشه مناسب به نظر می‌رسد. دلیل زیاد بودن واریانس ویژگی هفتم آن، شاید تفاوت بالارونده‌ها با هم باشد. بالا رونده‌ها در این خوشه شامل حروف " ک " و " ل " هستند. در شکل (۱۲) سه نمونه از زیر- کلمات این خوشه آمده است.

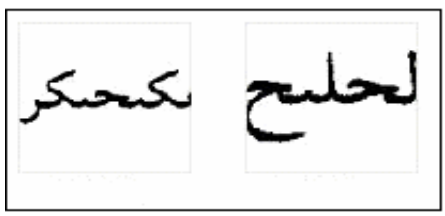
۴- متفاوت بودن شکل پایین رونده‌ها مانند "ح" و "ر":



شکل ۱۱- (الف) و (ب) نمودار فراوانی ویژگیهای مکان مشخصه و PCA برای زیر- کلمه "د" (ج) و (د) نمودار فراوانی ویژگیهای مکان مشخصه و PCA برای زیر- کلمه "لو"

شکل ۱۲- سه نمونه از زیر- کلمات یک خوشه با بالارونده‌های متفاوت

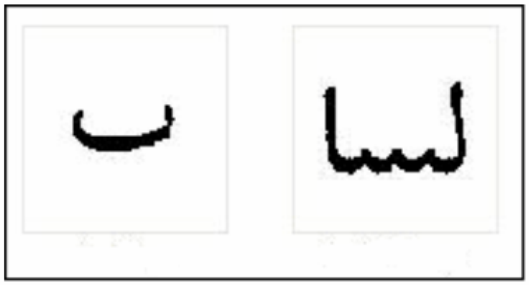
زیر- کلمات با دو بالارونده و یک پایین رونده در یک خوشه قرار گرفته‌اند. دلیل افزایش واریانس ویژگی ۵، متفاوت بودن حروف پایین رونده است. پایین رونده‌ها شامل حروف " و"، " ر " و "ح" هستند. در شکل (۱۳) دو زیر- کلمه از این خوشه آمده است.



شکل ۱۳- دو زیر- کلمه متعلق به یک خوشه، با پایین رونده‌های متفاوت

۵- خطای حاصل از نرمالیزه کردن به تعداد نقاط سفید

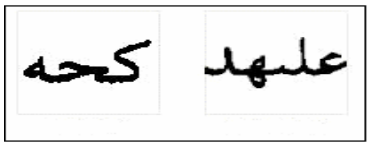
یک از خوشه‌ها مربوط به زیر- کلماتی است که در ابتدا و انتهای شکل آنها بالارونده وجود دارد. قرار گرفتن زیر- کلمه یک حرفی "ب" در این دسته باعث افزایش واریانس ویژگی ۹ شده است. به دلیل نرمالیزه کردن ویژگیهای مکان مشخصه به نقاط سفید تصویر، این خطا به وجود آمده است. دو زیر- کلمه نمونه در این خوشه در شکل (۱۴) آمده‌اند.



شکل ۱۴- دو زیر- کلمه با شکل متفاوت که به دلیل نرمالیزه کردن به نقاط سفید در یک خوشه قرار گرفته‌اند

۶- قرار گرفتن زیر- کلمات با انتهای " د " و " ه " در یک خوشه

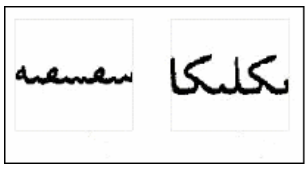
با آستانه‌زنی بر روی واریانس‌ها، در یکی از خوشه‌ها، ویژگی ۹ تغییرات زیادی نشان می‌دهد. در این خوشه زیر- کلمات با یک بالارونده، قرار دارند، ولی نواحی دیگر شکل آنها متفاوت هستند. در انتهای این زیر- کلمات حروف "د" یا "ه" قرار دارند. (شکل ۱۵).



شکل ۱۵- دو زیر- کلمه از یک خوشه با شکل کلی متفاوت

۷- قرار گرفتن زیر- کلمات با شکل کلی تخت در خوشه‌هایی شامل زیر- کلمات دارای بالارونده یا پایین رونده

در شکل (۱۶) دو زیر- کلمه متعلق به یک خوشه آمده است. اکثر زیر- کلمات این خوشه بیش از دو بالارونده دارند. ولی زیر- کلمه "شقشقیه" نیز در این خوشه قرار گرفته است. با توجه به ماهیت ویژگیهای مکان مشخصه، این زیر- کلمه نیز بیش از دو بالارونده کوچک دارد، که به خاطر نرمالیزه کردن به تعداد نقاط سفید با این زیر- کلمات در یک خوشه قرار گرفته است.



