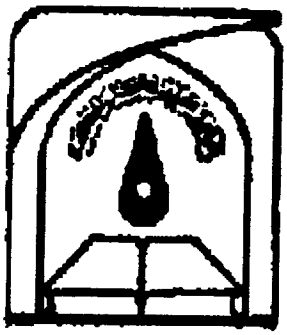
بسم الله

الرحمن

الرحيم



دانشگاه تربیت مدرس

دانشکده فنی و مهندسی

پایان نامه کارشناسی ارشدمهندسی برق

(الکترونیک)

پیش پردازش متون‌چاپی فارسی برا ی جداسازی حروف

حسین نظام آبادی پور

استاد راهنما

دکتر احسان‌اله کبیر

استاد مشاور

دکتر رضا عزمی

تابستان ۱۳۷۹

دانشگاه تربیت مدرس

تاییدیه هیات داوران

آقای حسین نظام‌آبادی‌پور پایان‌نامه ۶ واحدی خود را با عنوان پیش پردازش متون چاپی فارسی برای جداسازی حروف در تاریخ ۷۹/۶/۱۹ ارائه کردند. اعضای هیات داوران نسخه نهائی این پایان‌نامه را از نظر فرم و محتوی تایید و پذیرش آنرا برای تکمیل درجه کارشناسی ارشد رشته مهندسی برق باگرایش الکترونیک پیشنهاد می‌کنند.۹ب۶

اعضای هیات داوران	نام و نام خانوادگی	امضاء
۱ـ استاد راهنما :	آقای دکتر کبیر	
۲ـ استاد مشاور :	آقای دکتر عزمی	
۳ـ استادان ممتحن :	آقای دکتر لطفی‌زاد	
	آقای دکتر صفابخش	
۴ـ مدیر گروه :	آقای دکتر نبوی	
(یا نمایندهٔ گروه تخصصی)		

این نسخه به عنوان نسخه نهایی پایان‌نامه/رساله مورد تأیید است.

امضای استاد راهنما

بسمه تعالی

آیین‌نامهٔ چاپ پایان‌نامه (رساله)های دانشجویان دانشگاه تربیت مدرّس

نظر به اینکه چاپ و انتشار پایان‌نامه (رساله)های تحصیلی دانشجویان دانشگاه تربیت مدرّس، مبیّن بخشی از فعالیتهای علمی - پژوهشی دانشگاه است بنابراین به منظور آگاهی و رعایت حقوق دانشگاه، دانش‌آموختگان این دانشگاه نسبت به رعایت موارد ذیل متعهد می‌شوند:

مادّهٔ ۱ در صورت اقدام به چاپ پایان‌نامه (رساله)ی خود، مراتب را قبلاً به طور کتبی به «دفتر نشر آثار علمی» دانشگاه اطلاع دهد.

مادّهٔ ۲ در صفحهٔ سوم کتاب (پس از برگ شناسنامه)، عبارت ذیل را چاپ کند:

«کتاب حاضر، حاصل پایان‌نامهٔ کارشناسی ارشد/رسالهٔ دکتری نگارنده در رشتهٔ برق - الکترونیک است که در سال ۱۳۷۹ در دانشکدهٔ فنی و مهندسی دانشگاه تربیت مدرّس به راهنمایی سرکار خانم/جناب آقای دکتر احسان کلبیر، مشاورهٔ سرکار خانم/جناب آقای دکتر رضا عزمی و مشاورهٔ سرکار خانم/جناب آقای دکتر ——— از آن دفاع شده است.»

مادّهٔ ۳ به منظور جبران بخشی از هزینه‌های انتشارات دانشگاه، تعداد یک درصد شمارگان کتاب (در هر نوبت چاپ) را به «دفتر نشر آثار علمی» دانشگاه اهداکند. دانشگاه می‌تواند مازاد نیاز خود را به نفع مرکز نشر در معرض فروش قرار دهد.

مادّهٔ ۴ در صورت عدم رعایت مادّهٔ ۳، ۵۰٪ بهای شمارگان چاپ‌شده را به عنوان خسارت به دانشگاه تربیت مدرّس، تأدیه کند.

مادّهٔ ۵ دانشجو تعهد و قبول می‌کند در صورت خودداری از پرداخت بهای خسارت، دانشگاه می‌تواند خسارت مذکور را از طریق مراجع قضایی مطالبه و وصول کند؛ به علاوه به دانشگاه حق می‌دهد به منظور استیفای حقوق خود، از طریق دادگاه، معادل وجه مذکور در مادّهٔ ۴ را از محلّ توقیف کتابهای عرضه‌شدهٔ نگارنده برای فروش، تأمین نماید.

