تفکیک انسان از ماشین به کمک خط نستعلیق (Nastaliq CAPTCHA)

محمدحسن شیرعلی شهرضا و محمد شیرعلی شهرضا

چکیده: امروزه بسیاری از فعالیت‌های روزمره انسان‌ها همچون آموزش، تجارت، گفتگو و غیره از طریق اینترنت صورت می‌گیرد. در مواردی همچون ثبت‌نام در پایگاه‌های اینترنتی، بعضی افراد خرابکار با نوشتن نرم‌افزارهای رایانه‌ای، اقدام به ثبت‌نام جعلی در این پایگاه‌ها می‌کنند که موجب هدررفتن منابع پایگاه می‌شود. لذا باید به تفکیک کاربران انسانی از نرم‌افزارهای رایانه‌ای پرداخت. بدین منظور در این مقاله روشی برای تفکیک کاربران فارسی و عربی زبان از نرم‌افزارهای رایانه‌ای بر پایه خط نستعلیق ارائه شده است. در این روش تصویری از یک کلمه فارسی یا عربی که با خط نستعلیق نوشته شده انتخاب گردیده، برای کاربر نمایش داده شده و از وی خواسته می‌شود تا آن کلمه را تایپ کند. با توجه به عدم توانایی شناسایی این کلمات توسط نرم‌افزارهای تشخیص حروف فارسی و عربی موجود، امکان تشخیص این کلمات تنها توسط یک کاربر فارسی یا عربی زبان امکان‌پذیر می‌باشد. روش پیشنهادی توسط زبان جاوا پیاده‌سازی شده است.

کلید واژه: **آزمون تورینگ، امنیت شبکه، تفکیک کاربران انسانی از نرم‌افزارهای رایانه‌ای (CAPTCHA)، سیستم‌های تشخیص حروف (OCR).**

۱- مقدمه

با گسترش شبکه جهانی اینترنت، بسیاری از زوایای زندگی انسان نیز تحت تأثیر این پدیده قرار گرفته است. به طوری که در کشورهای صنعتی بسیاری از امور روزمره، از خریدهای روزانه گرفته تا آموزش و تجارت، همگی از طریق اینترنت صورت می‌گیرد.

یکی از موارد معمول در اکثر پایگاه‌ها، علی‌الخصوص پایگاه‌های تجاری و اداری، پرکردن فرم‌های ثبت‌نام به منظور انجام امور مورد نظر می‌باشد. در این فرم‌ها افراد با واردکردن اطلاعات مورد نیاز، اجازه برقراری ارتباط با پایگاه اینترنتی و انجام امور مورد نظر را پیدا می‌کنند.

اما متأسفانه امروزه برخی افراد قانون‌شکن و خرابکار، اقدام به نوشتن نرم‌افزارهای رایانه‌ای برای ثبت‌نام جعلی در این پایگاه‌ها می‌کنند. این نرم‌افزارها به طور خودکار وارد پایگاه اینترنتی شده و اقدام به پرکردن یک فرم با اطلاعات غیرصحیح و ثبت‌نام در پایگاه اینترنتی می‌کنند. به این ترتیب حجم زیادی از منابع این پایگاه‌ها به هدر رفته و در اختیار نویسندگان سودجوی آن نرم‌افزارها قرار می‌گیرد و یا اینکه کارآیی سیستم آن پایگاه کاهش می‌یابد که به این حملات، از کار انداختن خدمات[1] (DOS) می‌گویند.

به منظور جلوگیری از این گونه حملات، روش‌های مختلفی ارائه گردیده است. در این روش‌ها هدف، تفکیک کاربران انسانی از نرم‌افزارهای رایانه‌ای است. ویژگی اصلی این روش‌ها بایستی خودکاربودن آنها باشد، بگونه‌ای که بتوان به تنهایی به وسیله رایانه اقدام به پیاده‌سازی آنها نمود. زیرا بررسی حجم زیادی از ثبت‌نام‌ها در پایگاه‌های اینترنتی توسط نیروهای انسانی نیازمند زمان و هزینه زیادی می‌باشد؛ بگونه‌ای که در بعضی موارد مانند پایگاه‌های ارائه پست الکترونیک استفاده از نیروی انسانی به منظور بررسی فرم‌های ثبت‌نام عملاً غیرممکن می‌باشد. لذا استفاده از سیستم‌های خودکار تفکیک کاربران از نرم‌افزارهای رایانه‌ای ضروری می‌باشد.

