علوم اطلاع‌رسانی، دوره ۲۰، شماره ۳و۴ (بهار و تابستان ۱۳۸۴) ۳۳-۶۱

بازشناسی نوری حروف: مروری بر مباحث نظری و ملاحظات کاربردی با تأکید بر مسائل خاص زبان فارسی

اسماعیل فرامرزی

عضو هیئت علمی مرکز اطلاعات و مدارک علمی ایران

چکیده

در این مقاله مبانی نظری و جنبه‌های کاربردی مبحث بازشناسی نوری حروف (اُسی‌آر) بصورت جامع مورد بررسی قرار می‌گیرند و زیربخش‌ها و بلوک‌های پردازشی آن معرفی می‌گردند. همچنین خصایص و پیچیدگی‌های مختص نگارش زبان فارسی که یک نرم‌افزار «اُسی‌آر» باید آن‌ها را در عملیات پردازشی خود لحاظ نماید، بیان خواهند شد. تحقیقات داخلی انجام‌شده در زمینهٔ «اُسی‌آر» مورد اشاره قرار خواهند گرفت؛ نرم‌افزارهای معروف تجاری «اُسی‌آر» لاتین و فارسی معرفی، و قابلیت‌ها و نقاط قوت و ضعف آن‌ها تشریح می‌شوند. در آخر هم پیشنهادهایی در راستای انتخاب راهکارهای مناسب به منظور تسریع در حصول یک نرم‌افزار «اُسی‌آر» کارآمد برای زبان فارسی ارائه می‌گردد. مخاطب این مقاله، دانش‌آموختگان رشته‌های فنی و غیرفنی هستند که قصد دارند دربارهٔ این حوزه اطلاعات مقدماتی کسب نمایند. ازاین رو از جنبه‌های محاسباتی و ریاضیات مسئله چشم‌پوشی شده است.

کلیدواژه‌ها: بازشناسی نوری حروف («اُسی‌آر»[1])،تجزیه و تحلیل تصویر مدرک[2]، پردازش تصویر[3]، شناسایی الگوی آماری[4]، زبان فارسی

۱. مقدمه

قبل از آن که وارد مبحث «اُسی‌آر» شویم، لازم است اشارهٔ مختصری به حوزه‌های بازشناسی الگو و آنالیز تصویر اسناد (دی‌آی‌ای) داشته باشیم.

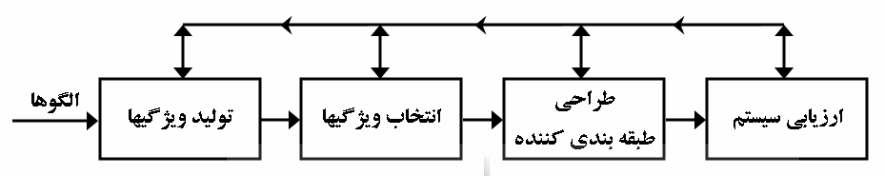
شناسایی الگو، شاخه‌ای از هوش مصنوعی[5] است که با طبقه‌بندی[6] و توصیف مشاهدات سروکار دارد. شناسایی الگو به ما کمک می‌کند داده‌ها (الگوها) را با تکیه بر دانش قبلی یا اطلاعات آماری استخراج‌شده از الگوها، طبقه‌بندی نماییم. الگوهایی که می‌بایست طبقه‌بندی شوند، معمولاً گروهی از سنجش‌ها یا مشاهدات هستند که مجموعه نقاطی را در یک فضای چند بعدی مناسب تعریف می‌نمایند.

نویسنده مکاتبه‌کننده. تلفن: ۶۶۴۹۴۹۸۰؛ نمابر: ۶۶۴۶۲۲۵۴

پست الکترونیکی: Faramarzi@irandoc.ac.ir

یک سیستم شناسایی الگوی کامل متشکل است از یک حسگر[۷] که مشاهداتی را که می‌بایست توصیف یا طبقه‌بندی شوند جمع‌آوری می‌نماید، یک سازوکار برای استخراج ویژگی‌ها[۸] که اطلاعات عددی یا نمادین را از مشاهدات، محاسبه می‌کند (این اطلاعات عددی را با یک بردار بنام بردار ویژگی‌ها[۹] نمایش می‌دهند)؛ و یک نظام طبقه‌بندی یا توصیف که وظیفهٔ اصلی طبقه‌بندی یا توصیف الگوها را با تکیه بر ویژگی‌های استخراج شده عهده‌دار است.

شکل ۱ نمودار بلوکی یک سیستم شناسایی الگو را نشان می‌دهد (Theodoridis، ۱۹۹۹). همانطوری که از پیکان‌های برگشتی مشخص است، این بلوک‌ها لزوماً مستقل نیستند و بسته به نتایج حاصله گاهی لازم است که بلوک‌های اولیه مجدداً طراحی گردند تا راندمان کلی سیستم بهبود یابد.

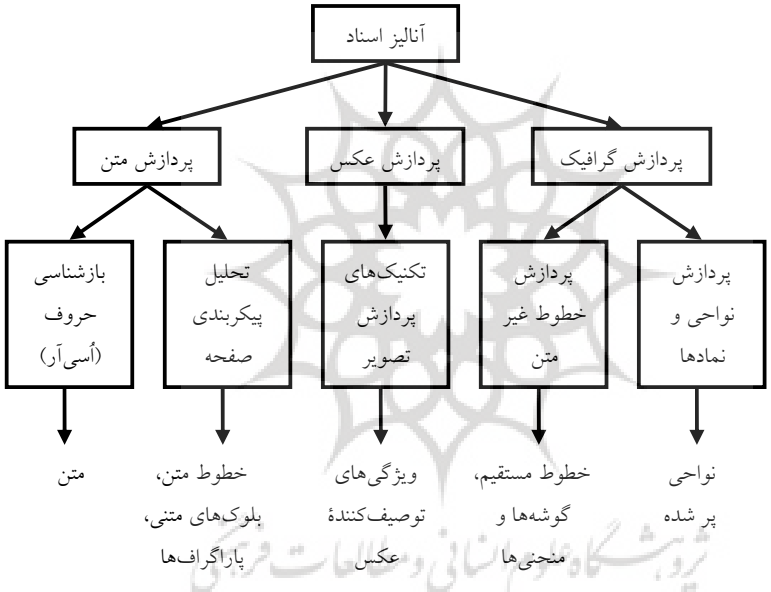


شکل ۱. نمودار بلوکی یک سیستم شناسایی الگو (Theodoridis، ۱۹۹۹)

امروزه حجم زیادی از اسناد کاغذی موجود، توسط اسکنرها یا دوربین‌ها به اسناد تصویری دیجیتالی تبدیل می‌شوند. ذخیره‌سازی، بازیابی و مدیریت کارآمد این آرشیوهای تصویری، در بسیاری از برنامه‌ها نظیر اتوماسیون اداری و کتابخانه‌های دیجیتالی اهمیت فراوانی دارند. در نتیجه دستیابی به الگوریتم‌های مؤثر به منظور آنالیز تصویری اسناد، یک نیاز اساسی به حساب می‌آید.

مبحث «آنالیز تصویر اسناد» (دی‌آی‌ای) از جمله شاخه‌های فعال در شناسایی الگو و پردازش تصاویر می‌باشد و مشتمل بر کلیهٔ مراحل پردازشی است که محتویات یک سند اسکن یا فکس شده را به یک فرم الکترونیکی مناسب، تبدیل می‌نماید. تکنیک‌های «دی‌آی‌ای» اجزای مختلف «ساختاری» سند، یعنی قسمت‌های متنی (پاراگراف‌ها، کلمات، حروف، ...)، قسمت‌های گرافیکی (خطوط، نمادها، نمودارها، ...) و قسمت‌های تصویری (تصاویر موجود در متن) را از یکدیگر تفکیک می‌کنند و پردازش مناسب را بر روی هر دسته از اجزا، اعمال می‌نمایند و نیز با توجه به ارتباط «منطقی» بین اجزای مختلف، نقش هر یک از این اجزا را در سند مشخص می‌سازند. شکل ۲ ساختار سلسله‌مراتبی «دی‌آی‌ای» را به نمایش می‌گذارد (O'Gorman، ۱۹۹۵).

همانگونه که شکل ۲ نشان می‌دهد، «دی‌آی‌ای» دربردارندهٔ دستهٔ بزرگی از تکنیک‌ها بنام تکنیک‌های «بازشناسی نوری حروف» (یا اُسی‌آر) است. این تکنیک‌ها در مورد اجزایی از تصویر سند که توسط تکنیک‌های تحلیل ساختاری در «دی‌آی‌ای» به عنوان متن تشخیص داده شده‌اند، اعمال می‌گردند و تصویر سند را به یک متن قابل ویرایش توسط رایانه تبدیل می‌نمایند. سیستم‌های اُسی‌آر با حذف نقش تایپیست‌ها در فرایند تبدیل اسناد کاغذی به قالب الکترونیکی، سرعت ورود اطلاعات به رایانه را ده‌ها برابر افزایش می‌دهند و روند انجام این فرایند را به میزان قابل توجهی تسهیل می‌کنند. امروزه بازار مصرف سیستم‌های «اُسی‌آر»، طیف بسیار وسیعی از مؤسسات (شامل مراکز نشر، دانشگاه‌ها، کتابخانه‌ها، بانک‌ها، ادارات پستی، شرکت‌های بیمه، و ...) را دربرمی‌گیرد. در نتیجه آشنایی اولیه با مبانی این سیستم‌ها برای کلیهٔ افرادی که به نحوی با اسناد و مدارک سر و کار دارند، ضروری به نظر می‌رسد.



شکل ۲. یک ساختار سلسله‌مراتبی از بخش‌های مختلف مبحث تحلیل

اکثر کارهای انجام‌شده در زمینهٔ «اُسی‌آر» در رابطه با متون لاتین، چینی و ژاپنی بوده است [عزمی، ۷۸؛ مسروری، ۷۹]و سیستم‌های تجاری «اُسی‌آر» لاتین در سال‌های اخیر پیشرفت

کیفی قابل ملاحظه‌ای داشته‌اند. اما «اُسی‌آر» فارسی با وجود حجم نسبتاً وسیع تحقیقات دانشگاهی و نیاز شدید بازار تجاری به آن، هنوز هم از جایگاه مورد نظر فاصلهٔ بسیاری دارد و تاکنون هیچ سیستم «اُسی‌آر» کارآمدی که از نظر دقت و کیفیت محیط نرم‌افزاری، قابل مقایسه با سیستم‌های «اُسی‌آر» لاتین باشد، عرضه نگردیده است. در نتیجه ضرورت انجام تحقیقات بیشتر در زمینهٔ متون فارسی و عربی کاملاً احساس می‌شود. بواسطهٔ وجود تفاوت‌های اساسی بین نحوهٔ نگارش فارسی و لاتین (نظیر چسبیده‌بودن حروف کلمه به یکدیگر، تغییر شکل حروف براساس موقعیت نسبی آن در کلمهٔ فارسی، و ...)، امکان اعمال مستقیم روش‌های بازشناسی متون لاتین به منظور شناسایی متون فارسی وجود ندارد.

