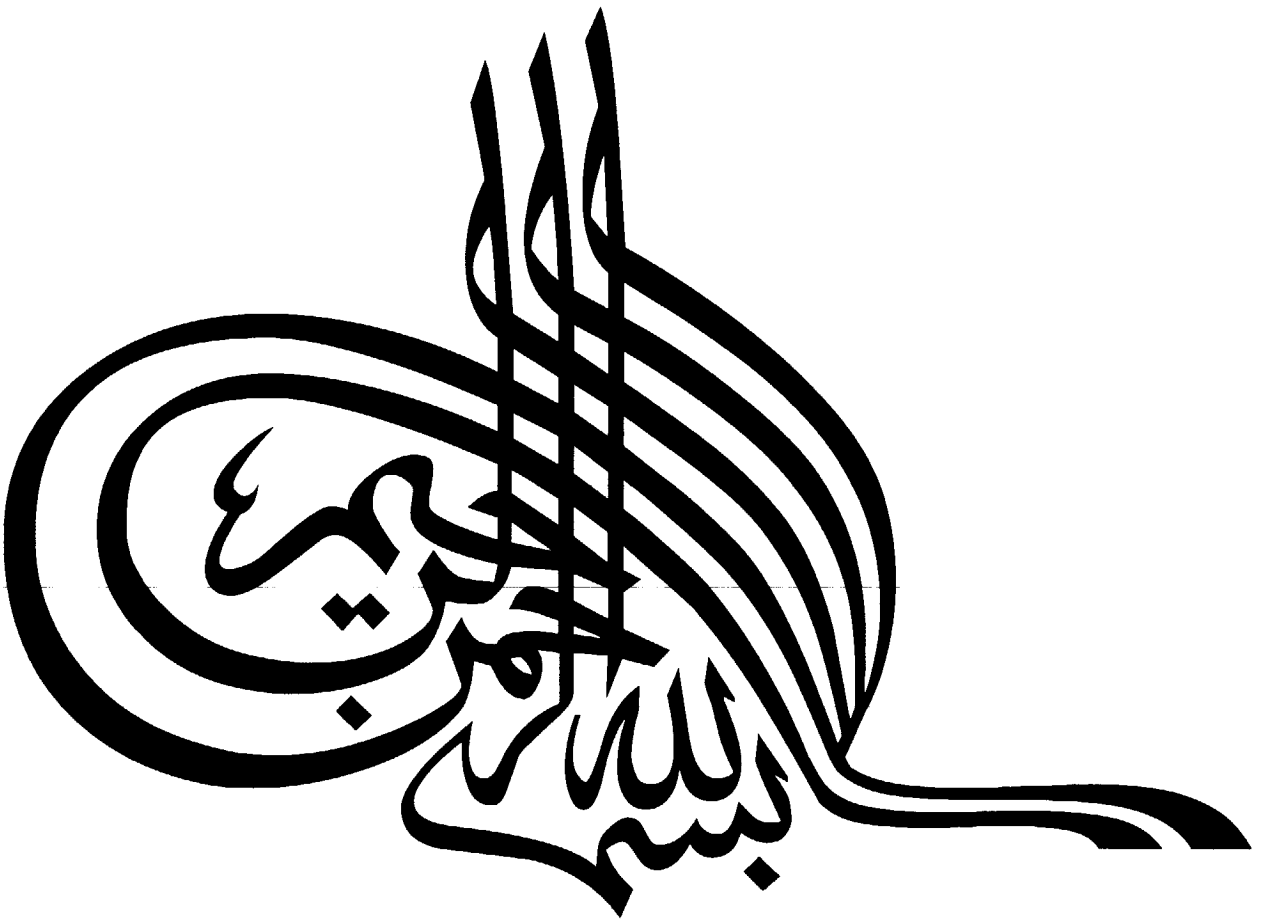


۱۲ /۷/ ۱۳۷۸

دانشگاه علم و صنعت ایران

دانشکده مهندسی کامپیوتر

بازشناسی حروف فارسی با استفاده از شبکه عصبی چند جمله ایهای جداکننده

سعید اسدی

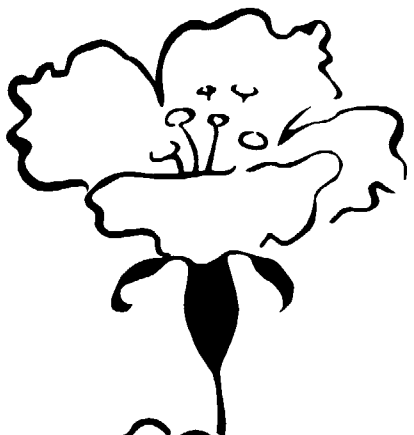
پایان نامه برای دریافت درجه کارشناسی ارشد

