



نقش فناوری ا.سی.آر (OCR) در ایجاد کتابخانه دیجیتالی

مهندس بهاره برهانی یزدی^۱

چکیده

امروزه افزایش استفاده از اطلاعات چندرسانه ای همانند متن، تصویر و تقاضا برای ذخیره، جستجو و بازیابی این اطلاعات، بسیار مورد توجه پژوهشگران قرار گرفته است و از این رو تحقیقات گسترده ای در زمینه بازشناسی نوشتار انجام شده و می شود. امروزه با توجه به پیشرفتهای حاصل شده در این زمینه، فناوری بازشناسی خودکار تحت عنوان بازشناسی نوری حروف (OCR)^۲ شکل گرفته است.

در مقاله حاضر سعی شده است با توجه به تحقیقات شرکت فناوری اطلاعات و ارتباطات رضوی در زمینه فناوری هوشمند ا.سی.آر، مبانی نظری و جنبه های کاربردی ا.سی.آر در نرم افزارهای کتابخانه های دیجیتالی و قابلیت ها، نقاط قوت و ضعف این فناوری در فارسی و الگوریتم های ارائه شده در این زمینه و در انتها شیوه انتخاب راهکارهای مناسب به منظور حصول یک نرم افزار ا.سی.آر فارسی در جهت تسهیل و عدم دوباره کاری در بسیاری از فرآیندهای کتابخانه های دیجیتالی ارائه گردد.
کلیدواژه های موضوعی: ا.سی.آر، سیستم بازشناسی حروف، کتابخانه دیجیتالی.

۱. مقدمه

در چند دهه گذشته مساله بازشناسی الگوهای نوشتاری شامل حروف ارقام و سایر نمادهای متداول در اسناد مکتوب، به زبانهای مختلف توسط گروههای مختلفی از محققان مورد مطالعه قرار گرفته است و نتیجه این تحقیقات منجر به پیدایش مجموعه ای از روشهای سریع مطمئن موسوم به ا.سی.آر شده که به معنای بازشناسی نوری حروف است. از لحاظ کاربردی، ا.سی.آر برنامه ای است که با استفاده از روشهای گوناگون هوش مصنوعی، اسناد، مدارک، کتابها و سایر مکتوبات چاپی یا تایپ شده و حتی دستنویس را به متن قابل ویرایش و جستجو تبدیل می کند. (C. Liu & K. Nakashima, 2004)

با توجه به آنکه حجم زیادی از اسناد کاغذی موجود، توسط اسکنرها یا دوربین ها به اسناد تصویری دیجیتالی تبدیل می شوند؛ ذخیره سازی، بازیابی و مدیریت کارآمد این آرشیوهای تصویری، در بسیاری از برنامه ها نظیر اتوماسیون اداری و کتابخانه های دیجیتالی اهمیت فراوانی دارند. در نتیجه دستیابی به الگوریتمهای مؤثر به منظور آنالیز تصویری اسناد، یک نیاز اساسی به حساب می آید. قابل ذکر است که ا.سی.آر به عنوان یکی از شاخه های فعال در شناسایی الگوی آنالیز تصویر اسناد و پردازش تصاویر می باشد که تصاویر حاوی متن را به متن قابل جستجو تبدیل می کند.

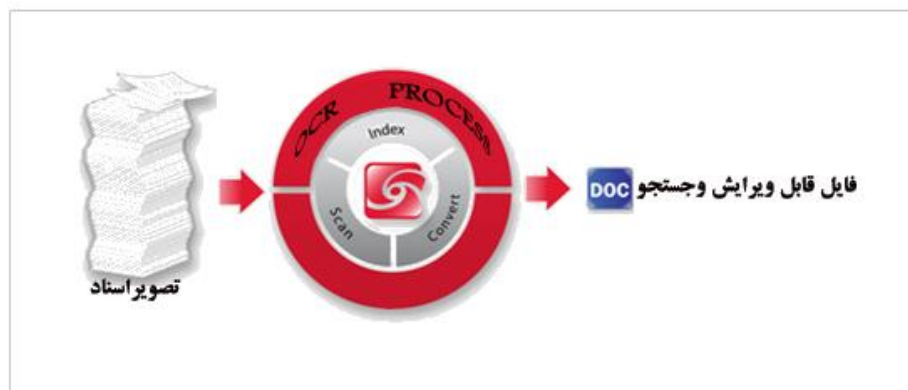
^۱ - بهاره برهانی یزدی، کارشناس ارشد شرکت فناوری اطلاعات و ارتباطات رضوی bahar_274_m@yahoo.com.ph