نحوهٔ ارائهٔ مطالب در این مقاله بدین شرح است: بخش ۲ به معرفی سیستم «اُسی‌آر» می‌پردازد. تاریخچهٔ تحقیقات انجام‌شده در زمینهٔ «اُسی‌آر» در بخش ۳ مطرح می‌شود. سابقهٔ مطالعات بر روی «اُسی‌آر» فارسی در بخش ۴ مورد اشاره قرار خواهد گرفت. در بخش ۵ تحقیقات صورت‌گرفته در داخل کشور در زمینهٔ «اُسی‌آر» فارسی معرفی می‌شوند. در بخش ۶ مهم‌ترین ویژگی‌های نگارشی زبان فارسی که در طراحی یک سیستم «اُسی‌آر» فارسی باید لحاظ گردند بیان می‌شوند. بخش ۷ انواع سیستم‌های «اُسی‌آر» را از لحاظ الگوی ورودی تشریح می‌کند. مرور جامعی بر عملیات پردازشی سیستم‌های «اُسی‌آر» در بخش ۸ انجام خواهد پذیرفت. بخش ۹ به بررسی معروف‌ترین نرم‌افزارهای تجاری «اُسی‌آر» فارسی و لاتین، نحوهٔ عملکرد، و نقاط قوت و ضعف آن‌ها می‌پردازد. و بالاخره در بخش ۱۰ راهکارهایی به منظور تسهیل در روند دستیابی به یک سیستم «اُسی‌آر» کارآمد فارسی ارائه خواهد شد.

۲. معرفی بازشناسی نوری حروف

در چند دههٔ گذشته بازشناسی الگوهای نوشتاری شامل حروف، ارقام و دیگر نمادهای متداول در اسناد نوشته‌شده به زبان‌های مختلف، توسط گروه‌های مختلفی از محققین مورد مطالعه و بررسی قرار گرفته است. نتیجهٔ این تحقیقات منجر به پیدایش مجموعه‌ای از روش‌های سریع و تا حد زیادی مطمئن موسوم به «اُسی‌آر» یا «بازشناسی نوری حروف» به منظور وارد نمودن اطلاعات موجود در اسناد، مدارک، کتاب‌ها و سایر مکتوبات تایپی و حتی دست‌نوشت[۱۰] به داخل رایانه شده است. اصطلاح «اُسی‌آر» به تکنیک‌هایی اطلاق می‌شود که در تصاویر اسکن یا فکس شده، نواحی متنی را تشخیص می‌دهند و سپس این نواحی (تصویری) را به متن قابل ویرایش تبدیل می‌نمایند (Trier، ۱۹۹۶).

با دستگاهی به نام اسکنر[۱۱] می‌توان تصویر یک صفحه کاغذ را به صورت یک فایل گرافیکی (تصویری)، به رایانه ارسال و در آن ذخیره نمود. بدین ترتیب کاربر می‌تواند با یک نرم‌افزار مناسب نمایش‌دهندهٔ تصاویر، تصویر صفحهٔ اسکن‌شده را بر روی نمایشگر رایانهٔ خود ملاحظه نماید یا آن را چاپ کند؛ اما قادر نخواهد بود که متن موجود در تصویر سند را ویرایش کند یا آن را مورد جستجو قرار دهد. یک نرم‌افزار «اُسی‌آر»، تصویر اسکن‌شده را می‌خواند، محتویات آن (شامل متن، خطوط، تصاویر، جداول، ...) را شناسایی می‌نماید، و سپس آن را به یک قالب قابل ویرایش (در واژه‌پردازها) تبدیل می‌کند. امروزه بیشتر دستگاه‌های اسکنر به نرم‌افزارهای «اُسی‌آر» مجهز گردیده‌اند و قادرند متن موجود در یک سند اسکن‌شده را تشخیص دهند و آن را با همان نحوهٔ قالب‌بندی، ستون‌بندی، جدول‌بندی و نوع فونت مطابق با سند کاغذی اصلی، در قالب یک فایل متنی با قالب‌بندی مناسب ذخیره نمایند.

استفاده از سیستم‌های «اُسی‌آر» دو مزیت عمده دارد:

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

ب. کاهش فضای ذخیره‌سازی؛ زیرا حجم فایل متنی استخراج‌شده از یک تصویر، معمولاً بسیار کمتر از حجم خود فایل تصویری است.

چنین قابلیتی امکان استفادهٔ گسترده از رایانه را در پردازش سریع حجم وسیعی از داده‌های مکتوب شرکت‌ها و مؤسسات مختلف (نظیر بانک‌ها، شرکت‌های بیمه، مؤسسات خدمات عمومی، ادارهٔ پست، و دیگر نهادهایی که سالانه با میلیون‌ها مورد پرداخت، دریافت و حسابرسی امور مشتریان خود مواجه‌اند فراهم می‌آورد (تیمساری، ۱۳۷۱).

۳. تاریخچهٔ سیستم‌های «اُسی‌آر»

از جنبهٔ تاریخی، سیستم‌های «اُسی‌آر» تا کنون سه مرحلهٔ تکاملی را پشت سر گذاشته‌اند (Arica، ۲۰۰۱):

الف. مرحلهٔ تکوین (از ۱۹۰۰ تا ۱۹۸۰): رد پای اولیهٔ اقدامات صورت‌گرفته در زمینهٔ بازشناسی حروف را در سال‌های اول دههٔ ۱۹۰۰ می‌توان یافت و آن زمانی است که «تیورینگ»[۱۲] دانشمند روسی بر آن بود که به افراد مبتلا به نارسایی‌های بینایی کمک نماید (Mantas، ۱۹۹۶). اولین اختراع‌های ثبت‌شده در این زمینه مربوط به سال‌های ۱۹۲۹ و ۱۹۳۳ میلادی هستند (Mori، ۱۹۹۲). این سیستم‌ها حروف چاپی را با روش تطابق قالبی[۱۳] شناسایی می‌کردند؛ به این صورت که ماسک‌های مکانیکی مختلفی از مقابل تصویر حرف عبور می‌کردند

(مکانیکی) و نور از یک سو به آن تابانده می‌شد و از سوی دیگر توسط یک آشکارساز نوری دریافت می‌گردید (اپتیکی). وقتی یک انطباق کامل صورت می‌گرفت، نور به آشکارساز می‌رسید و حرف ورودی بازشناسی می‌شد. این اختراع به دلیل فناوری اپتومکانیکی مورد استفاده در آن، کاربردی نبود. تصور دسترسی به دستگاهی برای بازشناسی حروف تا دهه ۱۹۴۰ میلادی و ظهور رایانه‌های دیجیتال، به صورت یک رؤیا باقی ماند.

اقدامات اولیه در زمینهٔ بازشناسی حروف، بر متون چاپی یا مجموعهٔ کوچکی از حروف و نمادهای دستنوشت که براحتی قابل تشخیص بودند، متمرکز گردیده بود. سیستم‌های بازشناسی حروف چاپی که در این مقطع زمانی عرضه شدند، عمدتاً از روش تطابق قالبی استفاده می‌نمودند که در آن، تصویر ورودی با مجموعهٔ بزرگی از تصاویر حروف، مورد مقایسه قرار می‌گرفت. در مورد متون دستنوشت نیز الگوریتم‌های پردازش تصویر که ویژگی‌های سطح پایین[۱۴] (ویژگی‌هایی که مستقیماً و بدون اعمال هیچ تبدیلی، از تصاویر استخراج می‌شوند) را از تصاویر استخراج می‌کنند، در مورد تصاویر دوسطحی[۱۵] اعمال می‌شدند تا بردارهای ویژگی استخراج گردند. سپس این بردارهای ویژگی به طبقه‌بندی‌کننده‌های آماری سپرده می‌شدند.

در این دوره، تحقیقات موفق اما مقید[۱۶] (منظور از مقید، مفروض دانستن شرایط و پیش‌فرض‌های خاص برای کاراکترهای ورودی است)، بیشتر بر روی حروف و اعداد لاتین انجام گرفت. با این حال مطالعات چندی نیز بر روی حروف ژاپنی، چینی، عبری، هندی، سیریلیکی، یونانی و عربی در هر دو زمینهٔ حروف چاپی و دستنوشت آغاز گردید. با ظهور صفحات رقومی‌کننده[۱۷] در دههٔ ۱۹۵۰ که قادر به تشخیص مختصات حرکتی نوک یک قلم مخصوص بودند، سیستم‌های «اُسی‌آر» تجاری نیز امکان عرضه یافتند. این نوآوری سبب شد که محققان بتوانند در زمینهٔ بازشناسی برخط[۱۸] حروف دستنوشت، فعالیت خود را آغاز نمایند. «Suen، ۱۹۹۲» یک منبع مناسب دربارهٔ اقدامات صورت‌گرفته بر روی بازشناسی برخط حروف تا سال ۱۹۸۰ می‌باشد.

ب. مرحلهٔ توسعه (از ۱۹۸۰ تا ۱۹۹۰): مطالعات صورت گرفته تا قبل از سال ۱۹۸۰ بدلیل فقدان سخت‌افزارهای رایانه‌ای قدرتمند و دستگاه‌های اخذ داده‌ها با مشکل همراه بودند. در این دهه بواسطهٔ رشد انفجارگونهٔ فناوری اطلاعات، وضعیت بسیار مناسبی برای تحقیقات مختلف از جمله بازشناسی حروف فراهم گردید. روش‌های ساختاری به همراه روش‌های آماری در بسیاری از سیستم‌ها استفاده شدند. تحقیقات در زمینهٔ «اُسی‌آر» اساساً به توسعه روش‌های بازشناسی معطوف گردید، بی آنکه مسئله استفاده از اطلاعات معناشناختی[۱۹] به منظور افزایش دقت بازشناسی مورد توجه قرار گیرد. این امر سبب گردید که دقت بازشناسی (نرخ بازشناسی)

از یک حد خاص فراتر نرود، که در بسیاری از کاربردهای «اُسی‌آر»، قابل قبول نبود. مروری بر تحقیقات و پیشرفت‌های حاصل‌شده در مورد «اُسی‌آر» در این برههٔ زمانی را می‌توان در «Mori، ۱۹۹۲» و «Suen، ۱۹۹۲»، بترتیب برای بازشناسی برخط و برون‌خط[۲۰]، جستجو نمود.

ج. مرحلهٔ بهبود (از ۱۹۹۰ به بعد): در این مقطع زمانی بود که با تکوین ابزارها و تکنیک‌های پردازشی جدید، پیشرفت واقعی در سیستم‌های «اُسی‌آر» محقق گردید. در اوایل دههٔ ۹۰، روش‌های پردازش تصویر و بازشناسی الگو با تکنیک‌های کارآمد هوش مصنوعی ادغام گشتند. محققان، الگوریتم‌های پیچیده‌ای را در بازشناسی حروف ابداع نمودند که قادر بودند داده‌های ورودی با تفکیک‌پذیری[۲۱] بالا را دریافت کنند و در مرحلهٔ پیاده‌سازی، محاسبات بسیار زیادی را بر روی داده‌ها انجام دهند. امروزه علاوه بر وجود رایانه‌های قدرتمندتر و تجهیزات الکترونیکی دقیق‌تر مانند اسکنرها، دوربین‌ها و صفحات رقومی‌کننده، استفاده از تکنیک‌های پردازشی مدرن و توانمند همچون شبکه‌های عصبی[۲۲]، مدل‌های مارکوف پنهان[۲۳]، منطق‌های مجموعهٔ فازی[۲۴] و مدل‌های پردازش زبان طبیعی[۲۵] امکان‌پذیر گشته است.

