یک روش دو مرحله‌ای برای بازشناسی زیر- کلمات چاپی

افشین ابراهیمی و احسان اله کبیر

چکیده: در این مقاله یک روش دو مرحله ای برای طبقه بندی زیر- کلمات چاپی فارسی ارائه شده است. زیر- کلمات چاپی با استفاده از ویژگیهای مکان مشخصه و روش k- میانگین، به ۳۰۰ خوشه تقسیم شده‌اند. از میانگین ویژگیهای زیر- کلمات هر خوشه به عنوان نماینده آن خوشه استفاده شده است. برای یک زیر- کلمه ورودی، در مرحله اول با استفاده از ویژگیهای مکان مشخصه و فاصله اقلیدسی از میانگین خوشه ها، طبقه بندی اولیه به ۳۰۰ خوشه انجام می‌شود و ۱۰ خوشه نزدیکتر تعیین می‌شوند. در مرحله دوم با استفاده از توصیفگرهای فوریه کانتور، زیر- کلمه ورودی به اعضای این ۱۰ خوشه طبقه بندی می‌شود.

مجموعه تمرین شامل زیر- کلمات متداول فارسی برای چهار قلم لوتوس، میترا، زر و یاقوت و سه اندازه ۱۰، ۱۲ و ۱۴ است. در این تحقیق از بدنه های بدون نقطه ۱۲۷۰۰ زیر- کلمه متداول فارسی به عنوان مجموعه تمرین استفاده شده است. در یک آزمایش برای ارزیابی طبقه بندی از مجموعه ای شامل ۵۰۰ زیر- کلمه استفاده شد. با احتساب اولین انتخاب، پنج انتخاب اول و ده انتخاب اول به ترتیب ۷۱/۴۰٪، ۹۵٪ و ۹۸/۲۰٪ از این زیر- کلمات به درستی طبقه‌بندی شدند. در مرحله پس پردازش از نوع و ترتیب نقاط زیر- کلمات برای بهبود بازشناسی آنها استفاده شد. در یک آزمایش برای بازشناسی یک مجموعه ۵۰۰ زیر- کلمه ای، در انتخاب اول ۹۲/۶۰٪ از آنها به درستی بازشناسی شدند.

کلید واژه: **متن چاپی، زیر- کلمه، خوشه بندی، طبقه بندی، بازشناسی، ویژگیهای مکان مشخصه، k - میانگین، توصیفگرهای فوریه.**

۱- مقدمه

اولین تحقیقات در زمینه بازشناسی متون مربوط به سالهای ۱۹۲۹ و ۱۹۳۳ میلادی است [۱]. این سیستمها حروف چاپی را با روش تطبیق کلیشه ای شناسایی می‌کردند و به دلیل استفاده از تکنولوژی اپتومکانیکی کاربردی نبودند. از اواسط دهه ۱۹۵۰، بازشناسی متن بصورت یک زمینه فعال برای تحقیق درآمد. امروزه نرم افزارهای تجاری برای بازشناسی مستندات چاپی با کیفیت مناسب و حروف مجزای دستنویس ساخته شده‌اند. تحقیقات فعلی در زمینه OCR مربوط به متون چاپی با قلمهای گوناگون[1] یا با کیفیت تصویری پایین و متون دستنویسی است که بدون محدودیت خاصی[2] نوشته شده باشند.

تحقیقات در زمینه بازشناسی متون فارسی و عربی نیز از سال ۱۹۸۰ شروع شده است [۲] تا [۴]. روشهای بازشناسی متون از دو رویکرد مبتنی بر جداسازی کلمات به حروف و زیر- حروف و رویکرد مبتنی بر شکل کلی کلمات، استفاده می‌کنند [۵] تا [۱۲]. در روشهای 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 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. در این تحقیق نیز از رویکرد مبتنی بر شکل کلی کلمات استفاده شده است.

در یک تحقیق، الگوریتمی برای خوشه بندی زیر- کلمات چاپی و ساختن دیکشنری تصویری برای کمک به بازشناسی آنها ارائه شده است [۱۳]. در مرحله ایجاد دیکشنری، کانتور بالایی تصویر هر زیر-کلمه استخراج می‌شود و به کمک مجموعه ای 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. زیر- کلمات مختلفی که اندیس یکسانی دارند، همسایگی خاص خود را در دیکشنری تشکیل می‌دهند. در مرحله طبقه بندی، 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.

در تحقیق مرجع [۱۳] 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 همسایگی های تصویری ۷۴/۳۷ زیر- کلمه و متوسط اندازه دسته های ایجاد شده با توجه به حروف شاخص ۴/۳ زیر- کلمه بوده است. میزان دسته بندی 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 نمونه ها، هیچگونه نویز یا اعوجاجی در تصاویر وجود ندارد.

