تشخیص متن چاپی فارسی با فونت ثابت با استفاده از شبه کلمات

کریم فائز

دانشیار دانشکده مهندسی برق، دانشگاه صنعتی امیرکبیر

محمدحسن شیرعلی شهرضا

دانشجوی دوره دکترای برق، دانشگاه صنعتی امیرکبیر

چکیده:

در روشهای معمول تشخیص متن، باید حروف هر کلمه جدا شده و سپس عمل تشخیص روی حروف مجزا انجام شود. استفاده از این روش در زبان فارسی مشکل است چون در زبان فارسی کلمات به صورت متصل نوشته می شوند و جدا کردن حروف یك کلمه بخصوص در مورد متون دست نویس به سادگی امکان پذیر نیست. در این مقاله روشی برای تشخیص متن با استفاده از «شبه کلمه» ارائه شده است. همچنین شبه کلمات مورد استفاده در زبان فارسی از نظر آماری بررسی شده و مشخص شده است که استفاده از روش تشخیص شبه کلمات عملاً امکان پذیر می باشد. مزیت عمده این روش سرعت زیاد آن می باشد، چون به جدا کردن حروف و سپس ترکیب حروف تشخیص داده شده نیازی ندارد. برای تشخیص شبه کلمات از روش تطبیق قالبها (Template matching) استفاده شده است. دقت تشخیص این روش ۹۶ درصد بوده وسرعت آن ۳ شبه کلمه (حدود ۵ حرف) در ثانیه می باشد.

لغات کلیدی: تشخیص الگو، تشخیص متن، تشخیص کلمات، تشخیص حروف عربی، تشخیص حروف فارسی.

Recognition of Fixed Font Printed Farsi Text Using Semi-Words

Karim Faez

Associate prof., Amirkabir Univ.

M. H. Shirali-Shahreza

Ph. D. Student, Amirkabir Univ.

Abstract:

In ordinary text recognition methods, characters of the text are first separated and then, the character recognition methods are applied on them. Since in Farsi texts, characters are connected to each other, separating the characters is a very difficult job. Therefore application of these methods will be very difficult and cumbersome, specially for handwritten texts.

This article introduces a new method for Farsi text recognition using "semi-words". Farsi semi-words are analyzed statistically and it is shown that appli-

cation of this method is much easier and faster than other methods. Finally, the accuracy and the speed of this method is compared with the other methods. The processing speed is 3 semi-words (5 characters) per second with a 96 percent recognition rate.

Key Words: *Pattern recognition, Text recognition, Word recognition, Arabic / Farsi character recognition.*

۱ ـ مقدمه

یکی از مسائل عمده در کامپیوتر، وارد کردن اطلاعات (Data Entry) می باشد. از آنجا که عمده اطلاعات ذخیره شده در کامپیوترها را اطلاعات نوشته شده (Text) تشکیل می دهد، برای وارد کردن این نوشته ها احتیاج به تایپ دوباره اطلاعات نوشته شده می باشد. برای اجتناب از تایپ مجدد متنهای نوشته شده این ایده مطرح شد که کامپیوتر بتواند یک متن نوشته شده را به طور اتوماتیک خوانده و تشخیص دهد. البته کاربرد تشخیص حروف منحصر به وارد کردن یک متن به کامپیوتر نبوده و کاربردهای زیاد دیگری نیز دارد. یکی ازکاربردهای عمده تشخیص حروف در اداره پست برای تفکیک اتوماتیک نامه ها می باشد. همچنین با استفاده از تشخیص حروف در سیستم بانکی می توان عملیات مربوط به چک و دیگر اسناد بانکی را سریعتر انجام داد. در ضمن اگر خروجی تشخیص حروف به یک سیستم تبدیل متن به صوت یا یک سیستم تبدیل متن به خط بریل [۷] متصل شود می تواند به اشخاص نابینا در خواندن کتاب و روزنامه کمک کند. استفاده دیگر تشخیص حروف در اتوماتیک کردن کارهای دفتری می باشد مثلاً می تواند در آرشیو کردن، ترجمه اتوماتیک و پاسخگویی اتوماتیک نامه ها استفاده شود.

گرچه بیش از ۳۰ سال از مطرح شدن ایده تشخیص حروف می گذرد [۱۲] هنوز این موضوع به عنوان موضوعی مهم مورد توجه متخصصین کامپیوتر بوده و تحقیق روی آن ادامه دارد. عمده کارهای انجام شده در زمینه تشخیص حروف مربوط به زبان لاتین می باشد. نتیجه این تحقیقات به صورت سیستمهای تجاری که قادرند یک متن تایپ شده را بخوانند به بازار عرضه شده اند.

متأسفانه در زمینه تشخیص حروف فارسی و عربی خیلی کم کار شده است. اولین مقاله مربوط به تشخیص حروف چاپی فارسی در سال ۱۳۵۹ چاپ شده است [۱۴]. در مورد تشخیص حروف چاپی عربی می توان به مقالات [۸]، [۹] و [۱۳] اشاره کرد، ولی تا کنون سیستم تجاری برای خواندن اتوماتیک متون فارسی و عربی به بازار نیامده است.