جدول ۱. جایگاه کنونی تحقیقات در زمینهٔ سیستم های «اُسی‌آر» لاتین

		متون چاپی			متون دستنوشت		
		یک نوع فونت	چند نوع فونت	همه نوع فونت[۲۶]	گسسته	پیوسته	مخلوط
برخط	مقید						
	نامقید						
برون‌خط	بدون نویز						
	نویزی						

در حد مطلوب — نیازمند بهبود — نیازمند تحقیقات بیشتر

سیستم‌های جدید «اُسی‌آر» برون‌خط متون چاپی و برخط متون دستنوشت با واژگان محدود و وابسته به نویسنده، در کاربردهای محدود به نحو کاملاً رضایت‌بخشی عمل می‌کنند (Arica، ۲۰۰۱). اما به منظور دستیابی به هدف نهایی در شبیه‌سازی ماشینی نگارش انسانی و متون چاپی، هنوز راه درازی در پیش است. «آریکا» (Arica، ۲۰۰۱) برخی از تکنیک‌های عرضه‌شده در این دوره را مرور کرده‌اند. جدول ۱ جایگاه کنونی پیشرفت‌های حاصل‌شده در

زمینهٔ سیستم‌های «اُسی‌آر» برای متون لاتین را به نمایش می‌گذارد (Arica، ۲۰۰۱). توجه شود که برای متون چاپی، پردازش برخط تعریف نمی‌شود.

۴. تحقیقات انجام‌شده در داخل کشور در زمینهٔ تولید «اُسی‌آر» فارسی

با توجه به اهمیت طیف وسیع کاربردها و نیاز شدید بازار تجاری، در سال‌های اخیر تحقیقات قابل ملاحظه‌ای در کشور در زمینهٔ «اُسی‌آر» 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 «اُسی‌آر» متون چاپی تاکنون هیچ نرم‌افزار کارآمد «اُسی‌آر» تجاری که محصول تحقیقات داخل کشور باشد، عرضه نگردیده است. در ادامه به برخی از تلاش‌های صورت گرفته در این زمینه اشاره می‌شود:

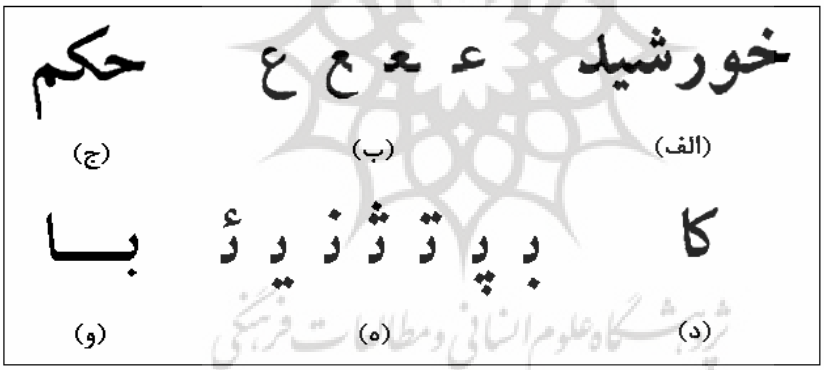
- در حوزهٔ تحقیقات دانشگاهی، تعداد نسبتاً زیادی پایان‌نامه (بخصوص در مقاطع کارشناسی ارشد و دکتری) و مقاله در این زمینه منتشر شده‌اند که نقطهٔ تمرکز بیشتر آن‌ها، ارائهٔ روش‌هایی به منظور قطعه‌بندی درونی[۲۷]، بازنمایی[۲۸] و بازشناسی[۲۹] حروف بوده است و سایر بخش‌ها شامل پیش‌پردازش[۳۰]، قطعه‌بندی بیرونی[۳۱] و پس‌پردازش[۳۲] کمتر مورد توجه قرار گرفته‌اند[۳۳]. بخش‌های مختلف پردازشی یک سیستم «اُسی‌آر» شامل پیش‌پردازش، قطعه‌بندی، بازنمایی، بازشناسی و پس‌پردازش در بخش ۷ مورد بررسی قرار خواهند گرفت.
- طرح ملی «بازشناسی متون چاپی و حجم محدودی از کلمات دست‌نویس» (کبیر، ۱۳۷۷) به سرپرستی دکتر «احسان‌الله کبیر» آغاز گردید که در آن تعدادی از دانشجویان و اساتید دانشگاه‌های تربیت مدرس و صنعتی امیرکبیر در قالب پایان‌نامه‌های کارشناسی ارشد و دکتری، به انجام تحقیق پرداختند. دکتر «کبیر» پروژه‌هایی نیز با عناوین «بازشناسی متون چاپی فارسی» (۱۳۷۱- ۱۳۶۹) و «بازشناسی حروف و ارقام فارسی دست‌نویس» (۱۳۷۴- ۱۳۷۲) برای «سازمان پژوهش‌های علمی و صنعتی ایران» انجام داده است.
- شرکت « اندیشهٔ نرم‌افزار پایا» در سال ۱۳۸۰ طرحی به منظور ارائهٔ یک سیستم «اُسی‌آر» برای شناسایی حروف فارسی گسستهٔ دست‌نوشت آغاز نمود. از محصول این طرح، در سال‌های ۱۳۸۱ و ۱۳۸۲ به منظور خواندن ۴۴۰ هزار فرم ثبت‌نام آزمون «سازمان ملی استعدادهای درخشان» استفاده گردید. در این محصول، حروف فارسی گسستهٔ دست‌نوشت پس از شناسایی، با یکدیگر ترکیب می‌شوند و کلمات به دست آمده از این طریق، در یک واژه‌نامهٔ لغات (شامل نام‌ها و نام‌های خانوادگی مصطلح در ایران) مورد جستجو قرار می‌گیرند که بدین صورت خطاهای بازشناسی تا حد زیادی کاهش پیدا می‌کند. به دلیل گسسته‌بودن حروف دست‌نوشت،

خوش‌خط‌بودن معمول فرم‌ها به واسطهٔ وسواس پرکنندگان آن‌ها (زیرا شرکت‌کنندگان در آزمون نمی‌خواهند که فرم آن‌ها قابل خواندن نباشد)، و نیز به دلیل استفاده از واژه‌نامهٔ لغات، دقت بازشناسی صحیح حروف، رضایت‌بخش (بالای ۹۰٪ بازشناسی صحیح) بود.

- دو شرکت دیگر نیز با حمایت دبیرخانه طرح «تکفا» (توسعه کاربرد فناوری اطلاعات و ارتباطات) مشغول پژوهش و آزمایش بر روی «اُسی‌آر» فارسی هستند. یکی از این شرکت‌ها «داده‌پردازان دوران نوین» (نشانی در وب: داده‌پردازان دوران) نام دارد. اخیراً اعلام گردیده که این شرکت موفق به ارائهٔ یک نرم‌افزار «اُسی‌آر» برای متون چاپی فارسی با دقت بیش از ۹۰٪ گردیده است (نشانی در وب: DOURAN OCR). البته این محصول هنوز به بازار عرضه نشده است.

۵. برخی ویژگی‌های متون چاپی فارسی از دیدگاه پردازش رایانه‌ای

نگارش فارسی، ویژگی‌های منحصر به فردی دارد که آن را کاملاً از نگارش لاتین متمایز می‌سازد. به منظور فعالیت در حوزهٔ «اُسی‌آر» فارسی، آگاهی از قوانین نگارشی و نحوهٔ چاپ حروف در این زبان، امری ضروری است. در اینجا به ویژگی‌های کلی نگارش فارسی اشاره می‌شود.



چهار شکل مختلف حرف «ع» با توجه به موقعیت آن در کلمه؛ ج) همپوشانی دو حرف «ح» و «ک» در کلمهٔ «حکم»؛ د) اتصال حروف«ک» و «ا» در دو محل؛ ه) حروف متفاوت با بدنهٔ مشابه؛ و) کشیدگی حرف «ب» در کلمهٔ «با».

- متون فارسی برخلاف متون لاتین از راست به چپ نوشته می‌شوند.

- در کلمات فارسی برخی از حروف از یک یا دو طرف به حروف مجاور خود اتصال دارند و برخی نیز به صورت مجزا نوشته می‌شوند. در نتیجه هر کلمه ممکن است شامل یک یا چند بخش متصل باشد که «زیرکلمه» نامیده می‌شوند (
شکل ۳- الف). چسبیده یا سرهم بودن حروف در نگارش فارسی، بازشناسی متون فارسی را برای سیستم‌های «اُسی‌آر»، نسبت به متون لاتین بسیار مشکل‌تر می‌سازد.

- حروف فارسی ممکن است چهار موقعیت مجزا و در نتیجه چهار شکل متفاوت نگارش داشته باشند: حروف ابتدایی، میانی، انتهایی و مجزا (شکل۳- ب).

- 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 (شکل۳- ج).

- در برخی از فونت‌ها بعضی از حروف، از یک سمت در دو محل به یکدیگر اتصال دارند (شکل۳- د).

- برخی از حروف بین یک تا سه نقطه دارند که ممکن است در بالا یا پایین بدنهٔ حرف واقع باشند (شکل ۳- هـ).

- بعضی از حروف بدنهٔ مشابه دارند و تفاوت آن‌ها تنها در تعداد و محل قرارگیری نقاط (شکل هـ) یا در وجود یک سرکش است (مانند «ک» و «گ») که در مقایسه با بدنهٔ حروف، اندازهٔ بسیار کوچکی دارند. این موضوع نیز یکی از مواردی است که بر پیچیدگی سیستم‌های «اُسی‌آر» فارسی می‌افزاید.

- حروف فارسی ممکن است در بالا یا پایین بدنه دارای اعراب باشند. سه اعراب ـَ ـِ ـُ در زبان فارسی، اعراب‌های اصلی‌اند و اعراب ـً در برخی کلمات عربی رایج در زبان فارسی دیده می‌شود (نظیر کلمات «عمداً» و «احتمالاً»). کلمات عربی دارای اعراب ـٌ و ـٍ در زبان فارسی عمومیت نیافته‌اند. هر چند کاربرد اعراب در زبان فارسی نسبت به زبان عربی بسیار محدودتر است، اما اگر کلمه‌ای نامتداول باشد یا به دلیل تشابه نگارشی آن با کلمهٔ دیگر، تأکید بر تلفظ صحیح آن باشد، از نشانه‌های اعراب استفاده می‌شود.

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

- برخی از حروف دارای علامت همزه هستند («ئ»، «أ»، «ؤ»، «ۀ»).

- حروفی که از طرف چپ قابلیت اتصال به حرف مجاور خود را دارند، ممکن است به صورت کشیده نوشته شوند (شکل ۳- و).