در رشته

مهندسی کامپیوتر

استاد راهنما: دکتر محمود فتحی

اردیبهشت ۱۳۷۷



تقدیم به مادرم

عزیزی که در نبود پدر وظیفه پدری و مادری را بنحو احسن جامه عمل پوشانید و با کوله بار سنگین مرارت و زحمت رنج به ثمر رسانیدن مرا به تنهایی بدوش کشید.

وتقدیم به همسر عزیزم

چکیده

در این پایان نامه یک روش بازشناسی حروف تایپی فارسی بصورت ساختاری و باپارامتری فرض کردن حروف ارائه شده است.در این روش ویژگیهای مناسبی که بتواند شکل و حالات هریک از حروف را نشان دهد ، استخراج میگردد. این ویژگیها بگونه ای مورد استفاده قرار میگیرند که بتوانند نمایانگر حروف مختلف بصورت یکتا باشند. با استفاده از جداول و درخت های تصمیم گیری ، روشی جهت بازشناسی حروف تایپی فارسی ارائه میگردد. مسائل مهم در این پایان نامه عبارتند از: جداسازی حروف که براساس موقعیت خط زمینه صورت می پذیرد، استخراج ویژگیهای مناسب وبازشناسی حروف فارسی که با استفاده از بانک های اطلاعاتی ودرخت های تصمیم گیری انجام میگیرد. روش دیگربازشناسی که در این پایان نامه مورد استفاده قرارگرفته است، استفاده از (*Padaline's*) شبکه عصبی چند جمله ایهای جداکننده است. در این روش تعدادی از ویژگیهای استخراج شده تبدیل به تابت های برای چند جمله ایهای مختلف میگرددو برای حالات مختلف حروف یک چند جمله ای ارائه میگردد. الگوریتم ها مستقل از اندازه حروف میباشند. روشهای اراته شده در این پایان نامه بدون هیچگونه تغییری میتواند برای بازشناسی حروف تایپی لاتین نیز بکار رود. برای طبقه بندی نهایی در هر دو روش بازشناسی نیز از بانک های اطلاعاتی استفاده شده است.

تقدیر و تشکر

از زحمات استاد ارجمند جناب آقای دکتر محمود فتحی که راهنمای اینجانب در انجام پروژه و همچنین تهیه و ارائه این پایان نامه بوده اند.

و همچنین از هیئت محترم ممتحن آقایان دکتر کبیر و دکتر رحمانی وکلیه اساتید محترم دانشکده کامپیوتر.

تشکر و امتنان خود را ابراز می نمایم

فهرست مطالب

فهرست شکلها

فهرست جداول

فهرست فرمولها

مقدمه

از زمانیکه کامپیوتر توسط بشر ساخته شده ، بشر همواره درآرزوی ایـن بـوده اسـت کـه بگونه ای خواندن ونوشتن را با کمک کامپیوتر سهل تر بنماید، در زمینه نوشـتن موفقیـت هـای چشمگیری برای بشر حاصل شده است، ولی برای خواندن هنوز راه طولانی در پیش است.

هـدف کلـی بازشناسـی حـروف(*optical character recognition*) تشـخیص حروف مختلف توسط کامپیوتر میباشد. نمونه های از کاربردهای بازشناسی، تشـخیص حـروف زبانهای مختلف بصورت تایپی و دستنویس ، تشخیص پلاک خودروهـا و غـیره میباشـد. بدیهـی است جواب مناسب در زمینه هرکدام از مواردشاره شده کاربردهای متعددوموثـری در زندگـی روزمره خواهد داشت. بعنوان مثال تشخیص پلاک خودروهـا، باتوجـه بـه تصـاویر ارائه شـده از دوربین های مختلف که در محل هـای مشـخص نصـب شـده اسـت،کمـک موثـری در ردیـابی رانندگان متخلف میباشد.هم چنین بازشناسی حروف تـایپی وچـاپی بـرای ورود اطلاعـات چـاپ شده به کامپیوتربا سرعت و دقت و درعین حال بـا هزینـه کمـتر کـاربرد دارد . بطـور متوسـط میتوان گفت که دقت یک تایپیست حرفه ای که حروف را تشخیص داده و سپس تـایپ میکنـد در حدود ۹۵٪ میتواند باشد، در حالیکه دقت نرم افزارهای حرفـه ای بازشناسـی حـروف تـایپی که اخیرا به بازار معرفی شده اند در حدود ۹۹٪ است[1]. سرعت نرم افزارهای مذکور برروی یک کامپیوتر ٤٨٦ بیشتر از ده برابر یک تایپیست حرفـه ای اسـت[2]. بدیهـی اسـت بـا بهینـه شدن روشهای بازشناسی و سرعت کامپیوترها، نسبت فوق نیز افزایش خواهد یافت.

