عنوان زیرپروژه:

بررسی نیازمندی‌های پروژه‌های زیرساختی نویسه‌خوان نوری فارسی به پیکره متنی زبان فارسی

فهرست مطالب

عنوان **شماره صفحه**

عنوان **شماره صفحه**

عنوان **شماره صفحه**

عنوان **شماره صفحه**

عنوان **شماره صفحه**

فهرست شکل‌ها

عنوان **شماره صفحه**

عنوان **شماره صفحه**

فهرست جدول‌ها

عنوان **شماره صفحه**

1 - مقدمه

نویسه‌خوان نوری تکنیکی برای تبدیل متون تصویری و غیر قابل ویرایش به متون قابل ویرایش و تشخیص توسط کامپیوتر می‌باشد. به عبارت دیگر اصطلاح نویسه‌خوان نوری[1] به تکنیک‌هایی اطلاق می شود که درتصاویر اسکن یا فکس شده، نواحی متنی را تشخیص می دهند و سپس این نواحی (تصویری) را به متن قابل ویرایش تبدیل می نماید [1] .

. یک نرم افزار نویسه‌خوان نوری تصویر اسکن شده را می خواند، محتویات آن (شامل متن، خطوط، تصاویر، جداول، ...) را شناسایی می نماید، و سپس آن را به یک قالب قابل ویرایش تبدیل می کند. نویسه خوان نوری، به عنوان تنها ابزار بازیابی اطلاعات متنی از تصویر، یکی از مهم‌ترین ابزارهای تبدیل اطلاعات موجود به صورت قابل استفاده و پردازش در رایانه‌ها و در نتیجه یکی از ارکان مهم تحقق محیط رایانه ای فارسی به شمار می آید.

مبحث نویسه‌خوان نوری سال‌هاست مورد توجه بسیاری از محققین در سراسر جهان قرار دارد و مقالات مرتبط ، نویسه خوان نوری یا با نویسه‌خوانی نوری همه ساله در بسیاری از کنفرانس‌های معتبر بینایی ماشین، شبکه‌های عصبی، هوش مصنوعی، پردازش تصویر و ... در سطح جهان ارائه می‌شود. با وجود تحقیقات فراوان و سرمایه گذاری‌های بسیاری در این زمینه، پیچیدگی ذاتی مسائل مطرح در این حوزه و گستردگی علوم و فناور یهای مرتبط با آن سبب شده است که راه زیادی تا حصول محصولات نویسه خوان نوری ایدهال باقی باشد.

[1] Optical Character Recognition

2 - مزایای و کاربردهای نویسه‌خوان نوری

امروزه بیشتردستگاه‌های اسکنر به نرم افزارهای نویسه‌خوان نوری مجهزگردیده اند و قادرند متن موجود در یک سند اسکن شده را تشخیص دهند و آن را با همان نحوه قالب بندی، ستون بندی، جدول بندی و نوع فونت مطابق با سند کاغذی اصلی، درقالب یک فایل متنی با قالب بندی مناسب ذخیره نمایند. سیستم‌های نویسه‌خوان نوری با حذف نقش تایپیست‌ها در فرآیند تبدیل اسناد کاغذی به قالب الکترونیکی، سرعت ورود اطلاعات به رایانه را ده‌ها برابر افزایش می دهند و روند انجام این فرآیند را به میزان قابل توجهی تسهیل می کنند. امروزه بازار مصرف سیستم‌های نویسه‌خوان نوری طیف بسیار وسیعی از موسسات (شامل مراکز نشر، دانشگاه‌ها، کتابخانه‌ها، بانک‌ها، ادارات پستی، شرکت‌های بیمه و..) را دربرمی‌گیرد.

استفاده ازسیستم‌های نویسه‌خوان نوری دو مزیت عمده دارد :

الف: افزایش چشمگیر سرعت دسترسی به اطلاعات؛ زیرا درمتن برخلاف تصویر، امکان جستجو و ویرایش وجود دارد.

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

چنین قابلیتی امکان استفاده گسترده از رایانه را در پردازش سریع حجم وسیعی از داده‌های مکتوب شرکت‌ها و موسسات مختلف (نظیر بانک‌ها، شرکتهای بیمه، موسسات خدمات عمومی، ادارات پست، و دیگر نهادهایی که سالانه با میلیون‌ها مورد پرداخت، دریافت و حسابرسی امورمشتریان خود مواجه اند فراهم می آورد [2].

نویسه‌خوان نوری یکی از بخش‌های مورد توجه در پردازش تصویر است .زیرا بسیاری از متون موجود بصورت چاپی بوده و می بایست برای جستجو و انجام پردازش‌های بعدی در کامپیوتر به صورت متن ذخیره شوند .نویسه‌خوان نوری در موارد دیگری نیز همچون پردازش خودکار اسناد و مدارک، خواندن آدرس‌های پستی چاپی و مرتب‌سازی خودکار نامه‌ها، کتابخانه دیجیتال، کمک به افراد نابینا برای خواندن، ذخیره و بازیابی متون، توسعه رابطه انسان و کامپیوتر کاربردهای زیادی دارد. هدف نهایی در بازشناسی

متون، تقلید از قابلیت انسان با سرعت بسیار بالاتر می باشد. یک سیستم نویسه‌خوانی باید حداقل دقت 99/9 % را داشته باشد [3]. در ادامه به طور جزئی‌تری هر یک از کاربردهای آن توضیح داده خواهد شد.

2-1- ورود داده

روش‌های ثبت حجم زیادی از اطلاعات با چهارچوب مشخص، در این حوزه قرار می‌گیرند. به عنوان مثال، دستگاه‌های چک‌خوان که در بانک‌ها برای ورود داده استفاده می‌شوند. این سیستم‌ها برای خواندن بخش‌های کوچکی از تصویر چک شامل نویسه‌های چاپی و دستنویس مانند شماره حساب، امضای مشتری، شماره برگ چک و مبلغ چک طراحی می‌شوند. ساختار کاغذ چک معمولاً ثابت است و نواحی مشخصی از آن توسط ماشین خوانده می‌شود.

دستگاه‌های چک‌خوان در حدود 150000 چک را در یک ساعت می‌خوانند و نرخ وازدگی و خطای آن‌ها کم است. این سیستم‌ها در کیفیت‌های پایین چاپ نیز به خوبی عمل می‌کنند. این دستگاه‌ها برای کاربردهای خاص طراحی می‌شوند و قیمت بالایی دارند.

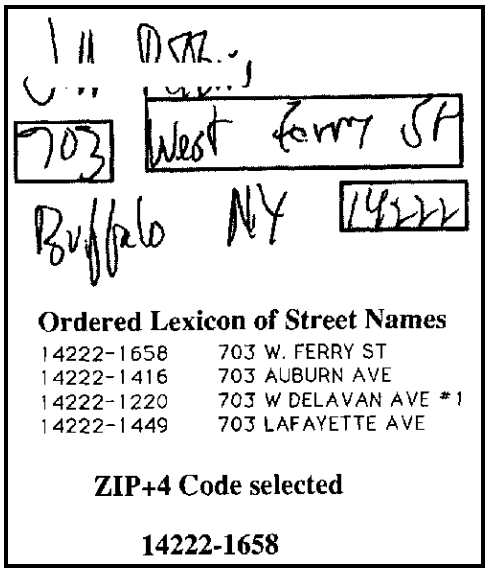
2-2- ورود متن

کاربرد دیگر سیستم‌های بازشناسی متن در بازشناسی سند برای استفاده در اتوماسیون اداری (ورود متن) است. این سیستم‌ها اسناد با قلم‌های محدود و کیفیت چاپ مشخصی را بازشناسی می‌کنند. این سیستم‌ها برای ورود حجم زیادی از متن به شکل قابل پردازش با نرم‌افزارهای *پردازشگر* متن‌استفاده می‌شوند. قابلیت این سیستم‌ها در ورود متن، رقیب قدرتمندی برای وارد کردن دستی اطلاعات است.

کارایی این سیستم‌ها به کیفیت چاپ اسناد وابستگی زیادی دارد. بطوری که در یک تصویر با کیفیت مناسب، نرخ بازشناسی بالای 99% تولید می‌کنند. سرعت خواندن این سیستم‌ها در حدود چند صد نویسه در دقیقه است.

2-3- خودکارسازی فرایند

در این کاربرد، کنترل یک فرایند مشخص به بازشناسی دقیق متن نوشته شده اولویت دارد. به عنوان مثال در خودکارسازی فرایند جداسازی نامه در مراکز پستی، بازشناسی نام شهر مقصد از بازشناسی دقیق کل متن نوشته شده بر روی پاکت نامه اولویت بالاتری دارد. نرخ بازشناسی این سیستم‌ها کاملاً به کیفیت تصویر پاکت نامه وابسته است. این سیستم‌ها معمولاً وازدگی بالا و نرخ خطای در حد صفر دارند. سرعت جداسازی نامه‌ها بر حسب آدرس، معمولاً در حدود چند ده هزار نامه در ساعت است. در شکل 1 نمونه‌ای از آدرس پستی دستنویس روی یک نامه و متن بازشناسی شده آن آمده است. با توجه به شکل ملاحظه می‌شود که برای آدرس نوشته شده 4 نامزد آدرس و کد پستی بر حسب شباهت به ورودی تولید شده است.



شکل 1- نمونه ای از بازشناسی آدرس و کد پستی[4]

2-4- بازیابی اسناد

هزینه بالای کار با پایگاه‌های داده بزرگ از تصاویر اسناد، نیاز به روش‌های خودکار و کارآمد برای دستیابی به اطلاعات داخل این تصاویر را بوجود آورده است. در تلاش برای حرکت به سوی ادارات بدون کاغذ، حجم بزرگی از اسناد کاغذی روبش شده و به صورت تصویری ذخیره می‌شوند، بدون اینکه برچسب اطلاعاتی مناسبی به آن‌ها زده شود. یک راه برای برچسب‌زنی و بازیابی پایگاه داده، تبدیل کامل اسناد به شکل الکترونیکی است که می‌توانند به طور خودکار برچسب زده شوند. موانعی مثل کیفیت پایین تصویر و هزینه بالای تبدیل در مقابل این روش وجود دارد. از طرفی چون بسیاری از اجزاء غیر متن را نمی‌توان به طور کامل در شکل تبدیل شده نمایش داد، داشتن یک کپی از سند در شکل تصویری لازم به نظر می‌رسد [5].

نگهداری و دسترسی به یک پایگاه داده متنی مسأله مهمی است. در چنین پایگاهی، مستند جدیدی که وارد می‌شود؛ لزوماً ساختاری مشابه مستندات قبلی ندارد، بنابراین تعریف ساختارهای مختلف از قبل، غیر ممکن است. توقعی که کاربران از یک سیستم پایگاه داده مستندات دارند، بازیابی دقیق مستند مورد نظر نیست، اما انتظار می‌رود که مکانیزمی برای بازیابی مستندات، به صورت مرتب شده بر اساس شباهت به مستند مورد پرس و جو، وجود داشته باشد. برای بازیابی اسناد مبتنی بر محتوا، می‌توان بلوک‌های متن تصویر سند را بازشناسی کرد و از آن‌ها برای بازیابی سند استفاده کرد [6].

2-5- کتابخانه دیجیتال

در یک کتابخانه دیجیتال برای دسترسی به کتاب مورد نظر از کلید واژه‌ها استفاده می‌شود. کلید واژه‌ها به صورت متن وارد می‌شوند و نرم‌افزار کتابخانه، کتاب‌های مرتبط با آن‌ها را بازیابی می‌کند. پیدا کردن کلمات کلیدی از تصاویر روبش شده کتاب‌ها هسته اصلی نرم‌افزار جستجوی کتابخانه است.

2-6- کنترل ترافیک

در کنترل ترافیک برای تشخیص پلاک خودروها می‌توان از بازشناسی متن استفاده کرد. صحنه ترافیک معمولاً با دوربین‌های سریع تصویربرداری می‌شود و تصاویر معمولاً دو سطحی نیستند. این ویژگی‌ها خواندن پلاک خودروها را پیچیده می‌کند. از محدود بودن حروف استفاده شده در پلاک‌ها می‌توان برای بهبود نتایج استفاده کرد. در شکل 2 رابط گرافیکی یک سیستم بازشناسی پلاک خودرو آمده است.

شکل 2- رابط گرافیکی یک سیستم بازشناسی پلاک خودرو

2-7- سیستم‌های امنیتی

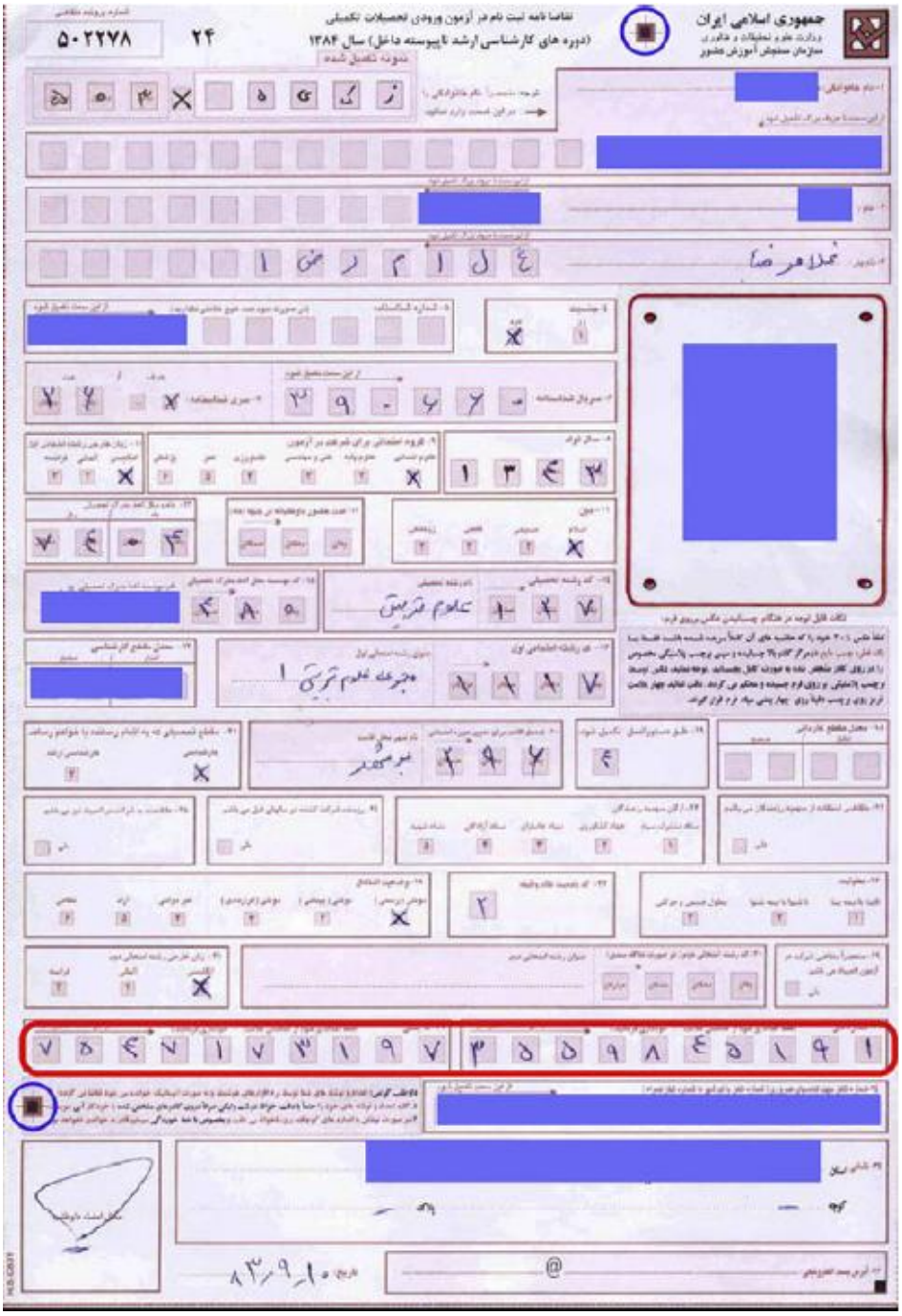
از سیستم‌های بازشناسی متن در سیستم بانکداری برای بازشناسی دست‌نوشته و امضاء به عنوان روشی برای تأیید هویت مشتری استفاده می‌شود.

2-8- تبدیل متن به صحبت

از یک سیستم بازشناسی متن در کنار یک سیستم پردازش صوت می‌توان در کاربردهای مختلف مانند کمک به نابینایان به منظور درک متون چاپی و خودکارسازی پاسخ به ارباب رجوع در ادارات استفاده کرد.

2-9- فرم‌خوان‌ها

این سیستم‌ها برای خواندن فرم‌های خاص طراحی می‌شوند. این سیستم‌ها معمولاً از روبشگرهایی که رنگ خاصی را حذف می‌کنند استفاده می‌کنند. حروف و ارقام بصورت مجزا در خانه‌هایی با رنگ مورد نظر نوشته می‌شوند و روبشگر هنگام روبش فقط این نواحی را اسکن می‌کند و بطور خودکار رنگ زمینه این خانه‌ها را حذف می‌کند. بنابراین این سیستم‌ها نرخ بازشناسی بالایی دارند. در شکل 3 نمونه ای از یک فرم ثبت نام کنکور کارشناسی ارشد ارائه شده است [7].



شکل 3- نمونه‌ای از یک فرم ثبت نام کنکورکارشناسی ارشد

2-10- نقشه‌خوانی خودکار

بازشناسی نویسه‌های روی نقشه‌ها با مسائلی مانند ادغام حروف با خطوط نقشه، چاپ شدن متن با زوایای مختلف، متون با قلم‌های متعدد، متون چاپ شده بر زمینه رنگی و وجود متن دستنویس روبرو است.

2-11- نت‌خوان‌ها

بازشناسی نت‌های موسیقی یکی دیگر از کاربردهای بازشناسی متن است. نت‌های موسیقی به دو صورت چاپی و دستنویس وجود دارند.

3 - عوامل موثر بر کیفیت سیستم نویسه‌خوان

دقت تشخیص در یک سیستم نویسه خوان نوری به شدت به کیفیت تصویر ورودی و میزان نویز تصویر بستگی دارد،عوامل اثر گذار بر کیفیت تصویر عبارتند از:

1) **عمر سند** :یک سند که فاکس شده و یا چند بار کپی شده است مشکل تر از تصویر اصلی خوانده می شود .متن نازک تر یا کلفت تر می شود و ممکن است نویز فلفلی روی آن بیفتد ویا از وضوح آن کاسته شود.
2) **فرآیند چاپ** :یک سند پرینت شده واضح تر از یک سند چاپ شده با ماشین تایپ وآن هم واضح تر از خروجی یک پرینتر ماتریسی است.
3) **وضوح فونت**: فونتهای عجیب و غریب، فونتهای کوچک، حروف کج :(*Italic*) و کلفت، حروف زیر نویس و بالا نویس و استفاده از چند فونت با چند اندازه تشخیص را پیچیده می کند.
4) **کیفیت کاغذ** :کاغذ اپک، صاف و هموار ساده تر از کاغذهای شفاف نازک (مانند روزنامه)خوانده می شود.
5) **شرایط سند** :وجود مهر ,آرم ,نشانه‌ها و عکس در سند باعث سخت شدن تشخیص متن می شود.
6) **کیفیت تصویر**: کیفیت تصاویر متون با وجود انحراف (Skew) کثیف بودن شیشهٔ پویشگر، دقت ، :پویشگر و ... ارتباط مستقیم دارد.

4 - تاریخچه سیستم‌های نویسه‌خوان نوری

4-1- تاریخچه نویسه‌خوان نوری لاتین

دردهه 1950، انقلاب تکنولوژی با سرعت بالایی رو به جلو حرکت می کرد و پردازش الکترونیکی داده‌ها، در حال تبدیل به یک زمینه مهم بود. ورود داده‌ها در این زمان با استفاده از کارت پانچ‌ها صورت می گرفت و یک روش کارا و کم هزینه برای کنترل افزایش داده‌ها مورد نیاز بود. به طورهمزمان، تکنولوژی خواندن ماشینی برای کاربردها در حال بلوغ بود و در میانه‌های دهه 1950 ماشین‌های نویسه‌خوان نوری به صورت تجاری فراهم شدند. اقدامات اولیه در زمینه بازشناسی حروف، بر متون چاپی با مجموعه کوچکی از حروف و نمادهای دست نویس که براحتی قابل تشخیص بودند، متمرکز گردیده بود. سیستم‌های بازشناسی حروف چاپی که در این مقطع زمانی عرضه شدند، عمدتا از روش تطابق قالبی استفاده می نمودند که در آن، تصویر ورودی با مجموعه بزرگی از تصاویر حروف، مورد مقایسه قرار می گرفت. در مورد متون دست نویس نیز الگوریتم‌های پردازش تصویر که ویژگی‌های سطح پایین (ویژگی‌هایی که مستقیما و بدون اعمال هیچ تبدیلی از تصاویر استخراج می شوند) را از تصاویر استخراج می کنند، در مورد تصاویر دو سطحی اعمال می شدند تا بردارهای ویژگی استخراج گردند. سپس این بردارهای ویژگی به طبقه بندی کننده‌های آماری سپرده می شدند.

دراین دوره، تحقیقات موفق اما محدود (منظوراز محدود، مفروض دانستن شرایط و پیش فرض‌های خاص برای کاراکترهای ورودی است)، بیشتر بر روی حروف و اعداد لاتین انجام گرفت. با این حال مطالعات چندی بر روی حروف ژاپنی، چینی، عبری، هندی، سیریلیکی، یونانی و عربی در هردو زمینه حروف چاپی و دستنوشت آغاز گردید. با ظهور صفحات رقومی کننده در دهه 1950 که قادر به تشخیص مختصات حرکتی نوک یک قلم مخصوص بودند، سیستم‌های نویسه‌خوان نوری تجاری نیز امکان عرضه یافتند.

اولین سیستم واقعی نویسه‌خوان نوری ماشین خواندنی بود که در سال 1954 در Reader ' s Digest نصب گردید. این تجهیزات برای تبدیل فاکتورهای فروش تایپ شده به کارت پانچ‌ها به عنوان ورودی کامپیوتر به کارمی رفت.

4-1-1- نویسه‌خوان نوری - نسل نخست

سیستم‌های نویسه‌خوان نوری تجاری که در دوره زمانی بین 1960 تا 1965 پدید آمدند را شاید بتوان نسل نخست نویسه‌خوان نوری نامید. ویژگی ماشین‌های این نسل نویسه‌خوان نوری توانایی خواندن اشکال حروف مشخص است. نمادها برای خواندن ماشین به طور ویژه طراحی می شدند و حتی موارد ابتدایی خیلی طبیعی هم به نظر نمی رسیدند. بعد از گذشت زمانی، ماشین‌های چند فونتی عرضه شدند که قابلیت خواندن ده نوع فونت مختلف را داشتند. تعداد فونت‌ها با محدودیت‌های حاصل از متد تشخیص الگوی به کار رفته مشخص می گردیدند. متد تشخیص الگوی مورد استفاده نیز تطابق قالبی بود که تصویر حرف را با کتابخانه ای از تصاویر طرح اصلی برای هر حرف با هر فونت، مقایسه می کرد..

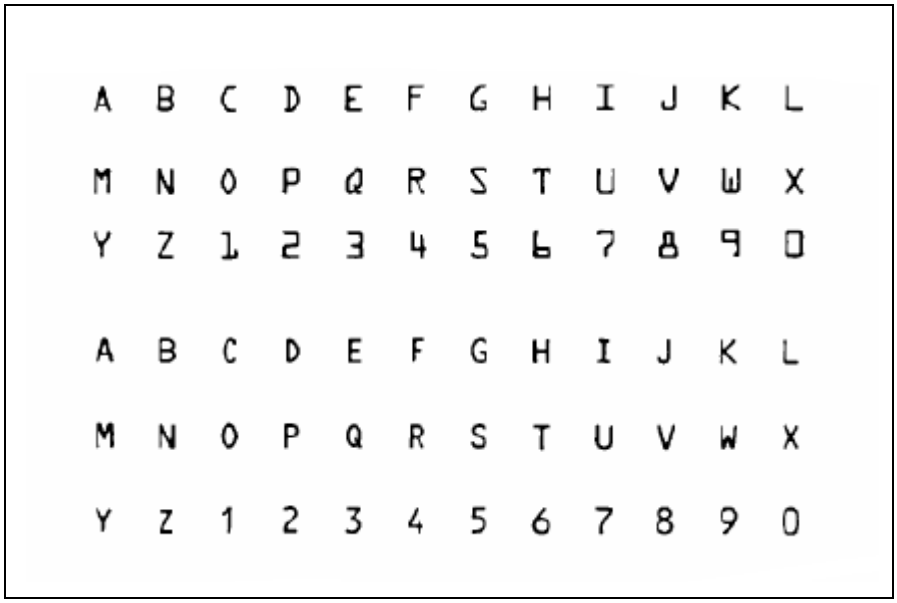
4-1-2- نویسه‌خوان نوری- نسل دوم

ماشین‌های خواندن نسل دوم در میانه‌های دهه 1960 و اوایل دهه 1970 پدیدار شدند. این سیستم‌ها توانایی شناسایی حروف چاپ شده توسط ماشین معمولی را داشتند. همچنین این ماشین‌ها قابلیت شناسایی حروف دست نویس خاص را داشتند. البته این حروف دست نویس به ارقام و برخی نمادهای خاص محدود می شدند.

نخستین و معروف ترین سیستم از این نوع، IBM 1287 بود که در نمایشگاه جهانی نیویورک در سال 1965 به نمایش گزارده شد. دراین دوره همچنین شرکت Toshiba نیز نخستین ماشین مرتب سازی خودکار حروف را برای اعداد کدهای پستی عرضه کرد و شرکت Hitachi نیز نخستین ماشین نویسه‌خوان نوری با کارآیی بالا و قیمت پایین را ساخت.

در این دوره، کار عظیمی در حوزه استاندارد سازی صورت گرفت. در سال 1966 یک مجموعه حروف استاندارد نویسه‌خوان نوری آمریکایی به نام نویسه‌خوان نوریA – تعریف شد. این فونت به شکل دقیقی قالب بندی شده بود و برای امکان پذیر ساختن بازشناسی نوری طراحی شده بود. همچنین یک فونت اروپایی به نام نویسه‌خوان نوریB – نیز طراحی شد که طبیعی تر از نمونه آمریکایی بود. تلاش‌هایی برای ادغام هر دو استاندارد صورت گرفت، اما به طور همزمان ماشین‌هایی طراحی شد که قابلیت خواندن هر دو نوع فونت را داشت.

جدول 1- نویسه‌خوان نوری A – (بالایی) و نویسه‌خوان نوری B – (پایینی)



4-1-3- نویسه‌خوان نوری نسل سوم

برای سیستم‌های نویسه‌خوان نوری نسل سوم، که در میانه‌های دهه 1970 ظاهر شدند، چالش‌های اصلی مستنداتی با کیفیت پایین و حجم بالا و مجموعه حروف دست نویس بودند. هزینه پایین و کارآیی بالا نیز هدف مهمی بود که پیشرفت‌های وسیع در سخت افزار به کمک آن می آمد. در دوره پیش از آنکه کامپیوترهای شخصی و چاپگرهای لیزری عرضه گسترده شوند، تایپ کردن یک زمینه سهولت برای نویسه‌خوان نوری بود. فضای چاپ منحصربه فرد و تعداد کم فونت‌ها، ماشین‌های ساده طراحی شده نویسه‌خوان نوری را خیلی مفید می ساخت.

چک نویس‌های خام به عنوان ورودی به ابزار نویسه‌خوان نوری داده می شد تا ویرایش نهایی روی آن انجام گیرد.

4-1-4- نویسه‌خوان نوری امروزی

اگرچه ماشین‌های نویسه‌خوان نوری به طور تجاری از دهه 1950 در دسترس قرارگرفتند، تنها چند هزار سیستم تا سال 1986 به فروش رفته بود. علت اصلی این امر قیمت بالای سیستم‌ها بود. اگر چه، با ارزان تر شدن سخت افزار و عرضه شدن سیستم‌های نویسه‌خوان نوری به صورت بسته‌های نرم افزاری، فروش به میزان بالایی افزایش یافت. در حال حاضر، تعداد سیستم‌هایی که هر هفته به فروش میرسند به چند هزار می رسد [8].

در اوایل دهه 90، روش‌های پردازش تصویر و بازشناسی الگو با تکنیک‌های کارآمد هوش مصنوعی ادغام گشتند. محققان، الگوریتم‌های پیچیده ای را در بازشناسی حروف ابداع نمودند که قادر بودند داده‌های ورودی با تفکیک پذیری بالا را دریافت کنند و در مرحله پیاده سازی، محاسبات بسیار زیادی را بر روی داده‌ها انجام دهند. امروزه علاوه بر وجود رایانه‌های قدرتمندتر وتجهیزات الکترونیکی دقیقتر مانند اسکنرها، دوربین‌ها و صفحات رقومی کننده، استفاده از تکنیک‌های پردازشی مدرن و توانمند همچون شبکه‌های عصبی، مدلهای مارکوف پنهان، منطق‌های مجموعه فازی و مدل‌های پردازش زبان طبیعی امکان پذیر گشته است [3].