برای تشخیص حروف، دو روش وجود دارد. در روش اول ابتدا حروف هر کلمه جدا شده و تشخیص، روی حروف جدا شده انجام می شود [۲] [۹]. در روش دوم به جای تشخیص حروف، تشخیص کلمه انجام می شود و کل کلمه مورد تشخیص قرار می گیرد. اشکال روش اول برای حروف فارسی این است که جدا کردن حروف در زبان فارسی مشکل می باشد. روش دوم نیز به خاطر تعداد زیاد کلمات قابل پیاده کردن نیست مگر تعداد کلمات را به طریقی محدود کنیم. در این مقاله یک روش میانه که تشخیص با استفاده از «شبه کلمات» می باشد، پیشنهاد شده است. منظور از شبه کلمه، تعدادی حرف متصل به هم می باشد.

ساختار مقاله به صورت زیر است: در بخش ۲ ویژگیهای زبان فارسی از نقطه نظر تشخیص حروف بررسی شده است. در بخش ۳ شبه کلمه تعریف شده و نتایج آماری بررسی شبه کلمات در زبان فارسی ارائه شده است. در بخش ۴ نحوه ورود متن به کامپیوتر و پردازشهای مقدماتی که روی آن انجام می شود، آمده است. در بخش ۵ روش استفاده شده برای تشخیص شبه کلمات توضیح داده شده و در بخش ۶ نتایج عملی و در بخش ۷ نتیجه گیری نهایی آمده است.

۲ـ ویژگیهای زبان فارسی از نظر تشخیص حروف

برای تشخیص حروف فارسی لازم است که ویژگیهای زبان فارسی را به خوبی شناخته و طراحی سیستم تشخیص حروف را متناسب با این ویژگیها انجام داد. ویژگیهای زبان فارسی از نقطه نظر تشخیص حروف در این بخش بیان می شود.

۲-۱ـ نوشتن از راست به چپ

در زبان فارسی نوشتن از راست به چپ انجام می شود، برخلاف زبان انگلیسی که در آن از چپ به راست و یا در زبانهای دیگری که از بالا به پایین می نویسند. بنابراین تشخیص حروف فارسی نیز باید از راست به چپ انجام شود تا عمل تصحیح نتایج به دست آمده یا تبدیل نتایج به دست آمده به صوت ساده تر باشد.

۲-۲ـ متصل بودن حروف

در زبان فارسی، حروف هنگام نوشتن به یکدیگر متصل می شوند بر خلاف زبان انگلیسی که در آن حروف به صورت مجزا نوشته می شوند. البته حروف دست نویس انگلیس نیز به صورت متصل نوشته می شوند.

۲-۳ـ نقطه دار بودن بعضی حروف

در زبان فارسی، نقطه اهمیت زیادی داشته و ۵۰ درصد از حروف الفبای فارسی نقطه دارند. اهمیت نقطه از این نظراست که تعدادی از حروف فقط در تعداد یا محل نقاط با یکدیگر اختلاف دارند. جدول (۱) حروفی که فقط در تعداد یا محل نقاط با یکدیگر اختلاف دارند را نشان می دهد. چنانچه در این جدول دیده می شود، حروف می توانند بدون نقطه بوده یا یک تا سه نقطه داشته باشند.

اهمیت نقطه در زبان فارسی از این جهت بیشتر است که برای تبدیل یک فعل امر به فعل نهی حرف «ب» به حرف

جدول (۱)
حروفی که فقط در تعداد یا محل نقاط با یکدیگر اختلاف دارند

ب	ج	د	ر	س	ص	ط	ع	ء	ه	ح	ف
پ	چ	ذ	ز	ش	ض	ظ	غ	ئ	ة	خ	ق
ت	ح		ژ								
ث	خ										

«نـ» تبدیل می شود یعنی با تغییر جای یک نقطه یک جمله مثبت به جمله منفی تبدیل می شود. مثلاً جمله «درس را بخوان» با جمله نهی آن یعنی «درس را نخوان» فقط در محل یک نقطه اختلاف دارند.

مشکل عمده نقطه در تشخیص حروف، این است که نقطه ممکن است با نویز موجود در تصویر یک متن اشتباه شود. مشکل دیگر نقطه این است که چون نقطه ها به یکدیگر می چسبند گاهی تشخیص بین دو نقطه یا سه نقطه امکان پذیر نیست.

۲-۴ـ علائم خاص

در زبان فارسی از علائم «تشدید»، «تنوین»، «همزه» و «مد» نیز استفاده می شود، گر چه بعضی از این علائم مختص زبان عربی می باشد ولی کمتر متن فارسی را می توان پیدا کرد که در آن این علائم موجود نباشد. این علائم روی حروف قرار می گیرند. مثالهایی برای این علائم، تشدید در کلمه «بچّه»، تنوین در کلمه «حتماً»، مد در کلمه «قرآن» و همزه در کلمه «سؤال» می باشند.

علامت دیگر، علامت «یای کوتاه شده» یا «یای میانجی» («ء») می باشد که روی حرف «ه» قرار می گیرد مثلاً در عبارت «خانهٔ دوست».

۲-۵ـ کشیدگی کلمات

در هنگام تایپ متون فارسی برای اینکه انتهای جملات در یک ستون قرار گیرند، در کلمات آخر جمله از علامت «ـ» برای کشیده نوشتن کلمه استفاده می شود، که این علامت هیچ معنای خاصی نداشته و فقط برای زیبائی متن استفاده می شود. مانند کمله «باشد» که به صورت «بـــاشد» نیز نوشته می شود.

۲-۶ـ شکلهای مختلف یک حرف

در زبان فارسی یک حرف می تواند تا چهار شکل مختلف داشته باشد. شکل هر حرف متناسب با محل قرار گرفتن آن حرف در کلمه می باشد. مثلاً حرف «عین» در اول کلمه به صورت «ع» در وسط کلمه به صورت «ـعـ»، در آخر کلمه به صورت «ـع» و در حالت منفرد به صورت «ع» نوشته می شود.