بازشناسی حروف توسط کامپیوتر را بطور کلی به دودسته *on line* و *off line* میتوان تقسیم بندی کرد. در مدل *on line* کاربر بااستفاده از وسایلی نظیر موشواره و قلم نوری حروف و یا کلمه مورد نظر خود را مینویسد ودر لحظه نوشتن کامپیوتر حروف نوشته شده را تشخیص خواهد داد.در مدل *off line* تصاویری که شامل حروف مورد نظر جهت بازشناسی میباشد در حافظه و یا دیسک ذخیره میشوند وسپس پردازش بازشناسی حروف بر روی این تصاویر صورت خواهد پذیرفت. در این پایان نامه مدل *off line* جهت تشخیص حروف تایپی مورد نظر است.

از سایر جنبه ها ، بازشناسی حروف تایپی را به دسته بندیهای مختلفی می توان تقسیم کرد. بطور مثال تقسیم بندی براساس اتصال یاانفصال حروف در کلمه هارا میتوان ذکر کرد. در زبانهای نظیر فارسی و عربی حروف با اتصال بهم تشکیل کلمه ها را میدهند ،در عین حال شکل بعضی از حروف درزمان اتصال تغییر خواهد کرد ودرحالیکه در زبان های نظیر لاتین و چینی حروف تشکیل دهنده کلمه ها جدا از هم میباشند. میتوان ساختار حروف را نیز از نظر ثابت یا متغییر بودن مدنظر قرار داد . ساختار ثابت مانند یک متن تایپی با اندازه ثابت حروف می باشدو ساختار متغییر مانند متون دستنویس و یا متون تایپی با اندازه و قلم های متفاوت میباشد.

در مقایسه با تشخیص حروف لاتین ، در زبان فارسی برای بازشناسی حروف در کلمه ها باید ابتدا عملیات جداسازی صورت پذیرد ، که این امر عملیات بازشناسی را پیچیده تر و طولانی تر خواهد کرد. بدیهی است صحت عملیات بازشناسی ارتباط مستقیم با جداسازی

جداسازی دارد.بطور کلی بازشناسی حروف فارسی و زبانهای مشابه از قسمت های زیر تشکیل شده است: ۱-ورود تصویر به کامپیوتر (*Digitizing*)

۲.پردازش تصویر(*Image Processing*)

۳.جداسازی حروف(*Segmentation*)

٤.استخراج ویژگیها(*Feature extraction*)

٥.طبقه بندی و بازشناسی حروف (*Classification & Recognition*)

در زبانهایی که حروف تشکیل دهنده کلمه ها جداز هم میباشند ، عملیات جداسازی صورت نخواهد پذیرفت و مراحل کلی بازشناسی به چهار مرحله خلاصه میگردد. در عین حال هرکدام از مراحل اشاره شده ممکن است به چندین قسمت دیگر تقسیم شوند.

۱- در مرحله اول ابتدا باید تصویر با ابزاری نظیر اسکنر، برنامه های نقاشی و غیره بصورت دیجیتالی وارد کامپیوتر گردد. البته در این پروژه تصاویر حاوی اطلاعات از برنامه گرافیکی (نقاشی) تحت ویندوز دریافت میشود بنابراین حداقل کیفیت مورد آزمایش در این پروژه *75dpi* خواهد بود.این تصاویر به صورت سیاه و سفید ذخیره میشوند . بنابراین در این پروژه مرحله پردازش تصویر مورد نظر نمی باشد.

