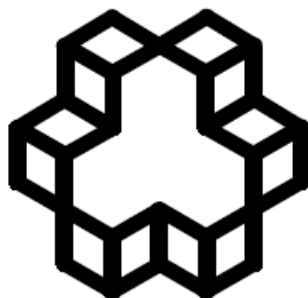


بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ



دانشگاه صنعتی خواجه نصیرالدین طوسی

تاسیس ۱۳۰۷

رساله دکترای الکترونیک

عنوان :

بازیابی متون چاپی فارسی بر اساس پرس و جوی کلمات

استاد راهنما : دکتر هوشنگ حسینی

نگارنده : یعقوب پوراسد

زمستان ۱۳۹۰

چکیده

در این پایان نامه روشی جدید برای بازیابی و جستجوی کلمات فارسی از میان متون تصویری فارسی ارائه شده است. همچنین روشی مبتنی بر اندازه و شکل نقاط موجود در مستند پس از آستانه گذاری، جهت تشخیص قلم و اندازه قلم مستندات تصویری فارسی ارائه شده است. روش ارائه شده برای تشخیص قلم از اولین روش هایی می باشد که می تواند با دقت بسیار بالایی اندازه قلم مستند فارسی را هم تشخیص دهد. برای ارزیابی روشهای ارائه شده برای تشخیص قلم و نیز سیستم بازیابی کلمات، چندین پایگاه تصویری با استفاده از کامپیوتر ایجاد شدند. پایگاه تصویری اصلی بکار رفته برای ارزیابی، شامل ۴۴۸ تصویر تمیز و بدون نویز بود که سیستم ارائه شده با دقت بیش از ۹۸٪ قلم و اندازه قلم آنها را تشخیص داد. همچنین سیستم کلی بازیابی کلمات، با دقت ۸۶٪ در نرخ بازیابی ۸۲٪ قادر به بازیابی کلمات از مستندات تصویری بود. این نرخ دقت و بازیابی با ارزیابی سیستم بر روی ۲۰۰ کلمه فارسی بدست آمده است. همچنین یک پایگاه مستندات تصویری کوچک از مستندات تصویری اسکن شده (شامل ۱۳ مستند تصویری نوشته شده در نرم افزار Ms Word که به صورت کاملا تمیز و بدون کجی اسکن شده اند) هم برای بررسی امکان پیاده سازی سیستم بر روی تصاویر اسکن شده واقعی ایجاد شد که مشاهده شد که هر دو روش ارائه شده قابل پیاده سازی بر روی تصاویر اسکن شده واقعی هم هستند. علاوه بر پایگاه تصاویر گفته شده، عملکرد سیستم بر روی یک پایگاه تصویری شامل تعدادی تصویر اسکن شده در شرایط غیر ایده آل دارای نویز و کجی هم مورد آزمایش قرار گرفت که نتایج حاصل نشان دهنده ضعف سیستم تشخیص قلم در تشخیص قلم مستندات نویزی و کج می باشد.

کلمات کلیدی: بازیابی مستندات تصویری فارسی، جستجوی کلمات کلیدی، تشخیص قلم، تشخیص

اندازه قلم، آستانه گذاری، هیستوگرام



اظهارنامه دانشجو

شماره:

تاریخ:

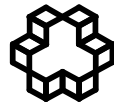
اینجانب **یعقوب پوراسد** دانشجوی دکترای رشته برق گرایش الکترونیک دانشکده مهندسی برق و کامپیوتر دانشگاه صنعتی خواجه نصیرالدین طوسی گواهی می نمایم که تحقیقات انجام شده در پایان نامه با عنوان

بازیابی متون چاپی فارسی بر اساس پرس و جوی کلمات

با راهنمایی استاد محترم **جناب آقای دکتر هوشنگ حسینی**، توسط شخص اینجانب انجام شده و صحت و اصالت مطالب نگارش شده در این پایان نامه مورد تأیید می باشد، و در مورد استفاده از کار دیگر محققان به مرجع مورد استفاده اشاره شده است. بعلاوه گواهی می نمایم که مطالب مندرج در پایان نامه تاکنون برای دریافت هیچ نوع مدرک یا امتیازی توسط اینجانب یا فرد دیگری در هیچ جا ارائه نشده است و در تدوین پایان نامه چارچوب (فرمت) مصوب دانشگاه را به طور کامل رعایت کرده ام.