۲-۷ـ اعراب

در زبان فارسی صداها موقع نوشتن کلمه نوشته نمی شوند ولی در محلهایی که احتمال اشتباه در موقع خواندن وجود دارد لازم است که برای بعضی از حروف، اعراب گذاشته شود. اعرابهایی که در زبان فارسی استفاده می شوند مشابه زبان عربی بوده و عبارتند از فتحه (ـَ) ضمه (ـُ) و کسره (ـِ). اعراب در بالا و پایین حروف گذاشته شده و

خواندن صحیح آن را امکان پذیر می کند. جدا کردن اعراب از خود حروف در موقع تشخیص حروف کارآسانی نیست.

۲-۸- هم پوشانی حروف

در موقع حروفچینی یا تایپ متون فارسی، بعضی از حروف روی حرف قبلی قرار می گیرند مثلاً در هنگام چاپ کلمه «را» حرف «الف» روی حرف «ر» قرار می گیرد، این مسأله نیز کار تشخیص حروف فارسی را سخت تر می کند.

۲-۹- شکل خاص «لا»

در زبان فارسی به خاطر زیبایی متن هر گاه حروف «لام» و «الف» پشت سرهم قرار گیرند به صورت «لا» نوشته می شوند، این مسأله در هنگام مرتب کردن یا جستجوی یک متن فارسی توسط کامپیوتر، مشکل ایجاد می کند. در تشخیص حروف می توان شکل «لا» را به عنوان یک حرف مستقل در نظر گرفت.

۲-۱۰- اندازه متفاوت حروف

در زبان فارسی حروف از نظر اندازه یکسان نیستند، مثلاً حرف «ب» هنگام چاپ جای بیشتری از حرف «د» اشغال می کند. یکسان نبودن اندازه حروف به پیچیدگی تشخیص حروف فارسی کمک می کند.

۲-۱۱- نبودن فاصله بین کلمات

در زبان فارس بر خلاف زبان انگلیسی که هر کلمه با کلمه بعد از آن به وسیله فاصله جدا می شود، بین کلمات فاصله وجود ندارد. به همین علت جدا کردن کلمات بدون توجه به تمامی جمله امکان پذیر نیست.

در تشخیص حروف، نبودن فاصله بین کلمات باعث می شود که عمل تصحیح متن تشخیص داده شده با استفاده از فرهنگ لغت، مشکل شود.

۳- شبه کلمه

۳-۱- تعریف

در این مقاله برای تشخیص کلمات فارسی از قسمتهای متصل هر کلمه، که شبه کلمه می نامیم، استفاده شده است. در اینجا شبه کلمه را به صورت زیر تعریف می کنیم.

شبه کلمه = مجموعه ای از حروف الفبای فارسی که به یکدیگر متصل باشند.

یک شبه کلمه می تواند خود یک کلمه مستقل باشد، مثل شبه کلمه «متصل» که یک کلمه کامل بوده و تمامی حروف آن به یکدیگر چسبیده اند، یا قسمتی از یک کلمه باشد مثل شبه کلمه «کا» در کلمه «کارخانه». چنانچه می دانیم الفبای فارسی از ۳۲ حرف تشکیل شده است. از نظر نگارش می توان این حروف را به دو دسته تقسیم کرد، دسته اول حروفی که موقع نوشتن می توانند به حرف بعد از خود متصل شوند و دسته دوم حروفی که نمی توانند به حرف بعد از خود متصل شوند و باید جدا نوشته شوند. چون در زبان فارسی حروف فقط روی خط زمینه به یکدیگر متصل می شوند، اگر انتهای حرفی روی خط زمینه قرار گیرد می تواند به حرف بعد از خودش متصل شود در غیراین صورت باید به صورت جدا نوشته شود، مثلاً حرف «ط» چون انتهای آن روی خط زمینه است به حرف بعد ازخودش متصل می شود ولی حرف «و» نمی تواند به حرف بعد از خودش متصل شود. جدول (۲) الفبای فارسی از نظر نگارش را نشان می دهد.

جدول (۲)

الفبای زبان فارسی از نظر نگارش

حروف متصله	حروف منفصله
حروفی که متصل نوشته می‌شوند	حروفی که به حرف بعد از خود متصل نمی‌شوند
ب ـ ـبـ ـت ـ ـث ـ ـج ـ ـچ ـ ـح ـ ـخ ـ ـس ـ ـش ـ ـص ـ ـض ـ ـط ـ ـظ ـ ـع ـ ـغ ـ ـف ـ ـق ـ ـک ـ ـگ ـ ـل ـ ـم ـ ـن ـ ـه ـ ـی	الف ـ ـد ـ ـذ ـ ـر ـ ـز ـ ـژ ـ ـو