مادّهٔ ۶ اینجانب حسین نطاق آبادی دانشجوی رشته برق - الکترونیک مقطع کارشناسی ارشد تعهد فوق و ضمانت اجرایی آن را قبول کرده، به آن ملتزم می‌شوم.

نام و نام خانوادگی: حسین نطاق آبادی

تاریخ و امضا:

تقدیم به:

پدر زحمتکش و مادر دلسوزم که موفقیت خود در تمامی مراحل زندگی به خصوص در امرتحصیل را مدیون آن ها میدانم؛ همچنین به اساتید دوران تحصیل به ویژه جناب آقای سید حسن جهان بین که رنج فراوانی را درراه تعلیم و تربیت اینجانب متحمل شده اند.

تشکر و قدردانی

لازم است ازجناب آقای دکتر احسان اله کبیر که در طول انجـام ایـن تحقیـق از راهنماییـهای مفید و ارزنده ایشان استفاده فراوان برده ام،صمیمانه تشکر وقدردانی کنم.همچنیـن از جنـاب آقای دکتر رضا عزمی به خاطررهنمودهای خردمندانـه شـان در طـول انجـام تحقیـق تشـکر وسپاسگزاری می کنم.

چکیده

پردازش مستندات یکی از جذاب ترین زمینه های بازشناسی الگو است و بازشناسی متون، محوری ترین بخش در پردازش مستندات است. یکی از مهمترین مراحل بازشناسی متون چاپی فارسی جداسازی حروف است.

ما در این پایان نامه با اصلاح الگوریتم عزمی که مبتنی بر کانتور بالایی است، الگوریتم جداسازی مناسبی برای متون چاپی قدیمی ارائه کرده ایم. برای حل مشکل نایکنواختی کرسی خط،روش مناسبی برای تعیین نوار زمینه پیشنهاد کرده ایم و با اصلاح روش بر چسب زنی کانتور بالایی وتکمیل قواعد جداسازی، دقت الگوریتم را برای متون قدیمی افزایش داده ایم. نرخ جداسازی درست حروف با استفاده از این الگوریتم برای یک مجموعه آزمایش که از متون قدیمی برگزیده شده است، بدون هیچ پس پردازشی ۹۷٪ است.برای جداسازی نقاط چسبیده به بدنه حروف نیز روشی ارائه شده است که نتایج رضایت بخشی داده است.

کلمات کلیدی: جداسازی حروف، متون چاپی فارسی، نوار زمینه محلی، کانتور بالایی، هیستوگرام، پروفایل، جداسازی نقاط

فهرست

فصل اول

مقدمه

اتوماتیک کردن کارها و سپردن فعالیت های انسانی به ماشین از جمله مباحثی است که در دنیای امروز توجه دانشمندان و محققان را به خود جلب کرده است . سیستم های قدیمی دیگر قادر به پاسخگویی به احتیاجات انسان ها و سرویس دهی به موقع و منظم نیستند . از دیگر عواملی که خودکار کردن سیستم ها را شتاب می دهد، مقرون به صرفه کردن آنها و جایگزینی ماشین خستگی ناپذیر به جای انسان است . در چنین شرایطی است که انسان خود را ملزم به بکار گیری ماشین در پردازش اطلاعات می داند. از جمله قسمتهای مهم و پیچیده یک سیستم پردازش اطلاعات ، بخش بازشناسی الگو است.

پردازش مستندات از جمله پرکاربرد ترین امور در زمینه بازشناسی الگو است. در این زمینه آنچه که قبل از هر چیز باید به آن پرداخته شود، درک هندسی ماشین از مستند تصویر برداری شده است . در این مرحله پس از پیش پردازش هایی مثل رفع نویز و بر طرف کردن کجی تصویر، محل گراف ها، جدول ها ، تصاویر و متون مشخص می شود. بازشناسی متون محوری ترین بخش در پردازش مستندات است. این بخش ، اطلاعات تصویری متن را به اطلاعات نمادین حروف و کلمات تبدیل می کند.