امضا دانشجو :

تاریخ :



حق طبع و نشر و مالکیت نتایج

شماره:

تاریخ:

- ۱- حق چاپ و تکثیر این پایان نامه متعلق به نویسنده آن می باشد. هرگونه کپی برداری بصورت کل پایان نامه یا بخشی از آن تنها با موافقت نویسنده یا کتابخانه دانشکده مهندسی برق و کامپیوتر دانشگاه صنعتی خواجه نصیرالدین طوسی مجاز می باشد.
ضمناً متن این صفحه نیز باید در نسخه تکثیر شده وجود داشته باشد.
- ۲- کلیه حقوق معنوی این اثر متعلق به دانشگاه صنعتی خواجه نصیرالدین طوسی می باشد و بدون اجازه کتبی دانشگاه به شخص ثالث قابل واگذاری نیست.
همچنین استفاده از اطلاعات و نتایج موجود در پایان نامه بدون ذکر مراجع مجاز نمی باشد.



تأییدیه هیأت داوران

شماره:
تاریخ:

هیأت داوران پس از مطالعه پایان نامه و شرکت در جلسه دفاع از پایان نامه تهیه شده تحت عنوان :

بازیابی متون چاپی فارسی بر اساس پرس و جوی کلمات

توسط آقای یعقوب پوراسد ، صحت و کفایت تحقیق انجام شده را برای اخذ درجه دکترای رشته برق گرایش
الکترونیک در تاریخ
مورد تأیید قرار می دهند.

-
- | | | |
|------|-------------------------------------|---------------------------|
| امضا | جناب آقای دکتر هوشنگ حسینی | ۱- استاد راهنما |
| امضا | جناب آقای دکتر محمد تشنه لب | ۲- ممتحن داخلی |
| امضا | جناب آقای حمید ابریشمی مقدم | ۳- ممتحن داخلی |
| امضا | جناب آقای دکتر احسان اله کبیر | ۴- ممتحن خارجی |
| امضا | جناب آقای دکتر افشین ابراهیمی | ۵- ممتحن خارجی |
| امضا | جناب آقای دکتر مسعود علی اکبر گلکار | ۶- نماینده تحصیلات تکمیلی |

.....	فهرس مطالب	أ
.....	فهرست جداول	د
.....	فهرست اشكال	ه
.....	فهرست علائم و اختصارات	ح
.....	فصل اول: بازياي مستندات تصويري	ا
.....	1-1 مقدمه	ب
.....	1-2 ويژگي هاي نوشتاري زبان فارسي	ب
.....	1-2-1 پيوستگي حروف	ب
.....	1-2-2 اشكال متفاوت يك حرف	ب
.....	1-2-3 همپوشاني بين حروف	ب
.....	1-2-4 نقطه دار بودن حروف	ب
.....	1-2-5 وجود اعراب	ب
.....	1-2-6 اندازه متفاوت حروف	ب
.....	1-2-7 ادغام حروف مجاور	ب
.....	1-2-8 تنوع فراوان در شيوه هاي نگارش	ب
.....	1-3 بازياي اطلاعات	ب
.....	1-4 معيارهاي فاصله	ب
.....	1-4-1 معيار فاصله مينكووسكي	ب
.....	1-4-2 معيار فاصله زاويه اي	ب
.....	1-4-3 معيار فاصله فو	ب
.....	1-4-4 معيار فاصله ماھالانوييس	ب
.....	1-4-5 معيار فاصله WMV	ب
.....	1-5 نتيجه گيري	ب
.....	فصل دوم: جستجوي كلمات كليدي از مستندات تصويري	ب
.....	2-1 مقدمه	ب
.....	2-2 زبان هاي مورد استفاده در مقالات جستجوي كلمات	ب
.....	2-3 سيستم هاي ارائه شده براي جستجوي كلمات	ب