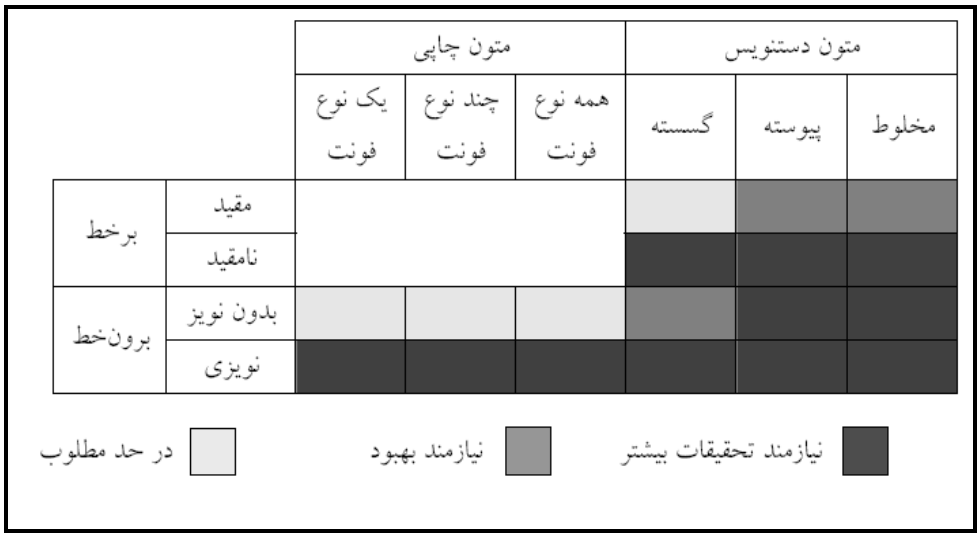
جدول 2 تاریخچه مختصر سیستم‌های نویسه‌خوان نوری و جدول 3 وضعیت کنونی تحقیقات در زمینه سیستم‌های نویسه‌خوان نوری لاتین را نشان می دهند.

جدول 2- تاریخچه مختصر نویسه‌خوان نوری

1870	تلاش‌های نخستین
1940	نسخه‌های نوین نویسه‌خوان نوری
1950	ظهور نخستین ماشین‌های نویسه‌خوان نوری
1960 -1965	نویسه‌خوان نوری نسل نخست
1965 -1975	نویسه‌خوان نوری نسل دوم

1975 -1985	نویسه‌خوان نوری نسل سوم

جدول 3- وضعیت کنونی تحقیقات در زمینه سیستم‌های نویسه‌خوان نوری لاتین



4-2- تاریخچه نویسه‌خوان عربی

حدود 350 میلیون نفر در سراسر جهان به زبان‌هایی صحبت می‌کنند که برای نوشتن آن‌ها از نویسه‌های عربی یا شبیه به آن استفاده می‌کنند. از زبان‌های غیر عربی می‌توان به فارسی، دری، پشتو، اردو، کردی و جاوی اشاره کرد.

سابقه اولین کار منتشرشده درباره نویسه‌خوان نوری عربی به پایان‌نامه کارشناسی ارشد نظیف در دانشگاه قاهره بر می‌گردد که در سال 1975 درباره بازشناسی نویسه‌های چاپی انجام داده است. پس از آن باید به کارهای بدیع و شیمورا درباره حروف دستنویس در سال‌های 1980 [9] و 1982 [10] و نوح و دیگران و امین و دیگران در سال 80 اشاره کرد [11]. پرهامی‌و ترقی در سال 1981 مقاله‌ای درباره بازشناسی متون چاپی منتشر کردند که برای نخستین بار در عنوان آن واژه فارسی دیده می‌شد[12].

به دلیل ماهیت پیوسته‌نویسی در خط عربی، برخلاف آنچه در تاریخچه نویسه‌خوان نوری برای زبان‌های لاتین ذکر شد، روند روشنی مثلاً از بازشناسی حروف و ارقام به بازشناسی کلمات در پژوهش‌های نویسه‌خوان نوری عربی دیده نمی‌شود. به همین دلیل سهم کارهای انجام شده درباره دستنوشته، به ویژه ارقام و حروف، بسیار بیشتر از کارهایی است که درباره نوشتار چاپی انجام شده است.

در نیمه اول دهه 80 میلادی، کارهای انگشت شماری درباره بازشناسی کلمات دستنویس، حروف و ارقام دستنویس، نویسه‌های چاپی و نویسه‌های برخط انجام شد. در نیمه دوم این دهه، پژوهش‌های نویسه‌خوان نوری عربی رونق بیشتری گرفت. مهمترین زمینه‌ها در این دوره عبارتند از: بازشناسی ارقام، حروف و مجموعه‌های بسیار محدود کلمات دستنویس، ارقام و حروف مجزای چاپی و کلمات چاپی با قلم‌های محدود. چند کار انگشت‌شمار نیز درباره بازشناسی نویسه‌های برخط گزارش شده است. در اواخر دهه 80، موضوعات بازشناسی متون دو و چند زبانه، بازشناسی متون با قلم‌های چاپی مختلف، جداسازی کلمات چاپی به حروف و زیر- حروف و بازشناسی نوع قلم مطرح شد.

در دهه 90 میلادی، به بازشناسی متون چاپی توجه بیشتری شد. پیش‌پردازش‌هایی مثل نازک‌سازی مورد توجه قرارگرفت و به پس‌پردازش‌هایی برای اصلاح خطاهای بازشناسی پرداخته شد. از طرفی روش‌هایی برای شکستن کلمات دستنویس به حروف و زیر- حروف مطرح شد. بازشناسی نویسه‌ها و کلمات دستنویس همچنان از مهم‌ترین موضوعات بود. چند کار نیز درباره تشخیص دستنویس برخط گزارش شد.

از سال 2000 میلادی تا کنون، پژوهش‌هایی در موارد زیر گزارش شده است: بازشناسی کلمات چاپی بدون جداسازی آن‌ها به حروف یا زیر- حروف، تعیین نوار زمینه یا خط کرسی در متون چاپی و دستنویس، بازشناسی متن چاپی بدون توجه به نوع قلم، بازشناسی اسامی‌شهرها و مبالغ مندرج در چک‌های بانکی به حروف، ترکیب طبقه‌بندها در بازشناسی کلمات و نویسه‌ها. البته کارهای مربوط به بازشناسی نویسه‌ها و کلمات دستنویس و شکستن کلمات چاپی و دستنویس به حروف و زیر- حروف و بازشناسی آن‌ها همچنان ادامه دارد.

4-3- تاریخچه نویسه‌خوان نوری فارسی

با توجه به اهمیت طیف وسیع کاربردها و نیاز شدید بازار جهانی تجاری، در سال‌های اخیر تحقیقات قابل ملاحظه ای درکشور در زمینه نویسه‌خوان نوری توسط دانشگاه‌ها، برخی نهادهای دولتی و شرکت‌های خصوصی صورت گرفته است که متاسفانه از آمار دقیق آنها اطلاعی در دست نیست. اما قدر مسلم این که برای نویسه‌خوان نوری متون چاپی تاکنون هیچ نرم افزار کارآمد نویسه‌خوان نوری تجاری که محصول تحقیقات داخل کشور باشد، عرضه نگردیده است. در ادامه به برخی از تلاش‌های صورت گرفته دراین زمینه اشاره می شود:

تعداد نسبتا زیادی پایان نامه (بخصوص در مقاطع کارشناسی ارشد و دکتری) و مقاله در این زمینه منتشر شده اند که نقطه تمرکز بیشتر آنها، ارائه روش‌هایی به منظور قطعه بندی درونی، بازنمایی، و بازشناسی حروف بوده است و سایر بخشها شامل پیش پردازش، قطعه بندی بیرونی و پس پردازش کمتر مورد توجه قرار گرفته اند.

در همین راستا طرح ملی بازشناسی متون چاپی و حجم محدودی ازکلمات دست نویس (کبیر، 1377) به سرپرستی دکتر احسان الله کبیر آغاز گردید که درآن تعدادی از دانشجویان و اساتید دانشگاههای تربیت مدرس و صنعتی امیرکبیر درقالب پایان نامه‌های کارشناسی ارشد و دکتری، به انجام تحقیق پرداختند. دکتر کبیر پروژه‌هایی نیز با عناوین "بازشناسی متون چاپی فارسی" (1371-1369) و "بازشناسی حروف و ارقام فارسی دستنویس" (1374-1372) برای سازمان پژوهش‌های علمی و صنعتی ایران انجام داده است.

نخستین تلاش‌های عملی در مقیاس تجاری وسیع، در حوزه ی ثبت نام داوطلبان آزمون (سازمان ملی پرورش استعدادهای درخشان) درسال 1380 آغاز شد. ثبت نام از روی فرم‌هایی که توسط دانش آموزان تکمیل می شد انجام می گرفت. دانش آموزان شرکت کننده در آزمون – مانند آزمون‌های سراسری – باید نام، نام خانوادگی، نام پدر، نام شهرستان محل تولد و سکونت، نام مدرسه و دین خود را داخل کادرهای مربعی شکل و به صورت حروف مقطع (یعنی هرحرف داخل یک کادر) می نوشتند. وقتی که همه فرم‌ها از طریق پست به سازمان مرکزی برگزارکننده آزمون می رسید، عده زیادی تایپیست آنها را دوباره وارد رایانه می کردند. در واقع همان حرف‌های داخل کادر را دوباره تایپ می کردند تا اطلاعات شناسنامه ای هر دانش آموز به صورت دیجیتالی درآید. این روش هم بسیار زمانبر بود و هم نیاز به تعداد

زیادی تایپیست داشت. احتمال داشت که تایپیست‌ها هم هنگام تایپ اشتباه کنند و با ثبت نادرست یک نام، مشخصات فردی در رایانه مرکزی وارد شود که اصلا متولد نشده است. افزون بر این، هزینه کار نیز بسیار زیاد بود. به علت همین مشکلات، شرکت اندیشه نرم افزار پایا در سال 1380 طرحی به منظور ارائه یک سیستم نویسه‌خوان نوری برای شناسایی حروف فارسی گسسته دست نوشته آغاز نمود.

در این محصول، حروف فارسی گسسته دست نوشته پس از شناسایی، با یکدیگر ترکیب می شوند و کلمات به دست آمده از این طریق، در یک واژه نامه لغات (شامل نام‌ها و نام‌های خانوادگی) مورد جستجو قرار می گیرند که بدین ترتیب خطاهای بازشناسی تا حد زیادی کاهش پیدا می کند. به دلیل گسسته بودن حروف دست نوشته، خوش خط بودن معمول فرم‌ها به واسطه وسواس پرکنندگان آنها (زیرا شرکت کنندگان در آزمون نمی خواهند که فرم آنها قابل خواندن نباشد)، و نیز به دلیل استفاده از واژه نامه لغات، دقت بازشناسی صحیح حروف، رضایت بخش (بالای 90% بازشناسی صحیح) بود [3].

در سال‌های 1381 و 1382 ثبت نام آزمون تیزهوشان به یاری این نرم افزار انجام شد. از طریق این طرح 440 هزار نفر به طور ماشینی ثبت نام شدند. این روش باعث شد که درسال 81 (نمونه اول) 45 درصد در هزینه‌ها و 25 درصد در زمان ثبت نام صرفه جویی شود.

درسال بعد (82) این رقم به 50 درصد رسید. نرم افزاری که دراین آزمون‌ها مورد استفاده قرار گرفت برای هر کدام از موارد صحت بازشناسی متفاوتی داشت و در مجموع کار آن خوب بود. جدول 3 صحت بازشناسی بخش‌های مختلف نرم افزار را نشان می دهد.

نرم افزار فوق مربوط به بازشناسی متون آف لاین دست نویس گسسته می باشد و قادر به تشخیص متون چاپی فارسی نیست. دو شرکت دیگر نیز با حمایت دبیرخانه طرح "تکفا" (توسعه کاربرد فناوری اطلاعات و ارتباطات) مشغول پژوهش و آزمایش بر روی نویسه‌خوان نوری فارسی هستند. یکی از این شرکت‌ها "داده پردازان دوران نوین" نام دارد. مدیریت این شرکت را دکتر "حسام فیلی" برعهده دارد. دکتر فیلی متخصص در رشته هوش مصنوعی، از دانشگاه صنعتی شریف است و شرکت "دوران نوین" را از سال 1381، با هدف کار تخصصی بر روی پروژه‌های هوش مصنوعی تاسیس کرده است. او درباره چگونگی پیوستن شرکتش به این طرح می گوید : " از تیرماه سال 82 با شروع فعالیت طرح (تکفا) وحمایت‌های مالی آنها، این شرکت تصمیم گرفت در زمینه طراحی نویسه‌خوان نوری فارسی پژوهش و فعالیت کند." این پروژه در شرکت "دوران نوین" با همکاری آقای "دکتر ابراهیم مقدم" که او هم از دانشجویان دوره دکتری هوش مصنوعی دانشگاه صنعتی شریف است، انجام می گیرد. اخیرا اعلام گردیده که این شرکت

موفق به ارائه یک نرم افزار نویسه‌خوان نوری برای متون چاپی فارسی با دقت بیش از 90% گردیده است (نشانی در وب DOURANOCR). البته این محصول هنوز به بازار عرضه نشده است[3].

جدول 4- صحت بازشناسی بخش‌های مختلف نرم افزار

موضوع	***صحت بازشناسی***
ارقام	96 درصد
حروف الفبا	91 درصد
نام	95 درصد
نام خانوادگی	79 درصد
محل (نام شهر، منطقه و مدرسه)	99.5 درصد
دین	99.9 درصد

5 - چالش‌های زبان فارسی با رویکرد نویسه خوان نوری

بسیاری از مسائل مطرح در نویسه خوانی نوری را می توان مرتبط با حوزه نوشتاری زبان (و حتی دیگر حوزه‌های زبان) دانست و از این رو نویسه‌خوانی نوری را می توان به عنوان یک مسئله بومی مرتبط با زبان قلمداد نمود .هر چند در کشور ما و برخی کشورهای دیگر بر روی نویسه‌خوان‌های نوری فارسی کارهای ارزشمندی انجام شده است، اما می‌توان گفت که نویسه خوان‌های نوری زبان‌های لاتین، چینی و ژاپنی بسیار بیش از نویسه‌خوان‌های نوری فارسی به ایدهال نزدیک شده‌اند.

آشنایی با قواعد نگارش فارسی در انتخاب روش‌های مناسب برای بازشناسی متون نقش اسا نگارش فارسی، ویژگی‌های منحصربه فردی دارد که آن را کاملا از نگارش لاتین متمایز می سازد. به منظور فعالیت در حوزه نویسه‌خوان نوری فارسی آگاهی از قوانین نگارشی و نحوه چاپ حروف در این زبان امری ضروری است. در اینجا به ویژگی‌های کلی نگارش فارسی اشاره می شود.

1- متون فارسی برخلاف متون لاتین از راست به چپ نوشته می شوند.

2- اندازه طول و عرض مستطیل‌های محیطی حروف بسیار متغیرند [13].

3- در حروف خطوط افقی و عمودی و انواع کمان یافت می گردد [13].

4- هجده حرف دارای نقطه و دوحرف نیز دو قطعه ای می باشند: حروف آ وگ. جدول 5 نشان دهنده ترکیب قرارگیری نقاط و تشابه حروف است [14].

5- درکلمات فارسی برخی ازحروف از یک یا دو طرف به حروف مجاور خود اتصال دارند و برخی نیز به صورت مجزا نوشته میشوند. درنتیجه هرکلمه ممکن است شامل یک یا چند بخش متصل باشد که "زیرکلمه" نامیده می شوند. (شکل 4) چسبیده یا سرهم بودن حروف درنگارش فارسی، بازشناسی متون فارسی را برای سیستم‌های نویسه‌خوان نوری نسبت به لاتین بسیار مشکل تر می سازد. محل اتصال حروف به هم چسبیده روی خط مبناست و حاوی ضخامت اندازه قلم است [13]. از این ویژگی می توان برای ارائه تکنیک‌های قطعه بندی استفاده کرد. افزون براین با توجه به اینکه از ضخامت خط مبنا می

توان برای تشخیص اندازه قلم استفاده کرد، این ویژگی در بخش پیش پردازش مورد استفاده قرارخواهد گرفت.

6-حروف فارسی ممکن است 4 موقعیت مجزا و در نتیجه چهار شکل متفاوت نگارش داشته باشند: حروف ابتدایی، میانی، انتهایی و مجزا. جدول 6 فهرست کامل نمادهای مختلف حروف الفبای فارسی را نشان می دهد [14].

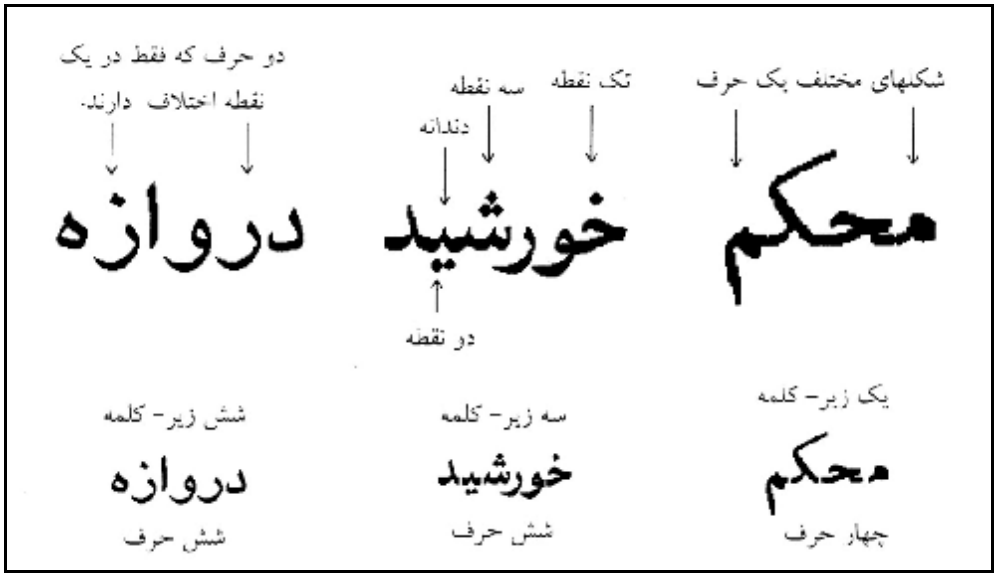
7- حروف واقع در یک کلمه ممکن است همپوشانی داشته باشند، بدین معنا که نتوان با رسم خطوط عمودی، حروف را به طور کامل از یکدیگر مجزا نمود (شکل 4).

8- در برخی از فونت‌ها بعضی از حروف، از یک سمت در دو محل به یکدیگر اتصال دارند مثل "کا"

9- برخی از حروف بین یک تا سه نقطه دارند که ممکن است در بالا یا پایین بدنه حرف واقع باشند همانند "ب"، "پ"، "ت"،"ث ".

جدول 5- ترکیب قرارگیری نقاط و تشابه حروف در زبان فارسی

گره	ترکیب حروف مشابه	گوه	ترکیب حروف مشابه
۱	آ ، ا	۱۰	ف
۲	ب ، پ ، ت ، ث	۱۱	ق
۳	ج ، چ ، ح ، خ	۱۲	ک ، گ
٤	د ، ذ	۱۳	ل
٥	ر ، ز ، ژ	۱٤	م
٦	س ، ش	۱٥	ن
۷	ص ، ض	۱٦	و
۸	ط ، ظ	۱۷	ه
۹	ع ، غ	۱۸	ی



شکل 4- برخی ویژگی‌های نگارش فارسی [15]

10- بعضی از حروف بدنه مشابه دارند و تفاوت آنها تنها در تعداد و محل قرارگیری نقاط یا در وجود یک سرکش است (مانند ک و گ) که در مقایسه با بدنه حروف، اندازه بسیار کوچکی دارند. این موضوع نیز یکی از مواردی است که بر پیچیدگی سیستم‌های نویسه‌خوان نوری فارسی می افزاید.

11- حروف فارسی ممکن است در بالا یا پایین بدنه دارای اعراب باشند. سه اعراب ـَ ـِ ـُ در زبان فارسی، اعراب‌های اصلی اند و اعراب ـً در برخی کلمات عربی رایج در زبان فارسی دیده می شوند (نظیر کلمات عمدا و احتمالا) کلمات عربی دارای اعراب ـٌ و ـٍ در زبان‌های فارسی عمومیت نیافته اند. هرچند کاربرد اعراب در زبان فارسی نسبت به زبان عربی بسیار محدودتر است اما اگرکلمه ای نامتداول باشد یا به دلیل تشابه نگارشی آن با کلمه دیگر، تاکید بر تلفظ صحیح آن باشد از نشانه‌های اعراب استفاده می شود.

12- در بالای بدنه یک حرف ممکن است علامت تشدید وجود داشته باشد.

13- برخی از حروف دارای علامت همزه هستند (ئ أ ؤ ئ).

14- حروفی که از طرف چپ قابلیت اتصال به حرف مجاور خود را دارند ممکن است به صورت کشیده نوشته شوند همانند "با" (شکل 4).

15- بیشتر حروف فارسی (مخصوصا حروف چسبیده) دندانه دار هستند. درمواردی که کیفیت سند اصلی یا دستگاه اسکنر پایین باشد، ارتفاع دندانه‌ها نسبت به خط زمینه کوتاه می شود و این امر، شناسایی صحیح این حرف درمرحله قطعه بندی یا بازشناسی را با مشکل مواجه می سازد.

16- در برخی از شیوه‌های نوشتاری زبان فارسی، دو یا چند حرف کنارهم می‌توانند به گونه‌ای با هم ترکیب شوند که شکل حاصل شباهتی به حروف تشکیل دهنده آن ندارد. به این ترکیب، حروف ادغام شده می‌گویند. چنین مواردی نه تنها در نوشتار دستنویس، بلکه در متون تایپی نیز وجود دارد. متداولترین ترکیب در متون تایپی، ادغام دو حرف (ل) و (ا) بصورت (لا) است. در نوشته‌های دستنویس فارسی نیز بخاطر زیبایی بصری نوشتار و همچنین سلیقه نویسنده، شکل بعضی از حروف کنار هم، بکلی تغییر می‌کند (شکل 5).

17- فاصله گذاری نادرست در متون

18- چسبیدگی‌های غیرمنتظره در برخی قلم‌ها

- چسبیدگی دو حرف
- چسبیدگی نقاط و اعراب به حروف
- چسبیدگی نقاط یا علائم دو حرف مجاور به همدیگر

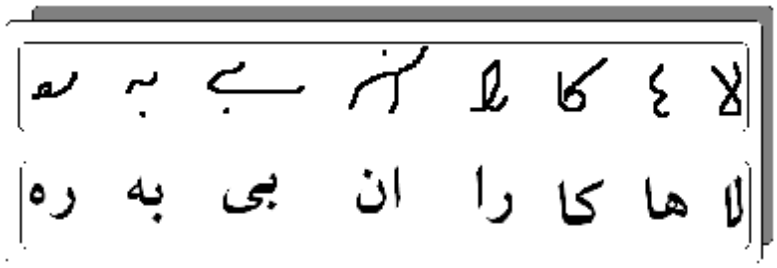
19- اختلاف زیاد ابعاد حروف

جدول 6- شکل‌های مختلف حروف الفبای فارسی با توجه به محل قرار گرفتن آن‌ها در زیر - کلمه

	مجزا	ابتدا	میان	انتها		مجزا	ابتدا	میان	انتها
1	ا یا آ	ا یا آ	---	ـا	17	ص	صـ	ـصـ	ـص
2	ب	بـ	ـبـ	ـب	18	ض	ضـ	ـضـ	ـض
3	پ	پـ	ـپـ	ـپ	19	ط	طـ	ـطـ	ـط
4	ت	تـ	ـتـ	ـت	20	ظ	ظـ	ـظـ	ـظ
5	ث	ثـ	ـثـ	ـث	21	ع	عـ	ـعـ	ـع

6	چ	جـ	ـجـ	ـچ	22	غ	ـغ	ـغـ	ـغ
7	چ	چـ	ـچـ	ـچ	23	ف	ـف	ـفـ	ـف
8	ح	حـ	ـحـ	ـح	24	ق	ـق	ـقـ	ـق
9	خ	خـ	ـخـ	ـخ	25	ک	ـک	ـکـ	ـک
10	د	---	---	ـد	26	گ	ـگ	ـگـ	ـگ
11	ذ	---	---	ـذ	27	ل	ـل	ـلـ	ـل
12	ر	---	---	ـر	28	م	ـم	ـمـ	ـم
13	ز	---	---	ـز	29	ن	ـن	ـنـ	ـن
14	ژ	---	---	ـژ	30	و	---	---	ـو
15	س	سـ	ـسـ	ـس	31	ه	ـه	ـهـ	ـه
16	ش	شـ	ـشـ	ـش	32	ی	ـی	ـیـ	ـی

20- اندازه تمام حروف فارسی یکسان نیستند. مثلاً حروف (ب) و (س) در حالت جدا اندازه بزرگتری نسبت به حروف (د) و (ه) دارند. این تنوع در اندازه حروف، کار قطعه بندی حروف را مشکل می‌کند.



شکل 5- نمونه‌هایی از ادغام حروف مجاور در متون دستنویس

6 - انواع سیستم‌های نویسه‌خوان نوری

سیستم‌های شناسایی متون را می‌توان بر اساس نحوه نگارش (تایپی و دستنوشته)، نوع داده ورودی (برخط و برون‌خط)، متصل بودن متن (کاراکترهای مجزا و کلمات پیوسته)، میزان اِعمال محدودیت در نحوه نگارش و خوانا بودن آن‌ها (مقید یا بدون قید و تک فونتی یا هرفونتی) به گروه‌های مختلفی تقسیم‌بندی کرد.

سیستم‌های نویسه‌خوان نوری را می توان از لحاظ نوع الگوی ورودی به دو گروه اصلی تقسیم کرد [16]:

1) سیستم‌های آن لاین.
2) سیستم‌های آف لاین.

در بازشناسی آن لاین، حروف درهمان زمان نگارش توسط سیستم تشخیص داده می شوند و دستگاه ورودی این سیستم‌ها یک قلم نوری است. دراین روش علاوه بر اطلاعات مربوط به موقعیت قلم، اطلاعات زمانی مربوط به مسیر قلم نیز دراختیار است. این اطلاعات معمولا توسط یک صفحه رقومی کننده اخذ می شوند. در این روش می توان از اطلاعات زمانی سرعت، شتاب، فشار و زمان برداشتن و گذاشتن قلم روی صفحه در بازشناسی استفاده کرد.

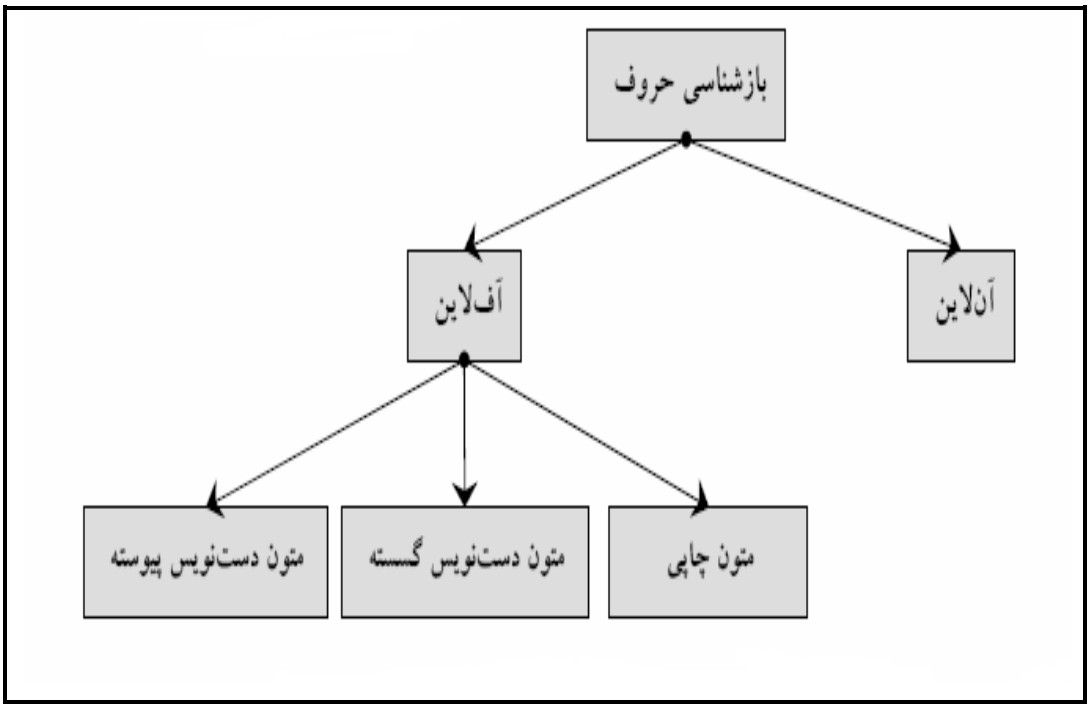
در بازشناسی آف لاین، از تصویر دوبعدی متن ورودی استفاده می شود. در این روش به هیچ نوع وسیله نگارش خاصی نیاز نیست و تفسیر داده‌ها مستقل از فرآیند تولید آنها و تنها بر اساس تصویر متن صورت می گیرد. این روش به نحوه بازشناسی توسط انسان شباهت بیشتری دارد.

بازشناسی آف لاین نیز ممکن است بر روی متون زیر اعمال شود:

1. متن چاپی
2. متون دست نویس محدود
3. متون دست نویس عمومی

متون چاپی متونی هستند که حروف آن شکل و قالب ثابت و از پیش تعیین شده ای دارند. در متون دست نویس گسسته (مانند فرم ثبت نام کنکور) حروف دست نویس اما محدود به اعداد یا حروف بزرگ

هستند. متون دست نویس شامل کلیه متن‌های نگارش یافته با دست هستند که حروف در آنها از قالب و شکل ثابتی پیروی نمی کنند. شکل 6 انواع سیستم‌های ذکر شده را نشان می دهد[8].

