١٧٣٩٩

بسم الله الرحمن الرحيم

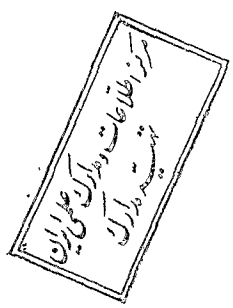
دانشگاه تربیت مدرس

دانشکده فنی و مهندسی

پایان نامه کارشناسی ارشد مهندسی الکترونیک

شناسایی ساختاری حروف دستنویس فارسی

غلامرضا اردشیر بهرستاقی



استاد راهنما:

دکتر احسان‌الله کبیر

پاییز ۱۳۷۲

دانشگاه تربیت مدرس

بسمه تعالی

چکیدهٔ پایان نامه

عنوان پایان‌نامه : شناسایی ساختاری حروف دستنویس فارسی

نام نویسنده : غلامرضا اردشیر بهرستاقی استاد راهنما : دکتر احسان الله کبیر

اساتید مشاور : .....

دانشکده : فنی و مهندسی رشتهٔ : برق - الکترونیک

تاریخ دفاع: ۷۲/۷/۲۱

تحقیق انجام شده در زمینه بازشناسی متون دستنویس است که یکی از شاخه‌های بازشناسی الگو می‌باشد. هدف بازشناسی حروف دستنویس فارسی است که بطور مجزا توسط افراد مختلف نوشته شده باشند. فرض بر این است که حروف با دقت ۳۰۰ نقطه در اینچ تصویربرداری و با کد MSP ذخیره شده‌اند.

الگوریتم طرح شده را می‌توان به دو بخش اساسی تقسیم نمود: در بخش اول، اقدام به استخراج ویژگیها می‌شود و در بخش دوم، حروف بر اساس یک درخت تصمیم به عنوان طبقه‌بندی‌کننده مورد بازشناسی قرار می‌گیرند. در مرحله استخراج ویژگیها، در گام اول یک مرحله پیش‌پردازش برای رفع پریدگی و جزء و نویز صورت می‌گیرد. سپس با اجرای الگوریتم برچسب‌زنی مولفه‌ها، نقطه و سرکش موجود از بدنه اصلی حرف جدا و نوع آن تشخیص داده می‌شود. در نهایت با گروه‌بندی بخشهای سطرهای مختلف ماتریس تصویر حرف، زیرحروف استخراج و مورد شناسایی قرار می‌گیرند.

برای یک مجموعه متشکل از ۶۱ نمونه از هر حرف، میزان بازشناسی ۹۱/۵ درصد بدست آمده است و برای مجموعه‌ای دیگر متشکل از ۴۰ نمونه از هر حرف ۹۲/۱۲ درصد حروف بطور صحیح بازشناسی شدند.

تحقیق انجام‌شده می‌تواند کاربردهای زیادی داشته باشد. برای مثال الگوریتم حاصل از این تحقیق را می‌توان به منظور بازشناسی خودکار حروف مجزایی استفاده نمود که در فرمهای گوناگون جمع‌آوری و ثبت اطلاعات و کدگذاریها به کار برده می‌شوند.

امضاء استاد راهنما

احسان الله کبیر

تقدیم به:

تمامی شهیدان راه حق

تقدیم به:

پدر و مادر وهمسر مهربانم که همواره

مشوقم بوده اند.

تشکر و قدردانی:

با تشکر از استاد گرامی وارجمند آقای دکتر احسان‌الله کبیر به خاطر راهنمائیهای ارزشمند ایشان و دوست عزیزم آقای مهندس رضا عزمی و سایر کسانی که در انجام این تحقیق مرا یاری نمودند.

این تحقیق با پشتیبانی مالی و تجهیزاتی بخش برق و الکترونیک سازمان پژوهشهای علمی صنعتی ایران انجام شده است.

چکیده :

تحقیق انجام شده در زمینه بازشناسی متون دستنویس است که یکی از شاخه‌های بازشناسی الگومی‌باشد .هدف،بازشناسی حروف‌دستنویس فارسی است‌که بطور مجزا توسط افراد مختلف نوشته‌شده‌باشند. فرض براین است‌که حروف با دقت ۲۰۰ نقطه در اینچ تصویربرداری و با کد MSP ذخیره شده‌اند.