۲۰	پیش پردازش.....
۲۳	۲-۵ ویژگیهای مورد استفاده در جستجوی کلمات کلیدی.....
۲۹	۲-۶ مقایسه تصاویر (اندازه گیری شباهت/ عدم شباهت).....
۳۰	۲-۶-۱ متدها و الگوریتم های مقایسه تصاویر.....
۳۱	۲-۶-۱-۱ مقایسه تصویر به تصویر.....
۳۵	۲-۶-۱-۲ روش های نگاشت تصویر به کد.....
۴۶	۲-۶-۲ روش های مقایسه مبتنی بر آموزش.....
۴۷	۲-۶-۳ سطوح مقایسه.....
۴۷	۲-۷ ارزیابی عملکرد سیستم.....
۴۹	۲-۸ نتیجه گیری.....
۵۰	فصل سوم: تشخیص قلم تصویر اسناد.....
۵۱	۳-۱ مقدمه.....
۵۳	۳-۲ روش ارائه شده برای تشخیص قلم مستندات فارسی.....
۵۷	۳-۳ پایگاه تصاویر.....
۶۰	۳-۴ استخراج ویژگی ها در سیستم تشخیص قلم.....
۶۴	۳-۵ مقایسه ویژگی ها در سیستم تشخیص قلم.....
۶۶	۳-۶ نتیجه گیری.....
۶۷	فصل چهارم: روش پیشنهادی جهت جستجوی کلمات کلیدی از مستندات تصویری فارسی.....
۶۸	۴-۱ مقدمه.....
۷۰	۴-۲ روش ارائه شده برای جستجوی کلمات.....
۷۲	۴-۲-۱ پیش پردازش.....
۷۳	۴-۲-۲ آستانه گذاری.....
۷۴	۴-۲-۳ بازنویسی کلمه و جستجوی آن در مستند تصویری.....
۷۵	۴-۲-۳-۱ جستجوی کلمات به صورت پیکسل به پیکسل با تابع XNOR.....
۸۱	۴-۲-۳-۲ جستجوی کلمات با استفاده از ویژگی عرض و ارتفاع شبه کلمات.....
۸۴	۴-۳ نتیجه گیری.....
۸۶	فصل پنجم: نتایج تجربی و بحث.....

۸۷	۵-۱ مقدمه.....
۸۸	۵-۲ بررسی عملکرد سیستم بر مستندات تمیز بدون نویز و کجی.....
۱۰۹	۵-۳ بررسی تاثیر نویز و کجی بر عملکرد سیستم ارائه شده.....
۱۰۹	۵-۳-۱ بررسی اثر نویز.....
۱۲۱	۵-۳-۲ بررسی اثر کجی.....
۱۲۶	۵-۴ بررسی عملکرد سیستم بر مستندات اسکن شده در شرایط غیر ایده آل.....
۱۴۰	۵-۵ بررسی تاثیر تغییرات رزولوشن تصاویر بر سیستم ارائه شده.....
۱۴۶	۵-۶ نتیجه گیری.....
۱۴۸	فصل ششم: نتیجه گیری و ارائه پیشنهادات.....
۱۵۸	لیست مقالات ارائه شده.....
۱۵۹	مراجع.....

فهرست جداول

- جدول ۱-۱: مجموعه حروف فارسی و شکل آنها در موقعیت‌های مختلف
- جدول ۲-۱: تعدادی از زبان‌هایی که مقالات جستجوی کلمات برای آنها ارائه شده است [۸۱]
- جدول ۲-۲: سطوح بخش‌بندی در مقالات مختلف [۸۱]
- جدول ۲-۳: ویژگی‌ها و فرمت‌های استفاده شده در مقالات مختلف [۸۱]
- جدول ۲-۴: ویژگی‌های در نظر گرفته شده برای استخراج کدها [۳۰]
- جدول ۲-۵: کدهای اختصاص یافته به کاراکترهای مختلف [۳۰]
- جدول ۲-۶: کدهای اختصاص یافته به کاراکترهای انگلیسی بر اساس روش مرجع [۵۰]
- جدول ۲-۷: مقادیر دقت بدست آمده برای تعدادی از سیستم‌های جستجوی کلمات کلیدی در زبان‌های مختلف [۸۱]
- جدول ۵-۱: حالت‌های در نظر گرفته شده برای تشخیص قلم که به صورت مخفف نوشته شده‌اند
- جدول ۵-۲: مثالی از مقادیر شباهت بدست آمده وقتی که مستند ورودی دارای قلم Lotus 8 بود
- جدول ۵-۳: مثالی از مقادیر شباهت بدست آمده وقتی که مستند ورودی دارای قلم Koodak 18 بود
- جدول ۵-۴: تعداد و درصد خطاها با فقط یک بار آستانه‌گذاری (تعداد کل آزمایشات: ۴۴۸)
- جدول ۵-۵: تعداد و درصد خطاها با سه بار آستانه‌گذاری
- جدول ۵-۶: تعداد و درصد خطاها با در نظر گرفتن مقدار آستانه ۶
- جدول ۵-۷: نرخ‌های بازشناسی قلم (بر حسب درصد) ارائه شده در چندین مقاله و نیز روش تشخیص قلم ما
- جدول ۵-۸: زمان تقریبی ارائه شده در مقالات مختلف تشخیص قلم و نیز روش ما
- جدول ۵-۹: مقادیر دقت و بازیابی سیستم ما در مقادیر مختلف آستانه شباهت
- جدول ۵-۱۰: نتایج بازیابی چند سیستم عربی و نتایج سیستم ما
- جدول ۵-۱۱: نرخ تشخیص قلم (بر حسب درصد) روش ارائه شده در مقادیر مختلف نویز یکنواخت
- جدول ۵-۱۲: نرخ تشخیص قلم (بر حسب درصد) در مقادیر مختلف نویز نمک و فلفل
- جدول ۵-۱۳: نرخ دقت و بازیابی بدست آمده با مقادیر مختلف نویز نمک و فلفل (با صرف‌نظر از خطاهای سیستم تشخیص قلم)
- جدول ۵-۱۴: نرخ دقت و بازیابی بدست آمده با مقادیر مختلف نویز نمک و فلفل (با در نظر گرفتن خطاهای سیستم تشخیص قلم)
- جدول ۵-۱۵: میانگین میزان دقت و بازیابی بدست آمده در مقادیر مختلف نویز یکنواخت (با صرف‌نظر از خطای تشخیص قلم)
- جدول ۵-۱۶: نرخ‌های دقت و بازیابی بدست آمده در مقادیر مختلف کجی (با صرف‌نظر از خطاهای تشخیص قلم)
- جدول ۵-۱۷: تاثیر کجی بر کل سیستم بازیابی ارائه شده
- جدول ۵-۱۸: مقادیر بیشینه شباهت بدست آمده پس از مقایسه کلمه "تفسیر" در ۱۶ حالت مختلف در یک مستند تصویری
- جدول ۵-۱۹: میانگین مقادیر دقت و بازیابی بدست آمده برای مستندات اسکن شده کتابی در مقادیر مختلف آستانه