در [۱۴] از ویژگیهای شبه زرنیکی برای بازشناسی متون تایپی و دستنویس با شبکه عصبی احتمالاتی استفاده شده است. روش ارائه شده در این تحقیق بر رویکرد مبتنی بر شکل زیر- کلمات استوار است. تعداد زیر- کلمات فارسی ۱۰۰۰ زیر- کلمه فرض شده است. شبکه عصبی استفاده شده، با در نظر گرفتن لایه ورودی، چهار لایه دارد.

در [۱۴] از ۲۸ ویژگی گشتاورهای شبه زرنیکی مرتبه ۴ برای توصیف شکل زیر- کلمات استفاده شده است. با آزمایش این روش بر روی مجموعه‌ای شامل ۵۰۰ زیر- کلمه دستنویس، نوشته شده با خط نسخ، و یک صفحه متن چاپی نرخ بازشناسی صحیح ۹۶٪ گزارش شده است.

این مقاله در تاریخ ۲۵ مرداد ماه ۱۳۸۳ دریافت و در تاریخ ۲۹ اسفند ماه ۱۳۸۳ بازنگری شد. این تحقیق از پشتیبانی مالی مرکز تحقیقات مخابرات بر اساس قرارداد شماره ۵۰۰/۸۲۱۱ت برخوردار بوده است.

افشین ابراهیمی، بخش مهندسی برق، دانشکده فنی و مهندسی، دانشگاه تربیت مدرس، تهران، ایران، (email:ebrah_af@modares.ac.ir).

احسان اله کبیر، بخش مهندسی برق، دانشکده فنی و مهندسی، دانشگاه تربیت مدرس، تهران، ایران، (email:kabir@modares.ac.ir).

1. Omnifont
2. Unconstrained Handwriting

در [۱۵] از ویژگیهای شکل کلمات چاپی در بازشناسی متون عربی، نوشته شده با سه قلم متداول، استفاده شده است. ویژگیهایی مانند نقاط، همزه، جهت پاره خطها، نقاط انتهایی و اتصالها، حفره ها، پایین رونده ها و فواصل درون کلمه ای، 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. 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 [۱۶]. تصویر کلمه به یک تصویر قطبی نرمالیزه شده تبدیل می‌شود، سپس از این تصویر تبدیل فوریه دوبعدی گرفته می‌شود. طیف حاصل نسبت به تغییرات اندازه، چرخش و جابجایی مقاوم است. از مجموعه ای از ضرایب فوریه، برای بازنمایی 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 [۱۷]. ویژگیهای استفاده شده در این تحقیق از اطلاعات کانتور کلمات استخراج شده‌اند. کدهای زنجیره ای جهتی کانتور کلمه دستنویس محاسبه می‌شود. این تصویر در راستای افقی به پنج قسمت مساوی و در راستای عمودی به قسمتهایی که ۵۰٪ با هم همپوشانی دارند تقسیم می‌شود. عرض تقسیمات عمودی دو برابر میانگین تکه های سیاه عمودی تصویر انتخاب می‌شود. برداری شامل هیستوگرام های جهتی کدهای زنجیره ای در هر یک از این پنجره ها بعنوان مدل کلمه انتخاب می‌شود.

برای کم‌کردن تعداد مشاهده هایی که به مدل مخفی مارکف گسسته اعمال می‌شود، فضای ویژگی با استفاده از یک شبکه عصبی خود سامانده کوهنن (SOM)، کوانتایز شده است. برای هر اسم شهر یک HMM گسسته با الگوریتم Baum Welch بطور 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 اسامی 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 بوسیله افراد مختلف نوشته شده‌اند.

در [۱۸] از خوشه بندی FCM و مدل مخفی مارکف برای بازشناسی اسامی دستنویس شهرهای فارسی استفاده شده است. مراحل پیش پردازش و استخراج ویژگی و مدل مخفی مارکف مانند مرجع [۱۷] است. خوشه بندی اولیه به روش FCM انجام شده است. با آزمایش این روش بر روی مجموعه ای شامل ۱۹۸ اسم شهر 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 بوسیله افراد مختلف نوشته شده‌اند.