الگوریتم طرح شده را می‌توان به دو بخش اساسی تقسیم نمود: در بخش اول، اقدام به استخراج ویژگیها می‌شود و در بخش دوم، حروف بر اساس‌یک درخت تصمیم به عنوان طبقه‌بندی‌کننده مورد بازشناسی قرار می‌گیرند. در مرحله استخراج ویژگیها، در گام اول یک مرحله پیش‌پردازش برای رفع بریدگی‌و حفره و نویز صورت می‌گیرد.سپس با اجرای الگوریتم برچسب‌زنی مولفه‌ها، نقطه و سرکش موجود از بدنه اصلی حرف جدا و نوع آن تشخیص داده می‌شود. در نهایت با گروه‌بندی بخشهای سطرهای مختلف ماتریس تصویر حرف، زیرحروف استخراج ومورد شناسایی قرار می‌گیرند.

برای مجموعه تمرین متشکل از ۱۶ نمونه از هرحرف،میزان بازشناسی۹۱/۵ درصد بدست آمده استوبرای مجموعه آزمایش‌متشکل از ۴۰ نمونه از هر حرف ۹۲/۱۲ درصد حروف بطور صحیح بازشناسی شده‌اند.

تحقیق انجام‌شده می‌تواند کاربردهای زیادی داشته باشد. برای مثال الگوریتم حاصل از این تحقیق را می‌توان به منظور بازشناسی خودکار حروف مجزایی استفاده نمود که در فرمهای گوناگون جمع‌آوری وثبت‌اطلاعات‌و کدگذاریها به کار برده می‌شوند.

فهرست مطالب

عنـــــوان	صفحـــه
فصل اول : مقدمه	
۱ـ۱ـ کلیات	۱
۱ـ۲ـ عمل بازشناسی	۲
۱ـ۳ـ وضعیت بازشناسی کاراکترها درزبانهای‌مختلف	۳
۱ـ۴ـ موضوع این تحقیق	۴
۱ـ۵ـ سازمان کلی رساله	۵
فصل دوم : بازشناسی الگو	
۲ـ۱ـ الگـــو	۶
۲ـ۲ـ بازشناسی الگو	۷
۲ـ۳ـ روشهای مختلف دربازشناسی الگو	۸
۲ـ۳ـ۱ـ روشهای تحلیلی	۸
الف) روشهای آماری	۸
ب) روشهای معین	۱۰
۲ـ۳ـ۲ـ روشهای نحوی یا ساختاری	۱۰
۲ـ۳ـ۳ـ مقایسه‌روشهای تحلیلی ونحوی	۱۲
۲ـ۴ـ بازشناسی کاراکترهای دستنویس	۱۴
۲ـ۵ـ نمونه‌هایی ازکارهای انجام شده	۱۶
۲ـ۵ـ۱ـ استخراج ویژگیها	۱۶
۲ـ۵ـ۲ـ طبقه‌بندی	۱۸
۲ـ۵ـ۳ـ نتایـــج	۱۹

فهرست مطالب

عنوان	صفحه
فصل سوم : مراحل بازشناسی حروف دستنویس فارسی	
۳-۱- مقدمه	۲۱
۳-۲- تعریف بخش	۲۲
۳-۳- پیش پردازش	۲۳
۳-۳-۱- رفع بریدگی	۲۴
۳-۳-۲- رفع حفره و نویز	۲۴
۳-۴- تشخیص نقطه و سرکش	۲۷
۳-۵- تعیین نوع نقطه و سرکش	۳۲
۳-۶- پیدا کردن ضخامت قلم	۳۵
۳-۷- زیرحرف	۳۵
۳-۸- تعریف گروه و گروه‌بندی	۳۶
۳-۹- اصلاح گروه‌بندی	۳۹
۳-۱۰- شناسایی نوع گروه یا زیرحرف	۴۰
۳-۱۰-۱- شناسایی کمان	۴۰
۳-۱۰-۲- شناسایی گروه " S "	۴۱
۳-۱۰-۳- شناسایی خط مورب	۴۳
۳-۱۰-۴- شناسایی گروه " T "	۴۳
۳-۱۰-۵- شناسایی خطوط افقی و عمودی	۴۴
۳-۱۰-۶- الگوریتم شناسایی نوع گروه	۴۵
۳-۱۱- شناسایی منحنی بسته	۴۵
۳-۱۲- شناسایی دندانه	۴۷