۲. معرفی بازشناسی نوری حروف

هنگامی که یک سند متنی اسکن می‌شود، رایانه این متن را به صورت یک تصویر گرافیکی تشخیص می‌دهد. در نتیجه کاربران قادر نخواهند بود که متن موجود در تصویر سند را ویرایش کرده، مطالعه نموده و یا مورد جستجو قرار دهند. اینجاست که با کمک فناوری ا.سی.آر محتویات تصاویر قابل شناسایی شده و به صورت یک فایل در رایانه ذخیره می‌گردد.

لذا چنین قابلیتی موجب حذف یک فرایند وقت گیر، یعنی وارد کردن دوباره اطلاعات موجود در نسخه چاپی می‌شود، ولی اگر نسخه‌ی اسکن شده ناقص باشد یا حاوی " نشانه‌های زیر و زبری " یا نویسه‌های غیرقابل تشخیص باشد، ممکن است نتیجه کار، قابل پیش بینی نباشد. (R.Plamondon&N.Srihari,2000) شکل (۱) بیانگر چرخه سیستم بازشناسی نوری حروف است. مطابق شکل در سیستم بازشناسی نوری حروف تصاویر، اسکن شده و سپس از طریق الگوریتم‌های پردازش تصویر - که در بخش بعدی مقاله به صورت کامل تشریح می‌شوند- سیستم بازشناسی نوری حروف، متن اسکن شده را خوانده محتویات آن را شناسایی نموده و در مرحله پایانی به صورت یک فایل دیجیتالی قابل ویرایش از اطلاعات سند (مثلاً به فرمت MS- Word) تبدیل می‌نماید.



شکل ۱- شمای کلی سیستم ا.سی.آر

۳. تاریخچه سیستم‌های ا.سی.آر

در سال ۱۹۲۹، تاوشچ^۱ در آلمان و در سال ۱۹۳۳ هندی در آمریکا ابداعاتی در زمینه سیستم بازشناسی حروف ثبت نمودند. اینها اولین ایده‌های شناسایی حروف هستند. این روش، تطبیق الگو نام دارد و به این صورت کار می‌کند که به هر حرف، نور تابیده می‌شود و نور باز تابیده شده از حروف، از قالب‌های مکانیکی عبور داده می‌شود. هرگاه نوری از قالب عبور نکند، حرف تشخیص داده می‌شود. از سال ۱۹۵۱، در آمریکا ایده ا.سی.آر به عنوان یک پدیده قابل پیاده‌سازی پذیرفته شد. از این زمان به بعد تحقیقات گسترده‌ای بر روی این فناوری

¹ Taushech



انجام گرفته است و در نتیجه سیستمهای تجاری عرضه شده اند که دارای قابلیت‌های خوبی هستند. سیستمهای تجاری عرضه شده را می توان به سه نسل زیر تقسیم بندی نمود: (Gonzalez & Woods, 2002)

۱-۳: سیستمهای نسل اول

سیستمهای نسل اول در اوایل دهه ۱۹۶۰ به بازار آمدند که نتیجه تحقیقات دهه ۱۹۵۰ بودند. این سیستم ها فقط قابلیت تشخیص کاراکترهای خاص با اندازه و فونت مشخصی را داشتند. در آنها از روشهایی استفاده شده بود که نسبت به تغییر جا کاراکترها، اندازه و دوران آنها فوق العاده حساس بودند. شاخص ترین سیستم این نسل، NCR 420 بود که می توانست اعداد و پنج نماد دیگر را تشخیص دهد. (Gonzalez & Woods, 2002)

۲-۳: سیستم های نسل دوم

سیستمهای نسل دوم در اواسط دهه ۱۹۶۰ به بازار آمدند و می توانستند بعضی از کاراکترهای دست نویس مانند اعداد را تشخیص دهند. RETINA یکی از این سیستم ها بود. (Gonzalez & Woods, 2002)