فهرست اشکال

- شکل ۱-۱: روند نمای بازیابی مستندات تصویری [۸۰]
- شکل ۱-۲: بازیابی مستندات بر اساس جستجوی کلمات کلیدی [۸۰]
- شکل ۲-۱: نواحی بالارونده، پایین رونده، و میانی [۸۱]
- شکل ۲-۲: نواحی بالارونده، پایین رونده، و میانی
- شکل ۲-۳: نمونه ای از ویژگی شبکه ای برای یک کلمه در ۳ قلم
- شکل ۲-۴: انواع متدها و الگوریتم های مقایسه تصاویر [۸۱]
- شکل ۲-۵: روند نمای یک سیستم جستجوی کلمات کلیدی مبتنی بر کدینگ شکل کلمه [۸۱]
- شکل ۲-۶: محل خطوط مختلف و تجزیه یک عبارت به پاره خط های آن [۳۰]
- شکل ۲-۷: تحلیل هیستوگرام افقی و عمودی جهت مکان یابی کلمه ها و خطوط [۳۰]
- شکل ۲-۸: دو ویژگی کلمه شامل نقاط اکسترمم کاراکتر و تعداد برش های افقی کلمه [۵۰]
- شکل ۳-۱: یک خط فارسی که با ۵ قلم متداول نوشته شده است
- شکل ۳-۲: یک خط فارسی و نقاط آن
- شکل ۳-۳: یک متن کوتاه فارسی و بعضی از نقاط موجود در آن
- شکل ۳-۴: تعدادی قطعه تک نقطه ای قبل و پس از آستانه گذاری و ابعاد آنها
- شکل ۳-۵: تعدادی قطعه دو نقطه ای قبل و پس از آستانه گذاری و ابعاد آنها
- شکل ۳-۶: تصویری از مستند فارسی در قلم Koodak 8
- شکل ۳-۷: تصویری از مستند فارسی در قلم Homa 8
- شکل ۳-۸: تصویری از مستند فارسی در قلم Lotus 20
- شکل ۳-۹: سه هیستوگرام مربوط به دو حالت مختلف
- شکل ۳-۱۰: هیستوگرامهای مربوط به همان حالت های شکل ۳-۹ در مقدار آستانه متفاوت (۰/۶)
- شکل ۴-۱: روند نمای روش جدید پیشنهادی برای جستجوی کلمات فارسی
- شکل ۴-۲: مراحل مختلف بازنویسی کلمه مورد جستجو با نرم افزار MATLAB
- شکل ۴-۳: عرض و فاصله خطوط واقعی و زائد
- شکل ۴-۴: مراحل جستجو با XOR
- شکل ۴-۵: میزان شباهت (XNOR) و نیز میزان شباهت نرمالیزه برای مقایسه تصاویر شکل ۴-۴
- شکل ۴-۶: چندین کلمه فارسی و زیر کلمات آنها