بیشتر حروف فارسی (مخصوصاً حروف چسبیده) دندانه‌دار هستند. در مواردی که کیفیت سند اصلی یا دستگاه اسکنر پایین باشد، ارتفاع دندانه‌ها نسبت به خط زمینه کوتاه

می‌شود و این امر، شناسایی صحیح این حروف در مرحلهٔ قطعه‌بندی یا بازشناسی را با مشکل مواجه می‌سازد.

۶. انواع سیستم‌های «اُسی‌آر» از لحاظ نوع الگوی ورودی

سیستم‌های «اُسی‌آر» را می‌توان از لحاظ نوع الگوی ورودی به دو گروه اصلی تقسیم کرد (عزمی، ۱۳۷۸):

الف. سیستم‌های برخط،

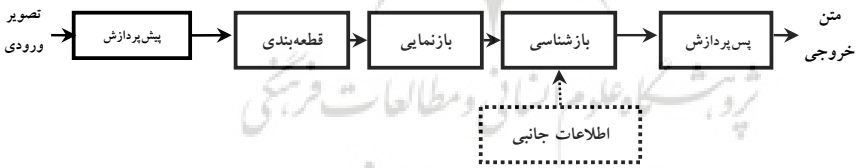
ب. سیستم‌های برون‌خط.

در بازشناسی برخط، حروف در همان زمان نگارش توسط سیستم تشخیص داده می‌شوند و دستگاه ورودی این سیستم‌ها یک قلم نوری است. در این روش علاوه بر اطلاعات مربوط به موقعیت قلم، اطلاعات زمانی مربوط به مسیر قلم نیز در اختیار است. این اطلاعات معمولاً توسط یک صفحهٔ رقومی‌کننده[۳۴] اخذ می‌شوند. در این روش می‌توان از اطلاعات زمانی سرعت، شتاب، فشار و زمان برداشتن و گذاشتن قلم روی صفحه در بازشناسی استفاده کرد.

در بازشناسی برون‌خط، از تصویر دوبعدی متن ورودی استفاده می‌شود. در این روش به هیچ نوع وسیله نگارش خاصی نیاز نیست و تفسیر داده‌ها مستقل از فرآیند تولید آن‌ها و تنها براساس تصویر متن صورت می‌گیرد. این روش به نحوهٔ بازشناسی توسط انسان شباهت بیشتری دارد.

۷. معرفی بخش‌های مختلف یک سیستم «اُسی‌آر»

شکل ۴ نمودار بلوکی یک سیستم «اُسی‌آر» را نمایش می‌دهد. لازم به ذکر است که بسته به الگوریتم کلی به کار رفته و سطح انتظارات از عملکرد نرم‌فزار، ممکن است برخی سیستم‌ها فاقد یک یا چند مرحله از مراحل فوق باشند.



شکل ۴. نمودار بلوکی دیاگرام یک سیستم «اُسی‌آر»

در ادامه هر یک از این بلوک‌ها مورد بررسی قرار می‌گیرند.

۷-۱. پیش‌پردازش

کلیهٔ اعمالی که روی تصویر صورت می‌گیرند تا موجب تسهیل در روند اجرای فازهای بعدی گردد؛ مانند دوگانی‌کردن[۳۵] تصویر، حذف نویز[۳۶]، هموارسازی[۳۷]، نازکسازی[۳۸]، تشخیص زبان و فونت کلمات، و نظایر این‌ها. از مجموعهٔ این پردازش‌ها، هدف‌های زیر دنبال می‌شود (Arica، ۲۰۰۱):

۱. کاهش نویز،

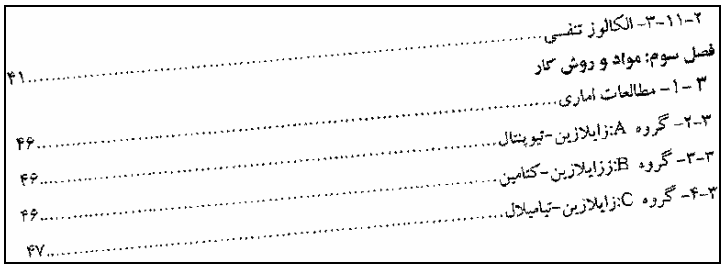
۲. نرمالیزه کردن داده‌ها،

۳. فشرده‌سازی میزان اطلاعاتی که می‌بایست محفوظ بماند.

کاهش نویز: نویز ایجادشده بواسطهٔ دستگاه‌های اسکنر نوری منجر به ایجاد نقطه نقطه‌های لک‌مانند[۳۹]، قطعه خط‌های گسسته[۴۰]، اتصال بین خطوط، فضاهای خالی در خطوط متن، پرشدن حفره‌های موجود در تصویر برخی حروف، و ... می‌گردد. همچنین اعوجاج‌های مختلف شامل تغییرات محلی، منحنی‌شدن گوشه‌های حروف، تغییر شکل[۴۱] یا خوردگی حروف[۴۲] را نیز باید در نظر داشت. قبل از مرحلهٔ بازشناسی حروف، لازم است که این نقایص برطرف شوند. مهم‌ترین دلیل برای کاهش نویز، کم‌کردن خطا در مراحل قطعه‌بندی و بخصوص بازشناسی می‌باشد. کاهش نویز همچنین سبب کم‌شدن اندازهٔ فایل تصویر می‌شود که به نوبهٔ خود، کاهش زمان مورد نیاز برای پردازش‌ها و ذخیره‌سازی‌های آینده را در پی خواهد داشت (O'Gorman، ۱۹۹۵).

نرمالیزه‌کردن: روش‌های نرمالیزه‌کردن داده‌ها به حذف تغییرات نگارشی کمک می‌کند و داده‌های استانداردشده‌ای را نتیجه می‌دهد. روش‌های پایهٔ نرمالیزه‌کردن عبارت‌اند از:

الف. نرمالیزه‌کردن کجی[۴۳] متن و استخراج خطوط زمینه[۴۴]: به دلیل بی‌دقتی در مرحلهٔ اسکن یا بی‌دقتی نویسنده در هنگام نگارش متن دست‌نوشت، ممکن است خطوط متن نسبت به تصویر، اندکی انحراف یا چرخش داشته باشند (شکل ۵). این وضع ممکن است کارآیی الگوریتم‌های به کار رفته در طبقات بعدی سیستم «اُسی‌آر» را تحت تأثیر قرار دهد؛ چرا که یکی از مفروضات در بیشتر روش‌های قطعه‌بندی، کج‌نبودن تصویر متن ورودی است و در نتیجه لازم است که این نقیصه، آشکار و تصحیح گردد. آشکارسازی خط زمینه در بسیاری از تکنیک‌های قطعه‌بندی و بازشناسی متون فارسی، عربی و لاتین، نقش اساسی دارد. علاوه بر این، برخی از کاراکترها مانند «9» و «g» در نگارش لاتین و «.» (نقطه) و «٠» (صفر) در نگارش فارسی را بواسطهٔ موقعیت نسبی‌شان نسبت به خط زمینه می‌توان آشکار ساخت.



شکل ۵. تصویر یک صفحه که کج اسکن شده است

کلیهٔ الگوریتم‌های توسعه داده شده برای آشکارسازی کجی صفحه، بر روی صفحات متنی با ترازبندی[۴۵] یکنواخت، دقیق عمل می‌کنند. الگوریتمی کارآتر است که به واسطهٔ حضور مواردی نظیر گرافیک، پاراگراف‌های دارای کجی متفاوت، اعوجاج‌های منحنی- خطی[۴۶] ظاهرشونده در کتاب‌های فتوکپی‌شده، نواحی وسیع پیکسل‌های سیاه نزدیک حاشیهٔ صفحه و خطوط متنی مختصر و کوتاه، دقت آن کمتر دستخوش تغییر شود.

روش‌های به کار رفته برای تصحیح کجی خطوط زمینه در متون لاتین را می‌توان به چهار گروه اصلی دسته‌بندی کرد که عبارت‌اند از (Arria ، ۲۰۰۱؛ «صفابخش»، ۱۳۷۸):

۱. به کارگیری هیستوگرام[۴۷] (پروفایل تصویرنمایی[۴۸]) تصویر،
۲. استفاده از روش خوشه‌بندی نزدیک‌ترین همسایه‌ها[۴۹]،
۳. روش همبستگی متقابل[۵۰] بین حروف،
۴. تبدیل هاف[۵۱].

اغلب پس از آشکارسازی کجی، تصویر صفحه در جهت اصلی چرخانده می‌شود تا عملیات تحلیل قالب‌بندی متن و «اُسی‌آر» با سهولت و دقت بیشتری انجام پذیرد. نمونه‌برداری مجدد[۵۲] مورد نیاز برای این منظور که باید بر روی صفحات دوگانی‌شده اعمال گردد، ممکن است الگوی کاراکترها را تغییر دهد. در این حالت به جای چرخاندن تصویر می‌توان الگوریتم‌های پردازشی را به نحوی اصلاح نمود که اثر چرخش در آن‌ها لحاظ گردد (Jain، ۱۹۹۸). همچنین می‌توان تصویر سند را قبل از دوگانی‌کردن، چرخش داد یا این که مقدار چرخش را از روی انتقال‌های کوچک و بدون اعوجاج کل بلوک‌های متنی، تقریب زد (Nagy، ۲۰۰۰).

ب. نرمالیزه‌کردن اریب‌شدگی[۵۳]: در متون چاپی فارسی و لاتین، کاراکترهای دارای قالب ایتالیک از راستای عمود، انحراف دارند. در متون دست‌نوشت نیز برخی از نویسنده‌ها حروف را به صورت زاویه‌دار می‌نویسند. این پدیده با عنوان «اریب‌شدگی» شناخته می‌شود و ممکن است دقت برخی از الگوریتم‌های قطعه‌بندی یا بازشناسی را تحت تأثیر قرار دهد و و از این رو در این

سیستم‌ها لازم است که در مرحلهٔ پیش‌پردازش، میزان اریب‌بودن کاراکترها شناسایی و تصحیح گردد.

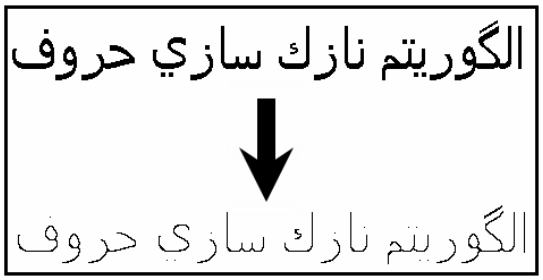
ج. نرمالیزه‌کردن (تغییر مقیاس دادن[۵۴]) اندازه: در سیستم‌های«اُسی‌آر»، اغلب تصاویر کلمات یا حروف خیلی کوچک یا خیلی بزرگ، به یک اندازهٔ استاندارد نرمالیزه می‌شوند تا بدین ترتیب عملیات بازشناسی، مستقل از اندازهٔ فونت متن گردد.

د. هموارسازی کانتور[۵۵]: خط تشکیل‌دهندهٔ مرز یک کاراکتر را کانتور آن کاراکتر گویند. در متون دست‌نوشت، به واسطهٔ لرزش یا حرکات ناخواستهٔ دست نویسنده در هنگام نگارش، ممکن است که کانتور حروف ناصاف شود. این وضع در سیستم‌های بازشناسی متون چاپی و دست‌نوشت نیز، به دلیل تغییر مقیاس حروف یا وجود نویز در مرحلهٔ اسکن تصاویر ممکن است ظاهر گردد. روش‌های هموارسازی کانتور، به منظور جبران این نقیصه مورد استفاده قرار می‌گیرند. به طورکلی هموارسازی کانتور، تعداد نقاط نمونهٔ مورد نیاز برای نمایش کاراکتر را کاهش می‌دهد و در نتیجه کارآیی مراحل پردازشی باقیمانده را بهبود می‌بخشد.