در مباحث هوش مصنوعی برای اثبات هوشمندی یک رایانه از آزمونی به نام تست تورینگ نام برده می‌شود. در این آزمون یک کاربر انسانی و یک رایانه در دو اتاق جداگانه قرار گرفته و یک ناظر انسانی در اتاقی دیگر از آنها سؤال می‌کند. چنانچه ناظر نتواند تشخیص دهد که در کدام اتاق رایانه و در کدام اتاق انسان قرار دارد، می‌گویند که رایانه در تست تورینگ موفق بوده است.

در تفکیک کاربران انسانی از نرم‌افزارهای رایانه‌ای باید روشی مشابه با تست تورینگ را اعمال نمود، با این تفاوت که در اینجا به جای ناظر انسانی، ناظر یک رایانه می‌باشد و رایانه بایستی با سؤالاتی که انجام می‌دهد، بتواند کاربر انسانی را از نرم‌افزار رایانه‌ای تفکیک کند. لذا تمرکز اصلی در این روشی بر روی سؤالاتی است که انسان به راحتی بتواند به آنها پاسخ گوید، ولی پاسخ‌دادن به آنها با برنامه‌های فعلی رایانه‌ای به سختی امکان‌پذیر باشد.

از روش‌های به کار رفته برای تفکیک کاربران انسانی از نرم‌افزارهای رایانه‌ای، استفاده از تصاویر کلمات می‌باشد. این روش بر پایه نقاط ضعف سیستم‌های تشخیص حروف[2] (OCR) استوار است.

سیستم‌های تشخیص حروف برای خواندن خودکار متون استفاده می‌شوند، ولی اینگونه نرم‌افزارها در خواندن متون چاپ‌شده با کیفیت پایین یا متون دست‌نویس با مشکل مواجه شده و فقط قادر به تشخیص متون تایپی که از کیفیت بالایی برخوردار بوده و از قالب‌های رایج و استاندارد پیروی می‌کنند، می‌باشند. پس می‌توان از این نقص نرم‌افزارهای تشخیص حروف استفاده کرده و تصویر یک کلمه را به شکلی تغییر داد که توسط انسان قابل خواندن باشد، ولی نرم‌افزارهای تشخیص حروف قادر به بازیابی و تشخیص آن نباشند. در مورد روش‌های به کار رفته بدین منظور در بخش ۲ به طور مفصل بحث خواهد شد.

این روش‌ها هم‌اکنون در پایگاه‌های اینترنتی بزرگ مانند یاهو و هات‌میل به منظور ثبت‌نام کاربران استفاده می‌شود. البته در کنار این امر، در سال‌های اخیر روش‌هایی برای شکستن این روش‌ها و تشخیص خودکار این گونه تصاویر کلمات تغیریافته [۱] و [۲] پیشنهاد شده است.

این مقاله در تاریخ ۲۰ اسفند ماه ۱۳۸۴ دریافت و در تاریخ ۶ آبان ماه ۱۳۸۵ بازنگری شد.

محمدحسن شیرعلی شهرضا، دانشکده مهندسی کامپیوتر، دانشگاه یزد، خیابان پژوهش، صفائیه، یزد، ایران (email: hshirali@yazduni.ac.ir).

محمد شیرعلی شهرضا، دانشکده علوم ریاضی و کامپیوتر، دانشگاه صنعتی شریف، خیابان آزادی، تهران، ایران (email: shirali@cs.sharif.edu).

1. Denial of Services

2. Optical Character Recognition

شکل ۱: نمونه‌ای از حروف مورد استفاده در پایگاه اینترنتی آلتاویستا به منظور تفکیک کاربران انسانی از نرم‌افزارهای رایانه‌ای.

در این مقاله با توجه به ویژگی‌های خاص زبان‌های فارسی و عربی، روشی جدید به منظور تفکیک کاربران فارسی و عربی‌زبان از نرم‌افزارهای رایانه‌ای با استفاده از تصویر کلمات فارسی و عربی ارائه شده است. در این روش تصویر یک کلمه فارسی یا عربی نوشته شده با خط نستعلیق در قالب یک پرونده تصویری ذخیره شده و بدون اعمال تغییراتی بر روی آن، به کاربر نشان داده شده و از او خواسته می‌شود تا کلمه نشان داده شده را تایپ کند. با توجه به عدم امکان شناسایی این کلمات توسط نرم‌افزارهای تشخیص حروف فارسی یا عربی موجود، کار تشخیص این کلمات فقط توسط یک کاربر فارسی یا عربی زبان امکان‌پذیر است.

