



دانشکده مهندسی برق و رباتیک

پایان نامه کارشناسی ارشد

شناسایی کلمات فارسی تاییبی در تصاویر درجه تفکیک پایین

حامد عرب یار محمدی

استاد راهنما:

دکتر علیرضا احمدی فرد

استاد مشاور:

دکتر حسین خسروی

زمستان ۱۳۹۲

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ



دانشکده مهندسی برق و رباتیک

گروه الکترونیک

شناسایی کلمات فارسی تایپی در تصاویر رزولوشن پایین

دانشجو: حامد عرب یار محمدی

استاد راهنما:

دکتر علیرضا احمد فرد

استاد مشاور:

دکتر حسین خسروی

پایان نامه ارشد جهت اخذ درجه کارشناسی ارشد

زمستان ۱۳۹۲

شماره : ۱۱۳۶ / ت س ب
تاریخ : ۹۲/۱۱/۲۸
ویرایش : —————

بسمه تعالی



فرم صورتجلسه دفاع پایان نامه تحصیلی دوره کارشناسی ارشد

با تأییدات خداوند متعال و با استعانت از حضرت ولی عصر (ع ج) جلسه دفاع از پایان نامه کارشناسی ارشد خاتم / آقای :
حامد عربیار محمدی رشته : برق گرایش : الکترونیک (سیستم)
تحت عنوان : شناسایی کلمات فارسی نابین در تصاویر درجه تکنیک پایین
که در تاریخ ۹۲/۱۱/۲۸ با حضور هیأت محترم داوران در دانشگاه صنعتی شاهرود برگزار گردید به شرح زیر است :

قبول (با درجه) : امتیاز : ۱۹/۵۰ دفاع مجدد مردود

۱- عالی (۲۰ - ۱۹) ۲- بسیار خوب (۱۸/۹۹ - ۱۸)

۳- خوب (۱۷/۹۹ - ۱۶) ۴- قابل قبول (۱۵/۹۹ - ۱۴)

۵- نمره کمتر از ۱۴ غیر قابل قبول

عضو هیأت داوران	نام و نام خانوادگی	مرتبه علمی	امضاء
۱- استاد راهنما	غفرین لهرستانی	استاد	
۲- استاد مشاور	سید محمد	استاد	
۳- نماینده شورای تحصیلات تکمیلی	امیررضا سرفراز	استادیار	
۴- استاد منتحن	حمید حسن پور	استاد	
۵- استاد منتحن	علیرضا	استادیار	

رئیس دانشکده :

تقدیم به تمام دوستان علم و دانش

رَأْسُ الْفَضَائِلِ الْعِلْمُ، غَايَةُ الْفَضَائِلِ الْعِلْمُ

سرآمد فضیلت‌ها، دانش است؛ نقطه پایان فضیلت‌ها دانش است

امیر المؤمنین علیؑ علیه السلام

تقدیر و تشکر

حمد بی حد و شنای بی حد و پاس بی قیاس خداوندی را که در سخن، ستودن او توان گفت و شمردن نعمت های او را نتوان بجای آورد و در دو بسیار و تحیات بی شمار بر بهترین آفریدگان حضرت محمد و خاندان پاکش...

مراتب پاس و قدردانی خویش را به

پدر دل سوز و مادر فداکارم، که همواره در تمام عرصه های زندگی یار و یاور بی چشم داشت برایم بوده اند

و

به جناب آقای دکتر احمدی فرد و جناب آقای دکتر خسروی که بارها بنیانی و مشاوره دقیق و مستمرشان، در شکل گیری این مجموعه نقش

مهمی داشتند

براز می دارم.

و از زحمات تمام اساتیدی که از محضر درس ایشان بهره جسته ام تشکر و قدردانی می کنم و پاسگزار یاری تمامی اعضای خانواده و

دوستان، هستم.

تعهد نامه

اینجانب حامد عرب یارمحمدی دانشجوی دوره کارشناسی ارشد رشته الکترونیک سیستم دانشکده

مهندسی برق و رباتیک دانشگاه صنعتی شاهرود نویسنده پایان نامه با عنوان :

شناسایی کلمات فارسی تایپی در تصاویر درجه تفکیک پایین

تحت راهنمایی جناب دکتر علیرضا احمدی فرد متعهد می شوم .

- تحقیقات در این پایان نامه توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است .
- در استفاده از نتایج پژوهشهای محققان دیگر به مرجع مورد استفاده استناد شده است .
- مطالب مندرج در پایان نامه تاکنون توسط خود یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ جا ارائه نشده است .
- کلیه حقوق معنوی این اثر متعلق به دانشگاه صنعتی شاهرود می باشد و مقالات مستخرج با نام « دانشگاه صنعتی شاهرود » و یا « Shahrood University of Technology » به چاپ خواهد رسید .
- حقوق معنوی تمام افرادی که در به دست آمدن نتایج اصلی پایان نامه تأثیرگذار بوده اند در مقالات مستخرج از پایان نامه رعایت می گردد.
- در کلیه مراحل انجام این پایان نامه ، در مواردی که از موجود زنده (یا بافتهای آنها) استفاده شده است ضوابط و اصول اخلاقی رعایت شده است .
- در کلیه مراحل انجام این پایان نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته یا استفاده شده است اصل رازداری ، ضوابط و اصول اخلاق انسانی رعایت شده است .

تاریخ

۱۳۹۳ / ۱۳ / ۱۲

امضای دانشجو

مالکیت نتایج و حق نشر

- کلیه حقوق معنوی این اثر و محصولات آن (مقالات مستخرج ، کتاب ، برنامه های رایانه ای ، نرم افزار ها و تجهیزات ساخته شده است) متعلق به دانشگاه صنعتی شاهرود می باشد . این مطلب باید به نحو مقتضی در تولیدات علمی مربوطه ذکر شود .
- استفاده از اطلاعات و نتایج موجود در پایان نامه بدون ذکر مرجع مجاز نمی باشد.

لیست مقالات مستخرج از پایان نامه :

- "Clustering Low Quality Farsi Sub-words For Word Recognition", Accepted At 12th Iranian Conference On Intelligent Systems, icis2014
- "پایگاه داده‌ی تصاویر کم کیفیت زیر-کلمات فارسی برای ارزیابی الگوریتم‌های بازشناسی"، پذیرفته شده در دوازدهمین کنفرانس سیستم‌های هوشمند ایران، زمستان ۱۳۹۲

برای بازشناسی زیرکلمات فارسی سه رویکرد مبتنی بر جداسازی به حروف، مبتنی بر شکل کلی زیرکلمه و ترکیبی از این دو وجود دارد. در بسیاری از سازمان‌ها، تصاویر نامه‌ها با درجه تفکیک ۱۵۰ نقطه بر اینچ و کمتر ذخیره می‌شوند. متنی با این درجه تفکیک برای خواندن توسط کاربر انسانی مناسب است اما برای بازشناسی توسط سیستم شناسایی نوری کلمات این درجه تفکیک بسیار کم به نظر می‌رسد. حتی سیستم‌های شناسایی نوری کلمات لاتین هم غالباً برای ۳۰۰ نقطه بر اینچ، توسعه یافته‌اند. در درجه‌ی تفکیک پایین، جداسازی به سختی امکان پذیر است و باید از روش‌هایی مانند شکل کلی استفاده نمود. در این پایان نامه به بررسی و بازشناسی زیرکلمات فارسی با درجه تفکیک ۹۶ نقطه بر اینچ می‌پردازیم که برای این منظور از شکل کلی زیرکلمات برای بازشناسی آنها بهره برده‌ایم.

سیستمی که در اینجا برای بازشناسی زیرکلمات فارسی ارائه شده مبتنی بر یک روش سه مرحله‌ای است. در مرحله نخست به کمک خوشه بندی، دامنه‌ی جستجوی تصاویر زیرکلمات موجود در فرهنگ لغت کاهش داده شده که این کار نه تنها سرعت سیستم را بالا می‌برد بلکه دقت را نیز افزایش می‌دهد، در مرحله دوم بازشناسی، با استفاده از یک طبقه‌بند، ۴ خوشه‌ی نزدیک به زیرکلمه‌ی آزمون ورودی تشخیص داده می‌شود و پس از آن با جستجو در میان آن خوشه‌های هدف، ۱۰ نزدیک‌ترین زیرکلمات موجود در فرهنگ لغت، به زیرکلمه‌ی آزمون را می‌یابیم، این روند برای تمام زیرکلمات یک کلمه تکرار شده سپس در مرحله‌ی سوم با استفاده از روش رخدادهای محتمل برای توالی زیرکلمات، کلمه‌ی آزمون تشخیص داده می‌شود.

دقت این الگوریتم بازشناسی بسیار مناسب تخمین زده می‌شود و قابلیت بازشناسی ۹۸/۰۱٪ در کلماتی که از بیش از یک زیرکلمه تشکیل شده‌اند را داراست و کلماتی که صرفاً، تک زیرکلمه‌ای اند را با دقت ۸۲/۵۳٪ بازشناسی می‌کند.

کلمات کلیدی:

بازشناسی زیرکلمه، شکل کلی زیرکلمه، تصاویر کم کیفیت، کاهش فرهنگ لغت، بلوک میانگین، k میانگین، k نزدیک‌ترین همسایه، الگوریتم POSS.

فصل ۱: مقدمه	۱
۱-۱- شناسایی الگو	۲
۲-۱- بازشناسی نویسه و تاریخچهی آن	۳
۱-۲-۱- مرحلهی تکوین: (از سال ۱۹۰۰ تا ۱۹۸۰)	۳
۲-۲-۱- مرحله توسعه: (از سال ۱۹۸۰ تا ۱۹۹۰)	۴
۳-۲-۱- مرحلهی بهبود: (از سال ۱۹۹۰ به بعد)	۴
۳-۱- تقسیم‌بندی سیستم‌های OCR	۵
۴-۱- چالش سیستم‌های OCR در وضوح پایین	۶
۵-۱- هدف کلی پایان نامه	۹
۶-۱- ساختار پایان نامه	۹
فصل ۲: مروری بر کارهای گذشته	۱۱
فصل ۳: مباحث نظری و استخراج ویژگی	۲۳
۱-۳- معیارهای فاصله	۲۴
۱-۱-۳- معیار فاصله مینکوسکی	۲۴
۲-۱-۳- معیار فاصله کسینوس	۲۵
۳-۱-۳- معیار فاصله چبی شف	۲۶
۴-۱-۳- معیار فاصله اسپیرمن	۲۶
۵-۱-۳- معیار فاصله همبستگی	۲۷
۲-۳- ویژگی‌ها	۲۷
۱-۲-۳- میانگین بلوکی	۲۷
۲-۲-۳- هیستوگرام گرادیان	۲۷

- ۳۰-۳-۲-۳- مکان مشخصه
- ۳۱-۴-۲-۳- سیگنال علامت
- ۳۱-۵-۲-۳- الگوهای دو دویی محلی
- ۳۲-۶-۲-۳- پروفایل بالا و پایین
- ۳۳-۷-۲-۳- افکنش
- ۳۴-۸-۲-۳- گذر سیاه به سفید
- ۳۴-۳-۳- دو سطحی سازی
- ۳۵-۱-۳-۳- روش پیشنهادی اُتسو
- ۳۸-۴-۳- تشخیص پهنای قلم
- ۳۸-۵-۳- عملگرهای شکل شناسی
- ۳۹-۱-۵-۳- عملگر ساییدگی و گسترش
- ۴۰-۶-۳- الگوریتم‌های خوشه یابی
- ۴۰-۱-۶-۳- الگوریتم K میانگین
- ۴۱-۲-۶-۳- الگوریتم سلسله مراتبی
- ۴۴-۳-۶-۳- الگوریتم ایزودیتا
- ۴۴-۷-۳- الگوریتم‌های طبقه بندی
- ۴۴-۱-۷-۳- نزدیک‌ترین همسایه k
- ۴۵-۲-۷-۳- شبکه‌های عصبی
- ۴۹- فصل ۴: الگوریتم پیشنهادی
- ۵۰-۱-۴- پایگاه داده
- ۵۰-۱-۱-۴- آماده سازی صفحات داده
- ۵۱-۲-۱-۴- تولید داده توسط روبشگر
- ۵۲-۳-۱-۴- روش مصنوعی تولید داده
- ۵۴-۴-۱-۴- پیش پردازش و بهبود کیفیت تصاویر

- ۴-۱-۵- جداسازی و برچسب گذاری زیر-کلمات ۵۶
- ۴-۱-۶- اطلاعات آماری ۵۶
- ۴-۱-۶-۱- پایگاه داده‌ی متنی شرکت هدی سیستم ۵۶
- ۴-۱-۶-۲- پایگاه داده‌ی تصاویر کم کیفیت زیرکلمات ۵۸
- ۴-۲- نرمال سازی برای استخراج ویژگی ۶۰
- ۴-۳- پارامترها در استخراج ویژگی ها ۶۱
- ۴-۳-۱- هیستوگرام گرادیان ۶۱
- ۴-۳-۲- میانگین بلوکی، مکان مشخصه، افکنش و LBP ۶۲
- ۴-۳-۳- سیگنال علامت ۶۳
- ۴-۴- انتخاب سازگارترین و بهترین ویژگی، یکسان ساز ابعاد، معیار فاصله و خوشه‌بند برای این گونه از تصاویر با وضوح پایین ۶۵
- ۴-۴-۱- گزارش نتایج ۶۶
- ۴-۴-۲- نتیجه گیری ۷۰
- ۴-۵- انتخاب طبقه بند ۷۱
- ۴-۵-۱- شبکه عصبی ۷۴
- ۴-۵-۲- K نزدیک ترین همسایه ۷۷
- ۴-۵-۲-۱- نزدیک ترین همسایه ۷۸
- ۴-۵-۲-۲- احتمال حضور زیرکلمه‌ی آزمون در m نزدیک ترین همسایه‌ی نزدیک به آن ... ۷۸
- ۴-۶- بازشناسی با معیارها و سنجش گره‌های انتخاب شده ۸۰
- فصل ۵: نتیجه گیری و پیشنهادات ۹۳
- ۵-۱- نتیجه گیری ۹۴
- ۵-۲- پیشنهادات ۹۵
- مراجع ۹۷

فهرست شکل‌ها

- شکل ۱-۱ : قسمتی از فرم درخواست کارت ملی، که با درجه تفکیک پایین ذخیره شده است. ۷.....
- شکل ۲-۱ : درصد خطای بازشناسی زیر-کلمه برای تفکیک‌پذیری‌های مختلف. ۸.....
- شکل ۱-۲ : طرز پیدا کردن نقاط جداساز [8]. ۱۳.....
- شکل ۲-۲ : فرایند بازشناسی با شکل کلی [۵]. ۱۴.....
- شکل ۳-۲ : معماری تخته سیاه برای بازشناسی متن بر پایه‌ی جداسازی [۵]. ۱۵.....
- شکل ۴-۲ : جداسازی رفت و برگشتی حروف در [۵] برای کاهش خطا در جداسازی. ۱۵.....
- شکل ۵-۲ : الگوریتم بازشناسی بر اساس [3]. ۱۶.....
- شکل ۶-۲ : تقسیم شدن بدنه‌ی کلمه به شبکه‌ی ۴۰ قسمتی در [9]. ۱۷.....
- شکل ۷-۲ : الگوریتم بازشناسی کلمات با شکل کلی در [9]. ۱۸.....
- شکل ۸-۲ : روش اول استفاده شده در [10]. ۱۹.....
- شکل ۹-۲ : روش دوم بازشناسی در [10]. ۲۰.....
- شکل ۱۰-۲ : روند بازشناسی روش سوم در [10]. ۲۰.....
- شکل ۱۱-۲ : فرایند بازشناسی در روش [۱۳]. ۲۲.....
- شکل ۱-۳ : در معیار فاصله مینکوسکی، برای $M = 1$ معیار فاصله بلوک شهر بدست می‌آید که در صفحه دارای مکان هندسی لوزی در فضای سه بعدی هرم و در فضاهای بالاتر یک ابر هرم است و به همین ترتیب $M = 2$ برای معیار اقلیدسی به ترتیب در فضای دو، سه و بالا تر شاهد مکان هندسی دایره، کره و ابر کره هستیم و در نهایت برای $M = \infty$ این مکان هندسی به مربع و مکعب و ابر مکعب ختم می‌شود [17]. ۲۵.....
- شکل ۲-۳ : مؤلفه‌های عملگرهای گرادیان سوبل و رابرت. ۲۸.....
- شکل ۳-۳ : نتیجه‌ی عملگر گرادیان با فیلتر رابرت بر روی تصویری با ابعاد 15×34 . ۲۹.....
- شکل ۴-۳ : نحوه استخراج ویژگی مکان مشخصه [۱۳]. ۳۰.....
- شکل ۵-۳ : استخراج ویژگی LBP [19]. ۳۱.....
- شکل ۶-۳ : شکل زیرکلمه و نمای پروفایل بالای آن. ۳۲.....
- شکل ۷-۳ : شکل زیرکلمه و نمای پروفایل پایین آن. ۳۳.....
- شکل ۸-۳ : شکل زیرکلمه و نمای افکنش افقی آن در سمت چپ و نمای افکنش عمودی آن در پایینش نشان داده شده است. ۳۴.....

- شکل ۳-۹ (الف) : یک تصویر نمونه است که روشنایی پس زمینه‌ی آن یکنواخت نیست، (ب) :
 هیستوگرام تصویر و آستانه‌ی بدست آمده با روش اُتسو [22] ۳۷
- شکل ۳-۱۰ : نتیجه‌ی باینری کردن شکل ۳-۹ با روش اُتسو از طریق آستانه گذاری سراسری [22] ۳۷
 شکل ۳-۱۱ (الف) : تصویر الف شکل ۳-۹ که جهت آستانه‌گذاری محلی قطعه‌بندی شده است، (ب) :
 نتیجه‌ی باینری کردن ، با روش اُتسو از طریق آستانه گذاری محلی است [22] ۳۷
- شکل ۳-۱۲ : تصویر (ج) نتیجه‌ی عملگر ساییدگی با عنصر ساختاری (ب) روی (الف) را نشان می‌دهد
 [24] ۳۹
- شکل ۳-۱۳ (الف) : نمونه ای از تصویر متن با کیفیت پایین که نویسه‌ها دچار شکستگی شده‌اند، (ب) :
 تصویر نتیجه‌ی اعمال عملگر گسترش روی تصویر (الف) با عنصر ساختاری لوزی [24] ۴۰
- شکل ۳-۱۴ (الف) : تعیین تصادفی مراکز اولیه به تعداد خوشه‌ها و مشخص کردن اینکه هر نقطه به
 کدام یک از این مراکز نزدیک‌تر است (ب) به روز کردن این مراکز (ج) یافتن مجدد نزدیک‌ترین مرکز
 به هر کدام از نقاط (د) با پایدار ماندن این مراکز الگوریتم پایان یافته است [25] ۴۱
- شکل ۳-۱۵ : سه دسته داده که با سلسله مراتبی خوشه بندی می‌شود [27] ۴۲
- شکل ۳-۱۶ : نمودار دندروگرام سه بعدی، که پیشرفت مراحل الگوریتم نمایان است (ابتدا خوشه
 بندی‌هایی که در ارتفاع پایین‌تر هستند انجام می‌شود) [27] ۴۳
- شکل ۳-۱۷ : نمونه‌ی نمودار دندروگرام دو بعدی که با متوقف کردن الگوریتم در ناحیه‌ی مشخص
 شده این سه دسته داده خوشه بندی شده‌اند [27] ۴۳
- شکل ۳-۱۸ : X نمونه‌ی وارد شده و هر شکل نماینده‌ی یک کلاس است جهت تعیین کلاس نقطه‌ی
 X در الگوریتم KNN شعاع همسایگی برای در بر گرفتن تعداد k همسایه در حال افزایش است. ... ۴۵
- شکل ۳-۱۹ : یک شبکه‌ی عصبی چهار لایه که دو لایه‌ی میانی و یک لایه‌ی ورودی و یک لایه‌ی
 خروجی دارد [30] ۴۶
- شکل ۳-۲۰ : تعیین مقدار خروجی در هر نود از شبکه [30] ۴۷
- شکل ۳-۲۱ : برخی از توابع فعال ساز به همراه روابطشان [30] ۴۷
- شکل ۴-۱ : دو تصویر با درجه تفکیک ۳۰۰ dpi و ۹۶ dpi ۵۱
- شکل ۴-۲ : یک تصویر نمونه از زیرکلمات مجموعه داده. ۵۱
- شکل ۴-۳ : نمایی از نرم افزار *doPDF* ۵۲
- شکل ۴-۴ : نمایی از نرم افزار *PDF2Image* ۵۳

- شکل ۴-۵: یک قطعه از تصویر تولیدشده‌ی زیر-کلمات با تفکیک‌پذیری ۹۶ نقطه در هر اینچ. (الف):
 به روش روبش از نسخه‌ی چاپی، (ب): به روش مصنوعی. ۵۴.....
- شکل ۴-۶: المان‌های ساختاری برای حذف نویز [23] ۵۵.....
- شکل ۴-۷: زیر-کلمه‌ی پانزدهم در خط ششم از صفحه شانزدهم، که با نام "۱۵_۶_۱۶" ذخیره شده است ۵۶.....
- شکل ۴-۸: آمارگان محل قرار گیری نقاط در زیر-کلمات نسبت به موقعیت خط کرسی ۵۸.....
- شکل ۴-۹: اطلاعات آماری پراکندگی زیر-کلمات برحسب اندازه‌ی آنها. (الف): برحسب طول، (ب): بر حسب عرض ۵۹.....
- شکل ۴-۱۰: یکسان‌سازی ابعاد زیر-کلمه‌ها. (الف): زیر-کلمه‌ی اصلی که برای نمایش بهتر با بزرگ‌نمایی نشان داده شده است، (ب): هم اندازه کردن زیر-کلمه به روش قرار دادن در مرکز قاب (ج): هم اندازه سازی با حفظ نسبت طول به عرض ۶۱.....
- شکل ۴-۱۱: نتیجه‌ی گرادیان با عملگر رابرت ۶۲.....
- شکل ۴-۱۲: بزرگ‌نمایی از یک قسمت از شکل ۴-۱۱ ۶۲.....
- شکل ۴-۱۳: ویژگی سیگنال علامت بدون وزن برای زیرکلمه‌ی "شکستگیها"، این ویژگی از سمت چپ از زیرکلمه استخراج شده است. ۶۴.....
- شکل ۴-۱۴: ویژگی سیگنال علامت وزن دار برای زیرکلمه‌ی "شکستگیها"، این ویژگی از سمت چپ از زیرکلمه استخراج شده است. ۶۴.....
- شکل ۴-۱۵: دیاگرامی از روش انجام کار ۶۶.....
- شکل ۴-۱۶: نمودار درصد، درست تشخیص دادن مراکز خوشه‌ها برای زیر-کلمه‌ها ۶۷.....
- شکل ۴-۱۷: پراکندگی زیر-کلمه‌ها در بین خوشه‌ها (الف): در روش خوشه‌بندی سلسله مراتبی، (ب): خوشه‌بندی به روش k میانگین ۶۹.....
- شکل ۴-۱۸: تعداد اعضای خوشه‌ها ۷۲.....
- شکل ۴-۱۹: تعداد کلاس هر یک از خوشه‌ها ۷۲.....
- شکل ۴-۲۰: تعدادی از اعضای خوشه‌ی یک ۷۳.....
- شکل ۴-۲۱: درصد حضور زیرکلمات آزمایش در m نزدیک‌ترین خوشه‌ی ادغام شده به زیرکلمه‌ها ۷۴.....
- شکل ۴-۲۲: نتایج انتخاب زیرکلمه توسط روش شبکه‌های مصنوعی، که در هر شکل بالاترین زیرکلمه سمت چپ زیرکلمه‌ی آزمون بوده و در سمت راست آن زیرکلمه‌ای که انتخاب شده آورده شده است. ۷۶.....

- شکل ۴-۲۳: افزایش زیرکلمات آموزش در بین خوشه‌ها..... ۷۷
- شکل ۴-۲۴: افزایش تعداد نمونه‌های جایگزین احتمالی به زیرکلمه‌ی آزمون برای بالا بردن احتمال تشخیص زیرکلمه ۷۸
- شکل ۴-۲۵: حضور نداشتن زیرکلمه‌ی نمونه‌ی آزمون در نزدیک‌ترین همسایه که با انتخاب نزدیک‌ترین همسایه در بازشناسی خطا رخ داده است..... ۸۰
- شکل ۴-۲۶: مثالی از الگوریتم POSS که زیرکلمه‌ی اول شناسایی شده و ترکیب‌های مختلف دو زیرکلمه‌ی پشت سر هم برای پیدا کردن معمول‌ترین حالت بررسی می‌شود. ۸۳
- شکل ۴-۲۷: مثال دیگری از الگوریتم POSS که زیرکلمه‌ی دوم باز شناسایی شده و ترکیب‌های مختلف دو زیرکلمه‌ی پشت سر هم برای پیدا کردن معمول‌ترین حالت بررسی می‌شود. ۸۴
- شکل ۴-۲۸: کلمه‌ای شامل ۳ زیرکلمه که ترکیب‌های دو به دوی زیرکلمات پشت سر هم در تصویر آمده است. ۸۶
- شکل ۴-۲۹: ترکیب دو به دوی زیرکلمات در کلمه‌ای با سه زیرکلمه. ۸۸
- شکل ۴-۳۰: ترکیب دو به دوی زیرکلمات در کلمه‌ای با سه زیرکلمه. ۹۰

فهرست جداول

- جدول ۱-۴ : اطلاعات آماری اولیه‌ی زیر-کلمات قبل از تصحیح [۳۱]..... ۵۷
- جدول ۲-۴ : تعدادی از پر تکرارترین و کم‌تکرارترین زیر-کلمات [۳۱]..... ۵۷
- جدول ۳-۴ : درصدهای پنج رتبه‌ی اول در تشخیص درست خوشه، تا انتخاب ده خوشه‌ی نزدیک به زیر-کلمه‌ها ۶۹
- جدول ۴-۴ : تعداد رخدادهای ترکیبی دو زیر کلمه در شکل ۴-۲۶ که بالاترین مقادیر مشخص شده‌اند..... ۸۳
- جدول ۵-۴ : تعداد رخدادهای ترکیبی دو زیر کلمه در شکل ۴-۲۷..... ۸۵
- جدول ۶-۴ : تعداد رخداد ترکیب زیر کلمات اول و دوم در شکل ۴-۲۸، (بالاترین مقادیر مشخص شده‌اند)..... ۸۷
- جدول ۷-۴ : تعداد رخداد ترکیب زیر کلمات دوم و سوم در شکل ۴-۲۸، (بالاترین مقادیر مشخص شده‌اند)..... ۸۷
- جدول ۸-۴ : تعداد رخداد ترکیب زیر کلمات اول و دوم در شکل ۴-۲۹، (بالاترین مقادیر مشخص شده‌اند)..... ۸۸
- جدول ۹-۴ : تعداد رخداد ترکیب زیر کلمات دوم و سوم در شکل ۴-۲۹، (بالاترین مقادیر مشخص شده‌اند)..... ۸۹
- جدول ۱۰-۴ : تعداد رخداد ترکیب زیر کلمات اول و دوم در شکل ۴-۳۰، (بالاترین مقادیر مشخص شده‌اند)..... ۹۰
- جدول ۱۱-۴ : تعداد رخداد ترکیب زیر کلمات دوم و سوم در شکل ۴-۳۰، (بالاترین مقادیر مشخص شده‌اند)..... ۹۱

فصل ۱:

مقدمه

۱-۱- شناسایی الگو

یک الگو^۱ در واقع مدل یا قالبی است که با تکرار عجین و سرشته شده باشد و این تکرارها طبق قوانینی است که دانش شناسایی آن، به شناسایی الگو معروف است؛ شناسایی الگو شاخه‌ای از هوش مصنوعی است که با دریافت داده‌های خام و طبقه‌بندی آنها به توصیف مشاهدات می‌پردازد. شناسایی الگو به ما کمک می‌کند تا داده‌ها را با تکیه بر اطلاعات آماری استخراج شده از الگوها، طبقه‌بندی نماییم.

یک سیستم شناسایی الگو متشکل است از یک حسگر که مشاهداتی را که باید توصیف یا کلاسه‌بندی گردند جمع‌آوری می‌نماید، یک مکانیزم استخراج ویژگی که اطلاعات عددی یا نمادین را از مشاهدات محاسبه می‌کند و یک طبقه‌بند که با تکیه بر ویژگی‌های استخراج شده عهده‌دار شناسایی است؛ پس از فراهم شدن مشاهدات، توسط درگاه‌های همچون روبشگر و دوربین و... دومین مسئله، استخراج ویژگی‌ها یا صفات خاصی از داده‌های ورودی دریافته شده و تشکیل بردارهای الگوست. این مورد اغلب به عنوان مسئله‌ی پیش‌پردازش و استخراج ویژگی معرفی می‌شود، مسئله‌ی سوم در طراحی سیستم تشخیص الگو تعیین رویه‌های تصمیم‌گیری است که در فرآیند شناسایی و دسته‌بندی مورد نیاز واقع می‌شود. پس از آنکه داده‌های مشاهده شده از الگوها جمع‌آوری شد و در فرم نقاط الگو یا بردارها در فضای الگو بیان شد، ماشین باید تصمیم بگیرد که این داده به کدام کلاس الگو تعلق دارد.

بازشناسی الگو در بسیاری از زمینه‌ها نقش کاربردی دارد. بازشناسی نویسه، بازشناسی نویسنده برای تصدیق امضا، طبقه‌بندی اثر انگشت و بازشناسی گفتار نمونه‌هایی از این کاربردها هستند.

^۱ Pattern

۱-۲- بازشناسی نویسه و تاریخچه‌ی آن

هزینه بالای نگهداری و کار با پایگاه‌های داده بزرگ از تصاویر اسناد و آرشیوها، نیاز به روش‌های خودکار برای دستیابی به اطلاعات داخل این تصاویر را بوجود آورده است. اهمیت بازشناسی نویسه‌ها در حرکت به سوی ادارات بدون کاغذ، که حجم بزرگی از اسناد کاغذی روبش^۱ شده دارند و آنها را به صورت تصویری ذخیره می‌کنند پیدا می‌شود.

نویسه‌خوانی نوری^۲ عبارت است از بازشناسی خودکار متون موجود در تصاویر اسناد و تبدیل آنها به متون قابل جستجو و ویرایش توسط رایانه که تصویر سند غالباً توسط روبشگر و یا دوربین دیجیتال تولید می‌شود و شامل تعدادی پیکسل با رنگ‌های مختلف و سطوح روشنایی گوناگون است. از دید انسان، یک سند ممکن است ارزش اطلاعاتی زیادی داشته باشد، اما از دید رایانه تصویر یک سند با تصویر یک منظره تفاوتی ندارد، چرا که هر دوی آنها مجموعه‌ای از پیکسل‌ها هستند. برای اینکه بتوان از اطلاعات نوشتاری تصویر سند استفاده کرد، باید به نحوی نوشته‌های موجود در سند را بازشناسی کرد. چنین کاری توسط سیستم‌های نویسه‌خوان نوری انجام می‌شود. از جنبه تاریخی سیستم‌های OCR تا کنون سه مرحله تکاملی را پشت سر گذاشته‌اند:

۱-۲-۱- مرحله‌ی تکوین: (از سال ۱۹۰۰ تا ۱۹۸۰)

رد پای اولیه‌ی اقدامات صورت گرفته در زمینه‌ی بازشناسی حروف را در سال‌های اول دهه‌ی ۱۹۰۰ می‌توان یافت و آن زمانی است که تیورینگ^۳ دانشمند روسی بر آن بود که به افراد مبتلا به نارسایی‌های بینایی کمک نماید و اولین اختراع‌های ثبت شده در این زمینه مربوط به سال‌های ۱۹۲۹ و ۱۹۳۳ هستند.

^۱ Scanned

^۲ Optical Character Recognition (OCR)

^۳ Turing

این سیستم‌ها، حروف چاپی را با روش تطابق قالبی^۱ شناسایی می‌کردند. ماسک‌های مکانیکی مختلفی از مقابل تصویر حرف عبور می‌کردند و نور از یک سو به آن تابانده شده و از سوی دیگر توسط یک آشکارساز نوری دریافت می‌شد. این اختراع به دلیل تکنولوژی اپتیومکانیکی مورد استفاده از آنها کاربردی نبود و تصور دسترسی به دستگاهی برای بازشناسی حروف تا دهه ۱۹۴۰ میلادی و ظهور کامپیوترهای دیجیتال به صورت یک رؤیا باقی ماند.

در این دوره تحقیقاتی موفق اما مقید بیشتر بر روی حروف و اعداد لاتین انجام گرفت با این وجود مطالعات چندی نیز در زمینه‌ی حروف ژاپنی، چینی، عبری، هندی و عربی آغاز گردید.

۱-۲-۲- مرحله توسعه: (از سال ۱۹۸۰ تا ۱۹۹۰)

مطالعات صورت گرفته تا قبل از سال ۱۹۸۰ از فقدان سخت افزارهای کامپیوتری قدرتمند و دستگاه‌های اخذ داده‌ها رنج می‌بردند. در این دهه به واسطه‌ی رشد انفجار گونه‌ی تکنولوژی اطلاعات، وضعیت بسیار مناسبی برای زمینه‌های تحقیقاتی مختلف از جمله بازشناسی حروف فراهم گردید. در این مرحله نرخ بازشناسی نسبت به دوره‌ی قبل رشد قابل ملاحظه‌ای داشت ولی نتوانست از یک حد فراتر رود، که در بسیاری از کاربردهای OCR قابل قبول نبود.

۱-۲-۳- مرحله‌ی بهبود: (از سال ۱۹۹۰ به بعد)

در این مقطع زمانی بود که با تکوین ابزارها و تکنیک‌های پردازشی جدید، پیشرفت واقعی سیستم‌های OCR محقق گردید. در اوایل دهه‌ی ۱۹۹۰ روش‌های پردازش تصویر و بازشناسی الگو با تکنیک‌های کارآمد هوش مصنوعی ادغام گشتند.

محققان الگوریتم‌های بازشناسی حروف پیچیده را ابداع نمودند که قادر بودند داده‌های ورودی با تفکیک پذیری بالا را دریافت کنند و در مرحله‌ی پیاده سازی، محاسبات بسیار زیادی را بر روی داده

^۱ Template Matching

انجام دهند.

۱-۳- تقسیم‌بندی سیستم‌های OCR

در یک تقسیم کلی می‌توان سیستم‌های OCR را به لحاظ نوع الگوی ورودی به دو گروه اصلی تقسیم کرد:

الف) سیستم‌های بازشناسی متون چاپی

ب) سیستم‌های بازشناسی متون دست‌نویس

که هر یک از این گروه‌ها تکنیک‌های خاص خود را دارند. همچنین از جنبه‌ی نحوه‌ی ورود اطلاعات، سیستم‌های OCR به دو دسته زیر تقسیم می‌شوند:

الف) سیستم‌های برخط^۱

در بازشناسی برخط، حروف در همان زمان نگارش توسط سیستم تشخیص داده می‌شوند. دستگاه‌های ورودی این سیستم‌ها یک قلم نوری و یا صفحات لمسی است. در این روش علاوه بر اطلاعات مربوط به موقعیت قلم، اطلاعات زمانی مربوط به مسیر قلم نیز در اختیار می‌باشد. در این روش می‌توان از اطلاعات زمانی، سرعت، شتاب، فشار و زمان برداشتن و گذاشتن قلم روی صفحه در بازشناسی استفاده کرد.

ب) سیستم‌های برون خط^۲

در بازشناسی برون خط از تصویر دو بعدی متن ورودی استفاده می‌شود. در این روش به هیچ نوع وسیله نگارش خاصی نیاز نیست و تفسیر داده مستقل از فرایند تولید آنها بر اساس تصویر متن صورت می‌گیرد. این روش به نحوه‌ی بازشناسی توسط انسان شباهت بیشتری دارد.