شکل ۱۶- دو زیر- کلمه با بالارونده‌های متفاوت از نظر تعداد و شکل که در یک خوشه افتاده‌اند

۶- نتیجه‌گیری

در این مقاله از ویژگیهای مکان مشخصه برای توصیف شکل کلی زیر- کلمات چاپی فارسی استفاده شد. محاسبه این ویژگیها با محدود کردن تعداد برخوردها با بدنه زیر- کلمه به ۲ انجام شد. با استفاده از روش PCA، ۸۱ ویژگی حاصل به ۱۲ ویژگی ناهمبسته تبدیل شدند. با استفاده از روش k-میانگین، تصاویر ۹۴۴۵ زیر- کلمه چاپی با قلم لوتوس ۱۲ به ۱۵۰ و ۳۰۰ خوشه تقسیم شدند. برای طبقه‌بندی یک مجموعه آزمایش شامل ۲۰۰ زیر- کلمه از کلماتی که در خوشه‌بندی شرکت نداشتند، از روش حداقل فاصله اقلیدسی استفاده شد. در انتخاب اول، پنج انتخاب اول و ده انتخاب اول به ترتیب ۸۰/۶۹٪، ۹۷/۵۲٪ و ۱۰۰٪ از این زیر- کلمات به درستی طبقه‌بندی شدند.

بیشتر کارهای انجام شده در زمینه بازشناسی متون فارسی با روشهای مبتنی بر جداسازی انجام شده‌اند. یکی از معدود کارهای انجام شده مربوط به کار آقای عزمی است [۲۵]. در این روش از داده‌هایی که در محیط کامپیوتری تولید شده‌اند (بدون چاپ و روبش)، استفاده شده است. این روش در مقابله با داده‌های واقعی، که چاپ و سپس روبش شده‌اند، مقاوم نیستند. با این حال با توجه به نتایج منتشر شده این روشها، نتایج روش این مقاله با آنها قابل مقایسه است. با این تفاوت که این کارها بر روی داده‌های کامپیوتری و روش ما بر روی داده‌های واقعی انجام شده است.

با توجه به نتایج بررسی خطاها، برای اصلاح آنها پیشنهاد می‌شود که در محاسبه ویژگیهای مکان مشخصه، تعداد برخوردها با بدنه زیر- کلمه به‌جای محدود کردن به ۲، به ۳ محدود شود. ویژگیهایی مانند تعداد بالارونده‌ها و پایین رونده‌ها، مکان تقریبی آنها، نسبت طول به عرض تصویر زیر- کلمه و ویژگیهای حروف شاخص به ویژگیهای مکان مشخصه اضافه شوند.

مراجع

[1] A. Amin, A. Kaced, J .P. Haton and R. Mohr, “Handwritten Arabic character recognition by the IRAC system”, Proceedings of the Fifth International Conference on Pattern Recognition, Miami Beach, FL, USA, pp. 729-731, 1980.

[2] K. Badie, M. Shimura, “Machine recognition of Arabic cursive scripts”, Proceedings of International Workshop Pattern Recognition in Practice., Amsterdam, Netherlands, pp. 315-323, 1980.

[3] M. Dehghan, K. Faez, M. Ahmadi and M. Shridhar, “Handwritten Farsi (Arabic) word recognition: a holistic approach using discrete HMM”, Pattern Recognition, vol. 34, no. 5, pp. 1057-1065, 2001.

[4] M. Dehghan, K. Faez, M. Ahmadi and M. Shridhar, “Unconstrained Farsi handwritten

word recognition using fuzzy vector quantization and hidden Markov models", Pattern Recognition Letters, vol. 22, no. 2, pp. 209-214, 2001.

[5] E. J. Erlandson, J .M. Trenkle, R .C. Vogt, "Word-level recognition of multifont Arabic text using a feature-vector matching approach" Proceedings of the SPIE, Document Recognition III, pp.63-71, San Jose, 1996.

[6] H. A. Glucksman, "Classification of mixedfont alphabets by characteristic loci", Proc. IEEE Comput. Conf., pp. 138-141, Sep. 1967.