۳-۳: سیستم های نسل سوم

سیستمهای نسل سوم مربوط به اواخر دهه ۱۹۶۰ هستند که باهدف تشخیص کاراکترهایی با کیفیت چاپ پایین و همچنین متون دست نویس لاتین ساخته شدند. این نسل تا به امروز هم ادامه دارد و فعالیتهای گسترده ای در این زمینه در حال انجام است. امروزه سیستمهای ا.سی.آر قادر به تشخیص دقیق کاراکترهای تایپی لاتین با انواع فونتها و در اندازه های متفاوت هستند (با دقتی در حدود ۹۹٪) ولی هنوز در مورد تشخیص متون دست نویس لاتین و یا فونتهایی که در آنها از خطوط خمیده استفاده می شود (مثل فارسی و عربی) مشکلات زیادی وجود دارد (Gonzalez & Woods, 2002)

۴. لزوم بکارگیری سیستم های ا.سی.آر در سازمانها

امروزه در اکثر کشورهای پیشرفته استفاده از سیستمهای ا.سی.آر در فرایندهای اداری کاملاً مرسوم است. حتی در بعضی از این کشورها به منظور تسریع بهینه سازی و کم کردن هزینه فرایند ورود اطلاعات به رایانه استفاده از این سیستمها به صورت قانون درآمده است و ادارات و سازمانهای دولتی ملزم به استفاده از این نوع نرم افزار هستند. البته با توجه به طبیعت پیدایش این سیستم لزوم استفاده از آن در سازمانهای دولتی و ادارات امری واضح و آشکار است؛ چرا که این سیستم یک نرم افزار پایه و اساسی برای تمام سیستمهای اتوماسیون اداری می باشد. از آنجا که تمام کشورهای پیشرفته به سوی ایجاد یک دولت الکترونیک پیش می روند لذا ورود چنین سیستمی را به بدنه دولت خود به راحتی پذیرفته اند. با وجود آنکه سیستمهای الکترونیکی امروزی تا حدودی بوروکراسی را در فرایندهای اداری کاهش داده است؛ ولی طبق پیش بینی های انجام شده تا قرن آینده استفاده از اسناد



مکتوب از فرایندهای اداری حذف نخواهد شد. به همین دلیل ورود اطلاعات اسناد مکتوب موجود فعلی و همچنین اسنادی که به روز تولید می شوند وجود چنین سیستمی را می طلبد. (Chang&Thouin,2006)

۵. سیستم های ا.سی.آر در کتابخانه های دیجیتالی

در عصر حاضر با ظهور شاخه های جدید علم و گسترش شاخه های موجود و نیز رشد تعداد دانشجویان و محققان، کتابخانه های سنتی دیگر به تنهایی گنجایش نگهداری حجم عظیم کتابهای موجود و نیز توان پاسخگویی به خیل افراد متقاضی استفاده کننده از این منابع را ندارند.

در کتابخانه های سنتی در زمان واحد تنها یک نفر می تواند از یک جلد کتاب استفاده نماید. به عبارت دیگر به ازای هر نفر متقاضی یک عنوان کتاب، می بایست یک جلد از آن در کتابخانه موجود باشد که این مساله در عمل غیر ممکن است. از سوی دیگر در کتابخانه های سنتی مشکلاتی از قبیل محدودیت فضای نگهداری و گسترش آن و خطر وقوع حوادثی مانند آتش سوزی، زلزله، جنگ و ... همواره وجود دارد یکی از کاربردهای فناوری اطلاعات و ارتباطات که می تواند نقش مهمی در زمینه گسترش ارائه و اشتراک منابع و اطلاعات داشته باشد، در ظهور کتابخانه دیجیتالی متجلی شده است. کتابخانه دیجیتالی مجموعه ای از اطلاعات سازماندهی شده در قالب الکترونیکی است که در آن :

- اطلاعات دیجیتالی باید مبتنی بر اصول مجموعه سازی، گردآوری یا تولید شده باشند.
 - اطلاعات دیجیتالی باید به صورت منسجم و مناسبی نظیر دیگر منابع اطلاعاتی در کتابخانه ها توسط استفاده کنندگان قابل دسترس و بازیابی باشند.
 - اطلاعات دیجیتالی باید در مدت زمان طولانی به مثابه منابع اطلاعاتی پایدار قابل دسترس باشند.
- اما در ایران همواره یکی از بزرگترین مسائل فراروی سیستم های کتابخانه های دیجیتالی انتقال اطلاعات مندرج از روی اسناد موجود در کتابخانه ها به سیستم های رایانه ای است. در بسیاری از موارد از روش داده آمایی دستی این فرایند به انجام می رسد که علاوه بر صرف زمان و هزینه زیاد "اطلاعات داده آمایی" شده نیز از دقت کافی برخوردار نبوده و به موقع در دسترس نمی باشند. در راستای رفع این مشکل سیستم ا.سی.آر ضمن کاهش زمان و هزینه داده آمایی، دقت اطلاعات وارد شده به سیستم های رایانه ای را نیز افزایش می دهد. بدین ترتیب با در اختیار داشتن اطلاعات دقیق و به موقع بسیاری از موارد از جهات مختلف در کتابخانه های دیجیتالی رفع می گردد.