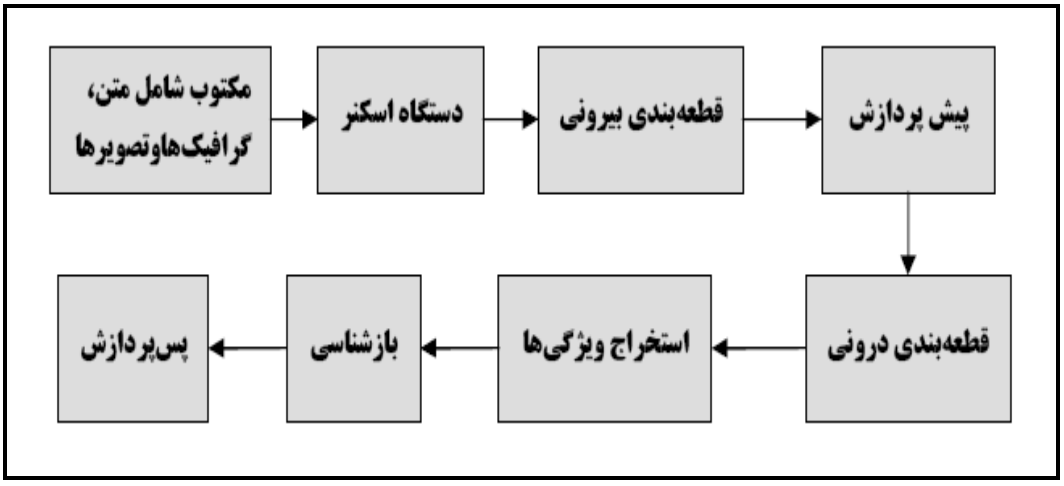


شکل 6- انواع سیستم‌های نویسه‌خوان نوری ازلحاظ نوع ورودی

7 - اجزای اصلی نویسه‌خوان نوری با رویکرد فارسی

یک سیستم نویسه‌خوان نوری معمولی از اجزای مختلفی تشکیل می شود. در شکل 7 مدل معمول یک سیستم نویسه‌خوان نوری مشاهده می گردد [8]. در ابتدا، مستند شامل متن تایپ شده، گرافیکها و تصویرها توسط دستگاه اسکنر ورودی به صورت نوری اسکن شده و به صورت یک فایل تصویری وارد حافظه کامپیوتر می شود. از اینجاست که کار واقعی یک سیستم نویسه‌خوان نوری آغاز میگردد. اولین مرحله، قطعه بندی بیرونی یا مکان یابی است که طی آن نواحی حاوی متن استخراج می گردند. مرحله بعد، مرحله پیش پردازش است که در این مرحله تکنیک‌های مختلفی برای بهینه کردن تصویر موجود از لحاظ کیفیت، فضا و غیره به کار گرفته می شود.

مرحله بعدی که تقریبا مختص سیستم نویسه‌خوان نوری فارسی است، قطعه بندی درونی است که در آن با تکنیک‌های خاص، حروف به هم چسبیده از هم جدا می شوند. استخراج ویژگی‌ها مرحله بعدی است. پس از استخراج ویژگی‌ها در مرحله دسته بندی و شناسایی، موجودیت هر نماد بدست آمده از مرحله قطعه بندی با مقایسه ویژگی‌های استخراج شده کلاس‌های نمادها که درمرحله آموزش در سیستم تعبیه شده اند، شناسایی می گردد. در نهایت با استفاده از اطلاعات متنی، کلمات و اعداد، متن اصلی بازسازی می گردد. در بخش‌های بعدی به بررسی تک تک مراحل یاد شده می پردازیم.



شکل 7- بخش‌های مختلف یک سیستم نویسه‌خوان نوری

7-1- قطعه بندی بیرونی

این مرحله، مکان یابی نیز نامیده می شود [8]. مکان یابی فرآیندی است که طی آن اجزای اصلی سازنده یک تصویر معین می گردد. لازم است که نواحی مختلف یک سند که داده‌ها در آن چاپ شده‌اند شناسایی شده و ازشکل‌ها و عکس‌ها تمایز یابند. برای مثال هنگامی که مرتب سازی خودکار نامه‌ها صورت می گیرد، مکان آدرس باید مشخص گردد و از دیگر چاپ‌های موجود روی پاکت نامه مانند لوگوها، تفکیک شود. چنانچه در مبحث آنالیز اسناد نیز اشاره گردید، اسناد معمولا همراه متن تایپی، نمودارها، جدول‌ها و تصاویری را دارند که در بازشناسی نوری حروف باید آنها را تشخیص داد. یک سیستم جامع نویسه‌خوان نوری باید این قابلیت را داشته باشد که بتواند به طورخودکار پیکربندی صفحات را تحلیل کرده و نواحی متن را مشخص گرداند.

تحلیل پیکربندی صفحات درسه مرحله انجام می گیرد: مرحله اول تفکیک نواحی متنی در تصویر از نواحی غیرمتنی (شامل گرافیک و خطوط) است. مرحله دوم تحلیل ساختاری است که با قطعه بندی نواحی متنی به بلوک‌هایی ازتصویرسند (ظیرپاراگراف، ردیف، کلمه و...) مرتبط می باشد.

مرحله سوم که تحلیل عملکردی نام دارد، با استفاده از اطلاعات مکانی، اندازه و قوانین مختلف صفحه بندی، عملکرد هریک از اجزای سند (نظیرعنوان، چکیده و ...) را تعیین می نماید [17].

این بخش از سیستم اگرچه نقش مهمی را ایفا می کند، به تکنیک‌هایی نیازمند است که خیلی در چارچوب بحث بازشناسی نوری حروف نیست و ازاین رو در اینجا تکیه زیادی روی آن صورت نمی گیرد.

7-2- پیش پردازش

تصویراسکن شده توسط اسکنر معمولی ممکن است شامل مقادیری نویز باشد. افزون براین حروف ممکن است شکل خاصی گرفته باشند که به صورت عادی قابل تشخیص نیستند [8].

این قسمت شامل کلیه اعمالی است که روی تصویر صورت می گیرد تا موجب تسهیل در روند اجرای فازهای بعدی گردد. از مجموعه این پردازش‌ها، هدف‌های زیر دنبال می شود. [3].

1. کاهش نویز و افزایش کیفیت تصویر

2. نرمالیزه کردن داده‌ها

3. فشرده سازی میزان اطلاعاتی که می بایست محفوظ بماند

4. بازشناسی خط، زبان و فونت

در ادامه به بررسی مهم ترین تکنیک‌های موجود برای برآورده کردن اهداف فوق می پردازیم.

7-2-1- کاهش نویز و افزایش کیفیت تصویر

نویز ایجاد شده بواسطه دستگاه‌های اسکنر نوری منجر به ایجاد نقطه نقطه‌های لک مانند، قطعه خطهای گسسته، اتصال بین خطوط، فضاهای خالی درخطوط متن، پرشدن حفره‌های موجود در تصویر برخی حروف و ... می گردد. همچنین اعوجاج‌های مختلف شامل تغییرات محلی، منحنی شدن گوشه‌های حروف، تغییر شکل یا خوردگی حروف را نیز باید درنظرداشت.

قبل ازمرحله بازشناسی حروف، لازم است که این نقایص برطرف شوند. مهم ترین دلیل برای کاهش نویز، کم کردن خطا درمراحل قطعه بندی و بخصوص بازشناسی می باشد. کاهش نویز همچنین سبب کوچک‌تر شدن اندازه فایل تصویر می شود که به نوبه خود، کاهش زمان مورد نیاز برای پردازش و ذخیره سازی آینده را درپی خواهد داشت [17].

برای کاهش نویز و بالا بردن کیفیت تصویر، روش‌های متعددی وجود دارد. این روش‌ها به دو دسته کلی تقسیم می شوند : 1) درفضای نقاط 2) درفضای فرکانس. نتایج برخی پژوهش‌ها نشان داده است که بهترین روش در بین روش‌های موجود، فیلترهای نرم کننده در فضای نقاط بوده است. [13]. در ادامه به بررسی یک نمونه ساده از این فیلتر می پردازیم.

در این روش از یک ماسک n*n استفاده می شود که تمامی ضرایب آن معمولا یک هستند. N میتواند 3،4، 5 و یا بیشتر باشد. هرچقدر n بیشتر باشد، تصویر پخش‌تر می شود. در واقع این ماسک سبب می شود که متوسط هر نقطه با نقاط همسایه آن حساب شده و به عنوان مقدار نقطه جدید در نظرگرفته شود. فرض کنید نقطه Z5 از تصویر با همسایه‌های Z1 تا Z9 و ماسک W مورد نظر باشد:

Z1	Z2	Z3
Z4	Z5	Z6
Z7	Z8	Z9

W1	W2	W3
W4	W5	W6
W7	W8	W9

آن گاه نتیجه اعمال این ماسک برروی نقطه Z5 برابر خواهد بود با :

$$R = W1Z1+W2Z2+...+W9Z9$$

درمورد یک ماسک 3 در 3 برای فیلتر پایین گذرW برابر خواهد بود با :

1/9	1/9	1/9
1/9	1/9	1/9
1/9	1/9	1/9

درعمل یک ماسک 3 در 3 برای فیلترپایین گذر، جواب قابل قبولی به دست می دهد. البته فیلترهای نرم کننده دیگری هم مثل فیلتر میانه وجود دارند که جواب مناسبی را برای تصاویر اسکن شده حروف نمی دهند. همچنین فیلترهای تیزکننده (بالاگذر) نیز جواب مناسب را نمی دهند [13].

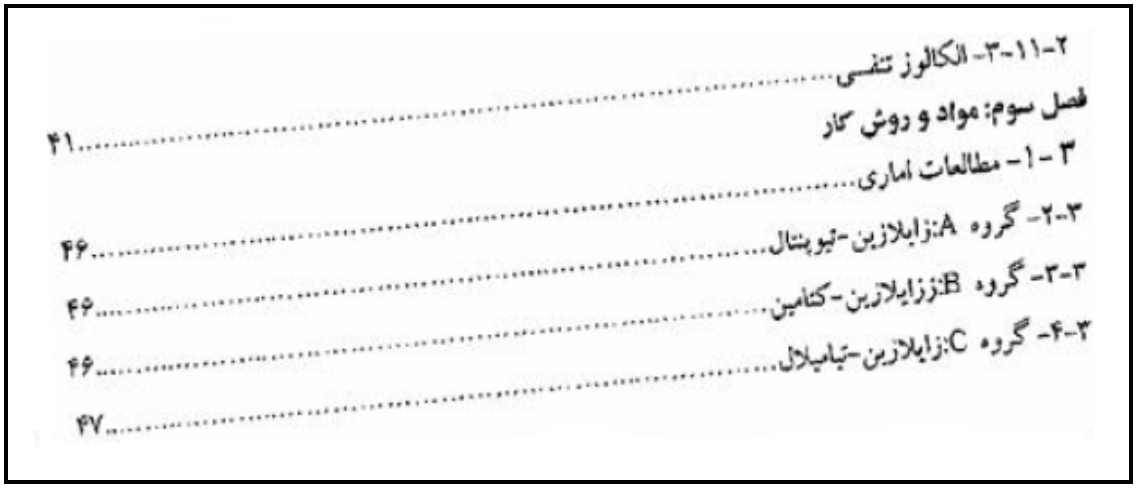
7-2-2- نرمالیزه کردن کجی متن و استخراج خطوط زمینه

به دلیل بی دقتی در مرحله اسکن یا بی دقتی نویسنده درهنگام نگارش متن دست نوشته، ممکن است خطوط متن همانند شکل 8 نسبت به تصویر، اندکی انحراف یا چرخش داشته باشند. این وضع ممکن

است کارآیی الگوریتم‌های به کار رفته درطبقات بعدی سیستم نویسه‌خوان نوری را تحت تاثیر قرار دهد؛ چرا که یکی از مفروضات در بیشتر روش‌های طبقه بندی، کج نبودن تصویر متن ورودی است و درنتیجه لازم است که این نقیصه، آشکار و تصحیح گردد. آشکارسازی خط زمینه در بسیاری از تکنیک‌های قطعه بندی و بازشناسی متون فارسی، عربی و لاتین نقش اساسی دارد. علاوه براین،

برخی از کاراکترها مانند (9) و (g) درنگارش لاتین و (.) نقطه و (0) صفر درنگارش فارسی را بواسطه موقعیت نسبی شان نسبت به خط زمینه می توان آشکار ساخت.

یکی از روش‌های مناسب در بدست آوردن خط زمینه، تراکم نقاط سیاه در یک جمله مورد پردازش می باشد. این تراکم با شمردن تعداد نقاط سیاه در هر سطر (در صورتی که قرارگیری نقاط تشکیل دهنده یک جمله را در یک ماتریس فرض کنیم) بدست می آید.



شکل 8- یک متن فارسی که به صورت کج اسکن شده است.

کلیه الگوریتم‌های توسعه داده شده برای آشکارسازی کجی صفحه، بر روی صفحات متنی با تراز بندی یکنواخت، دقیق عمل می کنند. الگوریتمی کارآتر است که به واسطه حضور مواردی نظیر گرافیک، پاراگراف‌های دارای کجی متفاوت، اعوجاج‌های منحنی، خطِ ظاهرشونده درکتاب‌های فتوکپی شده، نواحی وسیع پیکسل‌های سیاه نزدیک حاشیه صفحه وخطوط متنی مختصر و کوتاه، دقت آن کمتر دستخوش تغییر شود.

روش‌های به کار رفته برای تصحیح کجی خطوط زمینه در متون لاتین را می توان به چهار گروه اصلی دسته بندی کرد که عبارتند از [3]:

1. به کارگیری هیستوگرام (پروفایل تصویرنمایی) تصویر
2. استفاده از روش خوشه بندی نزدیک ترین همسایه‌ها
3. روش همبستگی متقابل بین حروف
4. تبدیل هاف

اغلب پس ازآشکارسازی کجی، تصویر صفحه در جهت اصلی چرخانده می شود تا عملیات تحلیل قالب بندی متن و نویسه‌خوان نوری با سهولت و دقت بیشتری انجام پذیرد. نمونه برداری مجدد مورد نیاز برای این منظور که باید برروی صفحات دوگانگی شده اعمال گردد، ممکن است الگوی کاراکترها را تغییر دهد. دراین حالت به جای چرخاندن تصویر می توان الگوریتم‌های پردازشی را به نحوی اصلاح نمود که اثرچرخش درآنها لحاظ گردد (Jain،1998). همچنین می توان تصویر سند را قبل از دوگانگی کردن، چرخش داد یا این که مقدار چرخش را از روی انتقالهای کوچک و بدون اعوجاج کل بلوک‌های متنی، تقریب زد [18].

7-2-3- نرمالیزه کردن اریب شدگی

در متون چاپی فارسی و لاتین، کاراکترهای دارای قالب ایتالیک از راستای عمود، انحراف دارند. درمتون دست نوشته نیز برخی از نویسنده‌ها حروف را به صورت زاویه دار می نویسند. این پدیده، تحت عنوان اریب شدگی شناخته می شود و ممکن است دقت برخی از الگوریتم‌های قطعه بندی یا بازشناسی را تحت تاثیر قرار دهد و از این رو در این سیستم‌ها لازم است که درمرحله پیش پردازش میزان اریب بودن کاراکترها شناسایی و تصحیح گردد.

اریب شدگی به صورت زاویه شیب بین طویل ترین زیرحرف دریک کلمه و جهت عمودی تعریف می گردد. معمول ترین روش در تخمین میزان اریب شدگی، محاسبه زاویه متوسط اجزای نزدیک به عمود است [19].

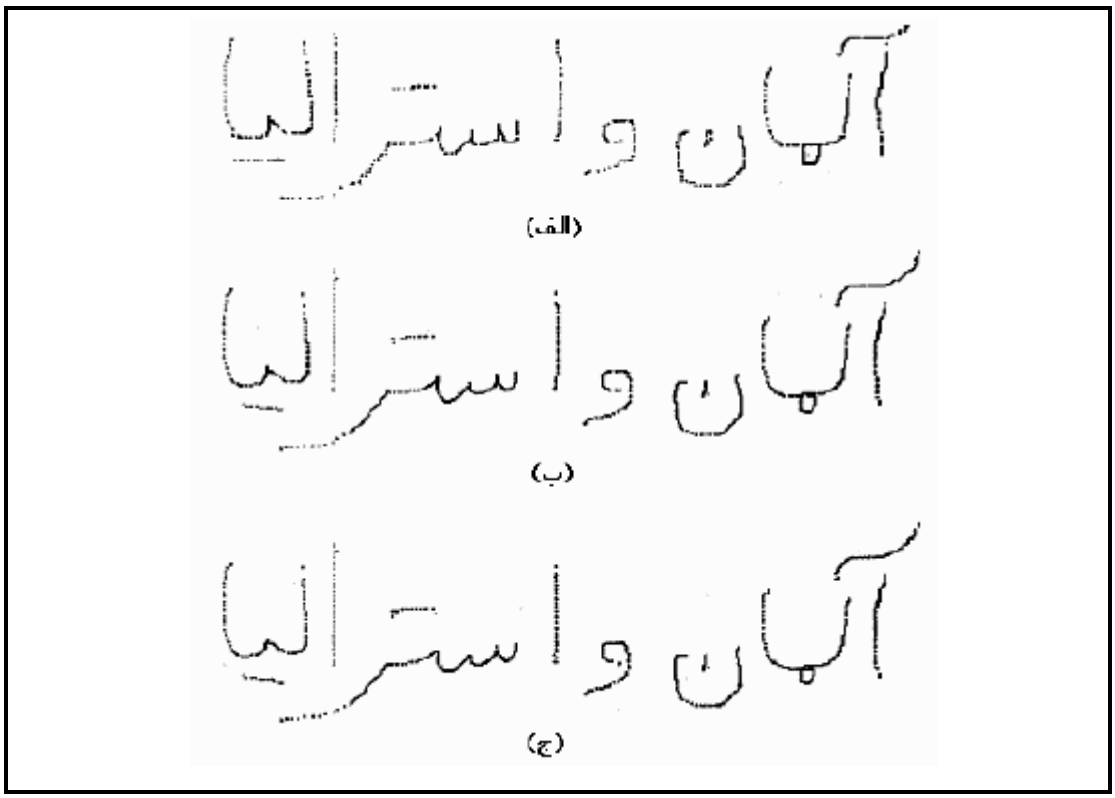
7-2-4- نرمالیزه کردن (تغییر مقیاس دادن) اندازه

درسیستم‌های نویسه‌خوان نوری اغلب تصاویر کلمات یا حروف خیلی کوچک یا خیلی بزرگ، به یک اندازه استاندارد نرمالیزه می شوند تا بدین ترتیب عملیات بازشناسی، مستقل ازاندازه فونت متن گردد. یک روش ساده برای نرمالیزه کردن اندازه‌ها، استفاده از ضخامت خط زمینه است. بدین صورت که پس از پیدا کردن خط زمینه و یافتن ضخامت آن برحسب پیکسل، خط مورد پردازش متناسب با اندازه استاندارد، بزرگ نمایی (یا کوچک نمایی) می گردد.

7-2-5- هموارسازی کانتور

خط تشکیل دهنده مرز یک کاراکتر را کانتور آن کاراکتر گویند. درمتون دست نوشته به واسطه لرزش یا حرکات ناخواسته دست نویسنده در هنگام نگارش، ممکن است که کانتور حروف ناصاف شود. این وضع در سیستم‌های بازشناسی متون چاپی و دست نوشته نیز، به دلیل تغییر مقیاس حروف یا وجود نویز در مرحله اسکن تصاویر ممکن است ظاهر گردد. روش‌های هموارسازی کانتور، به منظور جبران این نقیصه مورد استفاده قرار می گیرند. به طور کلی هموارسازی کانتور، تعداد نقاط نمونه مورد نیاز برای نمایش کاراکتر را کاهش می دهد و درنتیجه کارآیی مراحل پردازشی باقیمانده را بهبود می بخشد [3].

یک روش برای انجام این کارجایگزین کردن مقدارهر پیکسل از تصویرمتن با مقدارمیانگین وزنی پیکسل‌های همسایه آن است که با دو بارتکرار این عمل، تصویر هموارتری از متن دست نویس بدست می آید و در نتیجه اثر لرزش دست نویسنده کاهش پیدا می کند [20]. شکل 9 نمونه‌ای از این مساله را نشان می دهد.



شکل 9- اعمال یک الگوریتم هموارسازی کانتور بر متن.الف. متن اصلی ب. نتیجه اعمال یک مرحله هموارسازی ج. نتیجه اعمال دو مرحله هموارسازی

7-2-6- آستانه گیری و دوگانی (دوسطحی) کردن تصویر متن

تصاویر دیجیتالی به یکی از سه صورت – تصاویررنگی، تصاویرسطح خاکستری (مشابه تصویر یک تلویزیون سیاه و سفید که رنگ تصویر به صورت سیاه، سفید و طیفی از رنگ‌های خاکستری ظاهر می شود)، وتصاویر دوگانی یا دوسطحی (مشابه تصویر یک سند فکس شده که رنگ پیکسل‌های تصویر، تنها سیاه یا سفید است) – می باشند. به منظورکاهش حجم ذخیره سازی مورد نیاز و

افزایش سرعت و سهولت پردازش، اغلب مطلوب است که با انتخاب یک سطح آستانه، تصاویرسطح خاکستری یا رنگی را به تصاویر دوگانی تبدیل نمود [3]. تکنیک آستانه گیری، تصویر را از سطح خاکستری به سیاه و سفید تبدیل می کند. پس از اعمال این تکنیک، هر پیکسل فقط می تواند سیاه (صفر) یا سفید (255) باشد.

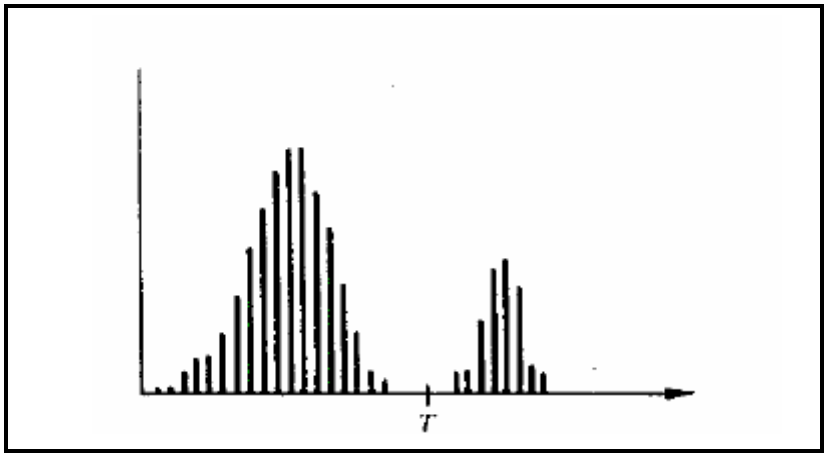
یک شیوه ساده جهت آستانه گیری استفاده ازفرمول زیر است:

$$g(x,y) = 1 \text{ if } f(x,y) > T \;\&\; g(x,y) = 0 \text{ if } f(x,y) <= T$$

معیار T حتی الامکان باید به گونه ای انتخاب شود که هیستوگرام تصویرخاکستری به دو قسمت کاملا مجزا تقسیم شود. جهت انتخاب این آستانه، روش‌های متعددی وجود دارد که آستانه گیری عمومی و آستانه گیری بهینه دو نمونه از این روش‌ها می باشد. در روش آستانه گیری بهینه، مقدارآستانه از فرمول زیر محاسبه می شود:

$$T = (\mu1 + \mu2) / 2 + (\sigma1 * \sigma2) / (\mu1 - \mu2) * \ln(P2/P1)$$

که μ1 و μ2 مقدارمتوسط دو سطح روشنایی، σ انحراف استاندارد از متوسط‌ها و p2 وp1 احتمال وجود دو سطح روشنایی است. این شیوه هر چند بهینه است اما محاسبات σ زمانبر است. شیوه دیگر آستانه گیری عمومی است. در این شیوه هیستوگرام تصویر بدست آمده و از روی هیستوگرام نقطه ای را که مابین دو قسمت اصلی هیستوگرام است به عنوان T در نظرمی گیریم. در عمل تصاویر اسکن شده به نحوی هستند که هیستوگرام به دو قسمت تقریبا مجزا تقسیم می شوند و پیدا کردن T ساده است و نیازی به محاسبات پیچیده نظیر آستانه گیری بهینه نیست. در عمل T=127 برای بسیاری از شرایط معمولی جواب مناسبی را می دهد [13] . شکل 10 طریقه یافتن مقدار T را از روی هیستوگرام نشان می‌دهد.



شکل 10- یافتن مقدار **T** از روی هیستوگرام

7-2-7- نازک سازی

با این عمل، تصویر کاراکترها به تصویری با عرض یک پیکسل تبدیل می شود؛ درست مثل این که کاراکترها با یک قلم نوک باریک نوشته شده باشند. نازک سازی در حالی که کاهش قابل ملاحظه ای درحجم داده‌ها ایجاد می کند، اطلاعات شکلی کاراکتر را نیز حفظ می نماید. دو روش پایه برای نازک سازی عبارتند از : (نازک سازی از طریق پیکسل) و (نازک سازی غیرازطریق پیکسل). نازک سازی از طریق پیکسل به صورت محلی و تکراری تصویر را مورد پردازش قرار می دهد، تا وقتی که از تصویر حروف تنها اسکلت آن به عرض یک پیکسل باقی بماند [3]. در ادامه این بخش یک الگوریتم نازک سازی از طریق پیکسل را مورد بررسی قرارمی دهیم.

فرض می کنیم حروف مقدار 1 و زمینه آنها مقدارصفر داشته باشد (یعنی فرض براین است که قبلا عملیات آستانه گیری صورت گرفته است). این روش با تکرار مرتب دو مرحله اساسی بر روی نقاط مرزی انجام می پذیرد. منظور از نقاط مرزی نقاطی است که مقدارشان یک بوده و حداقل یکی از هشت همسایه آنها صفر باشند.

P9	P2	P3
P8	P1	P4
P7	P6	P5

شکل 11- هشت همسایگی برای نقطه P1

با توجه به هشت همسایگی مطابق شکل 11 در مرحله اول نقطه مرزی P برای پاک شدن علامت گذاری می شود، اگر شرایط زیربرقرار باشد :

a) $2 \le N(P1) \le 6$

b) $S(P1) = 1$

c) $P2.P4.P6 = 0$

d) $P4.P6.P5 = 0$

S(P1) تعداد انتقالات از صفر به یک در حرکت دایره ای از P2 به P3 و بعد به P4، P5 و... تا P9 می باشد. N(P1) تعداد یک‌های حول P1 است. بـه عنـوان مثـال در شـکل 11، N(P1) = 4، S(P1) = 3 می باشد.

0	0	1
1	P1	0
1	0	1

شکل 12- نمونه ای از هشت همسایگی حول نقطه P1. N(P1) = 4 و S(P1) = 3

درمرحله دوم شرایط b،a ثابت می باشد ولی شرایط c،d به صورت زیر درمی آیند :

(c) p2. p4. p8 = 0

(d) p2. p6. p8 = 0

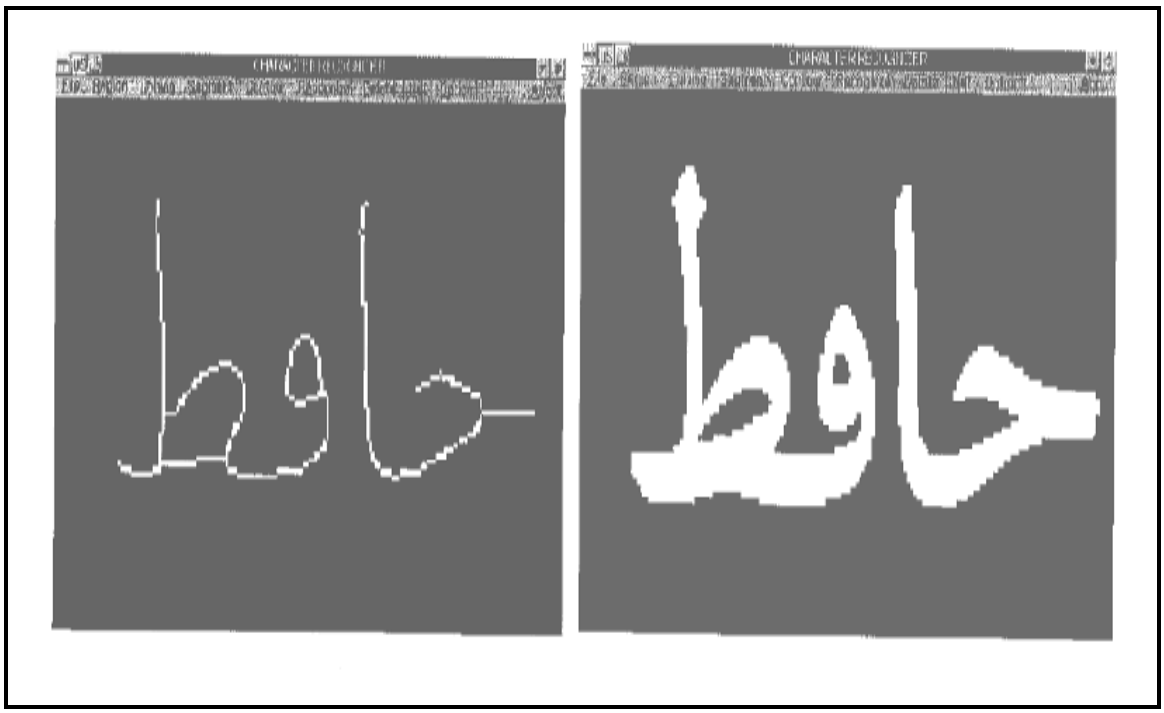
مراحل الگوریتم به صورت زیر خواهد بود :

1) در ابتدا، مرحله اول برای هر پیکسل مرزی انجام شده و نقاطی که شـرایط آن مرحلـه را داشـته باشند برای پاک شدن علامت گذاری می شوند ولی پاک نمی شوند.

2) پس از آنکه کل تصویر برای مرحله اول اسکن شد، نقاطی که علامت گذاری شده بودند پاک می شوند.

3) سپس مرحله دوم انجام شده و نقاطی که شرایط مرحله دوم را داشته باشـند بـرای پـاک شـدن علامت گذاری می شوند ولی پاک نمی شوند.

4) پس از آنکه کل تصویربرای مرحله دوم اسکن شد، نقاطی که علامت گذاری شده بودند پاک مـی شوند. مراحل فوق مرتبا تکرار می شوند و در هر بار تکرار به اندازه یک پیکسل از شکل تراشـیده می شود. این الگوریتم آن قدر ادامه می یابد تا دیگر نقطه ای بـرای پـاک شـدن وجـود نداشـته باشد. دراین حال ضخامت تصویر به اندازه یک پیکسل شده اسـت. سـرعت ایـن الگـوریتم قابـل قبول است و یک صفحه A4 را برروی یک کامپیوتر معمولی 486 در کمتـر از یـک دهـم ثانیـه نازک می سازد. شکل 13 و شکل 14 نمونه ای ازاعمال این الگوریتم را بر روی دو عبـارت نشـان می دهد [13].