فهرست مطالب

عنـــــوان	صفحـــه
۳ـ۱۳ـ شناسایی مجموعه‌گروههایی که‌به‌صورت کمان هستند	۵۱
۳ـ۱۴ـ شناسایی مجموعه‌گروههایی که‌به‌صورتخط مورب هستند	۵۲
۳ـ۱۵ـ شناسایی فرورفتگی در "ی" وحروف "ک ، گ" با سرکش متصل به‌بدنه	۵۳

فصل چهارم : الگوریتم بازشناسی حروف

۴ـ۱ـ مقدمه	۵۵
۴ـ۲ـ دسته‌بندی حروف	۵۵
۴ـ۲ـ۱ـ دسته‌بندی حروف براساس پایین ترین گروه	۵۵
۴ـ۲ـ۲ـ دسته‌بندی حروف براساس حرکت دردوطرف گروه g_0	۵۸
۴ـ۲ـ۳ـ دسته‌بندی حروف براساس تعداد گروهها دردوطرف گروه g_0	۶۱
۴ـ۲ـ۴ـ دسته‌بندی حروف براساس مرکزثقل	۶۳
۴ـ۲ـ۵ـ دسته‌بندی حروف براساس تعداد گروههای موجوددربدنه‌آنها	۶۴
۴ـ۳ـ الگوریتم بازشناسی حروف	۶۶
۴ـ۴ـ نتایـــــج	۸۹

فصل پنجم : خلاصه ،نتیجه‌گیری ، پیشنهادها

۵ـ۱ـ مقدمه	۹۴
۵ـ۲ـ شرایط تحقیق انجام شده‌ومشخصات مجموعه‌داده	۹۴
۵ـ۳ـ خلاصه‌ای ازمراحل اساسی الگوریتم ونتایج بدست‌آمده	۹۵
۵ـ۴ـ پیشنهادهـــــا	۹۶
منابـــــع	۹۸

فصل اول

۱-۱ کلیات

تحقیقات درزمینه بازشناسی الگو[1] تقریبا "ازسال ۱۹۵۰ شروع شده است. اولین تلاشها درزمینه بازشناسی الگو، برای تهیه برنامه های کامپیوتری جهت تصمیم گیری خودکار وتوسعهٔ سخت افزارهای خاص برای خواندن الگوهائی مانندکاراکترهای ارقام وحروف تایپ شده بوده است. دراواخردههٔ ۱۹۵۰، Resenblatt یک مدل اولیه درالگوریتم پرسپترون[2] برای آرایش وذخیره اطلاعات درمغز عرضه کرد. اکثرروشهای عرضه شده برای مسئله بازشناسی الگودرآن زمان براساس تئوری تصمیم آماری[3] واصول منطق آستانه ای[4] بنا شده بود. درطی دههٔ ۱۹۶۰ تحقیقات درزمینهٔ طراحی سیستمهای بازشناسی الگومقداری پیشرفت داشت. با استفاده ازتئوری زبان کامپیوتر[5] واستفاده ازتوانائیهای پردازش کامپیوتر، روش نحوی[6] به عنوان یک مکمل برای تکنیکهای تحلیلی[7] درحل مسائل بازشناسی الگوهای تصویری معرفی شد (Toe ۸۱). مفاهیم بازشناسی الگوبطورفزاینده ای درطراحی سیستمهای مدرن اطلاعاتی کامپیوتری به عنوان یک عامل مهم درنظرگرفته شده اند. جاذبیت دراین زمینه هنوزهم با سرعت زیاد درشدمی کند. درشاخه های مختلف مانندمهندسی، علوم کامپیوتر، علوم اطلاعات، آمار، فیزیک، شیمی، زبان، روانشناسی، زیست شناسی، فیزیولوژی، پزشکی وزبانشناسی موضوع بازشناسی الگومطرح است. هرکدام ازاین شاخه ها به جنبهٔ خاصی ازمسئله بازشناسی الگواهمیت می دهند.

---

1. Pattern Recognition
2. Perceptron
3. Statistical
4. Threshold Logic Principles
5. Computer Language Theory
6. Syntactic
7. Analytical

۱-۲- عمل بازشناسی

بازشناسی الگو از ویژگیهای انسان و دیگر موجودات زنده است که در تمام لحظاتی که در بیداری هستند، بطور خودکار آن را انجام می دهند. انسان از این نظر که از توانائی بالائی در بازشناسی برخوردار است یک سیستم کاملا "خبره" و ماهر است.