۶. مزایای بکارگیری سیستم های ا.سی.آر در کتابخانه های دیجیتالی

مزایای بکارگیری سیستمهای ا.سی.آر در کتابخانه های دیجیتالی طبق نظر تیمساری (۱۳۷۱) عبارتند از :



- افزایش چشمگیر سرعت دسترسی به اطلاعات: زیرا در متن بر خلاف تصویر، امکان جستجو و ویرایش وجود دارد.
- کاهش فضای ذخیره سازی: زیرا حجم فایل متنی استخراج شده از یک تصویر، معمولاً بسیار کمتر از حجم خود فایل تصویری است.
- امکان استفاده از قابلیت‌های پیش پردازشی آنها، ویرایشگری و تغییر فرمت داده ها
- عدم نیاز به مرتب سازی دستی
- بهبود سرویس دهی به مراجعین

۷. جنبه های کاربردی سیستم ا.سی.آر در نرم افزارهای جامع کتابخانه الکترونیکی

با توجه به ماهیت سیستم ا.سی.آر و قابلیت‌های بسیار آن، کتابخانه های دیجیتالی می توانند در حوزه فعالیت و چارچوب عملکرد خود از این سیستم بهره ببرند. طبق آمارگیری انجام شده در کشورهای پیشرفته ۶۵ درصد عملیات تایپ اسناد در سازمانها و ادارات، دوباره کاری می باشد (عزمی، ۱۳۷۸). فقط با توجه به این درصد بالا در کشورهای پیشرفته به راحتی به لزوم وجود این سیستم می توان پی برد. بدین منظور در زیر برخی از کاربردهای سیستم ا.سی.آر در کتابخانه های دیجیتالی به صورت موردی ذکر گردیده است:

۷-۱: مدیریت اسناد الکترونیکی

با بکارگیری ابزار های آرشیو، یک بایگانی از تصاویر با قابلیت های جستجو روی فیلدهای مورد نظر خواهید داشت و در واقع در سیستم مدیریت اسناد الکترونیکی با بکارگیری ا.سی.آر مشکلات مربوط به ویرایش و جستجوی اسناد حل خواهد شد و نیازی به صرف هزینه های زیاد برای ورود اطلاعات مجدد در سیستم نخواهید داشت. در ذیل برخی از مزایای بکارگیری فناوری ا.سی.آر در سیستم مدیریت اسناد الکترونیکی مطرح شده است که عبارتند از:

- دریافت و بازشناسی متن تصاویر اسناد با فرمت‌های مختلف به صورت تک قاب یا چند قاب
- مدیریت و بازشناسی دسته‌های تصویری بزرگ از اسناد
- تقسیم‌بندی خودکار تصویر سند به ناحیه‌های متنی، تصویری و جدولی
- ویرایشگر متن اسناد با امکانات لازم جهت تطبیق اصل تصویر با متن بازشناسی شده
- تولید خروجی های متفاوت با فرمت WORD و یا PDF



۲-۷: کتابهای الکترونیکی

- با فناوری اُ.سی.آر تصاویر صفحات کتاب به متن رایانه ای تبدیل شده و امکان ویرایش متن، جستجوی متن، چاپ مجدد متن با کیفیت بالا و انتقال الکترونیکی سریع آن (با توجه به آنکه حجم اطلاعات متنی بسیار کمتر از اطلاعات تصویری است) فراهم می شود. این مورد را نیز در نظر بگیرید که به منظور ایجاد نسخ الکترونیکی از کتاب های کتابخانه ای که ده ها هزار کتاب و میلیون ها صفحه مطلب در آن قرار دارد چه هزینه های مالی و زمانی هنگفتی را باید برای تایپ این منابع صرف نمود.
- اُ.سی.آر به ویژه برای دیجیتالی نمودن منابع و نسخ خطی بسیار مفید است، چرا که برای نسخ خطی اصالت منبع با حفظ شکل و قالب اصلی متون و نه به صورت تایپ شده باید حفظ شود. در مراحل بعد می توان با این برنامه مندرجات متون نسخ خطی را کاوش نمود. بنابراین می توان گفت در واقع سیستم اُ.سی.آر، کلید اصلی ایجاد یک کتابخانه دیجیتالی می باشد که ورود متون کتابها، مجلات، مقالات و منابع مکتوب را در این کتابخانه ممکن می سازد.

