شناسایی کلمات فارسی نستعلیق با استفاده از شبکه‌های عصبی مصنوعی

نفیسه صالحیان[۱]، محمدرضا یزدچی[۲]، علیرضا کریمیان[۳]
۱- دانشگاه آزاد اسلامی واحد نجف آباد، دانشکده تحصیلات تکمیلی، mmail25@yahoo.com
۲- عضو هیات علمی گروه مهندسی پزشکی دانشگاه اصفهان، yazdchi@eng.ui.ac.ir
۳- عضو هیات علمی گروه مهندسی پزشکی دانشگاه اصفهان، karimian@eng.ui.ac.ir

چکیده

در این مقاله یک سیستم کامل برای شناسایی کلمات فارسی نستعلیق با استفاده از شبکه‌های عصبی مصنوعی پیاده‌سازی شده است. در مرحله پیش‌پردازش پس از یافتن بخش‌های متصل، سرکش‌ها و زیرکش‌های حروف کشف و از تصویر حذف می‌گردند. با استفاده از یک الگوریتم تقطیع که بر اساس کانتور بالایی و پایینی کلمه عمل می‌کند، تصویر کلمه به دنباله‌ای از زیر کلمه‌ها شکسته می‌شود. پس از انجام عمل تقطیع، هشت ویژگی شامل سه توصیفگر فوریه و تعدادی ویژگی ساختاری گسسته برای نمایش زیرکلمه‌ها در فضای ویژگی بکار می‌روند. شناسایی با استفاده از یک شبکه عصبی پرسپترون چند لایه و بر اساس این بردارهای ویژگی انجام می‌شود. خطاهای احتمالی در تشخیص زیرکلمه‌ها با استفاده از یک الگوریتم جستجو در لغت‌نامه سیستم اصلاح می‌شود. آزمایش سیستم بر روی یک نمونه شامل ۳۲۰ کلمه عملکرد مناسبی (۹۷٪ صحت) را نشان می‌دهد.

واژه‌های کلیدی

تقطیع، شبکه‌های عصبی مصنوعی، شناسایی کلمات

۱– مقدمه

شناسایی کلمات دست‌نویس کاربرد زیادی در کارهای بانکی (پردازش چک‌های بانکی)، شناسایی آدرس‌های پستی و Zip code ها و .... دارد. روش‌های شناسایی خودکار متون با عنوان کلیOCR[۱]شناخته شده، برحسب اینکه متن به چه صورتی به سیستم ارائه می‌گردد، به دو دسته کلی تقسیم می‌شوند ۱- روش‌های بر خط[۲]: در این روش‌ها کلمات و حروف توسط کاربر روی یک صفحه مخصوص رسم و به‌صورت یک دنباله از نقاط که توسط مختصاتشان مشخص می‌شوند در اختیار سیستم قرار داده می‌شود. ۲- روش‌های برون خط[۳]: در این روش‌ها متن به‌صورت یک تصویر بدست آمده از Scanner یا دوربین‌های دیجیتال در اختیار سیستم قرار می‌گیرد. در این مقاله از روش برون خط به‌صورت کلمات نوشته شده با فونت نستعلیق که توسط نرم افزارPhotoshop CS به‌صورت عکس‌هایی در قالب JPG در آمده است استفاده شده است. یک عامل مهم در سیستم‌های شناسایی حروف روش‌های استفاده شده در تقطیع متن مورد نظر است. سه روش برای تقطیع در سیستم های مذکور وجود دارد[۱ و ۹]: ۱- راهکار کلاسیک[۴]: دراین روش‌کلمات به اجزاء قابل تشخیص شکسته می‌شوند. ۲- راهکار مبتنی بر شناسایی[۵]: این روش سعی در شکستن کلمات به حروف الفبای شناخته شده دارد ۳- راهکار کلی[۶]: در این روش سیستم، کلمه را به‌طور کامل و بدون نیاز به تقطیع شناسایی می‌کند. امتیاز روش کلی این است که مرحله شناسایی پیچیدگی چندانی ندارد. ولی سیستم قادر به شناسایی تعداد محدودی از کلمات است. دو روش دیگر بسیار قدرتمند عمل می‌کنند و محدودیتی در شناسایی کلمات ندارند. اما سیستم مربوط به آنها پیچیده‌تر است. تصویر

کلمات ابتدا تحت یک سری پیش‌پردازش برای مراحل بعدی آماده می‌شود. سپس درمرحله تقطیع براساس روش مبتنی بر شناسایی، کلمه به حروف یا زیرحروف تشکیل‌دهنده آن شکسته می‌شود. پس از آن یک مجموعه ویژگی از تصویر هر قطعه استخراج می‌شود. در مرحله بعدبا استفاده از شبکه‌های عصبی مصنوعی که با بردارهای ویژگی حاصل، آموزش دیده است، حروف شناسایی می‌شود. سپس با استفاده از یک الگوریتم سرچ در لغت‌نامه، کلمه شناسایی می‌شود.

۲– مشخصات نوشتار فارسی

نوشتار حروف در فارسی در چند جهت با نوشتار در زبان انگلیسی متفاوت است ۱- شکل حروف فارسی تابعی از موقعیت حرف در کلمه است. برای یک حرف ممکن است چهار شکل متفاوت بسته به موقعیت حرف در کلمه(ابتدای کلمه، وسط، آخر و یا حرف جدا) وجود داشته باشد. بعضی از حروف در سبک‌های نوشتاری مختلف و در موقعیت ثابت در کلمه به شکل‌های متفاوتی نوشته می‌شوند. ۲– تعدادی از حروف فارسی قابلیت متصل شدن به یکدیگر را دارند. درحالی که بعضی دیگر به هیچ حرفی وصل نمی‌شوند. ۳- به‌علت اینکه حروف دارای شکل‌های متفاوتی در موقعیت‌های مختلف در کلمه هستند، بنابراین تعداد کلاس‌های قابل تشخیص توسط سیستم به مقدار زیادی افزایش می‌یابد. بنابراین روش شناسایی سیستم پیچیده‌تر می‌شود. ۴- بعضی از حروف فارسی دارای۱، ۲ یا ۳ نقطه در بالا یا پایین خود هستند و گاهی تفاوت بین دو حرف در وجود یا عدم وجود و یا در تعداد این نقاط است. ۵- حروف فارسی بر خلاف حروف لاتین از راست به چپ نوشته می‌شوند. جدول(۱)

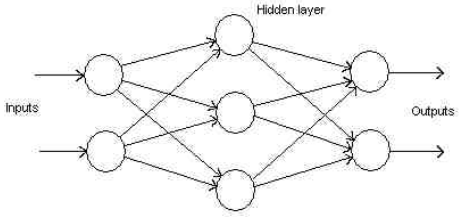
حروفی که تفاوت آن‌ها در وجود و یا عدم وجود سرکش‌ها و نقاط و تعداد آن‌ها است، در ابتدا جزء یک کلاس تشخیص داده می‌شوند. پس از تشخیص کلاسی که حرف مورد نظر به آن تعلق دارد و اضافه کردن اطلاعات مربوط به سرکش‌ها و نقاط، حروف کاملاً ازیکدیگر متمایز می‌شوند.