از الگوریتم DTW دوبعدی نیز برای بازشناسی اسامی دستنویس ۵۰۰ شهر ایران استفاده شده است [۱۹]. 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 تکه ای تقریب زده می‌شود. طول و زاویه هر یک از این پاره خطها نسبت به راستای افقی در بردار ویژگی زیر- کلمه ذخیره می‌شود. برای بازشناسی تصویر ورودی، میزان شباهت زیر- کلمات آن با زیر- کلمات دیکشنری با روش DTW دوبعدی محاسبه می‌شود و با روش فازی K همسایه نزدیکتر، طبقه بندی می‌شود. برای هر زیر- کلمه ۵ انتخاب اول در نظر گرفته می‌شود. بهترین تطابق ترکیب های مختلف این انتخاب ها با کلمات معتبر بعنوان کلمه بازشناسی شده معرفی می‌شود. در آزمایش این روش برای مجموعه اسامی ۱۰۰، ۳۰۰ و ۵۰۰ شهر ایران، نرخ های بازشناسی صحیح ۸۸، ۸۴ و ۷۵ درصد گزارش شده است.

هدف این مقاله استفاده از شکل زیر- کلمات برای طبقه بندی آنهاست. زیر- کلمات استفاده شده در این مقاله از پایگاه داده دو روزنامه کیهان و همشهری استخراج شده‌اند. از مستندات این پایگاهها، متداول ترین 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 آنها است. در استخراج زیر- کلمات، حروف با بدنه یکسان با یکی از این حروف، نماینده گروه، جاگذاری شده‌اند. 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 ۱۰، ۱۲و ۱۴ با یک چاپگر لیزری HP1200 چاپ و سپس با یک روبشگر HP ScanJet 5550c با درجه تفکیک ۴۰۰dpi روبش شده‌اند. بدین ترتیب پایگاه داده تصاویر شامل ۱۱۳۳۴۰ زیر- کلمه است.

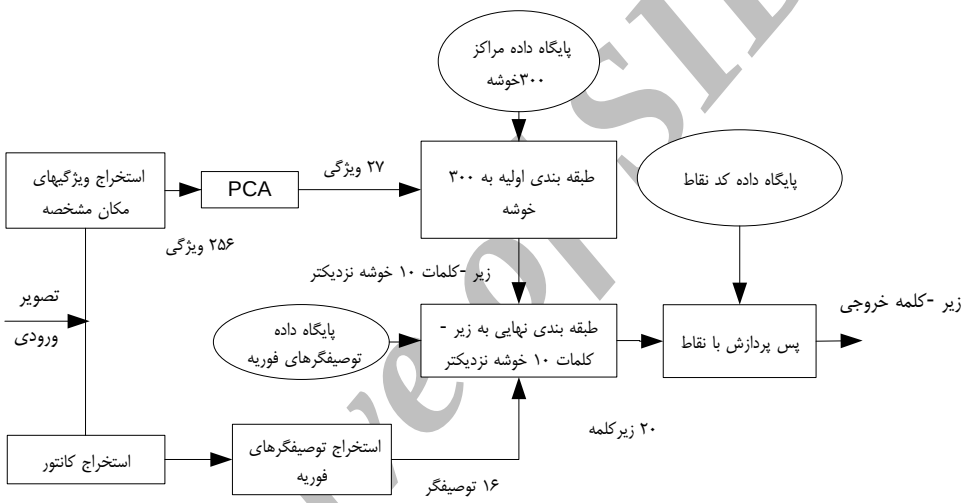
پس از استخراج خطوط با استفاده از هیستوگرام افقی، زیر- کلمات با برچسب زنی اجزای پیوسته استخراج شده‌اند. کجی تصاویر خطوط پس از استخراج برطرف شده است. محل خط زمینه استخراج شده و اجزای پیوسته‌ای که با آن تقاطع داشته باشند بعنوان بدنه زیر- کلمات در نظر گرفته شده‌اند. نقاط و علائم با توجه به قرار گیری آنها زیر یا روی هر بدنه به آن بدنه تخصیص داده شده‌اند. برای حذف نقاط، به اجزای پیوسته زیر- کلمه برچسب زده شده است. جزء پیوسته با تعداد نقاط سیاه بیشتر حفظ شده و بقیه اجزای پیوسته بعنوان نقاط یا نویز حذف شده‌اند.

در ادامه در بخش ۲ نحوه خوشه بندی زیر- کلمات آمده است. بخش ۳ به روش طبقه بندی دو مرحله ای زیر- کلمات می‌پردازد. در بخش ۴ پس پردازش با نقاط آمده است. در بخش ۵ نیز نتیجه گیری ارائه شده است.