^۱ On Line

^۲ Out Line

۱-۴- چالش سیستم‌های OCR در وضوح پایین

استفاده از سیستم‌های OCR دو مزیت عمده دارد :

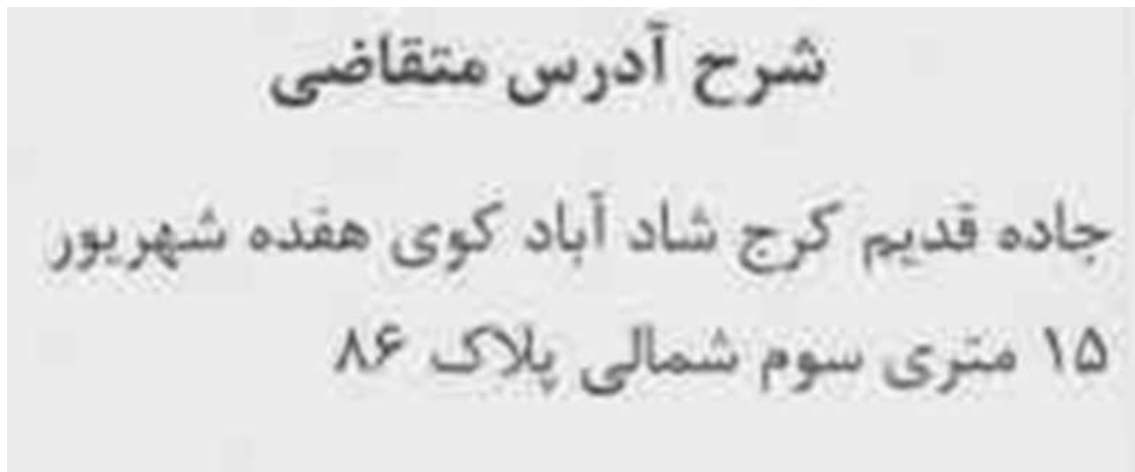
الف) افزایش چشمگیر سرعت دسترسی به اطلاعات؛ زیرا در متن بر خلاف تصویر، امکان جستجو و ویرایش وجود دارد.

ب) کاهش فضای ذخیره سازی؛ زیرا حجم فایل متنی استخراج شده از یک تصویر، معمولاً بسیار کمتر از حجم فایل خود تصویری است.

چنین قابلیت‌هایی امکان پردازش سریع حجم وسیعی از داده‌های مکتوب تولید شده توسط اداره‌ها، شرکت‌ها و مؤسسات مختلف را فراهم می‌آورد.

نیاز به بازشناسی نویسه‌ها در تصاویر با کیفیت پایین، از آنجا ناشی می‌شود که برای کاهش حجم ذخیره‌شده‌ی اسناد و بایگانی کردن آنها، نوعاً سازمان‌ها اسناد را با کیفیت پایین رویش و ذخیره می‌کنند. به طور مثال تصاویر مربوط به اطلاعات متقاضی کارت ملی یا نامه‌های اداری در سازمان‌ها با درجه تفکیک ۱۰۰ یا ۱۵۰ نقطه بر اینچ^۱ ذخیره می‌شوند. این درجه تفکیک برای خواندن متن توسط کاربر انسانی مناسب است اما برای بازشناسی توسط سیستم شناسایی نوری کلمات بسیار کم است، شکل ۱-۱ نمونه‌ی قسمتی از فرم درخواست کارت ملی یک متقاضی است.

^۱ Dots Per Inch (dpi)

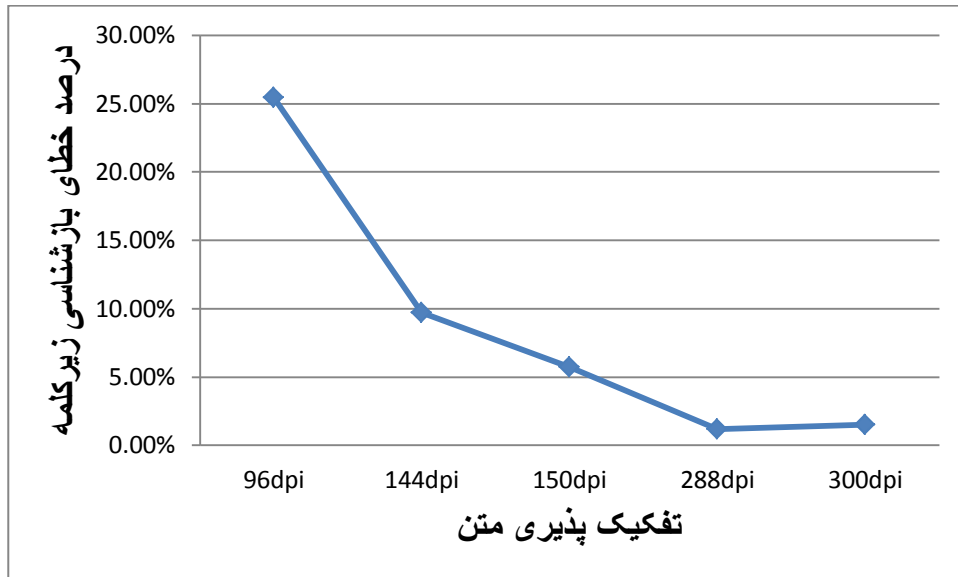


شکل ۱-۱: قسمتی از فرم درخواست کارت ملی، که با درجه تفکیک پایین ذخیره شده است.

برای تبدیل این اسناد به صورتی که قابلیت جستجو داشته باشند باید بتوان کلمات آنها را بازنشاسی کرد. در این صورت نیاز به سیستمی که قابلیت بازنشاسی در چنین شرایطی را دارا باشد ضرورت می‌یابد.

دو روش برای بازنشاسی کلمات وجود دارد، در بعضی روش‌ها، کلمه را به حروف آن شکسته و سپس هر یک از حروف را بازنشاسی می‌کنند [1,2]، یا بر مبنای شکل کلی تصویر، به بازنشاسی آن می‌پردازند [3,4]، برخی روش‌ها از ترکیب این دو استفاده می‌کنند.

در یک آزمایش، از روش ارائه‌شده در [۵] برای شناسایی کلمه‌های یک پایگاه داده در چند درجه تفکیک مختلف استفاده نمودیم. این روش مبتنی بر جداسازی حروف، و شناسایی آنها، زیر-کلمات را شناسایی می‌کند. خطا در شناسایی زیر-کلمه مطابق نمودار شکل ۱-۲ در ۳۰۰ dpi، حدود ۳ درصد و برای درجه تفکیک ۹۶ dpi، بیش از ۲۵ درصد است.



شکل ۱-۲: درصد خطای بازشناسی زیر-کلمه برای تفکیک پذیری‌های مختلف

مطالعه نتیجه‌ی حاصله نشان می‌دهد که تجزیه‌ی کلمه‌ها به حروف تشکیل‌دهنده‌شان در تصاویر با کیفیت پایین همراه با خطای زیاد است زیرا پیدا کردن محل تقاطع حروف دشوار است. لذا می‌بایست شناسایی را بر مبنای شکل کلی زیر-کلمه انجام داد، چرا که در زیر-کلمات چون که یا ترکیبی از حروف هستند که بدنه‌ی واحد دارند و یا حروف، بدون چسبندگی به یکدیگر ظاهر می‌شوند، بدون نیاز به جداسازی حروف، بازشناسی بر روی شکل زیرکلمه انجام می‌شود، به عنوان مثال کلمه‌ی "فرهیخته" از دو زیر-کلمه‌ی "فر" و "هیخته" تشکیل شده و یا کلمه‌ی "خوبی" شامل زیرکلمه‌ی "خو" و "بی" است.

همان‌طور که از شکل ۱-۲ پیداست، بازشناسی در درجه تفکیک پایین یک چالش محسوب می‌شود. و کمبود یک پایگاه داده مناسب در این زمینه نیز کاملاً احساس می‌شود. در اینجا با پیشنهاد یک پایگاه داده در وضوح پایین به ارائه‌ی روشی برای بازشناسی کلمات با این سبک می‌پردازیم.

۱-۵- هدف کلی پایان نامه

هدف از این پایان نامه بررسی استفاده از شکل کلی زیر-کلمات جهت بازشناسی آنهاست. برای توصیف شکل زیر-کلمات اطلاعات کلی و جزئیات شکل آنها بررسی می‌شود و از بین توصیفگرها کارآمدترینشان برگزیده شده تا در مراحل بعدی از آنها استفاده شود. برای تسریع مقایسه زیر-کلمه ناشناخته ورودی با دیکشنری، از روش خوشه بندی استفاده شده است که در هر خوشه نماینده‌ای برای اعضای آن در نظر گرفته می‌شود این کار برای کاهش فضای جستجو را صورت می‌گیرد.

در مرحله‌ی بعد برای بازشناسی زیرکلمه‌ی آزمون با استفاده از طبقه بندها، تعدادی از نزدیک‌ترین خوشه‌ها به زیرکلمه را یافته و نمونه‌ی ورودی در این فضای جدید جستجو می‌شود و در مرحله‌ی آخر با ترکیب طبقه بندها و روش نوین ارائه شده کلمات شناسایی می‌شوند.

۱-۶- ساختار پایان نامه

در فصل دوم به مرور مختصری بر کارهای انجام شده در این زمینه می‌پردازیم. فصل سوم به کلیات مربوط به روش‌های استخراج ویژگی، خوشه بندی و طبقه بندی اختصاص دارد، در این فصل مباحث نظری مربوط به فصل‌های آتی نیز آورده شده است. در فصل چهارم پایگاه داده ایجاد شده در این پایان نامه معرفی خواهد شد در این فصل به بررسی روش‌های خوشه بندی و طبقه بندی زیرکلمات و نتایج ارزیابی آنها پرداختیم که برای این‌گونه تصاویر خوشه بند k میانگین در دسته بندی داده‌ها و k نزدیک‌ترین همسایه در طبقه بندی آنها جواب بهتری دادند، و در نهایت این فصل با معرفی الگوریتم POSS و بازشناسی نمونه‌های آزمایش بر مبنای شکل کلی با این الگوریتم، پایان می‌یابد و فصل پنجم به نتیجه گیری و ارائه‌ی پیشنهادات اختصاص داده شده است.

فصل ۲:

مروری بر کارهای گذشته

در این فصل به بررسی برخی از سیستم‌های بازشناسی متون چاپی لاتین، عربی و فارسی ارائه شده می‌پردازیم.

سیستمی که توسط [1] ارائه شده تصاویر با درجه تفکیک ۳۰۰ dpi را بازشناسی می‌کند، این سیستم بر مبنای جداسازی به حروف سازنده‌ی یک کلمه اقدام به بازشناسی می‌کند این جداسازی با استفاده از افکنش عمودی انجام می‌شود و از ویژگی کدهای فریمنی استفاده می‌کند که از روی کانتور هر حرف استخراج می‌شود؛ سپس نزدیک‌ترین بردار ویژگی به آن انتخاب می‌شود دقت این روش ۹۰ درصد گزارش شده است.

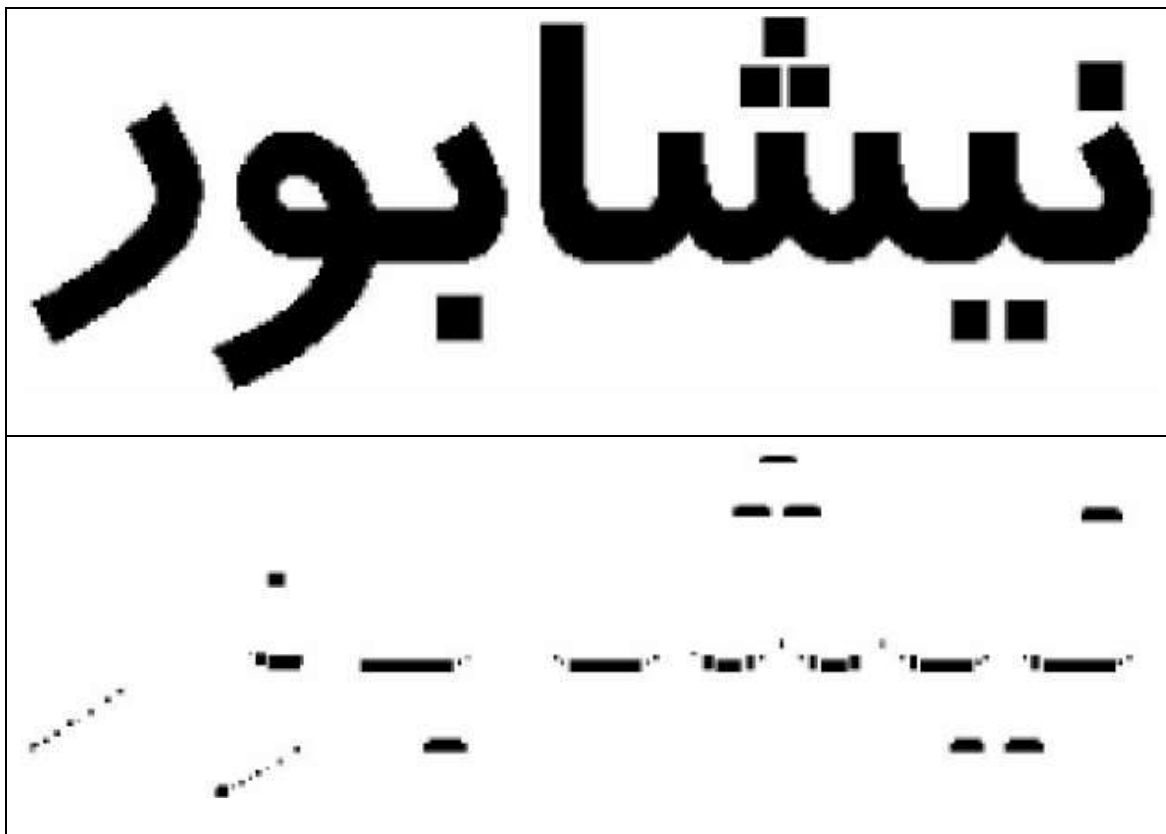
در یکی دیگر از سیستم‌ها [2] پس از جداسازی کلمه به حروف تشکیل دهنده‌ی آن، توسط شبکه‌های عصبی این بازشناسی حروف صورت می‌گیرد؛ در این روش نیز برای جداسازی از روش افکنش عمودی استفاده می‌کند و ویژگی مورد استفاده گشتاورهای پایدار در برابر چرخش و تغییر مقیاس هستند.

در [6] بازشناسی کلمات بر روی سه قلم مختلف صورت گرفته است. در این روش از ویژگی‌های شکل کلی کلمات چاپی در بازشناسی متون عربی استفاده شده است. این ویژگی‌ها شامل نقاط، همزه، جهت پاره خط‌ها، نقاط انتهایی و اتصال‌ها، حفره‌ها هستند و از تصویر کلمات استخراج و در یک دیکشنری ذخیره می‌شود. این دیکشنری شامل ۴۸۲۰۰ کلمه است. برای تصویر کلمه ورودی بردارهای ویژگی از آن استخراج شده 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.

در روش بازشناسی مطرح شده در [7] از روش جداسازی استفاده شده است. از توصیفگرهای فوریه به عنوان ویژگی و از شبکه‌ی عصبی چند لایه به عنوان طبقه‌بند استفاده کرده؛ برای جداسازی از برچسب زنی به کانتور استفاده شده است. این روش روی سه قلم تیترا، زر و تایمز نیو رومن

آزمایش شده است و برای تصاویر اسکن شده با درجه تفکیک ۴۰۰ dpi، دقت ۰.۹٪ گزارش شده است.

در تحقیق صورت گرفته در [8] از تبدیل موجک برای جداسازی حروف بهره برده شده است؛ از این تبدیل برای آشکارسازی لبه‌های افقی پایین کلمات و یافتن خط پایه استفاده شده و پس از یافتن لبه‌های پایین، با افکنش افقی گرفتن روی آنها موقعیت نقاط جداسازی تعیین می‌شود (شکل ۱-۲).



شکل ۱-۲: طرز پیدا کردن نقاط جداساز [8]

برای بازشناسی حروف پس از جداسازی از شبکه‌ی عصبی استفاده شده است. این سیستم روی

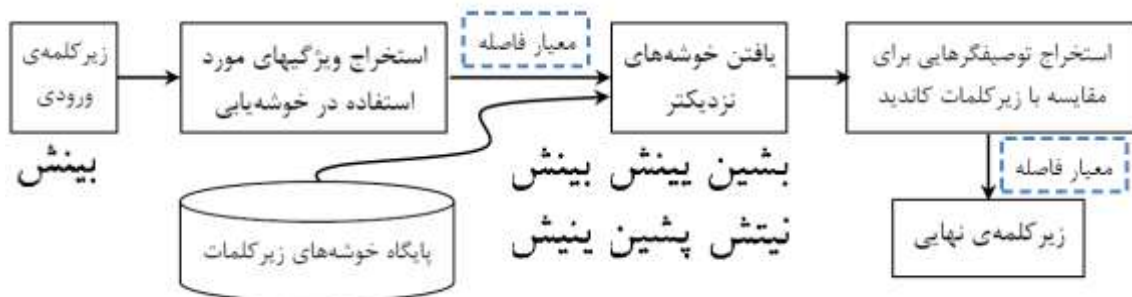
کلمه‌ی چاپی با قلم ترافیک آزمایش شده و دقت ۹۷/۸۳٪ گزارش شده است.

در [۵] یک سیستم یکپارچه برای بازشناسی برای ۱۰ قلم ارائه شده که در آن هم به بررسی با

روش جداسازی پرداخته و هم بازشناسی با شکل کلی زیرکلمه، در این رساله از ویژگی‌های مکان

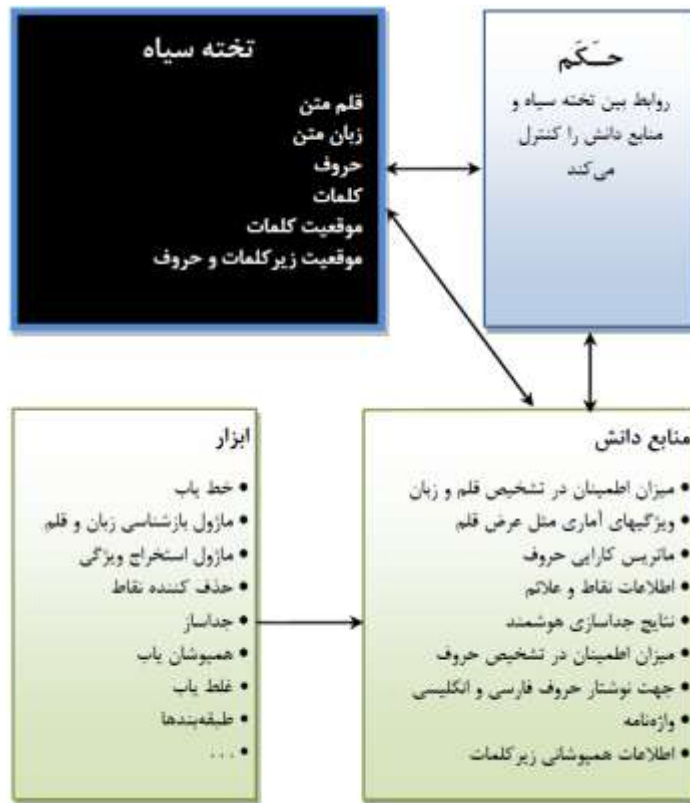
مشخصه، گرادیان، افکنش و بلوک بندی استفاده و معیار فاصله‌ی بلوک شهر بکار گرفته شده است. در

بازشناسی با شکل کلی از طبقه بند KNN استفاده شده و این بازشناسی در دو حالت در نظر گرفتن زیرکلمات، با نقطه و بدون نقطه مورد بررسی قرار گرفته است، این کار برای کاهش فضای حافظه انجام پذیرفته؛ در حالت با نقطه نتایج به دست آمده بهتر است، ضمن اینکه حافظه‌ی لازم برای اجرای برنامه نیز کمتر می‌شود. از مزایای شکل کلی با نقطه این است که قابلیت بازشناسی زیرکلماتی که نقاط آنها به بدنه چسبیده‌اند را دارد. از سوی دیگر در حالت بدون نقطه، تعداد زیرکلمات مجموعه داده، حدوداً نصف شده و فرایند خوشه‌یابی و بازشناسی آسان‌تر است. این بازشناسی برای تصاویر ۳۰۰ dpi برای هر دو الگوریتم در بازشناسی متون تمیز، دقت یکسانی، حدود ۹۳٪، دارند. در مورد تصاویری که نویز اندکی دارند نتایج بازشناسی به شدت افت می‌کند. دیاگرام بازشناسی به روش شکل کلی زیرکلمه در شکل ۲-۲ آورده شده است.

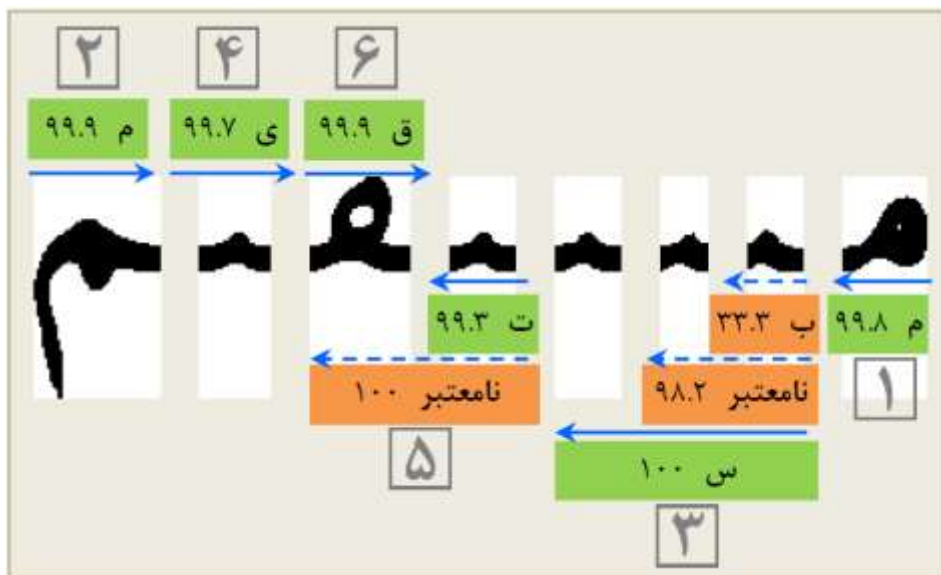


شکل ۲-۲: فرایند بازشناسی با شکل کلی [۵]

در بازشناسی به روش جداسازی به حروف از یک روش هماهنگ کننده به نام تخته سیاه کمک گرفته (شکل ۲-۳) و از طبقه بندی شبکه‌های عصبی استفاده کرده است برای جدا سازی حروف برای کاهش خطا در بازشناسی حروفی فارسی که تعداد زیاد دندانده دارند پس از یافتن نقاط جداساز با روش افکنش افقی تا سه مرحله امکان تصحیح خطا برای حالتی که درصد بازشناسی مناسب نباشد قرار داده شده که رفت و برگشتی است در شکل ۲-۴ این روش نشان داده شده است، این روش نیز برای اسناد ۳۰۰ dpi دقت ۹۷/۰۵٪ دارد.



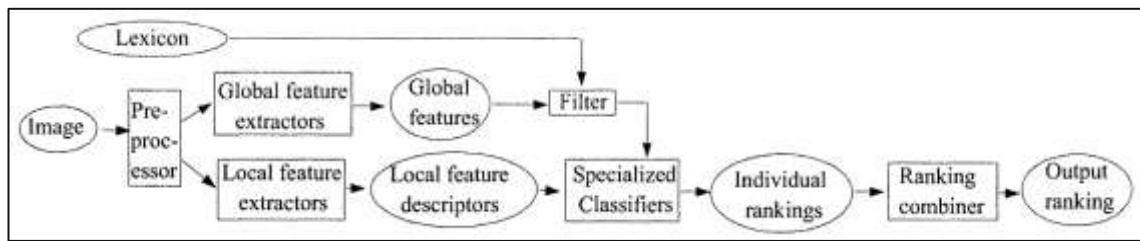
شکل ۲-۳: معماری تخته سیاه برای بازشناسی متن بر پایه‌ی جداسازی [۵]



شکل ۲-۴: جداسازی رفت و برگشتی حروف در [۵] برای کاهش خطا در جداسازی

در [3] از ویژگی‌های شکل کلمات برای بازشناسی آنها استفاده شده است. تصویر کلمه ورودی

ابتدا از یک الگوریتم پیش پردازش عبور داده می‌شود که در آن مراحل حذف خط زیر کلمه، حذف علائم، و تخمین خط مرجع^۱ انجام می‌شود. سپس ویژگی‌های کلی و محلی از تصویر کلمه استخراج شده و واژه نامه با ویژگی‌های کلی محاسبه شده فیلتر می‌شود. واژه نامه فیلتر شده و توصیفگرهای ویژگی‌های محلی^۲ محاسبه شده به مجموعه ای از طبقه بندهای داده می‌شوند. الگوریتم این روش در شکل ۲-۵ آمده است. ویژگی‌های کلی استفاده شده شامل تخمینی از طول کلمه، حفره‌ها و نحوه قرار گرفتن کلمه نسبت به خط زمینه است. ویژگی‌های محلی جزئیات شکل حروف کلمه را توصیف می‌کنند و شامل توزیع پاره خط‌ها^۳، ویژگی‌های لبه، ویژگی‌های نقاط انتهایی^۴ پاره خط‌ها هستند.



شکل ۲-۵: الگوریتم بازشناسی بر اساس [3]

در تحقیق دیگری [9] از شکل کلی کلمات و یک دیکشنری تصویری برای بازشناسی آنها استفاده شده است. تصویر هر کلمه به شبکه‌ی 4×10 قسمت، تقسیم می‌شود. برای تقسیم شدن به این شبکه ابتدا کلمه آغشته^۵ سازی شده و پس از افکنش افقی خط اصلی پایین^۶ و خط اصلی بالا^۷ تعیین می‌شوند. پس تعیین این دو خط کلمه به ۴ ناحیه‌ی افقی تقسیم شده و سپس از لحاظ عمودی کلمه به ۱۰ قسمت ثابت تقسیم می‌شود. طریقه تقسیم شدن کلمه به این ۴۰ قسمت در شکل ۲-۶ آورده شده است.

¹ Reference line estimation

² Local feature

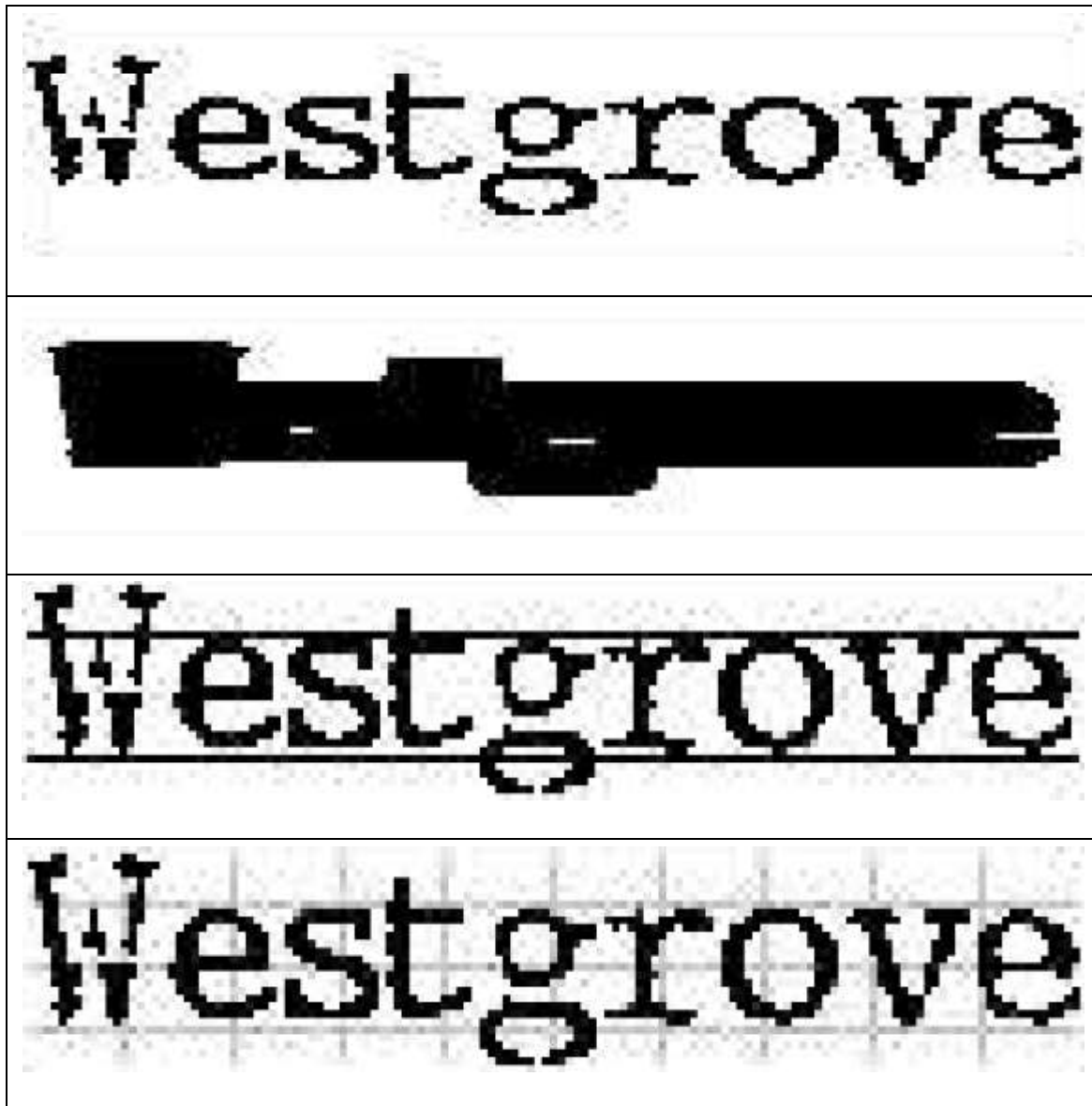
³ Storke distribution

⁴ End point featurese

⁵ Smeared

⁶ Base line

⁷ Top line



شکل ۲-۶: تقسیم شدن بدنه‌ی کلمه به شبکه‌ی ۴۰ قسمتی در [9]

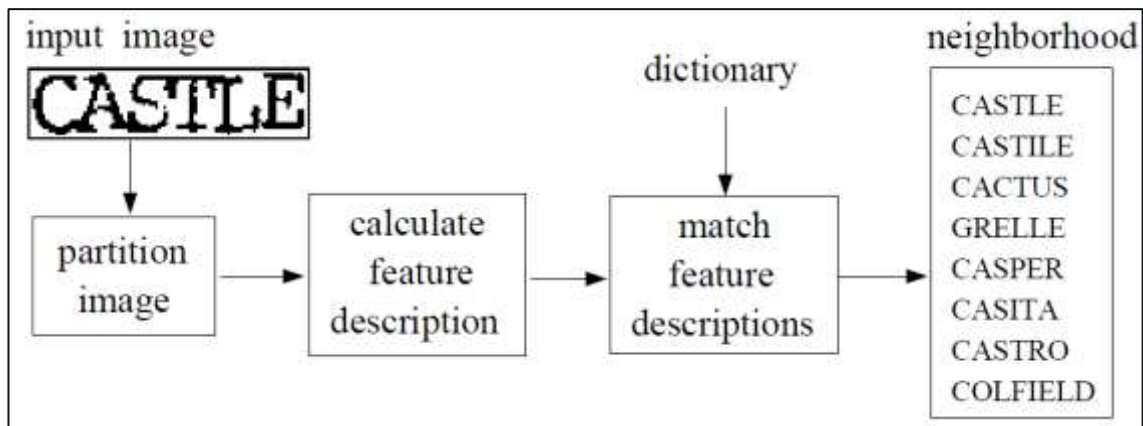
و سپس ویژگی‌های توزیع جهت پاره خط^۱ از این تصویر استخراج شده و مکان نسبی آنها در این شبکه در یک بردار ویژگی ذخیره می‌شود. این بردار ویژگی با بردارهای دیکشنری مقایسه شده و یک رتبه بندی از کلمات تولید می‌شود.

برای استخراج ویژگی هر نقطه سیاه از تصویر جهتی که تعداد نقاط سیاه در آن جهت، در داخل خانه‌ای که حول آن نقطه قرار گرفته، بیشترین است، به عنوان برجسب آن نقطه تعیین می‌شود. در هر

^۱ Storke direction distribution

کدام از ۴۰ خانه تصویر کلمه، تعداد نقاط سیاه با برچسب جهات شمارش شده و به تعداد نقاط تصویر تقسیم می‌شوند. بدین ترتیب از کنار هم قرار دادن این مقادیر، برداری ویژگی برای هر کلمه بدست می‌آید.

کلمات با رتبه های بالاتر به عنوان همسایه کلمه ورودی تعیین می‌شوند. اگر اندازه مناسب برای همسایگی تعیین شده باشد، کلمه ورودی حتما در این همسایگی قرار می‌گیرد. در یک دیکشنری شامل ۵۰۰ کلمه، بازشناسی صحیح کلمه ورودی با نرخ ۸۶/۳٪ در انتخاب اول، با نرخ ۹۰٪ در دو انتخاب اول و با نرخ ۹۴/۱٪ در ۱۰ انتخاب اول گزارش شده است. الگوریتم بازشناسی این روش در شکل ۷-۲ آورده شده است.

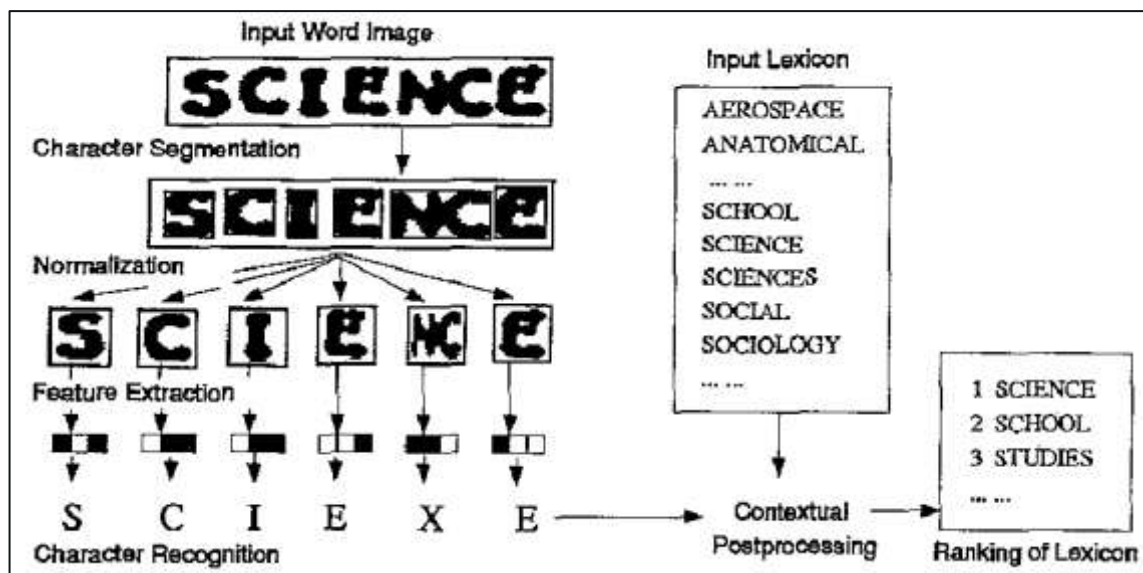


شکل ۷-۲: الگوریتم بازشناسی کلمات با شکل کلی در [9]

در [10] نیز از توزیع جهت پاره خط برای توصیف کلمات لاتین استفاده شده است. تصویر کلمه مانند [9] به یک شبکه ی 4×10 قسمتی، تقسیم می‌شود و همان ویژگی‌های توزیع جهت پاره خط، از این شبکه‌ها استخراج می‌شوند و موقعیت نسبی آنها در شبکه در یک بردار ویژگی ثبت می‌شود. این بردار ویژگی با یک واژه نامه برای تعیین همسایگی کلمه ورودی مقایسه می‌شود و کلماتی که در رتبه های بالاتر قرار می‌گیرند به عنوان همسایه کلمه ورودی تعیین می‌شوند برای تعیین فاصله نیز از معیار بلوک شهر استفاده شده است.