روش نوشتاری نستعلیق به خاطر زیبایی خاص آن معروف‌ترین روش نوشتاری فارسی است. در این مقاله، نستعلیق به‌عنوان روش نوشتاری کلمات برگزیده شده است.

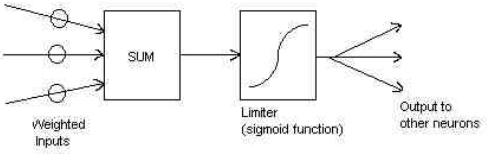
۳– شبکه عصبی

شبکه عصبی استفاده شده در این مقاله یک شبکه پرسپترون چند لایه است. با توجه به شکل‌های (۱) و (۲) شبکه فوق به این صورت عمل می‌کند: هر نرون یک سیگنال از نرون‌های لایه قبلی دریافت می‌کند، که هرکدام از این سیگنال‌ها در یک وزن جداگانه‌ای ضرب شده‌اند. سیگنال‌های حاصل، بایکدیگر جمع شده (شکل۲) و سپس به تابع محدودکننده تحویل داده می‌شوند. این تابع محدودکننده، خروجی نرون را در یک محدوده مشخص نگاه می‌دارد. خروجی این محدودکننده به تمام نرون‌های لایه بعد منتقل می‌شود. برای داشتن یک شبکه عصبی هوشمند باید وزن‌های مذکور محاسبه شوند. در این نوع شبکه خاص، سیستم با روش Back Propagation آموزش داده می‌شود و وزن‌های مذکور در این روش آموزشی محاسبه می‌شود.

برای آموزش شبکه، یک مجموعه از نمونه‌های ورودی و خروجی‌های صحیح مورد انتظار[۷] متناظر با هر ورودی تهیه شده سپس سیستم با این مجموعه آموزشی، آموزش داده می‌شود. به این صورت که با ورود یکی از نمونه‌ها به سیستم، خروجی براساس وزن‌های اولیه (در اولین مرحله وزن‌ها اعداد دلخواهی انتخاب می‌شوند) محاسبه می‌شود. سپس این خروجی با خروجی مورد انتظار مقایسه و یک سیگنال مربع خطا محاسبه می‌شود. از این مربع خطا برای اصلاح ضرائب (وزن‌ها) در هر لایه استفاده می‌شود. در هر مرحله از آموزش، مربع خطا کاهش می‌یابد. تمام این مراحل به طور چرخشی برای همه ورودی‌ها به ترتیب انجام می‌شود تا جائی که مربع خطا از یک حدی کمتر شود. در این صورت گفته می‌شود که سیستم آموزش دیده است. سیستم هیچ‌گاه کاملاً آموزش نمی‌بیند ولی می‌توان عملکرد آن را تا حد امکان بهینه کرد.



شکل ۱- ساختار یک شبکه عصبی



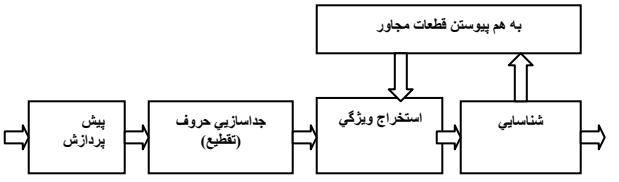
شکل۲- ساختار داخلی یک نرون

۴– سیستم پیاده سازی شده

شناسایی کلمات فارسی نستعلیق شامل مراحل زیراست. ابتدا یک سری پیش‌پردازش شامل حذف سرکش‌ها و زیرکش‌ها و نقاط، بر روی کلمات انجام می‌پذیرد. سپس کلمات حاصل به حروف سازنده آن، حروف و در برخی موارد به ترکیبی ازدو حرف شکسته می‌شود. پس از آن مجموعه‌ای از ویژگی‌ها از تصویر هر زیر کلمه استخراج می‌شود. شناسایی بر اساس بردار ویژگی حاصل انجام می‌پذیرد. شناسایی با استفاده از شبکه عصبی پرسپترون چند لایه انجام می‌شود. ابتدا بردارهای ویژگی حاصل از حروف برای تمام نمونه‌ها (حروف) روی هر ویژگی هنجارسازی[۸] می‌شود. سپس شبکه عصبی با این بردارهای ویژگی هنجارسازی شده، آموزش داده می‌شود. از این پس شبکه عصبی آموزش دیده برای شناسایی حروف مورد استفاده قرار می‌گیرد. مرحله شناسایی با بهره‌گیری از اطلاعات مربوط به لغت‌نامه و الگوریتم جستجو (شکل ۳) کامل می‌شود.