۲- خوشه بندی زیر- کلمات چاپی فارسی با ویژگیهای مکان مشخصه

در مرحله خوشه بندی از ویژگیهای مکان مشخصه برای توصیف شکل زیر- کلمات استفاده شده است [۲۰]. بردارهای مکانهای مشخصه به اینصورت 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 جهت های چهارگانه بالا، پایین، راست و چپ، بدنه زیر- کلمه را در چند نقطه قطع می‌کنند، محاسبه می‌شود. تعداد قطع بدنه را به ۳ محدود می‌کنیم، بنابراین یک عدد چهار رقمی در مبنای ۴ بدست می‌آید. برای نمایش مکانهای مشخصه از معادل مبنای ۱۰ این عدد استفاده می‌شود. بردارهای مکان مشخصه در این حالت ۲۵۶ عنصر دارند که هر کدام فراوانی عدد مربوط به خود یا به عبارتی سطح مکان

شکل ۱: شمای تعدادی از زیر- کلمات یک خوشه.



شکل ۲: دیاگرام بلوکی روش طبقه بندی زیر- کلمات.

مشخصه مربوطه را در زمینه تصویر نشان می‌دهند. برای نرمالیزه کردن این ویژگیها، عناصر بردار به تعداد نقاط سفید زمینه تصویر تقسیم می‌شوند. بدلیل صفر بودن بعضی از مؤلفه های این بردارها و همبسته بودن آنها، از این ویژگیها با استفاده از روش PCA، ۲۷ ویژگی ناهمبسته انتخاب شده است. با استفاده از این ۲۷ ویژگی و روش k – میانگین، ۱۱۳۳۴۰ زیر- کلمه موجود در پایگاه داده زیر- کلمات به ۳۰۰ خوشه تقسیم شده‌اند [۲۰].

تعداد مناسب خوشه ها با استفاده از معیار آنتروپی بدست آمده است. بدین ترتیب که زیر- کلمات را به تعداد خوشه های مختلف خوشه بندی کرده و در هر مرحله آنتروپی را محاسبه کردیم. با افزایش تعداد خوشه ها آنتروپی بیشتر شد و برای ۳۰۰ خوشه و بیشتر معیار آنتروپی تغییر محسوسی نکرد. بنابراین تعداد مناسب خوشه ها را ۳۰۰ انتخاب کردیم. میانگین تعداد نمونه های خوشه ها ۳۷۷/۸ زیر- کلمه است. از میانگین ویژگیهای زیر- کلمات هر خوشه به عنوان نماینده آن خوشه در مراحل بعدی استفاده شده است. در نتیجه این خوشه بندی، زیر- کلمات با شکل کلی مشابه در خوشه های یکسان قرار گرفته‌اند. در شکل ۱ تعدادی از زیر- کلمات یک خوشه آمده است. همانطور که ملاحظه می‌شود، برای زیر- کلمه "حیلها"، زیر- کلمات تمام قلم ها در این خوشه قرار گرفته‌اند.

۳- طبقه بندی زیر- کلمات با استفاده از روش دو مرحله‌ای

طبقه بندی زیر- کلمات در دو مرحله انجام شده است [۲۱]. در مرحله اول، زیر- کلمه ورودی با استفاده از ویژگیهای مکان مشخصه و معیار فاصله اقلیدسی از میانگین خوشه ها به ۳۰۰ خوشه طبقه بندی شده است. در مرحله دوم، این زیر- کلمه با استفاده از توصیفگرهای فوریه کانتور آن به زیر- کلمات ۱۰ خوشه نزدیکتر طبقه بندی شده است. دیاگرام بلوکی روش طبقه بندی در شکل ۲ آمده است.

۳-۱ مرحله اول

برای طبقه بندی زیر- کلمات از معیار فاصله اقلیدسی استفاده شده است. فاصله هر زیر- کلمه از میانگین هر خوشه با این معیار محاسبه شده و خوشه با کمترین فاصله به عنوان نزدیکترین خوشه انتخاب شده است. نتایج طبقه بندی نمونه های آموزش، برای زیرکلمات چاپی فارسی با چهار قلم لوتوس، میترا، یاقوت و زر و سه اندازه ۱۰، ۱۲ و ۱۴ به ۳۰۰ خوشه در جدول ۱ آمده است. در این آزمایش از مجموعه آموزش، شامل ۱۱۳۳۴۰ زیر- کلمه، استفاده شده است. همانطور که در جدول ۱

جدول ۱: طبقه بندی درست زیر- کلمات چاپی فارسی برای ۳۰۰ خوشه.

مجموعه	طبقه بندی درست در انتخاب اول	طبقه بندی درست در ۵ انتخاب اول	طبقه بندی درست در ۱۰ انتخاب اول
آموزش	٪۱۰۰	٪۱۰۰	٪۱۰۰
آزمایش	٪۹۳/۶۰	٪۹۹/۴۰	٪۱۰۰

جدول ۲: میزان طبقه بندی درست در ۱۰ خوشه اول با ۲۰ ضریب فوریه.