۳-۷: روزنامه ها و نشریات الکترونیکی

استفاده از اُ.سی.آر می تواند نقش عمده ای در کاهش هزینه ها و زمان بری ورود اطلاعات داشته باشد. اُ.سی.آر قادر به ایجاد بی واسطه فایل های استاندارد از مطبوعات و روزنامه های چاپ شده است.

۴-۷: سیستم اتوماسیون اداری

اُ.سی.آر در سیستم اتوماسیون امکان اصلاح، جستجو و یا استفاده دوباره از متن نامه های اسکن شده در بین مجموعه داده ها را به راحتی فراهم می آورد و از نیازهای اساسی سیستم اتوماسیون به شمار می آید.

۸. انواع سیستم های اُ.سی.آر

در یک تقسیم بندی کلی می توان سیستم های اُ.سی.آر را از لحاظ نوع الگوی ورودی به دو گروه اصلی تقسیم کرد:

الف- سیستم های بازشناسی متون چاپی

ب- سیستم های بازشناسی متون دست نویس

هر یک از این گروه ها تکنیک های خاص خود را دارند. همچنین از جنبه نحوه ورود اطلاعات، سیستم های اُ.سی.آر به دو دسته زیر تقسیم بندی می شوند:



۸-۱: سیستمهای برخط

بازشناسی برخط نوشتار بدلیل راحت تر بودن نوشتن از تایپ کردن، عدم امکان تایپ در بعضی مکان ها، عدم وجود یک صفحه کلید کامل روی رایانه های کوچک و سخت بودن تایپ نویسه ها در بعضی زبان ها به دلیل تعداد زیاد آنها مورد توجه خاصی قرار گرفته است.

در روشهای برخط از داده های یک بعدی استفاده می شود. در حالت برخط می توان قسمتهای مختلف متن را ذخیره کرد و سپس عملیات شناسائی را توسط روشهای برون خطی انجام داد. روشهای برخط دارای مزیت‌های فراوانی هستند که تعدادی از آنها عبارتند از:

۱- ترتیب نوشتن و

۲- جهت و نحوه حرکت قلم موجود که می تواند جهت شناسائی مورد استفاده قرار گیرد. همچنین برداشتن

و گذاشتن قلم نیز می تواند جهت شناسائی ضربه ها مورد استفاده قرار گیرد.

روشهای تشخیص کاراکترها به صورت برخط، به دودسته عمده تقسیم می شوند:

- تحلیل کاراکترها با استفاده از ترکیب که مبتنی بر مدل سازی متون تولید شده دست نویس است.
- و استفاده از روشهای برون خطی جهت شناسائی برخط.

۸-۲: سیستمهای برون خط

باز شناسی برون خط هم شامل بازشناسی نوشتار و هم نوشتار دست نویس می شود ولی بازشناسی برخط فقط در بازشناسی دست نوشته کاربرد دارد. در بازشناسی برون خط، تصویر اسکن شده به عنوان ورودی در نظر گرفته می شود.

روشهای شناسائی برون خطی بر روی تصاویر دوبعدی اعمال می شوند. از قبیل:

- استفاده از ویژگیهای آماری و طبقه بندی کننده های متداول آماری
- استفاده از ویژگیهای ساختاری و طبقه بندی کننده مناسب
- کاربرد منطق فازی
- کاربرد شبکه های عصبی (S.D.Connell, 2000)

۹. مراحل انجام کار سیستم ا.سی. آر در نرم افزار کتابخانه الکترونیکی

الف- پیش پردازش:

شامل کلیه اعمالی که روی سیگنال تصویری خام صورت می گیرند تا موجب تسهیل روند اجرای فازهای بعدی گردند؛ مانند باینری کردن تصویر، حذف نویز، هموارسازی، نازک سازی، تشخیص زبان و فونت کلمات و نظایر اینها.