جدول ۱- جدول حروف فارسی و شکل هر حرف در موقعیت‌های مختلف در کلمه

	Character	Isolated	First	Middle	Last
1	Alef	ا(آ)	ا(آ)	ـا	ـا
2	Be	ب	بـ	ـبـ	ـب
3	Pe	پ	پـ	ـپـ	ـپ
4	Te	ت	تـ	ـتـ	ـت
5	Se	ث	ثـ	ـثـ	ـث
6	Jim	ج	جـ	ـجـ	ـج
7	Che	چ	چـ	ـچـ	ـچ
8	He	ح	حـ	ـحـ	ـح
9	Khe	خ	خـ	ـخـ	ـخ
10	Dal	د	د	ـد	ـد
11	Zal	ذ	ذ	ـذ	ـذ
12	Re	ر	ر	ـر	ـر
13	Ze	ز	ز	ـز	ـز
14	Zhe	ژ	ژ	ـژ	ـژ
15	Sin	س	سـ	ـسـ	ـس
16	Shin	ش	شـ	ـشـ	ـش
17	Sad	ص	صـ	ـصـ	ـص
18	Zad	ض	ضـ	ـضـ	ـض
19	Ta	ط	طـ	ـطـ	ـط
20	Za	ظ	ظـ	ـظـ	ـظ
21	Ayn	ع	عـ	ـعـ	ـع
22	Ghayn	غ	غـ	ـغـ	ـغ
23	Fe	ف	فـ	ـفـ	ـف
24	Ghaf	ق	قـ	ـقـ	ـق
25	Kaf	ك	كـ	ـكـ	ـك
26	Gaf	گ	گـ	ـگـ	ـگ
27	Lam	ل	لـ	ـلـ	ـل
28	Mim	م	مـ	ـمـ	ـم
29	Noon	ن	نـ	ـنـ	ـن
30	Waw	و	و	ـو	ـو
31	He	ه	هـ	ـهـ	ـه
32	Ye	ي	يـ	ـيـ	ـي



شکل ۳- ساختار سیستم پیاده‌سازی شده

۴–۱– پیش پردازش

با توجه به این‌که کلمات ورودی با فونت نستعلیق، توسط نرم افزارAdobe Photoshop CS به‌صورت عکس‌هایی در قالب JPG در آمده است، پس به نویززدایی از تصویر کلمه ورودی نیاز نیست. در این مرحله بخش‌های متصل در تصویر کلمه ورودی از یکدیگر جدا می‌شوند. برای یافتن اجزائ متصل یک روند تکراری (و نه بازگشتی)، با میزان پیچیدگی محاسباتی پایین بکار رفته است. در این الگوریتم ابتدا بالاترین پیکسل سیاه تصویر کلمه ورودی پیدا می‌شود. سپس با استفاده از یک عنصر ساختاری ساده با ابعاد ۳×۳ ، هشت همسایگی پیکسل مورد نظر به‌صورت زیر پیدا شده، بررسی می‌شوند:

$$X_1 = imdilate(X, SE_1) \qquad (۱-۱)$$

که در آن SE_1 عنصر ساختاری با ابعاد ۳×۳ و X ماتریس شامل پیکسل سیاه یافت شده و X_1 هشت همسایگی پیکسل مورد نظر می‌باشند.

حال پیکسل‌های سیاه این هشت همسایگی مشخص می‌شوند. با در نظر گرفتن این پیکسل‌های سیاه به‌هم متصل یافت شده به‌صورت یک جزء واحد، هشت همسایگی آن بررسی شده، پیکسل‌های سیاه یافت شده به این جزء متصل اضافه می‌شوند. این روند ادامه داده می‌شود تا زمانی‌که دیگر در هشت همسایگی پیکسل‌های یافت شده هیچ پیکسل سیاهی یافت نشود. حال این جزء متصل یافت شده ذخیره شده، از تصویر کلمه حذف می‌شود. روند فوق برای تصویر حاصل تکرار می‌شود تا زمانی که همه اجزاء متصل کلمه یافت شده، به‌صورت مجزا ذخیره شوند و در تصویر حاصل هیچ پیکسل سیاهی باقی نماند. این الگوریتم تمام بخش‌های متصل را به‌درستی جدا می‌کند.

مختصات بالاترین نقطه سیاه از سمت راست‌ترین ستون هر جزء متصل در مراحل بعدی بسیار مفید و کارساز خواهد بود. یک الگوریتم پیمایش کانتور خارجی با شروع از این نقطه، نقاط روی

کانتور خارجی را بدست آورده، ذخیره می‌کند. با رسیدن به نقطه شروع، این پیمایش متوقف شده، طول بدست آمده ذخیره می‌گردد.

پیمایش پیکسل‌های بخش متصل از بالاترین پیکسل سیاه در تصویر کلمه شروع می‌شود. پس از یافتن آخرین پیکسل هر بخش متصل، این بخش به‌صورت یک عکس ذخیره شده، از تصویر کلمه حذف می‌شود. این کارمرتباً تکرار می‌شود تا زمانی که دیگر پیکسل سیاهی در تصویر کلمه موجود نباشد[۳]. بخش‌های متصل یک کلمه در شکل(۴) مشاهده می‌شود. در این سیستم چون شناسایی کلمه برمبنای شناسایی حروف است، برای بازسازی کلمه شناسایی شده، ترتیب راست به چپ حروف بسیار مهم است. اما سرکش‌های حروفی مثل "ک" و "گ" و زیرکش‌های حروفی مثل "ع" و"غ" و"ح" و"خ"و.... معمولاً زودتر از حروف قبل از خود شناسایی شده، ترتیب راست به چپ حروف را دچار خطا می‌کنند (شکل ۵). لذا برای تشخیص صحیح ترتیب راست به چپ حروف در این مرحله لازم است که سرکش‌ها و زیرکش‌ها حذف شوند. در این سیستم، شناسایی اولیه حروف بدون درنظرگرفتن نقاط حروف و سرکش‌ها و علائم اضافه‌ای مثل کلاه الف(~)و..... انجام می‌شود. پس از تعیین دسته‌ای که هر حرف به آن تعلق دارد، با اضافه کردن اطلاعات مربوط به تعداد نقاط حرف و وجود یا عدم وجود سرکش‌ها و علائم اضافه، شناسایی کامل و حرف تشخیص داده می‌شود. پس در مرحله پیش‌پردازش نقاط و علائم اضافه نیز حذف می‌گردند .