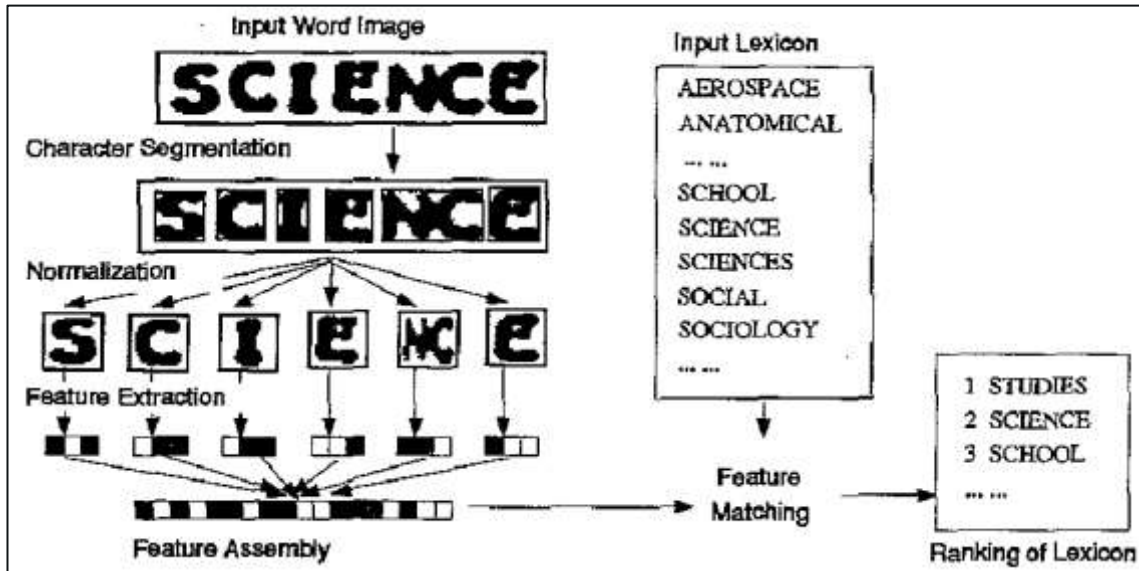
این تحقیق سه روش بر اساس بازشناسی حروف، بر اساس جداسازی و بر اساس شکل کلی را مورد بررسی قرار داده و از ترکیبی آنها برای بازشناسی در مرحله‌ی آخر استفاده کرده است. در روش اول، ابتدا تصویر کلمه به حروف آن جداسازی می‌شود. این حروف با استفاده از مقادیر دودویی نقاط تصویر و یک تطبیق‌گر فازی الگو، بازشناسی می‌شوند و با پس پردازش‌های محتوایی برای بازشناسی کلمه ورودی، با هم ترکیب می‌شوند.

این روش برای بازشناسی کلمات با کیفیت مناسب است. این الگوریتم در شکل ۸-۲ آمده است.



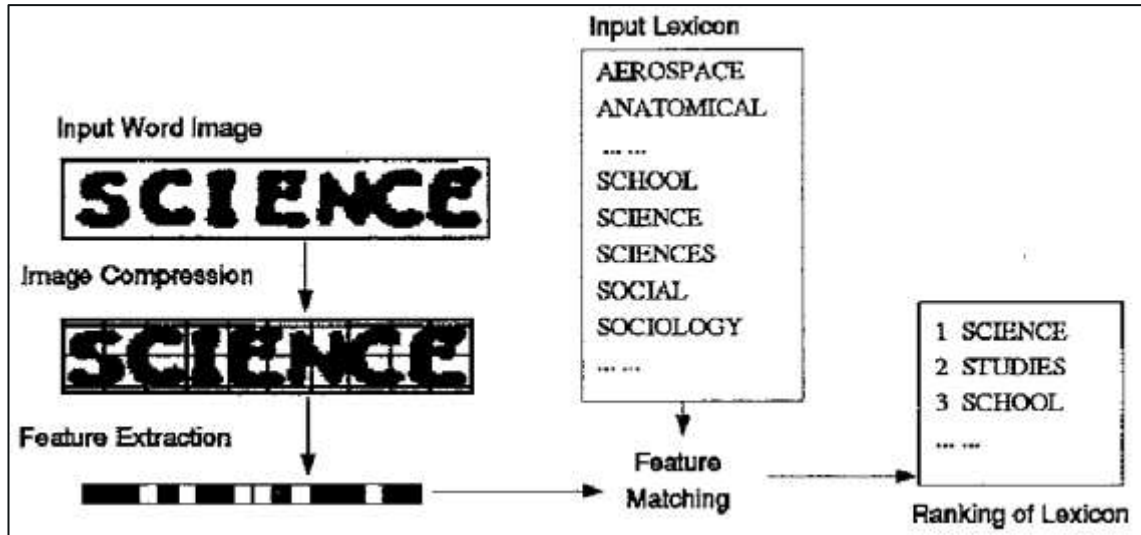
شکل ۸-۲: روش اول استفاده شده در [10]

در روش دوم نیز تصویر کلمه به حروف آن جداسازی می‌شود و ویژگی‌های مقادیر باینری نقاط تصاویر نرمالیزه شده به اندازه‌های 24×24 محاسبه می‌شوند. با کنار هم قرار دادن این بردارها، بردار توصیفگر کلمه تشکیل می‌شود. با مقایسه این بردار با الگوهای ذخیره شده در واژه نامه، رتبه بندی از واژه نامه تولید می‌شود. این روش برای تصاویری که جداسازی حروف آنها، با دقت بالایی انجام می‌شود ولی شناسایی حروف با دقت بالا ممکن نیست، مناسب است. الگوریتم این روش در شکل ۹-۲ ارائه شده است.



شکل ۲-۹: روش دوم بازشناسی در [10]

و در روش سوم، از روش ارائه شده در [9] استفاده شده است (شکل ۲-۱۰).



شکل ۲-۱۰: روند بازشناسی روش سوم در [10]

از ترکیب این روش‌ها برای بازشناسی نهایی استفاده شده است. با آزمایش این روش بر روی یک

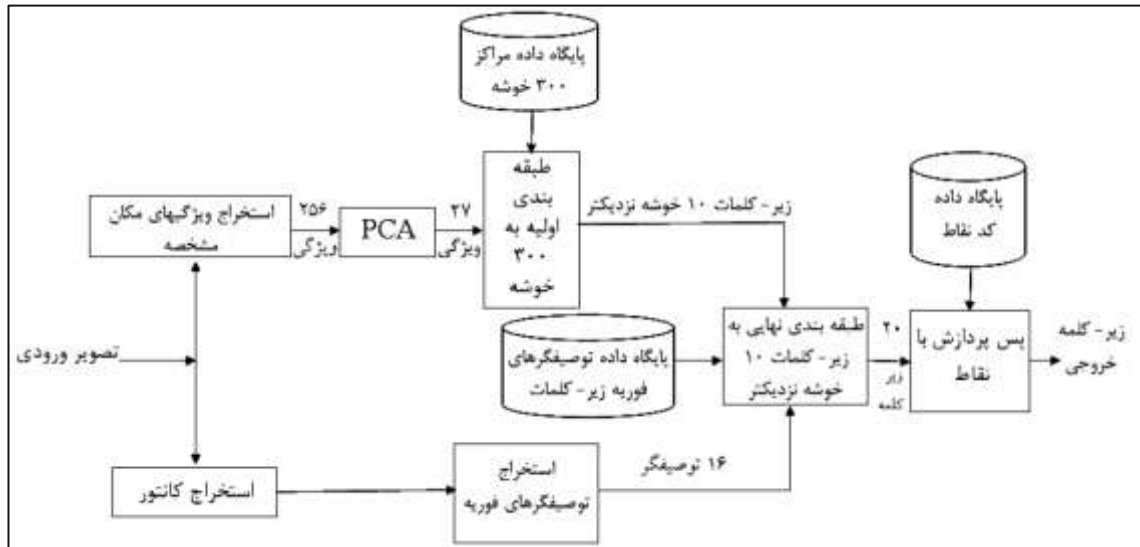
مجموعه ۱۶۷۱ کلمه‌ای، با استفاده از یک واژه‌نامه شامل ۳۳۸۵۰ لغت، دقت بازشناسی در انتخاب اول

۸۸/۹٪ و در ۱۰ انتخاب اول ۹۵/۵۳٪ گزارش شده است.

از تبدیل فوریه دوبعدی شکل کلمات نیز برای بازشناسی متون چاپی عربی با چهار قلم متداول استفاده شده است [11] تصویر کلمه نرمالیزه شده 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 از هر یک از لغات دیکشنری انجام می‌شود.

در [۱۲] الگوریتمی برای خوشه بندی زیرکلمات چاپی جهت بازشناسی آنها ارائه شده است در مرحله ایجاد دیکشنری، کانتور بالایی تصویر هر زیرکلمه استخراج شده و برچسب می‌خورد. جایگاه زیرکلمات در دیکشنری بر اساس این برچسب‌ها تعیین می‌شود. در مرحله طبقه بندی، مجموعه زیرکلمات موجود در دیکشنری که برچسب آنها با برچسب زیرکلمه‌ی ورودی یکسان است، مشخص کننده‌ی نمونه‌ی ورودی هستند.

در [۱۳] که مبتنی بر بازشناسی با شکل کلی است، ابتدا زیرکلمات، با استفاده از الگوریتم k میانگین و ویژگی مکان مشخصه به ۳۰۰ خوشه تقسیم شده‌اند. سپس نمونه ورودی پس از استخراج ویژگی مکان مشخصه، به یکی از این خوشه‌ها نسبت داده می‌شود. پس از آن با استفاده از توصیفگرهای فوریه، مجدداً استخراج ویژگی صورت گرفته و بازشناسی نهایی صورت می‌گیرد. فرایند کلی در شکل ۲-۱۱ آمده است.



شکل ۲-۱۱: فرایند بازشناسی در روش [۱۳]

در [14] که برای بازشناسی کلمات چاپی عربی با شبکه عصبی پیشنهاد شده است بازشناسی کلمات در سیستم پیشنهادی به سه مرحله تقسیم می‌شود. در مرحله پردازش‌های اولیه، تصویر توسط روبشگر با درجه تفکیک ۳۰۰ نقطه در اینچ تصویر برداری شده و تکه‌های جدای آن مشخص می‌شوند. سپس ویژگی‌های عمومی مانند تعداد تکه‌ها، تعداد بیشینه‌ها در هر تکه، تعداد و محل نقاط و علائم و ویژگی‌های هندسی دیگری از شکل کلی کلمه استخراج می‌شود. برای بازشناسی کلمات از یک شبکه عصبی سه لایه استفاده شده است. تعداد نورون‌ها در ورودی شبکه ۲۷۰ عدد گزارش شده است.

در [۱۵] 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 زیرکلمات استوار است. تعداد زیرکلمات فارسی ۱۰۰۰ عدد ذکر شده است. با آزمایش این روش بر روی مجموعه‌ای شامل ۵۰۰ زیرکلمه دست‌نویس نوشته شده و یک صفحه متن چاپی در هر دو مورد نرخ بازشناسی صحیح ۹۶٪ گزارش شده است.

فصل ۳:

مباحث نظری و استخراج ویژگی

در این فصل به توضیح مباحث نظری و استخراج ویژگی مورد استفاده در فصل‌های آتی می‌پردازیم. معیارهای فاصله روش‌های استخراج ویژگی، خوشه‌بندی و طبقه‌بندی داده‌ها از جمله‌ی این مباحث هستند.

۳-۱- معیارهای فاصله

در این بخش برخی معیارهای فاصله را که در فصل‌های بعدی برای محاسبه فاصله بین دو بردار در فضای داده‌ها استفاده خواهد شد، ارائه می‌شود. از این معیارها برای اندازه‌گیری فاصله بین یک نمونه و یک خوشه، فاصله بین دو نمونه و فاصله بین دو خوشه استفاده خواهد شد. با فرض اینکه بعد بردار ویژگی داده‌ها، n باشد و S_j و T_j عنصر J ام از دو بردار ویژگی S و T در این فضای ویژگی‌ها باشند، معیارهای زیر را داریم [16].

۳-۱-۱- معیار فاصله مینکوسکی^۱

فاصله مینکوسکی [17] از مشهورترین فواصلی است که از آن در بسیاری از کاربردها استفاده می‌شود.

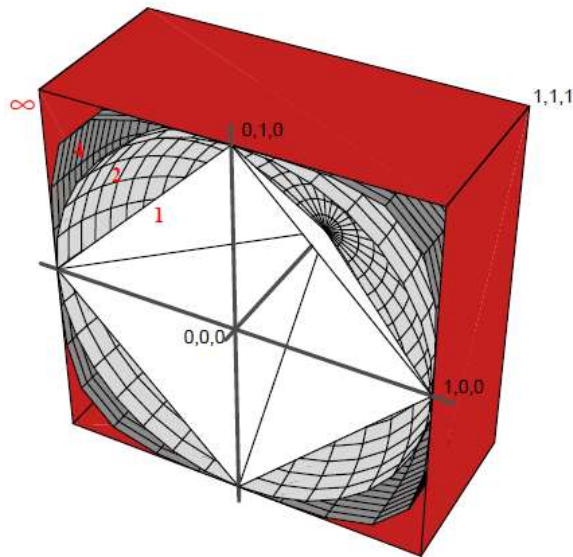
از این معیار فاصله که در رابطه (۳-۱) نشان داده شده است دو گونه مشهور از فاصله مینکوسکی، که به ازای $M = 1$ و $M = 2$ استخراج می‌شود به ترتیب فاصله ی بلوک شهر (۳-۲) و فاصله ی اقلیدسی (۳-۳) نام دارند در شکل ۳-۱ نمایی از این معیار فاصله نشان داده شده است. در این روابط f_i و g_i به ترتیب عناصر بردارهای f و g می‌باشند.

$$D_{minkowski}(F, G) = \left(\sum_{i=1}^n |f_i - g_i|^M \right)^{\frac{1}{M}}, \quad 1 \leq M \leq \infty \quad (1-3)$$

¹ Minkowski distance measure

$$D_{cityblock}(F, G) = \sum_{i=1}^n |f_i - g_i| \quad (2-3)$$

$$D_{euclidean}(F, G) = \left(\sum_{i=1}^n |f_i - g_i|^2 \right)^{\frac{1}{2}} \quad (3-3)$$



شکل ۳-۱: در معیار فاصله مینکوسکی، برای $M = 1$ معیار فاصله بلوک شهر بدست می‌آید که در صفحه دارای مکان هندسی لوزی در فضای سه بعدی هرم و در فضاهای بالاتر یک ابر هرم^۱ است و به همین ترتیب $M = 2$ برای معیار اقلیدسی به ترتیب در فضای دو، سه و بالا تر شاهد مکان هندسی دایره، کره و ابر کره^۲ هستیم و در نهایت برای $M = \infty$ این مکان هندسی به مربع و مکعب و ابر مکعب^۳ ختم می‌شود [17].

۳-۱-۲- معیار فاصله کسینوس

در معیار فاصله کسینوس این سنجش برابر با زاویه‌ی بین دو بردار ویژگی است و از رابطه‌ی

(۳-۴) به دست می‌آید [16].

$$D_{\cosine}(S, T) = 1 - \frac{S \times T'}{\sqrt{(S \times S')(T \times T')}} \quad (4-3)$$

S' و T' به ترتیب ترانهاده‌های بردارهای ویژگی S و T می‌باشند، و عملگر \times ضرب داخلی دو

¹ Hyperpyramid

² Hyperglobe

³ Hypercube

بردار است.

۳-۱-۳- معیار فاصله چبی شف^۱

فاصله‌ی چبی شف (رابطه (۳-۵))، برابر با بزرگ‌ترین مقدار قدر مطلق تفاضل اجزاء متناظر از دو

بردار ویژگی است [16].

$$D_{\text{chebychev}}(S, T) = \max_j |S_j - T_j| \quad (۵-۳)$$

۳-۱-۴- معیار فاصله اسپیرمن^۲

معیار فاصله‌ی اسپیرمن [16]، از رابطه‌ی (۳-۶) محاسبه می‌شود.

$$D_{\text{spearman}}(S, T) = 1 - \frac{(rs - \alpha)(rt - \alpha)'}{\sqrt{(rs - \alpha)(rs - \alpha)'} \sqrt{(rt - \alpha)(rt - \alpha)'}} \quad (۶-۳)$$

rs ، بردار متناظر با S می‌باشد؛ که جزء z ام آن، شماره‌ی اندیس نظیر با عنصر z ، S بعد از مرتب

شدن صعودی تمام عناصر S است، بدیهی است شماره‌ی اندیس یک، مربوط به اولین جزء مرتب شده

است و به همین ترتیب ادامه می‌یابد. برای مثال مرتب شده‌ی بردار $S = \{70, 115, 90, 40\}$ ، برابر با

$rs = \{40, 70, 90, 115\}$ شده و شماره‌ی اندیس‌ها به ترتیب $\{1, 2, 3, 4\}$ می‌شوند و $rs = \{2, 4, 3, 1\}$

حاصل می‌شود. rt ، بردار متناظر با T نیز به همین شکل به دست می‌آید.

و α نیز از رابطه‌ی (۳-۷) به دست می‌آید که n بُعد بردارهای T و S است.

$$\alpha = \frac{(n+1)}{2} \quad (۷-۳)$$

¹ Chebychev

² Spearman

۳-۱-۵- معیار فاصله همبستگی

معیار فاصله‌ی همبستگی^۱ با رابطه‌ی (۶) تعریف می‌شود [16].

$$D_{\text{comelation}}(S, T) = 1 - \frac{(S - \bar{S})(T - \bar{T})}{\sqrt{(S - \bar{S})(S - \bar{S})} \sqrt{(T - \bar{T})(T - \bar{T})}} \quad (۸-۳)$$

که \bar{S} و \bar{T} به ترتیب میانگین بردارهای S و T هستند.

۳-۲- ویژگی‌ها

هر مقدار کمی و کیفی که توصیف کننده و متمایز کننده‌ی الگوها باشد (در اینجا تصویر زیرکلمه الگوی مورد بازشناسی است) یک ویژگی محسوب می‌شود، این ویژگی‌ها در صورت کمی بودن به صورت بردار ظاهر می‌شوند و فضایی را که در آن هستند را فضای ویژگی‌ها می‌نامند.

۳-۲-۱- میانگین بلوکی^۲

ویژگی میانگین بلوکی بیان‌گر خصوصیت‌های کلی تصویر است و برای استخراج این ویژگی تصویر بلوک‌بندی شده و با میانگین گرفتن از هر کدام از ناحیه‌های تصویر بلوک‌بندی شده به دست می‌آید و نسبت به تغییرات ضعیف روشنایی پایداری دارد.

۳-۲-۲- هیستوگرام گرادیان

این ویژگی [۱۸] می‌تواند برای تصاویر دو سطحی و خاکستری استخراج شود ولی معمولاً برای تصاویر خاکستری نتایج بهتری حاصل می‌شود. عموماً این موضوع ناشی از این است که در حین دو سطحی سازی، نویز و حروف بریده‌ای حاصل می‌شود که سبب کاهش دقت بازشناسی می‌شود. این

^۱ Correlation

^۲ Zoning

ویژگی جدید، پیچیدگی محاسباتی کمی داشته و از این رو بسیار سریع است.

نحوه‌ی استخراج ویژگی به صورت زیر است.

۱. نرمال کردن تصویر به ابعادی مورد نظر

۲. گرادیان گرفتن از تصویر با اعمال فیلتر رابرت و یا سوبل برای یافتن مؤلفه‌های گرادیان که به

صورت زیر تعریف می‌شوند

مؤلفه‌های گرادیان در عملگر رابرت :

$$g_x(x, y) = f(x, y) - f(x + 1, y + 1) \quad (۹-۳)$$

$$g_y(x, y) = f(x + 1, y) - f(x, y + 1) \quad (۱۰-۳)$$

مؤلفه‌های گرادیان در عملگر سوبل :

$$g_x(x, y) = f(x + 1, y - 1) + 2f(x + 1, y) + f(x + 1, y + 1) - f(x - 1, y - 1) - 2f(x - 1, y) - f(x - 1, y + 1) \quad (۱۱-۳)$$

$$g_y(x, y) = f(x - 1, y + 1) + 2f(x, y + 1) + f(x + 1, y + 1) - f(x - 1, y - 1) - 2f(x, y - 1) - f(x + 1, y - 1) \quad (۱۲-۳)$$

-۱	۰	۱
-۲	۰	۲
-۱	۰	۱

مؤلفه g_x سوبل

۱	۲	۱
۰	۰	۰
-۱	-۲	-۱

مؤلفه g_y سوبل

۰	-۱
۱	۰

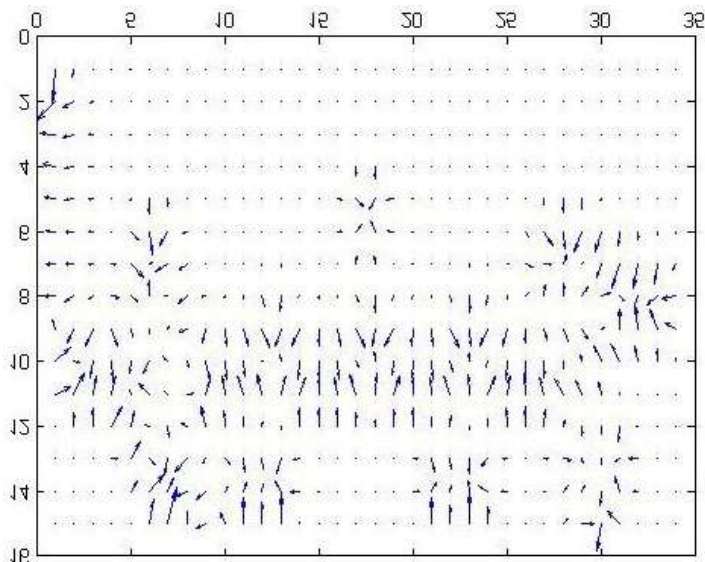
مؤلفه g_x رابرت

-۱	۰
۰	۱

مؤلفه g_y رابرت

شکل ۳-۲: مؤلفه‌های عملگرهای گرادیان سوبل و رابرت

در شکل ۳-۳ گرادیان زیرکلمه‌ی "چینیها" توسط عملگر رابرت مشاهده می‌شود.



شکل ۳-۳: نتیجه‌ی عملگر گرادیان با فیلتر رابرت بر روی تصویری با ابعاد 15×34

۳. برای هر پیکسل، با استفاده از مؤلفه‌های گرادیان اندازه و فاز گرادیان به صورت روابط (۳-۱۳)

و (۳-۱۴) بدست می‌آید

$$p(x, y) = \text{atan}(g_x, g_y) \quad (۳-۱۳)$$

$$A(x, y) = \sqrt{g_x^2 + g_y^2} \quad (۳-۱۴)$$

۴. زاویه‌ی گرادیان به تعداد مورد نظر کوانتیزه می‌شود

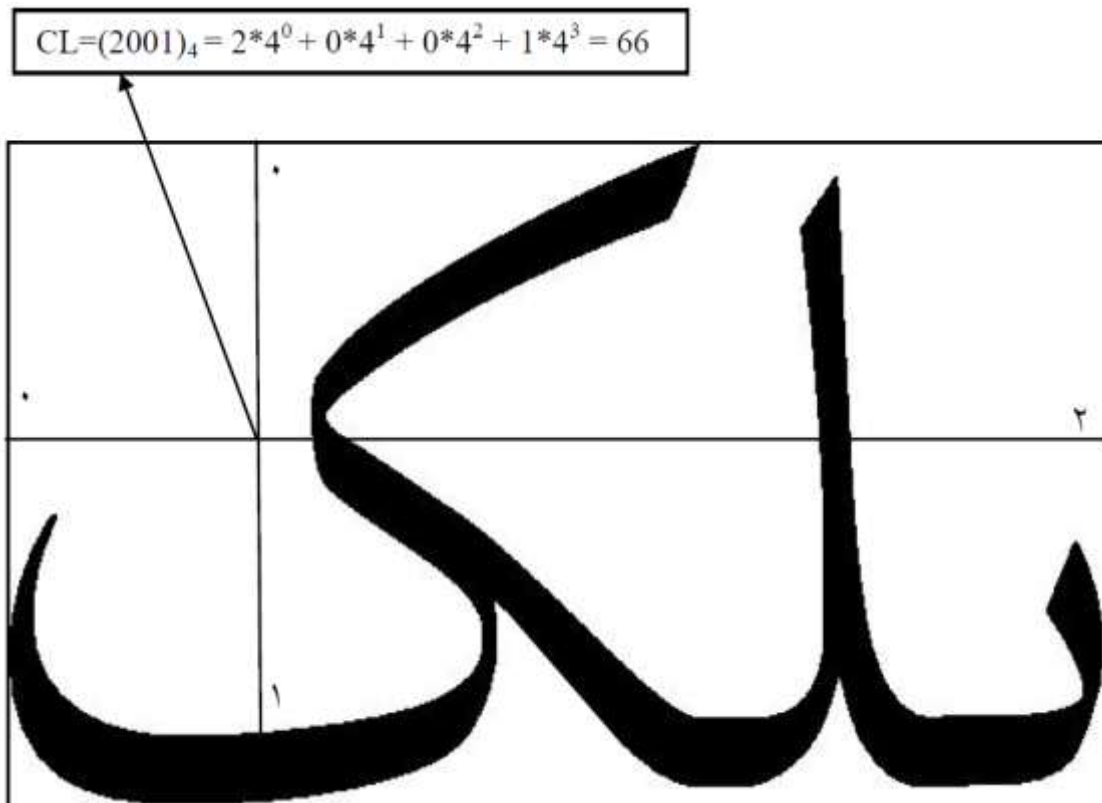
۵. برای هر زیر بلوک که به تعداد کوانتیزه‌ای که در مرحله‌ی چهار انجام شده است، در هر زاویه

با توجه به رابطه‌ی (۳-۱۵) ویژگی متناظر با آن زاویه‌ی بدست می‌آید [۱۸].

$$F_{\theta} = \sum_{x_{\theta}, y_{\theta}} A(x, y) \quad (۳-۱۵)$$

۳-۲-۳- مکان مشخصه

برای محاسبه‌ی این ویژگی [۱۳] به هر پیکسل پس زمینه یک عدد نسبت می‌دهیم این عدد با توجه به تعداد قطع کردن خطوطی که از آن نقطه در جهت‌های چهارگانه‌ی بالا، پایین، چپ، راست در بدنه‌ی زیرکلمه (گذر سیاه به سفید) اتفاق می‌افتد محاسبه می‌شود، در بازشناسی ارقام این عدد به دلیل ساده بودن شکل اعداد به ۲ محدود می‌شد اما در زیرکلمات با توجه به پیچیدگی ظاهری آنها این عدد را به ۳ محدود می‌کنند که در این صورت بردار ویژگی $(3 \times 3 \times 3 \times 3)$ ۸۱ حالت دارد، در [۵] این محدودیت را در سمت چپ و راست که تداخل بیشتری در بدنه‌ی زیرکلمه دارد را به ۳ و در راستای بالا و پایین را به ۲ کوآنتیزه کرده است که در نتیجه بُعد بردار ویژگی $(3 \times 4 \times 3 \times 4)$ دارای ۱۴۴ مؤلفه می‌شود.



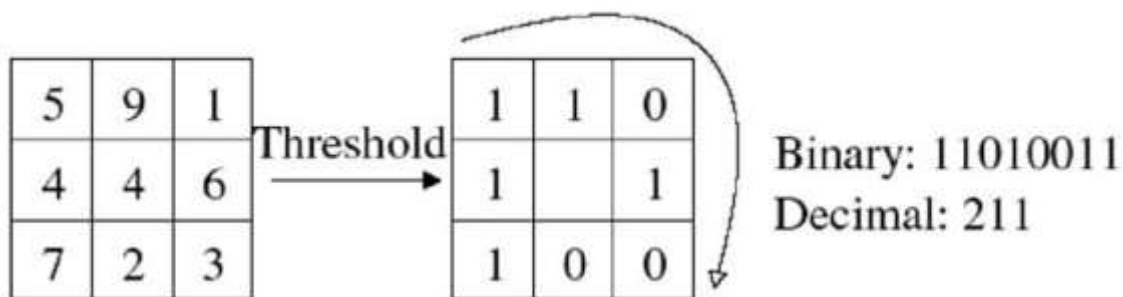
شکل ۳-۴: نحوه استخراج ویژگی مکان مشخصه [۱۳]

۳-۲-۴- سیگنال علامت^۱

ویژگی سیگنال علامت که در اینجا معرفی شده، متمرکز در نقاط زیر-کلمه است و پس از حذف بدنه از زیر-کلمه استخراج می‌شود، بدین صورت که اگر علامت در بالای خط کرسی باشد سیگنال، با +۱ و در صورتی که علامت در پایین خط کرسی باشد با -۱ ساخته می‌شود؛ البته به گونه ای دقیق‌تر نیز می‌توان عمل کرد به این صورت که وزن هر علامت در زیر-کلمه را به جای عدد یک قرار داد.

۳-۲-۵- الگوهای دو دویی محلی^۲

برای استخراج این ویژگی [19] برای هر پیکسل، از همسایه‌های آن استفاده می‌شود جهت یافتن یک کد دو دویی، که این همسایه‌ها روی دایره‌ای به مرکز این پیکسل قرار دارند، این کد که تعیین‌کننده‌ی مقدار ویژگی آن پیکسل است (معمولاً از ده دهی این کد استفاده می‌شود)؛ با توجه به اختلاف مقدار روشنایی همسایه‌ها با پیکسل مرکزی و آستانه گذاری روی این اختلاف‌ها، حاصل می‌شود؛ در نهایت با تشکیل هیستوگرام مقادیری حاصل شده از این کد در پیکسل‌ها، بردار ویژگی به دست می‌آید.



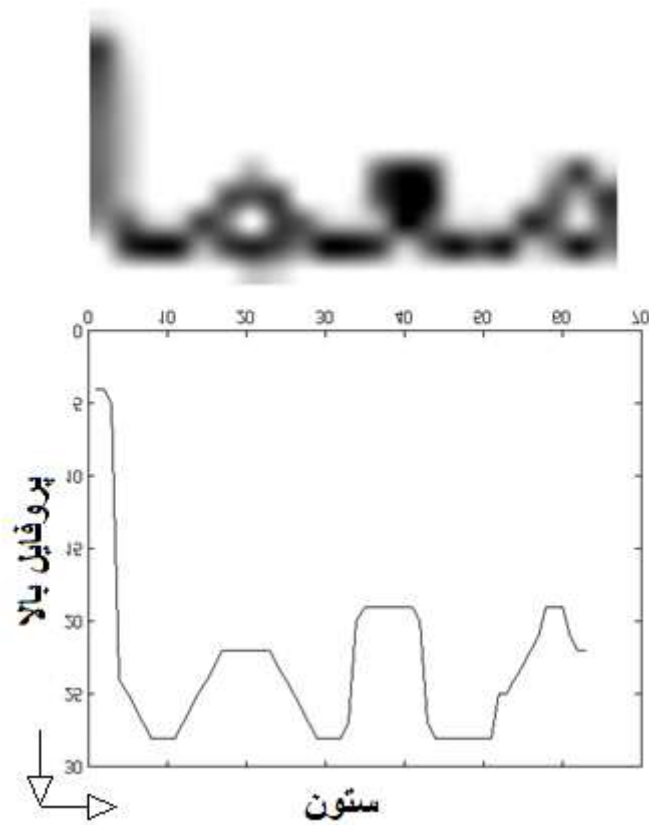
شکل ۳-۵: استخراج ویژگی LBP [19]

¹ Signe Signal

² Local Binary Pattern

۳-۲-۶- پروفایل^۱ بالا و پایین

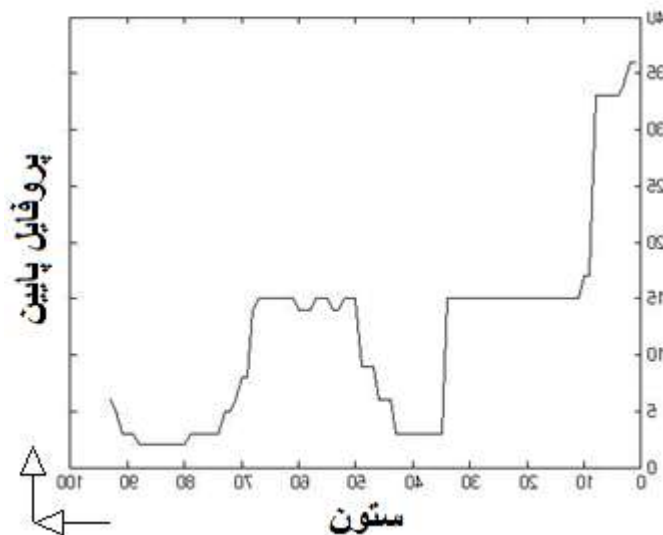
ویژگی پروفایل [۲۰] بالا و پایین، نمای دیده شده از بالا و پایین از تصویر کلمه را نشان می‌دهند، این ویژگی‌ها با اندازه گیری فاصله اولین پیکسل کلمه با مرز تصویر در هر ستون محاسبه می‌شود. ابعاد هر کدام از این بردار ویژگی‌ها برابر با تعداد ستون‌های تصویر زیرکلمه است. در شکل ۳-۶ و شکل ۳-۷ مثال‌هایی از استخراج این ویژگی در دو حالت پروفایل بالا و پروفایل پایین، برای زیرکلمات آورده شده است.



شکل ۳-۶: شکل زیرکلمه و نمای پروفایل بالای آن

^۱ Profile

گنجهش

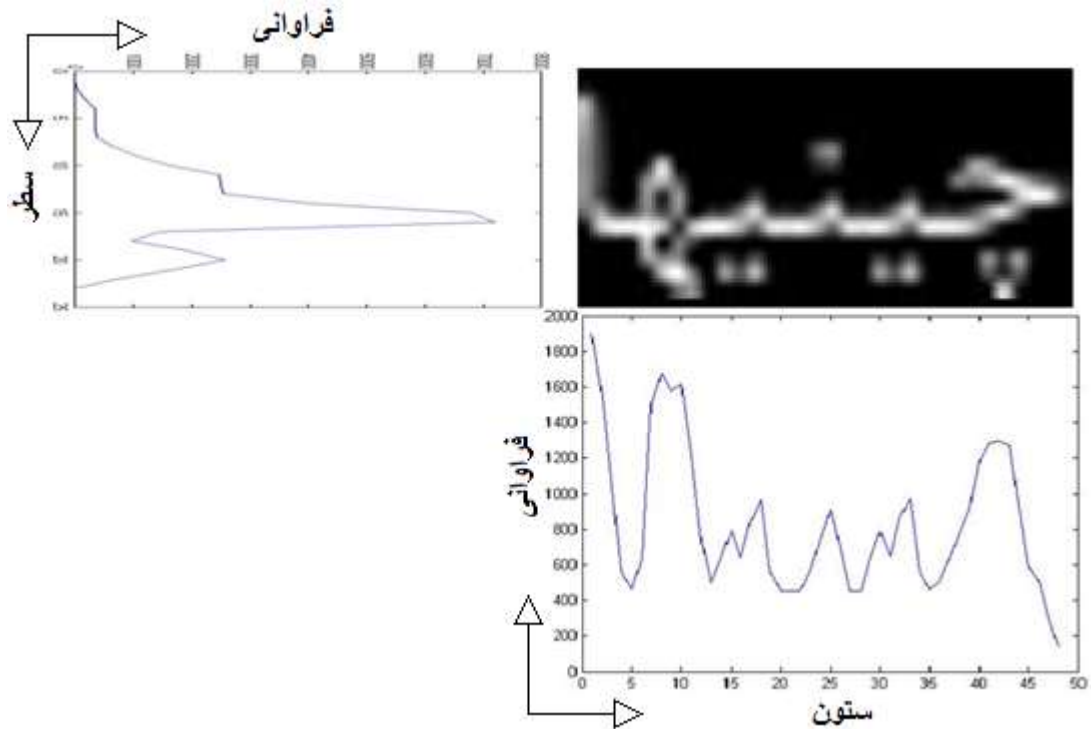


شکل ۷-۳: شکل زیرکلمه و نمای پروفایل پایین آن

۷-۲-۳- افکنش^۱

افکنش عمودی، توزیع پیکسل‌های زیرکلمه را در جهت عمودی نشان می‌دهد. [۲۰] این ویژگی از مجموع تعداد پیکسل‌های نویسه (در اینجا سفید) در هر ستون بدست می‌آید و بعد آن برابر با تعداد ستون‌های تصویر زیرکلمه است، همچنین افکنش افقی، توزیع پیکسل‌های زیرکلمه را در جهت افقی نشان می‌دهد و از مجموع تعداد پیکسل‌های نویسه در هر سطر بدست می‌آید، بعد این ویژگی نیز برابر با تعداد سطرهای تصویر زیرکلمه است، در شکل ۸-۳ روش استخراج ویژگی برای یک زیرکلمه مشاهده می‌شود.