- شکل ۴-۷: زیر کلمات استخراج شده برای یک مستند تصویری
- شکل ۴-۸: زیر کلمات یافته شده برای کلمه مورد جستجوی "جداکننده ها"
- شکل ۴-۹: نمونه یافته شده کلمه مورد جستجوی "جداکننده ها" در مستند تصویری
- شکل ۵-۱: قسمتی از یک مستند تصویری اسکن شده
- شکل ۵-۲: هیستوگرام مربوط به دو حالت مختلف از مستندات اسکن شده فارسی
- شکل ۵-۳: GUI مورد استفاده برای سیستم جستجوی کلمات
- شکل ۵-۴: یک مستند تصویری بدون نویز نوشته شده در قلم lotus 10
- شکل ۵-۵: مستند تصویری شکل ۴-۵ که با نویز یکنواخت با واریانس $0/005$ نویزدار شده است
- شکل ۵-۶: مستند تصویری شکل ۴-۵ که با نویز یکنواخت با واریانس $0/03$ نویزدار شده است
- شکل ۵-۷: مستند تصویری شکل ۴-۵ که با نویز یکنواخت با واریانس $0/1$ نویزدار شده است
- شکل ۵-۸: مستند تصویری شکل ۴-۵ که با نویز یکنواخت با واریانس $0/5$ نویزدار شده است
- شکل ۵-۹: یک مستند تصویری بدون نویز نوشته شده در قلم koodak 10
- شکل ۵-۱۰: مستند تصویری شکل ۹-۵ که با نویز نمک و فلفل با $d = 0/01$ نویزدار شده است
- شکل ۵-۱۱: مستند تصویری شکل ۹-۵ که با نویز نمک و فلفل با $d = 0/02$ نویزدار شده است
- شکل ۵-۱۲: مستند تصویری شکل ۹-۵ که با نویز نمک و فلفل با $d = 0/035$ نویزدار شده است
- شکل ۵-۱۳: مستند تصویری شکل ۹-۵ که با نویز نمک و فلفل با $d = 0/1$ نویزدار شده است
- شکل ۵-۱۴: جستجوی کلمه "صفحه" از مستندی با قلم Mitra 14 بدون نویز
- شکل ۵-۱۵: جستجوی کلمه "صفحه" از مستندی با قلم Mitra 14 دارای نویز فلفل و نمک با $d = 0/01$
- شکل ۵-۱۶: جستجوی کلمه "صفحه" از مستندی با قلم Mitra 14 دارای نویز فلفل و نمک با $d = 0/02$
- شکل ۵-۱۷: جستجوی کلمه "صفحه" از مستندی با قلم Mitra 14 دارای نویز فلفل و نمک با $d = 0/05$
- شکل ۵-۱۸: جستجوی کلمه "صفحه" از مستندی با قلم Mitra 14 دارای نویز فلفل و نمک با $d = 0/1$
- شکل ۵-۱۹: جستجوی کلمه "صفحه" از مستندی با قلم Mitra 14 دارای نویز فلفل و نمک با $d = 0/2$
- شکل ۵-۲۰: جستجوی کلمه "کلمه" از یک مستند تصویری بدون نویز با قلم Titr 12
- شکل ۵-۲۱: جستجوی کلمه "کلمه" از مستند تصویری شکل ۲۰-۵ با نویز یکنواخت با $var = 0/005$
- شکل ۵-۲۲: جستجوی کلمه "کلمه" از مستند تصویری شکل ۲۰-۵ با نویز یکنواخت با $var = 0/02$
- شکل ۵-۲۳: جستجوی کلمه "کلمه" از مستند تصویری شکل ۲۰-۵ با نویز یکنواخت با $var = 0/03$
- شکل ۵-۲۴: جستجوی کلمه "کلمه" از مستند تصویری شکل ۲۰-۵ با نویز یکنواخت با $var = 0/05$
- شکل ۵-۲۵: جستجوی کلمه "کلمه" از مستند تصویری شکل ۲۰-۵ با نویز یکنواخت با $var = 0/06$