تعداد انتخابها	انتخاب اول	۵ انتخاب اول	۱۰ انتخاب اول
میزان طبقه بندی درست	٪۷۰/۲۰	٪۹۴/۸۰	٪۹۹/۰۰

ملاحظه می‌شود، صد درصد نمونه ها حتما در ۱۰ خوشه اول بدرستی طبقه بندی می‌شوند. بنابراین می‌توان طبقه بندی نهایی را منحصر به زیر- کلمات ۱۰ خوشه نزدیکتر کرد.

۳-۲ مرحله دوم

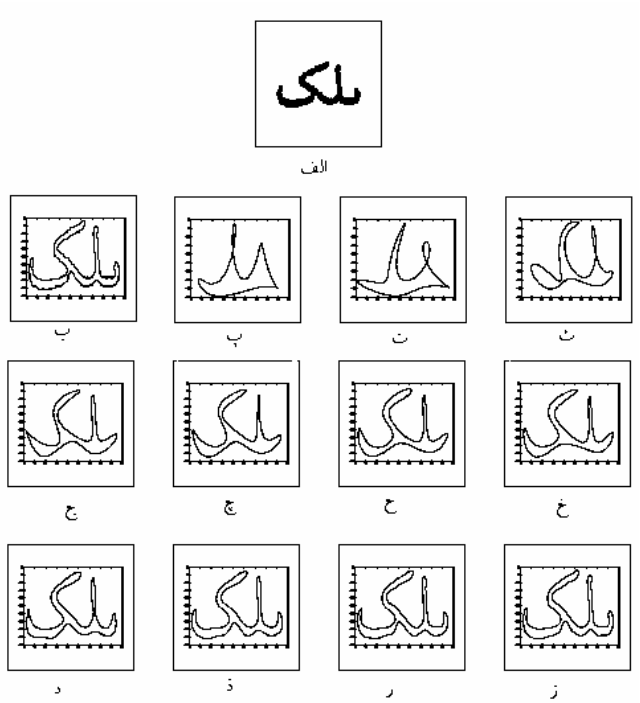
ویژگیهای مکان مشخصه، که برای خوشه بندی استفاده شدند، شکل کلی زیر- کلمه را توصیف می‌کنند. پس از خوشه بندی زیر- کلمات به ۳۰۰ خوشه، نمونه های داخل هر خوشه از نظر شکل کلی شبیه هم هستند. بنابراین برای مرحله دوم باید از ویژگیهایی که جزئیات شکل این زیر- کلمات را توصیف کنند، استفاده کنیم. در این تحقیق برای طبقه بندی در داخل خوشه ها از توصیفگرهای فوریه منحنی پیرامونی زیر- کلمات استفاده شده است [۲۲]. در شکل (۳-ب)، منحنی پیرامونی زیر- کلمه شکل (۳-الف) آمده است.

برای محاسبه توصیفگرهای فوریه یک زیر- کلمه، ابتدا منحنی پیرامونی آن را پیدا می‌کنیم. برای استخراج کانتور یک زیر- کلمه از روش دنبال کردن کانتور استفاده شده است. بدین ترتیب که ابتدا نقاط سیاه تصویر استخراج شده و سپس با روبش تصویر به صورت افقی و عمودی نقاط گذر از سیاه به سفید و برعکس پیدا شده‌اند. مختصات نقاط آن را به صورت رابطه زیر با هم ترکیب می‌کنیم.

$$z = x + iy \qquad (۱)$$

از این رشته موهومی تبدیل فوریه می‌گیریم و اندازه و فاز اعداد موهومی بدست آمده را محاسبه می‌کنیم. بدلیل اینکه اندازه این ضرایب، بر خلاف فازشان به چرخش حساس نیستند از آنها برای توصیف منحنی پیرامونی شکل زیر- کلمات استفاده شده است. اولین ضریب بدست آمده نشان دهنده مقدار ثابت سیگنال است. برای اینکه این ضرایب نسبت به اندازه حساس نباشند به ضریب اول تقسیم می‌شوند. در شکل ۳ تصویر یک زیر- کلمه و تصاویر بازسازی شده منحنی پیرامونی آن با استفاده از تعداد متفاوتی توصیفگر فوریه آمده است. همانطور که ملاحظه می‌شود، می‌توان منحنی پیرامونی را با ۲۰ توصیفگر فوریه به خوبی توصیف کرد. از ۲۰ ضریب اول برای توصیف منحنی پیرامونی شکل زیر- کلمه استفاده شده است. نتایج طبقه بندی نهایی در ۱۰ خوشه اول، برای مجموعه ای شامل ۵۰۰ زیر- کلمه در جدول ۲ آمده است.