ساختار این مقاله به صورت زیر می‌باشد: در بخش ۲ به مرور کارهای انجام‌شده در این زمینه پرداخته می‌شود. در بخش ۳ ویژگی‌های خاص زبان‌های فارسی و عربی و نیز ویژگی‌های خط نستعلیق که باعث مشکل‌شدن عملیات تشخیص حروف فارسی و عربی می‌شود، بیان می‌گردد. در فصل ۴ روش پیشنهادی ارائه‌شده و در فصل ۵ نتایج عملی پیاده‌سازی این روش گفته می‌شود. در فصل ۶ نیز نتیجه‌گیری نهایی انجام می‌شود.

۲- پیشینه کار

اولین بار در سال ۱۹۹۷ میلادی توسط آندری برودر[1] و همکارانش روشی برای تفکیک کاربران انسانی از نرم‌افزارهای رایانه‌ای ابداع شد و در همان سال نیز برای اولین بار پایگاه اینترنتی آلتاویستا از این روش استفاده کرد. در این روش تصویر یک کلمه انگلیسی همراه با اعوجاج به کاربر نمایش داده شده و کاربر بایستی آن کلمه را تایپ کند (شکل ۱). این اعوجاج به شکلی انجام می‌شود که سیستم‌های تشخیص حروف قادر به خواندن آن کلمات نبوده ولی انسان قادر به خواندن آنها باشد [۳].

این روش‌ها بعداً به CAPTCHA[2]، معروف گشته و امروزه توسط اکثر پایگاه‌های معروف مانند یاهو و مایکروسافت استفاده می‌شوند. در این بخش به بررسی کامل‌تر این روش‌ها می‌پردازیم.

۲-۱ روش اعوجاج

روش اعوجاج[3] در دانشگاه کارنگی‌ملون برای تفکیک کاربران انسانی از نرم‌افزارهای رایانه‌ای تهیه شده است [۴]. در این روش یک کلمه را از یک فرهنگ لغت انتخاب کرده و پس از اعمال تغییراتی همچون اضافه‌کردن خطوط سیاه و خطوط سفید، ایجاد تغییرات غیرخطی و غیره آن را در قالب یک تصویر نشان داده و کاربر بایستی آن کلمه را به طور صحیح تایپ کند. چون این روش کلمات خود را از یک فرهنگ لغت ۸۵۰ کلمه‌ای انتخاب می‌کند، لذا در مقابل حملات می‌تواند به آسانی شکسته شود [۱].

پایگاه اینترنتی یاهو از نسخه ساده‌ای از این روش به نام "اعوجاج ساده" برای تفکیک کاربران انسانی از نرم‌افزارهای رایانه‌ای به منظور

1. Anderi Broder
2. Completely Automated Public Turing Test to Tell Computers and Human Apart
3. Gimpy

شکل ۲: نمونه‌ای از حروف مورد استفاده در پایگاه اینترنتی PayPal به منظور تفکیک کاربران انسانی از نرم‌افزارهای رایانه‌ای.

جلوگیری از تعریف حساب‌های کاربری پی‌درپی توسط نرم‌افزارهای رایانه‌ای مخرب بهره می‌برد.

۲-۲ روش کلمات بی‌معنا

در روش کلمات بی‌معنا[4] واژه‌هایی که در فرهنگ لغت انگلیسی نباشند، تولید شده سپس تصویر این کلمات دستکاری می‌شوند [۵]. این شکل‌های تولیدشده می‌توانند درجات مختلفی از آسانی و سختی داشته باشند. در این روش گرچه می‌توان شکل‌هایی با درجه سختی زیاد تولید کرد، ولی در این موارد کاربران انسانی نیز به سختی می‌توانند کلمات ارائه‌شده را تشخیص دهند.