۳-۲- بررسی آماری شبه کلمات

برای بررسی امکان عملی بودن تشخیص متون فارسی با استفاده از شبه کلمات، مطالعه آماری روی ترجمه قرآن مجید به زبان فارسی [۱] انجام شد. نتایج این بررسی در جدول (۳) نشان داده شده است. در این جدول شبه کلمات بر اساس تعداد حروف موجود در آنها مرتب شده اند. با توجه به این نتایج می توان تعداد متوسط حروف یک شبه کلمه را به دست آورد. برای محاسبه تعداد متوسط حروف شبه کلمات از رابطه زیر استفاده شده است:

$$\bar{n}=\sum_{n=1}^{6} n * p(n)$$

که n تعداد حروف موجود در شبه کلمه و p (n) احتمال

وقوع آن می باشد. p (n)، با استفاده از رابطه زیر به دست آمده است:

$$p(n) = \frac{\text{تعداد شبه کلمات n حرفی}}{\text{تعداد کل شبه کلمات}}$$

با قرار دادن نتایج جدول (۳) در روابط فوق خواهیم داشت:

$$\bar{n} = ۱\times ۰/۵۰ + ۲\times ۰/۳۳ + ۳\times ۰/۱۲ + ۴\times ۰/۰۳ + ۵\times ۰/۰۱ + ۶\times ۰/۰۱ = ۱/۷۵ \text{ حرف}$$

یعنی در زبان فارسی به طور متوسط ۱/۷۵ حرف در یک شبه کلمه وجود دارد و این کوچک بودن شبه کلمات، یکی از دلایل استفاده از شبه کلمات در تشخیص حروف است. دلیل دیگر مشکل بودن جداسازی حروف می باشد.

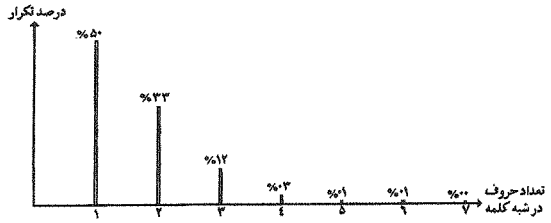
با توجه به نتایج به دست آمده (شکل ۱) حدود ۵۰٪ از شبه کلمات یک متن را شبه کلمات یک حرفی تشکیل می‌دهند. پس اگر سیستم تشخیص حروف فقط قادر به تشخیص شبه کلمات یک حرفی باشد می تواند ۵۰٪ از شبه کلمات متن را بخواند. اگر شبه کلمات ۲ و ۳ حرفی را نیز اضافه کنیم یعنی سیستم تشخیص حروف بتواند شبه کلمات یک ، دو و سه حرفی را تشخیص دهد ۹۵٪ از شبه کلمات یک متن را تشخیص داده است.

اگر برای شبه کلمات ۴، ۵، ۶ و ۷ حرفی فقط شبه کلمات پُراستفاده را انتخاب کنیم دقت تشخیص به ۹۸٪ می‌رسد. برای ۲٪ باقیمانده می توان از روش جدا کردن حروف شبه کلمه استفاده کرد.

جدول (۳)

شمارش شبه کلمات در ترجمه قرآن مجید به زبان فارسی

نوع شبه کلمه	تعداد	میزان تکرار	احتمال وقوع
۱ حرفی	۳۰	۱۴۳۷۰۵	۰/۵۰
۲ حرفی	۳۲۶	۹۵۰۷۸	۰/۳۳
۳ حرفی	۹۶۷	۳۴۵۵۶	۰/۱۲
۴ حرفی	۷۷۱	۸۵۹۳	۰/۰۳
۵ حرفی	۳۷۲	۲۶۴۳	۰/۰۱
۶ حرفی	۱۲۸	۵۴۳	۰/۰۱
۷ حرفی	۲۵	۸۲	۰/۰۰
جمع	۲۶۱۹	۲۸۵۲۰۰	۱/۰۰



شکل (۱)

نمودار تکرار شبه کلمات در زبان فارسی

۴ ـ پردازشهای مقدماتی

برای آماده شدن یک متن اسکن شده جهت تشخیص حروف لازم است که یک سری پردازشهای مقدماتی روی آن اعمال شود. در این بخش این پردازشها تشریح می شود.

۴-۱ ـ خواندن متن اسکن شده

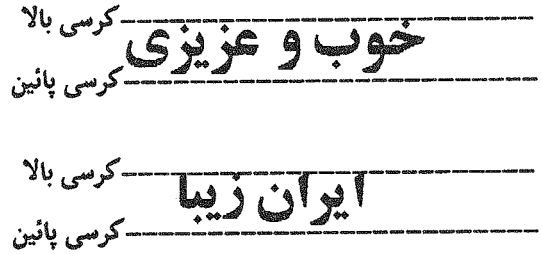
یک متن یا تصویر بعد از اسکن شدن توسط دستگاه اسکنر باید تحت یک فرمت خاص روی دیسک ذخیره شود. برای ذخیره تصاویر، فرمتهای مختلفی وجود دارد، از بین این فرمتها، فرمت TIFF (Tag Image File Format) انتخاب شد. دلیل انتخاب فرمت TIFF قابلیت آن برای ذخیره سازی تصاویر بزرگ بود. مزیت دیگر فرمت TIFF این است که اطلاعات زیادی در مورد تصویر دارد.

فرمت TIFF به دو صورت تصاویر را ذخیره می کند، حالت معمولی و حالت فشرده، که به علت سادگی کار حالت معمولی TIFF انتخاب شد.