فشرده‌سازی: این نکته پذیرفته شده است که تکنیک‌های کلاسیک فشرده‌سازی تصاویر که تصویر را از حوزهٔ مکانی به حوزه‌های دیگر منتقل می‌کنند، برای بازشناسی حروف مناسب نیستند. در بازشناسی حروف، عمل فشرده‌سازی نیازمند آن دسته از تکنیک‌های حوزهٔ مکانی است که اطلاعات شکلی را حفظ می‌نمایند. دو تکنیک متعارف فشرده‌سازی، یکی تکنیک اعمال سطح آستانه[۵۶] (به منظور دوگانی یا دوسطحی کردن تصاویر سطح خاکستری) و دیگری نازک‌سازی می‌باشند.

الف. دوگانی (دوسطحی) کردن تصویر متن: تصاویر دیجیتالی به یکی از سه صورت- تصاویر رنگی، تصاویر سطح خاکستری[۵۷] (مشابه تصویر یک تلویزیون سیاه و سفید که رنگ تصویر به صورت سیاه، سفید و طیفی از رنگ‌های خاکستری ظاهر می‌شود)، و تصاویر دوگانی یا دوسطحی (مشابه تصویر یک سند فکس‌شده که رنگ پیکسلهای تصویر، تنها سیاه یا سفید است)- می‌باشند. به منظور کاهش حجم ذخیره‌سازی مورد نیاز و افزایش سرعت و سهولت پردازش، اغلب مطلوب است که با انتخاب یک سطح آستانه، تصاویر سطح خاکستری یا رنگی را به تصاویر دوگانی تبدیل نمود.

ب. نازک‌سازی: با این عمل، تصویر کاراکترها به تصویری با عرض یک پیکسل تبدیل می‌شود؛ درست مثل این که کاراکترها با یک قلم نوک باریک نوشته شده باشند. نازک‌سازی در حالی که کاهش قابل‌ملاحظه‌ای در حجم داده‌ها ایجاد می‌کند، اطلاعات شکلی کاراکتر را نیز حفظ می‌نماید (شکل ۶).



شکل ۶. اعمال عملیات نازک‌سازی بر روی یک تصویر متنی نمونه

بازشناسی خط[۵۸]، زبان و فونت: بازشناسی خط، تعداد کلاس‌های مختلف نمادهایی را که باید مورد ملاحظه قرار گیرند کاهش می‌دهد. شناسایی زبان متن، به منظور به کارگیری مدل‌های متنی[۵۹] خاص ضرورت دارد. طبقه‌بندی فونت‌ها، کاهش تعدد شکل‌های مختلف حروف در هر کلاس که لازم است در فرایند بازشناسی لحاظ گردند را به دنبال دارد و سبب می‌شود که امر شناسایی، تنها به یک کلاس فونت محدود گردد. بازشناسی خط و زبان و فونت در کاربردهایی مانند نمایه‌سازی[۶۰] و دستکاری اسناد نیز مطلوب می‌باشد.

۷-۲. قطعه‌بندی (جداسازی)

قطعه‌بندی مرحله‌ای بسیار مهم برای سیستم‌های «اُسی‌آر» مخصوصاً «اُسی‌آر» فارسی و عربی (که حروف کلمات به صورت سرهم نوشته می‌شوند) می‌باشد. قطعه‌بندی به دو گونه تقسیم می‌شود (Arica، ۲۰۰۱):

۱. قطعه‌بندی بیرونی که عبارت است از تفکیک قسمت‌های مختلف تصویر نظیر متن، گرافیک و خطوط و نیز جدا کردن بخش‌های مختلف متن مانند پاراگراف‌ها، جملات یا کلمات؛

۲. قطعه‌بندی درونی که منظور از آن، جداسازی حروف کلمات مخصوصاً در مورد کلمات سر هم[۶۱] نوشته‌شده در متون لاتین، یا در رسم‌الخط‌های پیوسته نظیر فارسی و عربی است (شکل ۷). همچنین حروفی که در متن اصلی جدا بوده‌اند، امّا به خاطر کیفیت پایین دستگاه اسکنر به هم چسبیده‌اند، توسط این دسته از تکنیک‌ها از یکدیگر جدا می‌شوند.

شکل ۷. قطعه‌بندی یک کلمه به حروف

مرحلهٔ قطعه‌بندی بیرونی، بحرانی‌ترین و حساس‌ترین قسمت در حوزهٔ تحلیل تصویر اسناد می‌باشد و یک مرحلهٔ ضروری برای سیستم‌های «اُسی‌آر» برون‌خط محسوب می‌شود. گرچه مبحث تحلیل اسناد با روش‌ها و تکنیک‌های خاص خود یک حوزهٔ تحقیقاتی تا حدی متفاوت نسبت به «اُسی‌آر» است، امَا تقسیم‌بندی تصویر سند به نواحی متنی و غیرمتنی، یک بخش لاینفک در نرم‌افزارهای «اُسی‌آر» به حساب می‌آید. بنابراین برای افرادی که در زمینهٔ «اُسی‌آر» تحقیق می‌نمایند، داشتن دانش عمومی از تکنیک‌های آنالیز اسناد ضرورت دارد.

تحلیل پیکربندی صفحات در سه مرحله انجام می‌گیرد: مرحلهٔ اول تفکیک نواحی متنی در تصویر از نواحی غیرمتنی (شامل گرافیک و خطوط) است. مرحلهٔ دوم، تحلیل ساختاری[۶۲] است که با قطعه‌بندی نواحی متنی به بلوک‌هایی از تصویر سند (نظیر پاراگراف، ردیف، کلمه، و...) مرتبط می‌باشد. مرحلهٔ سوم که تحلیل عملکردی[۶۳] نام دارد، با استفاده از اطلاعات مکانی، اندازه و قوانین مختلف صفحه‌بندی، عملکرد هر یک از اجزای سند (نظیر عنوان، چکیده، و ...) را تعیین می‌نماید (O'Gorman، ۱۹۹۵).

نقطهٔ تمایز اصلی میان سیستم‌های «اُسی‌آر» لاتین و فارسی برای متون چاپی، در مرحلهٔ قطعه‌بندی درونی نهفته است؛ چرا که حروف کلمات در نگارش فارسی برخلاف نگارش رسمی لاتین به صورت سرهم نوشته می‌شوند و در نتیجه ضرورت انجام صحیح این مرحله برای متون فارسی و عربی نسبت به متون لاتین، اهمیت فوق‌العاده بیشتری دارد. با وجود فعالیت‌های نسبتاً چشمگیر دههٔ گذشته و تنوع تکنیک‌های عرضه‌شده، قطعه‌بندی متون پیوسته (بخصوص متون دست‌نوشت پیوسته) به حروف، هنوز هم یک مسئله قابل بررسی مانده است. روش‌های قطعه‌بندی حروف به سه دسته تقسیم می‌شوند (Mori، ۱۹۹۹):

۱. قطعه‌بندی صریح[۶۴]،
۲. قطعه‌بندی ضمنی[۶۵]،
۳. تکنیک‌های ادغام‌شده.

در مواردی همچون متون فارسی که حروف به صورت سرهم نوشته می‌شوند، سه رویکرد مختلف در بازشناسی برون‌خط متون کلمات یا زیرکلمات وجود دارد (عزمی ،۱۳۷۸):

۱. رویکرد مبتنی بر قطعه‌بندی کلمات[۶۶]،

۲. رویکرد مبتنی بر بازشناسی کلمه به عنوان یک الگوی واحد[۶۷]،

۳. رویکرد ترکیبی.

در رویکرد بازشناسی مبتنی بر قطعه‌بندی، ابتدا کلمه در مرحلهٔ جداسازی به حروف یا زیرحروف، شکسته می‌شود؛ آنگاه قطعات جداشده بازشناسی می‌شوند و از کنار هم قرارگرفتن آن‌ها، کلمه شناسایی خواهد شد. روش‌های به کار گرفته‌شده در این رویکرد به دو گروه مختلف تقسیم می‌شوند:

- تقطیع کلمه به حروف،
- تقطیع کلمه به زیرحروف.

در گروه اول، کلمه به حروف جداسازی می‌شود و با شناسایی حروف جداشده، کلمه بازشناسی می‌گردد.

در گروه دوم، کلمه به زیرحروف مثل پاره‌منحنی‌ها و ساختارهای پایهٔ دیگر جداسازی می‌شود و با شناسایی زیرحروف‌ها و ترکیب آن‌ها، کلمه بازشناسی می‌گردد. در این رویکرد نمی‌توان در ابتدا مرز حروف را به طور کامل مشخص کرد، بلکه حروف به ترتیب از ابتدا به انتهای کلمه، بازشناسی و جداسازی می‌شوند. در هیچک از دو رویکرد نخست که مبتنی بر جداسازی هستند، به شکل کلی کلمه توجهی نمی‌شود و سعی بر آن است که با بازشناسی حروف یک کلمه، آن کلمه شناخته شود.

در رویکرد بازشناسی کلمه به عنوان یک الگوی واحد، تلاشی برای تقطیع کلمه به حروف و بازشناسی حروف موجود در کلمه صورت نمی‌گیرد و کلمه در قالب یک الگو بررسی می‌گردد.

قطعه‌بندی غلط کاراکترها، عامل بسیاری از خطاهای «اُسی‌آر» است (مانند: rn→m یا m→rn). میزان دقت یک الگوریتم قطعه‌بندی به سبک نگارش حروف، کیفیت دستگاه چاپ (کاراکترهای ایتالیک لکه‌دار برای قطعه‌بندی دارای اشکال می‌باشند)، و نیز نسبت اندازهٔ فونت به قدرت تفکیک[۶۸] دستگاه اسکنر (تابع گسترش نقاط[۶۹] و نرخ نمونه‌برداری مکانی) بستگی دارد (Trier، ۱۹۹۶).

نتیجهٔ مطلوب مرحلهٔ قطعه‌بندی، تصویری است که تنها حاوی یک کاراکتر باشد و بجز پیکسل‌های پس‌زمینه، هیچ شیء دیگری را شامل نشود. اما هنگامی که اشیای چاپی، در تصویر ورودی خیلی نزدیک به هم ظاهر شوند (مانند نقشه‌های هیدروگرافی)، این منظور همواره قابل

حصول نخواهد بود. غالباً در چنین حالتی دیگر کاراکترها یا اشیای چاپی، به طور تصادفی در داخل تصویر کاراکتر قرار می‌گیرند و احتمالاً ویژگی‌های استخراج‌شده را تحریف می‌نمایند. این مورد یکی از دلایلی است که بیان می‌دارد چرا هر سیستم بازشناسی حروف، یک گزینهٔ وازدگی[۷۰] دارد (Trier، ۱۹۹۶). «عزمی» بررسی جامعی بر روی روش‌های عرضه‌شده برای قطعه‌بندی کلمات در متون فارسی انجام داده است (عزمی ،۱۳۷۸).