الگو توصیفی از یک موضوع است. از دیدگاههای مختلف، تقسیمات مختلفی می توان برای بازشناسی الگو انجام داد. بسته به طبیعت الگوهائی که باید بازشناسی شوند، ممکن است عمل بازشناسی را به دو نوع عمده بازشناسی نمونه های مجرد و انتزاعی[1] و نمونه های ملموس[2] تقسیم نمود. هنگامی که یک بحث قدیمی را به خاطر می آوریم، یا راه حل مسئله ای را پیدا می کنیم یک بازشناسی الگوی مجرد و انتزاعی انجام داده ایم. در بازشناسی نمونه های ملموس، الگوهای زمانی[3] و مکانی[4] بازشناسی می شوند. بازشناسی کاراکترها، بازشناسی آثار انگشت، خواندن نقشه های هواشناسی، موضوعات فیزیکی و تصاویر، نمونه هائی از بازشناسی الگوهای مکانی هستند. از نمونه های الگوهای زمانی میتوان به شکل موجهای صحبت، الکتروکاردیوگرامها و الکتروانسفالوگرامها اشاره کرد.

مطالعه در مورد بازشناسی الگو ممکن است به دو زمینه عمده تقسیم شود:

۱- مطالعه توانائی بازشناسی الگو توسط انسان و دیگر موجودات زنده.

۲- مطالعه تئوریها و تکنیکهای مناسب برای طراحی دستگاههائی که قادر به بازشناسی الگو باشند.

زمینه اول به روانشناسی، فیزیولوژی و زیست شناسی مربوط می شود

---

1. Abstract
2. Concrete
3. Temporal
4. Spatial

وزمینه دوم اصولا"با مهندسی ،کامپیوتروعلوم اطلاعات سروکار دارد .

موضوع این تحقیق درگروه دوم قرار دارد .به زبان ساده بازشناسی الگو می تواندبه عنوان دسته بندی داده های ورودی درکلاسهای قابل بازشناسی ازطریق استخراج ویژگیها وخواص با معنی آنها تعریف شود .سیستم بازشناسی کاراکتریک سیستم بازشناسی الگومی باشدکه تصویرکاراکترها ،به عنوان داده های ورودی آن ونام کاراکترها ،خروجی های آن محسوب می شود .

۱-۳- وضعیت بازشناسی کاراکترها درزبانهای مختلف

درزمینهٔ بازشناسی کاراکترهای انگلیسی تحقیقات قابل ملاحظه ای صورت گرفته ومقالات زیادی چاپ شده است .الگوریتمهائی که برای کاراکترهای تایپ شده یا دستنویس طراحی شده ،قادربه بازشناسی کاراکترها با دقت خوبی می باشد .درموردحروف چینی وژاپنی نیزتحقیقاتی انجام گرفته است .درزمینه بازشناسی کاراکترهای عربی ،فارسی وهندی تحقیقات کمی صورت گرفته است. تفاوت بین کاراکترهای لاتین وکاراکترهای عربی وفارسی اجازه استفاده مستقیم ازالگوریتمهای مربوط به بازشناسی کاراکترهای لاتین را برای کاراکترهای عربی وفارسی نمی دهد .

ازتحقیقات انجام شده برای شناسایی کاراکترهای عربی می توان موارد زیررا نام برد .کامبیزبدیع درسال ۱۹۸۰ حدود ۹۰ درصدازه ۱۲ حرف بزرگ (۱۰ تا ۱۲ نمونه ازهرحرف)دستنویس عربی را بازشناسی نمود (۸۰ Bad و ۸۲ Bad) .ادنان امین درسال ۱۹۸۰ تعداد ۷۳ کاراکتررا که توسط ۳ نویسنده نوشته شده بودند با دقت ۹۲ تا ۹۵ درصدبازشناسی کرد (۸۰ Ami و ۸۲ Ami) .یوسف درسال ۱۹۸۸ حدود ۹۹ درصدازکاراکترهای دستنویس عربی را بازشناسی نمودکه شامل ۵۰ نمونه درهرکلاس نوشته شده توسط ۲۵ نفربوده است (۸۸ You).