شکل 13- نتیجه اعمال الگوریتم برروی یک کلمه دست نویس.

شکل 14- نتیجه اعمال الگوریتم نازک سازی برروی یک عبارت تایپی.

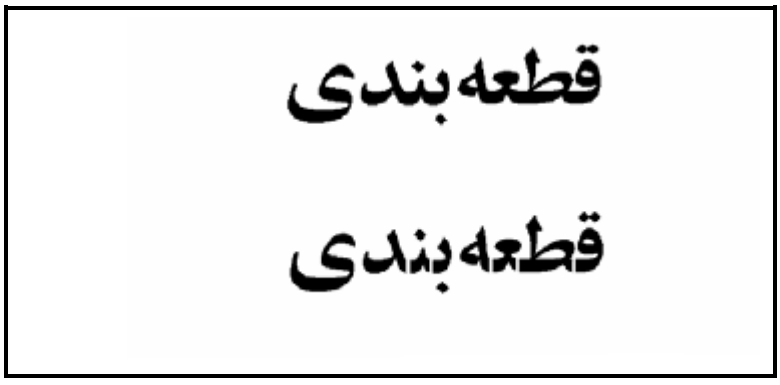
7-2-8- بازشناسی خط، زبان وفونت

بازشناسی خط، تعداد کلاس‌های مختلف نمادهایی را که باید مورد ملاحظه قرارگیرند کاهش می دهد. شناسایی زبان متن، به منظور به کارگیری مدل‌های متنی خاص ضرورت دارد. طبقه بندی فونت‌ها، کاهش تعداد شکل‌های مختلف حروف در هر کلاس که لازم است در فرایند بازشناسی لحاظ گردند را به

دنبال دارد و سبب می شود که امر شناسایی، تنها به یک کلاس فونت محدود گردد. بازشناسی خط و فونت درکاربردهایی مانند نمایه سازی و دستکاری اسناد نیزمطلوب می باشد [19].

7-3- جداسازی و قطعه بندی درونی

یکی از مراحل مهم و اساسی در یک سیستم نویسه‌خوان نوری فارسی، مرحله قطعه بندی درونی است که به آن جداسازی نیز می گویند. اصلی ترین تفاوت سیستم‌های نویسه‌خوان نوری لاتین با سیستم‌های نویسه‌خوان نوری فارسی در همین مرحله نهفته است. خروجی این مرحله، تاثیر مهمی بر سایر مراحل بازشناسی نوری خواهد داشت. در ایران تعداد زیادی مقاله و پروژه کارشناسی ارشد وجود دارد که تنها به این بخش پرداخته اند. برخی از روش‌های پیشنهادی برای بازشناسی حروف فارسی، البته این مرحله را حذف می کنند. در این روش‌ها بازشناسی بدون جداسازی و با استفاده از یک بانک اطلاعاتی کامل از کلمات، صورت می گیرد. در ادامه سه روش مواجهه با نحوه ارتباط جداسازی و بازشناسی را شرح می دهیم. شکل 15 نمونه‌ای از قطعه بندی یک کلمه به حرف را نشان می‌دهد. جداسازی و بازشناسی را به سه روش می توان انجام داد که در ادامه آورده شده‌است.



شکل 15- قطعه بندی یک کلمه به حروف

1) جداسازی همزمان با بازشناسی: این روش معمولا در بازشناسی هم زمان با ورود اطلاعات (مدل آن لاین) به کار می رود و سیستم مغز انسان نیز در خواندن متون از این روش استفاده می کند.

در این روش از شیوه مقایسه پویا استفاده می شود و در زمان بازشناسی، حروف از یکدیگر جدا می شوند. برای این منظور باید تمامی ترکیب‌های مختلف حروف گردآوری گردند و نمادهای مختلف حروف به عنوان مبنا قرارگیرند.

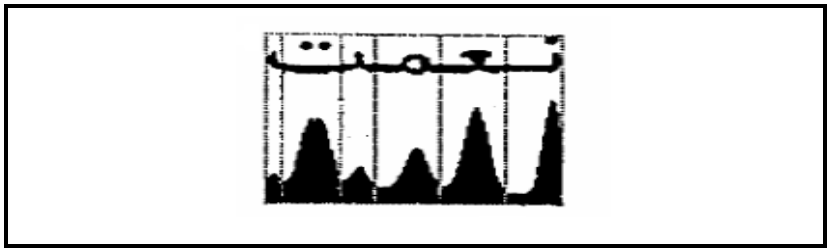
2) بازشناسی بدون جداسازی : در یک حالت از این روش باید مجموعه کاملی از کلمه‌ها وجود داشته باشد که گردآوری این مجموعه کار بسیارمشکلی است. تعداد کلمه‌های ممکن از ترکیب حروف فارسی تا پنج حرف در حدود پانزده میلیون حالت است که البته بسیاری از این ترکیب‌ها بی معنی می باشد و ترکیب‌های واقعی حدود 7500 کلمه و زیرکلمه است. البته همین مقدار هم با حذف نقاط حدود 4800 کلمه است [13]. استفاده از این روش در حالت معمولی بسیار پرهزینه و مشکل است ولی از این روش در بازشناسی حروف دست نویس استفاده شده است. روش‌های دیگری نیز بدون نیاز به فراهم آوری مجموعه کاملی از کلمه‌ها با استفاده از روش‌های دیگر مورد آزمایش و مطالعه قرارگرفته است. به طور کلی این روش چون ممکن است در شناسایی کلمات جدید (مثل اسامی خاص خارجی) دچار مشکل شود، رویکرد کارآمدی نیست.

3) جداسازی کلمه‌ها به زیر حروف و حروف : دراین روش ابتدا با استفاده از تکنیک‌های خاص، حروف از کلمات جدا شده و بازشناسی طی یک مرحله جداگانه بعد از استخراج ویژگی‌ها، صورت می گیرد. این روشی است که در اغلب تحقیقات انجام گرفته پیرامون زبان فارسی از آن استفاده شده است و در این پروژه نیز برخی از معمول ترین تکنیک‌های این روش مورد بررسی قرار می گیرد.

7-3-1- جداسازی با استفاده از مجموع فواصل

در این روش از مجموع فواصل برای جداسازی حروف استفاده می شود. بدین ترتیب که ابتدا موقعیت خط مبنا در کلمه تخمین زده می شود. آن گاه در هر ستون از ماتریس تصویر کلمه، مجموع فواصل پیکسل‌های سیاه کلمه از خط مبنا محاسبه شده و بدین صورت نمودار مجموع فواصل پیکسل‌های سیاه کلمه به دست می آید. پس از هموارسازی این نمونه با یک فیلتر میانگین، نقاط مینیمم موجود در این نمودار به عنوان محل‌های اتصال حروف درنظرگرفته میشوند. با استفاده از این الگوریتم محل‌های

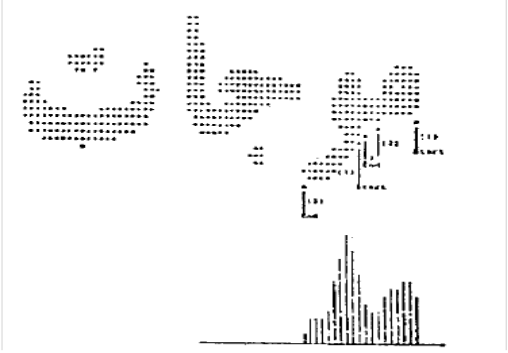
تخمینی جدا کننده حروف تعیین شده و برای جداسازی دقیق کلمه از یک باز شناسی کننده استفاده می شود. شکل 16 نمونه ای از اعمال این الگوریتم را نشان می دهد [14].



شکل 16- استفاده ازالگوریتم مجموع فواصل برای جداسازی حروف

7-3-2- جداسازی با استفاده از نمای افقی

این روش در سال 1994 در دانشگاه Cario برای جداسازی حروف تایپی عربی معرفی شد. دراین روش ابتدا نمای افقی تصویر از راست به چپ برای یافتن نقطه شروع تقطیع یک حرف پویش می شود و نقطه ای که درآن تعداد پیکسل‌های سیاه به میزان مشخصی کمتر از تعداد پیکسل‌های سیاه نقطه قبل باشد، جستجو میشود. پس از یافتن این نقطه، آن را نقطه شروع تقطیع مینامند و جستجو برای یافتن نقطه ای که در آن تعداد پیکسلهای سیاه به میزان مشخصی بیشتر از تعداد پیکسلهای نقطه قبل باشد ادامه می یابد و این نقطه به عنوان نقطه خاتمه تقطیع در نظر گرفته می شود. شکل 17 نمونه ای از به کارگرفتن این الگوریتم را نشان می دهد [13].

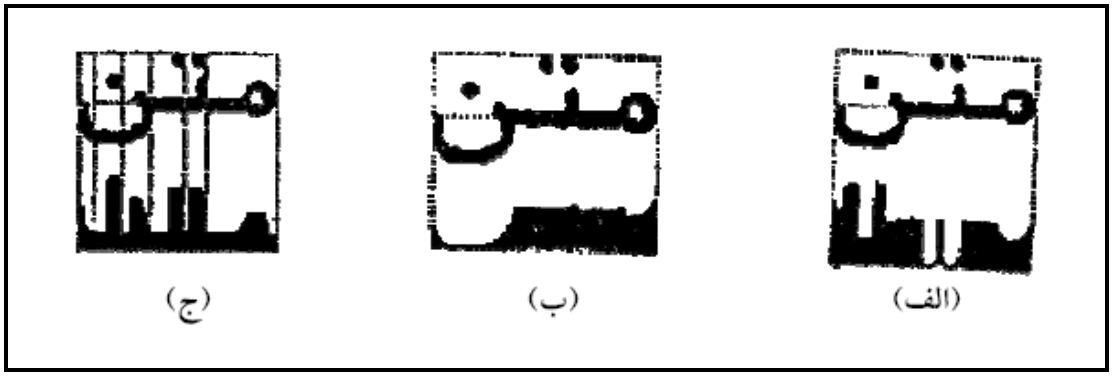


شکل 17- جداسازی با استفاده از نمای افقی

7-3-3- جداسازی با استفاده از نمای ضخامت

در این روش ابتدا محل‌های تخمینی اتصال حروف با به کارگیری یک نمای خاص از کلمه تعیین می شود. سپس با استفاده از نیم رخ‌ها بالا و پایین کلمه محل‌های دقیق اتصال حروف در همسایگی محل‌های تخمینی تعیین می گردد. منظور از نیم رخ بالا، نموداری است که مقدار هر نقطه از آن برابر با فاصله بالاترین نقطه کلمه تا اولین پیکسل سیاه در ستون متناظر با آن نقطه باشد. منظوراز نیم رخ پایین نموداری است که مقدار هر نقطه از آن برابر با فاصله پایین ترین نقطه کلمه تا اولین پیکسل سیاه در ستون متناظربا آن نقطه باشد. نمای ضخامت نموداری است که مقدارهر نقطه از آن از تفاضل ارتفاع کلمه و حاصل جمع نیم رخ‌های بالا و پایین آن در نقاط متناظر حاصل می شود. [13].

این روش مزیت استقلال از موقعیت خط مبنا را دارد ولی در عین حال تعداد زیادی نقطه تقطیع ایجاد می کند که البته می توان با روش‌های خاص آنها را کاهش داد. یک نقطه ضعف دیگر این روش هنگامی است که دو کلمه در یک سطر در کنار هم قرارگرفته و همپوشانی بین حروف مجاور وجود داشته باشد. شکل 18 نیمرخ‌های بالا و پایین و نمودار نمای ضخامت یک کلمه را نشان می دهد



شکل 18- الف. نیم رخ بالا ب. نیم رخ پایین ج. نمای ضخامت

7-3-4- جداسازی با استفاده از نمای قائم

در این روش از نمای قائم جمله استفاده می شود. درنمای قائم p1 مقدار هر نقطه نمودار، از تعداد نقاط سیاه در هر ستون عمودی بدست می آید. این نما، زیر حروف زیادی تولید می کند و از این رو روش مناسبی نیست. روش نمای p2 علاوه بر نقاط سیاه هر ستون، نقاط سفید را نیز در نظر می گیرد.

دراین روش مقدارهر نقطه از نمودار برابر با تعداد کل نقاط (اعم از سیاه و سفید) در یک ستون خاص مربوط به آن نقطه می باشد. این نما، تعداد زیرحروف کمتری ایجاد می کند ولی مثال‌های عملی نشان داده که نسبت به نویز حساسیت بالایی دارد. نمای عمودی p3 از مرز بالایی کلمات به عنوان مقدار نقاط استفاده می کند. ترکیب این سه نما می تواند مورد استفاده قرارگیرد، که دراین صورت تعداد کمتری حروف و زیرحروف ایجاد می گردد. شکل 19 نمونه ای از به کارگیری نماهای مختلف قائم را نشان می دهد [14].

شکل 19- نماهای قائم p1 و p2 برای یک جمله.

7-4- بازنمایی (استخراج ویژگی‌ها)

هدف استخراج ویژگی‌ها بدست آوردن مشخصه‌های اصلی نمادها است و عموما پذیرفته شده که این بخش یکی از دشوارترین مراحل سیستم نویسه‌خوان نوری است. سرراست ترین راه برای توصیف یک حرف، استفاده از تصویر واقعی آن است. این روش، از تکنیک تطابق قالبی [31] استفاده می کند.

رویکرد دیگر، استخراج ویژگی‌های معین که همچنان حروف را مشخص می سازند و رها کردن صفت‌های جزئی است. تکنیک‌های استخراج با توجه به محلی که ویژگی‌ها یافت می شوند به سه گروه اصلی تقسیم می شوند [8].

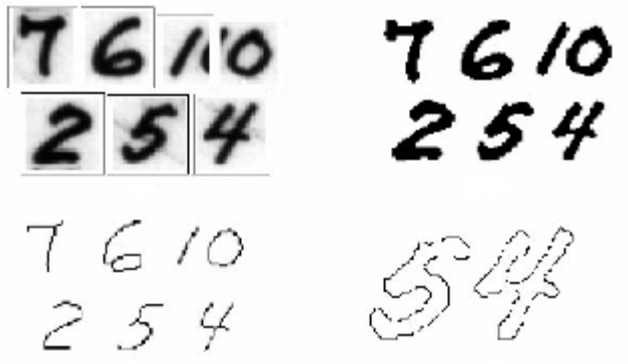
- توزیع نقاط
- تبدیلات و بسط‌های سری
- تحلیل سری

جدول 3-1 تعدادی از روش‌های معمول را با توجه به نوع تصویرمورد بازشناسی نشان می دهد.

جدول 7- روش‌های معمول استخراج ویژگی‌ها برای انواع مختلف تصویر

بردار (اسکلت)	تصویر دوسطحی		زیرتصویرخاکستری
	حروف توپر	کانتور مرزی	
تطابق قالبی	تطابق قالبی	تطابق قالبی	تطابق قالبی
الگوهای تغییرپذیر			الگوهای تغییرپذیر
توصیف گرافی	تبدیلات تکین		تبدیلات تکین
ویژگی‌های گسسته	هیستوگرام برآمدگی		
ناحیه بندی	ناحیه بندی	ناحیه بندی	ناحیه بندی
	ممان‌های هندسی	منحنی‌های اسپلاین	ممان‌های هندسی
توصیف کننده‌های فوریه	ممان‌های زرنیک	توصیف کننده‌های فوریه	ممان‌های زرنیک

شکل 20نمونه ای از چهار نوع تصویر ذکر شده در جدول 7 را نمایش می دهد [3].



شکل 20- الف. تصویر خاکستری قطعه بندی شده ب. تصویردو سطحی شده با حرف توپر ج. تصویر نازک سازی شده اسکلتی د. تصویر دو سطحی کانتور مرزی

درانتخاب بردارهای ویژگی لازم است موارد زیر مورد توجه قرار گیرند:[21] .

1) بردار ویژگی هرالگو باید تا حد امکان ازبردارهای ویژگی دیگر الگوها متمایز باشد (فاصله بین بردارهای ویژگی درفضای ویژگی‌ها، حداکثر باشد).
2) برداروِیژگی الگوها باید تا بیشترین حد ممکن، خصوصیات شکل و ساختارالگوها را از تصویر آنها استخراج نماید.
3) تا حد امکان نسبت به نویز، تغییر اندازه و نوع فونت، چرخش، و دیگر تغییرات احتمالی الگوها دارای ثبات باشد.
4) شرایط، نوع و خصوصیات الگوهای ورودی درانتخاب بردارهای ویژگی اثر می گذارند. به عنوان مثال باید تعیین نمود که آیا حروف یا کلماتی که می بایست تشخیص داده شوند جهت و اندازه مشخصی دارند یا خیر، دست نوشته هستند یا چاپی، یا این که تا چه حد بوسیله نویز، مغشوش شده اند. همچنین گاهی کفایت می کند که سیستم، تنها جوابگوی گروه محدودی از الگوها (مثلا الگوهایی با اندازه یا نوع فونت از پیش مشخص شده) باشد.

5) درمورد حروفی که به چندین شکل نوشته می شوند (مانند 'a' و 'α'، '4' و '4') لازم است بیش از یک کلاس الگو به یک کاراکتر خاص تعلق یابد.

گروه‌های مختلف ویژگی‌ها را می توان با درنظرگرفتن معیارهایی مانند حساسیت نسبت به نویز، تغییر شکل وآسانی پیاده سازی و استفاده، مورد ارزیابی قرارگیرند. نتیجه چنین ارزیابی ای در جدول 8 نمایش داده شده است. معیارهای مورد استفاده عبارتند از [8].

7-4-1- معیارهای شکلی

1) نویز: حساسیت نسبت به قطعه خطوط مقطوع، فاصله‌ها، ضربه‌ها و غیره.

2) تحریف (اعوجاج) [36] : حساسیت نسبت به تغییرهای محلی مانند گوشه‌های تاخورده، چروک شدگی، کشیدگی و غیره.

3) تغییرشکل: حساسیت نسبت به تغییرشکل مانند استفاده از شکل‌های گوناگون برای معرفی حروف یکسان یا اریب شدگی.

4) انتقال: حساسیت نسبت به حرکت تمام حرف و اجزایش.

5) دوران: حساسیت نسبت به تغییرمرکز ثقل حرف.

7-4-2- معیارهای کاربردی

1) سرعت شناسایی

2) پیچیدگی پیاده سازی

3) استقلال: عدم نیازبه تکنیک‌های اضافی

تکنیک استخراج ویژگی	معیارهای کاربردی ۱ ۲ ۳	معیارهای شکلی ۱ ۲ ۳ ۴ ۵
تطابق قالبی	○ ● ○	● ● ○ ○ ○
تبدیلات	○ ○ ●	○ ● ● ● ●
توزیع نقاط:ناحیه‌بندی	● ● ○	○ ● ○ ○ ●
ممان‌ها	○ ● ○	● ● ○ ● ●
چندتایی‌های n	● ● ●	● ○ ● ○ ●
مکان هندسی اساسی	● ● ○	○ ● ● ● ●
تقاطع ها	● ● ○	○ ● ● ● ●
ویژگی‌های ساختاری	● ○ ●	○ ● ● ● ●

پایین یا دشوار ○ متوسط ● بالا یا آسان ●

جدول 8- ارزیابی تکنیک‌های استخراج ویژگی‌ها.

7-4-3- تکنیک‌های تطابق قالبی و همبستگی

این تکنیک‌ها با سایر کنیک‌ها از ین جهت که هیچ ویژگی ای در واقع استخراج نمی گردد، متفاوتند. به جای استخراج ویژگی‌ها، ماتریسی که شامل تصویر حرف ورودی است، مستقیما با مجموعه ای ازحروف قالبی اصلی که کلاس‌های ممکن را معرفی می کنند، مقایسه می شود.

اختلاف بین الگو و هرقالب اصلی محاسبه می گردد و کلاس قالب اصلی ای که بهترین تطابق را عرضه می کند، به الگو تخصیص داده می شود. این تکنیک مرحله بعدی که بازشناسی است را نیز به طور ضمنی انجام می دهد و در نتیجه در صورت استفاده سیستم از این تکنیک، نیازی به

مرحله استخراج ویژگی‌ها و بازشناسی به طور جداگانه نیست. پیاده سازی سخت افزاری این تکنیک ساده و آسان است و در بسیاری از ماشین‌های تجاری نویسه‌خوان نوری مورد استفاده قرارگرفته است.

اگرچه این تکنیک به نویز و تغییر شکل حساس است و هیچ راهی برای مدیریت حروف چرخش یافته ندارد.[8] .

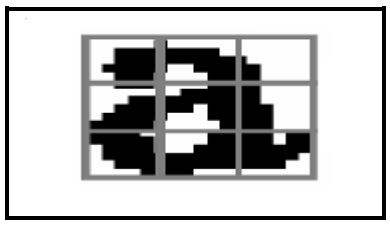
7-4-4- تکنیک‌های براساس ویژگی

دراین روش‌ها، محاسبات و اندازه گیری وسیعی از یک حرف انجام می گیرد و با توصیف‌های کلاس‌های حروف که در فاز آموزش بدست آمده اند، مقایسه می گردد. توصیفی که به نزدیک ترین شکل ممکن، تطابق یابد، بازشناسی می گردد. ویژگی‌ها به صورت اعدادی در بردار ویژگی قرار داده می شوند و این بردار ویژگی برای معرفی نماد به کار برده می شود .[8] .

7-4-4-1- توزیع نقاط

این گروه، تکنیک‌هایی را پوشش می دهد که ویژگی‌ها را بر اساس توزیع آماری نقاط استخراج میکند. این ویژگی‌ها اغلب نسبت به اعوجاج‌ها و تغییرشکل‌ها، مقاوم هستند. بعضی از معمول ترین تکنیک‌ها در ادامه بررسی می شود. .[8] .

1) **ناحیه بندی**: مستطیل محیطی حرف به چندین ناحیه دارای همپوشانی یا غیرهمپوشانی کننده تقسیم می شود و تراکم نقاط سیاه موجود درهر ناحیه محاسبه می گردد و به عنوان ویژگی مورد استفاده قرارمی گیرد. شکل 21 ناحیه بندی را نشان می‌دهد.



شکل 21- ناحیه بندی

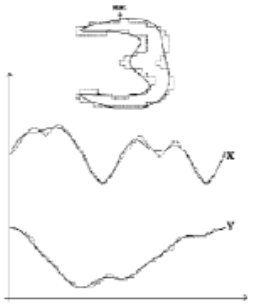
2) **ممان‌ها**: ممان‌های نقاط سیاه اطراف یک مرکز، برای مثال مرکز ثقل، یا یک سیستم هم پایه، به عنوان ویژگی‌ها استخراج می گردند.

3) **تقاطع‌ها و فاصله‌ها**: در تکنیک تقاطع‌ها، ویژگی‌ها از تعداد دفعاتی که شکل یک حرف با بردارهایی در طول جهت‌های معین، قطع می شود، استخراج می شوند. این تکنیک دراغلب سیستم‌های تجاری مورد استفاده قرارمی گیرد، چون سرعت بالا و پیچیدگی کمی دارد. هنگامی که از تکنیک فاصله‌ها استفاده می شود طول‌های معین بردارهای قطع کننده شکل حرف اندازه گیری می شود.

4) **مکان هندسی اساسی**: برای هر نقطه در زمینه یک حرف، بردارهای افقی و عمودی تولید می گردند. تعداد دفعاتی که قطعه خطوط توصیف کننده حروف با این بردارها تلاقی می یابند، به عنوان ویژگی‌ها استخراج می شود.

5)**تبدیلات و بسط‌های سری**: این تکنیک‌ها به کاهش ابعاد بردار ویژگی‌ها کمک می کند و ویژگی‌های استخراج شده می توانند نسبت به تغییرات سراسری مانند دوران و انتقال تغییر ناپذیرباشند.

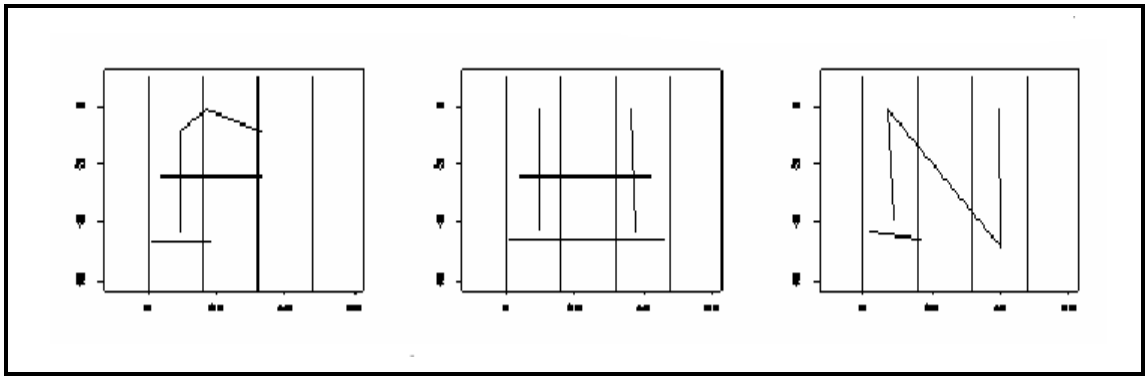
تبدیلات مورد استفاده ممکن است فوریه ،والش ،هار ،هادامارد ، کارهونن – لوئو ،هاف ، انتقال محور اصلی و غیره باشند. بسیاری از این تبدیلات مبتنی بر توصیف منحنی کانتور حروف هستند. این بدان معناست که این ویژگی‌ها نسبت نویزهایی که بر کانتور حروف مانند تاثیر می گذارند مانند فواصل ناخواسته در کانتور، بسیار حساس هستند. از این رو در جدول 5 این متدها به عنوان متدهای حساس نسبت به تاثیرات نویز درون حروف و اعوجاجات معرفی شده اند [8]. شکل 22 توصیف‌گرهای بیضوی فوریه نشان داده شده‌است.



شکل 22- توصیف گرهای بیضوی فوریه

7-4-4-2- تحلیل ساختاری

در تحلیل ساختاری، ویژگی‌هایی که ساختارهای توپولوژیکی و هندسی یک نماد را توصیف می کنند، استخراج می شوند. با این ویژگی‌ها، تلاش می شود که ساختار فیزیکی حروف توصیف شوند. بعضی از ویژگی‌های معمول مورد استفاده، پاره خط‌ها ، نقاط انتهایی و تلاقی بین خطوط و خمیدگی‌ها می باشند. در مقایسه با دیگر تکنیک‌ها، تحلیل ساختاری، ویژگی‌هایی با مقاومت بالا در برابر نویز و تغییر شکل‌ها تولید می کند. اگر چه، این ویژگی‌ها به میزان اندکی در برابر دوران و انتقال مقاوم هستند. متاسفانه استخراج این ویژگی‌ها ساده نیست و هنوز تا حد زیادی محل تحقیق است [8]. شکل 23 نمونه‌ای از تحلیل ساختاری آمده است.



شکل 23- پاره خط‌های حاصل از حروف بزرگ **H**، **N**، **F**.

7-5- بازشناسی و دسته بندی

این مرحله، آخرین مرحله پردازش در یک سیستم نویسه‌خوان نوری است. در این مرحله، هر حرف شناسایی می گردد و به کلاس حرفی مناسب ارجاع داده می شود. در ادامه، دو رویکرد مختلف برای دسته بندی در بازشناسی حروف مورد بحث قرار می گیرد. ابتدا، بازشناسی از طریق تئوری تصمیم گیری

مورد بررسی قرار می گیرد. این متدها زمانی مورد استفاده قرار می گیرند که مشخصه‌های یک کاراکتر بتوانند به صورت عددی توسط بردار ویژگی‌ها معرفی گردند.

ممکن است ما مشخصه‌های الگو داشته باشیم که از ساختار فیزیکی حروف مشتق شده اند که نمی توانند به آسانی به صورت عددی اندازه گیری شوند. در این حالات، ارتباط بین حروف در هنگام تصمیم گیری روی عضویت کلاس باید مورد توجه قرارگیرند. برای مثال، اگر ما بدانیم که یک حرف ازیک پاره خط عمودی و یک پاره خط افقی تشکیل شده است، ممکن است این حرف T یا L باشد. دراین صورت ارتباط بین دو پاره خط نیز برای تشخیص حرف لازم است. در این صورت یک رویکرد ساختاری مورد نیاز خواهد بود [8].

7-5-1- روش‌های تئوری تصمیم گیری

رویکردهای اصلی در بازشناسی تئوری تصمیم گیری عبارتند از : دسته بندی کننده‌های با فاصله مینیمم، دسته بندی کننده‌های آماری و شبکه‌های عصبی. هر یک از این تکنیک‌های دسته بندی در ادامه مورد بررسی مختصر قرار می گیرند.

7-5-1-1- تطابق

تطابق، گروهی از تکنیک‌هاست که مبتنی بر معیارهای تشابه در جایی است که فاصله بین برداررویژگی‌ها – که حروف استخراج شده را توصیف می کنند – و توصیف هر کلاس محاسبه می گردد.

معیارهای متفاوتی ممکن است مورد استفاده قرار گیرد، اما معمول ترین آنها (فاصله اقلیدوسی) است. این دسته بندی کننده با فاصله می نیمم هنگامی که کلاس‌ها به خوبی ازهم جدا شده اند، به خوبی کار می کند و این زمانی است که فاصله بین میانگین‌ها در مقایسه با گسترش هر کلاس بزرگ است.

هنگامی که کل حرف به عنوان ورودی دسته بندی کننده مورد استفاده قرارمی گیرد، و هیچ ویژگی ای استخراج نشده است (تطابق الگویی)، یک رویکرد همبستگی مورد استفاده قرارمی گیرد. دراینجا اختلاف بین تصویر حرف و تصویرقالب اصلی که هر حرف را معرفی می کند، محاسبه می گردد.