۴-۲ ـ جدا کردن خطوط متن

یک متن در زبان فارسی به صورت خط به خط و از راست به چپ نوشته می شود و معمولاً خطوط متوالی با فاصله از هم جدا می شوند. برای جدا کردن خطوط یک متن از نمای افقی (Horizontal Histogram) استفاده شده است. محلهایی که نمای افقی صفر باشد، نشان دهنده فاصله بین خطوط می باشد. البته به علت وجود نویز در تصویر اسکن شده، ممکن است فاصله بین دو خط کاملاً سفید نباشد، برای حل این مشکل می توان با توجه به کیفیت متن حدی برای سفید بودن یک سطر از متن تعیین کرد و اگر نمای افقی آن سطر از آن حد کمتر بود، آن سطر را سفید در نظر گرفت. شکل (۲) نحوه جدا کردن خطوط متن با توجه به فاصله بین خطوط را نشان می دهد. البته در موقع جدا کردن خطوط متن باید توجه

داشت که نقطه حروف جدا نشود، چون بین نقاط هر حرف و خود حرف نیز فاصله خالی وجود دارد. به دو خط فرضی که یک سطر از متن را جدا می کنند کرسی بالا و کرسی پایین می گویند [۵].



شکل (۲)

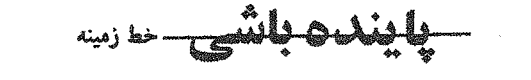
روش جدا کردن خطوط متن

البته در اینجا فرض شده است که خطوط متن، مستقیم و موازی با یکدیگر بوده و به صورت افقی نوشته شده باشند. اگر خطوط متن این شرایط را نداشته باشند باید به روشهای پیچیده تری خطوط متن را جدا کرد.

۳-۴ـ پیدا کردن خط زمینه

در زبان فارسی، خط زمینه اهمیت دارد چون حروف فقط روی خط زمینه به یکدیگر متصل می شوند و به کمک خط زمینه می توان حروف به هم چسبیده را جدا کرد. همچنین از این خط می توان در تشخیص حروف از یکدیگر نیز کمک گرفت مثلاً حرف «د» روی خط زمینه واقع می شود در صورتی که حرف «ر» پایین خط زمینه واقع می شود.

برای پیدا کردن خط زمینه از نمای افقی استفاده شده است. در این روش تعداد نقاط سیاه هر سطر از تصویر شمرده شده و سطری که تعداد نقاط سیاه آن حداکثر باشد به عنوان خط زمینه انتخاب می شود. شکل (۳) خط زمینه را نشان می دهد.



شکل (۳)

خط زمینه

۴-۴ـ جدا کردن شبه کلمات

برای جدا کردن شبه کلمات از فاصله موجود بین شبه کلمات استفاده شده است. در اینجا فرض شده که حروف یک شبه کلمه روی حروفِ شبه کلمه قبلی قرار نگرفته باشند و بین هر دو شبه کلمه حداقل یک ستون خالی وجود داشته باشد. در شکل (۴) روش استفاده شده برای جدا کردن شبه کلمات نشان داده شده است.

شکل (۴)

روش جدا کردن شبه کلمات

۵ـ تشخیص شبه کلمات

بعد از جدا کردن شبه کلمات، مهمترین قسمت کار یعنی تشخیص شبه کلمه انجام می گیرد. عمل تشخیص در دو مرحله انجام می شود در مرحله اول تشخیص مقدماتی شبه کلمه با توجه به تعدادی ویژگی خاص انجام شده و در مرحله دوم تشخیص نهایی با استفاده از روش تطبیق قالبها صورت می گیرد.

۱-۵ـ تشخیص مقدماتی شبه کلمات

در تشخیص حروف با استفاده از شبه کلمات، تعداد شبه کلمات زیاد می باشد (بیش از هزار شبه کلمه) به همین دلیل تطبیق یک شبه کلمه ورودی با تک تک شبه کلمات موجود در سیستم تشخیص حروف وقت گیر می باشد. بنابراین به جای مقایسه شبه کلمه ورودی با تمامی شبه کلمات از روش طبقه بندی چند سطحی استفاده شده است.

در طبقه بندی چند سطحی [۱۰] -(Multilevel Classifi cation)، طبقه بندی با توجه به تعدادی ویژگی و در چندین مرحله انجام می شود. در مرحله اول با توجه به چند ویژگی، کلاسهای مورد نظر به چند گروه تقسیم می شوند. در مراحل بعد این گروه ها با توجه به ویژگیهای دیگری به گروه های کوچکتر تقسیم خواهند شد.

در این طرح اولین ویژگی استفاده شده برای تشخیص شبه کلمات، ابعاد شبه کلمه می باشد. یعنی با توجه به عرض شبه کلمات، آنها به ۱۶ گروه تقسیم شده اند. مثلاً شبه کلماتی که عرض آنها بین یک تا هشت نقطه باشد در گروه

یک قرارداده شده‌اند. البته برای اینکه احتمال طبقه بندی غلط (Misclassification) از بین برود، برای هر شبه کلمه علاوه بر گروهی که شبه کلمه در آن واقع می‌شود، دو گروه همسایه آن گروه نیز در نظر گرفته می‌شوند. مثلاً اگر شبه کلمه ورودی در گروه ۳ واقع شود، در مراحل بعدی با شبه کلمات گروه‌های ۲، ۳ و ۴ مقایسه می‌شود. اگر اندازه حروف استفاده شده در چاپ متن برای تمام حروف یکسان باشد (مثلاً در تایپ کامپیوتری که حروف در یک ماتریس ۱۴×۸ نقطه‌ای چاپ می‌شوند)، طبقه بندی بر اساس ابعاد شبه کلمه باعث جدا شدن شبه کلمات بر اساس تعداد حروف موجود در شبه کلمه می‌شود یعنی شبه کلمات، یک حرفی در یک گروه و شبه کلمات دو حرفی در یک گروه جای می‌گیرند و به این ترتیب شبه کلمات موجود در هر گروه دارای تعداد حروف یکسانی هستند. در مورد متون چاپی معمولی، ابعاد شبه کلمه به طور تقریبی تعداد حروف آن شبه کلمه را مشخص می‌کند. مثلاً شبه کلمات موجود در گروه ۲ معمولاً یک یا دوحرفی می‌باشند.