(الف) (ب)

شکل ۴- جداسازی بخش‌های متصل کلمه الف-کلمه ورودی ب- بخش‌های متصل

(الف) (ب)

شکل ۵- سرکش‌ها و زیرکش‌ها در تعیین ترتیب راست به چپ حروف زودتر از حرف قبل از خود تشخیص داده می شوند(الف)- زیرکش. (ب)- سرکش

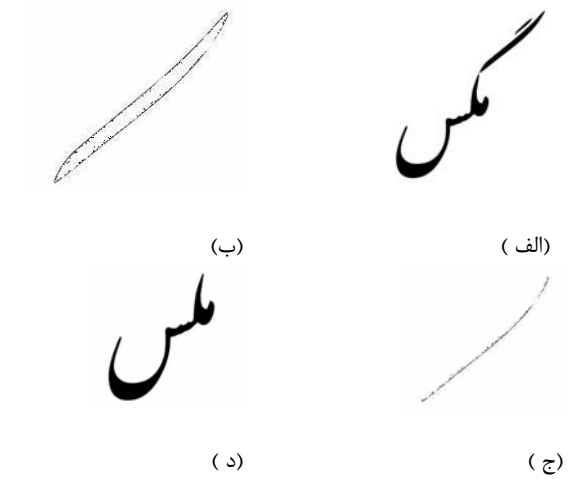
۴-۱-۱- حذف سرکش

برای حذف سرکش‌های حروفی مثل"ک" و"گ" از مشخصاتی که سرکش را از سایر بخش‌های متصل کلمه مجزا می‌کند، استفاده می‌شود. این مشخصات عبارتند از: شیب تقریبی ۴۵درجه، طول و کشیدگی بدون خمیدگی سرکش. با استفاده از این سه مشخصه، سرکش شناسایی و حذف می‌شود: ابتدا کانتور خارجی هر بخش متصل بدست می‌آید (شکل۶- ب). در حالی که این کانتور[۷و۳] در جهت عقربه‌های ساعت پیمایش می‌شود، به‌وسیله تعیین کد زنجیره‌ای[۹] [۷و۳] هر پیکسل ، کانتور پایینی هربخش متصل به دست می‌آید. پیکسل‌هایی که کد زنجیره‌ای آنها در جهت غرب، شمال غربی یا جنوب غربی قرار دارند، کانتور پایینی بخش متصل را تشکیل می‌دهند (شکل۶-ج). حال از بالاترین پیکسل کانتور پایینی شروع به پیمایش آن می‌نماییم. این پیمایش تا زمانی که شرط توقف غلط باشد، ادامه می‌یابد. شرط توقف وقتی برقرار می‌شود که حرکت در جهت قطری به سمت پایین و چپ (جنوب غربی) کمتر از تعداد دفعات مشخصی باشد (در این صورت شرط بدون خمیدگی بودن سرکش نقض می‌شود). با در نظر گرفتن طول و زاویه شیب سرکش پیدا شده، می‌توان آن را اعتبارسنجی کرد. حال با استفاده از مختصات نقاط شروع و پایان سرکش، می‌توان آن‌را از تصویر کلمه حذف کرد (شکل۶-د). این الگوریتم میزان صحت ۱۰۰٪ را درشناسایی و حذف سرکش‌ها نشان می‌دهد. در ادامه روش حذف سرکش به‌صورت الگوریتم آورده شده است:

روش حذف سرکش‌ها:

- ابتدا کانتور خارجی هر جزء متصل بدست می‌آید. برای بدست آوردن کانتور خارجی از یک نقطه (بالاترین پیکسل جزءمتصل) شروع کرده، در جهت عقربه‌های ساعت یا خلاف آن لبه تصویر جزء متصل پیمایش می‌شود و به‌این‌ترتیب با مشخص کردن پیکسل‌های موجود بر لبه تصویر، کانتور خارجی جزءمتصل مربوط مشخص می‌شود[۳] (شکل ۶- ب).
- حال با تعیین کد زنجیره‌ای [۳] برای پیکسل‌های کانتور خارجی، در حالی که کانتور خارجی در جهت عقربه‌های ساعت پیمایش می‌شود، می‌توان کانتور پایینی هر جزء متصل را مشخص کرد. به‌این‌صورت که وقتی کانتور خارجی در جهت عقربه‌های ساعت پیموده می‌شود، با تعیین پیکسل‌هایی که کد زنجیره‌ای آنها در جهت غرب، شمال غربی یا جنوب غربی قرار دارند، کانتور پایینی جزء متصل مشخص می‌شود (شکل ۶- ج).

- از بالاترین پیکسل کانتور پایینی جزءمتصل شروع به پیمایش کرده و این پیمایش تا زمانی که شرط توقف غلط باشد، ادامه داده می‌شود. شرط توقف وقتی برقرار می شود که حرکت بیش از تعداد دفعات مشخص در جهت قطری به سمت پایین و چپ (جنوب غربی) نباشد (در این صورت شرط بدون خمیدگی بودن سرکش نقض می‌شود).
- در حین این پیمایش تعداد انتقال‌های صحیح ،k۱ ، (یعنی در جهت جنوب غربی) و تعداد همه انتقال‌ها ، k ، (یعنی طول این قطعه از کانتوری که پیمایش شده است) شمارش می‌شود. پس از برقرار شدن شرط توقف اگر k نسبتاً کوچک باشد و $\frac{k_1}{k}$ نیز نسبتاً بزرگ باشد یک سرکش کوچک پیدا شده است. اگرK نسبتاً بزرگ و $\frac{k_1}{k}$ نیز نسبتاً بزرگ باشد، یک سرکش بزرگ پیدا شده است.
- برای هر جزءای که سرکش تشخیص داده شده است، تفاضل طول نقاط ابتدایی و انتهایی سرکش را بدست آورده و بر تفاضل عرض نقاط ابتدایی و انتهایی آن تقسیم می‌شود، به این ترتیب tg زاویه جزء سرکش تشخیص داده شده محاسبه می‌شود. اگر این مقدار بین ۰/۹ تا ۱/۴ باشد، سرکش تشخیص داده شده معتبر است.
- حال مختصات نقاط شروع و پایان سرکش ذخیره شده و در مرحله بعد با استفاده از این مختصات سرکش از تصویر کلمه حذف می‌شود. (شکل ۶- د)



(الف) (ب)

(ج) (د)

شکل ۶– مراحل حذف سرکش: الف – کلمه ورودی ب – کانتور خارجی سرکش ج- کانتور پایینی سرکش د- نتیجه حذف سرکش در کلمه