با توجه به این که توصیفگرهای فوریه به ترتیب از فرکانس‌های پایین به بالا مرتب می‌شوند، چند ضریب اول نشانگر فرکانس‌های پایین هستند. فرکانس‌های پایین نشانگر شکل کلی زیر- کلمات هستند. به دلیل این که در هر خوشه زیر- کلمات از نظر شکل کلی با هم شبیه هستند، برای



شکل ۳: (الف) تصویر زیر- کلمه "پلک"، (ب) منحنی پیرامونی آن، (پ-ز) بازسازی منحنی پیرامونی به ترتیب با استفاده از ۴، ۵، ۶، ۷، ۸، ۹، ۱۰، ۱۵، ۲۰، ۳۰ و ۴۰ توصیفگر فوریه.

طبقه‌بندی بهتر باید ویژگیها بیانگر جزییات شکل زیر- کلمه یا فرکانس های بالای توصیفگرهای فوریه باشند. بنابراین با استفاده از مجموعه آزمایش قبلی با حذف چند ضریب اول و طبقه بندی با بقیه ضرایب، تعداد مناسب توصیفگرهای فوریه بدست آمد. نتایج در جدول ۳ آمده است.

با توجه به جدول ۳، میزان طبقه بندی درست در انتخاب اول با حذف به ترتیب ضریب اول، دو، سه و چهار ضریب اول تغییر نمی‌کند و با حذف ضریب پنجم این میزان کاهش می‌یابد. بنابراین برای طبقه بندی داخل خوشه ها از ۱۶ توصیفگر فوریه استفاده شده است.

٤- بازشناسی نهایی با توجه به نقاط زیر- کلمات

در این مرحله نقاط زیر- کلمه ورودی به کمک یک شبکه عصبی پس انتشار خطا بازشناسی می‌شوند. علایم و نقاطی که در این مرحله بازشناسی می‌شوند، تک نقطه، دو نقطه، سه نقطه بالا، سه نقطه پایین، سرکش و همزه هستند. به همین دلیل در لایه خروجی شبکه عصبی ۶ گره به تعداد کلاسها قرار داده شده است. از ۶ ویژگی برای بازشناسی نقاط و علائم استفاده شده است. این ویژگیها با توجه به شکل و نحوه قرار گیری نقاط در کنار هم، تعریف شده‌اند.

f1 – نسبت تعداد نقاط سیاه به سطح مستطیل محیطی

f2 – نسبت تعداد نقاط سیاه در نیمه چپ مستطیل محیطی به تعداد نقاط سیاه در نیمه راست آن

f3 – نسبت تعداد نقاط سیاه در نیمه بالایی مستطیل محیطی به تعداد نقاط سیاه در نیمه پایینی آن

f4 – نسبت عرض به ارتفاع مستطیل محیطی

f5 – نسبت تعداد نقاط سیاه یا سفید مشترک در نیمه چپ و راست مستطیل محیطی به سطح آن

f6 – نسبت تعداد نقاط سیاه یا سفید مشترک در نیمه بالا و پایین مستطیل محیطی به سطح آن

جدول ۳: میزان طبقه بندی درست در ۱۰ خوشه اول با حذف چند توصیفگر اول فوریه.

میزان طبقه بندی درست	انتخاب اول	۵ انتخاب اول	۱۰ انتخاب اول
با حذف توصیفگر ۱	٪۷۲/۸۰	٪۹۵/۲۰	٪۹۹/۰۰
با حذف توصیفگرهای ۱ و ۲	٪۷۲/۶۰	٪۹۴/۶۰	٪۹۸/۸۰
با حذف توصیفگرهای ۱ تا ۳	٪۷۱/۶۰	٪۹۴/۸۰	٪۹۹/۰۰
با حذف توصیفگرهای ۱ تا ۴	٪۷۱/۴۰	٪۹۵/۰۰	٪۹۹/۰۰
با حذف توصیفگرهای ۱ تا ۵	٪۷۰/۲۰	٪۹۳/۲۰	٪۹۷/۴۰

لایه ورودی این شبکه نیز ۶ گره دارد. در لایه میانی ۹ گره قرار داده شده است. برای آموزش شبکه عصبی و تنظیم وزنها ۱۲۰۰ نمونه که شامل ۲۰۰ نمونه از هر کلاس بود را به شبکه نشان دادیم. پارامتر یادگیری این شبکه عصبی ۰/۲ بود.

پس از شناسایی تمام نقاط یک زیر- کلمه، آنها را از راست به چپ و با توجه به اینکه بالا یا پایین خط زمینه قرار دارند، مرتب کرده و کدی را به آن نسبت می‌دهیم. مثلا زیر- کلمه "محبت"، دو جزء نقطه ای دارد که کد نقاط آن "bt" است. کد "b" برای تک نقطه زیر و "t" برای دو نقطه بالا است. سپس این کد را با کد زیر- کلمات بدنه های شناخته شده - ۲۰ زیر- کلمه نزدیکتر- در مرحله دوم بازشناسی، مقایسه کرده و زیر- کلمهٔ بازشناسی شده استخراج می‌شود.