از زمینه های کاربردی بازشناسی حروف می توان به خواندن اتوماتیک چک های بانکی، خواندن اتوماتیک پرسشنامه ها، تبدیل متون معمولی به متون مخصوص نابینایان ، خواندن کدپستی و آدرس نامه ها و طبقه بندی اتوماتیک آنها اشاره کرد. شرکتها ،ادارات بیمه ، قضات ، مهندسان

گرافیک ، کارمندان دولت، روزنامه نگاران و تمامی کسانی که میل دارند اطلاعات را بدون آنکه لازم باشد آنرا تایپ کنند در کامپیوتر خود داشته باشند، از بازشناسی اتوماتیک متون کمک می گیرند. اولین نرم افزار بازشناسی حروف توسط شرکت ماشین های هوشمند[1] در سال ۱۹۵۹ ساخته شد[WWW.OCR.com].

تحقیقات درباره بازشناسی حروف با استفاده از تصویر برداری نوری را میتوان به دو دسته مهم تقسیم کرد:

بازشناسی حروف تایپ شده　　بازشناسی حروف دستنویس

در مورد بازشناسی حروف انگلیسی ، ژاپنی و چینی تحقیقات قابل ملاحظه ای صورت گرفته است [Bok,92][Mor,92] . الگوریتم های بکار رفته در بازشناسی متون تایپ شده و دستنویس برای این زبان ها دارای دقت بالایی هستند. درهر زبان روش بازشناسی متناسب با ویژگیهای خاص الفبای آن زبان است . به دلیل قواعد خاص نگارش متون فارسی، هیچیک از روش های موجود در بازشناسی متون به زبان های دیگر ، به طور مستقیم برای الفبای فارسی و عربی قابل استفاده نیست.

شروع تحقیقات درباره باز شناسی حروف فارسی و عربی به سال ۱۹۸۰ بر می گردد [par,81]. تحقیقات در این زمینه اگر چه دیر آغاز شده اما خوشبختانه در سالهای اخیر توجه شایانی به بازشناسی متون فارسی در دانشگاههای ایران شده است و نتایج الگوریتم های حاصل از این تحقیقات امیدوار کننده است.

بازشناسی متون چاپی فارسی را به سه روش کلی می توان انجام داد.

۱- شکستن کلمات به حروف و بازشناسی آنها (رویکرد بر اساس جداسازی)

۲- بازشناسی زیر کلمات به عنوان الگوهای پایه (رویکرد بدون جداسازی)

۳- رویکرد ترکیبی

---

[1] Intelligent machine

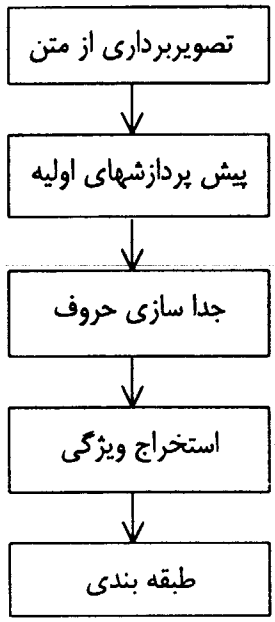
در روش اول، از حروف به عنوان الگوی بازشناسی استفاده می شود. بنابراین یک مرحله اساسی در این روش ، جداسازی حروف است. منظور از جداسازی حروف پیدا کردن چهار چوب هر حرف به صورت مجزا است. در این روش ابتدا زیر کلمات به حروف شکسته می شوند و سپس حروف جدا شده بازشناسی می شوند.

در روش دوم سعی بر این است که هر زیر کلمه بصورت یک واحد در نظر گرفته شود و بدون نیاز به جداسازی حروف آن شناسایی شود. به علت تعداد بیش از حد زیر کلمات فارسی، این روش فقط در بازشناسی حجم محدود کلمات کاربرد دارد[رفیعی ، ٧٤].

در روش سوم اطلاعات مربوط به جداسازی حروف کلمه و اطلاعات مربوط به شکل کلمه درقالب یک سیستم ترکیبی جدا سازی _ بازشناسی بکار گرفته می شود.

روش دوم که مبتنی بر بازشناسی بدون جداسازی است به علت حجم زیاد زیر کلمات فارسی از نظر حافظه و سرعت مقرون به صرفه نیست. روش اول و سوم در بازشناسی متون فارسی بیشترین کاربرد و استفاده را دارند . در این روش ها مرحله جداسازی حروف یکی از اساسی ترین مراحل است. به گونه ا ی که نتایج نهایی به شدت وابسته به نتایج این مرحله است.

مراحل مختلف یک سیستم بازشناسی حروف در بلوک دیاگرام شکل ١-١ آمده است.