^۱ Projection



شکل ۳-۸: شکل زیر کلمه و نمای افکنش افقی آن در سمت چپ و نمای افکنش عمودی آن در پایین نشان داده شده است.

۳-۲-۸- گذر سیاه به سفید^۱

سه ویژگی پروفایل بالا و پایین و افکنش عمودی ساختار بیرونی شکل را نشان می‌دهد، برای نشان دادن ساختار درونی کلمه از ویژگی گذر سیاه به سفید استفاده می‌شود، این ویژگی تعداد انتقال‌ها از پیکسل‌های پس زمینه (سفید) به پیکسل‌های پیش زمینه (سیاه) را محاسبه می‌کند [۲۰] برای استخراج این ویژگی ابتدا باید تصویر را دوسطحی^۲ کرد.

۳-۳- دو سطحی سازی

برخی از پردازش‌ها بر روی تصاویر دوسطحی انجام می‌شود. بنابراین یکی از مراحل پیش پردازش باینری کردن تصاویر رنگی رویش شده می‌باشد. تصاویر دوسطحی متون، تصاویری هستند که

¹ Zero Crossing

² Binary

پیکسل‌های مربوط به کلمه یا پیش زمینه دارای سطح صفر (سیاه) و پیکسل‌های مربوط به پس زمینه دارای سطح یک (سفید) می‌باشند که این کار توسط الگوریتم‌های آستانه گذاری انجام می‌شود. روش‌های آستانه گذاری به طور کلی به دو دسته تقسیم می‌شوند: آستانه گذاری کلی و محلی [۲۰]. در روش‌های کلی تنها یک آستانه برای کل تصویر استفاده می‌شود. این روش برای تصاویر خاکستری که هیستوگرام چند قلّه‌ای یا پس زمینه غیر یکنواخت دارند، مناسب نیستند. در روش‌های آستانه گذاری محلی که بیشتر برای تصاویری که در جای جای آن‌ها روشنی‌های متفاوت دارند مانند تصاویر گرفته شده با دوربین مناسب‌اند؛ با محاسبه آستانه‌های موضعی در محدوده پنجره‌های جداگانه و یا نواحی بخش بندی شده، کاستی‌های باینری سراسری اصلاح می‌شوند.

۳-۱- روش پیشنهادی اُتسو^۱

الگوریتم آستانه گذاری مذکور [21] بر اساس تقسیم بندی پیکسل‌های تصویر به دو کلاس شی و پس زمینه با یک آستانه بهینه می‌باشد. انتخاب آستانه بهینه با مینیمم کردن واریانس درون کلاس‌های بدست آمده توسط یک آستانه شناور انجام می‌شود. برای این منظور، ابتدا هیستوگرام روشنایی تصویر محاسبه می‌شود. در هیستوگرام روشنایی، مقدار $p(i)$ برابر با احتمال وجود سطح خاکستری i ام در تصویر است. برای تصویر خاکستری ۸ بیتی، ۲۵۶ سطح خاکستری ارزیابی می‌شود.

$$p(i) = \text{تعداد کل پیکسل‌ها} / \text{تعداد پیکسل‌ها با سطح خاکستری } i \quad (۱۶-۳)$$

اگر هیستوگرام دو قلّه‌ای باشد، بهترین آستانه، مقداری است که دو قلّه هیستوگرام را از هم جدا کند. در غیر این صورت آستانه، مقداری خواهد بود که مجموع وزن‌های واریانس‌های درونی دو کلاس تشکیل شده را حداقل کند.

اگر $\sigma_1^2(t)$ و $\sigma_2^2(t)$ به ترتیب واریانس کلاسی با مقدار کمتر یا مساوی آستانه t و واریانس کلاسی با مقدار بیشتر از t باشد و $q_1(t)$ و $\mu_1(t)$ به ترتیب احتمال و میانگین کلاس اول و

^۱ Otsu

$q_2(t)$ و $\mu_2(t)$ احتمال و میانگین کلاس دوم در نظر گرفته شود. واریانس درون کلاس‌ها، به صورت زیر محاسبه می‌شود.

$$\sigma_w^2(t) = q_1(t) \sigma_1^2(t) + q_2(t) \sigma_2^2(t) \quad (۱۷-۳)$$

که

$$q_1(t) = \sum_{i=0}^t p(i) \quad (۱۸-۳)$$

$$q_2(t) = \sum_{i=t+1}^{255} p(i) \quad (۱۹-۳)$$

$$\mu_1(t) = \sum_{i=0}^t ip(i) / q_1(t) \quad (۲۰-۳)$$

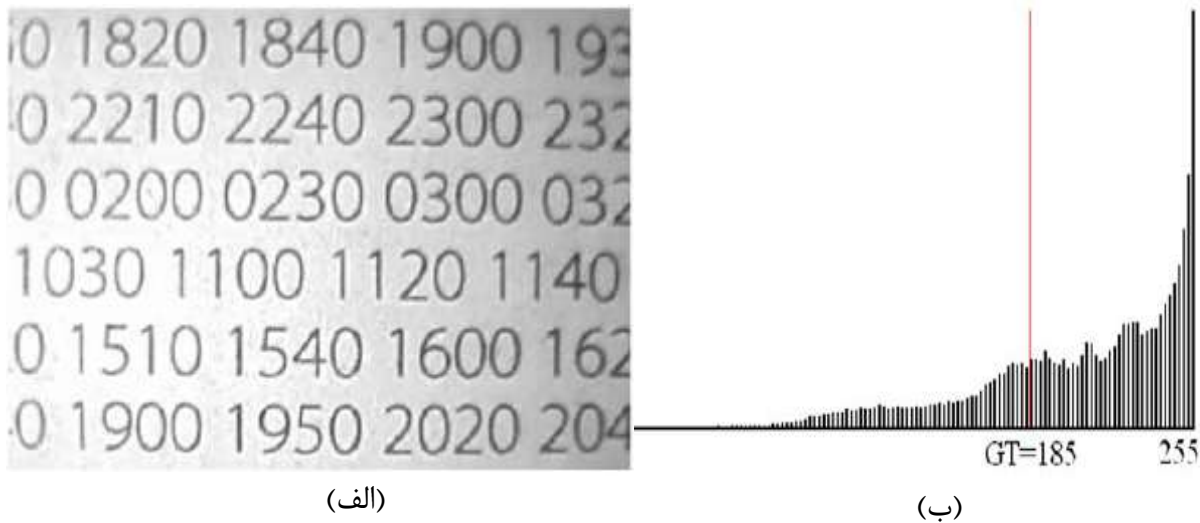
$$\mu_2(t) = \sum_{i=t+1}^{255} ip(i) / q_2(t) \quad (۲۱-۳)$$

و

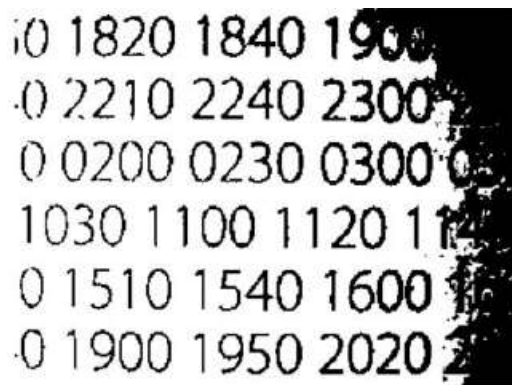
$$\sigma_1^2(t) = \sum_{i=0}^t (i - \mu_1(t))^2 p(i) / q_1(t) \quad (۲۲-۳)$$

$$\sigma_2^2(t) = \sum_{i=t+1}^{255} (i - \mu_2(t))^2 p(i) / q_2(t) \quad (۲۳-۳)$$

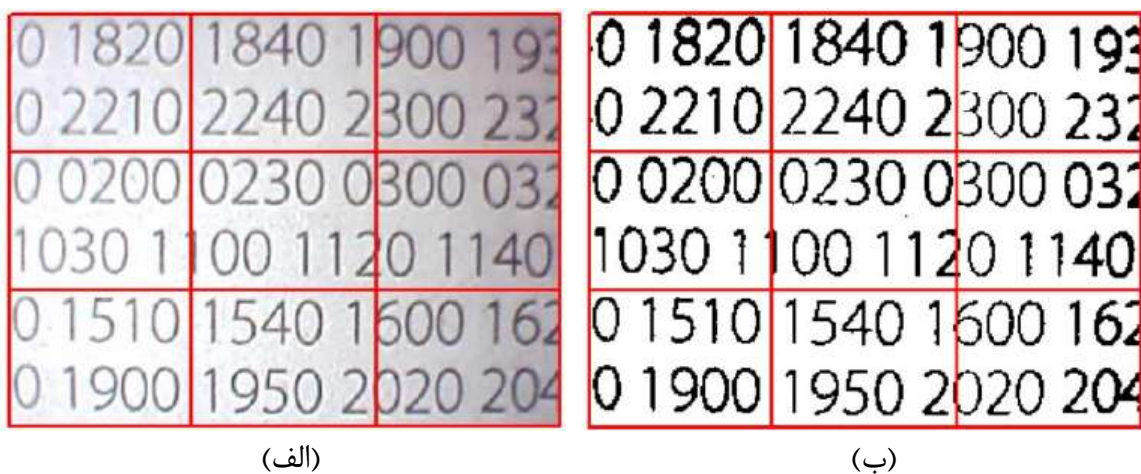
بهترین آستانه‌ی t مقداری است که $\sigma_w^2(t)$ را مینیمم کند. در شکل ۳-۹ یک نمونه‌ی تصویری که توسط دوربین گرفته شده یک بار با آستانه‌ی کلی (شکل ۳-۱۰) و بار دیگر با آستانه‌ی محلی (شکل ۳-۱۱) دوسطحی شده است تفاوت آستانه‌ی محلی و کلی در پنجره‌ی انتخاب شده برای تعیین آستانه است.



شکل ۹-۳: الف) : یک تصویر نمونه است که روشنایی پس زمینه آن یکنواخت نیست، ب) : هیستوگرام تصویر و آستانه‌ی بدست آمده با روش آتسو [22]



شکل ۱۰-۳: نتیجه‌ی باینری کردن شکل ۹-۳ با روش آتسو از طریق آستانه گذاری سراسری [22]



شکل ۱۱-۳: الف) : تصویر الف شکل ۹-۳ که جهت آستانه‌گذاری محلی قطع‌بندی شده است، ب) نتیجه‌ی باینری کردن ، با روش آتسو از طریق آستانه گذاری محلی است [22]

۳-۴- تشخیص پهناي قلم

در تخمین تشخیص پهناي قلم نویسه‌ها از تکنیک گذر سیاه/سفید بهره می‌برند [23]. چون که راستای اصلی نویسه‌ها در فارسی در جهت افق است پس بهتر است که به صورت افقی این جستجو صورت پذیرد و تعداد پیکسل‌های پیوسته^۱ در هر گذر شمارش شود.

پهناي قلم معمولاً در نوشته‌ها متفاوت است و به خاطر وجود غیر نویسه‌ها^۲ در یک متن باید آن پیکسل‌های پیوسته‌ای را که طول آنها کمتر و یا بیشتر از آستانه‌های از پیش تعیین شده است حذف شوند تا در رقابت هیستوگرام شرکت نکنند. بنابراین پس از بدست آوردن هیستوگرام، بیشترین فراوانی به عنوان عرض قلم شناخته می‌شود.

۳-۵- عملگرهای شکل شناسی^۳

از این عملگرها برای حذف نویز، آغشته سازی، بدست آوردن مرز بیرونی و اسکلت در نوشته استفاده می‌شود [24].

برای استفاده از آنها نیاز به جبر مجموعه‌ها داریم که این روابط زیر کارآمدند.

$$\hat{B} = \{\omega \mid \omega = -b, \text{ for } b \in B\} \quad (24-3)$$

معادله‌ی (۲۴-۳) به این معناست که تمام مجموعه B نسبت به مبدأ قرینه شوند

$$(A)_z = \{c \mid c = a + z, \text{ for } a \in A\} \quad (25-3)$$

معادله‌ی (۲۵-۳) به این معناست که تمام مجموعه A به Z منتقل شوند که $z = (z_1, z_2)$ است.

¹ Run Length

² Non-text

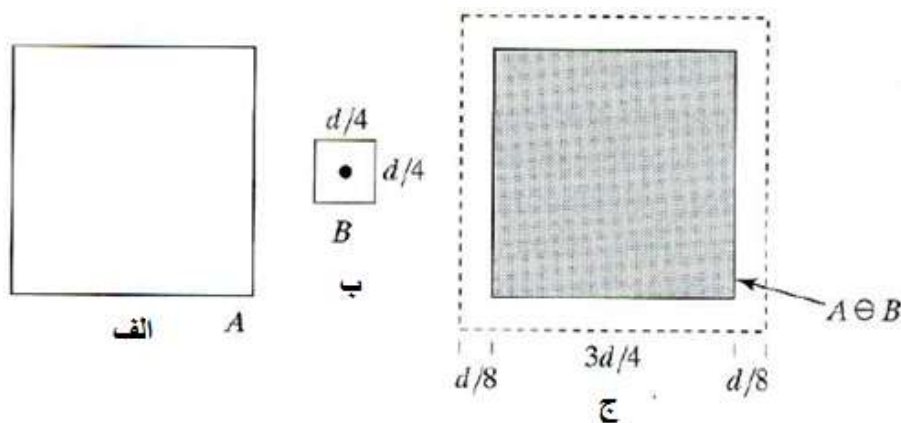
³ Morphology

۳-۵-۱- عملگر ساییدگی^۱ و گسترش^۲

اگر A و B مجموعه‌هایی در Z_2 باشند، ساییدگی A توسط B که با $A \ominus B$ نمایش داده می‌شود، به صورت معادله‌ی (۳-۲۶) تعریف می‌شود. (معمولا مجموعه A تصویر اصلی و مجموعه B به عنوان عنصر ساختاری معرفی می‌شوند)

$$A \ominus B = \{z \mid (\hat{B})_z \subseteq A\} \quad (۳-۲۶)$$

معادله (۳-۲۶) نشان می‌دهد که ساییدگی A توسط B ، مجموعه‌ای از تمام نقاط Z است که B ، یک قسمی که توسط Z انتقال یافته باشد در A قرار گیرد؛ نتیجه‌ی عملگر ساییدگی در شکل ۳-۱۲ آورده شده است.

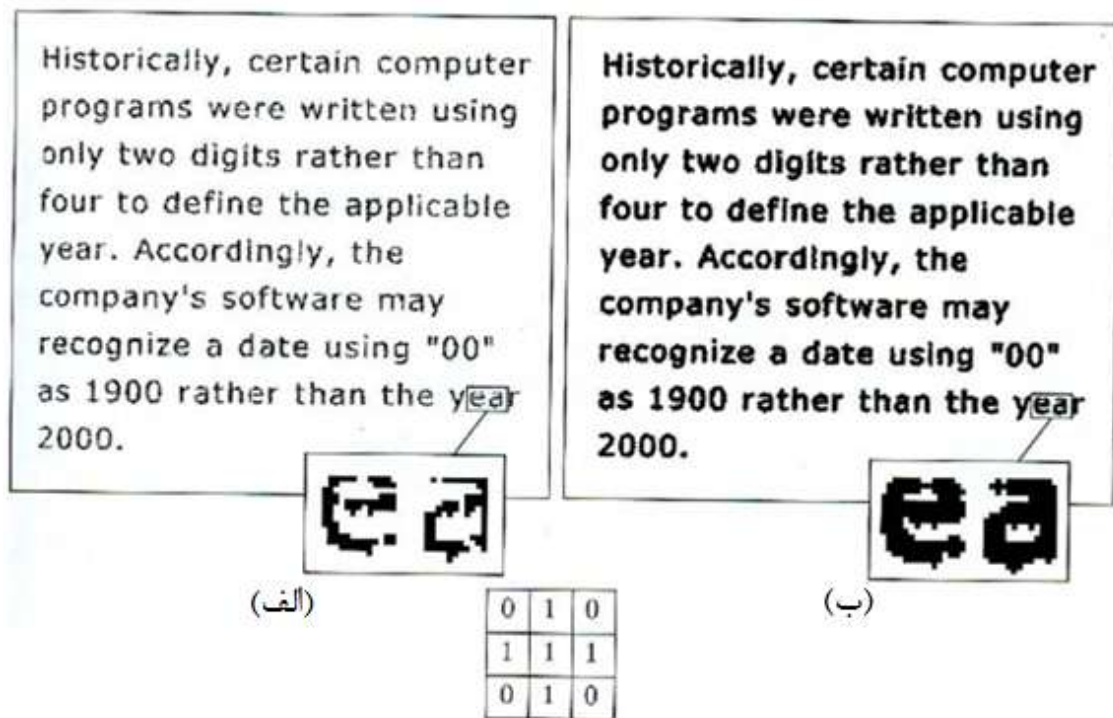


شکل ۳-۱۲: تصویر (ج) نتیجه‌ی عملگر ساییدگی با عنصر ساختاری (ب) روی (الف) را نشان می‌دهد [24]

به صورتی مشابه عملگر گسترش با رابطه (۳-۲۷) تعریف می‌شود. و در شکل ۳-۱۳ تصویر متنی با وضوح پایین و نتیجه‌ی اثر این عملگر بر روی آن دیده می‌شود.

$$A \oplus B = \{z \mid (\hat{B})_z \cap A \neq \emptyset\} \quad (۳-۲۷)$$

¹ Erosion
² Dilation



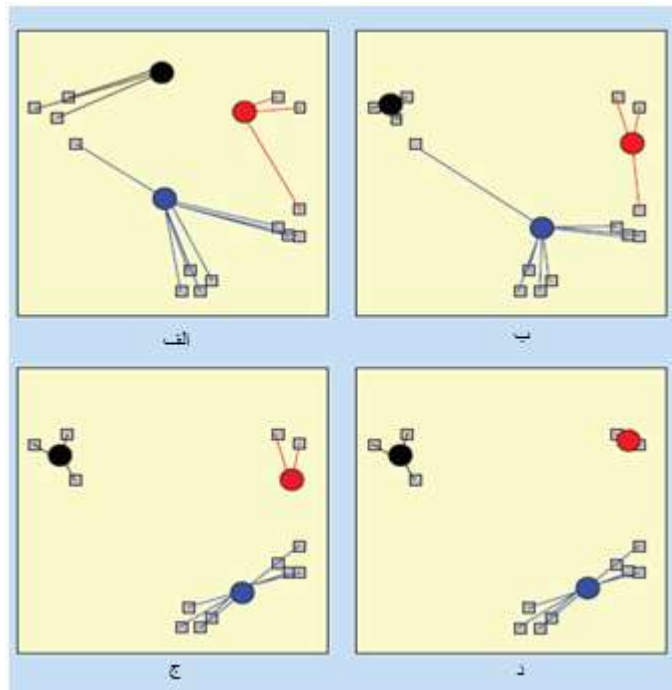
شکل ۳-۱۳: نمونه ای از تصویر متن با کیفیت پایین که نویسه‌ها دچار شکستگی شده‌اند، (ب): تصویر نتیجه‌ی اعمال عملگر گسترش روی تصویر (الف) با عنصر ساختاری لوزی [24]

۳-۶- الگوریتم‌های خوشه‌یابی

۳-۶-۱- الگوریتم K میانگین^۱

در الگوریتم k میانگین [25] تعداد خوشه‌ها مفروض و برابر با k است. الگوریتم با مراکز اولیه‌ی تصادفی شروع شده سپس هر نقطه از مجموعه داده به نزدیک‌ترین مرکز خوشه نسبت داده می‌شود. با مشخص شدن داده‌های متعلق به هر خوشه میانگین نمونه‌های هر خوشه محاسبه شده و لذا مراکز خوشه‌ها تغییر می‌یابند این فرایند تا زمانی که مراکز خوشه‌ها تغییر نکنند، ادامه پیدا می‌کند.

^۱ K means



شکل ۳-۱۴: (الف) تعیین تصادفی مراکز اولیه به تعداد خوشه‌ها و مشخص کردن اینکه هر نقطه به کدام یک از این مراکز نزدیک‌تر است (ب) به روز کردن این مراکز (ج) یافتن مجدد نزدیک‌ترین مرکز به هر کدام از نقاط (د) با پایدار ماندن این مراکز الگوریتم پایان یافته است [25]

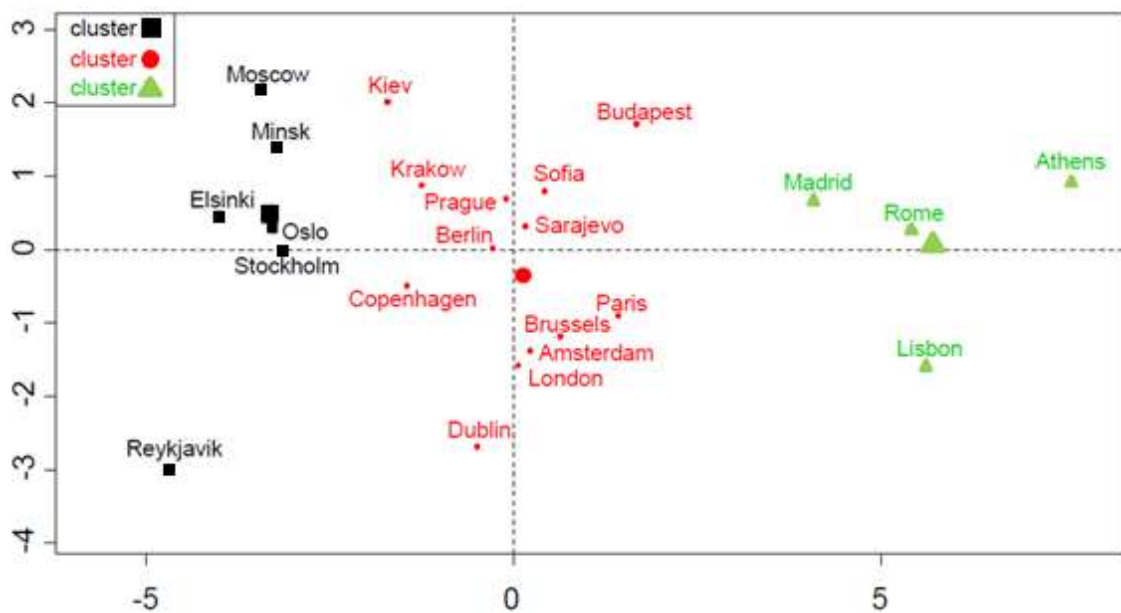
۳-۶-۲- الگوریتم سلسله مراتبی^۱

در مقایسه با k میانگین، الگوریتم سلسله مراتبی با تعداد خوشه‌ی برابر با داده‌ها آغاز می‌شود. پس از استخراج ویژگی از تمام نمونه‌ها، فاصله‌ی بین داده‌ها با هم در فضای ویژگی‌ها محاسبه شده [26]. سپس دو داده‌ی با نزدیک‌ترین فاصله با یکدیگر در یک خوشه قرار می‌گیرند و مرکز این خوشه، میانگین دو داده‌ی متعلق به آن است. در مرحله‌ی بعدی بردار میانگین خوشه‌های ترکیب‌شده به عنوان نمونه‌ی ورودی جدید در نظر گرفته‌شده و این روند تکرار می‌شود. باز هم خوشه‌های نزدیک به هم با یکدیگر ترکیب می‌شوند تا در نهایت تنها یک خوشه باقی بماند. در پس این فرایند نموداری به عنوان دندروگرام^۲ به دست می‌آید که شاخه‌های آن نشان‌دهنده‌ی خوشه‌هایی است که با هم ترکیب شده‌اند و ارتفاع هر یک از شاخه‌ها نشان‌دهنده‌ی میزان فاصله‌ی بین این خوشه‌ها است.

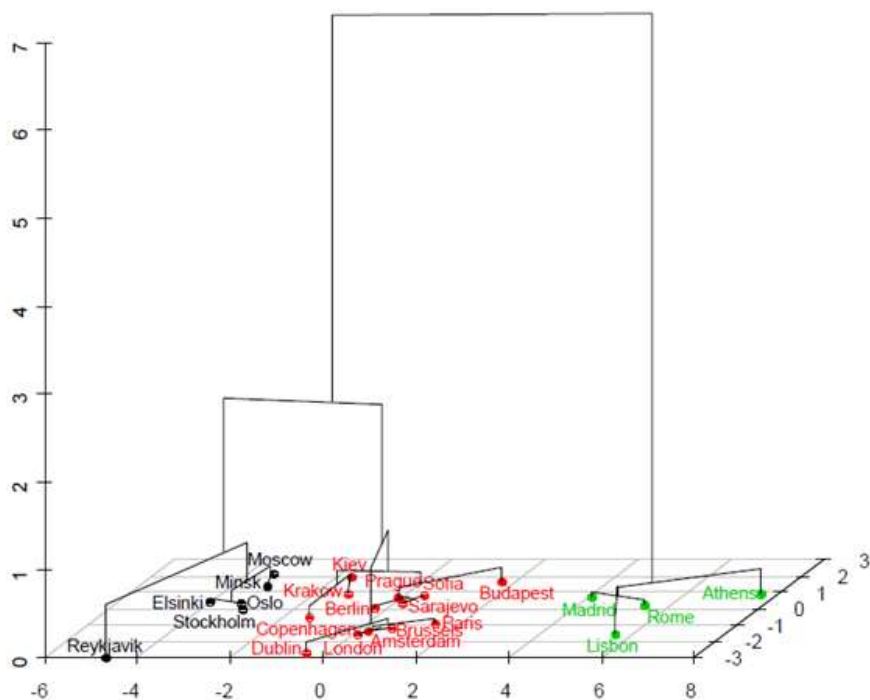
¹ Hierarchical
² Dendrogram

با قطع کردن نمودار دندروگرام در یک گره، مجموعه داده‌ها به تعدادی خوشه تقسیم می‌شوند؛ و برای اینکه متوجه شویم که بهتر است در چه گره‌ای از دندروگرام الگوریتم متوقف شود، باید افزایش‌های مختلف را بررسی کرده و یا با توجه به فاصله‌ی خوشه‌ها، واریانس در پخش‌شدگی نمونه‌های هر خوشه و تعداد خوشه‌ها، الگوریتم را متوقف کنیم [26].

در شکل ۱۵-۳ یک سری داده را مشاهده می‌کنید. در شکل ۱۶-۳ به صورت گرافیکی ادغام شدن داده‌ها مشخص شده است. در شکل ۱۷-۳ نمودار دندروگرام مربوطه و نتیجه‌ی متوقف کردن الگوریتم در قسمت مشخص شده را نشان می‌دهد که منجر به سه خوشه از داده‌ها شده است.

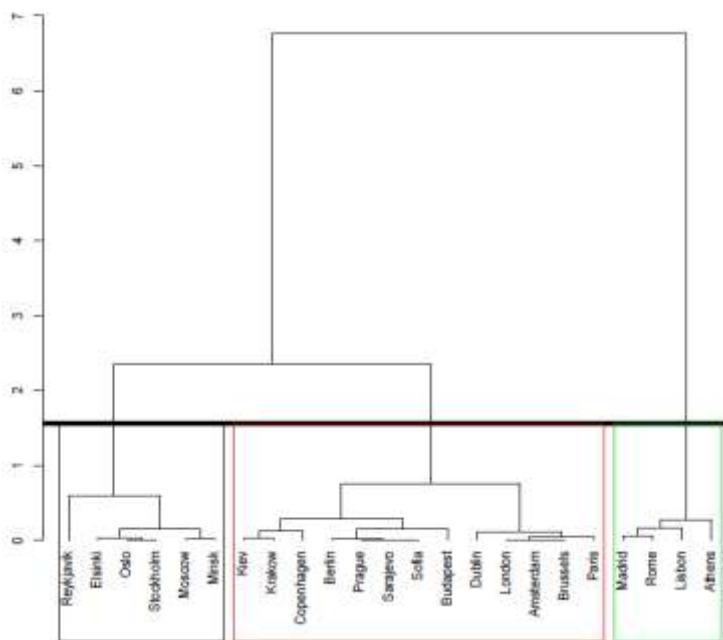


شکل ۱۵-۳ : سه دسته داده که با سلسله مراتبی خوشه بندی می‌شود [27]



شکل ۳-۱۶: نمودار دندروگرام سه بعدی، که پیشرفت مراحل الگوریتم نمایان است (ابتدا خوشه بندی‌هایی که در

ارتفاع پایین‌تر هستند انجام می‌شود) [27]



شکل ۳-۱۷: نمونه‌ی نمودار دندروگرام دو بعدی که با متوقف کردن الگوریتم در ناحیه‌ی مشخص شده این سه دسته

داده خوشه بندی شده‌اند [27].

۳-۶-۳- الگوریتم ایزودیتا

الگوریتم ایزودیتا از الگوریتم k میانگین ناشی شده است. [28,29] در این الگوریتم امکان تغییر تعداد خوشه‌ها در هر تکرار با حذف خوشه‌های کوچک و یا تقسیم خوشه‌های بزرگ ممکن می‌شود. در فاز اجرای الگوریتم زمان برترین بخش همچون k میانگین فاز نسبت دهی است، که در آن برای هر داده می‌بایست نزدیک‌ترین مرکز را انتخاب کنیم.

همان طور که ذکر شد این الگوریتم شبیه الگوریتم k میانگین است با این تفاوت که تعداد k خوشه همانند الگوریتم k میانگین تا آخر کار ثابت مانده و در صورتی که خوشه‌هایی، تعداد نمونه‌های کمی داشته باشند و یا فاصله‌ی دو خوشه از یک حدی کمتر شود مرکز خوشه‌ها را حذف و نمونه‌ها را در بین بقیه‌ی خوشه‌ها تقسیم می‌کند و در صورتی که پراکندگی داده‌ها در یک خوشه بیش از حد شود خوشه را جدا^۱ می‌کند. ورودی این الگوریتم علاوه بر مراکز اولیه‌ی خوشه‌ها، یک سری پارامترهای از پیش تعیین شده‌اند که روند الگوریتم را کنترل می‌کنند، که از آن جمله می‌توان به، حداکثر تعداد تکرار الگوریتم، حداکثر تعداد جفت خوشه‌ها برای ادغام در هر تکرار و حداکثر تعداد خوشه‌ها اشاره کرد.

۳-۷-۷- الگوریتم‌های طبقه بندی

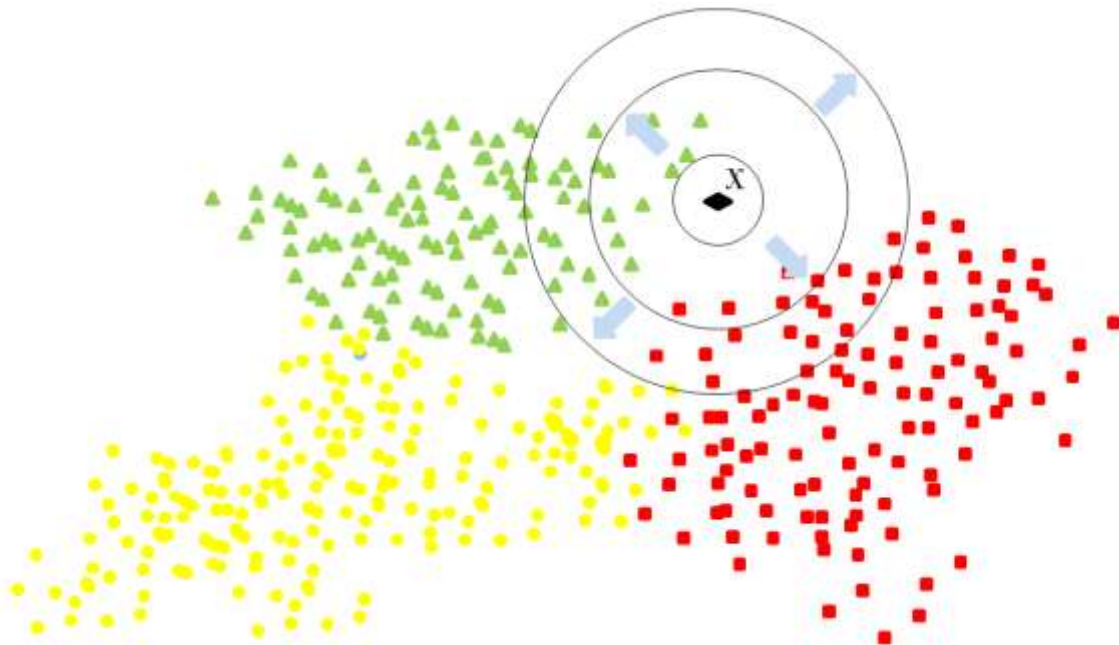
۳-۷-۱- k نزدیک‌ترین همسایه^۲

در طبقه‌بند k نزدیک‌ترین همسایه [17] همان‌طور که از اسم آن معلوم است هر نقطه از مجموعه داده با توجه به k نزدیک‌ترین همسایه آن طبقه‌بندی می‌شود، به این صورت که کلاس نقطه‌ی x با توجه به بیشترین تعداد آراء، از کلاس‌هایی که از k نزدیک‌ترین همسایه به آن جمع می‌شود، تخصیص می‌یابد. لذا شعاع همسایگی برای هر نقطه در فضا آنقدر افزایش می‌یابد تا فقط و فقط k

¹ split

² K Nearest Neighbor

داده در آن قرار گیرد (شکل ۳-۱۸).



شکل ۳-۱۸: X نمونه‌ی وارد شده و هر شکل نماینده‌ی یک کلاس است جهت تعیین کلاس نقطه‌ی X در الگوریتم KNN شعاع همسایگی برای در بر گرفتن تعداد k همسایه در حال افزایش است.

۳-۷-۲- شبکه‌های عصبی

اجزای سازنده یک شبکه عصبی نرون‌ها و وزن‌ها می‌باشند. رفتار شبکه نیز وابسته به ارتباط بین

اعضا است. در حالت کلی در شبکه‌های عصبی سه نوع لایه نورونی وجود دارد [30].

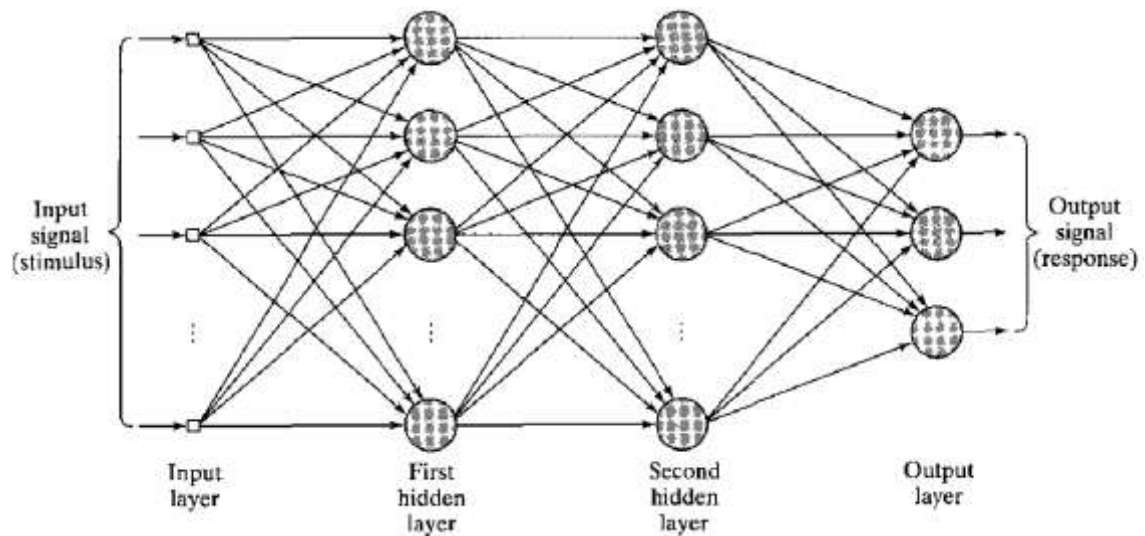
۱. لایه ورودی: دریافت اطلاعات خامی که به شبکه تغذیه شده است.