ب- قطعه بندی:

عبارت است از روشهایی که بخشهای مختلفی همچون پاراگرافها، جملات یا کلمات و حروف را از تصویر سند استخراج مینمایند.

ج- استخراج ویژگیها:



مجموعه کلیه محاسباتی است که روی الگوهای بدست آمده از مرحله پیش پردازش انجام می شود تا بردار ویژگیهای متناظر با هر الگو تعیین گردد.

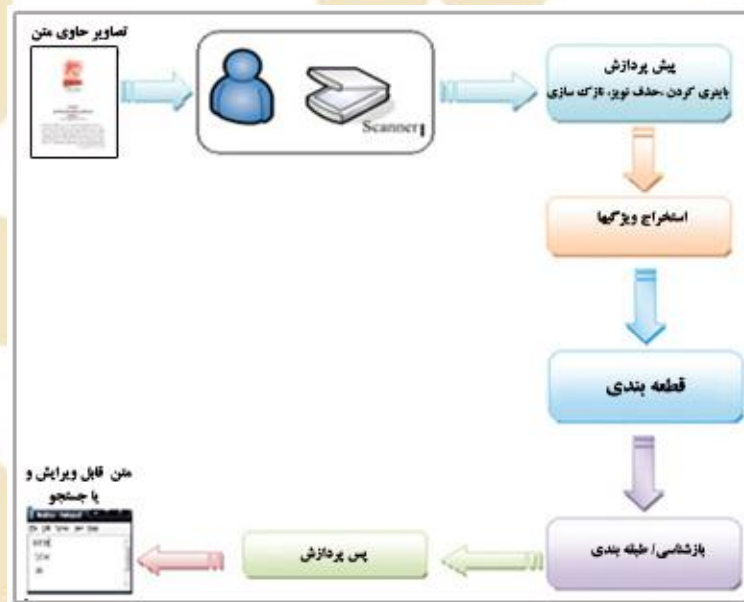
د- طبقه بندی و بازشناسی (با یک یا چند طبقه بندی کننده)

شامل روشهایی برای متناظر ساختن هر یک از الگوهای بدست آمده از مرحله استخراج ویژگیها با یکی از کلاسهای فضای الگوهای مورد بحث است که از طریق کمینه ساختن فاصله بردار ویژگیهای هر الگو نسبت به یکی از بردارهای مرجع موجود در پایگاه داده های سیستم انجام می گیرد.

ه- بکارگیری اطلاعات جانبی (پس پردازش):

مانند مجموعه لغات معتبر، اطلاعات آماری مربوط به رخداد حروف، اطلاعات دستوری و معنایی.

شکل ۲، یک سیستم اُسی.آر را نمایش می دهد. البته به این نکته توجه گردد که بسته به الگوریتم کلی بکار رفته و سطح انتظارات از عملکرد نرم افزار، ممکن است برخی سیستمها فاقد یک یا چندمورد از مراحل فوق باشند. (Arica، ۲۰۰۱)



شکل ۲- بلوک دیاگرام یک سیستم

۱۰. تحلیل ویژگیهای متون فارسی از دیدگاه پردازش سیستمی در اُسی.آر

نگارش فارسی ویژگیهای منحصر به فردی دارد که آن را کاملاً از نگارش لاتین متمایز می سازد . به منظور فعالیت در زمینه اُسی.آر 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:
الف- متون فارسی برخلاف متون لاتین از چپ به راست نوشته می شوند.



ب- در کلمات فارسی برخی از حروف از یک یا دو طرف به حروف مجاور خود اتصال داشته و برخی نیز به صورت مجزا نوشته می شوند. در نتیجه هر کلمه ممکن است شامل یک یا چند بخش متصل باشد که (زیرکلمه) نامیده می شود. شکل (۳-الف)

پ- حروف فارسی می توانند چهار موقعیت مجزا و در نتیجه چهار شکل متفاوت نگارش داشته باشند: حروف ابتدایی، میانی، انتهایی و مجزا. (شکل ۳-ب)

ت- حروف واقع در یک کلمه ممکن است همپوشانی داشته باشند، بدین معنی که نتوان با رسم خطوط عمودی، حروف را بطور کامل از یکدیگر مجزا نمود (شکل ۳-ج)

ث- در برخی از فونتها بعضی از حروف، در دو محل (از یک سمت) به یکدیگر اتصال دارند. (شکل ۳-د)

ج- برخی از حروف بین یک تا سه عدد نقطه دارند که ممکن است در بالا یا پایین بدنه حرف واقع باشند.