7-5-1-2- دسته بندی کننده‌های آماری

دردسته بندی کننده‌های آماری یک رویکرد احتمالاتی برای بازشناسی مورد استفاده قرار میگیرد. ایده این است که از یک طرح دسته بندی که از این نظرکه استفاده از آن کمترین احتمال ایجاد خطاهای دسته بندی را ایجاد می کند، بهینه است، مورد استفاده قرار می گیرد.

یک دسته بندی کننده که میانگین مجموع اتلاف را می نیمم می کند، یک (دسته بندی کننده بایس) [67] نامیده می شود. با یک حرف داده شده ناشناخته که توسط برداروِیژگی‌هایش توصیف شده است، احتمال اینکه نماد به کلاس c تعلق داشته باشد برای تمام کلاس‌های c = 1 … N محاسبه می شود. نماد سپس به کلاسی تخصیص داده می شود که حداکثر احتمال را بدهد.

برای اینکه این طرح بهینه باشد، توابع چگالی احتمال یک نماد از هرکلاس باید شناخته شده باشد، همچنین احتمال رخداد هر کلاس نیز باید معلوم باشد. نیازمندی دوم معمولا با اختصاص احتمال مساوی به تمام کلاس‌ها حل می شود. تابع چگالی نیز معمولا فرض می شود که به صورت نرمال توزیع شده است.

7-5-1-3- شبکه‌های عصبی

اخیرا کاربرد شبکه‌های عصبی برای شناسایی حروف و دیگر انواع الگو، مورد توجه قرارگرفته است. با در نظرگرفتن یک شبکه توزیع پشتیبان، این شبکه از چندین لایه از عناصر به هم متصل، ترکیب شده است. یک بردار ویژگی از لایه ورودی وارد شبکه می شود. هر عنصر لایه یک مجموع وزن دار از ورودی اش را محاسبه می کند و با یک تابع غیر خطی به یک خروجی منتقل می کند.

در طول فاز آموزش، وزن‌ها در هر اتصال تنظیم می شوند تا اینکه یک خروجی مناسب بدست آید. یک مشکل شبکه‌های عصبی در نویسه‌خوان نوری ممکن است محدودیت پیش بینی پذیری و عمومیت دادن آنها باشد، درعین حال یک مزیت آنها طبیعت وفق پذیرآنهاست.

7-5-2- روش‌های ساختاری

در حوزه شناسایی‌های ساختاری، روش‌های نحوی در میان متداول ترین رویکردها هستند. تکنیک‌های دیگری هم وجود دارند اما عمومیت کمتری دارند در اینجا معرفی نمی شوند.

7-5-3- روش‌های نحوی

معیارهای تشابه مبتنی بر ارتباط بین اجزای ساختاری، می تواند با استفاده از مفاهیم نحوی فرموله گردد. ایده این است که هر کلاس، تعریف نحوی خاص خود را دارد که ترکیب حروفش را تعریف می نماید. یک گرامر ممکن است به صورت رشته یا درخت تعریف شود و اجزای ساختاری استخراج شده از یک حرف نامعلوم، با گرامر هر کلاس مطابقت داده می شود. فرض کنید ما دو کلاس حروف متفاوت داریم که می توانند با دو گرامر G2، G1 تولید می شوند. با یک حرف داده شده، ما می گوییم که این حرف به کلاس اول بیشتر شبیه است اگر با گرامر G1 قابل تولید باشد و با G2 قابل تولید نباشد.

7-6- پس پردازش

با بازشناسی تمام تصاویر اجزای متن یک بلوک، متن متناظر با هر یک از آن‌ها مشخص شده است. برای بدست آوردن متن نهایی باید پس‌پردازش‌هایی روی این متن انجام شود. به عنوان مثال از به هم چسباندن حروف بازشناسی شده در مرحله قبل، زیر- کلمات و کلمات تشکیل می‌شوند. تشکیل این زیر-کلمات و کلمات با استفاده از موقعیت آن‌ها در تصویر، یک مرحله پس‌پردازش است. مرحله بعدی می‌تواند سنجش اعتبار این کلمات با استفاده از یک واژه نامه باشد. در ادامه هر یک از کارهایی که در این مرحله نجام می‌شود را توضیح خواهیم داد [8]..

7-6-1- گروه بندی

نتیجه بازشناسی ساده نمادها روی یک سند، مجموعه ای از نمادهای منفرد است. این نمادها در داخل خود اطلاعات کافی را ندارند. به جای این مجموعه منفرد از نمادها، ما باید نمادهایی را که به یک رشته واحد تعلق دارند، ترکیب کنیم و کلمات و اعداد را بسازیم. فرآیند انجام این ترکیب نمادها به رشته‌ها به

عنوان " گروه بندی" شناخته می شود. گروه بندی نمادها در رشته‌ها معمولا براساس اطلاعات مربوط به محل قرارگیری آنها درسند انجام می گیرد. نمادهایی که به اندازه کافی نزدیک به هم باشند با یکدیگر گروه بندی می شوند.

برای فونت‌های با زیر و بم ثابت، گروه بندی کارنسبتا ساده ایست چرا که مکان هر حرف شناخته شده است. برای حروفِ حروف چینی شده، فاصله بین حروف متغیر است. هرچند فاصله بین کلمات به مراتب بزرگتر از فاصله بین حروف است و گروه بندی همچنان امکان پذیر است. مشکل واقعی برای حروف دست نویس یا زمانی که متن اریب شده است رخ خواهد داد.

بازیابی و تصحیح خطا

دراین مرحله با استفاده از اطلاعات جانبی (نظیرمجموعه لغات معتبر، اطلاعات آماری مربوط به رخداد حروف، اطلاعات دستوری و معنایی) سعی در بهبود نتایج حاصل از مرحله بازشناسی می گردد. تاقبل از این مرحله، هیچگونه اطلاعات معناشناختی در طول مراحل پردازشی مورد استفاده قرارنگرفته بود. مرحله پیش پردازش سعی در " تمیز کردن " تصویرسند به نحو مقتضی دارد و به علت عدم دسترسی به اطلاعات مفهومی، ممکن است باعث حذف برخی از اطلاعات مهم تصویر گردد. فقدان اطلاعات معنایی در مرحله قطعه بندی می تواند به نتایج حادتر و غیرقابل برگشتی بینجامد؛ چرا که این مرحله، تعیین کننده مرز الگوهای ورودی می باشد. بنابراین واضح است که درصورت فراهم شدن اطلاعات معناشناختی، دقت نتایج بازشناسی به نحو چشمگیری افزایش می یابد. مروری بر تحقیقات اخیرا انجام شده درزمینه بازشناسی حروف نشان می دهد که در صورت استفاده از اطلاعات شکلی بدون به کارگیری اطلاعات معناشناختی، افزایش دقت قابل توجهی نخواهیم داشت. در نتیجه، یکپارچه کردن اطلاعات معنایی و شکلی در کلیه بلوک‌های سیستم‌های نویسه‌خوان نوری به منظور بهبود نرخ بازشناسی صحیح، ضرورت دارد.

ساده ترین راه برای این منظور، استفاده از یک فرهنگ لغات برای اصلاح خطاهای جزئی است. این کارتوسط یک برنامه واژه پرداز (با قابلیت خطایاب املایی) که املای کلمات را کنترل و چندین پیشنهاد برای اصلاح املای کلمات نامانوس ارائه می کند، انجام می گیرد [21].

8 - کارهای انجام شده

8-1- نرم افزارهای تجاری لاتین

جدول 2 معروف ترین نرم افزارهای تجاری نویسه‌خوان نوری لاتین را نشان می دهد. چنانکه در بخش تاریخچه نویسه‌خوان نوری اشاره شد، سیستم‌های تجاری نویسه‌خوان نوری از سال 1954 وارد بازار گردیدند. با این حال سیستم‌های اولیه اغلب دارای بخش سخت افزاری عمده ای بودند، اما سیستم‌های امروزی نویسه‌خوان نوری کاملا نرم افزاری هستند در جدول 9 مشخصات چند نرم افزار نویسه‌خوان نوری عمده لاتین را ملاحظه می کنیم.

از جمله مهم ترین ویژگی‌های این سه نرم افزار عبارتند از [21]:

1) دقت بازشناسی بالای 99 درصد.
2) پشتیبانی ازبیش از 140 زبان مختلف (فقط نسخه ای از Readiris، زبان عربی را پشتیبانی می کند).
3) پشتیبانی از قالب بندی‌ها مختلف فایلی شامل قالب‌های تصویری، قالب‌های pdf و...
4) تولید خروجی متنی در قالب‌های متنوع.
5) تشخیص خودکارمتن، گرافیک و جدول در تصویر ورودی.
6) قابلیت اضافه نمودن، ویرایش و حذف فریم‌های متنی، گرافیک و تصویری. به این ترتیب می توان درقسمت‌هایی از تصویرکه نرم افزار قادر به تشخیص درست نیست، به نرم افزارکمک نمود.
7) قابلیت حفظ قالب بندی نگارش تصویر ورودی (پاراگراف بندی، نوع و اندازه فونت‌ها، جدول بندی، ستون بندی و...) درتصویر خروجی.
8) بازشناسی خودکارمتن‌های چند زبانه.
9) تشخیص بارکد.

10)واژه پردازچند زبانه به منظورافزایش دقت بازشناسی.

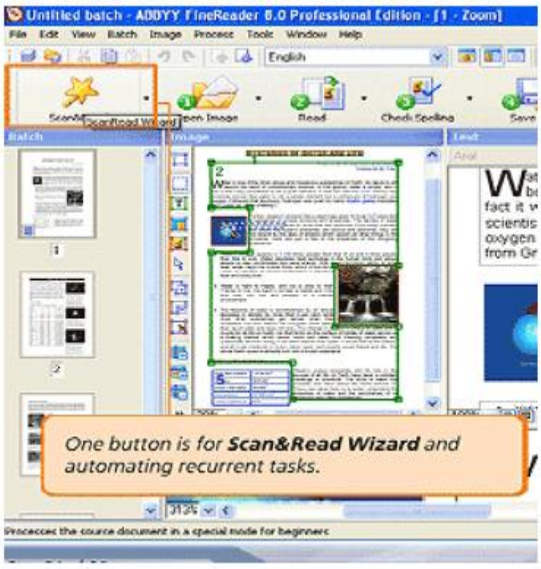
11)ویرایشگر متنی برای دیدن و اصلاح نتیجه عمل نویسه‌خوان نوری قبل از تولید فایل خروجی (تنها در نرم افزارهای 1 و 2).

12)قابلیت پردازش گروهی فایل‌ها.

جدول 9- معروف ترین نرم افزارهای تجاری نویسه‌خوان نوری لاتین.

شماره	نام محصول	شرکت سازنده	قیمت–نسخه	مرجع
۱	OmniPage	ScanSoft	Pro Edition v14.0 Office $600	(Web: ScanSost)
۲	FineReader	ABBYY	Professional Edition v7.0 $180 Corporate Edition v7.0.. $300	(Web: ABBYY)
۳	Readiris	I.R.I.S.	Pro Edition v10..... $130 Pro Corporate Editionv10..... $400	(Web: I.R.I.S.)

در شکل 24 نمایی از محیط نرم‌افزاری OmniPage و FineReader آمده‌است.

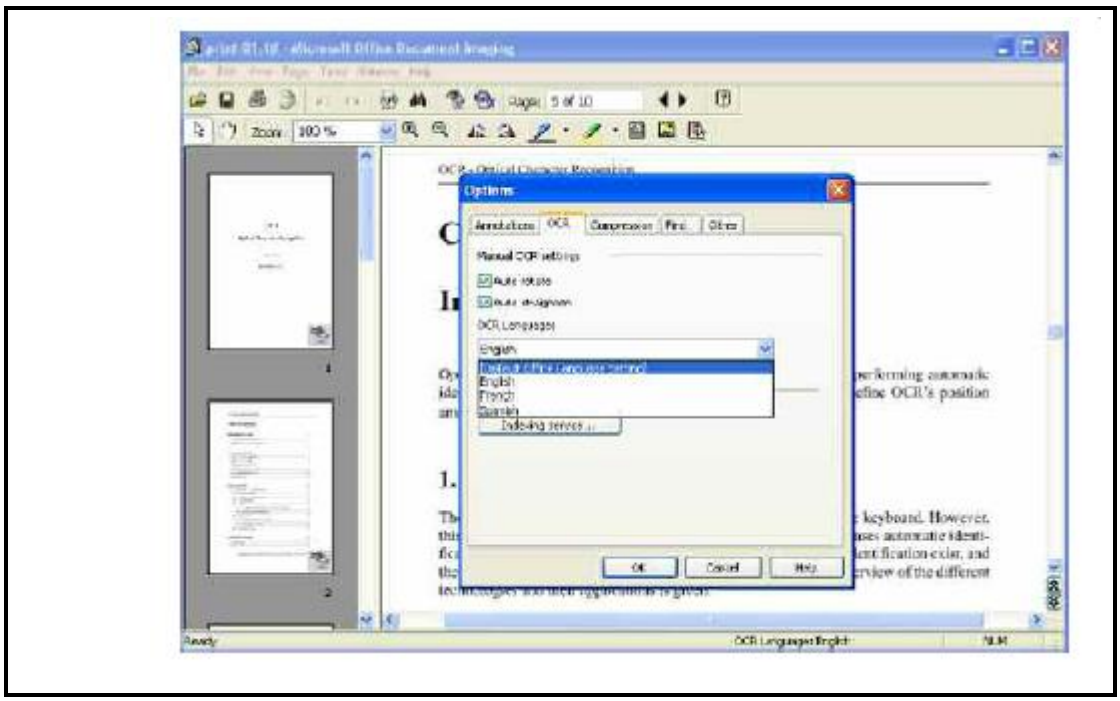
	
الف	ب

شکل 24- الف. نمایی از محیط نرم افزار Omnipage v.15، ب: .نمایی ازمحیط نرم افزار Fine Reader

علاوه بر نرم افزارهای تجاری فوق، می توان از نـرم افـزار Microsoft Office Document Imaging نـام برد که همـراه Microsoft Office توسـط شـرکت مایکروسـافت عرضـه مـی شـود و اگرچـه بسـیاری از قابلیت‌های نرم افزارهای قدرتمند تجاری را ندارد، با این حال به عنوان یک نرم افـزار خـانگی مـی توانـد مورد استفاده قرارگیرد. برخی از ویژگی‌های این نرم افزار را در ادامه مرور می نماییم :

1) قابلیت بازشناسی از روی تصاویر تنها یا چند صفحه ای (مانند تصاویر چند صفحه ای با فرمت.tif)
2) ارسال متون بازشناسی شده به Microsoft Word.
3) قابلیت تشخیص نوع و اندازه فونت ازتصویر متن و اعمال آن به متن بازشناسی شده.
4) قابلیت شناسایی تصاویر و جدول‌های ساده درون متن.
5) قابلیت شناسایی متون به زبان آسیای شرقی (چینی، ژاپنی، کره ای و...).
6) قابلیت ارسال فایل به سرویس فکس یا پست الکترونیکی.
7) استفاده از واژه نامه درونی برای تشخیص کلمات.
8) قابلیت دوران و تراز کردن خودکار تصویر
9) پشتیبانی از زبان‌های انگلیسی، فرانسوی و اسپانیایی به طور پیش فرض و هر زبان دیگری در محیط Microsoft Office به عنوان زبان پیش فرض استفاده می شود.

در شکل 25 نمایی از نرم افزار Microsoft Office Document Imaging، آمده است.



شکل 25- نمایی از محیط نرم افزار Microsoft Office Document Imaging.

8-2- نرم افزارهای تجاری عربی

در حال حاضر تنها نرم افزار نویسه‌خوان نوری تجاری که زبان فارسی را پشیبانی می نماید نرم افزار اتوماتیک ریدروی 8 محصول شرکت عربی صخره است.

صخره یک شرکت پیشگام در زمینه ارائه نرم افزار و آموزش به زبان عربی می باشد که موفقیت‌های چشمگیری را درعرصه فناوری پردازش زبان عربی حاصل نموده است. محصولات این شرکت طیف وسیعی ازکاربردها نظیر تحلیل مستندات، نویسه‌خوان نوری، ترجمه ماشینی، تولید صوت ماشینی، تبدیل متن به گفتار، موتور جستجوی وب، سیستم‌های مدیریت اطلاعات دروب و موارد دیگر را دربر می گیرد.

شرکت صخره درایران شعبه ندارد و به همین دلیل نیز دسترسی به این نرم افزار براحتی امکان پذیر نیست. این نرم افزار با یک قفل سخت افزاری که به پورت چاپگر وصل می شود قابل اجرا است. قابل

توجه است که نرم افزارواژه شناس که توسط شرکت هوش مصنوعی رایورز در بازار ایران عرضه می شود، درحقیقت همان نرم افزار تولیدی شرکت صخره است که تنها منوهای آن فارسی شده.

نرم افزار شناسا که در چند سال اخیر به وسیله شرکت جیحون افزار عرضه می شد نیز یکی از نسخه‌های پیشین همین نرم افزاربود که منوهای آن به فارسی ترجمه شده بود [21].

نرم افزار اتوماتیک ریدردارای دو نسخه گلد و پلاتینیوم می باشد. تنها تفاوت بین این دو نسخه در این است که نسخه پلاتینیوم، «SDK» نرم افزار را نیز شامل می شود. درحال حاضر قیمت نسخه‌های گلد و پلاتینیوم این نرم افزار به ترتیب درحدود 1400 و 4000 دلاراست.

نرم افزاراتوماتیک ریدر دارای ویژگی‌های زیر می باشد [22]:

1) سرعت بازشناسی بسیار بالا؛
2) پشتیبانی از 13 زبان مختلف (شامل عربی، فارسی، انگلیسی، دانمارکی، هلندی، فنلاندی، فرانسوی، آلمانی، ایتالیایی، نروژی، پرتغالی، اسپانیایی، سوئدی) ؛
3) بازشناسی متون دوزبانه عربی / انگلیسی، فارسی / انگلیسی و عربی / فرانسوی ؛
4) تشخیص اعراب و کشیدگی حروف.
5) تشخیص خودکارمتن، گرافیک و جدول در تصویر ورودی ؛
6) پشتیبانی ازقالب‌های گرافیکی TIFF، JPEG، PCX، BMP، ART ؛
7) تولید خروجی به قالب‌های ART، TXT، DTP، RTF و HTML ؛
8) قابلیت آموزش فونت به منظور افزایش دقت بازشناسی ؛
9) برخورداری از واژه پرداز عربی – انگلیسی؛

شکل 26 نمایی از محیط نرم افزار شرکت صخره را نشان می‌دهد.

شکل 26- نمایی ازمحیط نرم افزار Automatic Reader شرکت صخره

ازجمله نقطه ضعف‌های این نرم افزاربه این موارد می توان اشاره کرد [21] :

- هرچند ادعا شده که دقت بازشناسی نرم افزار اتوماتیک ریدر بالای 99 درصد است، اما آزمایش این نرم افزار صحت این ادعا را حتی برای صفحات فارسی اسکن شده با کیفیت عالی نیز تایید نمی کند. هرچند شرکت صخره با تولید این نرم افزار گام مهمی در زمینه نویسه‌خوان نوری فارسی و عربی برداشته، اما هنوز هم با جایگاه مورد انتظار فاصله زیادی دارد.
- پیکربندی صفحه و قالب بندی پاراگراف‌ها و حروف در فایل تصویری ورودی، در فایل متنی خروجی قالب (آرتی اف) با دقت بسیار پایینی رعایت می شود.
- محیط کاربری این نرم افزار، ضعیف تر از نرم افزارهای نویسه‌خوان نوری لاتین اشاره شده در بالا است.

- واژه پرداز این نرم افزاراز زبان فارسی پشتیبانی نمی کند.

8-3- نرم افزارهای تجاری فارسی

در ایران با توجه به جدید بودن نسبی بحث نویسه‌خوان نوری نرم افزارهای موجود همچنان در فاز تحقیقاتی و آزمایشی باقی مانده اند و با توجه به نقطه ضعف‌های زیاد، به بازار وارد نشده اند. اگرچه همچنان که در بخش دوم اشاره شد در حوزه نویسه‌خوان نوری فارسی گسسته، در حال حاضر نرم افزارهای در بخش آزمون ورودی استعدادهای درخشان مورد استفاده وسیع قرار گرفته است، با این حال در حوزه نویسه‌خوان نوری فارسی پیوسته (حتی از نوع تایپی) هنوز هیچ نرم افزار تجاری ای که تولید شرکت‌های داخلی باشد، وجود ندارد. در ادامه این بخش چند نرم افزار فارسی تحقیقاتی را با امکانات آنها مورد بررسی قرار می دهیم.

8-3-1- نرم افزار سپنتا

بنیاد پژوهشی رباتیک و هوش مصنوعی سپنتا از تیرماه 1383 تیمی پژوهشی در زمینه نویسه‌خوان نوری تشکیل داده است. اعضای تیم در ابتدا در یک فاز صفر مطالعاتی سه ماهه به مطالعه مقالات مرتبط با این موضوع پرداختند. در این رابطه مطالعات زیادی صورت گرفت و مقالات بسیاری خوانده شد. حاصل این تلاش‌ها را می توان در اتمام رسیدن مرحله پیش پردازش و تقسیم بندی یافت که با کیفیت بسیار بالایی یک صفحه را پردازش نموده و کجی آن را رفع می کند و بسیاری از نویزهای متداول را حذف می نماید و زیر کلمات فارسی را با دقت بسیار بالایی استخراج می کند.

نکته مهمی که در بازشناسی متون چاپی فارسی باید مد نظر قرارگیرد وجود یک بانک زیر کلمات فارسی متداول است تا از روی آن بتوان مدل زیرکلمه ورودی را مورد مقایسه قرار داد. این کار بسیار ارزشمند که کمتر مورد توجه قرار گرفته است در این مرکز انجام شده و اکنون مجموعه ای از زیر کلمات متداول فارسی تهیه شده است. با اتمام مرحله تشخیص که مقدمات آن به طور کامل آماده شده است و اکنون در دست مطالعه است می توان یک سیستم کامل بازشناسی متون چاپی را ارایه داد [21].

این بنیاد نسخه آزمایشگاهی نرم افزار خود را بر روی پایگاه وب خود قرار داده است. در راهنمای این نسخه آزمایشی شرایط زیر برای بازشناسی متن تایپی مشخص شده است ؛

- فونت بایستی در حالت Lotus Bold و به اندازه 16 یا 20 قرار داده شود.
- جملات باید کامل و نسبتا طولانی باشند.
- تصویر ایجاد شده باید در فرمت.bmp و بدون نویز باشد.

در صورت رعایت شرایط فوق، این نرم افزار آزمایشی قادر به شناسایی تصاویر با ویژگی‌های فوق می باشد. در شکل 27 نمایی از محیط آزمایشگاهی این نرم افزار مشاهده می‌شود.

شکل 27- نمایی از محیط نرم افزار آزمایشگاهی سپنتا

8-3-2- نرم افزار خودنگار

این نرم افزار درسال 1381 به بازار آمد. قابلیت تشخیص حروف و اعداد و علائم به زبان فارسی را از تصاویر آن‌ها دارا می‌باشد. این کار را از طریق خواندن فایل‌های گرافیکی و یا با استفاده از یک اسکنر می‌توان انجام داد.

8-3-3- تشخیص دهنده متن روژاوه

این محصول در سال 1384 به ثبت رسد. از ویژگی‌های آن به موارد ذیل می‌توان اشاره نمود:

1) دقت بازشناسی بیش از 98% برای تصاویر با کیفیت مناسب با فونت سایز 12 به بالا
2) قابلیت اصلاح زاویه چرخش تصویر بصورت اتوماتیک بدون محدودیت (تا 360)
3) قابلیت پردازش دسته‌ای بر روی مجموعه‌ای از فایل‌های ورودی
4) پس پردازش مبتنی بر خطاهای متداول در بازشناسی متون فارسی و امکان دریافت اصلاحات پیشنهادی از واژه نامه
5) قابلیت تشخیص متون چند فونتی و ایجاد فایل با قابلیت txt یا rtf
6) قابلیت آموزش دوگانه (از طریق تصویر اسکن شده و از طریق فونت)

9 - تعیین مشخصات مجموعه داده‌های لازم برای ارزیابی

مجموعه داده تصویری از ملزومات اساسی پژوهش‌های پردازش تصویر است. جمع‌آوری مجموعه داده برای تک‌تک پژوهشگران به دلیل پرهزینه بودن و وقت‌گیربودن کار سنگینی است و معمولاً در حداقل اندازه و برای مقاصد خاص انجام می‌پذیرد. مجموعه داده به پژوهشگران اجازه می‌دهد تا کارایی الگوریتم‌های خود را در کاربردهای واقعی آزمایش نمایند. این مسئله ایجاب می‌کند که مجموعه داده تا حد امکان به فضای کاربردی نزدیک باشد. مجموعه داده به پژوهشگران امکان می‌دهد که الگوریتم‌های خود را روی تعداد قابل توجهی از نمونه‌ها آموزش داده و تست نمایند. این خود باعث افزایش بهره‌وری پژوهشگران در تولید و توسعه الگوریتم‌ها و سیستم‌ها می‌گردد، چرا که بدون جمع‌آوری داده امکان تجربه محیط‌های واقعی برای آن‌ها میسر نخواهد بود. پژوهش در زمینه پردازش تصویر بسیار متنوع است و محتوا و خصوصیات تصاویر گردآوری شده در مجموعه داده با توجه به زمینه پژوهش و شاخص‌های در نظر گرفته شده متغیر خواهد بود. بطور کلی سه دسته کلی تحقیقاتی و کاربردی در زمینه نویسه‌خوان نوری وجود دارد (دستنویس برون‌خط، دستنویس برخط و اسناد تایپی) که هر یک نیاز به مجموعه داده ویژه‌ای دارند. مجموعه داده دستنوشته برون‌خط معمولاً با نوشتن متون مختلف توسط اشخاص گوناگون بر روی کاغذ، اسکن نمودن و نهایتاً ذخیره سازی آن‌ها در قالب تصاویر رقومی تهیه می‌گردد. دستنوشته‌های برخط معمولاً به شکل مستقیم توسط قلم‌های خاص مستقیماً رقومی می‌گردند. اطلاعات رقومی شده شامل مختصات x-y مسیر حرکت قلم همراه با اطلاعات دیگری نظیر فشار قلم و بالا و پایین رفتن قلم می‌باشد. اسناد چاپی اسناد کاغذی هستند که بوسیله دستگاه‌های چاپ افست، چاپ لیزری، جوهر افشان، چاپگر سوزنی دستگاه‌های تایپ و غیره منتشر می‌شوند. هر یک از سه دسته فوق ویژگی‌های خاصی دارند که در طراحی مجموعه داده تأثیر گذار است.

در اینجا سعی شده تمامی نیازمندی‌های مختلف بخش‌های مختلف نویسه خوان در نظر گرفته شود. هر یک از این واحدها ممکن است با اهداف متفاوتی از مجموعه داده استفاده کنند. این اهداف عمدتاً عبارتند از: آموزش (Training)، آزمون (Test)، ارزیابی کارائی (Performance Evaluation)، تنظیم

پارامترها (Parameter Tuning)، محک‌زنی (Benchmarking)، رتبه‌بندی (Ranking)، برگزاری مسابقات (Competition) و یا موارد دیگر.

مجموعه داده به چند بخش مختلف تقسیم می‌گردد که هر یک ویژگی‌های خاص خود را دارا می‌باشند. مجموعه داده‌ها می‌بایست ویژگی‌های عمومی زیر را دارا باشد:

1) وجود همه حروف، اعداد و علائم رایج در زبان فارسی
2) عدم وابستگی به الگوریتم خاص در تهیه مجموعه داده
3) برآورده نمودن نیازمندی‌های رویکردهای شناخته‌شده نویسه‌خوان نوری
4) قابلیت ارتقاء
5) پوشش منابع مختلف (مانند روزنامه، کتاب، مجله، نامه اداری، فاکس و ...)

9-1- نیازمندی‌های یک مجموعه داده خوب

یکی از مسائل مهم در ارائه مجموعه داده، سهل الوصول و قابل استفاده بودن آن برای کاربران می‌باشد. لازمهٔ این امر رعایت چند نکته می‌باشد که استفاده کردن از رویه‌های جاری و استانداردهای موجود در این زمینه یکی از آن موارد است. هر اندازه رویه‌ها و استانداردها بیشتر رعایت شوند، استفاده از مجموعه داده نیز راحت‌تر خواهد بود و این راحتی باعث می‌گردد تا تعداد کاربران مجموعه داده بیشتر شده و اضافه شدن تعداد کاربران باعث بالا رفتن اعتبار مجموعه داده می‌گردد. برهمین اساس لازم است استانداردهای لازم در این زمینه شناسایی گردند و در صورت مطلوب بودن حداکثر استفاده از آن‌ها به عمل آمده و از وضع قوانین دست و پاگیر و استانداردهای جدید حتی‌الامکان جلوگیری به‌عمل آید. استفاده از رویه‌ها، ساختارهای اطلاعاتی رایج و استانداردها به دلایل ذکر شده یک امتیاز محسوب می‌گردد. ولی باید این امر را نیز مدنظر داشت که زبان فارسی تفاوت‌هایی با زبان‌های رایج دنیا دارد. از این‌رو بومی‌سازی رویه‌ها و ساختارها یک امر ضروری به‌نظر می‌رسد و استانداردها نیز باید برای ویژگی‌ها و نیازمندی‌های زبان فارسی مورد بررسی قرار گیرند.

از مسائل مهم در تهیه و ارائه یک مجموعه داده می‌توان به موارد زیر اشاره نمود.

9-1-1- منطبق بودن بر نیاز کاربران

کاربران دو دسته هستند: یکی کاربران فنی و پژوهشگران که از دیدگاه علمی و مسائل فنی اسناد را بررسی می‌کنند و دسته دیگر کاربرانی که اسنادشان نیاز به پردازش داشته و آن‌ها را *کاربران نهایی* می‌نامیم. از دیدگاه کاربران نهایی مسئله تطابق نیاز، پوشش دادن مجموعه داده به تمامی اسنادشان است. اما از دیدگاه فنی- پژوهشی، مجموعه داده علاوه بر پوشش مناسب به اسناد واقعی باید دارای ویژگی‌هایی باشد که انتظار این دسته از کاربران را برآورده نماید. این ویژگی‌ها هم مربوط به تصاویر و هم مربوط به اطلاعات توصیفی تصاویر می‌باشند که از این میان می‌توان به میزان دقت و رنگ اسناد و همچنین محتوای توصیف‌گرها اشاره نمود.