۴–۱–۲– حذف زیرکش

برای شناسایی زیرکش‌ها از ویژگی‌های زیر استفاده می‌شود: تصویر زیرکلمه شامل زیرکش، بدین‌صورت است که اگر از پایین‌ترین قطعه سیاه در اولین ستون سمت راست شروع کرده، ستون به ستون این قطعه سیاه به سمت چپ تا انتهای زیرکش دنبال شود، طول هم‌پوشانی این قطعه سیاه (زیرکش) با قطعه‌های سیاه بالایی وطول زیرکش نسبتاً بزرگ است، ثانیاً در اکثر ستون‌های زیرکش تعداد قطعه‌های سیاه بیشتر از یک می‌باشد. این ویژگی‌ها زیرکش را از سایر بخش‌های کلمه به خوبی مجزا می‌کند. بنابراین برای حذف زیرکش: از اولین ستون سمت راست که دارای بیش از یک قطعه سیاه است شروع کرده، مختصات پیکسل‌های ابتدا و انتهای پایین‌ترین قطعه سیاه این ستون ذخیره می‌شود (x_1, x_2). ستون مجاور سمت چپ (($k+1$) ام) ستون فعلی ((k)ام) بررسی می‌شود، اگر تعداد تداوم‌های موجود در ستون ($k+1$) ام در مجاورت قطعه سیاه فعلی (متعلق به ستون k ام) بیش از یک باشد، تداوم سیاه فعلی متعلق به زیرکش نیست (شکل ۷- ج). حال تعداد تداوم های ستون قبلی ($k-1$)ام، در محدوده x_1, x_2 مربوط به ستون k ام بدست آورده می‌شود، اگر این تعداد برابر یک باشد، بررسی ستون به ستون به سمت چپ به همین روال ادامه داده می‌شود. اما اگر این تعداد برابر دو باشد، همپوشانی زیرکش با قطعات بالا تمام شده است و به آخر زیرکش رسیده‌ایم. (شکل۷-د)

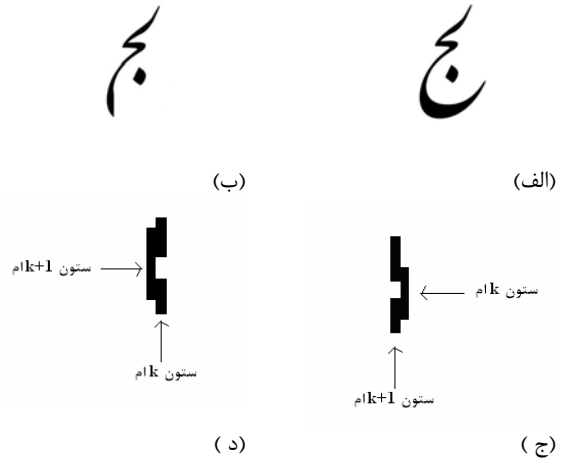
تذکر: با انتقال از یک ستون به ستون دیگر به سمت چپ، متغییر k که تعیین کننده تعداد ستون‌های پیموده شده است، یک واحد اضافه می‌شود (طول زیرکش). اگر k در یک محدوده منطقی باشد، زیرکش معتبر است. پس از تعیین زیرکش مختصات شروع و پایان آن ذخیره شده، در مرحله بعدی با استفاده از این مختصات زیرکش از تصویر کلمه حذف می‌شود. میزان صحت حذف زیرکش‌ها با این روش برابر ۱۰۰٪ است. در ادامه روش حذف زیرکش به‌صورت الگوریتم آورده شده است:

روش حذف زیرکش‌ها:

- از آنجایی که زیرکش با قطعات بالایی خود همپوشانی دارد بنابراین از سمت راست‌ترین ستونی که دارای بیش از یک قطعه سیاه است شروع کرده، مختصات پیکسل‌های ابتدا و انتهایی پایین‌ترین قطعه سیاه این ستون را بدست می‌آوریم.(x2,x1)
- ستون سمت چپ (k+1 ام) مجاور ستون فعلی (k۱م) بررسی می‌شود، اگر در مجاورت قطعه سیاه فعلی (متعلق

به ستونk ۱م) فقط یک قطعه سیاه در ستون k+1م۱ وجود داشته باشد، مختصات نقاط بالا و پایین این قطعه سیاه را بدست می‌آوریم(x4,x3)در غیراین‌صورت تداوم سیاه ستون k۱م متعلق به زیرکش نیست (شکل ۷- ج).

- حال تعداد تداوم‌های ستون قبلی (k)، در محدوده x4,x3 مربوط به ستون k+1م را بدست می‌آوریم، اگر این تعداد برابر ۱ باشد، بررسی ستون به ستون به سمت چپ به همین روال ادامه داده می‌شود. اما اگر این تعداد برابر ۲ باشد یعنی همپوشانی زیرکش با قطعات بالا تمام شده است و به آخر زیرکش رسیده‌ایم. بنابراین مختصات ابتدای تداوم سیاه ستون فعلی (k+1) بدست می‌آوریم و از این نقطه برای برش زیرکش استفاده می‌کنیم (شکل ۷ - د).
- اگرk از حد مشخصی بزرگتر باشد، زیرکش معتبر می‌باشد.



شکل ۷– حذف زیر کش :الف – تصویر کلمه ورودی ب – نتیجه حذف زیرکش ج- حالتی که قطعه سیاه متعلق به زیرکش نیست د– حالتی که به انتهای زیرکش رسیده است

۴–۱–۳– حذف نقاط و علائم اضافه

تشخیص نقاط و علائم اضافه از روی ویژگی‌هایی مانند کوچکی اندازه و قرار گرفتن آن‌ها در محدوده زیرکلمه‌های بزرگتر امکان‌پذیراست. این مشخصات به گونه‌ای در نظر گرفته می‌شود که هیچ زیرکلمه دیگری غیر از زیرکلمه مورد نظر در این محدوده قرار نگیرد. الگوریتم حذف این نقاط و علائم اضافه در ادامه آورده شده است:

الگوریتم حذف نقاط و علائم اضافه:

۱ـ اجزاء متصل بدست آیند.

۲ـ برای هر جزء متصل:

۲-۱- اگر $Width1 \le tw1 * Est$ و $Height1 \le th1 * Est$ یا $Width1 \le tw2 * Est$ و $Height1 \le th2 * Est$ و $Width1 \le tw3 * Est$ و $Height1 \le th3 * Est$ باشد

۲ـ۱ـ۱ـ اگر جزء متصل دیگری وجود داشت که پهنای جزء متصل جاری به‌طور کامل در محدوده پهنای آن قرار گیرد،جزء متصل جاری، یک cs است و باید حذف شود.

۲-۱-۲- جزء متصل جاری از تصویر حذف گردد

Width1 و Height1 به‌ترتیب پهنا و ارتفاع جزء متصل جاری را نشان می‌دهند.