در یک آزمایش برای ارزیابی روش ارائه شده از یک مجموعه شامل ۵۰۰ زیر- کلمه استفاده شد. از این ۵۰۰ زیر- کلمه ٪۹۲/۶۰ از آنها در انتخاب اول به درستی بازشناسی شدند.

تمام مراحل این تحقیق در نرم افزار MATLAB 7.0 پیاده سازی شده‌اند. سرعت بازشناسی زیر- کلمات در حدود ۳۰۰ زیر- کلمه در دقیقه است.

۵- نتیجه گیری و پیشنهادها

با توجه به نتایج بدست آمده، طبقه بندی اولیه به ۳۰۰ خوشه با استفاده از ویژگیهای مکان مشخصه با محدود کردن تعداد برخوردها با بدنه زیر- کلمه به ۳ و طبقه بندی نهایی در ۱۰ خوشه نزدیکتر به تصویر ورودی با ۱۶ توصیفگر فوریه منحنی پیرامونی شکل زیر- کلمه انجام شد. سنجش کارایی طبقه‌بندی به این صورت بوده است که اگر یک زیر- کلمه به یکی از ۱۲ زیر- کلمه مربوط به آن، معادل آن در پایگاه داده تصاویر زیر- کلمات چهار قلم و سه اندازه، نسبت داده شد؛ طبقه‌بندی، درست انجام شده است.

خطاهای طبقه بندی بیشتر مربوط به زیر- کلمات با اندازه ۱۰ بوده است. بدلیل اینکه زیر- کلمات مجموعه آزمایش با کیفیت نامطلوب چاپ شده‌اند، زیر- کلمات با اندازه ۱۰ نازکتر از اندازه واقعی چاپ شده‌اند. بنابراین محاسبه توصیفگرهای فوریه کانتور این زیر- کلمات دقیق نبوده است.

در مرحله پس پردازش با نقاط از یک مجموعه شامل ۵۰۰ زیر- کلمه با اندازه ها و قلم های مختلف،٪۹۲/۶۰ از زیر- کلمات به درستی طبقه بندی شدند. در شکل ۴ تصاویر تعدادی از زیر- کلمات که به اشتباه

شکل ۴: نمونه ای از خطاهای الگوریتم بازشناسی.

بازشناسی شده‌اند، آمده است. با توجه به شکل ۴، می‌توان خطاهای بوجود آمده را به چند دسته تقسیم کرد. نوع خطا در شکل ۴ با برچسبهای "الف" تا "خ" نشان داده شده است.

الف- بریدگی در بدنه، بدنه تعدادی از زیر- کلمات کامل نبودند.
ب- چسبیدن سرکش "گ"
ت- عدم تشخیص صحیح سه نقطه هایی که جدا از هم بودند.
ث- بدون نقطه بودن زیر- کلمه، مانند زیر- کلمه "لکل"
ج- اشتباه در تشخیص حرف "آ"،
ح- اشتباه در تشخیص نقاط یا علائم.
خ- اشتباه در تطبیق نقاط با بدنه.

با توجه به خطاهای بوجود آمده، می‌توان با آموزش نمونه های بیشتری از علائم به شبکه عصبی، قرار دادن کلاه حرف "آ" به عنوان یکی از کلاسها و اصلاح الگوریتم بازشناسی نقاط بیشتر این خطاها را اصلاح کرد.