۷-۳. بازنمایی (استخراج ویژگی‌ها)[۷۱]

این مرحله یکی از مراحل بسیار با اهمیت در سیستم‌های «اُسی‌آر» است؛ چرا که نتایج حاصل از این مرحله، مستقیماً بر روی کیفیت مرحلهٔ بازشناسی اثر می‌گذارد. در مرحلهٔ بازنمایی، به هر الگوی ورودی (کاراکتر یا کلمه- بر حسب آن که رویکرد سیستم، مبتنی بر کدامیک از دو گروه «قطعه‌بندی کلمات» یا «شناسایی کلمه به عنوان یک الگوی واحد» باشد)، یک کد یا بردار ویژگی نسبت داده می‌شود که معرف آن الگو در فضای ویژگی‌ها است و آن را از دیگر الگوها متمایز می‌سازد. در انتخاب بردارهای ویژگی لازم است موارد زیر مورد توجه قرار گیرند:

۱. بردار ویژگی هر الگو باید تا حد امکان از بردارهای ویژگی دیگر الگوها متمایز باشد (فاصلهٔ بین بردارهای ویژگی در فضای ویژگی‌ها، حداکثر باشد).

۲. بردار ویژگی الگوها باید تا بیشترین حد ممکن، خصوصیات شکل و ساختار الگوها را از تصویر آن‌ها استخراج نماید.

۳. تا حد امکان نسبت به نویز، تغییر اندازه و نوع فونت، چرخش، و دیگر تغییرات احتمالی الگوها دارای ثبات باشد.

۴. شرایط، نوع و خصوصیات الگوهای ورودی در انتخاب بردارهای ویژگی اثر می‌گذارند. به عنوان مثال، باید تعیین نمود که آیا حروف یا کلماتی که می‌بایست تشخیص داده شوند جهت و اندازهٔ مشخصی دارند یا خیر، دست‌نوشت هستند یا چاپی، یا این که تا چه حد بوسیلهٔ نویز، مغشوش شده‌اند. همچنین گاهی کفایت می‌کند که سیستم، تنها جوابگوی گروه محدودی از الگوها (مثلاً الگوهایی با اندازه یا نوع فونت از پیش مشخص شده) باشد.

۵. در مورد حروفی که به چندین شکل نوشته می‌شوند (مانند ‘a’ و ‘a’، «۴» و «۴») لازم است که بیش از یک کلاس الگو به یک کاراکتر خاص تعلق یابد.

همانطور که عنوان شد، بازنمایی یک مرحلهٔ بسیار مهم در حصول راندمان مناسب برای سیستم‌های بازشناسی حروف است؛ ولی برای دستیابی به عملکرد بهینه، لازم است که دیگر مراحل نیز بهینه گردند و باید توجه نمود که این مراحل، مستقل از هم نیستند. یک روش خاص

استخراج ویژگی‌ها، طبیعت خروجی مرحلهٔ پیش‌پردازش را به ما دیکته می‌کند یا حداقل ما را در انتخابمان محدود می‌سازد.

مراحل قطعه‌بندی و بازشناسی، دو وجه تمایز عمده میان سیستم‌های «اُسی‌آر» فارسی و لاتین می‌باشند. بواسطهٔ وجود تفاوت‌های اساسی بین نحوهٔ نگارش فارسی و لاتین، امکان اعمال مستقیم تکنیک‌های قطعه‌بندی و بازشناسی مربوط به سیستم‌های «اُسی‌آر» لاتین، برای متون فارسی وجود ندارد. پیچیدگی‌های مختص نگارش فارسی، بر پیچیدگی الگوریتم‌های این دو مرحله می‌افزاید. درست به همین دلیل است که بیشتر نرم‌افزارهای «اُسی‌آر» تجاری لاتین، قادر به پشتیبانی زبان فارسی و عربی نمی‌باشند.

۷-۴. طبقه‌بندی و بازشناسی[۷۲] (با یک یا چند طبقه‌بندی‌کننده)

این مرحله شامل روش‌هایی برای متناظرساختن هر یک از الگوهای به دست‌آمده از مرحلهٔ استخراج ویژگی‌ها، با یکی از کلاس‌های فضای الگوهای مورد بحث است که از طریق کمینه ساختن فاصلهٔ بردار ویژگی‌های هر الگوی ورودی نسبت به یکی از بردارهای مرجع، انجام می‌گیرد. بردارهای مرجع، بردارهایی هستند که قبلاً از نمونه‌های آموزشی اخذ شده‌اند. تکنیک‌های عرضه‌شده برای این مرحله را می‌توان در روش‌های مربوط به چهار گروه عمومی مبحث شناسایی الگو، جستجو نمود (Arica، ۲۰۰۱):

- تطابق قالبی،
- تکنیک‌های آماری،
- تکنیک‌های ساختاری،
- شبکه‌های عصبی.

چهار گروه فوق لزوماً مستقل یا مجزا از یکدیگر نمی‌باشند و گاهی می‌توان تکنیک‌های یک گروه را در میان تکنیک‌های مربوط به دیگر گروه‌ها یافت.

۷-۵. به کارگیری اطلاعات جانبی (پس‌پردازش)

در این مرحله با استفاده از اطلاعات جانبی (نظیر مجموعه لغات معتبر، اطلاعات آماری مربوط به رخداد حروف، اطلاعات دستوری و معنایی) سعی در بهبود نتایج حاصل از مرحلهٔ بازشناسی می‌گردد. تا قبل از این مرحله، هیچگونه اطلاعات معناشناختی در طول مراحل پردازشی مورد استفاده قرار نگرفته بود. مرحلهٔ پیش‌پردازش سعی در «تمیز کردن» تصویر سند 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 مرحلهٔ قطعه‌بندی می‌تواند به نتایج

حادتر و غیرقابل برگشتی بینجامد؛ چرا که خروجی این مرحله، تعیین‌کنندهٔ مرز الگوهای ورودی می‌باشد. بنابراین واضح است که در صورت فراهم‌شدن اطلاعات معناشناختی، دقت نتایج بازشناسی به نحو چشمگیری افزایش می‌یابد. مروری بر تحقیقات اخیراً انجام‌شده در زمینهٔ بازشناسی حروف نشان می‌دهد که در صورت استفاده از اطلاعات شکلی بدون به‌کارگیری اطلاعات معناشناختی، افزایش دقت قابل توجهی نخواهیم داشت. در نتیجه، یکپارچه‌کردن اطلاعات معنایی و شکلی در کلیهٔ بلوک‌های سیستم‌های «اُسی‌آر» به منظور بهبود نرخ بازشناسی صحیح، ضرورت دارد.

ساده‌ترین راه برای این منظور، استفاده از یک فرهنگ لغات برای اصلاح خطاهای جزئی است. این کار توسط یک برنامهٔ واژه‌پرداز (با قابلیت خطایاب املایی[۷۳]) که املای کلمات را کنترل و چندین پیشنهاد برای اصلاح املای کلمات نامأنوس ارائه می‌کند، انجام می‌گیرد.

۸. نرم‌افزارهای تجاری «اُسی‌آر» لاتین و فارسی

جدول ۲. معروف‌ترین نرم‌افزارهای تجاری «اُسی‌آر» لاتین را نشان می‌دهد.

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

جدول ۲. معروف‌ترین نرم‌افزارهای تجاری «اُسی‌آر» لاتین

از جمله مهم‌ترین ویژگی‌های این سه نرم‌افزار عبارت‌اند از:

۱. دقت بازشناسی بالای ۹۹ درصد؛

۲. پشتیبانی از بیش از ۱۴۰ زبان مختلف (فقط نسخه‌ای از Readiris، زبان عربی را پشتیبانی می‌کند)؛

۳. پشتیبانی از قالب‌بندی‌های مختلف فایلی شامل قالب‌های تصویری[۷۴]، قالب‌های پی‌دی‌اف، و ...؛

۳. تولید خروجی متنی در قالب‌های متنوع[۷۵] ؛

۴. تشخیص خودکار متن، گرافیک و جدول در تصویر ورودی[۷۶]؛

۵. قابلیت اضافه‌نمودن، ویرایش و حذف فریم‌های متنی، گرافیکی و تصویری[۷۷]. به این ترتیب می‌توان در قسمت‌هایی از تصویر که نرم‌افزار قادر به تشخیص درست نیست، به نرم‌افزار کمک نمود؛

۶. قابلیت حفظ قالب‌بندی نگارش تصویر ورودی (پاراگراف‌بندی، نوع و اندازهٔ فونت‌ها، جدول‌بندی، ستون‌بندی، و ...) در تصویر خروجی؛

۷. بازشناسی خودکار متن‌های چندزبانه؛

۸. تشخیص بارکد؛

۹. واژه پرداز[۷۸] چندزبانه به منظور افزایش دقت بازشناسی؛

۱۰. ویرایشگر متن برای دیدن و اصلاح نتیجهٔ عمل اُسی‌آر قبل از تولید فایل خروجی (تنها در نرم افزارهای ۱ و ۲)؛

۱۱. قابلیت پردازش گروهی فایل‌ها[۷۹].

در حال حاضر تنها نرم‌افزار «اُسی‌آر» تجاری که زبان فارسی را پشتیبانی می‌نماید، نرم‌افزار «اتوماتیک ریدر وی۸»[۸۰] محصول شرکت عربی «صخر»[۸۱] است. صخر یک شرکت پیشگام در زمینهٔ ارائهٔ نرم‌افزار و آموزش به زبان عربی می‌باشد که موفقیت‌های چشمگیری را در عرصهٔ فناوری پردازش زبان عربی حاصل نموده است. محصولات این شرکت طیف وسیعی از کاربردها نظیر تحلیل مستندات، «اُسی‌آر»، ترجمهٔ ماشینی، تولید صوت ماشینی، تبدیل متن به گفتار، موتور جستجوی وب، سیستم‌های مدیریت اطلاعات در وب و موارد دیگر را دربرمی‌گیرد. شرکت «صخر» در ایران شعبه ندارد و به همین دلیل نیز دسترسی به این نرم‌افزار براحتی امکان‌پذیر نیست. این نرم‌افزار با یک قفل سخت‌افزاری[۸۲] که به پورت چاپگر وصل می‌شود قابل اجرا است. قابل توجه است که نرم‌افزار «واژه‌شناس» که توسط شرکت «هوش‌مصنوعی رایورز» در بازار ایران عرضه می‌شود، در حقیقت همان نرم‌افزار تولیدی شرکت «صخر» است که تنها منوهای آن فارسی شده. نرم‌افزار «شناسا» که در چند سال اخیر به وسیلهٔ شرکت «جیحون‌افزار» عرضه می‌شد نیز یکی از نسخه‌های پیشین همین نرم‌افزار بود که منوهای آن به فارسی ترجمه شده بود.

نرم‌افزار «اتوماتیک ریدر» دارای دو نسخهٔ «گلد»[۸۳] و «پلاتینیوم»[۸۴] می‌باشد. تنها تفاوت بین این دو نسخه در این است که نسخهٔ «پلاتینیوم»، «اس‌دی‌کا»ی نرم‌افزار را نیز شامل می‌شود[۸۵]. در حال حاضر قیمت نسخه‌های «گلد» و «پلاتینیوم» این نرم‌افزار به ترتیب در حدود ۱۴۰۰ و ۴۰۰۰ دلار است.