منظور از CS (Characteristic Strokes) : تکه نوشته‌های مشخصه) نقطه‌های حروف و کلاه‌الف(~) و همزه می‌باشد.
ثابت‌های عددی تعریف شده عبارتند از:

th1=2.5 , tw1=2.5 (نقطه)
th2=2.5 , tw2=6 (~)(کلاه الف)
th2=2 , tw2=4 (~)(همزه)

پارامترهای الگوریتم به‌گونه‌ای در نظر گرفته شده‌اند که تا جایی که ممکن است زیرکلمه دیگری به‌جای نقطه‌ها یا علائم حذف نشود. این الگوریتم تمام نقاط و علائم اضافه را به‌درستی شناسایی و حذف می‌کند.

۴–۲– تقطیع و تعیین ترتیب راست به چپ

در اینجا چند روش تقطیع[۱۰] معرفی می‌شود: روشی که برمبنای هیستوگرام عمودی[۱۱] و یا خط زمینه[۵-۶] است. این روش برای تقطیع کلمات نوشته شده به‌خصوص کلمات فارسی (نستعلیق) به‌خاطر تنوع نوع نوشتار آن مناسب نیست. دو روش دیگر که برای تقطیع کلمات فارسی(نستعلیق) مناسب‌تر به‌نظر می‌رسند عبارتند از: روش تقطیع با استفاده از اجزاء منفرد و منظم [۲] که اجزاء منظم را به‌عنوان محل‌های اولیه تقطیع در نظر می‌گیرد. روش دیگر روشی است که براساس آنالیز کانتور بالایی عمل می‌کند. از آنجایی که در روش‌های مبتنی بر آنالیز کانتور بالایی احتمال عدم تقطیع و تقطیع

زیادی (ایجاد بیش از یک برش بین دو حرف) کمتر از سایر روش‌ها است، پس این روش مناسب‌تر می‌باشد. در ادامه روش تقطیع و تعیین ترتیب راست به چپ حروف بیان می‌شود.

۴-۲-۱- تقطیع با استفاده از مینیموم‌های کانتور بالایی

در این روش [۴] کمینه‌های محلی کانتور بالایی به‌عنوان محل‌های اولیه انجام برش در نظر گرفته می‌شوند. سپس در صورت برقرار بودن مجموعه‌ای از شرایط(که در ادامه توضیح داده خواهد شد (بخش۴-۲-۴))، تقطیع در این نقاط انجام می‌شود. با انجام یک سری اصلاحات در این روش می‌توان آن را برای تقطیع کلمات فارسی (نستعلیق) بهبود داد.از آنجا که در خط نستعلیق حرف "ر" چسبان و مواردی مثل"ح"موجود در وسط کلمه(شکل (۸)) بدون داشتن کمینه محلی پیش از خود نوشته می‌شوند، این روش قادر به تقطیع آنها نیست. بنابراین ابتدا یک الگوریتم جدید برای شناسایی و تقطیع این موارد بکار می‌رود. سپس کانتور بالایی با یک روش ساده بدست آمده و کمینه‌های محلی آن به‌عنوان نقاط اولیه تقطیع بدست می‌آید. آنگاه این نقاط اولیه اعتبار سنجی شده، در محل‌های معتبر به طریق مناسب تقطیع انجام می‌شود. این مراحل در ادامه توضیح داده شده است.

شکل ۸- -حرف" ح" و" م" در وسط کلمه

۴-۲-۲- شناسایی و تقطیع "ر" چسبان

برای کشف "ر"چسبان از روشی مشابه با روش کشف سرکش‌ها استفاده شده است. در پیدا کردن "ر" چسبان برخلاف آنچه در یافتن سرکش انجام شد، از پایین ترین نقطه کانتور پایینی در جهت خلاف عقربه ساعت شروع به پیمایش آن می‌نماییم. شرط توقف وقتی برقرار می‌شود که حرکت در جهت قطری و به سمت بالا و راست (شمال شرقی) کمتر از تعداد دفعات مشخصی باشد (در این صورت شرط بدون خمیدگی بودن نسبی"ر"چسبان نقض می‌شود). یعنی زمانی که به تعداد مشخصی انتقال در جهت‌هایی غیر از جهت شمال شرقی رسیدیم، پیمایش متوقف می‌شود. زمانی" ر" پیدا شده معتبر است که طول آن دریک محدوده منطقی باشد. در این صورت حرف"ر" از قسمت‌های دیگر زیرکلمه جدا می‌شود. میزان صحت این الگوریتم ۱۰۰٪ می‌باشد.

شکل ۹- نمونه‌هایی از نتایج الگوریتم جداسازی " ر" چسبان در کلمه

۴-۲-۳- شناسایی همپوشانی و انجام تقطیع مناسب

تقطیع با استفاده از کمینه‌های کانتور بالایی علاوه بر ناتوانی در تقطیع "ر" چسبان، برای مواردی مثل"ح" و "م" موجود در وسط کلمه (شکل (۸)) نیز قادر به انجام عمل تقطیع نمی‌باشد. این مشکل با یافتن همپوشانی‌ها و تقطیع مناسب در آن‌ها رفع می‌شود. برای شناسایی همپوشانی: در هر زیرکلمه از اولین ستون سمت راست شروع کرده، تعداد، طول و محل تداوم‌های سیاه موجود در ستون جاری بدست می‌آید. اگر تعداد تداوم‌ها بیش از یک باشد و درستون مجاور سمت راست پایین‌ترین تداوم سیاه آن هیچ پیکسل سیاهی یافت نشود، این ستون به‌عنوان شروع همپوشانی در نظر گرفته می‌شود. سپس به سمت چپ حرکت می‌کنیم تا دو قطعه همپوشانی به‌هم بپیوندند. به این ترتیب دوستونی که بین آنها همپوشانی وجود دارد بدست می‌آید. حال با پیمایش کانتور خارجی، اگر همپوشانی یافت شده مربوط به یک حلقه نباشد، در محلی از بخش بالای همپوشانی که دارای کمترین پهنا است، برش زده می‌شود. میزان صحت این الگوریتم ۱۰۰٪ می‌باشد.

شکل۱۰- نمونه‌هایی از نتایج الگوریتم کشف همپوشانی و انجام تقطیع مناسب

۴-۲-۴- یافتن کمینه‌های کانتور بالایی و انجام تقطیع

در هر زیرکلمه، با پیمایش کانتور خارجی در جهت خلاف قربه‌های ساعت و با استفاده از کد زنجیره‌ای، پیکسل‌هایی که در

جهت‌های غرب، شمال غربی و جنوب غربی می‌باشند مشخص می‌شوند. به‌این‌ترتیب کانتور بالایی بدست می‌آید.