در مرحله دوم با توجه به تعداد اجزای هر شبه کلمه، آنها را طبقه بندی می‌کنیم.

جدول (۴) نحوه طبقه بندی چند شبه کلمه با توجه به تعداد اجزاء را نشان می‌دهد. این اجزاء شامل بدنه اصلی و نقاط و علائم می‌باشد.

جدول (۴)

طبقه‌بندی شبه کلمات با توجه به تعداد اجزای آنها

تعداداجزا	۱	۲	۳	۴	۵	۶	۷	۸	۹
	ا	ز	ق	ف	توجه	شیر	شش		بیش
مثالهایی	ما	قد	یا	شا	نیز	بیش	چیز		
از هر	م	چه	ست	شد	شب	پیر	ثالث		
گروه	صلا	بر	یم	بند	شتند	شید			

اگر عرض شبه کلمه بیشتر از ۱۲۸ نقطه (تقریباً ۱ سانتیمتر) یا تعداد اجزای آن بیش از ۷ جزء بود، شبه کلمه توسط این روش تشخیص داده نشده و کنار گذاشته می‌شود (Rejection). این شبه کلمات را باید توسط روشهای دیگر مثلاً روش جدا کردن حروف تشخیص داد.

۵-۲ ـ تشخیص نهائی شبه کلمات

بعد از تشخیص مقدماتی، تشخیص نهایی شبه کلمه انجام می‌گیرد. برای انجام عمل تشخیص از روش تطبیق الگوها [۱۱] (template Matching) استفاده شده است. در این روش الگوی ورودی با الگوهای مرجع تطبیق داده شده والگوی ورودی به الگویی که بهترین تطبیق (Best match) را با ورودی داشته باشد، طبقه بندی می‌شود.

برای تشخیص شبه کلمات، شبه کلمه ورودی را با شبه کلمات موجود در سیستم که در مرحله تشخیص مقدماتی مشخص شده‌اند. تطبیق داده و میزان مشابهت ورودی با شبه کلمات مرجع اندازه‌گیری می‌شود. الگوی ورودی به الگوی مرجعی که حداکثر تشابه را با آن دارد نسبت داده می‌شود. در واقع طبقه بندی به روش نزدیکترین همسایه (Nearest Neighbour) صورت می‌گیرد.

روش تطبیق الگوها، یکی از روشهای اولیه [۱۲] برای تشخیص حروف می‌باشد. این روش برای تشخیص اعداد دست نویس استفاده شده [۱۱] و نتایج به دست آمده در مقایسه با سایر روشهای تشخیص اعداد رضایت بخش بوه است [۱۵]. معایب عمده روش تطبیق الگوها عبارتند از:

۱ـ نیاز به حافظه زیاد، چون در این روش لازم است که تمامی الگوهای مرجع در سیستم ذخیره شوند.

۲ـ حساس بودن به اندازه حروف، البته این مشکل را می‌توان با مقیاس کردن شکلهای ورودی حل کرد [۱۱].

از مزایای این روش، حذف مرحله استخراج ویژگی و سادگی طراحی می‌باشد. همچنین یادگیری در این روش ساده می‌باشد. یعنی می‌توان الگوهای جدیدی را به راحتی به سیستم اضافه کرده یا الگوهایی که متناسب نیستند را حذف کرد.

در روشهای دیگر تشخیص حروف ابتدا یکسری ویژگی استخراج شده و سپس تشخیص با توجه به آن ویژگیها انجام می‌شود. مثلاً در مرجع [۴] از گشتاورهای زرنیکی برای تشخیص اعداد فارسی استفاده شده است. در آن مرجع از شبکه عصبی برای طبقه بندی اعداد استفاده شده است. مشکل روش فوق این است که به تعداد زیادی نمونه از هر کلاس برای یادگیری شبکه عصبی نیاز بوده و به علت زیاد بودن تعداد کلاسها، یادگیری شبکه فوق العاده طولانی است.

در مقاله حاضر چون تشخیص متن با فونت ثابت، مورد نظر بوده است مشکل حساس بودن روش تطبیق الگوها

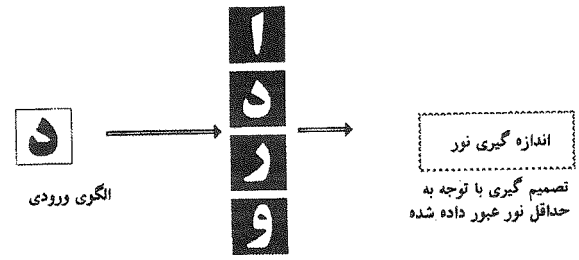
به اندازه حروف وجود ندارد. همچنین از هرشبه کلمه فقط یک شکل به عنوان الگوی مرجع در سیستم نگهداری شده است. برای تطبیق شبه کلمات از معیار فاصله ژاکارد (Jaccard) استفاده شده است [۱۱].

معیار ژاکارد به صورت زیر محاسبه می شود.

$$J=\frac{N11}{N11+N10+N01}$$

که در این رابطه Nij نشان دهنده تعداد نقاطی است که در الگوی ورودی مقدار i داشته و در الگوی مرجع مقدار j داشته اند. این معیار، برای هر الگو مقداری بین صفر تا یک دارد. برای بازشناسی الگوی ورودی، ابتدا نزدیکترین همسایه به الگوی ورودی یعنی الگوی مرجعی که معیار ژاکارد آن از بقیه الگوها بیشتر است انتخاب می شود. اگر این معیار از حدی کمتر بود، الگوی ورودی کنار گذاشته می شود (Rejection).