۲. لایه‌های پنهان: عملکرد این لایه‌ها به وسیله ورودی‌ها و وزن ارتباط بین آنها و لایه‌های

پنهان تعیین می‌شود.

۳. لایه خروجی: عملکرد واحد خروجی بسته به فعالیت واحد پنهان و وزن ارتباط بین

واحد پنهان و خروجی می‌باشد.



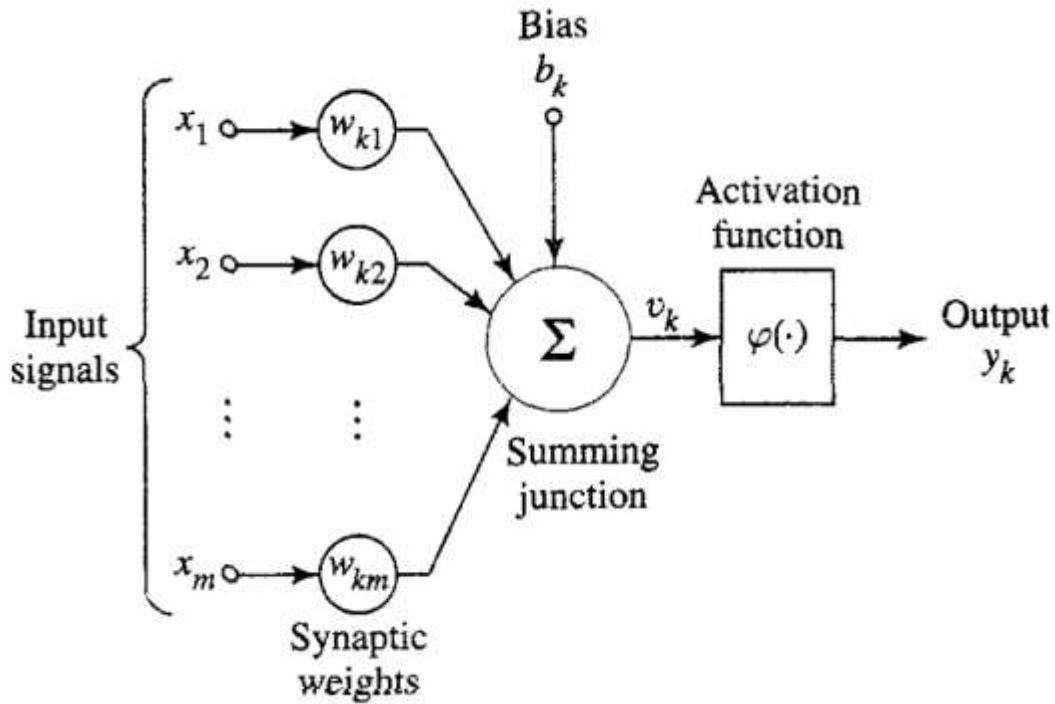
شکل ۳-۱۹: یک شبکه‌ی عصبی چهار لایه که دو لایه‌ی میانی و یک لایه‌ی ورودی و یک لایه‌ی خروجی دارد [30]

همان‌طور که در شکل ۳-۲۰ نشان داده شده است، در هر نود، تمام ورودی‌ها پس از ضرب شدن در وزن مربوط به خودشان، با هم جمع شده و پس از گذشتن از تابع فعال ساز^۱ (شکل ۳-۲۱)، مقدار خروجی آن نود را با توجه به رابطه‌ی (۳-۲۸) تعیین می‌کند [30].

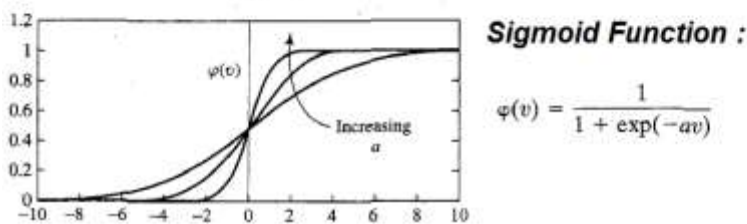
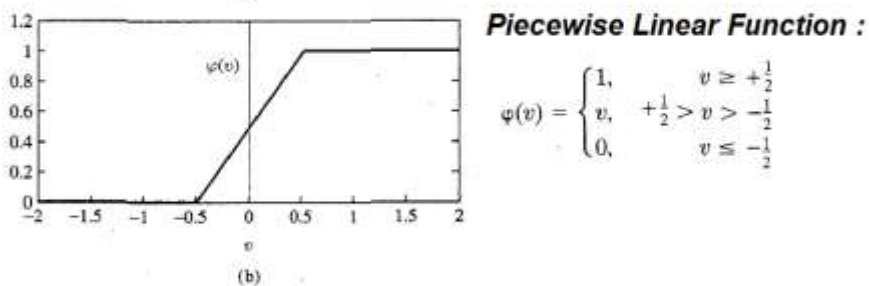
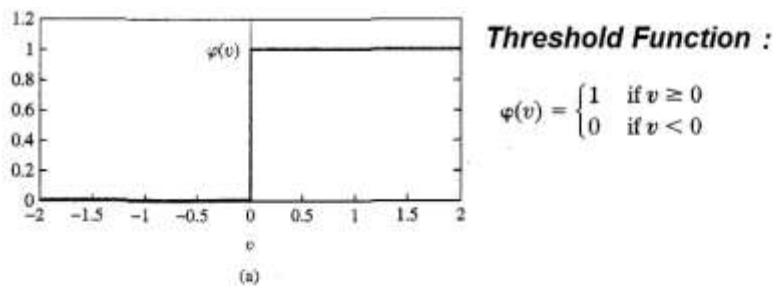
$$y_k = \varphi\left(\sum_j w_{kj} \times x_j\right) \quad (3-28)$$

در این رابطه w_{kj} وزن بین گره‌ی k ام و گره‌ی j ام می‌باشد، φ تابع فعال ساز و y_k خروجی گره‌ی k ام است.

^۱ Activation Function



شکل ۳-۲۰: تعیین مقدار خروجی در هر نود از شبکه [30]



شکل ۳-۲۱: برخی از توابع فعال ساز به همراه روابطشان [30].

یادگیری در شبکه‌های عصبی از این قرار است که اطلاعات بین نرون‌ها از طریق ارتباطات آنها رد و بدل می‌شود. هر یک از این روابط بین نرون‌ها دارای وزن مختص به خود هستند که در مقدار اطلاعات رد و بدل شده با سایر نرون‌ها ضرب شده و به مرور زمان این وزن‌ها تنظیم می‌گردند، یا به اصطلاح شبکه آموزش می‌بینند. در آموزش به روش عکس‌العمل انتشار به عقب^۱، وزن‌های شبکه عصبی به گونه‌ای تغییر داده می‌شوند تا رابطه‌ی (۲۹-۳) که تابع هدف حداقل مربعات خطا نامیده می‌شود، کمینه گردد، با کمینه شدن این تابع، خروجی‌های شبکه با مقادیر مورد انتظار، کمترین اختلاف را به لحاظ حداقل مربعات خطا خواهند داشت [30].

$$J(w) = \frac{1}{2} \sum_{k=1}^{\text{Class Number}} (t_k - y_k)^2 \quad (29-3)$$

که در این رابطه t_k هدف در کلاس k ام و y_k خروجی بدست آمده از گرهی k ام لایه‌ی خروجی است.

^۱ Resilient backpropagation (RP)

فصل ۴:

الگوریتم پیشنهادی

۴-۱- پایگاه داده

۴-۱-۱- آماده سازی صفحات داده

برای تهیه پایگاه داده، از مجموعه داده‌ی متنی که توسط شرکت هدی سیستم^۱ تهیه و ارائه شده استفاده کردیم. این پایگاه داده، از طریق منابع موجود در اینترنت، و نسخه‌های بر خط و دیجیتال از صدها کتاب فارسی، وبلاگ فارسی و ده‌ها روزنامه استخراج و تهیه شده است.

این مجموعه شامل حدوداً شش و نیم میلیون کلمه‌ی فارسی (۶۶۶۰۸۱۰) است که البته در این مجموعه اشتباهات نوشتاری زیادی هم وجود داشته که به روش‌های مختلف دستی و خودکار اصلاح شده است. نسخه‌ی تصحیح شده‌ی این داده‌ها شامل مجموعه‌ای از ۱۰۷۴۶ زیر-کلمه‌ی یکتای فارسی است [۳۱]، سپس با این مجموعه داده از دو طریقِ روبش اسناد چاپی و روش مصنوعی تولید تصویر، داده‌های با تفکیک پذیری ۹۶ نقطه بر اینچ، را تولید کردیم.

در این داده‌ها نسبت به داده‌های با تفکیک پذیری ۳۰۰ نقطه بر اینچ، محوشدگی در دندانها، مرزها و حفره‌ها، به هم چسبیدن نقطه‌ها، اتصال سرکش و در مواردی نقطه‌ها به بدنه‌ی اصلی زیر-کلمه مشاهده می‌شود. دو نمونه تصویر زیر کلمه با تفکیک پذیری ۳۰۰ و ۹۶ dpi، در شکل ۴-۱ نشان داده شده است.



(الف)



(ب)

^۱ www.Hodasystem.com

گنجکشاها

(ج)

گنجکشاها

(د)

شکل ۴-۱: دو تصویر با درجه تفکیک ۳۰۰ dpi و ۹۶ dpi

۴-۱-۲- تولید داده توسط روبشگر

پس از تهیه پایگاه داده‌ی خام از مجموعه‌ی زیر-کلمات، آن‌ها را با استفاده از چاپگر Hp LaserJet P2035، با تفکیک‌پذیری ۳۰۰ نقطه در اینچ چاپ کرده و با روبشگر HP ScanJet G3110 آن‌ها را با تفکیک‌پذیری ۹۶ نقطه در اینچ روبش کردیم، پایگاه داده‌ی مذکور نسبت به روش مصنوعی تولید داده نویز مختصری داشته و روشنایی متفاوتی دارد. شکل ۴-۲ یک تصویر نمونه‌ی روبش شده را نمایش می‌دهد.



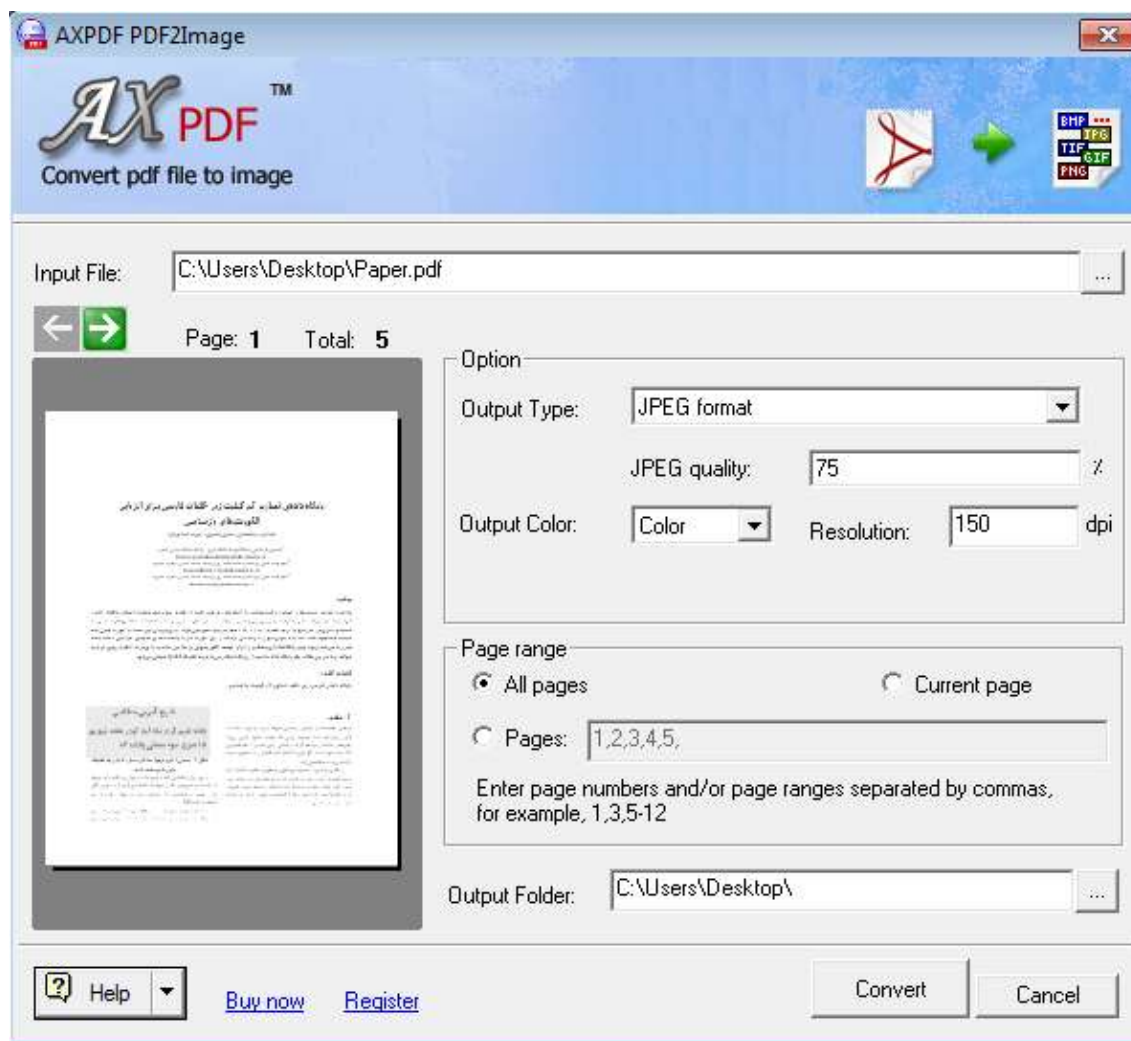
شکل ۴-۲: یک تصویر نمونه از زیرکلمات مجموعه داده.

۴-۱-۳- روش مصنوعی تولید داده

بار دیگر با دادن همان پایگاه مجموعه‌ی زیر-کلمات به نرم‌افزار doPDF که به شکل یک چاپگر مجازی عمل می‌کند و قابلیت تعیین تفکیک‌پذیری تصویر خروجی را دارد، تصاویری به دست می‌آوریم که از نظر قالب آماده ورود به نرم‌افزار PDF2Image که آن نیز قابلیت تنظیم تفکیک‌پذیری را داراست شده، و در نهایت تصاویری با تفکیک‌پذیری ۹۶ نقطه در هر اینچ به دست آوردیم. نمایی از این دو نرم‌افزار ذکر شده در شکل ۳-۴ و شکل ۴-۴ آورده شده است.



شکل ۳-۴: نمایی از نرم‌افزار doPDF



شکل ۴-۴: نمایشی از نرم افزار PDF2Image

یک نمونه از تصاویر تولیدشده به روش روبش از نسخه‌ی چاپی و روش مصنوعی، در شکل ۴-۵ نشان داده شده است. بعد از به دست آوردن صفحه‌های داده از زیر-کلمه‌ها با تفکیک‌پذیری مورد نظر، زیر-کلمات را از آن‌ها استخراج کرده و برای هر زیر-کلمه برچسب مناسب اختصاص یافته و سرانجام در پایگاه داده ذخیره می‌شود.

قلک شمشک م
چهلمین مستضه
هیخته متشنج

(الف)

قلک شمشک م
چهلمین مستضه
هیخته متشنج

(ب)

شکل ۴-۵: یک قطعه از تصویر تولیدشده‌ی زیر-کلمات با تفکیک‌پذیری ۹۶ نقطه در هر اینچ. (الف): به روش روبش از نسخه‌ی چاپی، (ب): به روش مصنوعی.

۴-۱-۴- پیش پردازش و بهبود کیفیت تصاویر

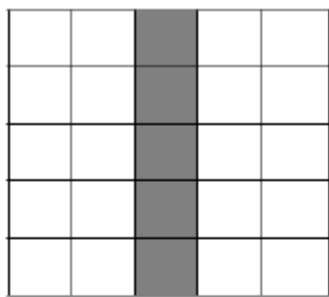
بهرتر است بر روی زیر-کلمات، عمل پیش‌پردازش را برای حذف نویز از جمله فلفل نمکی^۱ اعمال

کنیم [23]، در این صورت باید برای زیرکلمه‌های مورد آزمون‌ی که از روبشگر می‌گیریم این عملیات به

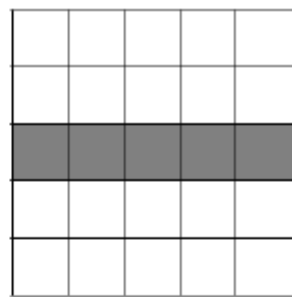
^۱ Salt & Pepper

طور مشابه انجام شود؛ در این اینجا چون که این نویز (فلفل نمکی) در اندازه‌هایی با پیکسل‌های کوچک که از عرض قلم کوچک‌تر و حداکثر در ابعاد ۵ در ۵ پیکسل، هستند، از طریق عملیات شکل شناسی، با استفاده از عنصرهای ساختاری که در شکل ۴-۶ ارائه شده حذف می‌شوند البته حذف آنها از طریق مساحت‌های کوچکتر از 5×5 نیز امکان پذیر است. شیوه‌ی اعمال الگوریتم برای تصویر A به صورت معادله‌ی (۱-۴) است که I نتیجه‌ی نهایی از اجتماع این فرسایش‌ها بدست می‌آید.

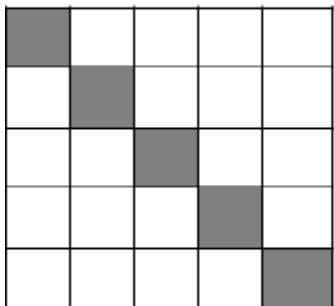
$$I = (A \ominus R_1) | (A \ominus R_2) | (A \ominus R_3) | (A \ominus R_4) | (A \ominus R_5) | (A \ominus R_6) \quad (1-4)$$



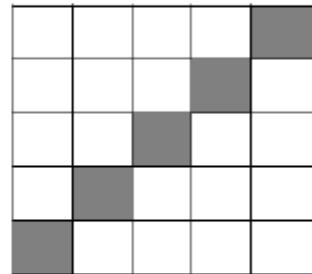
R_1



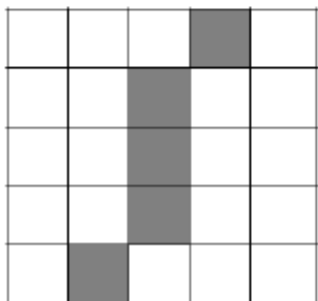
R_2



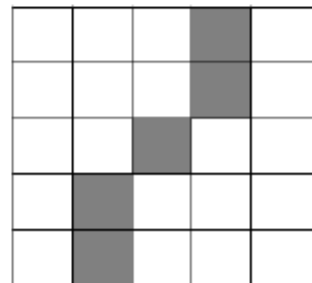
R_3



R_4



R_5



R_6

شکل ۴-۶: المان‌های ساختاری برای حذف نویز [23]

۴-۱-۵- جداسازی و برچسب گذاری زیر-کلمات

پس از تهیه‌ی صفحات شامل زیر-کلمات، هر یک از زیر-کلمه‌ها برچسب زنی شدند تا برای استفاده در مرحله‌ی آموزش و آزمون قابل تشخیص باشند، به این صورت که برای جداسازی آنها از افکنش‌های افقی و عمودی استفاده شد و مختصات سلول‌های مستطیل شکل حاوی زیر-کلمات را بدست آوردیم و پس از جدا سازی تقریبی زیر-کلمات حاشیه‌های سفید اطراف آنها حذف شده و هر سلول به صورت یک فایل تصویری مستقل با پسوند bmp ذخیره شدند.

به روش‌های مختلفی می‌تواند این برچسب زنی انجام شود، در اینجا برچسب زنی هر تصویر زیر-کلمه با یک رشته نام‌گذاری شده به صورتی که از چپ عدد بیان شده شماره‌ی صفحه‌ای که زیر-کلمه از آن استخراج شده و پس از زیر خط اول، عدد مذکور شماره‌ی خطی که زیر کلمه در آن صفحه بوده و پس از زیر خط دوم شماره‌ی زیر-کلمه، در آن سطر را نشان می‌دهد.

شکل ۴-۷ تصویر کلمه ای که از صفحه‌ی ۱۶ استخراج شده را نشان می‌دهد این تصویر ۱۵ امین زیر کلمه در خط ۶ ام از این صفحه می‌باشد و با اسم "۱۵_۶_۱۶" ذخیره شده است.



شکل ۴-۷: زیر-کلمه‌ی پانزدهم در خط ششم از صفحه شانزدهم، که با نام "۱۵_۶_۱۶" ذخیره شده است

۴-۱-۶- اطلاعات آماری

۴-۱-۶-۱- پایگاه داده‌ی متنی شرکت هدی سیستم

اطلاعات آماری اولیه از زیر-کلمات قبل از تصحیح سازی در جدول ۴-۱ آمده است، سطر اول بیان کننده‌ی تعداد زیر-کلمات یکتا با تعداد حروف مختلف است و سطر دوم تعداد زیر-کلمات یکتایی را که فقط یک بار در تمام متون آمده‌اند، بیان می‌کند، از این سطر می‌توان برای تصحیح کردن زیر-

کلمات استفاده کرد زیرا که آنها غالباً آسامی خاص و یا اشتباه تایپی بوده‌اند.

جدول ۴-۱: اطلاعات آماری اولیه‌ی زیر-کلمات قبل از تصحیح [۳۱]

طول زیر کلمه	۲ حرفی	۳ حرفی	۴ حرفی	۵ حرفی	۶ حرفی	۷ حرفی	۸ حرفی	۹ حرفی	مجموع
تعداد یکتا	۶۱۶	۴۱۵۷	۸۹۵۸	۶۳۶۲	۲۲۶۹	۴۹۱	۴۶	۹	۲۲۹۰۸
یکبار تکرار	۵۱	۷۱۸	۲۷۶۵	۲۷۶۸	۱۲۴۶	۲۹۹	۳۰	۷	۷۸۸۴
جمعیت	۴۷۰۴۶۲۸	۱۷۵۰۹۳۴	۵۸۴۵۲۵	۱۴۷۱۳۷	۲۵۳۷۴	۲۳۳۸	۱۳۸	۱۴	۷۲۱۵۰۸۸

در جدول ۴-۲ نیز تعدادی از پر تکرارترین و کم تکرارترین زیر-کلمات بدست آمده آورده شده است.

جدول ۴-۲: تعدادی از پر تکرارترین و کم تکرارترین زیر-کلمات [۳۱]

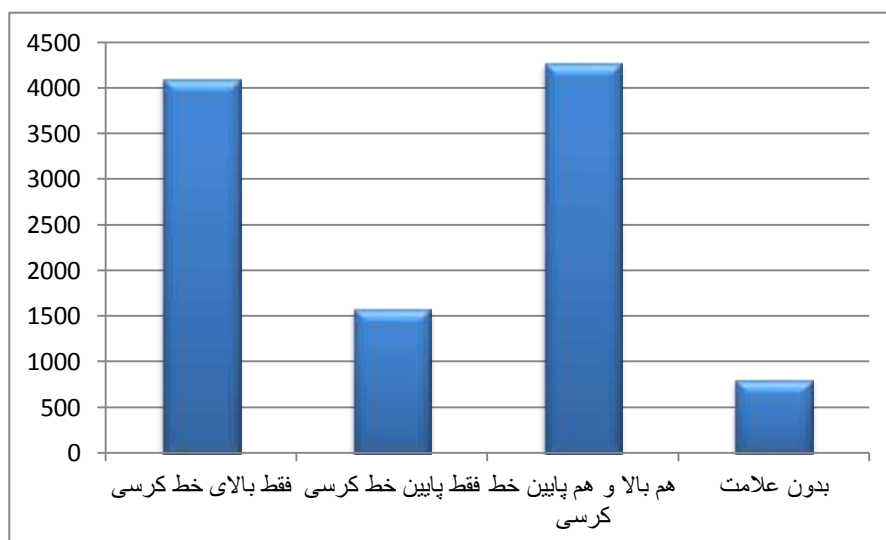
تکرار	زیر کلمه	تکرار	زیر کلمه
۲	سنسن	۱۸۱۲۲۵	به
۲	طلم	۱۷۶۳۱۰	با
۲	متبلو	۱۶۷۴۱۵	که
۲	میبلر	۱۶۶۹۸۰	بر
۲	تبخشی	۱۳۷۹۱۳	ند
۲	مضیقا	۱۲۳۷۸۶	ست
۲	نتیینگتو	۱۲۰۵۳۴	خو
۲	خشککنها	۱۱۰۲۹۵	تو
۲	بنهفتمی	۲۱۳۷۷	گفت
۲	لمستنیر	۲۰۰۲۶	کند

۲	تکنسینها	۱۹۳۷۹	همه
۲	فمینیستی	۱۸۸۶۰	جها
۱	ظز	۲۰۰۲۶	کند
۱	بغ	۱۷۱۲۴	نیست

۴-۱-۶-۲- پایگاه داده‌ی تصاویر کم کیفیت زیر کلمات

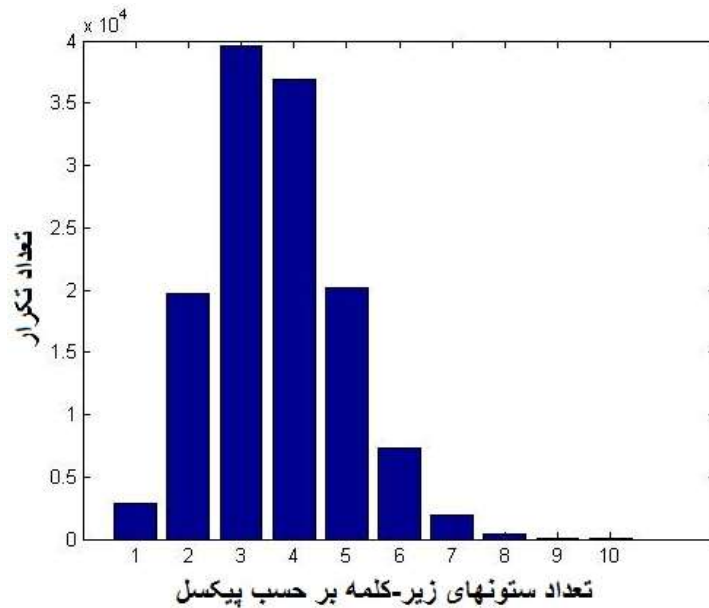
این پایگاه داده شامل ۱۲۸۹۵۲ تصویر از ۱۰۷۴۶ زیر-کلمه است، که در چهار اندازه‌ی ۱۱ تا ۱۴ آورده شده‌اند. از این تعداد، ۸۵۹۶۸ تصویر با روبش نسخه‌ی چاپی و ۴۲۹۸۴ تصویر با روش مصنوعی تولید شده‌اند.

یکی از ویژگی‌های مناسب که برای خوشه بندی زیر-کلمات می‌توان از آن استفاده کرد، توجه به نقاط آنهاست؛ اما به علت محوشدگی لبه‌ها و نویزی شدن زیر-کلمات خصوصاً قسمت‌های کوچک‌تر مثل نقاط، تفکیک کردن یک، دو و یا سه نقطه از هم بسیار دشوار است ولی دانستن اینکه علامتی همچون نقاط در چه موقعیتی از خط کرسی است این خود یک ویژگی تفکیک کننده است، در شکل ۴-۸ اطلاعات آماری از این قسم آورده شده است.

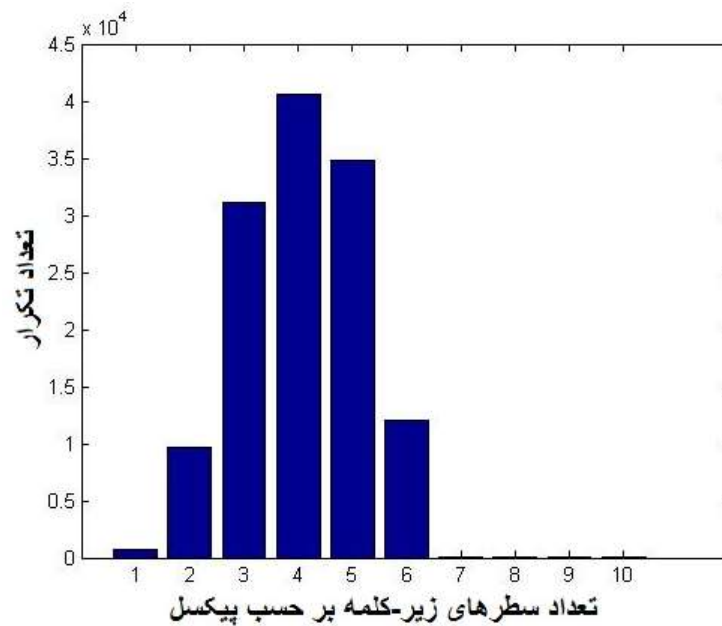


شکل ۴-۸: آمارگان محل قرار گیری نقاط در زیر-کلمات نسبت به موقعیت خط کرسی

در شکل ۹-۴ پراکندگی زیر-کلمات برحسب اندازه‌ی طول و عرض آنها در این پایگاه داده آورده شده است، از قسمت الف در این شکل اطلاعاتی همچون کشیدگی زیر-کلمات در راستای افق، مثل "مستغنیست" و از قسمت ب اطلاعات مربوط به زیر-کلماتی که کشیدگی در طول دارند بدست می‌آید مانند "با".



(الف)



(ب)

شکل ۹-۴: اطلاعات آماری پراکندگی زیر-کلمات برحسب اندازه‌ی آنها. (الف): برحسب طول، (ب): بر حسب عرض

از زیرکلمه‌های پایگاه داده در چهار اندازه‌ی ۱۱-۱۴ و از هر اندازه دو نمونه‌ی حاصل از روبشگر و یک نمونه‌ی مصنوعی تولید شد. سپس زیر-کلمات از این پایگاه داده استخراج شده و برچسب زده می‌شوند از این تعداد ۷۵٪ زیر-کلمات برای آموزش و ۲۵٪ برای آزمایش تقسیم می‌شوند، با توجه به اینکه از هر زیرکلمه ۱۲ نمونه تولید شده از هر زیرکلمه ۹ تا برای آموزش و ۳ تا برای آزمایش در نظر گرفته می‌شود.

۴-۲- نرمال‌سازی برای استخراج ویژگی

برای اینکه بتوان از تمام زیر-کلمه‌ها اعم از بزرگ‌ترین زیر-کلمه تا کوچک‌ترین آن‌ها به تعداد مساوی ویژگی استخراج کرد، باید همه تصاویر را به ابعادی یکسان درآورد. به همین منظور دو روش برای یکسان کردن ابعاد همه‌ی زیر-کلمه‌ها بکار می‌بریم و این روش‌ها را با هم در بخش نتایج مقایسه می‌کنیم.

در روش اول زیر-کلمه را تغییر اندازه داده و آن را در مرکز قابی که برای همه‌ی زیر-کلمه‌ها یکسان است قرار می‌دهیم. در روش دوم با حفظ نسبت طول به عرض، زیر-کلمه را تغییر اندازه داده و این آماده سازی انجام می‌شود [۵] در این حالت اگر زیر-کلمه از نظر ابعاد خیلی کوچک‌تر از قالب باشد شاهد کشیدگی، اگر خیلی بزرگ‌تر از قالب باشد دچار جمع شدگی می‌شود، نتیجه‌ی این دو روش در شکل ۴-۱۰ نشان داده شده است.

فلسفه

(الف)

فلسفه

(ج)

فلسفه

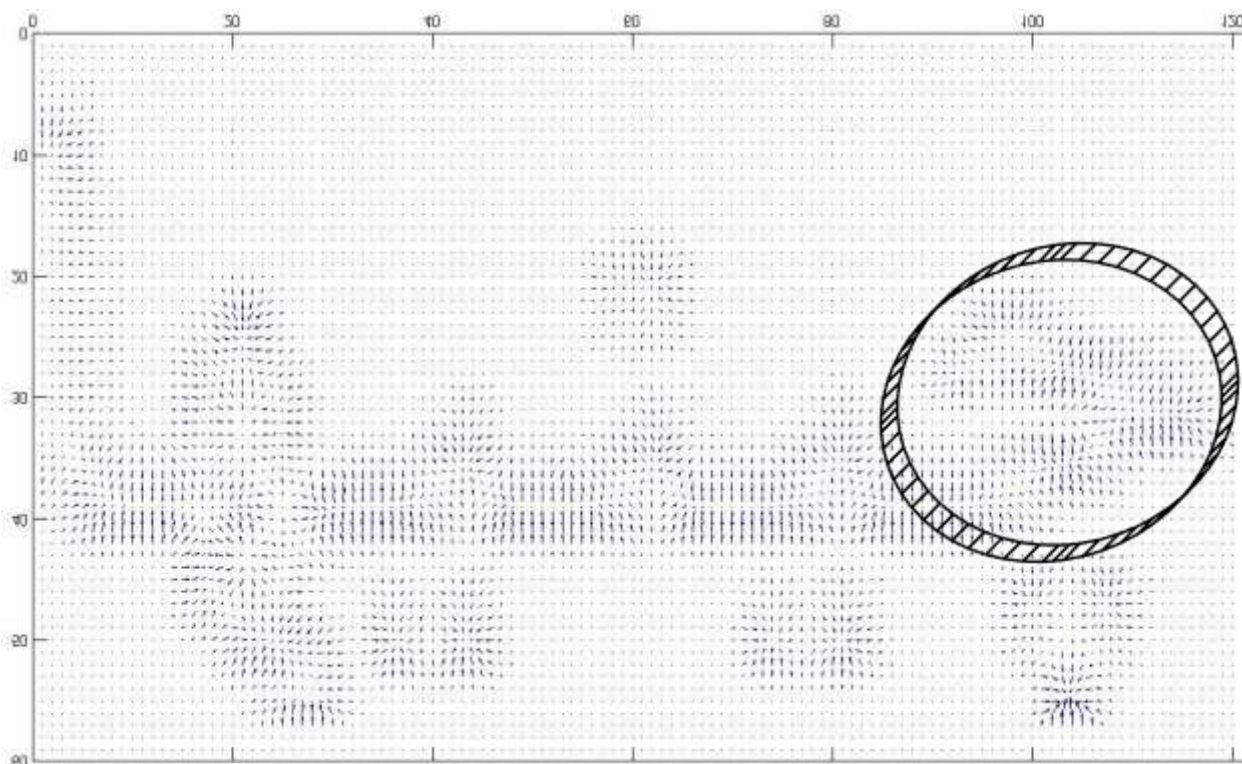
(ب)

شکل ۴-۱۰: یکسان‌سازی ابعاد زیر-کلمه‌ها. (الف): زیر-کلمه‌ی اصلی که برای نمایش بهتر با بزرگ‌نمایی نشان داده شده است، (ب): هم‌اندازه کردن زیر-کلمه به روش قرار دادن در مرکز قاب (ج): هم‌اندازه سازی با حفظ نسبت طول به عرض

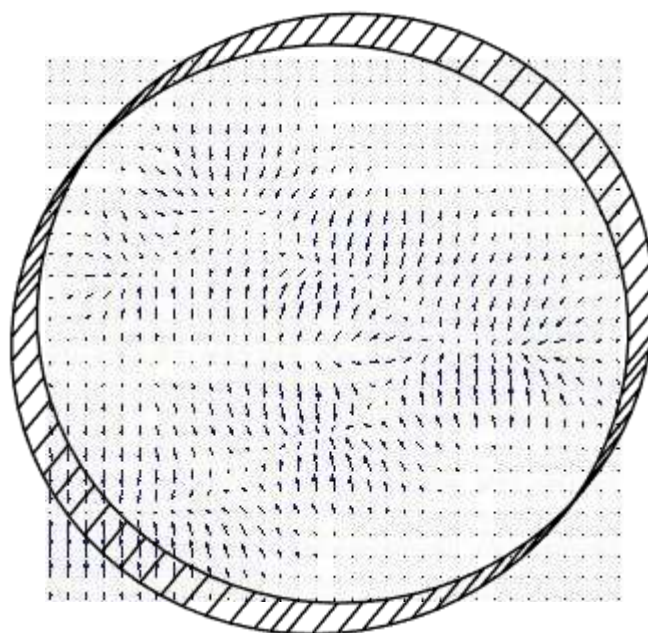
۳-۴- پارامترها در استخراج ویژگی‌ها

۴-۳-۱- هیستوگرام گرادیان

زیرکلمه‌ها را برای استخراج این ویژگی به ابعاد 60×120 نرمال کرده و تصویر نرمال شده را به پنجره‌هایی با ابعاد 20×20 تقسیم کردیم این کار برای حفظ مختصات نویسه‌ها انجام شد سپس روی هر یک از این پنجره‌ها، با فیلتر رابرت عمل گرادیان‌گیری صورت گرفت. در شکل ۴-۱۱ نتیجه گرادیان مشاهده می‌شود. دایره‌ی مثلثاتی به ۸ قسمت کوانتیزه شد؛ با این وصف طول بردار ویژگی‌ها برای هر زیرکلمه ۱۴۴ بدست آمد.