شکل۱۱-کانتور بالایی کلمه" مگس"

حال برای هر بخش متصل در حین پیمایش کانتور بالایی، با استفاده از کد زنجیره‌ای، نقاطی که در آن‌ها یک روند کاهشی تبدیل به یک روند افزایشی می‌شود به‌عنوان کمینه‌های محلی کانتور بالایی ذخیره می‌شود. این نقاط به‌عنوان محل‌های اولیه تقطیع بکار می‌روند. برای جلوگیری از تقطیع زیادی مانند ایجاد برش در بین دندانه‌های حروفی مثل"س" و "ش" شرایط زیر بررسی می‌شود: فاصله مجاز بین دو تقطیع متوالی، فاصله مجاز محل تقطیع با شروع و پایان بخش متصل، حداقل پهنای مجاز خط در آن محل و شکل کانتور بالایی حروف. پس از تعیین قطعی محل‌های مجاز، برش در آن محل‌ها انجام می‌شود. میزان صحت تقطیع با این روش ۹۵٪ است.

شکل ۱۲- نمونه‌هایی از نتایج تقطیع کلمات با استفاده از روش یافتن کمینه‌های محلی کانتور بالایی

۴-۲-۵- یافتن ترتیب را ست به چپ

برای یافتن ترتیب راست به چپ حروف (زیر کلمه‌ها)، از کلمه تقطیع شده فاقد سرکش‌ها و زیرکش‌ها و نقاط استفاده می‌شود. در این کلمه تقطیع شده، زیر کلمه‌ها بر اساس سمت راست‌ترین ستون آنها مرتب می‌شوند. به‌این‌صورت ترتیب زیرکلمات مشخص می‌شود.

۴-۳- استخراج ویژگی

مهمترین عامل در شناسایی حروف، روش استخراج ویژگی است. در میان روش‌های استخراج ویژگی گوناگون مناسب‌ترین آن‌ها بر اساس میزان صحت شناسایی انتخاب می‌شود[۲]. در شناسایی حروف، هشت ویژگی شامل سه توصیفگر فوریه (توصیفگرهای۱تا۳)، تعداد حلقه‌ها، نسبت ارتفاع به پهنا، نسبت نقاط سیاه به کل نقاط و اتصال چپ و راست حروف بیشترین میزان صحت را نشان می‌دهد. این ویژگی‌ها برای تمام حروف کلمات مورد آزمایش محاسبه می‌شود. حال هر ویژگی بر روی تمام نمونه‌های حاصل از کل کلمات مورد آزمایش هنجارسازی می‌شود. در نتیجه میانگین هر ویژگی برابر صفر و واریانس آن برابر یک است.

۴-۴- شناسایی

سیستم شناسایی شامل یک شبکه عصبی پرسپترون چند لایه می‌باشد[۹]. این شبکه عصبی در اینجا شامل یک لایه پنهان و یک لایه خروجی است. باروش سعی وخطا و درنظرگرفتن سیزده نرون در لایه پنهان و تابع غیر خطی تانژانت هیپربولیک برای لایه‌های پنهان‌وخروجی و تابع آموزشی(Resilient backpropagation) [۱۰] نتیجه آزمون نمونه‌های آموزشی ۹۶ ٪ و نتیجه آزمون نمونه‌های آزمایشی ۹۳ ٪ به‌دست می‌آید. لازم به ذکر است که در این مرحله از شناسایی نقاط و سرکش‌ها خود به‌عنوان دسته‌های خروجی مجزا در نظر گرفته می‌شوند. سایر دسته‌ها شامل حروف یا بعضی از ترکیبات حروف که قابل تقطیع از یکدیگر نیستند (شکل ۱۳)، بدون نقطه و سرکش‌های آنها در نظر گرفته می‌شوند. تعداد کل نمونه‌های (حرف‌ها یا زیرکلمه‌ها) حاصل از همه کلمات تحت آزمایش و آموزش سیستم(۳۲۰ کلمه)برابر ۱۸۰۵ نمونه است. شبکه عصبی با این بردارهای ویژگی، آموزش داده می‌شود و به‌عنوان سیستم شناسایی اولیه بکار می‌رود. حروفی که تفاوت آنها فقط در نقاط و سرکش‌ها می‌باشد مانند" ت "و" ب " یا " ک " و " گ " و " لام کوچک" (مثل لبه) در یک گروه قرار می‌گیرند. حال مرحله بعدی اضافه کردن اطلاعات مربوط به نقاط و سرکش‌ها به دسته‌بندی اولیه و جداسازی کامل حروف از یکدیگر است .

شکل ۱۳- زیرکلمه " کا "در کلمه "کاه" قابل تقطیع نیست

برای اضافه کردن نقاط به حروف ، از میزان همپوشانی نقطه با حرف استفاده می‌شود. برای اضافه کردن سرکش‌ها به این نکته توجه می‌شودکه آخرین پیکسل مربوط به سرکش، کمترین فاصله را با بالاترین پیکسل حرف مربوط به خود دارد.

اکنون با استفاده از دو روش می‌توان شناسایی حروف را کامل کرد. در روش اول یک بردار ویژگی جدید شامل موارد ذیل تشکیل داده می‌شود: هشت ویژگی استفاده شده در آموزش شبکه عصبی مرحله قبل، ویژگی‌های مربوط به تعداد نقاط هر حرف و موقعیت نقطه نسبت به حرف (بالاوپایین)، وجود یا عدم وجود سرکش‌ها وکلاه الف، وجود یا عدم وجود دنباله" ت"و " ن" (چون در مرحله تقطیع در کمینه‌های محلی کانتور بالایی برش انجام می‌شود، وسط حروف بزرگی مثل " ت" و" ن" نیز برش داده می‌شود (شکل (۱۴)). به این ترتیب یک بردار ویژگی جدید شامل ۱۳ ویژگی تشکیل داده می‌شود. این بردار ویژگی پس از هنجارسازی طبق روش ذکرشده در بخش(۴-۳) به همراه خروجی‌های مورد انتظار، برای تعلیم یک شبکه عصبی پرسپترون چند لایه جدید بکار می‌رود. آزمایش این روش میزان صحت تشخیص ۸۳٪ را نشان می‌دهد. این شبکه عصبی جدید نیز دارای ۱۳ نرون در لایه پنهان و تابع غیر خطی تانژانت هیپربولیک برای لایه‌های پنهان و خروجی و تابع آموزشی (Resilient back propagation) می‌باشد.