شکل١-١ بلوک دیاگرام یک سیستم بازشناسی حروف

ابتدا از متن ورودی ، تصویری با درجه تفکیک مناسب(حدود dpi ۳۰۰) با سطوح خاکستری گرفته می شود . درمرحله بعد، عمل دو سطحی کردن تصویر با استفاده از یک آستانه کلی یا آستانه های محلی انجام می شود . این مرحله از اهمیت ویژه ای برخوردار است زیرا خطاهای این مرحله موجب از بین رفتن اطلاعاتی می شود که در مراحل بعد، سیستم های جداساز حروف و بازشناسی را به اشتباه می اندازد. درمرحله جداسازی ، ابتدا باید خطوط متن از یکدیگر جدا و نوار زمینه تعیین شود. بعد از آن در هر خط متن تصاویر زیر کلمات از یکدیگر جدا می شوند و الگوریتم جدا سازی هر زیر کلمه را به حروف آن می شکند. سپس ویژگیهای مناسب استخراج می شوند و در مرحله طبقه بندی نوع حروف مشخص می شود. آنچه گفته شد، شرح کوتاهی بر یک سیستم بازشناسی حروف چاپی بود. اما در عمل سیستم های بازشناسی متون فارسی با مشکلات فراوانی روبرو می شوند.

از جمله مشکلات مشترک میان بازشناسی حروف فارسی و سایر زبان ها میتوان وجود نویز و بریدگی حروف را نام برد. چسبیدن نقاط به بدنه و یا چسبیدن حروف در غیر نقاط متعارف به یکدیگر از دیگر عواملی است که بازشناسی و جداسازی را با مشکل روبرو می کند. جد اسازی حروف در این متون بدون پشتیبانی الگوریتم های بازشناسی مشکل می شود. باید در این موارد بازشناسی و جداسازی با یکدیگر توام باشند[Lu,95,96].

حال مشکلاتی را مطرح می کنیم که به قواعد و ساختار زبان و نگارش فارسی در ارتباط است.

الف) اندازه هندسی حروف فارسی مختلف است.

ب) اندازه و شکل هر حرف در قلم های مختلف، متفاوت است.

ج) حروف با توجه به موقعیت آن ها در کلمه شکل های مختلفی به خود می گیرند. برخی حروف مثل "ع" چهار شکل مختلف به خود می گیرند.

د) در بعضی از قلم های حروف ممکن است با یکدیگرهمپوشانی عمودی داشته باشند که جداسازی آنها مشکل می شود. مثل کلمات "محمد، اسلام"

ذ) بعضی از حروف در دو محل به یکدیگر می چسبند، مثل "کا".

ر) حروف مجزای فارسی دارای ارتفاع یکسانی نیستند ، بنابراین تشخیص نوع حروف مشکل می شود.

و) بعضی از حروف بدنه مشابه دارند و تنها در تعداد نقاط متفاوت هستند. این موضوع شناسایی را مشکل می کند.

همانگونه که قبل از این نیز اشاره کردیم، جداسازی حروف در میزان بازشناسی نقش تعیین کننده ای دارد. اگر چه الگوریتم هایی که تا کنون ارائه شده اند از توانایی بالایی برخوردارند، اما هر کدام به نحوی خاص دارای نقاط ضعف هستند. یکی از معایب الگوریتم های جداسازی که قبل از این ارائه شده است عدم توجه به جابجایی کرسی خط می باشد. این موضوع باعث شده است تا در متون قدیمی تر که به دلیل ابتدایی بودن ماشین های تایپ کلمات بروی یک خط زمینه ثابت نوشته نمی شدند، عمل جداسازی و به تبع آن بازشناسی با نرخ مناسبی انجام نشود.

هدف پایان نامه

مهمترین موضوعی که ما در این پایان نامه دنبال می کنیم ، مسئله جابجایی کرسی نوار زمینه است. سعی کرده ایم تا این موضوع را به نحو شایسته و ممکن قبل از مرحله بازشناسی یعنی در مراحل پیش پردازش و جداسازی حل کنیم . توجه زیادی به نوار زمینه محلی کرده ایم و چگونگی تعیین این نوار به نحو شایسته ای بررسی شده است. در ضمن تا حدی نیز به تصحیح و بهبود کیفیت الگوریتم جداسازی خاصی که مبتنی به نوار زمینه بوده است پرداخته ایم [عزمی، ۷۸]. خوشبختانه نتایج الگوریتم نهایی امیدوار کننده است و نشان ازپیشرفت الگوریتم جداساز دارد. این موضوع امکان استفاده از روش اول در بازشناسی متون چاپی فارسی که مبتنی به جداسازی است را افزایش می دهد.