روش تطبیق الگوها معادل این است که شکل الگوهای مرجع را روی صفحه شفافی به صورت منفی چاپ کرده و این صفحه را روی الگوی ورودی قرار داده و میزان نوری که از تصویر ورودی منطبق شده با تصویر مرجع عبور می کند اندازه گیری شده و الگویی که حداقل نور را داشته باشد، به عنوان شکل الگوی ورودی تشخیص داده شود. در شکل (۵) این روش نشان داده شده است. [۱۲]



تصویر منفی الگوهای موجود در سیستم

شکل (۵)

روش تطبیق الگوها

۵-۳ - سرعت و دقت در تشخیص حروف

دو عامل عمده باعث شده اند که در تشخیص متن از روشهای پیشنهادی تشخیص حروف استفاده نشود. این دو عامل دقت کم و سرعت پایین روشهای تشخیص حروف می باشند. منظور از دقت در تشخیص این است که سیستم چند در صد حروف را به درستی تشخیص می دهد. مثلاً اگر سیستم از هر ۱۰۰ حرف، ۹۸ حرف را درست تشخیص دهد و دو حرف راتشخیص ندهد یا اینکه اشتباه تشخیص دهد، می گویند دقت سیستم ۹۸٪ است. مشکل تشخیص حروف این است که به دقت خیلی زیادی نیاز دارد، حتی دقت ۹۹٪ نیز قابل قبول نیست چون اگر سیستم فقط یک در صد از حروف را اشتباه تشخیص دهد، تعداد اشتباهات در یک صفحه از متن (اگر متن را ۲۴ سطری و هر سطر را ۷۵ حرفی در نظر بگیریم) ۱۸ حرف می باشد، که این تعداد اشتباه در یک صفحه خیلی زیاد است. البته دقت اکثر سیستمهای تشخیص حروف از ۹۵٪ نیز کمتر است. در مورد زبان عربی دقتهای گزارش شده از ۹۰٪ [۹] تا ۹۸٪ [۱۳] می باشد. در مورد زبان فارسی دقت گزارش شده برای جدا کردن حروف یک کلمه ۹۹٪ [۲] می باشد.

مسأله دیگر در تشخیص حروف سرعت سیستم می باشد، یعنی سیستم در هر دقیقه قادر به تشخیص چند حرف می باشد. سرعت خواندن یک انسان که در خواندن ضعیف بوده و مجبور به هجا کردن حروف در هنگام خواندن باشد حداقل ۱۵۰ کلمه در دقیقه می باشد (البته افراد ماهر قادرند تا ۱۰۰۰ کلمه را در یک دقیقه بخوانند.) [۳]، پس سیستم تشخیص حروف باید لااقل سرعتی حدود ۱۵۰ کلمه در دقیقه داشته باشد. اگر به طور متوسط کلمات را ۴ حرفی فرض کنیم سرعت مورد نظر ۶۰۰ حرف در دقیقه یا ۱۰ حرف در ثانیه می باشد. سرعتهای گزارش شده در مورد تشخیص حروف چاپی عربی ۳۰ کلمه در دقیقه [۸] و ۴ حرف در ثانیه [۱۳] می باشد.

البته اگر زمان مورد نیاز برای وارد کردن متن به سیستم (مثلاً اسکن کردن متن) و زمان مورد نیاز برای ذخیره کردن یا چاپ نتایج به دست آمده نیز در نظر گرفته شوند سرعت این روشها به مراتب کمتر می باشد. مزیت عمده تشخیص حروف با استفاده از تشخیص شبه کلمات به جای شکستن کلمات به حروف، بالا رفتن سرعت تشخیص می باشد چون در زمان مورد نیاز برای شکستن کلمات و سپس ترکیب حروف تشخیص داده شده صرفه جویی می شود.

۶- نتایج عملی

برنامه های مورد نیاز برای تشخیص حروف به روش تشخیص شبه کلمات به زبان c نوشته شده و روی یک کامپیوتر PC مدل ۴۸۶ اجرا شد، سپس برنامه های نوشته

شده روی قسمتی از کتاب داستان راستان [۶] امتحان شد. شکل (۶) یکی از متنهای استفاده شده را نشان می‌دهد. دقت به دست آمده ۹۶٪ می باشد. یعنی این سیستم توانست ۹۶٪ از شبه کلمات را به درستی تشخیص دهد. از ۴٪ شبه کلمات باقیمانده ۲٪ اشتباه تشخیص داده شدند (Misclassification) و ۲٪ دیگر تشخیص داده نشدند(Rejection). علت اکثر اشتباهات در تشخیص نقطه های حروف می باشد، مثلاً «یا» و «با» با یکدیگر اشتباه می شوند. دلیل عدم تشخیص بعضی از شبه کلمات وجود اعراب روی آنها بوده است. باید توجه داشت که دقت به دست آمده در این تحقیق برای شبه کلمات بوده در صورتی که در نتایج قبلی [۹] و [۱۳] دقت گزارش شده مربوط به حروف می باشد. بدیهی است از آنجا که غلط بودن یک حرف در یک شبه کلمه باعث غلط شدن کل آن شبه کلمه می گردد، لذا دقت به دست آمده در این تحقیق از دقتهای گزارش شده قبلی بیشتر است.