شکل ۴-۱۱: نتیجه‌ی گرادیان با عملگر رابرت



شکل ۴-۱۲: بزرگ‌نمایی از یک قسمت از شکل ۴-۱۱

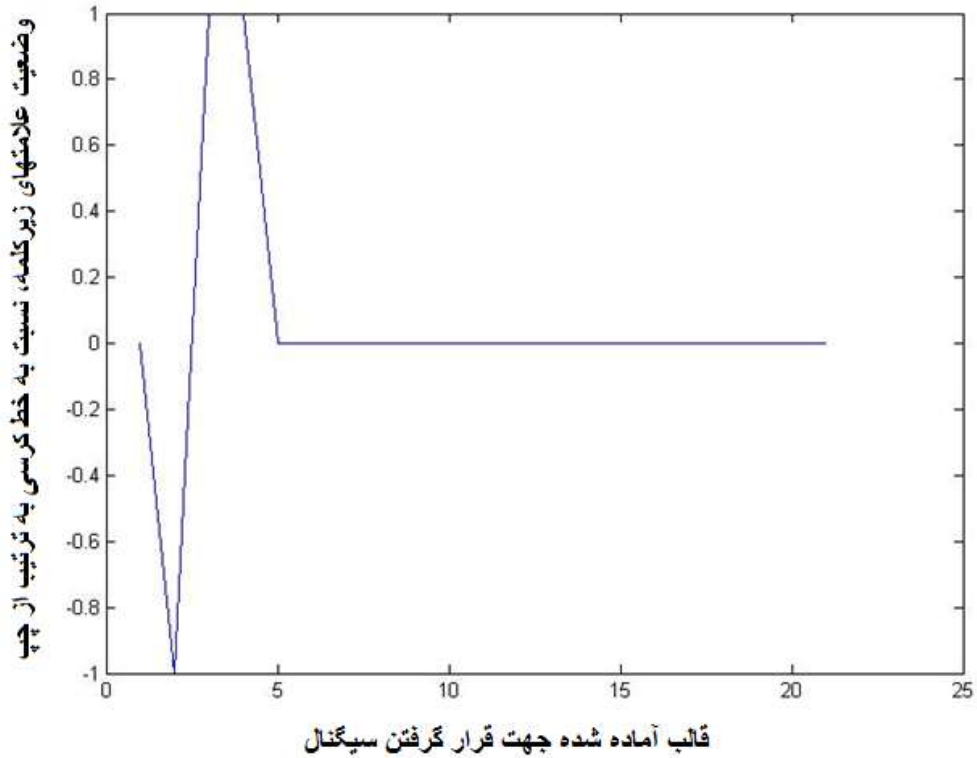
۴-۳-۲- میانگین بلوکی، مکان مشخصه، افکنش و LBP

برای استخراج هر یک از این ویژگی‌ها ابعاد تصاویر را به 24×48 نرمال کردیم، پنجره انتخابی در

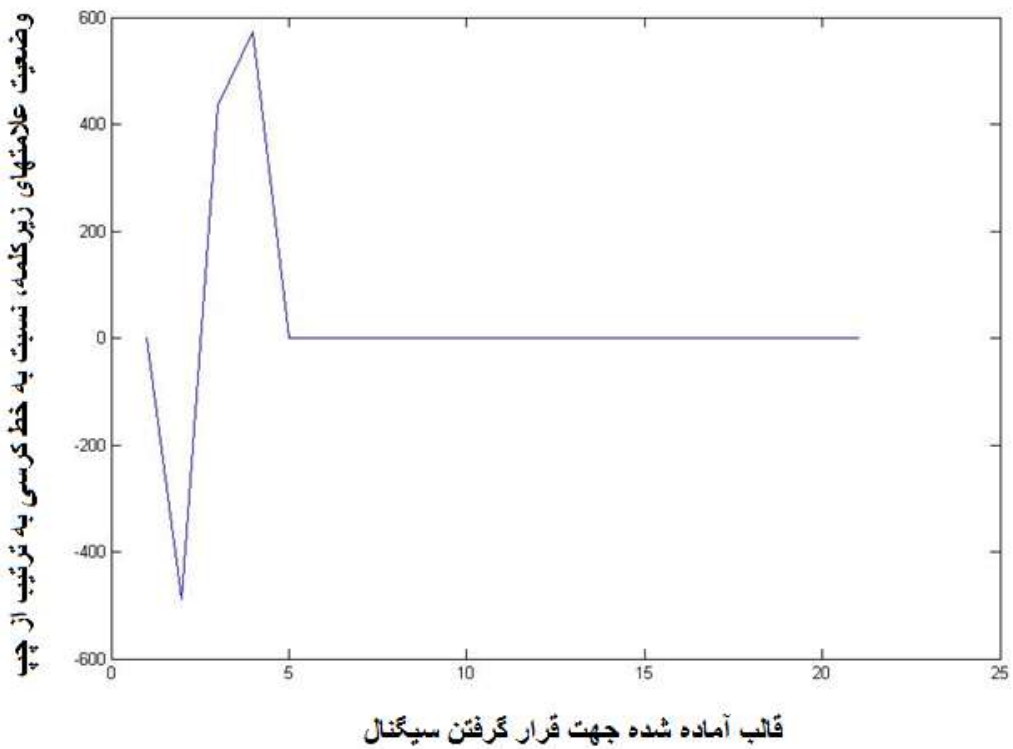
میانگین بلوکی 4×4 انتخاب شد و برای استخراج ویژگی مکان مشخصه، محدودیت را در راستای افقی در جهت‌های چپ و راست به ۳ و در جهت‌های بالا و پایین به ۲ قرار دادیم، برای استخراج ویژگی افکنش نیز در دو راستای عمودی و افقی تصویر سازی صورت گرفت و تعداد ۸ همسایه برای استخراج ویژگی LBP در نظر گرفته شد.

۴-۳-۳- سیگنال علامت

به دلیل اینکه کیفیت تصویر زیر-کلمه بسیار کم و احتمال چسبیدگی نقاط آن به هم بسیار زیاد است و از طرفی شمارش آنها از طریق مساحت نیز امکان پذیر نیست (زیرا مساحت سه نقطه و دو نقطه در خیلی از این تصاویر با این وضوح تقریباً برابر است) لذا بعد از حذف بدنه‌ی زیرکلمه و قبل از استخراج این ویژگی، بر روی نقاط زیرکلمه که به فاصله‌ی یک یا دو پیکسل از هم هستند آغشته سازی کرده، سپس با توجه به حضور علامت‌ها در زیرکلمات نسبت به خط کرسی از چپ این ویژگی استخراج شده و در قالب آماده شده قرار می‌گیرد؛ صفر شدن قالب به معنای اتمام علامت‌ها در آن زیرکلمه هست. در شکل ۴-۱۳ حالت بدون وزن سیگنال علامت و در شکل ۴-۱۴ حالت وزن دار این ویژگی را برای زیرکلمه‌ی "شکستگیها" مشاهده می‌کنید.



شکل ۴-۱۳: ویژگی سیگنال علامت بدون وزن برای زیرکلمه‌ی "شکستگیها"، این ویژگی از سمت چپ از زیرکلمه استخراج شده است.



شکل ۴-۱۴: ویژگی سیگنال علامت وزن دار برای زیرکلمه‌ی "شکستگیها"، این ویژگی از سمت چپ از زیرکلمه استخراج شده است.

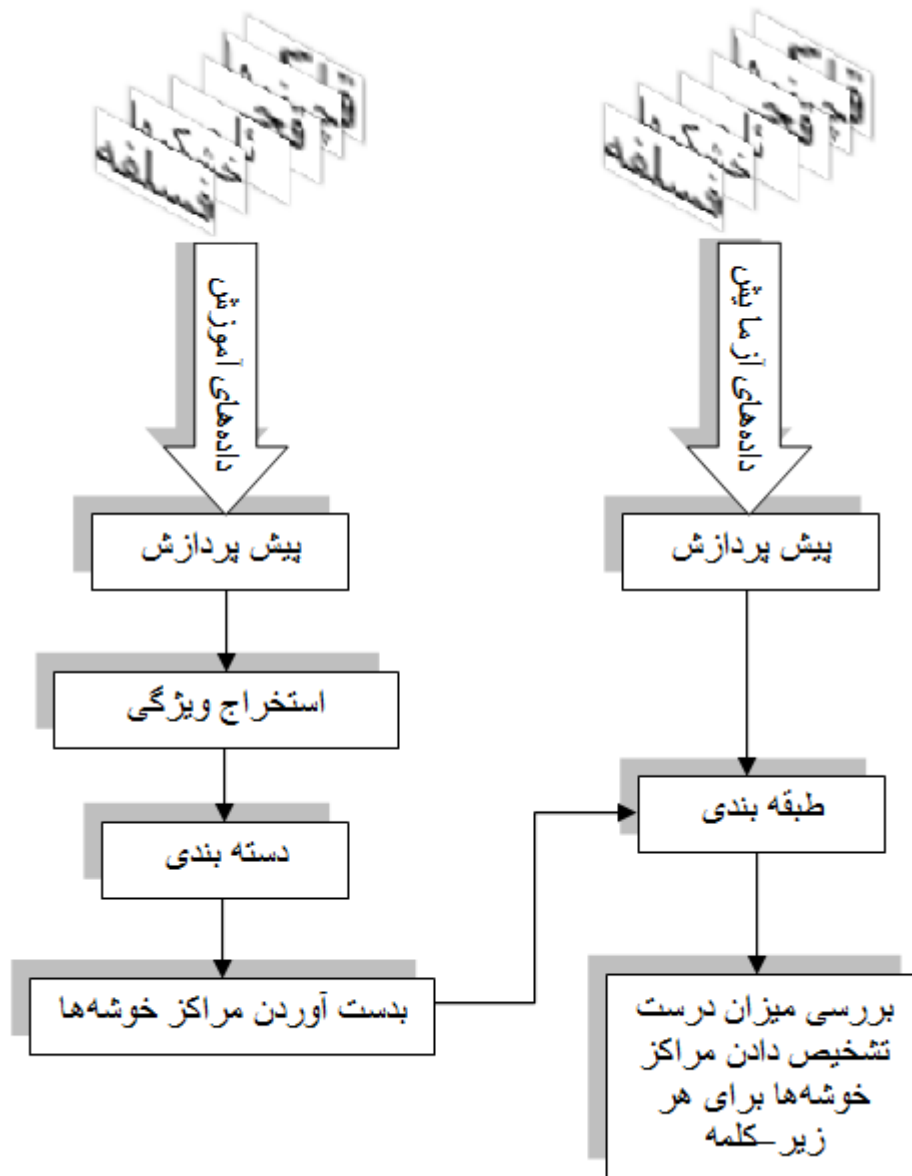
۴-۴- انتخاب سازگارترین و بهترین ویژگی، یکسان‌ساز ابعاد، معیار

فاصله و خوشه‌بند برای این‌گونه از تصاویر با وضوح پایین

برای پیدا کردن مناسب‌ترین معیارها و سنجش‌گرها برای این نوع خاص از تصاویر با وضوح پایین ترکیب آنها را با هم بر قسمتی از زیرکلمات مورد آزمایش قرار دادیم برای این منظور بعد از استخراج ویژگی از زیر-کلمات آموزش آنها را در دو روش جدا یک بار از روش سلسله مراتبی و بار دیگر از روش k میانگین خوشه‌بندی می‌کنیم و از خوشه‌های به دست آمده در هر روش مراکز خوشه‌ها، به دست می‌آیند.

برای ارزیابی هر یک از روش‌های خوشه‌بندی، از داده‌های آزمایش استفاده می‌کنیم. پس از استخراج ویژگی از هر داده، فاصله مراکز خوشه‌ها با آن را بدست می‌آوریم. انتظار از یک خوشه‌بندی صحیح آن است که مدل زیرکلمه‌ی مربوط به داده‌ی آزمایش، درون خوشه‌ای باشد که مرکز آن نزدیک‌ترین فاصله تا داده آزمایش را داشته باشد. بر این اساس خوشه‌ها را به ترتیب فاصله‌ی مرکز آنها از داده‌ی آزمون مرتب می‌کنیم، اگر خوشه‌ی صحیح در لیست مرتب شده، اولین باشد تشخیص خوشه کاملاً صحیح است. اگر خوشه صحیح در لیست خوشه‌ها، m امین باشد تشخیص خوشه در رتبه‌ی m نامیده می‌شود.

دیگرام کلی روش پیشنهادی کار در شکل ۴-۱۵ آورده شده است.

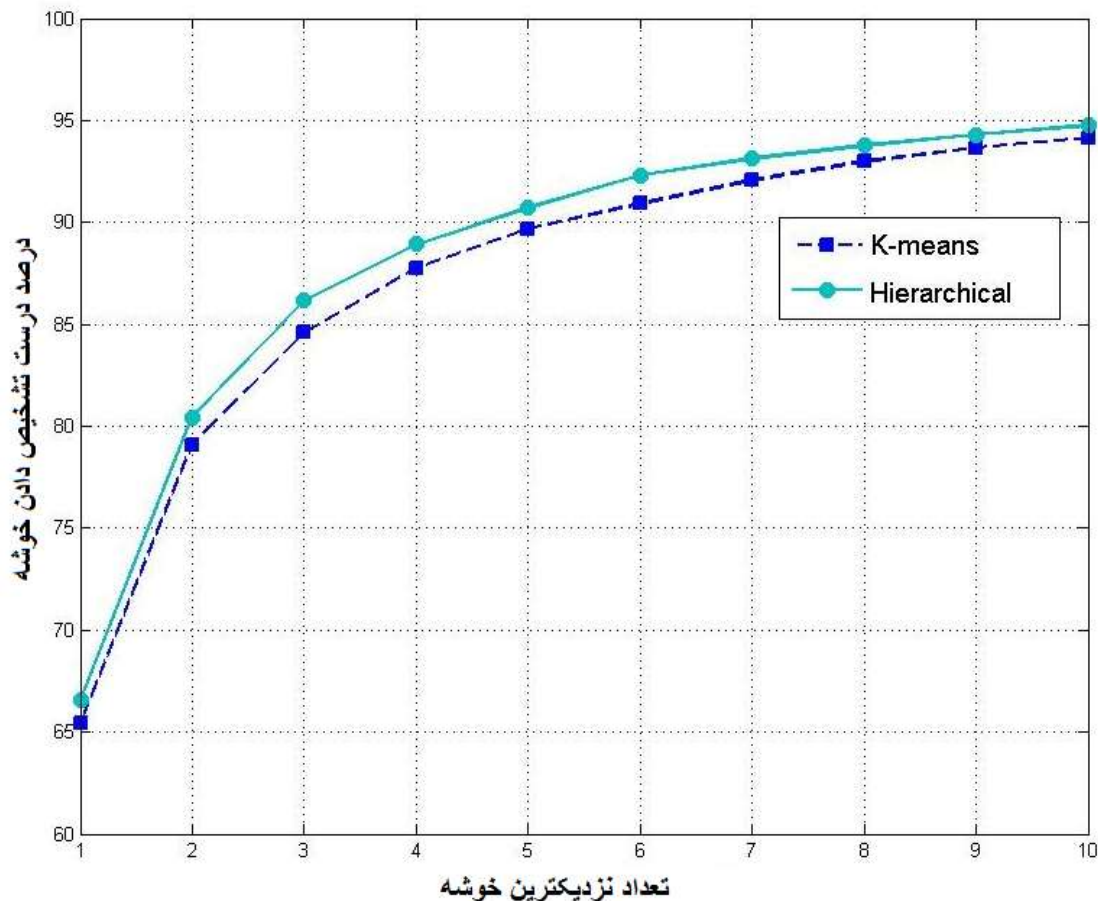


شکل ۴-۱۵: دیاگرامی از روش انجام کار

۴-۴-۱- گزارش نتایج

درصد تشخیص خوشه‌ها در روش خوشه‌بندی سلسله مراتبی به کمک داده‌های آزمایش تا رتبه‌ی $m = 10$ در شکل ۴-۱۶ نشان داده شده است، که از بردار ویژگی میانگین بلوکی، معیار فاصله‌ی همبستگی و از نظر قالب، یکسان‌سازی ابعاد با حفظ نسبت طول به عرض استفاده شده است. البته تا پنج رتبه‌ی اول از بهترین نتایج از نظر درصد تشخیص خوشه‌ی درست تقریباً شبیه به

هم هستند (مشابه شکل ۴-۱۶) ولی از نظر پراکندگی داده‌ها در خوشه‌ها تفاوت‌های چشم‌گیری در دو قسم روش خوشه‌بندی سلسله‌مراتبی و k میانگین دیده می‌شود.



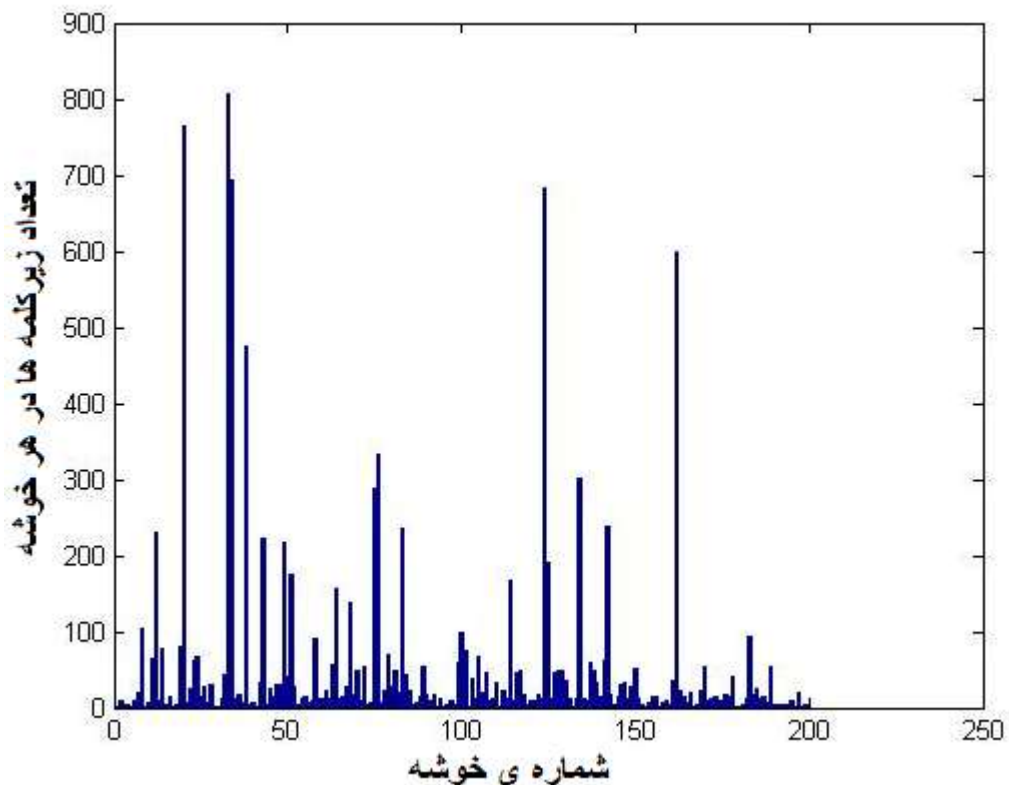
شکل ۴-۱۶: نمودار درصد، درست تشخیص دادن مراکز خوشه‌ها برای زیر-کلمه‌ها

همان طور که بیان شد پراکندگی داده‌ها در بین خوشه‌ها با توجه به روش خوشه‌بندی متفاوت است ولی این پراکندگی در یک نوع خوشه‌بندی مستقل از تغییر سایر پارامترها، از نظر تغییر ویژگی، تغییر قالب هم اندازه سازی و یا معیار فاصله، شبیه به هم هستند.

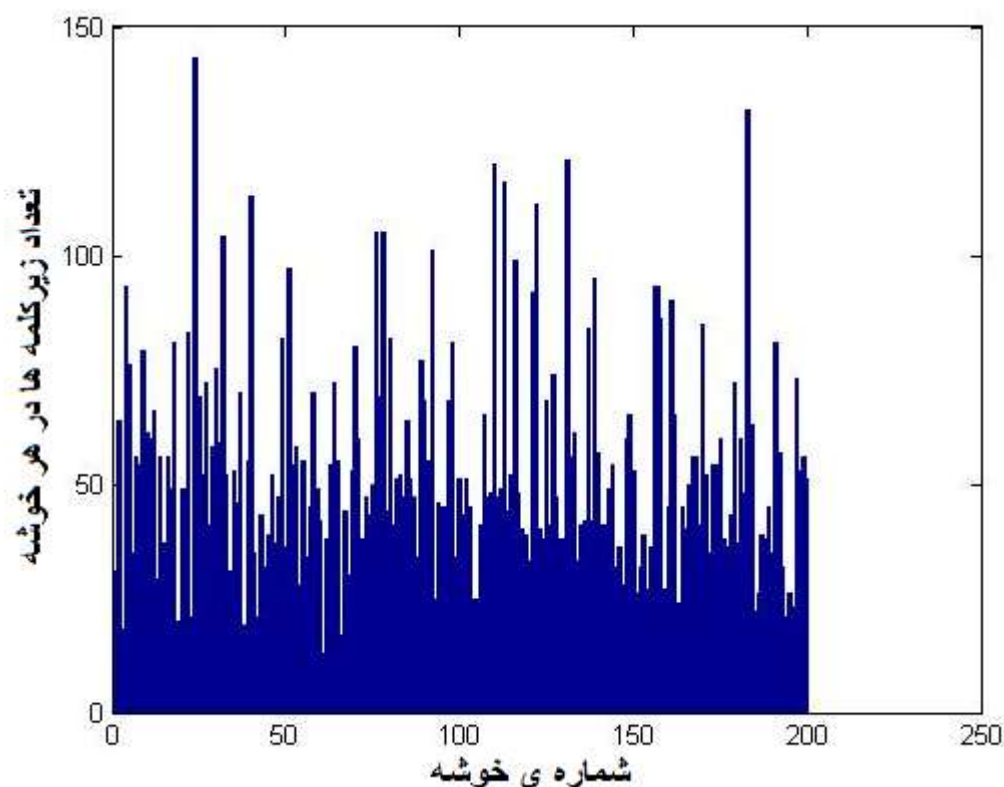
در خوشه‌بندی به روش k میانگین زیر-کلمه‌ها پخش‌شدگی بیشتری داشته و پراکندگی داده‌ها یکنواخت‌تر است، ولی در روش سلسله‌مراتبی داده‌ها به صورت ناهمگن در بین خوشه‌ها پخش‌شده‌اند، در بعضی از خوشه‌ها تعداد زیادی از زیر-کلمه‌ها و در بعضی دیگر تعداد کمی قرار گرفته

است.

البته پراکندگی خوشه بند سلسله مراتبی را با استفاده از الگوریتم ایزودیتا بهبود دادیم که به پراکندگی شبیه به خوشه بند k میانگین رسیدیم، ولی با کاهش حدود یک درصدی، در تشخیص درست خوشه مواجه شدیم که این درصدها بسیار نزدیک به نتایج k میانگین می‌شد لذا مقرون به صرفه است که از همان روش k میانگین استفاده شود. نتایج پراکندگی زیرکلمه‌ها در خوشه‌ها در دو گونه‌ی خوشه‌بندی سلسله‌مراتبی و k میانگین، در شکل ۴-۱۷ نشان داده شده است.



(الف)



(ب)

شکل ۴-۱۷: پراکندگی زیر-کلمه‌ها در بین خوشه‌ها (الف): در روش خوشه‌بندی سلسله مراتبی، (ب): خوشه‌بندی به روش k میانگین

نتایج پنج رتبه‌ی اول که درصد‌های بهتری در درست تشخیص دادن خوشه‌ها، احراز کرده‌اند در جدول ۳-۴ آورده شده است، که قالب همه، یکسان‌سازی ابعاد با حفظ نسبت طول به عرض است و ویژگی استخراج‌شده در هر پنج تای آن‌ها ویژگی میانگین بلوکی می‌باشد.

جدول ۳-۴: درصد‌های پنج رتبه‌ی اول در تشخیص درست خوشه، تا انتخاب ده خوشه‌ی نزدیک به زیر-کلمه‌ها

conrelation	cosine	cosine	spearman	conrelation	معیار فاصله
K میانگین		سلسله مراتبی			خوشه‌بندی
۶۵/۲۸	۶۵/۴۱	۶۵/۸۲	۶۴/۹۸	۶۶/۵۷	انتخاب اولین خوشه نزدیک

۷۸/۶۲	۷۹/۱۱	۷۹/۸۷	۸۰/۰۹	۸۰/۴۴	انتخاب دو خوشه نزدیک
۸۴/۵۲	۸۴/۵۹	۸۵/۳۸	۸۷/۰۳	۸۶/۱۵	انتخاب سه خوشه نزدیک
۸۷/۴۹	۸۷/۷۵	۸۸/۶۴	۹۰/۶۵	۸۸/۹۲	انتخاب چهار خوشه نزدیک
۸۹/۷۴	۸۹/۶۸	۹۰/۵۸	۹۳/۰۱	۹۰/۷۱	انتخاب پنج خوشه نزدیک
۹۱/۳۵	۹۰/۹۴	۹۲/۲۱	۹۴/۵۴	۹۲/۳۱	انتخاب شش خوشه نزدیک
۹۲/۳۶	۹۲/۰۵	۹۳/۱۹	۹۵/۵۰	۹۳/۱۲	انتخاب هفت خوشه نزدیک
۹۳/۰۴	۹۲/۹۷	۹۴/۰۷	۹۶/۴۳	۹۳/۷۷	انتخاب هشت خوشه نزدیک
۹۳/۷۱	۹۳/۶۵	۹۴/۷۱	۹۶/۸۹	۹۴/۳۱	انتخاب نه خوشه نزدیک
۹۴/۲۲	۹۴/۱۶	۹۵/۱۸	۹۷/۳۵	۹۴/۷۵	انتخاب ده خوشه نزدیک

هرچند درصدهای مربوط به روش خوشه‌بندی سلسله مراتبی بیشتر است ولی همان طور که در پراکندگی زیر-کلمه‌ها در خوشه‌ها دیده شد، در این روش خوشه‌بندی، داده‌ها، به پخش‌شدگی روش k میانگین بین خوشه‌ها تقسیم نمی‌شوند و چون در بعضی از خوشه‌ها، تعداد زیر-کلمه‌ها زیاد است با در نظر گرفتن اینکه در این مرحله، خوشه به درستی تشخیص داده شود احتمال خطا در مرحله‌ی بعد زیاد است.

پس با فرض اینکه، پراکندگی بهتر زیر-کلمه‌ها، دقت تشخیص زیر-کلمه را پس از تعیین خوشه در مراحل بعد افزایش داده و خطا در این مرحله را کاهش می‌دهد، می‌توان از این اختلاف درصد کوچک در تشخیص خوشه چشم‌پوشی کرد و روش خوشه‌بندی k میانگین را برگزید.

۴-۴-۲- نتیجه گیری

نتیجه این آزمایش نشان داد که استفاده از خوشه‌بندی سلسله مراتبی با معیار فاصله‌ی همبستگی بالاترین درصد تشخیص درست خوشه‌ها را به دنبال دارد.

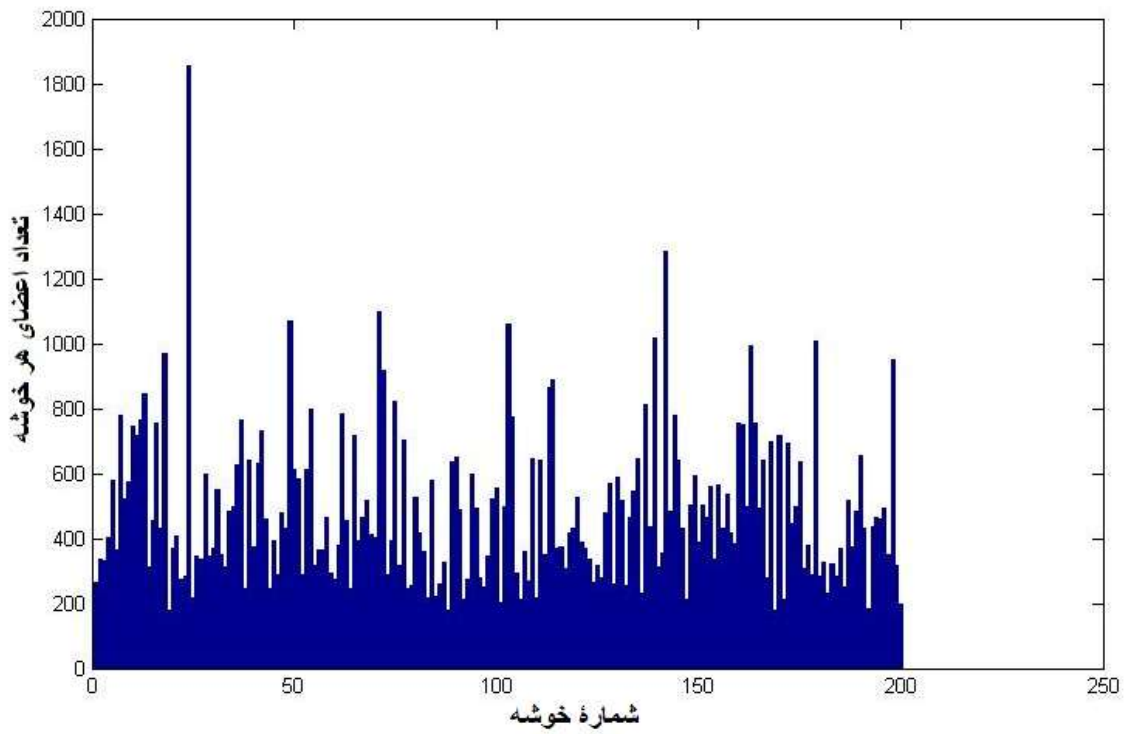
هرچند روش خوشه‌بندی سلسله مراتبی در رتبه بالاتری نسبت به روش k میانگین است ولی به علت

پخش نشدن یکنواخت داده‌ها در خوشه‌ها در روش سلسله مراتبی که سبب قرار گرفتن تعداد زیادی از زیرکلمات در بعضی از خوشه‌ها می‌شود، ضمن دور شدن از هدف خوشه بندی برای کاهش فضای انتخاب، باعث افزایش احتمال خطا در مراحل بعد نیز می‌شود، لذا با چشم پوشی از اختلاف ناچیز درصد خوشه‌بندی، روش k میانگین، بیشتر مورد توجه قرار گرفت. از لحاظ یکسان‌سازی ابعاد، تغییر اندازه با حفظ نسبت طول به عرض، نتایج بهتری نشان داد، در بهترین شرایط برای سایر پارامترها و برای شرایط مساوی برای هر دو، در روش یکسان‌سازی ابعاد با حفظ نسبت طول به عرض به درصد $66/57$ و با قرار دادن در مرکز قاب به 55 درصد در تشخیص اولین خوشه‌ی درست دست یافتیم.

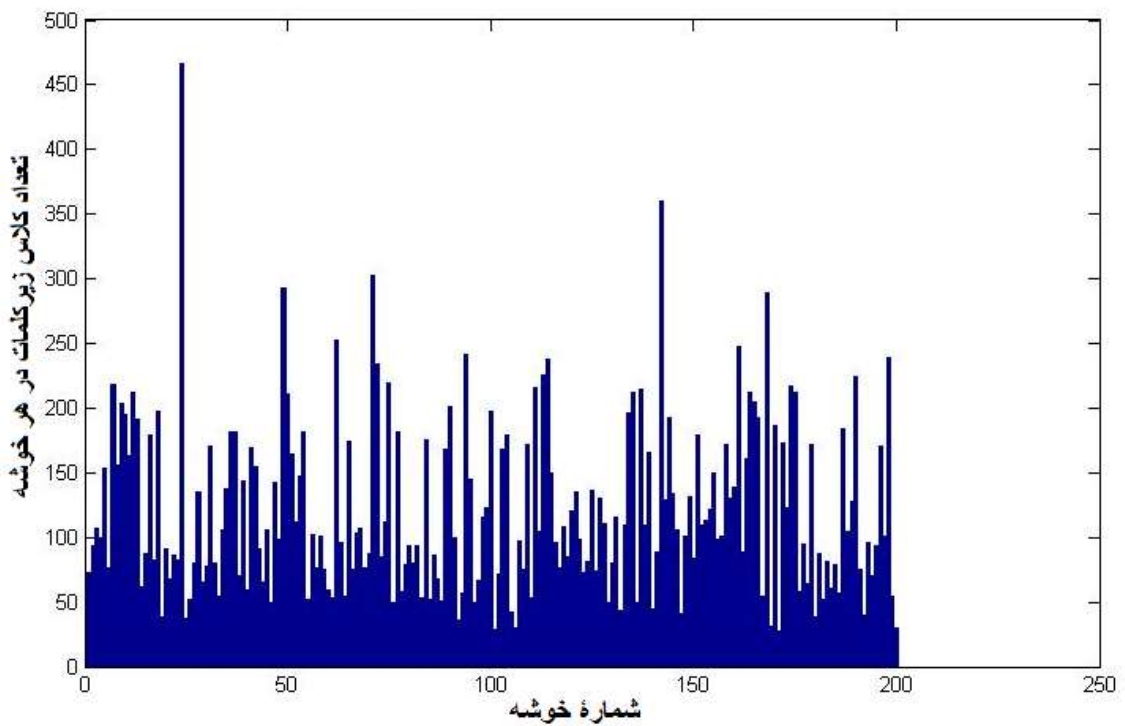
۴-۵- انتخاب طبقه بند

طبقه بندی پهنای قلم، تعدد شکل‌های مختلف زیرکلمه در هر کلاس که می‌بایست در فرایند بازشناسی لحاظ گردند را کاهش می‌دهد و سبب می‌شود که امر شناسایی، تنها به یک کلاس با یک پهنای قلم محدود گردد، لذا پس از بررسی این امر، مشاهده شد که تفاوت پهنای قلم در فونت ۱۱ با ۱۴ کمتر از یک پیکسل است و طبقه بندی از این لحاظ صورت نگرفت.

با توجه به نتایج قسمت قبل، با در نظر گرفتن ویژگی میانگین بلوکی و معیار فاصله کرولیشین و خوشه بندی به روش k میانگین ($k = 200$)، به خوشه بندی زیرکلمات پرداختیم که منجر به نتایج زیر شد، در شکل ۴-۱۸ پراکندگی زیرکلمات در هر خوشه نمایش داده شده است و تعداد کلاس‌های زیرکلمات در هر خوشه در شکل ۴-۱۹ قابل مشاهده است.



شکل ۴-۱۸ : تعداد اعضای خوشه‌ها.



شکل ۴-۱۹ : تعداد کلاس هر یک از خوشه‌ها.