[7] T. K. Ho and J. J. Hull and S. N. Srihari, "A Computational Model for Recognition of Multifont Word Images", Machine Vision and Applications, vol. 5, 3, pp. 157-168, 1992.

[8] T. K. Ho, J. J. Hull and S. N. Srihari, "Word recognition with multi-level contextual knowledge", Proceedings of the First International Conference on Document Analysis and Recognition, Saint-Malo, France, pp. 905-915, 1991.

[9] M. S. Khorsheed and W. F. Clocksin, "Multi-Font Arabic word recognition using spectral features", Proc. of ICPR 2000, vol. 4, pp. 543-546, 2000.

[10] T. K. Ho, J. J. Hull and S. N. Srihari, "A Word Shape Analysis Approach to Recognition of Degraded Word Images", Proc. of the 4th USPS Advanced Technology Conference, pp. 217-231, 1990.

[11] T. K. Ho, J. J. Hull and S. N. Srihari, "A Hypothesis Testing Approach to Word Recognition Using Dynamic Feature Selection", Proc. of the 11th Int'l Conference on Pattern Recognition, pp. 586-589, 1992.

[12] W. Huang, C. Tan, S. Sung and Y. Xu, "Word Shape Recognition for Image-based Document Retrieval", Proceedings of International Conference on Image Processing (ICIP01), pp. 1114-1117, 2001.

[13] J. J. Hull and S. N. Srihari, "A Computational Approach to Visual Word Recognition: Hypothesis Generation and Testing", Proceedings of IEEE-CS Conference on Computer Vision and Pattern Recognition, Miami Beach, pp. 156-161, Jun. 1986.

[14] J. J. Hull, "Hypothesis Testing in a Computational Theory of Visual Word Recognition", Proceedings of the Sixth National Conference on Artificial Intelligence (AAAI), Washington, pp. 718-722, 1987.

[15] S. Mori, C. Y. Suen and K. Yamamoto, "Histogram Review of OCR Research and Development", Proc. of IEEE, vol. 80, no. 7, pp. 1029-1058, Jul. 1992.

[16] J. O'Connor and A. F. Smeaton, "A Statistical Refinement Method for Word Shape Token Querying of Document Images", Proceedings of the Workshop on Document Analysis and Understanding for Document Databases, in conjunction with DEXA'99, Florence, Italy, pp. 572-576, Aug. 30-Sep. 3, 1999.

[17] S. K. Pal and D. K. Dutta, "Fuzzy Mathematical Approach to Pattern Recognition", Wiley Eastern Limited, 1986.

[18] R. K. Powalka, N. Sherkat and R. J. Whitrow, "Word shape analysis for a hybrid recognition system", Pattern Recognition, vol. 30, no. 3, pp. 421-445, 1997.

[19] A. F. Smeaton and A. L. Spitz, "Using character shape coding for information retrieval", 4th International Conference on Document Analysis and Recognition (ICDAR97), vol. 2, p. 974, 1997.

[20] A. F. Smeaton and J. O'Connor, "User-Mediated Word shape Tokens for Querying Document Images", Proceedings of the 3rd Australian Document Computing Symposium, University of Sydney, J. Kay (Ed), Aug. 1998.

[21] A. L. Spitz, "Shape-based word recognition", International Journal of Document Analysis and Recognition, vol. 1, no. 4, pp. 178-190, 1999.

[22] O. V. Trier, A. K. Jain and T. Taxt, "Feature extraction method for character recognition: A survey", Pattern Recognition, vol. 29, no. 4, pp. 641-662, 1996.

[23] J. S. Zhu, T. Hong, and J. J. Hull, "Image-based keyword recognition in oriental language document images", Pattern Recognition, vol. 30, no. 8, pp. 1293-1300, 1997.

[۲۴] م. رضوی، "گزارش داخلی طرح ملی بازشناسی متون چاپی فارسی و حجم محدودی از کلمات دستنویس"، دانشگاه تربیت مدرس، ۱۳۸۱.

[۲۵] ر. عزمی، "بازشناسی متون چاپی فارسی"، رساله دکتری مهندسی برق- الکترونیک، دانشگاه تربیت مدرس، ۱۳۷۸.

[۲۶] ک. مسروری، "شناسایی برون خط کلمات دستنویس فارسی در یک مجموعه محدود"، رساله دکتری مهندسی برق- الکترونیک، دانشگاه تربیت مدرس، تابستان ۱۳۷۹.