سرعت به دست آمده برای تشخیص، ۳ شبه کلمه در ثانیه می باشد که با در نظر گرفتن تعداد متوسط ۱/۷۵ حرف برای هر شبه کلمه حدود ۵ حرف در ثانیه می باشد که سرعت مناسبی برای تشخیص حروف است.

عرق کار

امام کاظم ، در زمینی که متعلق به شخص خودش بود ، مشغول کار و اصلاح زمین بود. فعالیت زیاد، عرق امام را از تمام بدنش جاری ساخته بود. علی بن ابی حمزهٔ بطائنی، در این وقت رسید و عرض کرد:

« قربانت گردم، چرا این کار را به عهدهٔ دیگران نمی‌گذاری؟ »

– «چرا به عهدهٔ دیگران بگذارم ؟ افراد از من بهتر همواره از این کارها می‌کرده‌اند. »

– «مثلا چه کسانی ؟ »

– «رسول خدا و امیرالمؤمنین و همهٔ پدران و اجدادم. اساساً کار و فعالیت در زمین از سنن پیغمبران و اوصیای پیغمبران و بندگان شایستهٔ خداوند است. »

شکل (۶)

نمونه‌ای از متن استفاده شده برای آزمایش سیستم [۶]

۷ـ نتیجه گیری:

در این مقاله تشخیص متون چاپی فارسی با استفاده از شبه کلمات انجام شده است. همچنین شبه کلمات از نظر آماری بررسی شده‌اند. طبقه بندی شبه کلمات در دو قسمت انجام شده است. در قسمت اول با استفاده از اندازه و تعداد اجزای هر شبه کلمه تشخیص مقدماتی داده شده و در قسمت بعد تشخیص نهایی به روش تطبیق الگوها و با استفاده از معیار تشابه ژاکارد انجام شده است.

در این مقاله نشان داده شد که استفاده از روش ارائه شده برای تشخیص متون فارسی با فونت ثابت مناسب است. در حالتی که متن شامل چندین فونت مختلف باشد یا در مورد متون دست نویس لازم است که قبل از تشخیص، شبه کلمات را مقیاس کرد.

منابع:

[۱] آیتی، عبدالمحمد-ترجمه قرآن مجید، انتشارات سروش، تهران ۱۳۷۱.

[۲] احمدزاده، محمدرضا-کبیر، احسان اله، «شکستن کلمات تایپ شده فارسی به حروف» گزارش اولین کنفرانس بین المللی کامپیوتر در علوم، فنون و پزشکی در ایران، دانشگاه اصفهان، ۵ تا ۷ دیماه ۱۳۷۰، صفحات ۱ الی ۶.

[۳] باطنی، محمدرضا-مسائل زبانشناسی نوین، انتشارات آگاه، تهران، چاپ سوم ۱۳۷۰.

[۴] ختن زاد، علیرضا-شیرعلی شهرضا، محمدحسن، «تشخیص اعداد چاپی فارسی مستقل از اندازه و جابجایی با استفاده از گشتاورهای زرنیکی و به کمک شبکه های عصبی»، مجموعه مقالات کنفرانس مهندسی برق ایران، تهران-دانشگاه تربیت مدرس، ۲۴ الی ۳۰ اردیبهشت ۱۳۷۳، جلد ۵، صفحات ۴۱۷ الی ۴۲۴.

[۵] فضائلی، حبیب اله-تعلیم خط، انتشارات سروش، چاپ ششم، ۱۳۷۰.

[۶] مطهری، مرتضی-داستان راستان، جلد اول، چاپ پانزدهم، انتشارات صدرا-۱۳۷۰.

[۷] نعیمی، محمود جعفری نژاد-«خودآموز دستگاه بریل برای معلولین بصری»، گزارش سمینار طراحی و کاربرد ریزپردازنده ها، دانشگاه صنعتی امیرکبیر، تهران، دوم الی چهارم خرداد ۱۳۶۸، صفحات ۱ الی ۱۴.

[8] H. Y. Abdelazim and M. A. Hashish, «Arabic typeset: an OCR approach," In Proceedings of EUSIPCO-90, Fifth European Signal Processing Conference, Vol. 2, pp.1019-1022, 1990.

[9] A. Amin and J. F. Mari, "Machine recognition and correction of printed Arabic text," IEEE Trans. Syst. Man Cybern., vol. 19, no. 5, pp. 1300-1305, 1989.

[10] K. S. Fu, "Applications of pattern recognition, to remote sensing," In K. S. Fu, Editor, Applications of Pattern recognition, CRC Press, Florida, pp. 65-105, 1982.

[11] P. Gader, et al., "Recognition of handwritten digits using template and model matching," Pattern Recognition, vol. 24, no. 5, pp. 421-431, 1991.

[12] A. W. Holt, "Comparative religion in character recognition machines," IEEE Comput. Group News. vol. 2, pp. 3-11, Nov. 1968.

[13] V. Margner, "SARAT-A system for the recognition of Arabic printed text," Proceedings of 11th IAPR International Conference on Pattern Recognition, vol. II, pp. 561-564, 1992.

[14] B. Parhami and M. Taraghi," Automatic recognition of printed Farsi texts," Pattern Recognition, vol. 14, pp. 395-403, 1981.

[15] R. A. Wilkinson, et al., The first optical character recognition systems conference, Technical Report NISTIR 4912, National Institute of Standards and Technology, August 1992.