9-1-2- قابلیت استفاده آسان

فایل‌های تصویری و توصیف‌گرها باید به گونه‌ای ارائه شوند که بتوان به سهولت از آن‌ها استفاده نمود. قالب فایل‌های گرافیکی باید به گونه‌ای باشد که به راحتی بتوان اطلاعات آن را بازیابی نمود. در اینجا استفاده نکردن از سربارهای پیچیده و فشرده‌سازی‌ها می‌تواند مزیت محسوب گردد. اطلاعات توصیف‌گر نیز حتی اگر دارای پیچیدگی هستند باید به گونه‌ای ارائه شوند که بتوان به ساده‌ترین روش آن‌ها را بازیابی نمود. اغلب اطلاعات توصیفی دارای چند سطح می‌باشند که احتمالاً باعث بوجود آمدن یک ساختار در ارائه آن‌ها می‌گردد که در نظر گرفتن سادگی استفاده در آن‌ها حائز اهمیت است.

9-1-3- در دسترس بودن

از مسائل مهم دیگر در مجموعه داده، اطلاع‌رسانی در مورد ویژگی‌های مجموعه داده و همچنین نحوه دسترسی به آن می‌باشد. به علت وجود تصاویر در مجموعه داده معمولاً حجم این نوع از مجموعه داده زیاد است و نمی‌توان آن‌ها را به راحتی در دسترس قرار داد. با در نظر گرفتن کاربردها، شاید بتوان مجموعه داده را به چند بخش تقسیم نمود. با تقسیم نمودن مجموعه داده به بخش‌های کوچکتر با کاربردهای متفاوت می‌توان از انتقال بی‌مورد همهٔ مجموعه داده جلوگیری نموده و بخش‌های مورد علاقه و نیاز کاربران را ساده‌تر در دسترس قرار داد. مجموعه داده را باید به طرق گوناگون (مثلاً از طریق لوح‌های فشرده یا اینترنت) در اختیار قرار داد. وجود نرم‌افزارهای لازم جهت استفاده از مجموعه داده نیز باعث سهولت در استفاده از مجموعه داده می‌گردد. قیمت مجموعه داده نیز از دیگر عواملی است که ممکن است سدی را در برابر کاربران ایجاد نماید. از این رو قیمت پایین و یا رایگان بودن می‌تواند در استفاده از آن تأثیر داشته باشد.

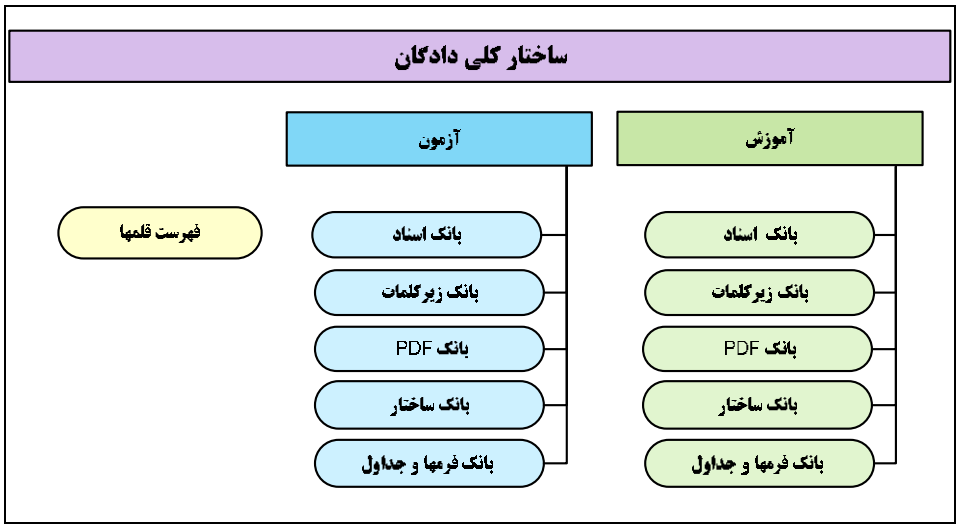
با در نظر گرفتن مسائل ذکر شده به بررسی چند مجموعه داده موجود در این زمینه پرداخته و تا حد امکان ویژگی‌های آن‌ها بررسی می‌گردد. این بررسی‌ها علاوه بر کسب تجربه از نتیجه کارهای انجام شده قبلی در دنیا و احتمالاً جلوگیری از اشتباهات گذشته گروه‌های قبلی، می‌تواند به جامع‌نگری و تبیین ویژگی‌های مجموعه داده مناسب برای خط فارسی کمک نماید.

برای یافتن ساختارهای اطلاعاتی جاری و مرسوم و شاخص‌های مد نظر، در اینجا به بررسی چند مجموعه داده موجود در این زمینه می‌پردازیم. در میان مجموعه داده‌های موجود به نظر می‌رسد مجموعه داده دانشگاه واشنگتن بیشتر مورد توجه قرار گرفته است و اطلاعات توصیف‌گر تصاویر آن جامع‌تر و کامل‌تر از دیگر مجموعه داده‌های موجود می‌باشد. از این رو این مجموعه داده با جزئیات بررسی می‌گردد.

9-2- ساختار عمومی مجموعه داده‌ها

ساختار مجموعه داده و محتوای آن باید به‌گونه‌ای باشد که بتواند نیازمندی‌های مختلف تحقیق و توسعه نرم‌افزارهای نویسه‌خوان نوری را فراهم آورد. از این نیازمندی‌ها، می‌توان به برآورده نمودن اطلاعات لازم برای قسمت‌های مختلف آموزش‌پذیر و آزمون قسمت‌های آموزش‌پذیر و غیر آموزش‌پذیر اشاره کرد. این

قبیل نیازمندی‌ها را با تفکیک مجموعه داده به مجموعه‌های آموزش و آزمون می‌توان برآورده نمود (شکل 28).



شکل 28- ساختار کلی مجموعه داده

از نظر محتوا بخش‌های مختلف نویسه‌خوان نوری احتیاج به اطلاعات مختلفی دارند. یکی از قسمت‌های مهم، واحد بازشناسی نویسه‌ها می‌باشد. از مهمترین اطلاعاتی که واحد بازشناسی به آن احتیاج دارد، می‌توان به متن معادل تصویر یا رونوشت و نوع و حالت قلم اشاره نمود که یک بانک مجزا به‌نام بانک نویسه‌ها برای آن درنظر گرفته شده است. از دیگر بخش‌های مهم نویسه‌خوان نوری، واحد تحلیل ساختار سند می‌باشد که از اطلاعات لازم برای این واحد می‌توان به مکان نواحی متنی، تصویری، جدول و غیره اشاره نمود. برای برآورده نمودن نیاز این قسمت یک بانک جداگانه به نام بانک ساختار در نظر گرفته شده است.

اشاره به این نکته ضروری است که در بانک نویسه‌ها اطلاعات مربوط به ساختار سند نیز وجود دارد ولی در اینجا این سؤال مطرح می‌شود که با توجه به وجود اطلاعات ساختار سند در بانک نویسه‌ها چه نیازی به بانک ساختار به شکل جداگانه است؟ علت جدا شدن این بانک به این دلیل است که هر صفحه و فایل بانک نویسه‌ها ممکن است حاوی هزاران نمونه از نویسه‌های مختلف باشد ولی از لحاظ ساختار سند فقط یک رکورد محسوب می‌گردد. با بالا بردن تعداد رکوردهای ساختار سند تعداد رکوردهای نویسه‌ها به

شکل نامتناسبی افزایش می‌یابد. با رساندن تعداد رکوردهای ساختار به حد مطلوب، باید تعداد زیادی اطلاعات در سطح نویسه تولید نمود که این امر باعث نامتقارن شدن مجموعه داده می‌گردد. از طرفی تولید کردن اطلاعات مشخصه در سطح تحلیل ساختار سند ساده‌تر از تولید اطلاعات مشخصه در سطح نویسه‌ها می‌باشد. دلیل ذکر شده از دلایل عمده برای در نظر گرفتن یک بانک مجزا برای بانک ساختار سند است. توجه به این نکته ضروری است که در بانک نویسه‌ها هم اطلاعات مشخصه ساختار وجود دارد و هم اطلاعات مشخصه نویسه‌ها ولی در بانک ساختار فقط اطلاعات مشخصه در سطح ساختار وجود دارد.

یکی دیگر از اهداف تهیه این مجموعه داده برآورده نمودن نیازهای مختلف رویکردهای مختلف واحد بازشناسی نویسه‌ها است. یکی از رویکردهای مورد توجه محققین استفاده کردن از واحدهای کوچکتر از کلمه می‌باشد که به نام‌های مختلفی مانند زیرکلمه، شبه‌کلمه یا همبند نامگذاری شده است. برای برآورده نمودن اطلاعات مورد نیاز این دسته نیز یک بانک به نام بانک زیرکلمات در نظر گرفته شده است. ابتدا زیرکلمات از یک بانک بزرگ متنی استخراج شده و با استفاده از قلم‌ها و اندازه‌های مختلف، بعد از چاپ بر روی کاغذ، تصاویر آن‌ها تهیه می‌گردد.

یکی از مسائل مورد توجه در بازشناسی اسناد، بازشناسی جداول و استخراج اطلاعات از انواع فرم‌ها می‌باشد. به‌دلایلی که قبلاً در مورد تفکیک بانک ساختار از بانک نویسه‌ها ذکر شد، یک بانک مجزا برای انواع جداول و فرم‌ها در نظر گرفته شده است که از منابع مختلف واقعی تهیه می‌گردند.

هر چند تمام بانک‌های ذکر شده دارای اطلاعات غنی از مشخصه‌های مختلف هستند ولی اغلب آن‌ها به شکل دستی تهیه شده و به همین دلیل برخی اطلاعات دقیق مانند چهارچوب دقیق یک نویسه از آن‌ها قابل استخراج نیست. برای بدست آوردن اطلاعات مشخصه دقیق، به اصل منبع دیجیتالی اسناد نیاز است. این منابع دیجیتالی همان قالب‌های ذخیره‌سازی اطلاعات در فایل‌های رایانه است که تنوع زیادی دارند. ولی در میان آن‌ها و طبق اطلاعات ما، تقریباً از هیچ یک غیر از PDF نمی‌توان اطلاعات مشخصه لازم را استخراج نمود. این منابع دیجیتالی باید دارای ویژگی‌های خاصی می‌باشند. برای استخراج اطلاعات مشخصه دقیق تصاویر، یک بانک به نام بانک PDF در نظر گرفته شده است .

یکی از مسائل دیگر در تولید مجموعه داده، مشخص کردن و نام‌گذاری قلم‌ها در اطلاعات مشخصه می‌باشد. فونت‌های زیادی در اسناد چاپی وجود دارند که نام‌گذاری آن‌ها کار ساده‌ای نیست و شاید بسیاری از آن‌ها اسم خاصی نداشته و اگر هم اسم خاصی داشته باشند ممکن است اتفاق نظر برای نام

فونت‌ها وجود نداشته باشد. البته این مسئله در مورد قلم‌های رایانه‌ای کمتر است. برای حل کردن این مسئله، در مجموعه داده یک فهرست از قلم‌ها جمع‌آوری می‌گردد. در مجموعه فهرست قلم‌ها، یک نمونه از هر یک از قلم‌های موجود در مجموعه داده قرار می‌گیرد. به هر یک از قلم‌های موجود در فهرست یک کد به عنوان شناسهٔ قلم تخصیص می‌یابد. بنابراین با داشتن شناسه فونت، می‌توان مشخصات فونت، ازجمله نام فونت را از مجموعه فهرست فونت‌ها استخراج نمود. سپس در هر کجای مجموعه داده که لازم باشد نوع قلم مشخص گردد، به جای نام قلم از شناسهٔ قلم استفاده می‌شود .

همانگونه که در شکل 28 مشخص شده است کل مجموعه داده به دو بخش آموزش و آزمون تفکیک می‌گردد. تقریباً دو سوم از حجم کل مجموعه داده در مجموعه آموزش و یک سوم باقیمانده در مجموعه آزمون قرار خواهد گرفت.

9-3- مجموعه داده‌های مطرح

9-3-1- مجموعه داده دانشگاه واشنگتن (UW-III)

UW-III سومین دسته از سری مجموعه داده تولید شده توسط دانشگاه واشنگتن می‌باشد. این مجموعه داده توسط یک تیم از دانشجویان و فارغ‌التحصیلان به سرپرستی چند استاد دانشگاه تهیه شده است. این مجموعه داده شامل موارد زیر است :

25 صفحه از فرمول‌های شیمیایی همراه با اطلاعات مشخصه در XFIG، 25 صفحه فرمول‌های ریاضی همراه با اطلاعات مشخصهٔ در XFIG و LATEX، 40 صفحه ترسیم همراه با اطلاعات مشخصه در AUTUCAD و IGES، 44 صفحه تصاویر TFF از ترسیمات مهندسی همراه با اطلاعات مشخصه در AUTOCAD، 33 صفحه تصاویر TIFF متون انگلیسی که از منبع LaTeX تولید شده‌اند همراه با اطلاعات مشخصه آن‌ها در سطح نویسه‌ها، 33 صفحه چاپ شده و اسکن شده متن انگلیسی همراه با اطلاعات مشخصه در سطح نویسه‌ها، 979 صفحه از مجموعه داده UW-I در فرمت DAFS که زاویه اریب آن‌ها اصلاح شده است همراه با چارچوب کلمات ، 623 صفحه از UW-II در فرمتDAFS که زاویه اریب آن‌ها اصلاح شده است همراه با چارچوب کلمات برای 607 صفحه از آن‌ها.

در این مجموعه داده نرم افزارهای زیر وجود دارند:

1) ایجاد اطلاعات مشخصه برای اسنادی که به شکل مصنوعی تولید می‌گردند.
2) ایجاد افت کیفیت مصنوعی در اسناد
3) منطبق کردن تصاویر ایده‌ال روی تصاویر واقعی اسناد
4) انتقال اطلاعات مشخصه نویسه‌ها بین دو تصویر منطبق شده
5) تخمین پارامترهای مدل افت کیفیت اسناد
6) معتبر ساختن مدل افت کیفیت
7) قالب بندی اسناد عربی، هندی و موزیک در LaTeX
8) تست فرضیه‌ها دربارهٔ پارامترهای جمعیت گوسی چند متغیره
9) توابع کتابخانه‌های TIFF

همانند مجموعه داده‌های دیگر UW، اطلاعات چارچوب نواحی هر صفحه، ویژگی‌های کلی سند و اطلاعات کیفی هر صفحه ارائه شده است.

9-3-2- مجموعه داده عربی ERIM (Environmental Research Institute)

تصاویر این مجموعه داده از کتاب‌ها و مجلات چاپی عربی استخراج شده‌اند. بنابر اطلاعات ارائه شده توسط [23] ، این مجموعه تنوع زیادی از قلم‌ها با کیفیت کم تا کیفیت زیاد را شامل می‌شود. تمامی تصاویر با دقت 300dpi اسکن شده‌اند. ویژگی‌های مجموعه داده آن به قرار ذیل می‌باشد:

1) بیش از 750 صفحه عربی
2) همهٔ صفحات با دقت 300dpi توسط یک روبشگر صفحه تخت اسکن شده‌اند.
3) برخی صفحات تصاویر سطوح خاکستری نیز دارند.

4) تمام اطلاعات نوشتاری به شکل یونی‌کد ذخیره شده‌اند.

5) در جمع در حدود 1.000.000 کاراکتر در مجموعه داده وجود دارد.

6) توصیف‌گر متنی برای تمامی کارکترها و کارکترهای ترکیب شده ارائه شده است.

7) بیش از 200 کارکتر ترکیبی عربی مجزا در مجموعه داده وجود دارد.

8) در مجموعه داده تنوع زیادی از قلم و کیفیت وجود دارد (طبق ادعای ارائه‌کنندگان).

9) مجموعه داده به مجموعه‌های آموزش، آماری و آزمون تقسیم شده است.

10)اسناد مربوط به قالب تصاویر ارائه شده است.

9-3-2-1- فرمت قالب تصویر

فایل‌های تصویر با نام img.* ارائه شده‌اند و به دوگونه هستند: فایل‌های دو سطحی و سطوح خاکستری. فایل‌های سطوح خاکستری با چهار بایت شروع می‌شوند که دو بایت اول یک عدد صحیح است که عرض تصویر را مشخص می‌نماید. دو بایت بعدی ارتفاع تصویر را مشخص می‌نماید که این ارتفاع یک عدد صحیح است. بعد از این چهار بایت یک رشته از بایت‌ها (به اندازهٔ طول x ارتفاع) وجود دارد که نقاط تصویر را مشخص می‌نماید. یک بایت برای هر نقطه.

در تصاویر دو سطحی نیز دو بایت اول عرض و دو بایت بعدی طول تصویر می‌باشد که با رشته‌ای از بایت‌های تصویر ادامه پیدا می‌کند. هر 8 نقطه از تصویر در یک بایت کد شده‌اند (هر بیت برای یک نقطه از تصویر).

9-3-2-2- فرمت فایل‌های توصیف‌گر

فایل‌های توصیف‌گر با نام "tru.*" ارائه شده‌اند. تمامی فایل‌های دو سطحی و توصیف‌گرها به شکل سلسله مراتبی سازمان‌دهی شده‌اند. تمامی اطلاعاتی که از یک منبع گرفته شده‌اند در یک گروه به نام

سند گردآوری شده‌اند. هر سند دارای تعداد مشخصی صفحه می‌باشد و هر صفحه تعداد مشخص *ناحیه* دارد. هر تصویر دو سطحی (فایل img.) یک فایل متناظر توصیف‌گر (tru.) دارد. هر زوج img. و tru.، تصویر دو سطحی تصویر ناحیه و فایل توصیف‌گر متناظر ناحیه می‌باشد. نام هر زوج به گونه‌ای است که نام سند، شماره صفحه و شمارهٔ ناحیه را مشخص می‌نماید.

هر ناحیه تعداد مشخصی خط متنی دارد که هر خط تعداد مشخصی کلمه و هر کلمه تعداد مشخصی کاراکتر دارد. این ساختار چند سطحی برای هر ناحیه در فایل tru. متعلقه ارائه شده است. برای هر ناحیه از تصویر یک فایل tru. وجود دارد.

9-3-3- مجموعه داده اولو

این مجموعه داده توسط گروه تحقیقاتی *مدیاتیم* که در دانشگاه اولو قرار دارند، تهیه شده است و یک مجموعه از اسناد اسکن شده همراه با توصیف‌گرهای ساختار فیزیکی و منطقی آن‌ها می‌باشد. این مجموعه داده شامل 500 صفحه از اسناد انتشار یافته سال 1978 میلادی و قبل از آن می‌باشد که زبان اسناد اغلب انگلیسی و فنلاندی است.

در توصیف‌گرهای این مجموعه داده می‌توان اطلاعاتی از قبیل نوع سند (مقاله، آگهی، فرم، اسناد دیگر)، نوع صفحه (فهرست مندرجات، صفحه عنوان، صفحه منابع)، نوع ناحیه (متنی، تصویری، ترسیم، دیگر) و ترتیب منطقی نواحی را پیدا کرد. قابل ذکر است که توصیف‌گر نواحی متنی در این مجموعه داده وجود ندارد. در این مجموعه داده ابزاری برای مرور مجموعه داده و مشاهدهٔ اسناد انتخاب شده قرار داده شده است. تصاویر اصلی این مجموعه داده در قالب tiff فشرده نشده (جمعاً 7/2 GB) نگهداری می‌شوند، ولی به دلیل محدودیت حجم لوح‌های فشرده، تصاویر در قالب jpeg ارائه شده‌اند. تصاویر رنگی و با دقت 300dpi اسکن شده‌اند. تصاویر این مجموعه داده در وبگاه [24] وجود دارد و قابل دانلود کردن است. فایل‌های توصیف‌گر به شکل ASCII ارائه شده‌اند. جدول 10 انواع سندهای این مجموعه داده را نشان می‌دهد.

جدول 10- تنوع و حجم مجموعه داده اولو

نوع سند	تعداد سند	تعداد صفحه

لیست آدرس	4	6
آگهی	24	24
مقاله	58	233
کارت ویزیت	10	11
چک	2	3
تصاویر جداسازی رنگ‌ها	10	10
فرهنگ لغت	3	12
فرم	18	23
ترسیم	10	10
دفترچه راهنما (دستورالعمل)	4	35
ریاضی	2	17
موزیک	5	10
روزنامه	5	42
رئوس مطالب	6	19
دفترچه تلفن	2	7
لیست برنامه رایانه‌ای	4	12
نقشه‌خیابان	5	6
نقشه عوارضی زمین	8	8

جمع	198	512

9-3-4- مجموعه داده Al-Isra

مجموعه داده Al-Isra در مقاله کنفرانس IEEE مهندسان برق و کامپیوتر کانادا در سال 1999 میلادی ارائه شد [25]. این مجموعه داده شامل کلمات دستنویس عربی، اعداد و امضاء است که در دانشگاه Al-Isra کشور اردن جمع‌آوری شده است. این داده‌ها از 500 دانشجوی این دانشگاه که بطور تصادفی انتخاب شده‌اند بدست آمده است.

این مجموعه داده شامل موارد زیر است:

1) مجموعه داده برون‌خط.
2) 37,000 کلمه عربی دستنویس.
3) 10,000 نمونه از اعداد.
4) 2,500 نمونه امضاء.
5) 500 جمله دلخواه عربی.

تمام تصاویر این مجموعه بصورت سیاه سفید و خاکستری است که با فرمت BMP ذخیره شده‌اند. اگر چه این مجموعه داده در سال 1999 ارائه شد، اما تا کنون هیچ داده‌ای از آن بر روی اینترنت منتشر نشده است.

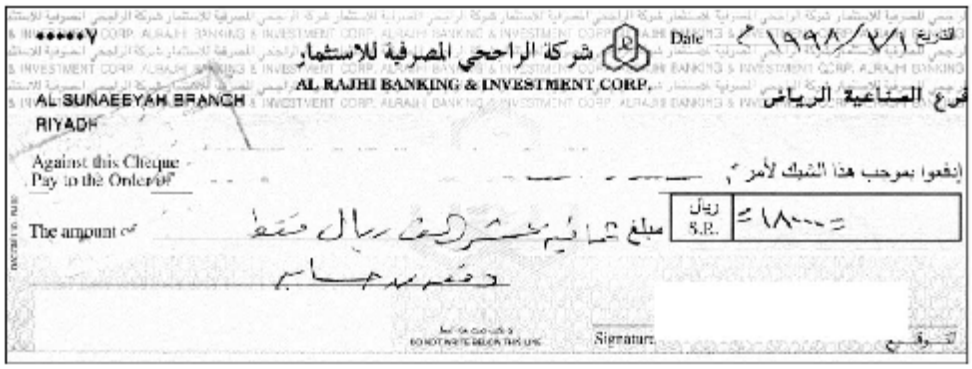
9-3-5- مجموعه داده Arab_CENPARMI

در سال 2000 میلادی و در کارگاه IWFHR ، *مرکز شناسایی الگو و هوش ماشین* در مونترال کانادا، مجموعه داده‌ای مشتمل بر تصاویر 3,000 چک بانکی را منتشر نمود [26]. این تصاویر واقعی، از چک‌های استفاده شده در سیستم بانکی انتخاب شده‌اند که قسمت‌های مبلغ مندرج در چک بصورت

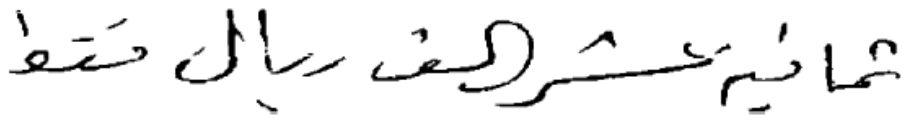
عدد و حروف از تصویر چک جدا شده و با کدهای ASCII برچسب خورده‌اند. شکل 42 نمونه‌ای از این مجموعه داده را نشان می‌دهد. قسمت‌های مبلغ عددی و حروفی چک در شکل 43و شکل 44 نشان داده شده‌اند.

مهمترین خصوصیات این مجموعه داده عبارتند از:

1) مجموعه داده برون‌خط.
2) شامل 29,498 زیرکلمه است.
3) شامل 15,175 نمونه عدد است.
4) شامل 2,499 مبلغ نوشته شده بصورت عدد و حروف است.
5) از تصاویر موجود در کاربرد واقعی جمع‌آوری شده است.



شکل 29- مثالی از یکی از چک‌های موجود در مجموعه داده CENPARMI



شکل 30- قسمت مبلغ چک که بصورت حروف نوشته شده است

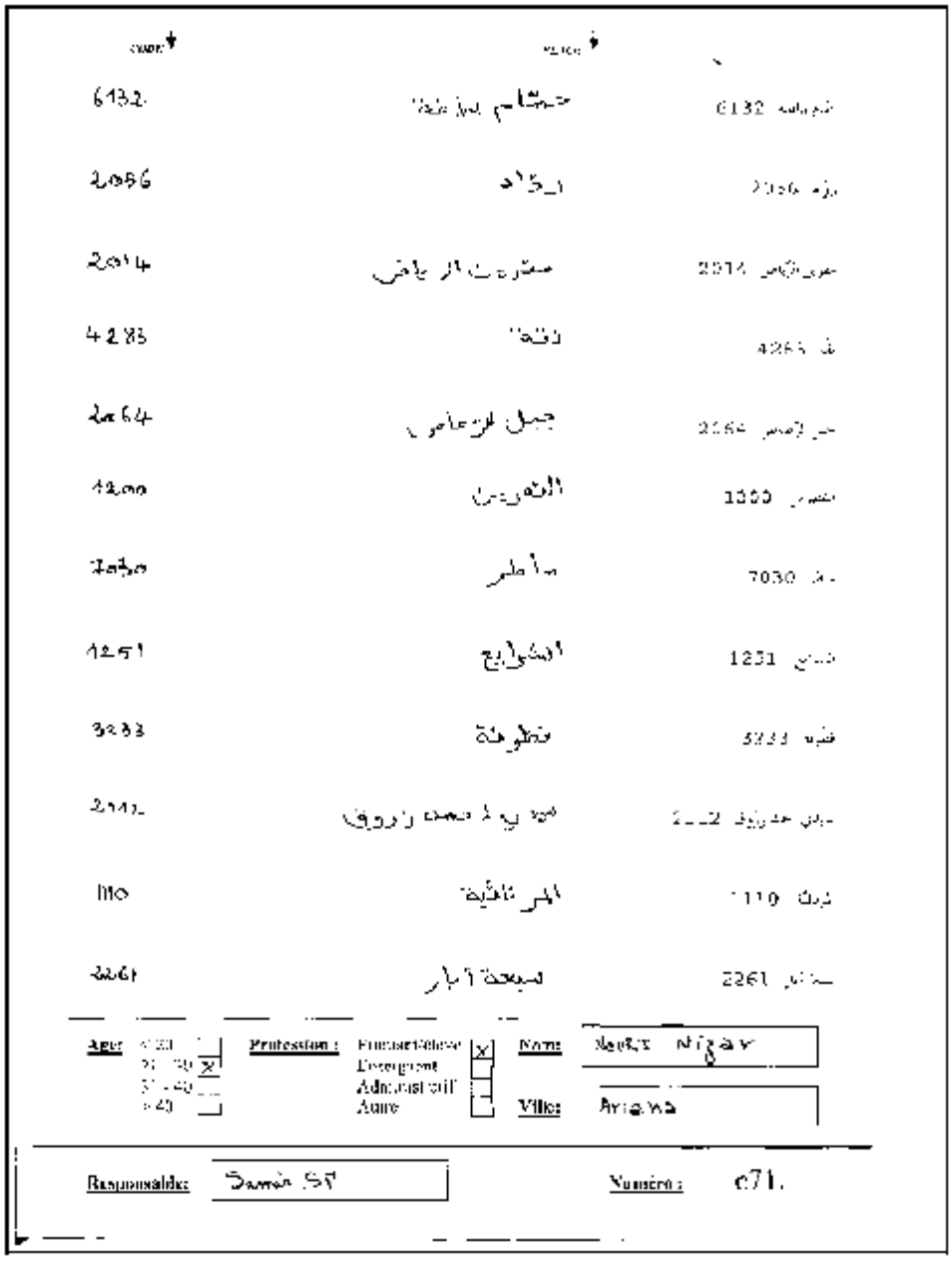
شکل 31- قسمت مبلغ چک که بصورت عددی نوشته شده است

9-3-6- مجموعه داده IFN/ENIT

مجموعه داده IFN/ENIT در سال 2002 میلادی با همکاری موسسه IFN در آلمان و موسسه ENIT در کشور تونس ایجاد شد [27]. این مجموعه داده شامل اسامی‌شهرها و روستاهای کشور تونس است که بوسیله فرم‌های خاصی که مرحله برچسب زنی را تسهیل می‌کنند(شکل 32) ، جمع‌آوری شده است. مهمترین ویژگی‌های این مجموعه عبارتند از:

1) مجموعه داده برون‌خط.
2) شامل 937 کلمه از اسامی‌شهرها و روستاهای تونس است.
3) دارای 26,459 کلمه دستنویس عربی است.
4) دارای 115,585 زیرکلمه و 212,211 کاراکتر است.
5) داده‌ها از 411 نویسنده با سنین و تحصیلات مختلف جمع‌آوری شده‌اند.
6) مجموعه داده به 4 زیرمجموعه مجزا تقسیم شده است.

در دو مسابقه معتبر کنفرانس ICDAR در سال‌های 2005 و 2007 به عنوان مجموعه داده محک زنی، بکار گرفته شده است.