چ- بعضی از حروف بدنه مشابه دارند و تفاوت آنها تنها در تعداد و محل قرارگیری نقاط در آنهاست. (شکل ۳-ه)

ح- حروف فارسی می توانند در بالا یا پایین بدنه خود دارای اعراب باشند. سه اعراب - ، - ، - در زبان فارسی، اعرابهای اصلی بوده و اعراب - در برخی کلمات عربی رایج در زبان فارسی دیده می شود. کلمات عربی دارای اعراب - و - در زبان فارسی عمومیت نیافته اند. هر چند کاربرد اعراب در زبان فارسی نسبت به زبان عربی بسیار محدودتر است، اما در مواردی که کلمه ای نامتداول باشد و یا بدلیل تشابه نگارشی آن با کلمه دیگر، تأکید بر تلفظ صحیح آن باشد، مورد استفاده قرار می گیرند.

خورشید	ع ع ع ع	حکم
(الف)	(ب)	(ج)
کا	پ پ ت ت ز ز ی ی ء	با
(د)	(ه)	(و)

شکل ۳- برخی از ویژگیهای نگارش زبان فارسی (یگانه، ۱۳۸۳)

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

د- برخی از حروف شامل همزه هستند ("تا"، "ا"، "و"، "به").

ذ- حروفی که از طرف چپ قابلیت اتصال به حرف مجاور خود را دارند، می توانند به صورت کشیده نوشته شوند.

(شکل ۳- و) (یگانه، ۱۳۸۳)



۱۱. تحقیقات انجام شده در مورد پیاده سازی سیستم های ا.سی.آر فارسی

عواملی چون پیچیدگیهای مربوط به متون فارسی از دیدگاه پردازش سیستمی در ا.سی.آر و علاوه بر آن وجود جمعیت کم کاربر زبان فارسی سبب شده سامانه‌های نویسه‌خوان قدرتمندی برای زبان فارسی نداشته باشیم. البته قابل ذکر است در سالهای اخیر تلاشهای قابل توجهی از سوی برخی شرکتهای فعال در زمینه پردازش تصویر انجام شده که برخی از آنها منجر به محصولات قابل قبولی شده است.

برخی از تحقیقاتی که در زمینه ا.سی.آر فارسی انجام شده به شرح زیر می باشد:

- سیستم بازشناسی حروف دست نویس فارسی بر اساس الگوریتم آموزش فعال با استفاده از کمیته ای از دسته بندها (خیرخواه، ۱۳۸۶)
- پیاده سازی سیستم ا.سی.آر فارسی با استفاده از عملگرهای مورفولوژی (محمد جواد فدایی اسلام، ۱۳۸۵)
- بازشناسی بر خط حروف مجزای فارسی با شبکه فارسی (رضوی؛ کبیر، ۱۳۸۳)
- سیستمی جهت بازشناسی متون چاپی فارسی که با استفاده از رویکرد ترکیبی (بکارگیری هر دو رویکرد مبتنی بر جداسازی کلمات و مبتنی بر بازشناسی کلمه به عنوان یک الگوی واحد) ارائه نموده است. (عزیمی، ۱۳۷۸)
- الگوریتمی برای قطعه بندی برخط کلمات دستنویس فارسی. منظور از قطعه بندی در اینجا یافتن قطعاتی است که بوسیله آنها بتوان کلیه کلمات را ایجاد نمود. بدین معنی که مجموعه کلمات نوشته شده توسط افراد مختلف به اجزاء ساده تری تقسیم می شوند تا برای بازشناسی مورد استفاده قرار گیرند. (مرتضی پور، ۱۳۷۸)
- روشهایی برای شناسایی کلمات دستنویس فارسی در یک مجموعه محدود که بدون قطعه بندی کلمات ارائه شده است. این روشها در سطح کلمات یا زیر کلمات عمل می کنند. (مسروری، ۱۳۷۳)
- تشخیص حروف دست نویس فارسی به وسیله سیستم هایبیرید نور و فازی. (بنیادی، ۱۳۷۳)