۲- یکی از مطالب مهم در بازشناسی حروف فارسی عملیات جداسازی حروف است. اولین قدم در جداسازی حروف تایپی فارسی در این پروژه پیداکردن خط زمینه است .بطوریکه مشخص است بیشترین تراکم نقاط سیاه در یک جمله در خط زمینه میباشد. یکی از روشهای مناسب در بدست آوردن خط زمینه ، تراکم نقاط سیاه در یک جمله مورد پردازش میباشد، این

تراکم با شمردن تعداد نقاط سیاه در هر سطر (در صورتیکه قرارگیری نقاط تشکیل دهنده یک جمله را در یک ماتریس فرض کنیم) بدست می آید. بیشترین تراکم نقاط سیاه در واقع در خط زمینه یا در نزدیکی خط زمینه است. با استفاده از موقعیت و قطر خط زمینه روشی جهت جداسازی ارائه میگردد. این بخش از پروژه در فصل "جداسازی حروف " شرح داده شده است.

۳- مهمترین قسمت بازشناسی ، شناسایی و استخراج ویژگیهای مناسب است ودر بخش "استخراج ویژگیها " درباره آن بحث خواهیم کرد.این ویژگیها میتوانند بصورت کمیت های عددی و یا توصیفی بیان شوند و بصورت کلی و عمومی روش ثابت و تعریف شده ای برای شناسایی و استخراج ویژگیهای مناسب وجود ندارد. نوع ویژگی های که انتخاب میشوند تا حد زیادی دقت و سرعت بازشناسی را نشان میدهند. در بخش "استخراج ویژگیها " ، خصوصیات مختلفی که در مقالات داخلی و خارجی برای شناسایی حروف تایپی فارسی معرفی شده ، ارائه میگردد. در این پروژه ویژگی های جدیدی انتخاب شده است که برخی از آنها شامل : درصد نقاط سیاه در بالای خط زمینه، تعداد حفره ها در هر حرف، تعداد قوسها یا دهانه ها در جهت های اصلی و در جهت های اصلی با ٤٥ در جه چرخش حرف ، تعداد قطعه ها در هر حرف ، میباشند . ویژگی های مذکور پس از پردازش اولیه ، در مراحل آموزش و بازشناسی مورد استفاده قرار میگیرند.

٤- شیوه های مختلفی نظیر شبکه های عصبی ، الگوریتم های آماری ، سیستم های خبره ، بانک های اطلاعاتی و غیره برای کلاسه بندی وجود دارد .در این پروژه از دو روش "

شبکه های عصبی چند جمله ای های جداکننده (*Padaline's*) و بانک های اطلاعاتی(*Microsoft Access*) برای کلاسبندی و بازشناسی استفاده شده است . البته روش شبکه های عصبی به همراه ترکیبی از بانک های اطلاعاتی (*Microsoft Access*) مورد استفاده قرار گرفته اند . این بخش از پایان نامه در فصل "بازشناسی حروف" مورد بحث قرار گرفته است.

در انتهای پایان نامه الگوریتم های بکار برده شده در بخشهای جدا سازی ، استخراج ویژگی ها و بازشناسی شرح داده میشوند. در بخش ضمائم، جداول تقسیم بندی حروف و نتایج حاصل از استخراج ویژگی ها از یک نمونه آزمایشی ارائه شده است .همچنین لیست توابع مهم برنامه که به زبان *Visual Basic*و تحت ویندوز میباشد،در "ضمیمه ب" آورده شده است.

فصل دوم

جداسازی حروف

Character Segmentation

مقدمه:

یکی از مهمترین قسمت های بازشناسی حروف فارسی و یا زبان های مشابه ، جداسازی حروف در کلمه ها میباشد . صحت بازشناسی ارتباط مستقیم با عملیات جداسازی دارد و چنانچه جداسازی صحیح انجام نپذیرد ، نمی توان انتظار نتایج درست از پردازش های بازشناسی داشت . بصورت کلی روش های مختلفی برای جداسازی وجود دارد که در این فصل به شرح آنها پرداخته میشودو همچنین روش پیشنهادی در این پروژه که مورد آزمایش قرارگرفته است ، بطورکامل شرح داده میشود.چند ویژگی مهم حروف فارسی که تاثیر بسیار زیادی در عملیات جداسازی و همچنین در بازشناسی دارند بصورت مختصر در زیر بیان میشود:

- حروف فارسی دارای دو ،سه ویا چهار نماد میباشند .شکلهای مختلف یک حرف را اصطلاحاً نماد می نامند ، این نمادها بستگی به محل قرارگیری در کلمه ها مشخص میشوند.برخی از نمادهای حروف کاملا جداهستند و بعضی از دوطرف و بعضی فقط از راست به حروف دیگرمتصل میشوند و محل اتصال حروف به یکدیگر