شکل 32- فرم تهیه شده برای جمع‌آوری مجموعه داده IFN/ENIT

9-3-7- مجموعه داده ARABASE

در سال 2005 میلادی، مجموعه داده ARABASE توسط Ben Amara و همکارانش معرفی شد [28]. نویسندگان ادعا می‌کنند که این مجموعه داده که شامل متون دستنویس و تایپی است می‌تواند به محققان نویسه‌خوان نوری برخط و برون‌خط عربی کمک کند. این مجموعه شامل عبارات، کلمات، حروف، اعداد و امضاء است. همچنین ابزارهایی برای کار با آن نیز آماده شده است. اما متاسفانه علیرغم ادعاهای انجام شده، هنوز داده‌ای در اختیار پژوهشگران قرار نگرفته است.

9-3-8- مجموعه داده IFHCDB

بر اساس دانسته‌های ما، تا قبل از سال 2006 هنوز هیچ مجموعه داده استانداردی در زمینه نویسه‌خوان نوری فارسی در مجامع بین‌المللی ارائه نشده بود و بنظر می‌رسید که پژوهشگران علاقه‌مند به نویسه‌خوان نوری عربی گوی سبقت را از همتایان خود در زبان فارسی ربوده‌اند. در این سال، دو مجموعه داده فارسی (IFHCDB و Farsi_CENPARMI) در کارگاه IWFHR بطور همزمان در اختیار محققان قرار گرفتند که هر دو از لحاظ اندازه و تنوع از اهمیت خاصی برخوردار بودند.

مجموعه داده IFHCDB بخشی از مجموعه داده بزرگ حدف است که به سفارش *دبیرخانه شورای عالی اطلاع‌رسانی* و توسط شرکت *اندیشه نرم‌افزار پایا* تهیه شد. سپس مجموعه داده IFHCDB بوسیله دانشکده مهندسی برق دانشگاه صنعتی امیرکبیر تدوین و در دهمین کنفرانس بین‌المللی IWFHR ارائه شد [26].

این مجموعه شامل ارقام و حروف گسسته فارسی است که از فرم‌های مربوط به ثبت نام دانش آموزان متقاضی ورود به دبیرستان‌های استعدادهای درخشان جمع‌آوری شده است (شکل 33). شکل 34چند نمونه از داده‌های این مجموعه را نشان می‌دهد.

مهمترین مشخصات این مجموعه داده عبارتند از:

1) مجموعه داده برون‌خط.

2) تمام تصاویر خاکستری هستند که با درجه تفکیک 300dpi روبش شده‌اند.

3) شامل 52,380 نمونه حرف است.

4) شامل 17,740 نمونه عدد است.

5) درصد فراوانی هر یک از حروف و ارقام مطابق درصد فراوانی آن‌ها در کاربرد واقعی است.

6) تعداد نویسندگان بسیار زیاد.

7) محدوده سنی و تحصیلات نویسندگان آن محدود است.

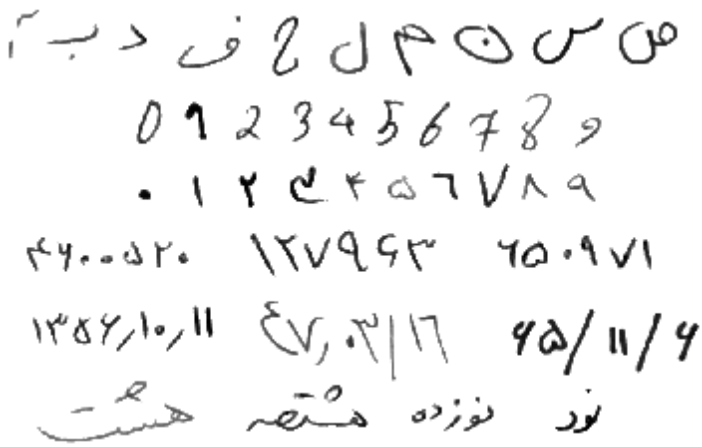
8) به زیرمجموعه‌های آموزش و آزمایش تقسیم شده است.

شکل 33- نمونه فرم مورد استفاده برای جمع‌آوری اطلاعات در مجموعه داده IFHCDB

شکل 34- چند نمونه از حروف و ارقام مجموعه داده IFHCDB

9-3-9- مجموعه داده Farsi_CENPARMI

همزمان با مجموعه داده IFHCDB مجموعه داده Farsi_CENPARMI نیز ارائه شد[29]. گروه تحقیقاتی CENPARMI با استفاده از تجربیات قبلی در جمع‌آوری مجموعه داده عربی (Arabic_CENPARMI)، این بار به جمع‌آوری ارقام مجزا، رشته‌های عددی، مبلغ نوشته شده به حروف بر روی چک‌های بانکی و تاریخ نوشته شده با اعداد، پرداختند. شکل 35 چند نمونه از داده‌های مجموعه Farsi_CENPARMI را نشان می‌دهد.



شکل 35- چند نمونه از داده‌های مجموعه Farsi_CENPARMI

سطر اول: حروف مجزا، سطر دوم: ارقام لاتین، سطر سوم: ارقام فارسی.

سطر چهارم: مبلغ نوشته شده به عدد، سطر پنجم: تاریخ، سطر ششم: مبلغ نوشته شده به حروف.

مهمترین مشخصات این مجموعه داده عبارتند از:

1) مجموعه داده برون‌خط.
2) تمام تصاویر هم بصورت خاکستری و هم بصورت سیاه و سفید موجود هستند.
3) مجموعه داده به سه زیرمجموعه آموزش، آزمایش و تصدیق ارزیابی تقسیم شده است.
4) داده‌ها از 175 نفر نویسنده تهیه شده است.
5) نویسندگان در خارج از ایران بوده‌اند.
6) شامل 18,000 نمونه رقم است.
7) شامل 11,900 نمونه حرف است.
8) شامل 3,500 نمونه رقم لاتین است.
9) شامل 7,350 رشته ارقام است.

10)شامل 175 تاریخ است.

11)شامل 7,875 مبلغ نوشته شده بصورت حروف است.

12)شامل 700 مبلغ نوشته شده بصورت عددی است.

9-3-10- مجموعه داده On line_TMU

جمع‌آوری داده‌های برخط، بدلیل استفاده از قلم مخصوص و صفحه دیجیتال کننده بجای قلم و کاغذ معمولی، اغلب دشوارتر از جمع‌آوری داده‌های برون‌خط است. برای جمع‌آوری نوشتار برون‌خط می‌توان صفحات مستندات تایپی را بوسیله روبشگر روبش کرد. همچنین برای جمع‌آوری داده‌های برون‌خط دستنویس، می‌توان یک قلم و کاغذ در اختیار افراد قرار داد و از آن‌ها خواست که نویسه‌های مورد نظر را روی کاغذ بنویسند. سپس این کاغذها را بوسیله روبشگر روبش نمود. اما برای جمع‌آوری داده‌های برخط، نویسنده باید به کمک قلم مخصوص و صفحه دیجیتال کننده، در مقابل کامپیوتر نویسه‌ها را بنویسد. شاید به همین علت باشد که تعداد مجموعه‌های داده برخط برای نویسه‌خوان نوری فارسی و عربی بسیار کم و دور از دسترس عموم است و بیشتر گروه‌های تحقیقاتی با استفاده از وسائل خاص، تعداد محدودی داده را جمع می‌کنند. اما امروزه با پیشرفت تکنولوژی و ظهور وسایلی مانند *کامپیوترهای* جیبی و تلفن‌های همراه هوشمند، موارد کاربرد داده‌های برخط افزایش چشمگیری داشته است. بنابراین وجود چنین مجموعه داده‌هایی اهمیت بیشتری یافته‌اند.

شاید بتوان مجموعه داده On line_TMU را یکی از معدود مجموعه داده‌های بزرگ جمع‌آوری شده برای شناسایی نویسه‌های برخط فارسی دانست. این مجموعه توسط دانشکده مهندسی برق دانشگاه تربیت مدرس جمع‌آوری شده است [30]. از میان کلمات رایج فارسی که از بایگانی روزنامه‌های همشهری و کیهان استخراج شده‌اند، 7,317 زیرکلمه انتخاب شده است. جدول 11 و جدول 12 بترتیب نمونه‌هایی از کلمات و زیر-کلمات استخراج شده و تعداد تکرار آن‌ها را نشان می‌دهند. برای جمع‌آوری داده‌ها از یک برنامه رابط بین نویسنده و کامپیوتر استفاده شده که بوسیله آن نویسنده بتواند مطالب

نوشته شده را ببینند. این روش کمک می‌کند تا فرآیند نوشتن بصورت طبیعی‌تری انجام شود. شکل 36 چند نمونه از داده‌های این مجموعه را نشان می‌دهد.

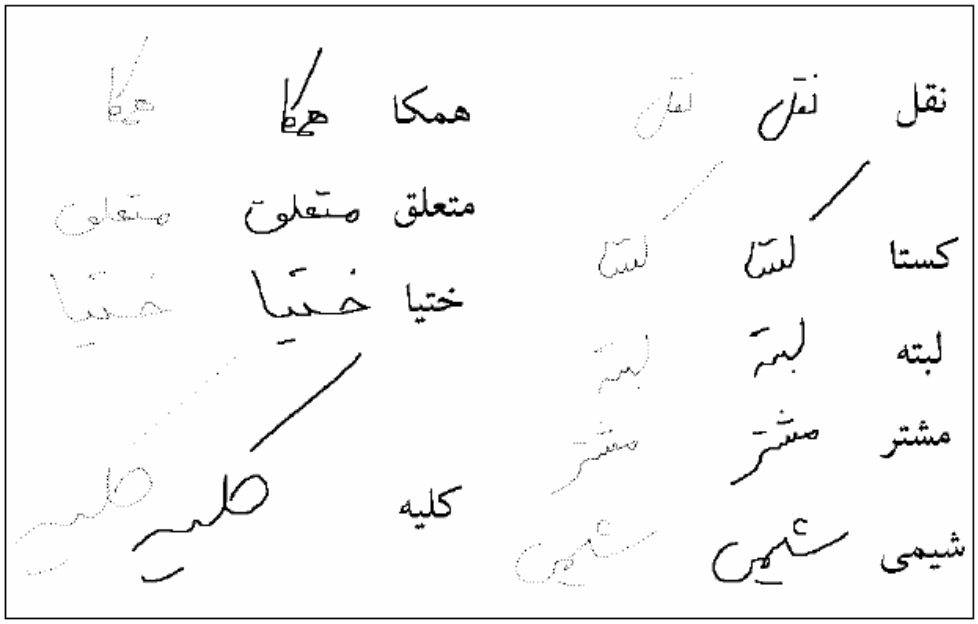
از این داده‌های جمع‌آوری شده می‌توان برای بازشناسی برخط ارقام، حروف مجزا و کلمات فارسی و همچنین شکستن کلمات به زیر-کلمات و حروف استفاده کرد.

و	۲۰۹۰۸۲۵	کسانی	۱۲۲۱۵	روان	۳۳۷۷	بتوانیم	۳۰۳	روندپرونده	۴
در	۱۵۴۶۰۲۶	خرداد	۱۲۲۰۷	اینگونه	۲۷۷۶	قادرند	۳۰۳	روشنفکرترین	۴
به	۱۲۴۳۵۴۰	بیماری	۱۱۱۹۲	دانشکده	۲۲۷۴	وزرا	۳۰۳	روندصعودی	۴

جدول 11- نمونه‌ای از کلمات و تعداد تکرار آن‌ها که از بایگانی روزنامه استخراج شده‌اند

ا	۱۳۳۳۸۲۳۸	عمو	۴۰۹۸۰	طمه	۱۹۵۶
ر	۸۵۸۶۹۷۹	قی	۴۰۷۱۴	منظر	۱۹۵۳
د	۷۹۵۵۳۳۳	نشگا	۴۰۶۶۳	عیل	۱۹۴۹

جدول 12- نمونه‌ای از زیر-کلمات و تعداد تکرار آن‌ها که از بایگانی روزنامه استخراج شده‌اند



شکل 36- چند نمونه از زیر-کلمات جمع‌آوری شده در مجموعه داده TMU_online

ستون اول: زیرکلمه بصورت تایپی. ستون دوم: شکلی که نویسنده موقع نوشتن روی صفحه کامپیوتر می‌بیند

ستون سوم: نقاطی که مختصات آن‌ها به عنوان داده ذخیره می‌شوند

مهمترین ویژگی‌های این مجموعه داده عبارتند از:

- مجموعه داده برخط.
- شامل 1,000 زیرکلمه است.
- از هر زیرکلمه بطور متوسط 12 نمونه جمع‌آوری شده است.
- شامل ارقام، حروف مجزا و چهار علامت { ! ؟ ، . } است.
- 124 نفر در جمع‌آوری دستنوشته‌ها شرکت داشته‌اند (98 نفر مرد و 26 نفر زن).
- هر نفر ارقام صفر تا نه، حروف مجزا ،چهار علامت و همچنین 100 زیرکلمه از مجموعه 1000 زیرکلمه را نوشته است.

9-3-11- مجموعه داده HODA Farsi Digit

چندی پس از انتشار مجموعه داده‌های IFHCDB و Farsi_CENPARMI ،مجموعه داده بسیار بزرگ HODA که شامل ارقام فارسی است، توسط دانشگاه تربیت مدرس ایجاده شده و در اختیار محققان قرار گرفت. مشخصات و روند جمع‌آوری این مجموعه داده در سال 2007 میلادی در مجله Pattern Recognition Letters منتشر خواهد شد [7]. این مجموعه شامل 102,352 نمونه عدد فارسی است که از 12,000 فرم مربوط به آزمون ورودی کاردانی به کارشناسی و کارشناسی ارشد جمع‌آوری شده است .

مهمترین ویژگی‌های این مجموعه عبارتند از:

- مجموعه داده برون‌خط.
- فقط شامل ارقام دستنویس فارسی است.
- تصاویر بصورت سیاه و سفید هستند.

- شامل 102,352 رقم است (تعداد نمونه‌های بسیار زیاد).
- محدوده سنی و تحصیلات نویسندگان آن محدود است.
- توزیع نمونه‌ها در هر کلاس تقریباً یکسان است (حدود 10,000 نمونه در هر کلاس).

مجموعه داده به سه زیرمجموعه آموزش، آزمایش و یک مجموعه دیگر تقسیم شده است.

9-4- مجموعه داده دستنویس برای نویسه‌خوان نوری زبان فارسی

9-4-1- انواع مجموعه داده‌های دستنویس

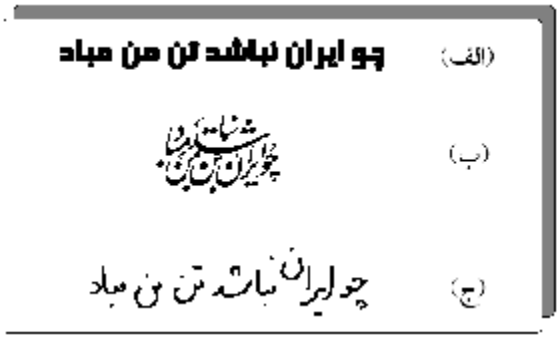
9-4-1-1- داده تایپی و دستنویس

متون به دو دسته کلی تایپی و دستنویس تقسیم می‌شوند. متون دستنویس به آسانی و با ساده‌ترین ابزارها مانند کاغذ و قلم نوشته می‌شوند و در نتیجه از نوشتارهای تایپی متداول‌تر و فراگیرتر هستند. باید توجه داشت که جمع‌آوری و شناسایی متون تایپی در مقایسه با متون دستنویس ساده‌تر می‌باشد. شکل 37 چند نمونه از متون تایپی و دستنویس را نشان می‌دهد.

9-4-1-2- داده برخط و برون‌خط

اولین مرحله در هر سیستم شناسایی متن، دریافت داده و تبدیل آن به شکل دیجیتالی قابل استفاده در کامپیوتر است. این سیستم‌ها می‌توانند متن ورودی را به دو شکل برخط و یا برون‌خط دریافت کنند. در سیستم‌های شناسایی متون برخط، کاربر با استفاده از ابزارهای مخصوص مانند قلم و صفحه دیجیتال

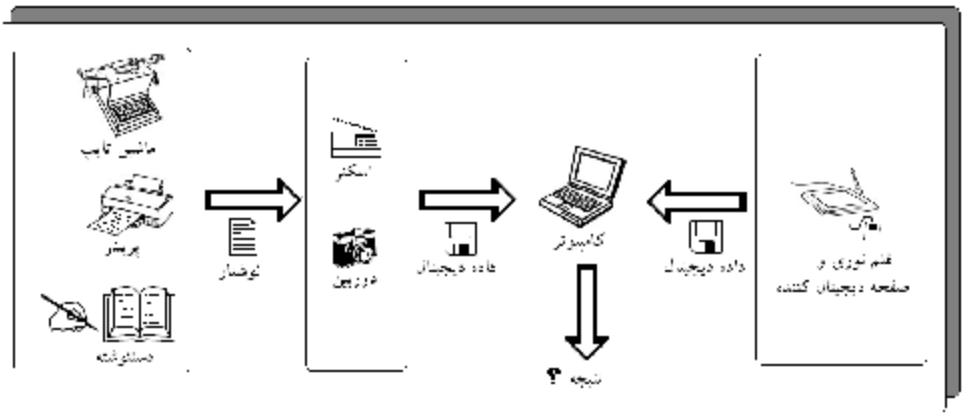
کننده اقدام به نوشتن متن می‌کند. در این گونه از نوشتار، اطلاعات زیادی در مورد متن، مثل تعداد *پاره‌خط‌های کلمه*، ترتیب ترسیم پاره‌خط‌ها و جهت و سرعت نوشتن در دسترس است. در روش برخط، عمل شناسایی نوشتار بصورت همزمان با عمل نوشتن انجام می‌شود. با توجه به خصوصیت ذاتی سیستم‌های برخط در نحوه جمع‌آوری دادگان ورودی، کاربرد آن‌ها به شناسایی متون دستنویس مانند کاراکترهای مجزا، کلمات پیوسته، فرمول‌ها یا عبارات ریاضی محدود می‌شود.



شکل 37- نمونه‌هایی از گونه‌های مختلف نوشتارهای فارسی

(الف): تایپی. (ب): دستنویس نستعلیق. (ج): دستنویس معمولی

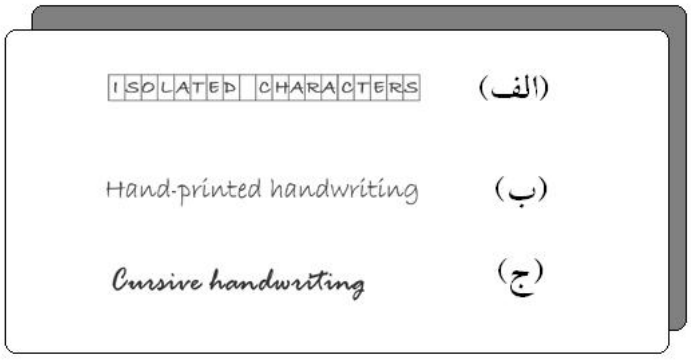
در مقابل، سیستم‌های برون‌خط وظیفه شناسایی متونِ از قبل نوشته و یا تایپ شده را برعهده دارند. بنابراین، این سیستم‌ها به اطلاعات زمانی موجود در دادگان برخط دسترسی نخواهند داشت. در این سیستم‌ها، از یک قلم و کاغذ معمولی برای نگارش استفاده می‌شود. سپس این متون نوشته شده و یا تایپی بوسیله یک دوربین یا پویشگر به صورت دیجیتالی تبدیل می‌شوند. شکل 38 وسایل مورد نیاز و چگونگی جمع‌آوری اطلاعات را در دو سیستم شناسایی برخط و برون‌خط نمایش می‌دهد.



شکل 38- نمای کلی از سیستم‌های شناسایی برخط و برون‌خط

9-4-1-3- داده مقید و بدون قید

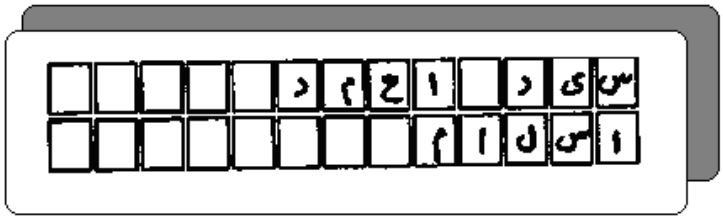
نوشتار دستنویس بر حسب میزان پیروی نویسنده از قواعد رسم‌الخط زبان و خوانا بودن، خود به دو دسته نوشتار دستنویس مقید و بدون قید تقسیم می‌شود. در برخی از کاربردها به منظور سهولت در تقطیع یک کلمه و یا کاهش تنوع در نحوه نگارش، نویسنده مجبور است که حروف را بطور جداگانه در محل‌های مخصوص (برای داده برون‌خط) و یا با رعایت قوانین خاص (برای داده برخط) بنویسد. شکل 39 سه نمونه متفاوت از کلمات دستنویس انگلیسی را نشان می‌دهد.



شکل 39- تقسیم‌بندی یک متن لاتین برحسب پیوستگی و میزان اعمال محدودیت در نحوه نگارش

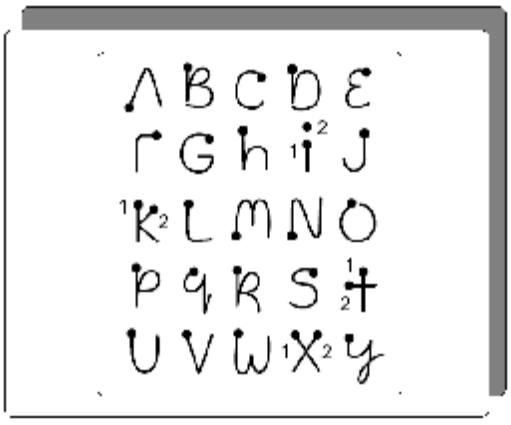
(الف): نوشتار دستنویس مقید با حرف مجزا. (ب): نوشتار دستنویس مقید با حرف گسسته(ج): نوشتار دستنویس بدون قید با حروف پیوسته

با توجه به پیوسته بودن ذاتی نوشتار فارسی، متداولترین محدودیت قابل اعمال برای کلمات دستنویس برون‌خط، نوشتن حروف آن‌ها بطور مجزا در داخل محل‌های خاص است(شکل 40).



شکل 40- نمونه‌ای از نوشتار مقید فارسی

در مورد داده‌های برخط، محدودیت‌های اعمالی بیشتر بر روی شروع نگارش از یک نقطه آغازین خاص و یا رعایت ترتیب ترسیم اجزای یک حرف مرکب مانند (t) که از دو بخش جداگانه تشکیل شده است، می‌باشد. در نوشتار مقید، گاهی محدودیت‌های نوشتاری بقدری زیاد است که کلمه و یا حرف نوشته شده کمترین شباهتی به نمونه واقعی آن ندارد. همانطور که در شکل 41 مشخص است، حروف (A) و (F) به نمونه‌های واقعی خود شباهت زیادی ندارند.



شکل 41- نمونه‌های قابل قبول در یک سیستم مقید شناسایی حروف برخط لاتین

9-4-1-4- ارقام و حروف گسسته

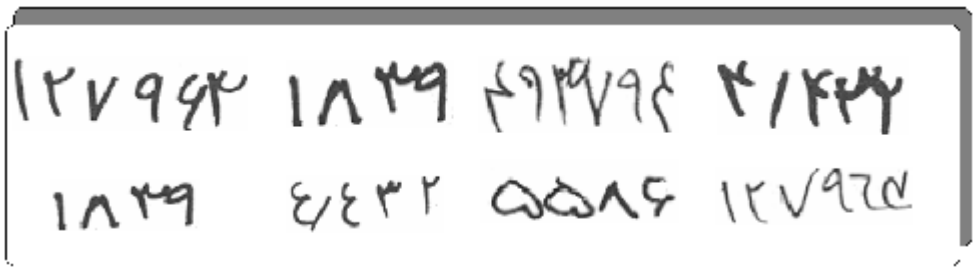
اعداد در نوشتار فارسی بطور ذاتی گسسته هستند. داشتن یک مجموعه داده از ارقام دستنویس برون‌خط در کاربردهایی مانند شناسایی کدهای پستی، تشخیص مبلغ چک‌های بانکی و غیره مفید خواهد بود.

اعداد برخط نیز در بسیاری از کاربردهای ورود اطلاعات به کامپیوتر از جمله تشخیص فرمول‌ها و عبارات ریاضی می‌توانند بکار روند. با توجه به اینکه بعضی از ارقام مانند رقم چهار، به دو صورت (۴ و 4) و یا شش به صورت (6 و ۶) نوشته می‌شوند، تعداد کلاس‌های درنظر گرفته شده برای ارقام معمولاً بین 12 تا 14 خواهد بود.

همانطور که قبلاً توضیح داده شده، علیرغم اینکه کلمات در نوشتار فارسی پیوسته هستند، برای بالا بردن درصد تشخیص یک سیستم شناسایی، می‌توان آن‌ها را با اعمال برخی محدودیت‌ها بصورت گسسته و جدا نوشت. این نوع داده مقید در کاربردهایی مانند درج کدپستی در نامه‌ها و یا نوشتن اسامی‌در فرم‌های ثبت نام رواج زیادی دارد. همانطور که قبلا توضیح داده شد، برخی از حروف فارسی می‌توانند بر حسب موقعیت، تا چهار شکل مختلف داشته باشند. همچنین با درنظر گرفتن کلاس‌های مجزا برای ادغام حروف مجاور، تعداد کلاس‌های لازم برای نویسه در حدود 160 کلاس خواهد بود.

9-4-1-5- رشته‌ای از ارقام

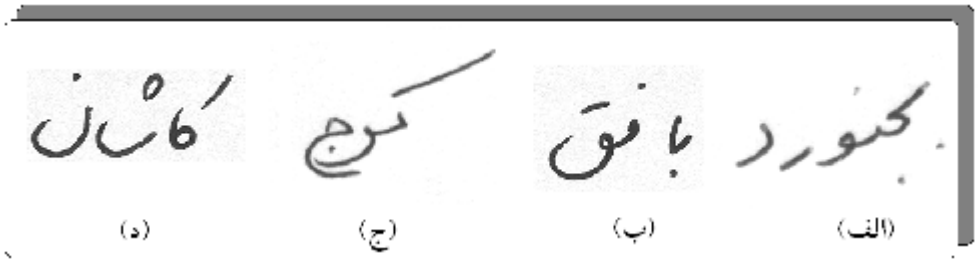
اگرچه ارقام بطور طبیعی جدا از هم نوشته می‌شوند، اما در برخی کاربردها مانند تشخیص کدهای پستی بدون قید که شامل 10 رقم هستند، با رشته‌ای از ارقام مواجه هستیم. تجربه نشان می‌دهد که در این کاربردها، ارقام انتهای رشته ممکن است به هم متصل شوند (شکل 42). این اتصال ارقام، محققان را برای جداسازی ارقام به مبارزه می‌طلبد. وجود مجموعه‌ای از رشته‌های عددی در بسیاری از کاربردهای عملی مفید خواهد بود.



شکل 42- نمونه‌هایی از ارقام به هم چسبیده در رشته‌ای از ارقام

9-1-4-6- زیر-کلمات

در نوشتار فارسی هر کلمه از یک یا چند زیرکلمه تشکیل می‌شود. زیر-کلمات نیز به نوبه خود از حروف تشکیل می‌شوند. با درنظر گرفتن زیرکلمه به عنوان یکی از اجزاء کلمه، برخی از محققان بر این باور هستند که در صورت داشتن مجموعه‌ای از زیر-کلمات رایج و ترکیب آن‌ها، می‌توان محدوده وسیعی از کلمات را در اختیار داشت. در ارتباط با مجموعه داده‌های متشکل از زیر-کلمات، باید دو مورد را مد نظر داشت. اولاً، سیستم‌های شناسایی مبتنی بر مجموعه داده‌های متشکل از زیر-کلمات، این پیش فرض را دارند که الگویتم‌های تقطیع می‌توانند کلمه ورودی را بدرستی به زیر-کلمات آن تجزیه کنند. ثانیاً، باید توجه داشت که شکل حروف در یک کلمه تابعی از حروف مجاور آن است. به عبارت دیگر، یک زیرکلمه خاص در دو کلمه متفاوت می‌تواند دارای شکل‌های کاملاً مختلفی باشد (شکل 43). با توجه به این مطالب، می‌توان گفت که مجموعه داده‌های متشکل از زیر-کلمات در اولویت دوم در تهیه مجموعه داده‌ها می‌باشد.



شکل 43- شکل‌های مختلف حروف یکسان در کلمات متفاوت

(الف) و (ب): شکل‌های متفاوت حرف "ب" در کلمات (بجنورد) و (بافق)

(ج) و (د): شکل‌های متفاوت حرف "ک" در کلمات (کرج) و (کاشان)

9-1-4-7- کلمات

از آنجائیکه کلمات یکی از کوچکترین واحدهای نوشتاری با معنی محسوب می‌شوند، مجموعه داده‌های مشتمل بر آن‌ها، دارای کاربرد بیشتری از حروف مجزا و یا زیر-کلمات هستند. مجموعه کلمات در کاربردهایی مانند شناسایی اسامی‌افراد و یا تشخیص نام شهرها می‌تواند مفید باشد.

9-4-1-8- جملات و پاراگراف‌ها

از ترکیب کلمات، جمله و از کنار هم قرار گرفتن جملات، پاراگراف ایجاد می‌شود. هدف نهایی سیستم‌های شناسایی، تشخیص جملات و پاراگراف‌های نوشته شده توسط هر نویسنده‌ای است. اگرچه سیستم‌های شناسایی متون فارسی قادر به تشخیص کلمات با درصد صحت قابل قبول هستند، اما هنوز با این هدف نهایی فاصله زیادی دارند. شاید بتوان گفت که محققان این رشته هنوز از پس شناسایی کلمات فارسی بخصوص در یک مجموعه بزرگ بطور کامل بر نیامده‌اند. بنابراین، جمع‌آوری مجموعه‌ای از جملات و پاراگراف‌ها، پس از کلمات، در درجه دوم اهمیت قرار دارد.