- شکل ۲۶-۵: یک مستند تصویری که ۱ درجه کج شده است
- شکل ۲۷-۵: یک مستند تصویری که ۳ درجه کج شده است
- شکل ۲۸-۵: کلمات یافته شده برای کلمه مورد جستجوی "محلول" در یک مستند با ۰ درجه کجی
- شکل ۲۹-۵: کلمات یافته شده برای کلمه مورد جستجوی "محلول" در یک مستند با ۱ درجه کجی
- شکل ۳۰-۵: کلمات یافته شده برای کلمه مورد جستجوی "محلول" در یک مستند با ۲ درجه کجی
- شکل ۳۱-۵: کلمات یافته شده برای کلمه مورد جستجوی "محلول" در یک مستند با ۶ درجه کجی
- شکل ۳۲-۵: کلمات یافته شده برای کلمه مورد جستجوی "محلول" در یک مستند با ۷ درجه کجی
- شکل ۳۳-۵: یک مستند تصویری کتابی دارای کمی نویز اما تقریباً بون کجی
- شکل ۳۴-۵: یک مستند تصویری کتابی دارای کجی و نویز
- شکل ۳۵-۵: یک مستند تصویری کتابی دارای شکل مدار، سیگنال، زیرنویس اشکال، تقریباً بدون نویز و کجی
- شکل ۳۶-۵: یک مستند تصویری کتابی، دارای کمی نویز، شکل مدار، نمودار، سیگنال، بالانویس، فرمول و ...
- شکل ۳۷-۵: یک نامه اداری دارای یک مجموعاً متنوعی از قلم ها، آرم ها و ...
- شکل ۳۸-۵: یک مستند تصویری مربوط به روزنامه
- شکل ۳۹-۵: مستند تصویری مربوط به یک مجله
- شکل ۴۰-۵: انتخاب یک کلمه دلخواه از یک مستند و تلاش برای یافتن قلم و اندازه قلم آن مستند
- شکل ۴۱-۵: یک مستند تصویری اسکن شده در رزولوشن ۴۰۰ dpi
- شکل ۴۲-۵: بردار هیستوگرام مستند شکل ۴۱-۵ اسکن شده با رزولوشن ۱۵۰ dpi
- شکل ۴۳-۵: بردار هیستوگرام مستند شکل ۴۱-۵ اسکن شده با رزولوشن ۲۰۰ dpi
- شکل ۴۴-۵: بردار هیستوگرام مستند شکل ۴۱-۵ اسکن شده با رزولوشن ۳۰۰ dpi
- شکل ۴۵-۵: بردار هیستوگرام مستند شکل ۴۱-۵ اسکن شده با رزولوشن ۴۰۰ dpi
- شکل ۴۶-۵: حروف حفره دار موجود در فارسی
- شکل ۴۷-۵: یک مستند تصویری و حفره های استخراج شده آن
- شکل ۴۸-۵: تعدادی از حروف فارسی که دارای یکی از ۴ ویژگی حفره ها، بالارونده، پایین رونده و نقاط هستند
- شکل ۴۹-۵: مثالی برای بیان ویژگیهای ۴ کلمه فارسی

OCR: Optical Character Recognition

CBD: City Block Distance

CORR: Correlation (Used to represent Correlation Similarity Measures)

DTW: Dynamic Time Warping

EDM: Euclidean Distance Measure

K-NN: K-Nearest Neighbor

NSHP-HMM: Non-Symmetric Half Plane Hidden Markov Model

P2D-HMM: Pseudo 2 Dimensional Hidden Markov Model

PHMM: Planar Hidden Markov Model (Another name of P2D-HMM)

SLH: Scot and Longuet Higgins algorithm

SSD: Sum of Squared Distances

SC: Shape Context algorithm.