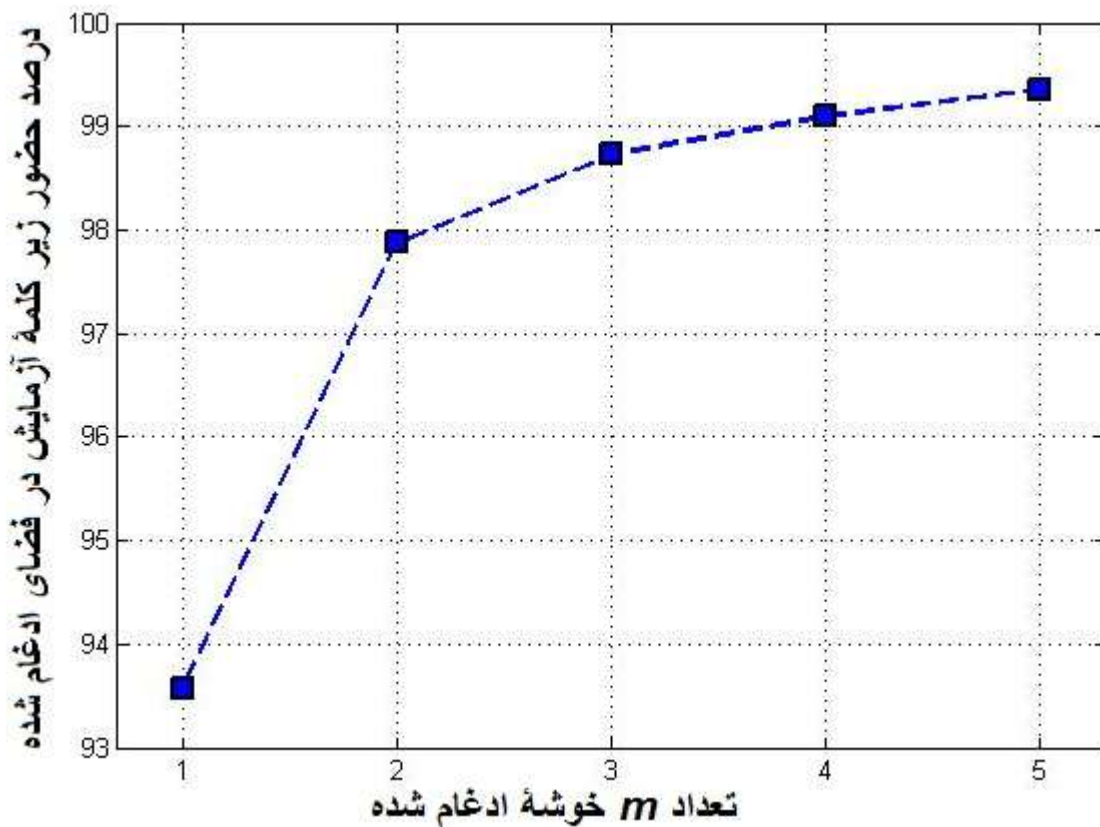
تعدادی از اعضای خوشه‌ی اول به عنوان نمونه از نتیجه‌ی این خوشه بندی در شکل ۴-۲۰ آورده

شده است.

ینکم ینکم ینکم ینکم ینکم ینکم ینکم ینکم
 لنگم لنگم لنگم لنگم لنگم لنگم لنگم لنگم
 ننگها ننگها ننگها ننگها ننگها ننگها ننگها ننگها
 تنگبا تنگبا تنگبا متکثر ینگلا ینگلا شگبا شگبا تنگها
 ملکبا ملکبا ملکبا نمکم چکم چکم چکم چکم
 مکم مکم مکم مکم مکم مکم مکم مکم
 شقبا شقبا شقبا بیقد بیقد چکها نیکتر نیکتر نیکتر مکم
 بیگم بیگم بیگم بیگم بیگم بیگم بیگم بیگم
 شکبو شکبو شکبو شکبو شکبو شکبو شکبو شکبو

شکل ۴-۲۰: تعدادی از اعضای خوشه‌ی یک.

پس از دسته‌بندی زیرکلمات آزمایش با خوشه‌بند k میانگین و بدست آوردن مراکز خوشه‌ها، این مراکز را نسبت به نزدیک بودن به هر یک از زیرکلمات آزمایش در فضای ویژگی‌ها مرتب کردیم و m خوشه‌ی نزدیک به هر زیرکلمه با یکدیگر ادغام شدند سپس حضور زیرکلمه‌ی آزمایش در این فضای جدید ادغام شده بررسی شد که به نتیجه‌ی شکل ۴-۲۱ منجر شد.



شکل ۴-۲۱: درصد حضور زیر کلمات آزمایش در m نزدیک‌ترین خوشه‌ی ادغام شده به زیر کلمه‌ها

همان‌طور که ملاحظه می‌شود با ادغام کردن تعداد مناسب از خوشه‌ها می‌توان حد نهایی درصد بازشناسی را تعیین نمود اما باید این نکته را هم مد نظر داشت که با افزایش تعداد خوشه‌های ادغام شده فضای جستجو نیز افزایش می‌یابد و هدف ما از خوشه‌بندی کاهش فضای جستجو بود لذا با در نظر گرفتن ۴ خوشه‌ی نزدیک به زیر کلمه‌ی آزمون به حد نهایی درصد ۹۹.۱٪ در بازشناسی بسنده می‌کنیم.

۴-۵-۱- شبکه عصبی

چنانچه بخواهیم از شبکه‌های عصبی برای طبقه‌بندی زیر کلمات استفاده کنیم باید به تعداد ویژگی‌ها، نود ورودی و به تعداد کلاس‌های زیر کلمات باید نود خروجی برای شبکه در نظر بگیریم، برای تعداد لایه‌های میانی نیز از آنجایی که نتیجه‌ی یک شبکه با یک لایه‌ی میانی با توابع فعال ساز

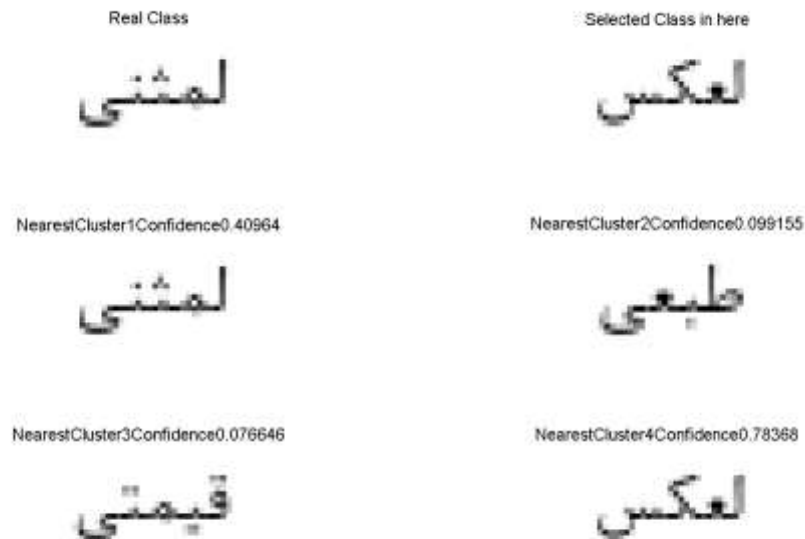
غیر خطی با نتیجه‌ی نهایی همین شبکه با در نظر گرفتن چند لایه یکی است [32]؛ لذا برای شبکه، یک لایه‌ی میانی در نظر می‌گیریم و از تابع فعال ساز سیگموئید با ثابت $a = 2$ استفاده می‌کنیم، تعداد نودها لایه‌ی میانی نیز به صورت تجربی و آزمون و خطا ۳۰ در نظر گرفته شده است.

در صورتی که بخواهیم برای تمام زیرکلمات یک شبکه عصبی را به صورت یکپارچه و کلی، تعریف و آموزش بدهیم چون باید برای هر کلاس زیرکلمه یک نود خروجی در شبکه در نظر بگیریم به تعداد نودهای خروجی زیادی در شبکه احتیاج است (از آنجایی که تعداد ویژگی‌ها به تعداد معمولی است تعداد نودهای ورودی شبکه متعارف است) که این باعث غیر قابل تعریف شدن شبکه می‌شود و یا قابلیت آموزش دادن به شبکه را ندارد لذا ابتدا زیرکلمات را خوشه بندی کرده و برای هر خوشه یک شبکه عصبی به صورت جداگانه تعریف می‌کنیم که نودهای ورودی به تعداد ویژگی‌ها است و به تعداد کلاس‌های زیرکلمات موجود در هر خوشه نود خروجی در نظر می‌گیریم؛ با در نظر گرفتن خوشه بندی قسمت قبل، برای مثال برای خوشه‌ی اول که ۷۳ نوع زیرکلمه در آن وجود دارد، تعداد نودهای خروجی در آن شبکه ۷۳ است و با فرض در نظر گرفتن ویژگی میانگین بلوکی، با توجه به نتیجه‌ی قسمت قبل، ۷۲ نود ورودی در شبکه داریم.

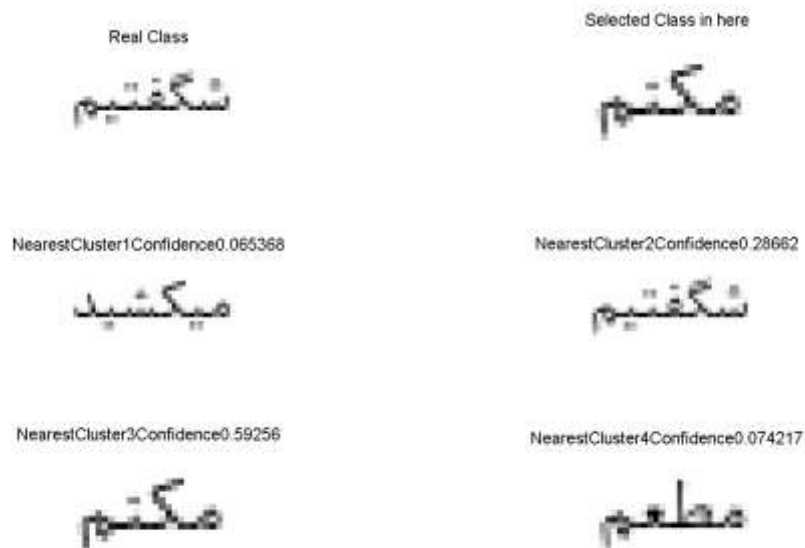
بعد از آموزش شبکه‌های عصبی به روش RP برای هر خوشه، در هر بار با دادن زیرکلمه‌ی آموزش به شبکه‌ی متناظر با خوشه‌ای که مرکز آن به زیرکلمه‌ی آزمون نزدیک‌ترین است یک نود در خروجی شبکه دارای بیشترین مقدار می‌شود که تعیین کننده‌ی کلاس برنده در خوشه است البته با توجه به شکل ۴-۲۱ چون همیشه زیرکلمه‌ی مورد نظر در نزدیک‌ترین خوشه نیست این کار برای ۴ خوشه‌ی نزدیک به زیرکلمه‌ی آزمون تکرار شد. که در هر خوشه نود برنده شده به عنوان کلاس منتخب در آن خوشه در نظر گرفته شده و این کار برای چهار خوشه‌ی نزدیک تکرار شد و مقدار نود خروجی برنده در شبکه به عنوان درصد اطمینان^۱ برای آن خوشه ذخیره شد، سپس برای انتخاب زیرکلمه از بین این چهار برنده در ۴ خوشه‌ی نزدیک به زیرکلمه‌ی آزمون، زیرکلمه‌ای که اطمینان

^۱ Confidence

بیشتری داشت، انتخاب شد، درصد بازشناسی زیرکلمه توسط روش ذکر شده ۱۳/۴٪ محاسبه شد؛ دو نمونه از نتیجه‌ی این انتخاب در شکل ۴-۲۲ آورده شده است.



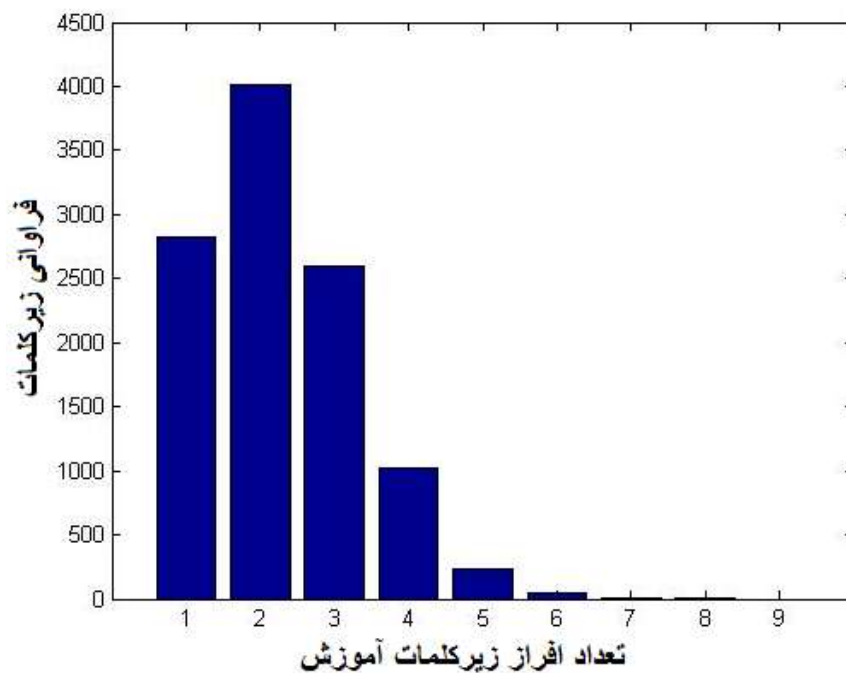
(الف)



(ب)

شکل ۴-۲۲: نتایج انتخاب زیرکلمه توسط روش شبکه‌های مصنوعی، که در هر شکل بالاترین زیرکلمه سمت چپ زیرکلمه‌ی آزمون بوده و در سمت راست آن زیرکلمه‌ای که انتخاب شده آورده شده است.

علت پایین بودن درصد بازشناسی با این طبقه بند را می توان این گونه توضیح داد که چون تمام زیر کلمات آموزش در یک خوشه قرار نمی گیرند، (افراز زیر کلمات آموزش در شکل ۴-۲۳ نشان داده شده است) برای هر زیر کلمه تعداد مقبولی برای آموزش شبکه وجود نداشته و شبکه آن طور که باید آموزش نمی بیند.



شکل ۴-۲۳: افراز زیر کلمات آموزش در بین خوشه ها

۴-۵-۲- K نزدیک ترین همسایه

در این طبقه بند نیز برای بالا بردن درصد بازشناسی چهار خوشه ی نزدیک به زیر کلمه ی آزمون در نظر گرفته شده است؛ قبل از استفاده از این طبقه بند، این چهار خوشه را با هم ادغام کرده و نتیجه را به عنوان فضای جدید در اختیار این طبقه بند قرار می دهیم؛ با توجه به نتایجی که تا کنون از آزمایش هایی که انجام شده بدست آمد؛ ویژگی میانگین بلوکی و معیار فاصله کرویشین و خوشه بندی به روش k میانگین با $k = 200$ را انتخاب کردیم و ۱۰ همسایه را برای جمع کردن آرا ($k = 10$) گرفتیم که با این تفاسیر به درصد بازشناسی $40/03\%$ رسیدیم.

۴-۵-۲-۱- نزدیک ترین همسایه

با در نظر گرفتن همان ویژگی، معیار فاصله و همان روش خوشه بندی ولی با در نظر گرفتن

نزدیک ترین همسایه به زیرکلمه‌ی آزمون به درصد بازشناسی با $82/53\%$ رسیدیم.

۴-۵-۲-۲- احتمال حضور زیرکلمه‌ی آزمون در m نزدیک ترین همسایه‌ی نزدیک به آن

با در نظر گرفتن تعداد بیش از یکی از نزدیک ترین نمونه‌ها (همسایه‌ها) به زیرکلمه‌ی آزمون

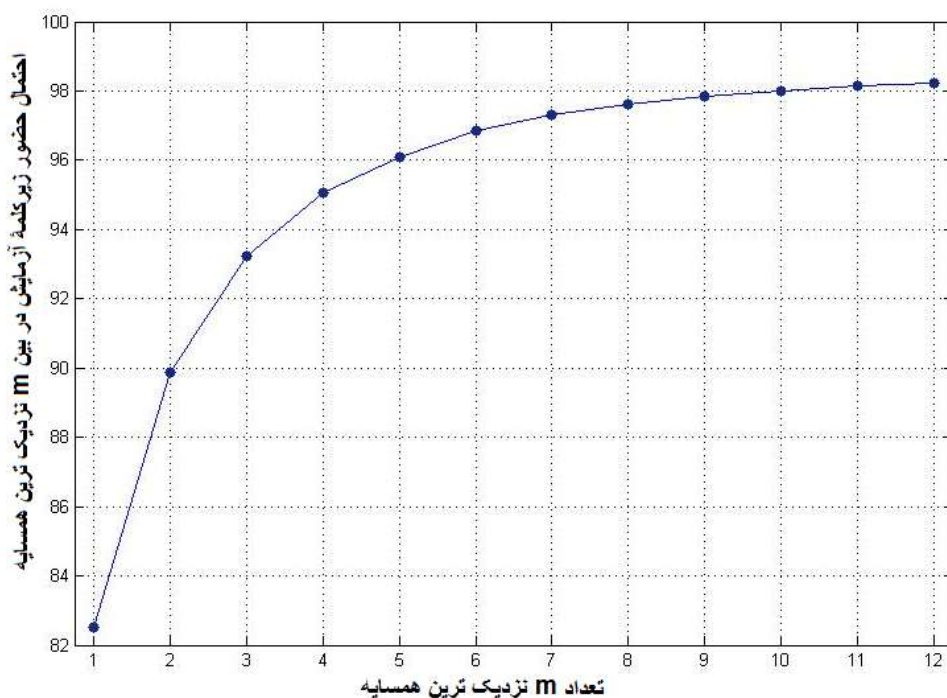
ورودی می‌توان زیرکلمات احتمالی برای جایگزینی زیرکلمه‌ی آزمون را افزایش داد، که از کاندیدهای

احتمالی در آینده استفاده می‌کنیم و با این کار احتمال تشخیص زیرکلمه‌ی مورد آزمون ارتقا می‌یابد

برای یافتن بهترین تعداد، از نزدیک ترین زیرکلمات چند حالت را مورد بررسی قرار دادیم که در هر

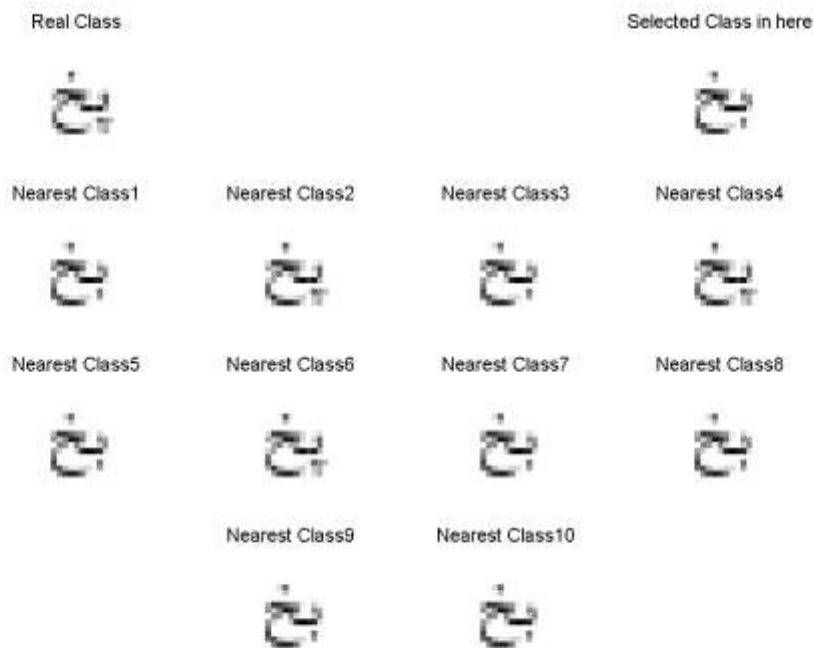
حالت با فرض اینکه m تعداد از نزدیک ترین همسایه‌ها را در نظر بگیریم، احتمال حضور زیرکلمه‌ی

آزمون در بین این m همسایه محاسبه و منجر به شکل ۴-۲۴ شد.

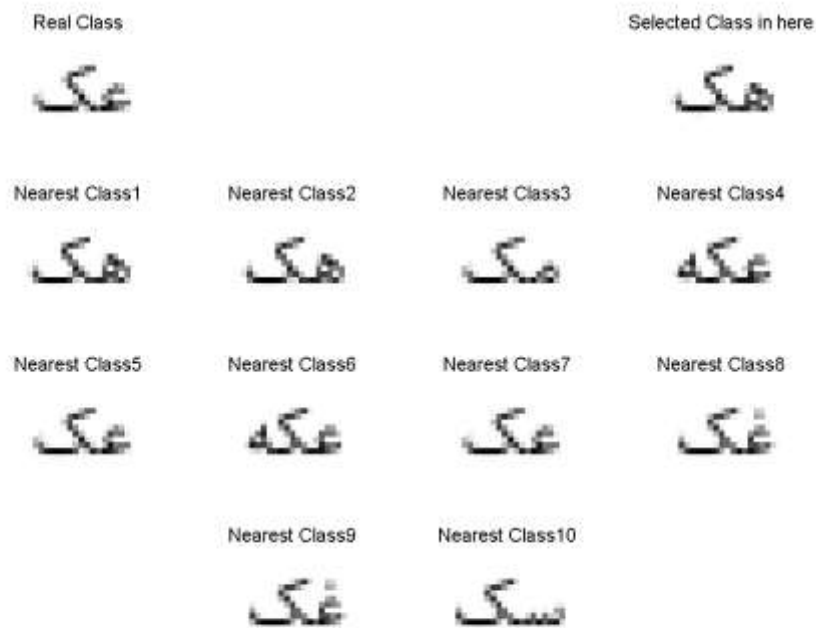


شکل ۴-۲۴: افزایش تعداد نمونه‌های جایگزین احتمالی به زیرکلمه‌ی آزمون برای بالا بردن احتمال تشخیص زیرکلمه

همان‌طور که مشاهده می‌شود با در نظر گرفتن ۱۰ همسایه‌ی نزدیک می‌توان به ۹۸ درصد رسید، بعداً نیز بیان می‌شود که این درصد تعیین‌کننده‌ی حد نهایی در بازشناسی است؛ از طرفی با افزایش همسایه‌ها از این بیشتر رشد چندانی وجود ندارد لذا $m = 10$ گرفته شد. البته این را هم باید مد نظر داشت که این درصد، صرفاً احتمال حضور زیرکلمه‌ی آزمایش در m نزدیک‌ترین همسایه هست و دست‌یابی به زیرکلمه‌ی مورد نظر در طی فرایند بازشناسی هدف ما در قسمت بعد است. چند نمونه از تصاویر نزدیک‌ترین زیرکلمات به نمونه‌ی ورودی آزمون و انتخاب نزدیک‌ترین زیرکلمه به آن را در شکل ۴-۲۵ مشاهده می‌کنید که به دلیل انتخاب نزدیک‌ترین همسایه در بازشناسی خطا رخ داده است، همان‌طور که پیداست این زیرکلمات بسیار به هم شبیه‌اند.



(الف)



(ب)

شکل ۴-۲۵: حضور نداشتن زیرکلمه‌ی نمونه‌ی آزمون در نزدیک‌ترین همسایه که با انتخاب نزدیک‌ترین همسایه در بازشناسی خطا رخ داده است.

۴-۶- بازشناسی با معیارها و سنجش‌گرهای انتخاب شده

با جمع بندی نتایجی که از قسمت‌های قبل بدست آمد، معیارها و سنجش‌گرها به این صورت انتخاب شدند؛ یکسان‌سازی ابعاد زیر-کلمه‌ها با هم اندازه سازی با حفظ نسبت طول به عرض، استخراج ویژگی میانگین بلوکی و معیار فاصله‌ی کرولیشن، هم برای خوشه بندی و هم برای طبقه بندی، خوشه بند K میانگین و طبقه بند نزدیک‌ترین همسایه، همان‌گونه که دیده شد با بررسی و ترکیب روش‌های مختلف بالاترین درصد بازشناسی که در این‌گونه از تصاویر بدست آوردیم ۸۲/۵۳٪ بود ولی با استفاده از روش نوین رخدادهای محتمل برای توالی زیرکلمات^۱ می‌توان جهش قابل

^۱ Probable Occurrences for Sub-words Sequence (POSS)

از درصد بازشناسی را مشاهده کرد، این روش که پیشنهاد ما در این پایان نامه می‌باشد، به این صورت است که احتمال حضور زیرکلمه‌ی قبل و بعد از هر زیرکلمه را تخمین می‌زند؛ برای بالا بردن سرعت و دقت بازشناسی استفاده از این روش این صورت پیشنهاد می‌شود که از بین تعداد محدودی از زیرکلمات این گزینش صورت گیرد، که تفصیل آن در قسمت‌های بعد آورده شده است.

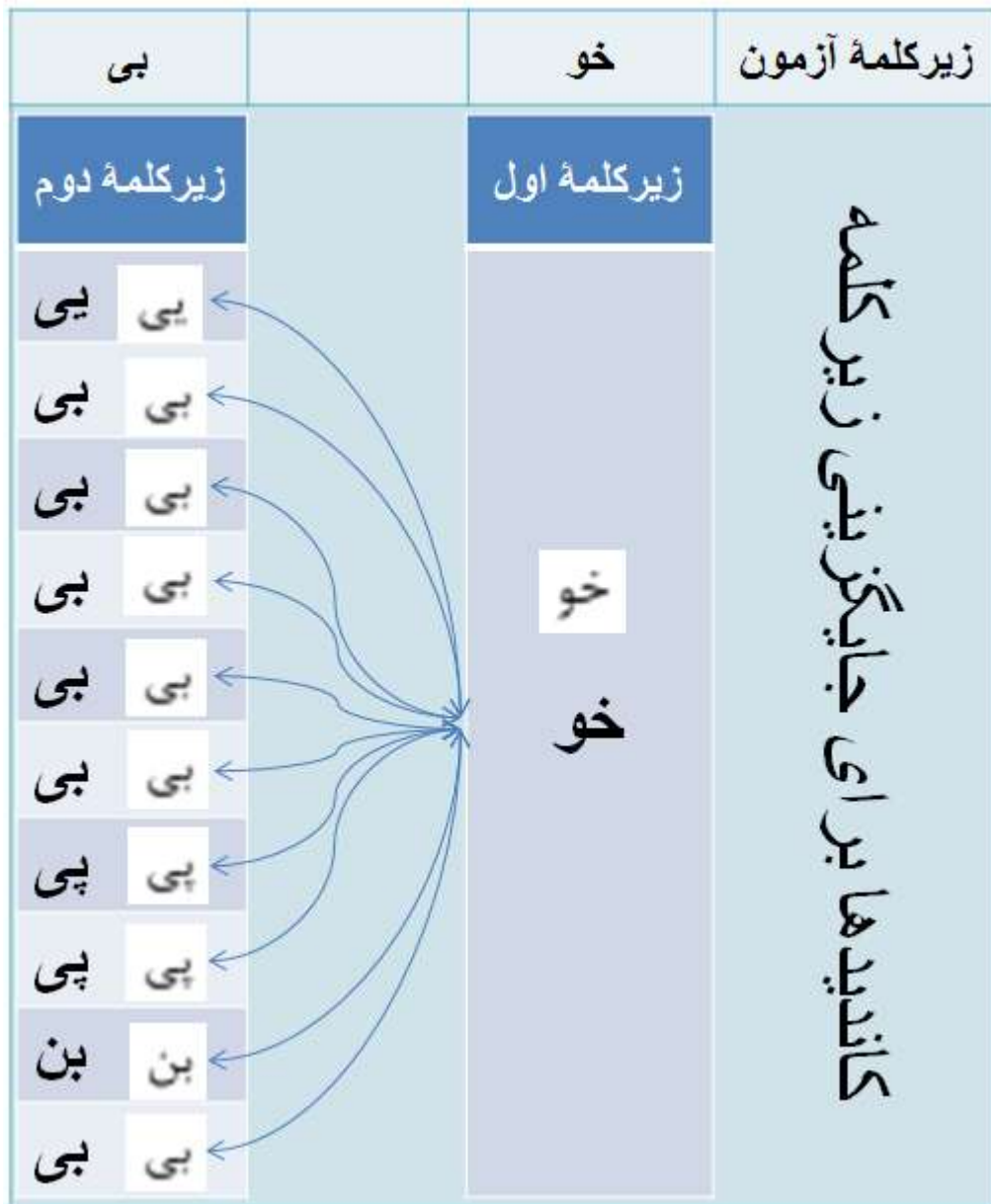
همان‌گونه که از قسمت طبقه بند k نزدیک‌ترین همسایه دیدیم با در نظر گرفتن ۱۰ همسایه‌ی نزدیک به زیرکلمه‌ی مورد آزمایش احتمال حضور این زیرکلمه در این بین افزایش می‌یابد پس با در نظر گرفتن این کاندیدها برای زیرکلمه در صورتی که بتوانیم زیرکلمه‌ی آزمون را از میان آنها با اطمینان بالا پیدا کنیم می‌توان به درصد بازشناسی ۹۸.۰۱٪ رسید از طرفی با توجه به شکل ۴-۲۵ زیرکلماتی که به عنوان کاندیدا برای جایگزینی نمونه‌ی ورودی آزمون انتخاب می‌شوند زیرکلماتی هم شکل و همانندی به زیرکلمه‌ی ورودی هستند که با انتخاب محتمل‌ترین حالت این دقت دور از انتظار نیست، همان‌طور که ذکر شد هر کلمه از زیرکلمات تشکیل شده است که ما از اهمیت قرار گرفتن این زیرکلمات برای تشکیل کلمات استفاده کرده و در تمام دیتابسی از کلمات که جمع آوری کرده بودیم برای تک تک زیرکلمات تمام این احتمال‌ها را محاسبه کردیم البته در اینجا فقط پیش بینی دو زیرکلمه‌ی پشت سر هم را در نظر گرفتیم که منجر به بردار $10746 \times 10746 = 115476516$ تایی شد، حتی می‌توان این بردار احتمالاتی برای بیش از دو زیرکلمه و یا فراتر از زیرکلمات و حتی برای کلمات نیز محاسبه نمود.

با این وصف در فرایند بازشناسی برای هر زیرکلمه دو حالت اتفاق می‌افتد یا زیرکلمه با اطمینان بالایی شناسایی می‌شود و یا در غیر این صورت ۱۰ زیرکلمه‌ی نزدیک به آن کاندیدی برای جایگزینی با آن در نظر گرفته می‌شوند قابل ذکر است که این اطمینان بالا، ناشی از آرایی است که از ۱۰ نزدیک‌ترین همسایه‌ی هر زیرکلمه بدست می‌آید، با این تفاوت که تاثیر آرا برای تأیید شدن رای را بجای ۶ عدد، ۷ رای در نظر گرفتیم زیرا زیرکلماتی که در یک خوشه هستند به هم شبیه‌اند و این شباهت در ۱۰ نزدیک‌ترین زیرکلمات به نمونه‌ی ورودی حتی شدیدتر هم هست که در مواردی با

وجود ۶ رای بر تایید یک زیرکلمه به خطا زیرکلمه‌ی دیگری که شباهت زیاد به آن را دارد انتخاب می‌شود برای جلوگیری از این دست از اشتباهات تعداد آرا برای تایید زیرکلمه، ۷ در نظر گرفتیم با این کار هرچند که درصد بازشناسی پایین بیاید، با POSS قابل جبران است؛ ولی در عوض مقدار اطمینان از بازشناسی بالاتر می‌رود.

در روش POSS دو جایگاه زیرکلمه در نظر گرفته می‌شود از آنجایی که ترکیب دو زیرکلمه با ترتیب مختلف با هم متفاوت است لذا ترتیب از اهمیت بالایی در این روش برخوردار است، در نتیجه برای تخمین هر دو زیرکلمه چهار حالت رخ می‌دهد در صورتی که زیرکلمه‌ی اول با اطمینان بالا شناسایی شود و زیرکلمه‌ی دوم دقت مطلوبی نداشته باشد، یا بالعکس و یا هر دو زیرکلمه با دقت بالایی شناسایی نشوند که در هر حالت تمام احتمالات برای ترکیب‌های بین دو زیرکلمه‌ی پیاپی محاسبه شده و دو ترکیبی پیروز است که احتمال رخ داد آن از همه بیشتر باشد در حالت چهارم هم که هر دو زیرکلمه با دقت مطلوب شناسایی شوند نیازی به ادامه کار نیست هر چند استفاده از روش POSS در هر صورت، حتی زمانی که زیرکلمه با دقت مطلوب شناسایی شود قابل استفاده است و حتی درصد یقین به بازشناسی صحیح را بالا می‌برد ولی هدف ما علاوه بر بازشناسی صحیح با دقت بالا، کاهش زمان بازشناسی نیز هست، لذا زمانی که دقت در حد مطلوب باشد به همان بسنده می‌کنیم.

در شکل ۴-۲۶ و شکل ۴-۲۷ به ترتیب دو مثال از کلماتی آورده شده که در اولی زیرکلمه‌ی اول و در شکل بعد زیرکلمه‌ی دوم با دقت مناسب شناسایی شده، در شکل اول، به دلیل شناسایی نشدن زیرکلمه در مرحله‌ی نخست، ۱۰ جایگزین آن به همراه زیرکلمه‌ی قبلی به الگوریتم POSS برای شناسایی برده می‌شوند و در شکل بعد عکس این عمل انجام می‌شود؛ تعداد رخ داده‌ها برای ترکیب این زیرکلمات برای دو به ترتیب در جدول ۴-۴ و جدول ۴-۵ آمده است.



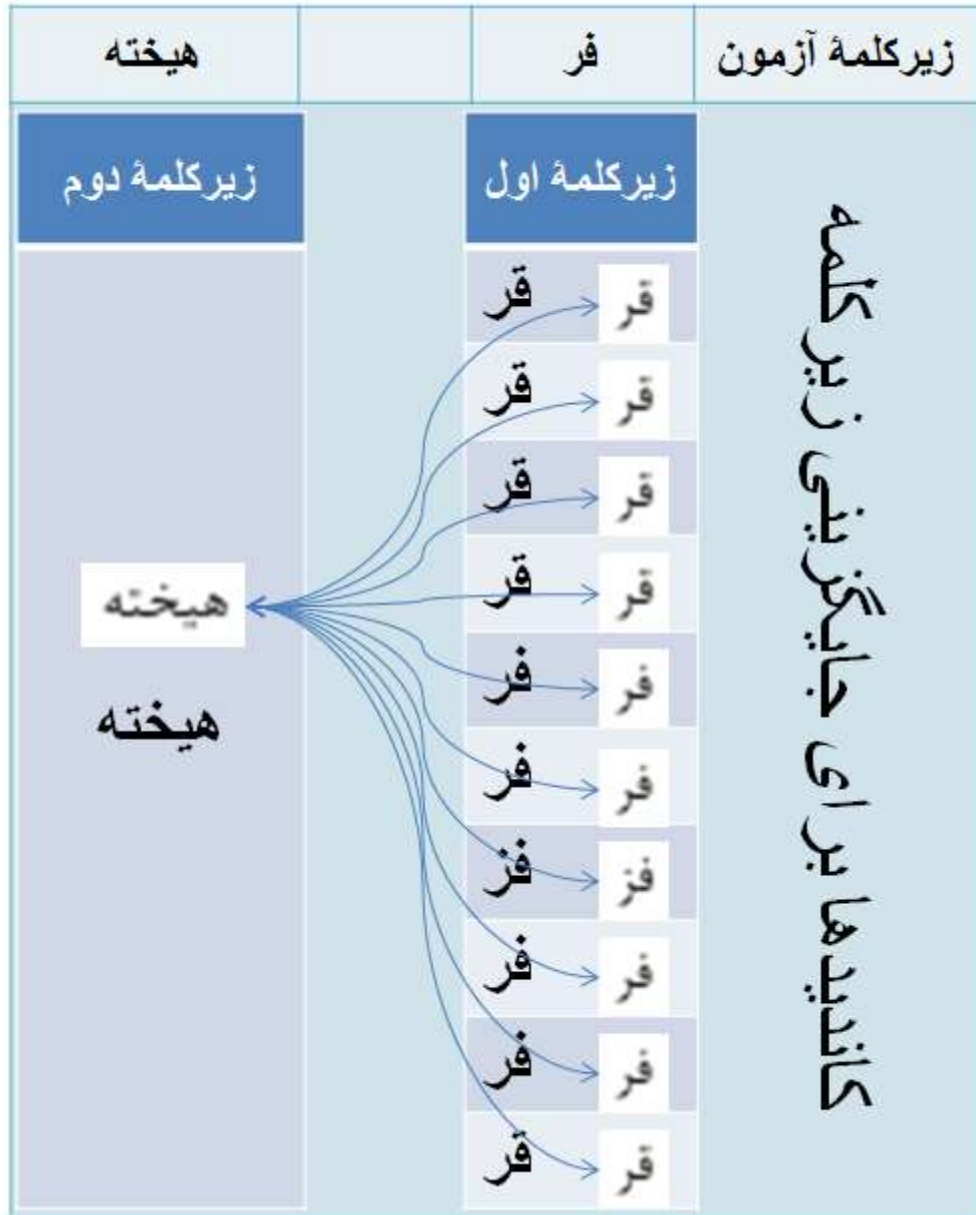
شکل ۴-۲۶: مثالی از الگوریتم POSS که زیرکلمه‌ی اول شناسایی شده و ترکیب‌های مختلف دو زیرکلمه‌ی پشت سر هم برای پیدا کردن معمول‌ترین حالت بررسی می‌شود.