در مورد داده‌های برخط، با توجه به اینکه نویسنده در هنگام نوشتن، متن نوشته شده را ممکن است بدرستی مشاهده ننماید، احتمال اینکه جملات و پاراگراف‌های نوشته شده با هم تداخل داشته باشند و یا با شکل عادی نوشتار وی تفاوت زیادی داشته باشند، بسیار زیاد است. بنابراین، مجموعه داده‌های برخط تاکنون بیشتر به مجموعه اعداد، حروف گسسته و یا کلمات محدود شده‌اند.

9-4-2- اطلاعات مشخصه داده‌ها

میزان مفید بودن یک مجموعه داده به دو پارامتر اساسی زیر بستگی دارد:

1) داده‌ها
2) اطلاعات مشخصه آن‌ها

علاوه بر کیفیت داده‌های ذخیره شده و میزان تنوع آن‌ها، ساختار اطلاعات مشخصه این داده‌ها نیز می‌تواند در کیفیت یک مجموعه داده تأثیر گذار باشد.

اولین قدم در تهیه یک مجموعه داده، جمع‌آوری داده‌های مرتبط با کاربرد مورد نظر است. اما قبل از انجام عملیات برچسب‌زنی که معمولاً بسیار زمان‌بر است، باید ساختار داده‌ها برای اطلاعات مشخصه و برچسب‌ها از قبل تعیین گردد.

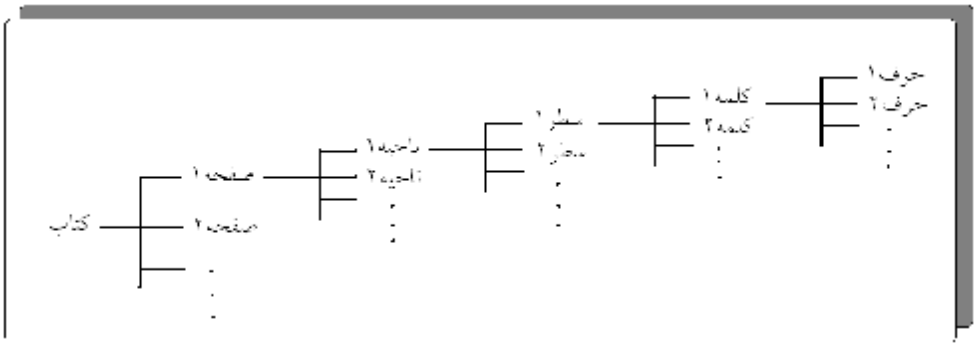
در کاربردهای شناسایی الگو، اصطلاح Ground Truth را اطلاعات مشخصه می‌نامیم. این اطلاعات نشان می‌دهند که هر یک از الگوها به کدام کلاس تعلق دارند. از اطلاعات مشخصه برای کاهش خطا در هنگام استفاده از الگوریتم‌های آموزش با نظارت استفاده می‌شود. همچنین برای تعیین دقت یک سیستم شناسایی در مرحله آزمایش، از این اطلاعات استفاده می‌شود.

ساختار اطلاعات مشخصه باید بطور اجمالی دارای ویژگی‌های زیر باشد:

- خواندن آن تا حد امکان ساده باشد.
- منطبق با روند تولید داده برای تمام مراحل یک سیستم شناسایی باشد.
- درصورت نیاز ساختار منطقی و فیزیکی متن را توصیف کند.
- قابلیت ذخیره سازی تمام اطلاعات لازم را داشته باشد.
- انعطاف پذیر بوده و در دو زمینه ساختار داده و اندازه مجموعه داده‌ها قابل تعمیم باشد.
- نباید حجم زیادی را اشغال کند.

ساختار *سلسله مراتبی* یا درختی به دلیل سادگی و قابلیت گسترش آسان، یکی از متداول‌ترین ساختارها برای ذخیره سازی و ارائه داده‌های متنی است. در این ساختار، متن به زیرمجموعه‌هایی با نوع و سطح اهمیت مختلف تقسیم می‌شود. سپس هر یک از این مجموعه‌ها خود به مجموعه‌های کوچکتری تقسیم می‌شوند. فرض کنید که می‌خواهیم یک مجموعه داده از متون موجود در یک کتاب را با ساختار سلسله مراتبی توصیف کنیم. با توجه به اینکه کتاب از چند صفحه، هر صفحه از چند ناحیه، هر ناحیه از چند سطر، هر سطر از چند کلمه و هر کلمه از چند حرف تشکیل شده است، در اینجا کتاب به عنوان ریشه اصلی این درخت محسوب خواهد شد. صفحات، نواحی، سطرها، کلمات و حروف به عنوان شاخه‌های بعدی این درخت خواهند بود (شکل 44).

در ادامه جزئیات مربوط به فایل‌های مشخصه برای داده‌های برون‌خط و برخط بطور جداگانه بررسی می‌شوند.



شکل 44- توصیف سلسله مراتبی یک کتاب

9-4-2-1- اطلاعات مشخصه داده‌های برون‌خط

مجموعه داده‌های برخط اغلب در سطح کلمه یا جمله هستند اما مجموعه داده‌های برون‌خط می‌توانند حتی تا سطح صفحه و کتاب نیز پیش بروند. با توجه به اینکه مجموعه داده‌های برون‌خط معمولاً نسبت به مجموعه داده‌های برخط گستردگی بیشتری دارند، ساختار داده‌های ذخیره شده در آن‌ها پیچیده‌تر است.

با توجه به مثال شکل 44 اطلاعات مشخصه برای هر یک از سطوح فوق (صفحه، ناحیه، سطر، ...) از داده‌های برون‌خط می‌تواند شامل موارد زیر باشد.

1) اطلاعات مشخصه در سطح کتاب
   - تعداد صفحات
2) اطلاعات مشخصه در سطح صفحه
   - تعداد نواحی
   - نوع هر ناحیه (متن، جدول، شکل و...)
   - مختصات هر ناحیه در صفحه
3) اطلاعات مشخصه در سطح ناحیه

- تعداد خطوط در هر ناحیه
- نوع هر ناحیه (دستنویس، تایپی، ترکیبی)
- اطلاعات مشخصه در سطح سطر
- تعداد کلمات
- مختصات خط کرسی
- نوع سطر (دستنویس، تایپی، ترکیبی)
- اندازه قلم در حالت تایپی

4) اطلاعات مشخصه در سطح کلمه

- تعداد حروف
- توالی حروف
- تعداد زیر-کلمات
- موقعیت و شماره کلمه در سطر
- نوع کلمه (دستنویس، تایپی)
- نوع فونت در حالت تایپی (معمولی، *ایتالیک*، پر رنگ ، *مایل-ضخیم* و ...)
- اندازه فونت در حالت تایپی
- موقعیت خط کرسی

5) اطلاعات مشخصه در سطح حرف

- شماره حرف در کلمه
- نوع حرف (مجزا، ابتدا، وسط، انتها)
- کد حرف

9-4-2-2- اطلاعات مشخصه داده‌های برخط

1) اطلاعات مشخصه داده‌های برخط معمولاً شامل موارد زیر است:
2) ابعاد صفحه دیجیتال کننده
3) نرخ نمونه برداری صفحه دیجیتال کننده
4) نام و مدل صفحه دیجیتال کننده
5) کلمه نوشته شده
6) توالی حروف موجود
7) مشخصه نویسنده
8) تعداد پیکسل‌های نویسه
9) مختصات پیکسل ابتدا و انتهای هر *پاره خط*
10) تعداد پاره خط‌ها
11) مختصات نقاط بدون برخورد قلم با صفحه
12) فشار و زاویه قلم
13) محل گذاشت و برداشت قلم
14) فاصله زمانی بین گذاشت و برداشت قلم

9-5- مجموعه داده تایپی برای نویسه‌خوان نوری زبان فارسی

9-5-1- مجموعه داده اسناد

تصاویر موجود در این بانک، تصاویر اسکن شده از نمونه‌های کاغذی چاپ شدهٔ واقعی موجود است. مهمترین اصل در تهیهٔ این بانک واقعی بودن نمونه‌ها است که واقعی بودن نمونه‌ها معمولاً به الگوریتم‌ها کمک می‌کند که در فضاهای کاربردی بهتر عمل کرده و در عمل کارائی بیشتری داشته باشند. از طرفی کار جمع‌آوری نمونه‌های واقعی و تهیهٔ اطلاعات مشخصه آن‌ها مشکل‌تر است. یکی از تفاوت‌های این بانک با بانک‌هائی که تصاویر آن‌ها به شکل مصنوعی بوجود می‌آیند این است که در تصاویر این بانک تأثیر رنگ، جنس کاغذ، نوع مرکب، نوع چاپ و دیگر نویزهای معمول وجود داشته و لازم نیست برای به وجود آوردن آن‌ها کار خاصی انجام شود. تنها کاری که لازم است انجام شود جمع‌آوری نمونه‌ها با دقت است.

مواردی که در جمع‌آوری نمونه‌های این مجموعه داده اهمیت دارند عبارتند از :

1) تنوع فونت
2) اندازه و حالت فونت
3) وجود پیچیدگی‌های معمول تصویری (مانند کثیف بودن، چرخش، رنگ زمینه و غیره)
4) تنوع منبع اصلی مانند کتاب‌های جدید، کتاب‌های قدیمی، روزنامه، مجله و غیره
5) تنوع در ابزارهای مختلف چاپ مانند (چاپ افست، چاپ لیزری، چاپ سربی، چاپ سوزنی، فاکس، تایپ، کپی و غیره).

نوع چاپ سند نیز باید در فایل‌های مشخصه آورده شود. برای دسترسی بهتر به این مجموعه، فایل‌های این بانک باید با توجه به نوع چاپ (چاپ‌های جدید، سربی، تایپ و غیره) تفکیک گردد.

9-5-2- مجموعه داده زیرکلمات

زیرکلمات در زبان فارسی اصطلاحاً به حروف به هم چسبیدهٔ معتبر همراه با اجزاء وابستهٔ آن‌ها مانند نقاط اطلاق می‌گردد. برای مثال کلمهٔ "بانک" از چهار حرف تشکیل شده و دارای دو زیرکلمه "با" و "نک" می‌باشد.

برای تهیهٔ این بانک از زیرکلمات 1 تا 9 حرفی استفاده می‌گردد که این زیرکلمات باید از یک پیکره متنی بزرگ استخراج گردد. بزرگترین پیکره متنی شناخته شده فارسی [31] است که حجم تقریبی آن 100 میلیون کلمه است. با توجه به حجم بالای این پیکره تقریباً تمام زیرکلمات موجود زبان فارسی از آن قابل استخراج خواهد بود. توسط زیرکلمات استخراج شده، متنی که شامل همه زیرکلمات باشد تهیه می‌شود. این متن یک رشته از تمام زیرکلمات است. سپس با توجه به تنوع قلم‌ها، اندازه و حالت‌های مختلفِ معمولی، پررنگ، ایتالیک و پررنگ ایتالیک توسط روش‌های مختلف چاپ، روی کاغذ چاپ می‌شوند. بعد از تهیهٔ نمونه‌های کاغذی، توسط روبشگر، تصاویری تهیه می‌گردد.

9-5-3- مجموعه دادهPDF

در بین منابع الکترونیکی موجود، PDF بهترین انتخاب است. چون قالب آن یک قالب باز است و ویژگی‌های این قالب به گونه‌ایی است که اطلاعات مناسبی از محتوا را در اختیار قرار می‌دهد. از میان ویژگی‌های خاص PDF می‌توان به امکان استخراج کادر دور نویسه‌ها از این قالب اشاره نمود. ساختار فایل‌های PDF به گونه‌ای هستند که می‌توان اطلاعات با ارزش و جامعی را در مورد تصویر متناظرشان به‌دست آورد. از این رو برخلاف دیگر بانک‌ها، فایل‌های مشخصه تصاویر این بانک فایل‌های PDF هستند تا کاربرانی که به اطلاعات بیشتر احتیاج دارند، بتوانند اطلاعات مورد نیاز خود را از فایل‌های PDF استخراج نمایند. این بانک شامل تصاویر اسکن شده از چاپ‌های لیزری همراه با منبع‌های PDF آن‌ها است.

ابتدا فایل‌های PDF از منابع مختلف تهیه می‌گردند. سپس این فایل‌ها ویرایش شده و علامت‌هایی در مکان‌های مشخص به آن‌ها اضافه می‌گردد. این علامت‌ها کمک خواهند کرد تا تابع تبدیل (مانند ماتریس

تبدیل Affine) اطلاعات موجود در فایل‌های PDF به تصویر اسکن شده آن‌ها، قابل محاسبه شود. بعد از ویرایش فایل‌ها، نسخهٔ چاپی آن‌ها توسط چاپگر لیزری تهیه می‌گردد. پس از آن اسناد کاغذی بوجود آمده اسکن شده و تصاویر متناظر فایل‌های PDF بوجود می‌آیند. البته از فایل‌های PDF می‌توان به سادگی تصاویر بدون نویز با dpiهای متفاوت را نیز استخراج نمود که در این صورت تصاویر بدون اعوجاج بدست می‌آید و دیگر به علامت‌های اضافه شده در تصاویر نیاز نخواهد بود.

9-5-4- مجموعه داده ساختار

هدف اصلی از تهیه بانک نویسه‌ها تأمین نمونه‌های لازم جهت آموزش نویسه‌های فارسی است. اگر چه در این بانک اطلاعات مربوط به ساختار فیزیکی نیز موجود می‌باشد ولی به دو دلیل ممکن است جوابگوی الگوریتم‌های تحلیل ساختار اسناد نباشد. دلیل اول حجم اطلاعات است. هر صفحه از بانک نویسه‌ها شامل هزاران نمونه از نویسه‌های فارسی می‌باشد ولی هر صفحه آن فقط ارزش یک رکورد را برای تحلیل ساختار سند دارد. اگر برای بالا بردن نمونه‌های لازم تحلیل ساختار به حجم این بانک اضافه شود، هم زمان و هزینه زیادی صرف خواهد شد و هم تعادل بانک به هم خواهد خورد. دلیل دوم پوشش دادن به پیچیدگی‌های لازم در ساختار سند است. هدف اصلی در بانک نویسه‌ها، تأمین نمونه‌های مختلف از نویسه‌ها است، نه نمونه‌های مختلف ساختار سند. بنابراین ممکن است ساختار اسناد بکار رفته در بانک نویسه‌ها از پیچیدگی‌های لازم برخوردار نباشند.

بنابراین برای تأمین نمونه‌های لازم جهت تحلیل ساختار اسناد، بانک ساختار در نظر گرفته شده است که تصاویر این بانک مانند بانک واقعی از نمونه‌های کاغذی چاپ شدهٔ موجود تهیه می‌گردند. با این تفاوت که فایل‌های مشخصه آن‌ها شامل اطلاعات متنی نمی‌شود و فقط اطلاعات مربوط به نواحی مختلف، نوع آن‌ها و دیگر مشخصات لازم نواحی را دارا می‌باشد.

9-5-5- مجموعه داده‌های جداول و فرم‌ها

با توجه به تنوع جداول و فرم‌ها و پراکندگی آن‌ها در اسناد و اهمیت بازشناسی آن‌ها از یک طرف و کمبود تنوع و تعداد آن‌ها در دیگر بانک‌های تشریح شده از طرف دیگر، بنظر می‌رسد وجود این بانک ضروری باشد. در جمع‌آوری این بانک سعی خواهد شد انواع جداول و فرم‌های رایج در اسناد، که به شکل واقعی چاپ شده و وجود دارند (مانند بانک‌های نویسه‌ها و ساختار) جمع‌آوری گردند. پس از جمع‌آوری، تصاویر آن‌ها توسط روبشگر تهیه و در بانک ذخیره می‌گردد. فایل‌های مشخصه این نوع از تصاویر در حد بانک ساختار است. اطلاعات مشخصه مربوط به محتوای جداول در این فاز آورده نمی‌شود و اطلاعات مشخصه محدود به تعیین چهارچوب جداول می‌باشد.

9-5-6- حجم مجموعه داده

حجم مجموعه داده مطابقجدول 13 درنظر گرفته شده است.

جدول 13- حجم مجموعه داده

ردیف	نام بانک	نوع فایل‌های مشخصه‌ها	حجم (صفحه)
1	بانک اسناد	*RGT*[1], *StdGT*[2] ,*StrGT*[3]	10000
2	بانک ساختار	RGT, StdGT, StrGT	2000
3	بانک زیرکلمات	RGT, StdGT, StrGT	10.000
4	بانک PDF	PDF[4]	2000

[1]. Raw Ground Truth

[2]. Standard Ground Truth

[3]. Structural Ground Truth

[4]. Portable Document Format

10 - نرم افزارهای ارزیابی

10-1- نرم‌افزار PinkPanther

PinkPanther [32] نرم‌فزاری است برای تولید فایل‌های توصیف‌گر ساختار سند و همچنین سنجش کارایی تحلیل ساختار سند. کارآیی تحلیل ساختار سند، با اجرا کردن الگوریتم تحلیل ساختار سند روی یک مجموعه از تصاویر اسناد و مقایسه خروجی آن‌ها با توصیف‌گر متناظرشان صورت می‌گیرد. PinkPanther از دو قسمت مجزا تشکیل شده است: Ground-Keeper و Cluzo. Ground-Keeper ابزاری برای ایجاد فایل‌های توصیف‌گر است. Ground-Keeper تصویر سند و فراداده‌های آن را نمایش داده و اجازه می‌دهد که نواحی مختلف آن را مشخص و ترسیم کرده و اطلاعات مربوط به هر ناحیه را وارد نمود. فایل‌های توصیف‌گر این نرم افزار در قالب ASCII ذخیره می‌گردد. Cluzo یک ابزار ارزیابی است که مکان، نوع و دقت نواحی را جمع‌آوری کرده و مقدار خطا و کارآئی را محاسبه می‌نماید. PinkPanther در محیط Unix/X Windows به زبان C پیاده‌سازی شده است. نکته قابل توجه در مورد این نرم‌افزار این است که این نرم‌افزار امکان وارد نمودن اطلاعات را در سطح نواحی می‌دهد اما امکان وارد نمودن اطلاعات متنی نواحی در آن فرآهم نیست.

10-2- نرم‌افزار Illuminator

Illuminator [33] یک ویرایشگر است که توسط شرکت RAF برای تهیه کردن مجموعه‌های آموزش و آزمون فهمیدن سند برای تصحیح خطاهای نویسه‌خوان نوری (نویسه‌خوان نوری)، و برای جمع‌آوری و نگهداری اطلاعات و ویژگی‌های تصویر اسناد تهیه شده است. Illuminator یک نمایشگر و ویرایشگر تصاویر و توصیف‌گرهای آن می‌باشد. این نرم‌افزار قابلیت کار با متن اغلب زبان‌های اروپایی و ژاپنی را دارا است و برای ذخیره‌سازی تصویر و توصیف‌گرها از قالب *DAFS* [34] استفاده می‌نماید. Illuminator در محیط Unix / X Windows و با زبان C پیاده‌سازی شده است.

10-3- مرورگر مجموعه داده دانشگاه Oulu

دانشگاه Oulu یک مجموعه داده شامل تصاویر اسکن شده همراه با فایل‌های توصیف‌گر ساختار فیزیکی و منطقی را تهیه نموده است. در کنار این مجموعه داده ابزاری برای مرور مجموعه داده وجود دارد که امکان مشاهده هم‌زمان فایل‌های تصویری مجموعه داده و فایل‌های توصیف‌گر را داده و در یک پنجره ویژگی‌های سند را نمایش می‌دهد. این نرم‌افزار به زبان java تهیه شده است. فایل‌های تصویری در قالب jpeg و فایل‌های توصیف‌گر در قالب ASCII ارائه میگردند.

PinkPanther و Illuminator فقط در محیط Unix اجرا می‌شوند و از این بابت دارای محدودیت هستند ولی ابزار Oulu با زبان Java برنامه‌نویسی شده است و در محیط‌های مختلفی اجرا می‌گردد. اما از لحاظ قالب فایل‌های گرافیکی این نرم‌افزار فرمت jpeg را پشتیبانی می‌کند. این درحالی است که فرمت اکثر تصاویر مربوط به اسناد، tiff است. علاوه براین فایل‌های توصیف‌گر این ابزارها غیر استاندارد است. در حقیقت همگی آن‌ها اطلاعات توصیفی را در یک ساختار خاص نگهداری می‌نمایند که این کار زیاد مطلوب نبوده و بهتر است برای ایجاد سازگاری بیشتر با محیط‌ها و نرم‌افزارهای دیگر از یک قالب استاندارد استفاده شود.

11 - ارزیابی نویسه‌خوان نوری

برای ارزیابی کارایی سیستم‌های بازشناسی متن معیارهای متداولی وجود دارد که مهم‌ترین آن‌ها به قرار ذیل می‌باشد:

1) نرخ بازشناسی: تعداد حروف، زیر- کلمات یا کلماتی که به درستی بازشناسی شده‌اند به تعداد کل آن‌ها در یک بلوک متن را نرخ بازشناسی می‌گویند.
2) نرخ وازدگی: تعداد حروف، زیر- کلمات یا کلماتی که سیستم اعلام می‌کند قادر به بازشناسی آن‌ها نیست، به تعداد کل آن‌ها در یک بلوک متن را نرخ وازدگی می‌گویند.
3) نرخ خطا: تعداد حروف، زیر- کلمات یا کلماتی که اشتباه بازشناسی شده‌اند به تعداد کل آن‌ها در یک بلوک متن را نرخ خطا می‌گویند.

12 - نتیجه گیری

زمانی در میانه‌ها و اواخر دهه 70 تلاش‌های گسترده‌ی تحقیقاتی و تجاری در این خصوص انجام گرفت ولی پس از مدتی بسیاری از تلاش رها شد. طرح‌های تحقیقاتی زیادی دراین مدت درعرصه‌ی آکادمیک ایران ارائه شد ولی هیچ یک توفیق تجاری شدن نیافته است. برخی شرکت‌های تجاری هم که با حمایت‌های دولتی – نظیر طرح تکفا – به فعالیت دراین خصوص مشغول بودند پس از مدتی با قطع حمایت‌ها , این فعالیت‌ها را متوقف کردند. نمونه‌ی آن شرکت دوران نوین بود که درسال 82 با حمایت طرح تکفا به مطالعه درخصوص نویسه‌خوان نوری مشغول شد و چندی بعد اعلام کرد که به نرم افزاری با دقت بیش از 90% دست یافته است ولی اکنون در پایگاه اطلاع رسانی این شرکت هیچ نشانی – حتی ازتحقیقات صورت گرفته – نیست . به هرحال در حال حاضر نرم افزار تجاری نویسه‌خوان نوری فارسی توسط شرکت‌های Sakhr و Zylab عرضه شده است اما این نرم افزارها نیزهنوز تا حالت بهینه فاصله‌ی زیادی دارند و البته هزینه‌ی آنها نیزبسیار بالاست (برای مثال نرم افزار اتوماتیک شرکت صخره با قیمت‌های 1400 دلار عرضه می شوند و استفاده از آنها مقرون به صرفه نیست. ضمن اینکه هیچ یک از این نرم افزارها از متون دست نویس پشتیبانی نمی‌کنند. به هر حال عرصه‌ی نویسه‌خوان نوری با توجه به ضعف‌های نرم افزارهای موجود که ظاهرا به ضعف تحقیقاتی و تئوریک آنها برمی‌گردد و همچنین عدم پشتیبانی از این نرم افزارها از متون دست نویس، عرصه ایست که همچنان می‌بایست مسیر دستیابی به یک نرم افزار بهینه و کارا و با هزینه مناسب را طی نماید. از سوی دیگر ارتباط هوش مصنوعی با تکنیک‌های نویسه‌خوان نوری و رشد روز به روز این شاخه, این امکان را فرآهم آورده که هر روز تحقیقات جدیدتری دراین خصوص صورت بگیرد.

مراجع

[1] O. D Trier, A. K. Jain, and T. Taxt, "Feature extraction methods for character recognition - Asurvey," *Pattern Recognition*, vol. 29, pp. 641-662, 1996.

[2] ب. . تیمساری " بازشناسی حروف در کلمات تایپ شده فارسی با استفاده از روش مورفولوژی," دانشکده مهندسی برق , دانشگاه صنعتی اصفهان, 1371.

[3] B. Badr, and A. M. Saberi, "Survey and Bibligraphy of Arabic Optical Text Recognition," *Elsvier Signal Processing*, pp. 49-77, 1995.

[4] S. R. Srihari, "High Performance Reading Machine," in Proceedings of IEEE 1992, pp. 1120-1132.

[5] ا. ابراهیمی, ا. کبیر, "بازیابی تصاویر نامه های اداری "،سومین کنفرانس ماشین بینایی و پردازش تصویر ایران, 1383،، صفحه 24-29.

[6] D. Doermann, "The Indexing and Retreival of document Images: a Survey," *Computer Vision and Image Understanding*, vol. 70, no. 3, pp. 287-298, 1998.

[7] H. Khosravi, and E. Kabir, "Introducing a very large dataset of handwritten Farsi digits and a study on their varieties," *Pattern Recognition Letters*, vol. 28, pp. 1133-1141, 2007.

[8] L. Eikvil, "OCR: Optical Character Recognition," *Norsk Regnesentral*, 1993.

[9] K. Badie, and M. Shimura, "Machine recognition of Arabic cursive scripts," *Pattern Recognition in Practice*, pp. 315-323, 1980.

[10] K. Badie, and M. Shimura, "Machine recognition of Arabic handprinted scripts," *Trans. of IECE*, vol. E65, no. 2, pp. 107-114, 1982.

[11] B. Al-Badr, and S. A. Mahmoud, "Survey and bibliography of Arabic optical text recognition," *Signal Processing*, vol. 41, pp. 49-77, 1995.

[12] B. Parhami, and M. Taraghi, "Automatic recognition of printed Farsi Texts," *Pattern Recognition in Practice*, vol. 14, no. 1-6, pp. 395-401, 1981.

[13] ح. مقسمی, " تشخیص حروف تایپی فارسی با روش ساختاری " پایان نامه کارشناسی ارشد, دانشکده مهندسی کامپیوتر, دانشگاه علم و صنعت ایران, ۱۳۷۶.

[14] س .اسدی, ". بازشناسی حروف فارسی با استفاده از شبکه عصبی چند جمله ای های جداکننده " پایان نامه کارشناسی ارشد دانشکده مهندسی کامپیوتر, دانشگاه علم و صنعت ایران 1377.

[15] R. Azmi, and E. Kabir, "A New Segmentation Technique for Omnifont Farsi Tex," *Pattern Recognition Letters*, vol. 22, pp. 97-104, 2001.

[16] ر. عزمی, "پایان نامه دکتری بازشناسی متون چاپی فارسی," دانشکده مهندسی برق , دانشگاه تربیت مدرس, 1378.

[17] L. O'Gorman, and R. Kasturi, "Document Image Analysis," *IEEE Computer Society Press*, 1995.

[18] G. N. Nartker, A. Thomas, S. V. Rice *et al.*, "Optical Character Recognition: An illustrated guide to the frontier," Dept. of Electrical, Computer and Systems Engineering, Rensselaer Polytechnic Institute, 1999.

[19] م. رئیسی : "OCR ,آموزش الفبای فارسی به رایانه," ماهنامه دانشگر, 1383.

[20] ح. م. پور, " قطعه بندی برخط کلمات دست نویس فارسی," پایان نامه کارشناسی ارشد, دانشکده مهندسی برق , دانشگاه تربیت مدرس 1378.

[21] ا. فرامرزی, " بازشناسی نوری حروف : مروری بر مباحث نظری و ملاحظات کاربردی با تاکید بر مسائل خاص زبان فارسی," ماهنامه علوم اطلاع رسانی ، شماره 20، 1384

[22] D. Marcu, "The Theory and Practice of Discourse Parsing and Summarization.," *MIT Press*, 2000.

[23] "ERIM,http://documents.cfar.umd.edu/resources/database/ERIM_Arabic_DB.html.

[24] J. Sauvola, and H. Kauniskangas. "MediaTeam Document Database II, a CD-ROM collection of document images," http://www.mediateam.oulu.fi/downloads/MTDB/download.htm.

[25] N. Kharma, M. Ahmed, and R. Ward, "A new comprehensive database of handwritten Arabic words, numbers, and signatures used for OCR testing," *Proceedings of IEEE Canadian Conference on Electrical and Computer Engineering*, pp. 766-768, 1999.

[26] S. Mozaffari, K. Faez, F. Faradji *et al.*, "A Comprehensive Isolated Farsi/Arabic Character Database for Handwritten OCR Research," in Proceedings of 10th International Workshop on Frontiers in Handwriting Recognition, 2006, pp. 385-389.

[27] M. Pechwitz, S. Snoussi, V. Märgner *et al.*, "IFN/ENIT-database of handwritten Arabic words," in Proceedings of CIFED, 2002, pp. 129-136.

[28] N. B. Amara, O. Mazhoud, N. Bouzrara *et al.*, "ARABASE: A relational database for Arabic OCR systems," *in the Int. Arab Journal of Information Technology*, pp. 259-266, 2005.

[29] F. Solimanpour, J. Sadri, and C. Suen, "Standard databases for recognition of handwritten digits, numerical string, legal amounts, letters and dates in Farsi language," in Proceedings of 10th International Workshop on Frontiers in Handwriting Recognition, 2006, pp. 3-7.

[30] س. م. رضوی,, "بازشناسی برخط کلمات دستنویس فارسی," دانشکده مهندسی برق، دانشگاه تربیت مدرس, 1385.

[31] م. ب. ج. خان, "پیکرهٔ متنی زبان فارسی," اولین کارگاه پژوهشی زبان فارسی 1383.

[32] B. A. Yanikoglu, and L. Vincent, "Pink panter: A complete environment for ground-truthing and benchmarking document page segmentation," *Pattern Recognition*, vol. 31, pp. 1191-1204, 1998.

[33] T. Fruchterman, "Dafs: A standard for document and image understanding," in in Proceeding of Symposium on Document Image Understanding Technology, 1995, pp. 94-100.

[34] *DAFS Document Attribute Format Specification*, RAF Technology, Inc., Washington, 1994.