SRF : Sobel and Robert Filters

WED : Weighted Euclidean Distance

SVM : Support Vector Machine

GSC : Gradient Structural Concavity Features

DP : Dynamic Programming

CSC : Character Shape Coding

WSC : Word Shape Coding

BCT : B Classification Tree

فصل اول

بازیابی مستندات تصویری

از زمان پیدایش نوشتار برای برقراری ارتباط میان افراد کاغذ به عنوان یک ابزار مهم برای درج اطلاعات مورد استفاده قرار گرفته است. اما امروزه محیط‌های الکترونیکی به دلیل نگهداری آسان و دسترسی سریع رواج فراوانی یافته و به عنوان جایگزینی برای کاغذ مطرح شده اند. از سوی دیگر سهولت استفاده، فراگیر بودن و وجود حجم عظیمی از اطلاعات کنونی به روی کاغذ، محققان را بر آن داشته است تا به دنبال روش‌هایی برای خواندن اتوماتیک این اطلاعات از روی کاغذ و تبدیل آن به شکل الکترونیکی باشند. امروزه کاربردهای فراوانی برای شناسایی متون وجود دارد از آن میان می‌توان به پردازش چک‌های بانکی شناسایی آدرس و کدهای پستی، کمک به نابینایان برای خواندن و... اشاره کرد. پردازش نوشتار یکی از شاخه‌های مربوط به شناسایی الگو است که هدف نهایی آن تقلید از انسان برای خواندن و درک متون است. اما با وجود پیشرفت‌های به دست آمده، هنوز قدرت این سیستم‌های پردازش‌گر در خواندن حروف با توانایی بشر فاصله دارد. کاربردهای عملی فراوان و طبیعت جذاب این رشته توجه پژوهشگران زیادی را به خود جلب کرده و پیشرفت‌های محسوسی در این زمینه به دست آمده است. با پیدایش کامپیوتر تلاش برای حل مسئله شناسایی نوشتار از دهه ۱۹۴۰ میلادی با تحقیقات در مورد شناسایی حروف و ارقام مجزای چاپی لاتین شروع شد، اولین ماشین‌های پردازش‌گر حروف که سیستم شناسایی کننده آن‌ها به طور مکانیکی کار می‌کرد، در اوایل دهه ۱۹۵۰ میلادی به بازار عرضه شدند. این ماشین‌ها فقط قادر به شناسایی اعداد لاتین تایپ شده با یک قلم خاص بودند. در دهه ۱۹۷۰ میلادی، سیستم‌های شناسایی کننده نرم‌افزاری با توانایی تشخیص متون تایپی تک قلمی پدید آمدند. در دهه ۱۹۸۰ میلادی پیشرفت زیادی در زمینه شناسایی متون لاتین صورت گرفت به طوری که سیستم‌های

موجود می‌توانستند متون تایپ شده با چندین قلم را نیز با سرعت بالا تشخیص دهند. امروزه سیستم‌های تجاری بازشناسی متون لاتین علاوه بر تشخیص متون تایپی قادر به درک متون دستنویس نیز هستند. در چند دهه اخیر کارهای زیادی برای شناسایی حروف زبان‌های لاتین، چینی و ژاپنی انجام شده است. اما برای شناسایی حروف عربی و فارسی، با اینکه تعداد زیادی از مردم جهان برای نوشتن از این حروف استفاده می‌کنند، کارهای انجام شده نسبتاً کم و پراکنده بوده است و نتایج به دست آمده چندان رضایت‌بخش نیستند. شاید بتوان ویژگی‌های نوشتاری منحصر به فرد این زبان‌ها را به عنوان یکی از موانع موجود در این راه نام برد. در ادامه تعدادی از ویژگی‌های نوشتاری زبان فارسی به طور خلاصه ذکر می‌گردد.

۱-۲ ویژگی‌های نوشتاری زبان فارسی

۱-۲-۱ پیوستگی حروف

بر خلاف نوشتار لاتین، متون فارسی هم در حالت تایپی و هم در حالت دستنویس به صورت پیوسته از راست به چپ نوشته می‌شوند. اگر چه به صورت عادی در هر یک کلمه، هر حرف به حرف قبلی و بعدی متصل است، اما تعدادی از حروف هستند (حروف ا. د. ذ. ر. ز. ژ. و) که فقط می‌توانند از سمت راست به حروف ماقبل خود وصل شوند. وجود یک یا چند حرف از این مجموعه باعث می‌شود یک کلمه به تعدادی بخش پیوسته به نام شبه کلمه تقسیم شود. به عنوان مثال کلمه (کتاب) از دو زیر کلمه (کتا) و (ب) تشکیل شده است که این بخشها با فضای خالی از یکدیگر جدا می‌شوند.

۲-۲-۱ اشکال متفاوت یک حرف

شکل حروف در زبان فارسی تابعی از محل قرار گرفتن آن در کلمه است و هر حرف بر حسب موقعیت های مختلف در کلمه (اول، وسط، آخر و جدا) می تواند تا چهار شکل مختلف داشته باشد به عنوان مثال شکل حرف (ه) در حالت جدا باید به شکل (ه) در ابتدای کلمه، شکل (ه) در وسط کلمه و یا شکل (ه) در آخر کلمه تغییر پیدا کند.

۲-۳-۱ همپوشانی بین حروف

حروف در یک کلمه فارسی می توانند حتی بدون اتصال با هم، همپوشانی عمودی داشته باشند. معمولاً حروف (چ، ر، ژ، و) با حروف بعدی خود دارای همپوشانی هستند. علاوه بر این حروف قسمت های مکمل حروف (آ، ک) نیز می توانند بخشی از حروف مجاور خود را بپوشانند. این ویژگی در نوشتار فارسی، عمل تقطیع کلمات به حروف تشکیل دهنده آنها را بسیار مشکل کرده است.