مراجع

[1] S. Mori, C. Y. Suen, and K. Yamamoto, "Histogram review of OCR research and development," in *Proc. of IEEE*, vol. 80, no. 7, pp. 1029-1058, Jul. 1992.
[2] A. Amin, A. Kaced, J. P. Haton, and R. Mohr, "Handwritten Arabic character recognition by the IRAC system," in *Proc. of the Fifth Int. Conf. on Pattern Recognition*, pp. 729-731, Miami Beach, FL, US, 1980.
[3] K. Badie and M. Shimura, "Machine recognition of Arabic cursive scripts," in *Proc. of Int. Workshop on Pattern Recognition in Practice*, pp. 315-323, Amsterdam, Netherlands, 1980.
[4] B. Parhami and M. Taraghi, "Automatic recognition of printed Farsi texts," *Pattern Recognition*, vol. 14, no. 1-6, pp. 395-403, 1981.
[5] T. K. Ho, J. J. Hull, and S. N. Srihari, "A word shape analysis approach to recognition of degraded word images," in *Proc. of the 4th USPS Advanced Technology Conference*, pp. 217-231, 1990.
[6] T. K. Ho, J. J. Hull, and S. N. Srihari, "A hypothesis testing approach to word recognition using dynamic feature selection," in *Proc. 11th Int. Conf. on Pattern Recognition*, pp. 586-589, 1992.
[7] W. Huang, C. Tan, S. Sung, and Y. Xu, "Word shape recognition for image-based document retrieval," in *Proc. of Int. Conf. on Image Processing (ICIP01)*, pp. 1114-1117, 2001.
[8] J. J. Hull and S. N. Srihari, "A computational approach to visual word recognition: hypothesis generation and testing," *Computer Vision and Pattern Recognition, IEEE*, pp. 156-161, 1986.
[9] J. J. Hull, "Hypothesis testing in a computational theory of visual word recognition," in *Proc. of the Sixth National Conf. on Artificial Intelligence (AAAI)*, pp. 718-722, Washington, 1987.

[10] T. K. Ho, J. J. Hull, and S. N. Srihari, "A computational model for recognition of multifont word images," *Machine Vision and Applications*, vol. 5, no. 3, pp. 157-168, Summer 1992.
[11] T. K. Ho, J. J. Hull, and S. N. Srihari, "Word recognition with multi-level contextual knowledge," in *Proc. of the First Int. Conf. on Document Analysis and Recognition*, pp. 905-915, Saint-Malo, France, 1991.
[12] A. L. Spitz, "Shape-based word recognition," *Int. Journal of Document Analysis and Recognition*, vol. 1, no. 4, pp. 178-190, May 1999.
[۱۳] ر. عزمی، *بازشناسی متون چاپی فارسی*، رساله دکتری مهندسی برق- الکترونیک، دانشگاه تربیت مدرس، ۱۳۷۸.
[۱۴] م. شیرعلی شهرضا و ک. فائز، *تشخیص کلمات و ارقام دستنویس فارسی بوسیله شبکه های عصبی(خط نسخ)*، رساله دکترای مهندسی برق- کامپیوتر، دانشگاه صنعتی امیر کبیر، ۱۳۷۴.
[15] E. J., Erlandson, J. M., Trenkle, and 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.
[16] M. S. Khorsheed and W. F, Clocksin, "Multi-Font Arabic word recognition using spectral features," in *Proc. of ICPR2000*, vol. 4, p. 4543, 2000.
[17] 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, May 2001.
[18] 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, Feb. 2001.
[۱۹] ک. مسروری، *شناسایی برون خط کلمات دستنویس فارسی در یک مجموعه محدود*، رساله دکتری مهندسی برق- الکترونیک، دانشگاه تربیت مدرس، تابستان ۱۳۷۹.
[۲۰] ا. ابراهیمی و ا. کبیر، "خوشه بندی تصاویر زیر- کلمات چاپی فارسی با استفاده از ویژگیهای مکان مشخصه و الگوریتم k- میانگین،" ارسال شده به مجله دانشکده فنی دانشگاه تبریز.
[۲۱] ا. ابراهیمی و ا. کبیر، "استفاده از یک روش دو مرحله ای برای طبقه بندی زیر- کلمات چاپی فارسی،" ششمین کنفرانس سیستمهای هوشمند، دانشگاه شهید باهنر کرمان، ۴-۵ آذرماه ۱۳۸۳.
[22] R. C. Gonzalez, *Digital Image Processing*, Addison-Wesley, 1972.

افشین ابراهیمی کارشناسی و کارشناسی ارشد خود را به ترتیب در مهندسی مخابرات از دانشگاه تبریز و مهندسی الکترونیک از دانشگاه تربیت مدرس در سالهای ۱۳۷۸ و ۱۳۸۰ دریافت کرد.

او هم اکنون در مقطع دکترای مهندسی الکترونیک دانشگاه تربیت مدرس مشغول به تحصیل است. زمینه های پژوهشی مورد علاقه او، بازشناسی الگو و پردازش تصویر است.

احسان اله کبیر کارشناسی ارشد پیوسته خود را در مهندسی برق و الکترونیک از دانشکده فنی دانشگاه تهران و دکترای خود را در مهندسی سیستمهای الکترونیک از دانشگاه اسکس در انگلستان، به ترتیب در سالهای ۱۳۶۴ و ۱۳۶۹ دریافت کرد.

او اکنون دانشیار بخش مهندسی برق دانشگاه تربیت مدرس است. زمینه های پژوهشی مورد علاقه او بازشناسی الگو، بویژه بازشناسی متون چاپی و دستنویس و بینایی ماشین است.