شکل ۱۴– تقطیع در وسط حروف " ن" و" ت"

در روش دوم براساس مشخصات حروف و مشاهده خروجی مربوط به شبکه عصبی اول، قوانینی بکار برده می‌شود که میزان صحت شناسایی سیستم را ارتقاء می‌دهد. این قوانین به‌همراه اطلاعات مناسب دیگر می‌توانند در تشخیص نهایی حروف به‌گونه‌ای بکار روند که خطاهای مربوط به خروجی شبکه عصبی اول نیز تا حد امکان اصلاح شود.

در اینجا بردار ویژگی شامل خروجی شبکه عصبی اول، تعداد نقاط و موقعیت آن‌ها نسبت به حرف، وجود یا عدم وجود کلاه الف وسرکش‌ها و دنباله حروف، چگونگی اتصال چپ و راست حروف و وجود یا عدم وجود حلقه کامل در حرف می‌باشد. سپس این بردار ویژگی را با بردار ویژگی تمام زیرکلمه‌ها (۸۴ زیرکلمه) مقایسه می‌کنیم. اگر بردار ویژگی مورد نظر با بردار ویژگی یکی از زیرکلمه‌ها کاملا مشابه باشد، حرف مورد نظر شناسایی شده است. اما اگر بردار ویژگی مورد نظر با هیچ کدام از آن‌ها کاملاً مشابه نباشد، زیرکلمه‌هایی که حداقل اختلاف را با برداریژگی مورد نظر دارند در نظر می‌گیریم. سپس با استفاده از قوانین مناسب زیرکلمه‌ای که مناسب‌تر است، به عنوان نتیجه شناسایی مشخص می‌شود. این قوانین شامل شباهت‌های ظاهری زیرکلمه‌ها مانند کشیدگی، بلندی، میزان انحنا و حلقه‌های موجود در زیرکلمه است. عملکرد سیستم با این روش ۹۵٪ صحت در شناسایی زیرکلمه‌ها را نشان می‌دهد.

۴–۵– الگوریتم جستجو

در این قسمت زیرکلمه‌های شناسایی شده توسط سیستم به ترتیب در یک بردار قرار داده می شود (معادل هر زیرکلمه یک عدد تعریف شده است). حال یک ماتریس که هر سطر آن شامل زیرکلمه‌های یک کلمه از لغت نامه سیستم می‌باشد، تهیه کرده، بردار خروجی سیستم با سطر به سطر این ماتریس مقایسه می‌شود. کلمه‌ای (سطری از ماتریس) از لغت نامه که حداقل خطا را نسبت به کلمه شناسایی شده توسط سیستم دارد به‌عنوان خروجی صحیح در نظر گرفته شده، بر اساس آن زیرکلمه‌هایی که اشتباهاً شناسایی شده‌اند اصلاح می‌شوند. میزان صحت شناسایی سیستم با استفاده از الگوریتم جستجو به ۹۷٪ افزایش می‌یابد.

۵– نتیجه گیری

با استفاده از مجموعه نمونه‌های آموزشی، سیستم برای شناسایی کلمات فارسی(نستعلیق) آموزش داده شد. سپس عملکرد آن بر روی مجموعه نمونه‌های آزمایشی ارزیابی شد. نتایج آزمایش مراحل مختلف سیستم شناسایی در جدول (۲) آورده شده است. بررسی نشان می‌دهد که الگوریتم جستجو با استفاده از دانش واژگانی افزایش نسبی خوبی در بازشناسی کلمه بوجود می‌آورد.

جدول ۲- نتایج آزمایش مراحل مختلف سیستم شناسایی

	حذف سرکش	حذف زیرکش	تقطیع
میزان صحت	٪۱۰۰	٪۱۰۰	٪۹۵

	شناسایی حروف بدون نقطه وسرکش	شناسایی حروف با نقطه وسرکش	الگوریتم جستجو
میزان صحت	٪۹۳	٪۹۵	٪۹۷

۵– تقدیر و تشکر

با سپاس از آقای دکترپیمان ادیبی، به خاطر هم‌فکری‌هایی که در پیشبرد این مقاله داشتند.

۶– مراجع

[1] G.Nicchiotti, C.Scagliola; **"A simple and effective cursive word segmentation method"** hwr.nici.kun.nl/iwfhr-7-pp 499-503 2000

[2] R. Safabakhsh, P.Adibi; **"Nastaaligh handwritten word recognition using a continuous-density variable-duration HMM"**, The Arabian Journal for Science and Engineering, Vol 30, No 1 B. April 2005

[3] R.C.Gonzalez, R.E.Woods; **"Digital Image Processing"**, Prentice-Hall,Inc,pp,644-646,534-536,2004

[4] C.Olivier, H.Miled; **"Segmentation and Coding of Arabic Handwritten",** IEEE proc of 13 th Int .Conf. on Pattern Recognition , Vol.3 , pp.264-268,1996

[5] C.Dunn, P.Wang; **"Character Segmention Techniques for Handwritten Text---A survey",** Proc. 11th Int'l Conf. Pattern Recognition, Vol. 2, p.577, Aug. 2001.

[6] S. Tsujimoto, H. Asada; **"Major Components of Complete Text reading System",** Proc. IEEE, Vol. 80, No. 7, pp. 1133-1149, July 1992.

[7] M. Yamani, M. Noorzaily**; " Parking System Using Chain Code & A-Star Algorithm"**, myais.fsktm.um.edu.my 2006

[8] A.L.Betker1, T.Szturm2**; " Application of Feedforward Backpropagation Neural Network",** IEE ,2003

[9] R. G. Casey, E. Lecolinet; **"A Survey of Methods and Strategies in Character Segmentation"** IEEE Trans. Pattern Anal.Mach. Intell., 18 (7) (1996), pp. 690–706.

[10] S.Fritsh, F.Guenther; **"The neuralnet package"**, October 2008

۷– پی‌نوشت‌ها

1- Optical Character Recognition
2- Online
3- Offline
4- Classic Strategy
5- Recognition Based Strategy
6- Holistic Strategy
7- Target
8- Normalization
9- Chain Code
10- Segmentation
11- Vertical Histogram