جدول ۴-۴: تعداد رخداد‌های ترکیبی دو زیر کلمه در شکل ۴-۲۶ که بالاترین مقادیر مشخص شده‌اند.

بی	بی	بی	بی	بی	بی	پی	پی	بن	بی	
۱۳۷	۱۲۷۳	۱۲۷۳	۱۲۷۳	۱۲۷۳	۱۲۷۳	۰	۰	۵	۱۲۷۳	خو

همان‌طور که از جدول ۴-۴ پیداست زیرکلمه‌ی "بی" از لیست ۱۰ زیرکلمه‌ی نزدیک به زیرکلمه‌ی

آزمون انتخاب می‌شود.



شکل ۴-۲۷: مثال دیگری از الگوریتم POSS که زیرکلمه‌ی دوم باز شناسایی شده و ترکیب‌های مختلف دو زیرکلمه‌ی پشت سر هم برای پیدا کردن معمول‌ترین حالت بررسی می‌شود.

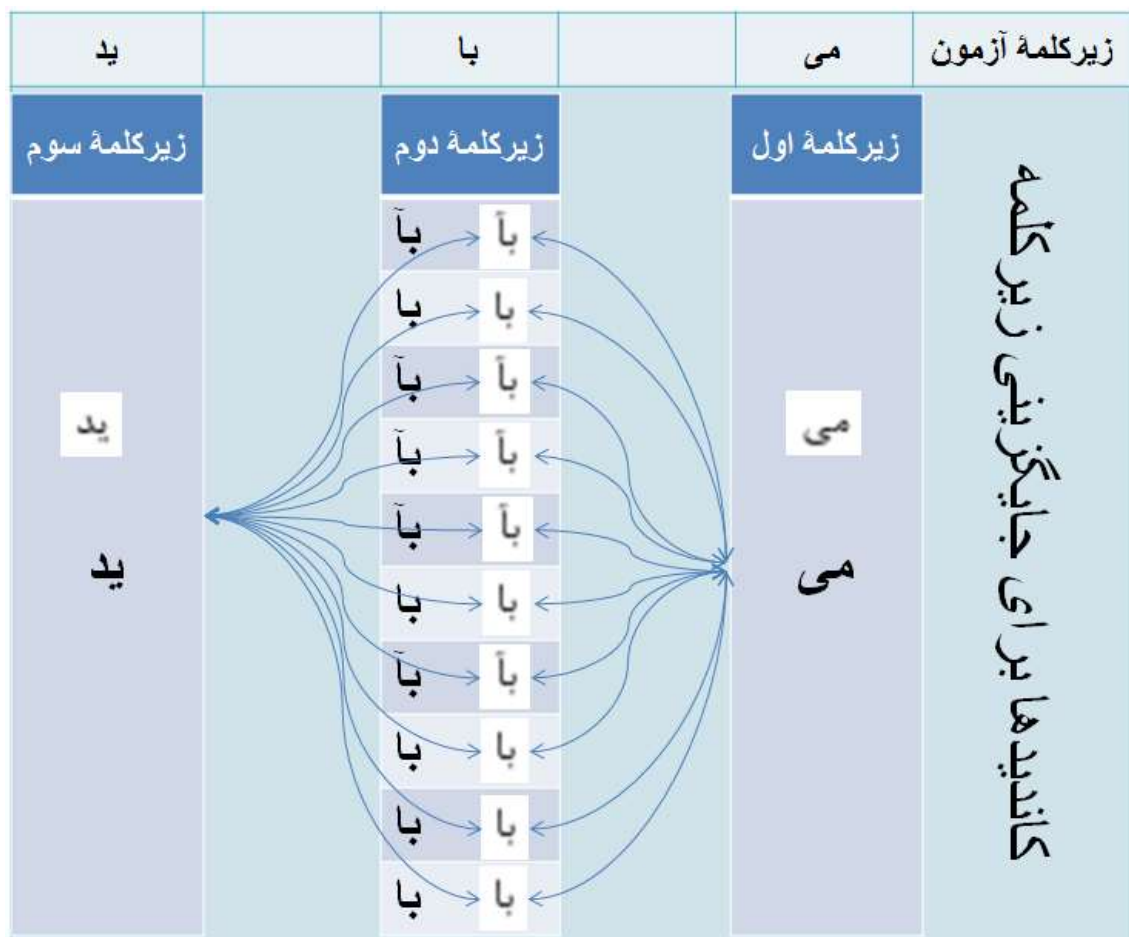
جدول ۴-۵: تعداد رخداد‌های ترکیبی دو زیرکلمه در شکل ۴-۲۷.

هیخته	
۰	قر
۰	قر
۰	قر
۰	قر
۱	فر
۱	فر
۰	فز
۱	فر
۱	فر
۰	قر

همان‌گونه که پیداست روش POSS برای کلماتی کاربرد دارد که از بیش از یک زیرکلمه تشکیل شده باشند، در صورتی که کلمه‌ی مورد نظر بیش از دو زیرکلمه داشته باشد برای دو به دو زیرکلمات این فرایند از سمت راست ادامه می‌یابد با این تفاوت که تمام ترکیب‌های بین دو احتمال‌های مخالف صفر در دو زیرکلمه‌ی پیاپی با یکدیگر جمع شده و این کار تا پایان کلمه‌ی مد نظر ادامه پیدا می‌کند و از بین آن زیرکلمات پیاپی با احتمال مخالف صفر، آنهایی انتخاب می‌شوند که جمع بین احتمال‌های ترکیب‌های مختلف از همه بیشتر باشد.

سه مثال از کلمات شامل ۳ زیرکلمه که ترکیب‌های هر دو زیرکلمه‌ی پشت سر هم در تصاویر نمایش داده شده و تعداد رخداد هر دو ترکیب از زیرکلمات در جداول آورده شده است.

با توجه به جدول ۶-۴ و جدول ۷-۴ در زیرکلمات شکل ۴-۲۸، کلمه‌ی "می‌باید" انتخاب می‌شود، با توجه به جدول ۸-۴ و جدول ۹-۴ در زیرکلمات شکل ۴-۲۹، کلمه‌ی "می‌بایست" انتخاب می‌شود و در نهایت در کلمه‌ای با ترکیب زیرکلماتی که در شکل ۴-۳۰ آمده با استفاده از جدول ۴-۱۰ و جدول ۴-۱۱، کلمه‌ی "نمی‌باشد" انتخاب می‌شود.



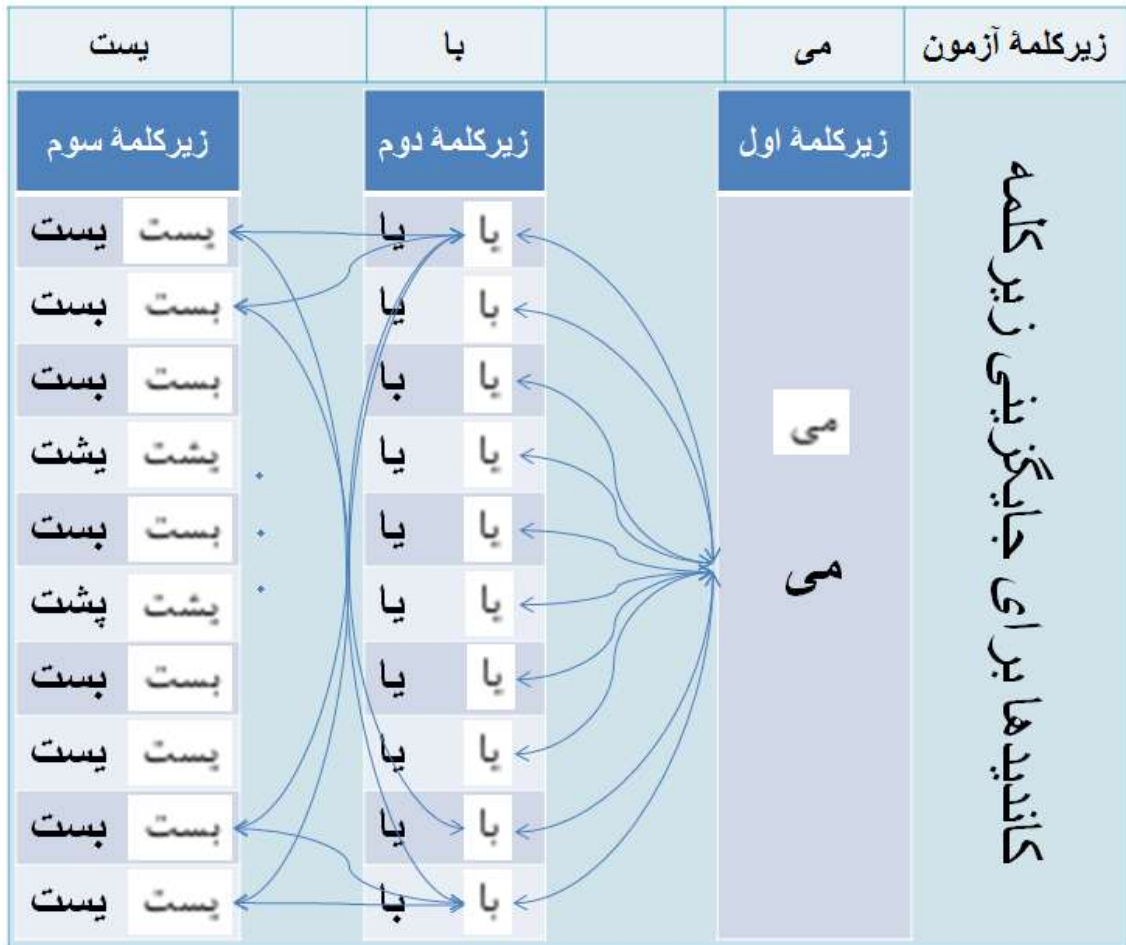
شکل ۴-۲۸: کلمه‌ای شامل ۳ زیرکلمه که ترکیب‌های دو به دوی زیرکلمات پشت سر هم در تصویر آمده است.

جدول ۴-۶: تعداد رخداد ترکیب زیرکلمات اول و دوم در شکل ۴-۲۸، (بالاترین مقادیر مشخص شده‌اند).

با	با	با	با	با	با	با	با	با	با	
۰	۷۸۲	۰	۰	۰	۷۸۲	۰	۷۸۲	۷۸۲	۷۸۲	می

جدول ۴-۷: تعداد رخداد ترکیب زیرکلمات دوم و سوم در شکل ۴-۲۸، (بالاترین مقادیر مشخص شده‌اند).

ید	
۰	با
۹۴۸۲	با
۰	با
۰	با
۰	با
۹۴۸۲	با
۰	با
۹۴۸۲	با
۹۴۸۲	با
۹۴۸۲	با



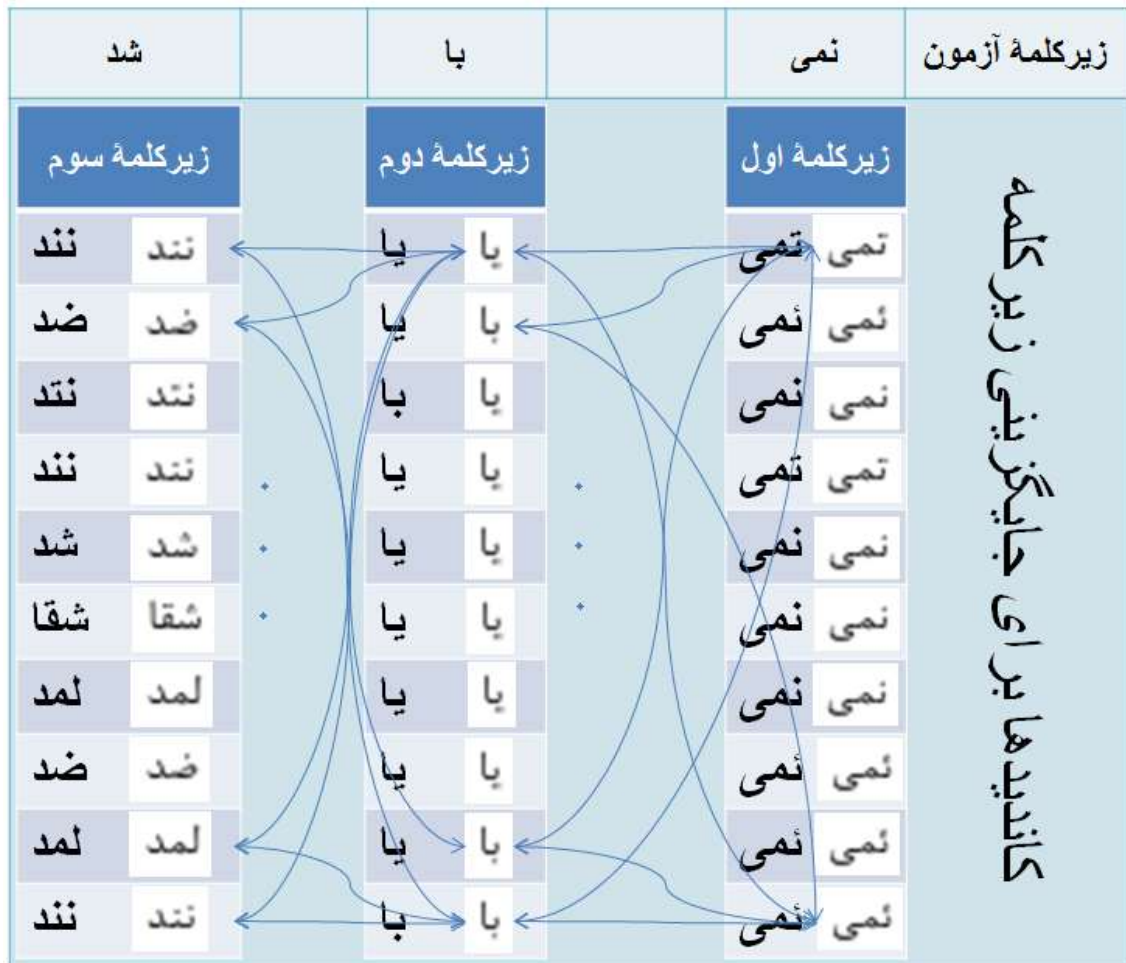
شکل ۴-۲۹: ترکیب دو به دوی زیرکلمات در کلمه‌ای با سه زیرکلمه.

جدول ۴-۸: تعداد رخداد ترکیب زیرکلمات اول و دوم در شکل ۴-۲۹، (بالاترین مقادیر مشخص شده‌اند).

یا	یا	با	یا	یا	یا	یا	یا	یا	با	
۸۷	۸۷	۷۸۲	۸۷	۸۷	۸۷	۸۷	۸۷	۸۷	۷۸۲	می

جدول ۹-۴ : تعداد رخداد ترکیب زیرکلمات دوم و سوم در شکل ۴-۲۹، (بالاترین مقادیر مشخص شده‌اند).

یست	بست	بست	یشت	بست	پشت	بست	یست	بست	یست	
۱۷	۱۶	۱۶	۰	۱۶	۰	۱۶	۱۷	۱۶	۱۷	یا
۱۷	۱۶	۱۶	۰	۱۶	۰	۱۶	۱۷	۱۶	۱۷	یا
۹۳۵	۳۵	۳۵	۰	۳۵	۱	۳۵	۹۳۵	۳۵	۹۳۵	با
۱۷	۱۶	۱۶	۰	۱۶	۰	۱۶	۱۷	۱۶	۱۷	یا
۱۷	۱۶	۱۶	۰	۱۶	۰	۱۶	۱۷	۱۶	۱۷	یا
۱۷	۱۶	۱۶	۰	۱۶	۰	۱۶	۱۷	۱۶	۱۷	یا
۱۷	۱۶	۱۶	۰	۱۶	۰	۱۶	۱۷	۱۶	۱۷	یا
۱۷	۱۶	۱۶	۰	۱۶	۰	۱۶	۱۷	۱۶	۱۷	یا
۱۷	۱۶	۱۶	۰	۱۶	۰	۱۶	۱۷	۱۶	۱۷	یا
۱۷	۱۶	۱۶	۰	۱۶	۰	۱۶	۱۷	۱۶	۱۷	یا
۹۳۵	۳۵	۳۵	۰	۳۵	۱	۳۵	۹۳۵	۳۵	۹۳۵	با



شکل ۴-۳۰: ترکیب دو به دوی زیر کلمات در کلمه‌ای با سه زیر کلمه.

جدول ۴-۱۰: تعداد رخداد ترکیب زیر کلمات اول و دوم در شکل ۴-۳۰ (بالاترین مقادیر مشخص شده‌اند).

یا	یا	با	یا	یا	یا	یا	یا	یا	با	
۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	تمی
۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	نمی
۸	۸	۳۱	۸	۸	۸	۸	۸	۸	۳۱	نمی
۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	تمی
۸	۸	۳۱	۸	۸	۸	۸	۸	۸	۳۱	نمی

۸	۸	۳۱	۸	۸	۸	۸	۸	۸	۳۱	نمی
۸	۸	۳۱	۸	۸	۸	۸	۸	۸	۳۱	نمی
۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	نمی
۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	نمی
۰	۰	۰	۰	۰	۰	۰	۰	۰	۰	نمی

جدول ۴-۱۱ : تعداد رخداد ترکیب زیر کلمات دوم و سوم در شکل ۴-۳۰، (بالاترین مقادیر مشخص شده‌اند).

نند	ضد	نند	نند	شد	شقا	لمد	ضد	لمد	نند	
۵۵	۰	۰	۵۵	۲	۰	۰	۰	۰	۵۵	یا
۵۵	۰	۰	۵۵	۲	۰	۰	۰	۰	۵۵	یا
۵۳	۰	۰	۵۳	۱۶۷۲۵	۰	۱	۰	۱	۵۳	با
۵۵	۰	۰	۵۵	۲	۰	۰	۰	۰	۵۵	یا
۵۵	۰	۰	۵۵	۲	۰	۰	۰	۰	۵۵	یا
۵۵	۰	۰	۵۵	۲	۰	۰	۰	۰	۵۵	یا
۵۵	۰	۰	۵۵	۲	۰	۰	۰	۰	۵۵	یا
۵۵	۰	۰	۵۵	۲	۰	۰	۰	۰	۵۵	یا
۵۵	۰	۰	۵۵	۲	۰	۰	۰	۰	۵۵	یا
۵۳	۰	۰	۵۳	۱۶۷۲۵	۰	۱	۰	۱	۵۳	با

فصل ۵:

نتیجه گیری و پیشنهادات

۵-۱- نتیجه گیری

در این پایان نامه از شکل کلی زیر کلمات برای بازشناسی کلمات استفاده شد. برای این کار مجموعه زیر کلمات معتبر فارسی تعیین شدند سپس از صفحات شامل این زیر کلمات از دو طریق روبشگر و روش مصنوعی تصاویری با وضوح ۹۶ dpi تولید کردیم، در روش تولید دیتابیس با روبشگر، ابتدا این صفحات شامل زیر کلمات با درجه تفکیک ۳۰۰ dpi چاپ و با ۹۶ نقطه در اینچ روبش شدند و در روش مصنوعی این ترتیب با نرم افزارهای شبیه ساز صورت گرفت پس از حذف نویز به جداسازی و شماره گذاری زیر کلمات پرداختیم.

پس از بررسی توصیف‌گرهای مختلف، مناسب‌ترین ویژگی برای این‌گونه از تصاویر میانگین بلوکی شناخته شد؛ همچنین معیارهای فاصله‌ی مختلف کرولیشن بهترین گزینه بود.

برای کاهش دامنه جستجو در بازشناسی یک زیر-کلمه ناشناخته ورودی، از خوشه بندی k میانگین استفاده شد، برای بدست آوردن تعداد مناسب خوشه‌ها، از یک روش اعتبارسنجی خوشه بندی استفاده شد. با استفاده از این معیار ملاحظه شد که مناسب‌ترین تعداد خوشه‌ها ۲۰۰ در نظر گرفته شد، نتایج خوشه بندی نشان داد که زیر-کلماتی که شکل کلی مشابهی داشتند، در خوشه‌های یکسان قرار گرفته‌اند و از مراکز هر یک از خوشه‌ها به عنوان نماینده‌ی هر خوشه استفاده شد.

در بازشناسی زیر کلمات از یک روش دو مرحله ای استفاده شده است. در مرحله اول فاصله هر زیر کلمه از مراکز خوشه‌های زیر کلمات محاسبه می‌شود و زیر کلمات ۴ خوشه نزدیک‌تر تعیین و در یک خوشه‌ی جدید جمع می‌شوند. در مرحله دوم با استفاده از توصیفگر اشاره شده، زیر کلمات این فضای جدید بر اساس شباهت به زیرکلمه ورودی مرتب می‌شوند. با بررسی تعداد زیرکلمات نزدیک به نمونه‌ی ورودی ملاحظه شد که در ۱۰ نزدیک‌ترین همسایه به زیرکلمه‌ی آزمایش با اطمینان ۹۸٪ حتماً کلمه‌ی ناشناخته‌ی ورودی شامل آنها هست.

در یک آزمایش برای کارایی سیستم از هر زیرکلمه ۳ نمونه تولید و بازشناسی شد در این آزمایش، در حالتی که نزدیک‌ترین زیرکلمه دیتابیس به نمونه‌ی ورودی را گرفتیم ۸۲/۵۳٪ بازشناسی موفق داشتیم؛ در حالتی که با در نظر گرفتن ۱۰ نزدیک‌ترین همسایه و با تایید ۶ همسایه به یک زیرکلمه ۴۰/۰۳ درصد بازشناسی داشتیم.

در انتها الگوریتم POSS را معرفی کردیم و جهت سرعت بخشیدن به این الگوریتم کاهش ترکیب‌های زیرکلمات را پیشنهاد دادیم برای این منظور ۱۰ نزدیک‌ترین همسایه‌ی به زیرکلمه‌ی آزمون ورودی را به عنوان کاندیدی برای جایگزینی آن انتخاب کردیم، و روش ترکیبی KNN و POSS را برای تسریع در بازشناسی برگزیدیم؛ در صورتی که تصویر ورودی با ۷ زیرکلمه‌ی نزدیک به خودش مورد تایید قرار بگیرد همان زیرکلمه انتخاب و در غیر این صورت ۱۰ نزدیک‌ترین همسایه به آن نامزدی برای جایگزینی با نمونه‌ی ورودی می‌شوند، برای کلمات بیش از دو زیرکلمه نیز این فرایند برای دو به دو زیرکلمات این فرایند از سمت راست ادامه می‌یابد با این تفاوت که تمام ترکیب‌های بین دو احتمال‌های مخالف صفر در دو زیرکلمه‌ی پیاپی با یکدیگر جمع شده و این کار تا پایان کلمه‌ی مد نظر ادامه پیدا می‌کند و از بین آن زیرکلمات پیاپی با احتمال مخالف صفر، آنهایی انتخاب می‌شوند که جمع بین احتمال‌های ترکیب‌های مختلف از همه بیشتر باشد.

همان‌طور که از قسمت طبقه بند k نزدیک‌ترین همسایه دیدیم با در نظر گرفتن ۱۰ همسایه‌ی نزدیک به زیرکلمه‌ی ورودی احتمال حضور این زیرکلمه در این بین افزایش می‌یابد و به دلیل شباهت زیاد این ۱۰ کاندید به زیرکلمه‌ی ورودی، در صورتی که بتوانیم زیرکلمه‌ی آزمون را از میان آنها پیدا کنیم می‌توان به درصد بازشناسی ۹۸.۰۱٪ رسید.

۵-۲- پیشنهادات

با توجه به اهمیت موضوعات مطرح شده در قالب پایان‌نامه حاضر و به منظور پیشرفت سریع‌تر در

این حوزه‌ها، پیشنهادهای برای انجام پژوهش‌های جدید ارائه می‌شود:

- (۱) تهیه نرم افزار خود سازمانده^۱ برای بروز رسانی زیرکلمات جدیدی که وارد زبان فارسی می‌شوند و محاسبه‌ی برخط احتمال ترتیب زیرکلمات برای استفاده در الگوریتم POSS.
- (۲) ارتقا الگوریتم POSS با بدست آوردن احتمال رخداد بیش از دو زیرکلمه‌ی مجاور به هم همچنین دخالت دادن کلمات مجاور به زیرکلمات برای انتخاب.
- (۳) هرچند بازشناسی در dpi های بالا با دقت بالا گزارش شده است (۰/۹۷) به نظر می‌رسد با پیاده سازی این الگوریتم بازشناسی برای تصاویر با وضوح بالا حتی بتوان این درصد را از این هم فراتر برد.
- (۴) با فراگیر شدن تلفن‌های همراه و تبلت‌ها از آنجایی که وضوح تصاویر در دوربین این گونه دستگاه‌ها هر چند باکیفیت باشد نسبت به تصاویری که توسط روبشگر دریافت می‌کنیم، پایین‌تر است لذا در صورت گسترش دادن این روش بازشناسی، برای OCR کردن این گونه تصاویر، می‌توان قدم بزرگی در راستای بازشناسی نویسه برداشت.
- (۵) یکپارچه سازی و تکامل این الگوریتم بازشناسی مانند اضافه کردن قابلیت‌های رفع چرخش صفحه، رفع منحنی شدگی صفحه در اثر قرار دادن روی روبشگر و ... و در نهایت تجاری سازی محصول.

¹ Self Organization

- [1] Amin, A. and J. F. Mari, "Machine Recognition and Correction of Printed Arabic Text," IEEE Transactions on Systems, Man and Cybernetics 16(5), pp. 1300-1306, 1989 .
- [2] Altuwaijri, M. M. and M. A. Bayoumi, "Arabic text recognition using neural networks ," IEEE International Symposium on Circuits and Systems, pp. 415-418, 1994 .
- [3] Ho T. K., Hull J. J., and Srihari S. N, "A Word Shape Analysis Approach to Recognition of Degraded Word Images," Proc. of the 4th USPS Advanced Technology Conference, pp. 217-231, 1990 .
- [4] Hull, J.J., Srihari, S.N, "A Computational Approach to Visual Word Recognition: Hypothesis Generation and Testing," Computer Vision and Pattern Recognition, IEEE, pp. 156-161, 1986 .
- [5] ح. خسروی و ا. کبیر، " رویکرد یکپارچه برای بازشناسی متون چاپی فارسی،" رساله دکترای مهندسی برق - الکترونیک، دانشگاه تربیت مدرس، زمستان ۱۳۸۷.
- [6] Erlandson, E. J., Trenkle, J.M., Vogt, R.C, "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 .
- [7] Menhaj, M. B. and M. Adab, "Simultaneous segmentation and recognition of Farsi/Latin printed texts with MLP," International Joint Conference on Neural Networks, pp. 1534-1539, 2002 .
- [8] Broumandnia, A., J. Shanbehzadeh, et al, "Segmentation of Printed Farsi/Arabic Words ," IEEE/ACS International Conference on Computer Systems and Applications, pp. 761-766, 2007 .

- [9] Ho T. K., Hull J. J. and Srihari S. N, "A word shape analysis approach to lexicon based word recognition," Pattern Recognition Letters, Vol. 13, pp. 821-826, 1992 .
- [10] Ho T. K. and Hull J. J. and Srihari S. N "A Computational Model for Recognition of Multifont Word Images ," Machine Vision and Applications, Vol. 5, 3, pp. 157-168, 1992 .
- [11] Khorsheed M.S. and Clocksin W.F, "Multi-Font Arabic word recognition using spectral features ," Proc. of ICPR2000, Vol. 4, p. 4543, 2000 .
- [۱۲] ر. عزمی، احسان ا. کبیر، ک. بدیع، "بازشناسی متون چاپی فارسی"، رساله دکتری، دانشکده فنی و مهندسی، دانشگاه تربیت مدرس، ۱۳۷۸.
- [۱۳] ا. ابراهیمی و ا. کبیر، "استفاده از شکل کلی زیر - کلمات چاپی در بازیابی تصویر مستندات و بازشناسی متون فارسی"، رساله دکترای مهندسی برق - الکترونیک ، دانشگاه تربیت مدرس ، تابستان ۱۳۸۴.
- [14] Amin, A. and Mansoon, W, "Recognition of Arabic printed text using neural networks," Proc. 4th Int. Conf. on Document Analysis Recognition, Ulm, Germany, August 1997 .
- [۱۵] م. شیرعلی 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.
- [16] Flora S. Tsai "A visualization metric for dimensionality reduction," Expert Systems with Applications ,Vol. 39,No. 2, pp. 1747-1752, February 2012 .
- [17] Richard O.Duda, Peter E.Hart and David G.Stork, Pattern Classification, 2rd Edition, John Wiley&Scons, Inc, United State, America, 2001 .
- [۱۸] ح. خسروی و ا. کبیر، "معرفی دو ویژگی سریع و کارآمد برای بازشناسی ارقام دستنویس فارسی"، کنفرانس ماشین بینایی و پردازش تصویر، مشهد، صفحه ۱۱۲۶-۱۱۳۱، ۱۳۸۵.

[19] C. Shan, S. Gong, P. W. McOwan "Facial expression recognition based on Local Binary Patterns: A comprehensive study," sciencedirect Image and Vision Computing, Vol. 27, No. 6, pp. 803-816, 2009 .

[۲۰] ا. بایسته تاشک، ع. احمدی فرد، ح. خسروی، "بازشناسی برون خط کلمات دست نوشته فارسی در یک مجموعه محدود لغات"، پایان‌نامه‌ی کارشناسی ارشد، دانشگاه صنعتی شاهرود، ۱۳۹۰.

[21] Otsu, N "A Threshold Selection Method from Gray Level Histograms," IEEE Trans. on Systems, Man and Cybernetics, Vol. 9, pp. 62-66, Jan. 1979

[22] Chien H. Chou, Wen H. Lin, and F. Chang "A binarization method with learning-built rules for document images produced by cameras, " Pattern Recognition, Vol. 43, No. 4, pp. 1518-1530, November 2009 .

[23] Huy P. Le, and Guee S. Lee "Noise removal from binarized text images," Computer and Automation Engineering The 2nd International Conference ,Vol. 3,pp. 586-589, February 2010, Singapore .

[24] Rafael C. Gonzalez, and Richard E. Woods, Digital Image Processing, 2rd Edition, Prentice Hall ,Upper Saddle River, New Jersey, 2002 .

[25] G. Bradski and A. Kaehler "Robot-Vision signal processing primitives," IEEE Signal Processing Magazine, Vol. 25, No. 1, pp. 130-133, January 2008 .

[26] F. Murtagh "A survey of recent advances in hierarchical clustering algorithms," Computer Journal. Vol. 26, No. 4, pp. 354-359, November 1983 .

[27] F. Husson, J. Joose and J. Pages "Principal component methods – hierarchical clustering - partitional clustering clustering: why would we need to choose for visualizing data," Applied Mathematic department Agrocampus Ouest, September 2010 .

[28] Anil K. Jain "Data clustering: 50 years beyond k-means," Pattern Recognition Letters, Vol. 31, No. 8, pp. 651-666, June 2010 .

[29] N. Memarsadeghi, David M. Mount, Nathan S. Netanyahu, and Jacqueline L. Moigne "A Fast Implementation of the ISOCLUS Algorithm," IEEE International

Geoscience and Remote Sensing Symposium, Vol. 3, pp. 2057-2059, July 2003, Toulouse, France .

[30] Simon Haykin, Neural Networks A Comprehensive Foundation, 2nd Edition, Prentice Hall Inc, Tom Robbins, United States Of America, 2001 .

[۳۱] ح. خسروی و ا. کبیر "ارزیابی روشهای بازشناسی متون فارسی بر مبنای شکل کلی زیر- کلمات"، نشریه مهندسی برق و کامپیوتر ایران، جلد ۷، شماره ۴، ص ۲۶۷-۲۸۰، زمستان ۱۳۸۸.

[32] V. Kůrková "Kolmogorov's theorem and multilayer neural networks," sciencedirect Neural Networks, Vol. 5, No. 3, pp. 501-506, September 1992 .

واژه نامه فارسی - انگلیسی

Backpropagation	انتشار به عقب
Hyperglobe	ابر کره
Hypercube	ابر مکعب (چیزی شبیه مکعب در فضاهاى بیش از سه بعد)
Hyperpyramid	ابر هرم (چیزی شبیه هرم در فضاهاى بیش از سه بعد)
Contrast	اختلاف شدت فروزندگی بین اجزاء مختلف یک عکس
Extraction	استخراج
Confidence	اطمینان
Projection	افکنش
Pattern	الگو، طرح
Smearing	آغشته سازی
Continous	پیوسته
Preprocessing	پیش پردازش
Investigate	تجسس کردن
Combinational	ترکیبی
Resolution	تفکیک پذیری
Segmentation	جداسازی
Character	حرف
Secretariat	دبیر خانه
Classifier	دسته کننده
Binary	دوسطحی
Sequence	رشته
Scanner	رویشگر
Illumination	روشنایی
Sub word	زیر کلمه
Global	سراسری
Gray level	سطح خاکستری
Hierachical	سلسله مراتبی
Holistic shape	شکل کلی
Recognition	شناسایی

Classify	طبقه بندی کردن
Resilient	عکس العمل
City block distance	فاصله ی بلوک شهر
Erosion	فرسایش
Portable	قابل حمل
Template	قالب
Run lenth	قطع گذر
Holistic	کل نگر
Dilation	گسترش
Discrete	گسسته
Local	محلی
Stage	مرحله
Contour	مرز بیرونی
Minkowski distance measure	معیار فاصله ی مینکوسکی
k-mean	k میانگین
Optical	نوری
Histogram	هیستوگرام
Feature	ویژگی

Abstract

There are three approaches for word recognition: segmentation, holistic shape, and combination of them. At the first approach a word is broken into its letters, then each letter is recognized and the results combined together. In another scenario, the word image is given to the system for holistic recognition; i.e., recognizing the word as a whole shape. In many personal and official systems, document images are stored in low-resolution. This resolution is suitable for human reading, but it is hardly recognizable by machine. Even, most of Latin optical character recognition systems have been developed for 300 dots per inch. In low resolution documents, it is difficult to segment the word into its letters, so for this situation, holistic approach is adopted. In this thesis, recognition of the sub-words image in 96 dpi, based on holistic shape is presented.

In this presentation, the proposed system is based on a three step method. In the first stage, the number of investigated sub-words in the dictionary is reduced by using clustering method; this strategy enhances the recognition speed as well as recognition performance. In the second stage, 4 nearest clusters of the test sub-word is found with a classifier and then by searching among that clusters, the 10 nearest sub-words to the test sample will be found. This process is repeated for all the sub-words of a word then at the last stage, the word will be recognized with probable occurrences sub-words sequence method.

The accuracy which is estimated in this recognition algorithm is so convenient, it has the ability to recognize the words with more than one sub-word with recognition rate of 98.01% and recognize words in single sub-word with 82.53% rate.

Keywords— Sub-word recognition, holistic shapes of Sub-words, images with low resolution, lexicon reduction, zoning feature, k mean, k nearest neighbor, POSS algorithm.



Shahrood University of Technology

Faculty of Electrical & Robotic Engineering

Electrical and Robotic Department

Thesis Submitted for the Degree of Master of Science

Recognizing Farsi words in low resolution images

Hamed ArabYarmohammadi

Supervisor:

Dr. Alireza AhmadiFard

Advisor:

Dr. Hossein Khosravi

Winter 2014