نتیجه گیری

با توجه به مطالب ذکر شده و براساس پیچیدگیهای خاص ا.سی.آر و نوشتار و الفبای فارسی و تعدد سبک های نوشتاری فارسی برای اشخاص مختلف که حروف گوناگون را به اشکال متنوع تحریر می کنند و تحقیقات و مطالعات انجام شده، مدتی است نرم افزار ا.سی.آر گسسته فارسی طراحی و نوشته شده و مورد استفاده قرار گرفته است. با بررسی وضعیت کنونی ا.سی.آر در جهان و وضعیت کنونی آن در ایران، ا.سی.آر پیش بینی کننده نیازهای آتی کتابخانه های دیجیتالی است. به این منظور برخی از زیرساختها به منظور ایجاد ساز و کار لازم جهت ارزیابی و کنترل کیفی برنامه‌های نویسه‌خوانی نوری، ایجاد انگیزه در بخش خصوصی و دانشگاهی برای رقابت و حضور فعال تر، پوشش دادن به نیازهای علمی کاربران در سطوح مختلف و بررسی جایگاه و نقش منابع در نویسه‌خوانی نوری و تسهیل حرکت به سمت منابع باز ا.سی.آر مورد نیاز است.



منابع و مآخذ

تیمساری، بیژن (۱۳۷۱). "بازشناسی حروف در کلمات تایپ شده فارسی با استفاده از روش مورفولوژی". پایان نامه کارشناسی ارشد، به راهنمایی: فهیمی، حمید، دانشکده مهندسی برق، دانشگاه صنعتی اصفهان.

خیرخواه، احمد رضا، سعید راحتی (۱۳۸۵). "سیستم بازشناسی حروف با استفاده از قاب بندی در استخراج ویژگیها و بر اساس الگوریتم یادگیری فعال و شبکه های عصبی". ارائه شده در: چهارمین کنفرانس ماشین بینایی و پردازش تصویر ایران، مشهد ۱۳۸۵. مشهد: دانشگاه فردوسی.

رضوی، س.م.؛ کبیرا (۱۳۸۳). "بازشناسی بر خط حروف مجزای فارسی با شبکه فارسی". ارائه شده در: سومین کنفرانس بینایی و پردازش تصویر ایران دانشگاه تهران، تهران اسفند ۱۳۸۳. تهران: دانشگاه تهران.

امید یگانه، مونا (۱۳۸۳). "جداسازی حروف تایی ی مستقل از نوع فونت و اندازه". پایان نامه کارشناسی ارشد، دانشکده مهندسی برق، دانشگاه صنعتی شریف.

مرتضی پور، حمیدرضا (۱۳۷۸). "قطعه بندی بر خط کلمات دستنویس فارسی". پایان نامه کارشناسی ارشد، به راهنمایی: کبیر، احسان الله، دانشکده مهندسی برق، دانشگاه تربیت مدرس.

مسروری، کیوان (۱۳۷۹). "شناسایی برون خط کلمات دست نویس فارسی در یک مجموعه محدود، پایان نامه دکتری، به راهنمایی: کبیر، احسان الله، دانشکده فنی و مهندسی"، دانشگاه تربیت مدرس

فدایی اسلام، محمد جواد (۱۳۸۵)، "سیستم ا.سی.آر فارسی با استفاده از عملگرهای مورفولوژی"، پایان نامه کارشناسی دانشگاه علم و صنعت ایران

عزمی، ر. (۱۳۷۸). "بازشناسی متون چاپی فارسی". پایان نامه دکتری، دانشکده مهندسی برق، دانشگاه تربیت مدرس.

Arica Nazif, Yamin-Vural Fatos T(2001). "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,

C. Liu, K. Nakashima, and H. Sako(2004), "Handwritten digit recognition: investigation of normalization and feature extraction techniques". 2004, vol. 37, n^o2, pp. 265-279 [15 page(s) (article)]

C.-I. Chang, Y. Du, J. Wang, S.-M. Guo, and P. Thouin(2006). "Survey and comparative analysis of entropy and relative entropy thresholding techniques". **Vision, Image and Signal Processing, IEE Proceedings**, vol. 153, no. 6, pp. 837 –850

Gonzalez, Rafael C., Woods, Richard E(2002)"Digital Image Processing". Second Edition, New Jersey: Prentice-Hall.



R.Plamondon,N.Srihari(2000),On_Line and Off_Line Hanwriting Recognition:A Comprehensive Survery,IEEE Transsactions on Pattern Analysis and Machine Intelligence,vol22,pp63-84
S.D.Connell, (2000)"Online Handwriting Recognition Using Multiple Pattern Class Models",Phd Thesis MichiiganState University