۲-۴-۱ نقطه دار بودن حروف

بیش از نیمی از حروف فارسی نقطه دار هستند. به عبارتی دقیق تر ۱۰ حرف دارای یک نقطه ۳ حرف دارای دو نقطه و ۵ حرف دارای سه نقطه هستند. در بعضی از موارد، وجود و یا عدم وجود، تعداد و محل قرار گرفتن نقطه ها تنها عامل متمایز کننده بین حروف متشابه (ب، پ، ت، ث، ن، ی) (ج، چ، ح، خ)، (د، ذ)، (ر، ز، ژ)، (ص، ض) (س و ش)، (ع، غ)، (ف ق) است. جدول ۱-۱ مجموعه حروف فارسی و اشکال مختلف هر حرف در موقعیت های مختلف را نشان می دهد.

۵-۲-۱ وجود اعراب

در نوشتار فارسی، هنگامی که احتمال تلفظ اشتباه یک کلمه وجود داشته باشد صدا به صورت اعراب به بالا یا پائین برخی از حروف اضافه می‌شود. علاوه بر اعراب در نوشتار فارسی، علائمی از قبیل تشدید، تنوین، همزه و مد نیز وجود دارند. با توجه به شباهت ظاهری اعراب و این علائم مکمل با نقاط، جدا کردن آن‌ها به خصوص در کلمات دستنوشته کار مشکلی است.

۶-۲-۱ اندازه متفاوت حروف

اندازه و عرض تمام حروف فارسی یکسان نیستند. مثلاً حروف (ب)، (س) در حالت جدا اندازه بزرگ‌تری نسبت به حروف (د) و (ه) دارند. این تنوع در اندازه حروف کار قطعه‌بندی حروف را مشکل می‌کند.

۷-۲-۱ ادغام حروف مجاور

در برخی از شیوه‌های نوشتاری زبان فارسی، دو یا چند حرف کنار هم می‌توانند به گونه‌ای با هم ترکیب شوند که شکل حاصل شباهتی به حروف تشکیل دهنده آن نداشته باشد. چنین مواردی در نوشتار دستنویس، بلکه در متون تایپی نیز وجود دارد. متداول‌ترین ترکیب در متون تایپی، ادغام دو حرف (ل) و (ا) به صورت (لا) است. در نوشته‌های دستنویس فارسی نیز بخاطر کمک به زیبایی نوشتار و همچنین سلیقه نویسندگان، شکل بعضی از حروف کنار هم به کلی تغییر می‌کند.

۸-۲-۱ تنوع فراوان در شیوه‌های نگارش

در مقایسه با نوشتار لاتین، قلم‌های فارسی در حالت چاپی دارای واحدهای نوشتاری (شامل اعداد، اشکال مختلف حروف و علائم خاص) بسیار زیاد و در حالت دستنویس دارای سبک‌های متعدد می‌باشند. جدول ۱-۱ مجموعه حروف فارسی را در موقعیت‌های مختلف نشان می‌دهد.

جدول ۱-۱ مجموعه حروف فارسی و شکل آن‌ها در موقعیت‌های مختلف

مجزا	ابتدا	وسط	انتهای	مجزا	ابتدا	وسط	انتهای
ا	-	-	ا	ض	ض	ض	ض
ب	ب	ب	ب	ط	ط	ط	ط
ت	ت	ت	ت	ظ	ظ	ظ	ظ
ث	ث	ث	ث	ع	ع	ع	ع
ج	ج	ج	ج	غ	غ	غ	غ
ح	ح	ح	ح	ف	ف	ف	ف
خ	خ	خ	خ	ق	ق	ق	ق
د	-	-	د	ك	ك	ك	ك
ذ	-	-	ذ	ل	ل	ل	ل
ر	-	-	ر	م	م	م	م
ز	-	-	ز	ن	ن	ن	ن
س	س	س	س	ه	ه	ه	ه
ش	ش	ش	ش	و	-	-	و
ص	ص	ص	ص	ي	ي	ي	ي

۳-۱ بازیابی اطلاعات

امروزه با ازدیاد کتابخانه‌های دیجیتال و نیز در راستای ایجاد ادارات بدون کاغذ، تعداد فزاینده‌ای از مستندات تایپی و دست‌نویس شامل کتاب‌ها، روزنامه‌ها، مجلات، مقالات و . . . با کیفیت‌های