نرم‌افزار «اتوماتیک ریدر» دارای ویژگی‌های زیر می‌باشد:

سرعت بازشناسی بسیار بالا؛

پشتیبانی از ۱۳ زبان مختلف (شامل عربی، فارسی، انگلیسی، دانمارکی، هلندی، فنلاندی، فرانسوی، آلمانی، ایتالیایی، نروژی، پرتغالی، اسپانیایی و سوئدی)؛

بازشناسی متون دوزبانهٔ عربی/انگلیسی، فارسی/انگلیسی و عربی/فرانسوی؛

تشخیص اعراب[۸۶] و کشیدگی[۸۷] حروف؛

تشخیص خودکار متن، گرافیک و جدول در تصویر ورودی؛

پشتیبانی از قالب‌های گرافیکی BMP، PCX، JPEG،TIFF و ART؛

تولید خروجی به قالب‌های RTF، DTP، TXT، ART و HTML؛

قابلیت آموزش فونت به منظور افزایش دقت بازشناسی؛

برخورداری از واژه‌پرداز عربی- انگلیسی.

از جمله نقطه‌ضعف‌های این نرم‌افزار به این موارد می‌توان اشاره کرد:

- هرچند ادعا شده که دقت بازشناسی نرم‌افزار«اتوماتیک ریدر» بالای ۹۹ درصد است، اما آزمایش این نرم‌افزار، صحت این ادعا را حتی برای صفحات فارسی اسکن‌شده با کیفیت عالی نیز تأیید نمی‌کند. هر چند شرکت «صخر» با تولید این نرم‌افزار گام مهمی در زمینهٔ «اُسی‌آر» فارسی و عربی برداشته، اما هنوز هم با جایگاه مورد انتظار فاصلهٔ زیادی دارد.

- پیکربندی صفحه و قالب‌بندی پاراگراف‌ها و حروف در فایل تصویری ورودی، در فایل متنی خروجی قالب «آرتی‌اف»با دقت بسیار پایینی رعایت می‌شود.

- محیط کاربری این نرم‌افزار، ضعیف‌تر از نرم‌افزارهای «اُسی‌آر» لاتین اشاره‌شده در بالا است.

- واژه‌پرداز این نرم‌افزار از زبان فارسی پشتیبانی نمی‌کند.

۹. پیشنهادهایی در جهت تسریع دستیابی به یک سیستم اُسی‌آر فارسی

همانطور که ملاحظه شد، فرآیند تبدیل یک تصویر سند به متن قابل ویرایش یا همان «اُسی‌آر»، پیچیده است و دستیابی به یک سیستم «اُسی‌آر» کارآمد فارسی، در درازمدت و با انجام تحقیقات وسیع و متمرکز محقق خواهد شد. در ادامه، راهکارهایی به منظور تسریع در دستیابی به یک سیستم «اُسی‌آر» فارسی پیشنهاد می‌گردد:

- برخی از عملیات پردازشی سیستم «اُسی‌آر» وابستگی کمتری به نوع زبان متن دارند و روش‌های پردازشی مورد استفاده در سیستم‌های «اُسی‌آر» لاتین، مستقیماً یا با انجام تغییرات جزئی برای آن‌ها قابل اعمال هستند. به عنوان مثال، بسیاری از الگوریتم‌های پیشنهادشده در

مقالات برای عملیات پیش‌پردازش (نظیر دوگانی‌کردن تصویر، کاهش نویز، تصحیح کجی صفحه، و ...) برای هر متنی صرف نظر از نوع زبان آن قابل استفاده می‌باشند. همچنین بسیاری از ایده‌های ارائه‌شده برای قطعه‌بندی بیرونی، بازشناسی، آموزش و پس‌پردازش متون لاتین را می‌توان با انجام یکسری تغییرات، برای متون فارسی نیز استفاده نمود.

- گلوگاه یک سیستم «اُسی‌آر» فارسی و نقطهٔ تمایز آن با «اُسی‌آر» لاتین، عملیات قطعه‌بندی درونی است که چنانچه بتوان بخوبی آن را انجام داد، بجرأت می‌توان عنوان نمود که نیمی از راه را طی نموده‌ایم. بنابراین تیمی که قرار است در زمینهٔ «اُسی‌آر» فعالیت نماید، بجا است که تمرکز اصلی خود را بر این قسمت معطوف دارد.

- یک سیستم «اُسی‌آر» حجم وسیعی از پردازش‌ها و بلوک‌های عملیاتی را می‌طلبد؛ هم از جنبهٔ طراحی الگوریتم و هم از لحاظ برنامه‌نویسی. بنابراین یک راهکار مناسب برای حصول هرچه سریع‌تر یک خروجی قابل قبول این است که تا حد امکان از کیت‌های برنامه‌نویسی (اس‌دی‌کا) آمادهٔ موجود استفاده گردد. به عنوان مثال کیت‌هایی (با عناوین Raster Imaging و Document Imaging) موجود هستند که امکان به‌کارگیری طیف بسیار وسیعی از عملیات پردازشی مورد نیاز برای یک سیستم «اُسی‌آر» از قبیل پشتیبانی از انواع مختلف قالب‌بندی فایل، قابلیت بارگذاری و نمایش و ویرایش و ذخیره‌سازی تصاویر، اخذ تصاویر از اسکنرها و دستگاه‌های تصویربرداری، بسیاری از تکنیک‌های پرکاربرد برای پیش‌پردازش تصاویر (نظیر تکنیک‌های زدودن زوائد و نویز از تصاویر، پردازش تصاویر با انواع فیلترهای دیجیتالی، تصحیح کج‌شدگی، ...)، امکانات مختلف برنامه‌نویسی برای کار با تصاویر (نظیر شیءنگاری روی تصاویر[۸۸]، تسهیلاتی به‌منظور کار با تصاویر در پایگاه‌های داده[۸۹]، مدیریت تصاویر تحت اینترنت[۹۰] و شبکه، امکانات مختلف برای چاپ‌گرفتن از تصاویر، ...)، تحلیل ساختاری تصویر، «اُسی‌آر» لاتین و بسیاری از امکانات دیگر را شامل می‌شوند. این کیت‌ها عمدتاً با صرف هزینه، نیروی انسانی و وقت بسیار زیاد توسط شرکت‌های نرم‌افزاری سرشناس و به صورت بهینه تولید می‌شوند و در نتیجه در صورتی که کیت‌های آمادهٔ موجود قادر باشند بخشی از نیازهای ما را مرتفع سازند، به‌کارنگرفتن آن‌ها و سعی در پیمودن مسیر تولید این کیت‌ها از نقطهٔ صفر، چندان عاقلانه به نظر نمی‌رسد. به عنوان مثال شرکت «صخر» نیز در نرم‌افزار «اُسی‌آر» خود، از کیت برنامه‌نویسی ImageGear ساخت شرکت «اکیوسافت» برای انجام برخی عملیات پردازشی بهره گرفته است. شرکت‌های «لیدتکنولوجیز» و «اکیوسافت» (Web: AccuSoft) معروف‌ترین محصولات را در این زمینه ارائه می‌دهند.

منابع

Arica Nazif ,Yamin-Vural Fatos T., "An overview of character recognition based focused on off-line handwriting", **IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews,** Vol. 31, No. 2, May 2001.

Jain, A. K. and Yu, B., "Document representation and its application to page decomposition", **IEEE Trans. Pattern Anal. Machine Intell.**, vol. 20, pp. 294–308, Mar. 1998.

Mantas, J., "An overview of character recognition methodologies," **Pattern Recognit**., vol. 19, no. 6, pp. 425–430, 1986.

Mori, S., Suen, C. Y., and Yamamoto, K., "Historical review of OCR research and development", **Proc. IEEE**, vol. 80, pp. 1029–1057, July 1992.

Mori, Shunji; Nishida, Hirobumi; Yamada Hiromitsu, **Optical Character Recognition**, Wiley-Interscience, 1999.

Nagy, G., "Twenty years of document image analysis in PAMI", **IEEE Transactions on Pattern Analysis and Machine Intelligence,** Vol. 22, pp. 38-62, Jan 2000.

O'Gorman, Lawrence; Kasturi, Rangachar, *Document Image Analysis*, IEEE Computer Society Press, Los Alamitos, 1995.

Suen, C. Y., Tappert, C. C., and Wakahara, T., "The state of the art in online handwriting recognition", **IEEE Trans. Pattern Anal. Machine Intell**., vol. 12, pp. 787–808, Aug. 1990.

Theodoridis, Sergios; Koutroumbas, Konstantinos; Smith Ricky, **Pattern Recognition**, Academic Press, 1st edition, January 15, 1999.

Trier, O. D.; Jain, A. K. and Taxt, T., "Feature extraction methods for character recognition - A survey", **Pattern Recognition** 29, pp. 641-662, 1996.

Web Site: ScanSost company, http://www.omnipage.com/

Web Site: ABBYY company, http://www.abbyy.com/

Web Site: I.R.I.S company, http://www.irislink.com/

Web Site: Sakhr company, http://www.sakhr.com/

Web Site: LEAD Technologies company, http://www.leadtools.com/

Web Site: AccuSoft company, http://www.accusoft.com/

اردشیربهرستاقی، غلامرضا. **شناسایی ساختاری حروف دست‌نویس فارسی**، پایان‌نامهٔ کارشناسی ارشد، به راهنمایی: کبیر، احسان‌الله ، دانشکدهٔ مهندسی برق، دانشگاه تربیت‌مدرس، ۱۳۷۲.

اسدی، سعید. **بازشناسی حروف فارسی با استفاده از شبکه عصبی چند جمله‌ایهای جداکننده**، پایان‌نامهٔ کارشناسی ارشد، به راهنمایی: فتحی، محمود، دانشکدهٔ مهندسی کامپیوتر، دانشگاه صنعتی اصفهان، ۱۳۷۷.

امیری، امیررضا. **استخراج متن چاپی از تصاویر گرافیکی**، پایان‌نامهٔ کارشناسی ارشد، به راهنمایی: کبیر، احسان‌الله، دانشکدهٔ مهندسی برق، دانشگاه تربیت‌مدرس، ۱۳۸۱.

بحری، پیمان. **شناسایی حروف دست‌نویس فارسی به کمک شبکه عصبی فازی**، پایان‌نامهٔ کارشناسی ارشد، به راهنمایی: فتحی، محمود، دانشکدهٔ مهندسی کامپیوتر، دانشگاه علم و صنعت ایران، ۱۳۷۶.

بنی‌اسدی، امیرعلی. **تشخیص حروف دست‌نویس فارسی به وسیله سیستم هایبرید نور و فازی،** پایان‌نامهٔ کارشناسی ارشد، به راهنمایی: ساداتی، ناصر، دانشکدهٔ مهندسی برق، دانشگاه صنعتی شریف، ۱۳۷۳.