از تحقیقات انجام شده در زمینه حروف فارسی می توان به کار عزمی اشاره نمود که حدود ۲۰۰ نمونه از ۳۳ کلاس در نظر گرفته شده برای حروف دستنویس فارسی را که توسط افراد مختلف نوشته شده بودند با دقت ۹۰% بازشناسی کرده است (عزمی ۷۲).

۱-۴- موضوع این تحقیق

موضوع این تحقیق ،بازشناسی حروف دستنویس فارسی است .حروف به صورت مجزا فرض شده و توسط افراد مختلف نوشته شده اند .به منظور تصویربرداری از یک روبشگر[1] نوری استفاده شده است که از حروف با دقت ۲۰۰ dpi تصویربرداری نموده و با کد MSP در حافظه ذخیره می نماید .

برای بازشناسی حروف از ویژگیهای ساختاری آنها استفاده شده است . ویژگیها یی اصلی عبارتند از :کمان،خطوط مورب ، خطوط عمودی و افقی و شکلهای "S" و "T".بعضی از ویژگیهای ساختاری حروف فارسی نیز مورد توجه قرار گرفته اند . دندانه در حروف "س" و "ش" ،منحنی بسته در حرف "ص" وغیره و فرورفتگی در حروف "ی" ،"ک" و "گ" از این جمله اند .

برای بازشناسی از یک درخت تصمیم به عنوان طبقه بندی کننده استفاده شده است .این درخت تصمیم به صورت تجربی طرح شده است .

در این تحقیق ،از ۶۱ نمونه برای هرکلاس به عنوان مجموعه تمرین استفاده شده که میزان بازشناسی ۹۱/۵ درصد حاصل شده است .برای یک مجموعه دیگر از حروف (مجموعه آزمایش) شامل ۴۰ نمونه از هرکلاس میزان بازشناسی ۹۲/۱۲ بدست آمده است .

1.Scanner

۱-۵- سازمان کلی رساله

درفصل دوم این رساله ،ابتددرموردالگووبازشناسی الگوتوضیح داده میشود. سپس درموردروشهای مختلف بازشناسی الگوتقسیم‌بندی آنها مطالبی بیان میشود. درآخرفصل دوم‌دربارهٔ‌بازشناسی‌کاراکترهای دستنویس ونمونه‌کارهای انجام‌شده توضیحات کوتاهی آورده میشود.

درفصل سوم ،مراحل مختلف استخراج ویژگیهای ساختاری بیان می شـــود. درفصل چهارم الگوریتم‌بازشناسی حروف راشرح خواهیم‌داد.

فصل پنجم‌به‌بررسی نتایج و پیشنهادها اختصاص دارد.

فصل دوم : بازشناسی الگو

در این فصل ابتدا مفهوم الگو و بازشناسی الگو مطرح می شود . سپس درمورد روشهای مختلف بازشناسی الگو و تقسیم بندی آنها به روشهای آماری و نحوی (ساختاری) مطالبی بیان می شود،در آخر درباره بازشناسی کاراکترهای دستنویس و نمونه کارهای انجام شده توضیحات کوتاهی آورده می شود .

۲-۱- الگو[1] :

در متون مهندسی بازشناسی الگو[2] ، توصیف های مختلفی از الگو ارائه شده است (Kab ۹۰) :"یک الگو توصیف یک شیا ست " ، " درکلی‌ترین عبارت الگوها وسائلی هستند که ماجهان را بوسیله آنها تفسیر می‌کنیم " ، " بسیاری از اطلاعاتی که در زندگی با آنها سروکار داریم به شکل الگوهای پیچیده هستند " ، " یک الگو یک توصیف کمی یا ساختاری از یک شئی است " ، " یک الگو ترکیبی از n مورد است که n می تواند برابر با ۲ یا بزرگتر از آن باشد ".

در یک بررسی دقیق ، کلمه الگو به دوصورت تعریف می شود : در تعریف اول الگو به‌عنوان یک پدیده یا موضوع الگو شده تعریف می شود . به عبارت دیگر پدیده یا موضوع خودش به عنوان یک الگو درنظر گرفته می شود . بنابراین بازشناسی الگو همان طبقه بندی است . درتعریف

1.Pattern

2.Pattern Recognition