تیمساری، بیژن. **بازشناسی حروف در کلمات تایپ‌شده فارسی با استفاده از روش مورفولژی،** به راهنمایی: فهیمی، حمید، دانشکدهٔ مهندسی برق، دانشگاه صنعتی اصفهان، ۱۳۷۱.

ثانی، رویا. **بازشناسی حروف دست‌نویس فارسی با استفاده از شبکه عصبی مصنوعی،** پایان‌نامهٔ کارشناسی ارشد، به راهنمایی: فهیمی، مهرداد، دانشکدهٔ مهندسی کامپیوتر، دانشگاه صنعتی شریف، ۱۳۷۳.

رادمهر، مهدی. **درک پیکربندی هندسی یک صفحه متن چاپی،** پایان‌نامهٔ کارشناسی ارشد، به راهنمایی: کبیر احسان‌الله ، دانشکدهٔ مهندسی برق، دانشگاه تربیت‌مدرس، ۱۳۷۲.

دهقانی، علیرضا. **بازشناسی حروف مجزای دست‌نویس فارسی با استفاده از مدل پنهان مارکف با چگالی پیوسته و ایده ترکیب چند سیستم خبره،** پایان‌نامهٔ کارشناسی ارشد، به راهنمائی: مسندی‌شیرازی، محمدعلی، دانشکدهٔ مهندسی برق، دانشگاه شیراز، ۱۳۷۹.

رضوی، محمد. **خواندن اتوماتیک فرمهای انتخاب درس،** پایان‌نامهٔ کارشناسی ارشد، به راهنمایی: کبیر، احسان‌الله، دانشکدهٔ مهندسی برق، دانشگاه تربیت‌مدرس، ۱۳۷۵.

رفیعی‌کراچی، شعبانعلی. **شکستن کلمات تایپ شده به حروف در رسم‌الخطهای مختلف،** پایان‌نامهٔ کارشناسی ارشد، به راهنمایی: کبیر، احسان‌الله، دانشکدهٔ مهندسی برق، دانشگاه تربیت‌مدرس، ۱۳۷۳.

شاه‌حسینی، علی. **شناسایی حروف دست‌نویس فارسی با استفاده از شبکه عصبی،** پایان‌نامهٔ کارشناسی ارشد، به راهنمایی: کبیر، احسان‌الله، دانشکدهٔ مهندسی برق، دانشگاه تربیت‌مدرس، ۱۳۷۴.

صدوقی‌یزدی، هادی. **پیش پردازش برای بازشناسی متون فارسی،** پایان‌نامهٔ کارشناسی ارشد، به راهنمایی: کبیر، احسان‌الله، دانشکدهٔ مهندسی برق، دانشگاه تربیت‌مدرس، ۱۳۷۵.

صفابخش، رضا «آنالیز اسناد - تصحیح کج‌شدگی اسناد»، طرح ملی بازشناسی متن فارسی، گروه (ب)، بهار ۱۳۷۸.

عباسیان، کریم. **بازشناسی برخط نویسه‌های فارسی،** پایان‌نامهٔ کارشناسی ارشد، به راهنمایی: کبیر، احسان‌الله، دانشکدهٔ مهندسی برق، دانشگاه تربیت‌مدرس، ۱۳۷۶.

عزمی، رضا. **بازشناسی متون چاپی فارسی،** پایان‌نامهٔ دکتری، به راهنمایی: کبیر، احسان‌الله، دانشکدهٔ فنی و مهندسی، دانشگاه تربیت‌مدرس، ۱۳۷۸.

کبیر، احسان‌الله؛ صفابخش، رضا؛ منهاج، محمدباقر؛ عزمی، رضا؛ احسانی، محمدسعید؛ مسروری، کیوان؛ عبدالله‌زاده، احمد؛ رضوی، محمد؛ **طرح ملی بازشناسی متون چاپی فارسی و حجم محدودی از کلمات دست‌نویس،** ۱۳۷۷.

مرتضی‌پور، حمیدرضا. **قطعه‌بندی بر خط کلمات دست‌نویس فارسی**، پایان‌نامهٔ کارشناسی ارشد، به راهنمایی: کبیر، احسان‌الله، دانشکدهٔ مهندسی برق، دانشگاه تربیت‌مدرس، ۱۳۷۸.

مسروری، کیوان. **بازشناسی حروف دست‌نویس فارسی با روش فازی**، پایان‌نامهٔ کارشناسی ارشد، به راهنمایی: کبیر، احسان‌الله، دانشکدهٔ فنی و مهندسی، دانشگاه تربیت‌مدرس، ۱۳۷۳.

مسروری، کیوان. **شناسایی برون خط کلمات دست‌نویس فارسی در یک مجموعه محدود**، پایان‌نامهٔ دکتری، به راهنمایی: کبیر، احسان‌الله، دانشکدهٔ فنی و مهندسی، دانشگاه تربیت‌مدرس، ۱۳۷۹.

مقدم‌تبریزی، کاوه. **تشخیص الگوی تصویری در محیط پیچیده**، پایان‌نامهٔ کارشناسی ارشد، به راهنمایی مشیری، بهزاد، دانشکدهٔ فنی، دانشگاه تهران، ۱۳۷۵.

مقسمی، حمیدرضا. **تشخیص حروف تایپی فارسی با روش ساختاری**، پایان‌نامهٔ کارشناسی ارشد، به راهنمایی: فتحی، محمود، دانشکدهٔ مهندسی کامپیوتر، دانشگاه علم و صنعت ایران، ۱۳۷۶.

نحوی، منوچهر. **استخراج کدپستی از آدرسهای تایپی و دست‌نویس**، پایان‌نامهٔ کارشناسی ارشد، به راهنمایی: کبیر، احسان‌الله، دانشکدهٔ فنی و مهندسی، دانشگاه تربیت‌مدرس، ۱۳۷۶.

نظام‌آبادی‌پور، حسین. **پیش‌پردازش متون چاپی فارسی برای جداسازی حروف**، پایان‌نامهٔ کارشناسی ارشد، به راهنمایی: کبیر، احسان‌الله، دانشکدهٔ فنی و مهندسی، دانشگاه تربیت‌مدرس، ۱۳۷۹.

نمازی، مهدی. **شناسایی حروف تایپی فارسی با قلمهای متفاوت به کمک شبکه عصبی فازی**، پایان‌نامهٔ کارشناسی ارشد، به راهنمایی: فائز، کریم، دانشکدهٔ مهندسی برق، دانشگاه صنعتی امیرکبیر، ۱۳۷۳.

پایگاه وب: شرکت داده‌پردازان دوران: http://www.douran.com

پایگاه وب: نرم‌افزار تشخیص متون تایپی فارسی دوران.
http://www.douran.com/DesktopDefault.aspx?TabID=3526&Alias=DouranPortal&Lang=fa-IR

پی‌نوشت

[1]. Optical Character Recognition (OCR
[2]. Document Image Analysis (DIA)
[3]. Image processing
[4]. Statistical pattern recognition
[5]. Artificial intelligence
[6]. Classification
[7]. Sensor
[8]. Feature extraction
[9]. Feature vector
[10]. Handwritten
[11]. Scanner
[12]. Tyuring

پژوهشگاه علوم انسانی و مطالعات فرهنگی
پرتال جامع علوم انسانی

[13]. Template matching
[14]. Low level features
[15]. binary
[16]. Constrained
[17]. Digitizers
[18]. Online
[19]. Semantics
[20]. Offline
[21]. Resolution
[22]. Neural Networks
[23]. Hidden Markoff Model
[24]. Fuzzy Set reasoning
[25]. Natural Language Processing
[26]. Omni font
[27]. Internal Segmentation
[28]. Representation
[29]. Recognition
[30]. Preprocessing
[31]. External Segmentation
[32]. Postprocessing
[33] به عنوان نمونه می‌توان به پایان‌نامه‌های «اردشیربهرستاقی» (۱۳۷۲)، «اسدی» (۱۳۷۷)، «امیری» (۱۳۸۱)، «بحری» (۱۳۷٦)، «بنی‌اسدی» (۱۳۷۳)، «تیمساری» (۱۳۷۱)، «ثانی» (۱۳۷۳)، «رادمهر» (۱۳۷۲)، «دهقانی» (۱۳۷۹)، «رضوی» (۱۳۷۵)، «رفیعی کراچی» (۱۳۷۳)، «شاه‌حسینی» (۱۳۷٤)، «صدوقی‌یزدی» (۱۳۷۵)، «عباسیان» (۱۳۷٦)، «عزمی» (۱۳۷۸)، «مرتضی‌پور» (۱۳۷۸)، «مسروری» (۱۳۷۳)، «مسروری» (۱۳۷۹)، «مقدم تبریزی» (۱۳۷۵)، «مقسمی» (۱۳۷٦)، «نحوی» (۱۳۷٦)، «نظام‌آبادی‌پور» (۱۳۷۹) و «نمازی» (۱۳۷۳) مراجعه کرد.
[34]. Digitizer
[35]. Binerization
[36]. Denoising
[37]. Smoothing
[38]. Thinning
[39]. Speckle
[40]. Disconnected line segments
[41]. Dilation
[42]. Erosion
[43]. Skew
[44]. Baselines
[45]. Alignment
[46]. Curvilinear Distortion
[47]. Histogram
[48]. Projection profile
[49]. Nearest neighbors clustering
[50]. Cross correlation
[51]. Hough transform
[52]. Resampling

پژوهشگاه علوم انسانی و مطالعات فرهنگی
پرتال جامع علوم انسانی

[53]. Slant Normalization
[54]. Scaling
[55]. Contour Smoothing
[56]. Thresholding
[57]. Gray level
[58]. Script
[59]. Context models
[60]. Indexing
[61]. Cursive
[62]. Structural analysis
[63]. Functional analysis
[64]. Explicit Segmentation
[65]. Implicit Segmentation
[66]. Segmentation-based approach
[67]. Segmentation-free approach
[68]. Resolution
[69]. Point-spread Function
[70]. Reject Option
[71]. Feature Extraction
[72]. Classification and Recognition
[73]. spelling Checker
[74]. PNG, BMP, TIFF, JPEG 2000, JPEG, etc.
[75]. Txt(text), RTF(Microsoft Word), XLS(Microsoft Excel), PPT(Microsoft Power Point), HTML,XML, PDF(Adobe Acrobat), etc.
[76]. Automatic page layout analysis
[77]. Manual page layout analysis
[78]. Word Processor
[79]. Batch processing
[80]. Automatic Reader v8
[81]. Sakhr
[82]. Dongle
[83]. Gold
[84]. Platinum
[۸۵]. «اس‌دی‌کا» عبارت است از یک کیت برنامه‌نویسی شامل کدها، توابع کتابخانه‌ای و فایل‌های راهنما که توسط برخی شرکت‌های تولیدکنندهٔ نرم‌افزار ارائه می‌شود و از طریق برنامه‌نویسی، امکان استفاده از قابلیت‌های آن نرم‌افزار را در دیگر محصولات نرم‌افزاری فراهم می‌آورد. «اس‌دی‌کا»ی نرم‌افزار «اتوماتیک ریدر» در سه فرمت مختلف تولید شده است: OLE automation، COM object و DLL
[86]. Diacritics
[87]. Cashida
[88]. Image Annotation
[89]. Database Imaging
[90]. Internet Imaging

پرتال جامع علوم انسانی