۲-۳ استفاده از کلمات دست‌نویس

روش دیگر، استفاده از کلمات دست‌نویس می‌باشد [۶]. در این روش بانکی از کلمات دست‌نویس از نام شهرهای آمریکا (که از نامه‌های پست‌شده توسط مردم استخراج شده است) تهیه شده است. به منظور تفکیک کاربران انسانی از نرم‌افزارهای رایانه‌ای، تصویر کلمه یکی از شهرها انتخاب شده و به کاربر نشان داده می‌شود و کاربر بایستی نام آن شهر را به طور صحیح تایپ نماید. در این روش تصویر کلمات دارای کیفیت مناسبی نبوده و خواندن بعضی از آنها حتی برای کاربران انسانی نیز مشکل می‌باشد.

۲-۴ روش پایگاه اینترنتی PayPal

پایگاه اینترنتی PayPal به ارائه خدماتی در زمینه‌های پرداخت الکترونیکی پول می‌پردازد. این پایگاه نیز برای تفکیک کاربران انسانی از نرم‌افزارهای رایانه‌ای از روش نمایش یک کلمه با اعوجاج، مانند شکل ۲ استفاده می‌کند [۷].

متأسفانه پایگاه PayPal هیچ توضیحی در مورد جزئیات این روش منتشر نکرده است ولی با توجه به فاصله زیاد حروف به نظر می‌رسد که جداکردن این حروف توسط سیستم‌های تشخیص حروف ساده باشد.

۲-۵ استفاده از الگوهای بصری پویا

در این روش، کلمات در زمینه‌ای از الگوهای بصری چاپ می‌شوند [۸]. مثلاً متن در زمینه‌ای از دایره‌های سیاه چاپ می‌شود. سپس این تصویر به منظور تفکیک کاربران انسانی از نرم‌افزارهای رایانه‌ای نمایش داده می‌شود. علی‌رغم مشکل‌بودن تشخیص این کلمات برای نرم‌افزارهای رایانه‌ای، خواندن این شکل‌ها توسط کاربران انسانی نیز مشکل می‌باشد.

۲-۶ روش پایگاه اینترنتی هات‌میل

در ثبت‌نام سرویس پست الکترونیک هات‌میل (Hotmail) مربوط به شرکت مایکروسافت [۹]، یک رشته از حروف انگلیسی به طور تصادفی انتخاب شده و پس از دستکاری شکل آن حروف؛ تصاویر حروف به کاربر نشان داده می‌شوند و کاربر بایستی این حروف را به درستی تایپ کند.

4. Baffletext

شکل ۳: نمونه‌ای از حروف مورد استفاده در پایگاه اینترنتی هات‌میل به منظور تفکیک کاربران انسانی از نرم‌افزارهای رایانه‌ای.

در این روش از تحقیقات مربوط به تشخیص حروف استفاده شده و با توجه به اینکه در سیستم‌های تشخیص حروف، جداسازی حروف[1] از تشخیص حروف[2] مشکل‌تر می‌باشد، لذا سعی شده تا حروف به شکلی تغییر داده شوند که جداکردن آنها از یکدیگر به آسانی امکان‌پذیر نباشد. در شکل شماره ۳ مشاهده می‌شود که با قراردادن کمان‌هایی سعی شده است تا حد امکان جداسازی کلمات مشکل گردد [۱۰]. در نتیجه اگرچه جداسازی این حروف برای انسان‌ها ساده می‌باشد، ولی جداسازی آنها توسط نرم‌افزارهای موجود امکان‌پذیر نیست.

در این روش به علت قراردادن کمان‌هایی میان حروف، گاهی بعضی حروف به صورت دیگری خوانده شده و گاهی نیز حروف اضافه‌ای ایجاد می‌شوند.

۲-۷ روش مثله‌کردن حروف

در روش مثله‌کردن حروف[3] نیز مانند روش ۲-۶، تکیه اصلی بر روی جداسازی حروف می‌باشد. یعنی سعی می‌گردد حروف به گونه‌ای تغییر داده شوند که امکان جداسازی آنها نباشد [۱۱]. بدین منظور هر یک از حروف تکه‌تکه شده، سپس این تکه‌ها جابجا می‌شوند، این عملیات باعث می‌شود که جداسازی حروف با روش‌های موجود تشخیص حروف مشکل باشد زیرا در این روش هر حرف به تعداد زیادی تکه‌های ریز شکسته شده است. از طرف دیگر حروف به صورت تصادفی انتخاب می‌شوند تا نتوان از فرهنگ لغت برای پیش‌بینی کلمات استفاده کرد.

۲-۸ روش چاپ با کیفیت بد

چاپ با کیفیت بد[4] بر پایه یکی از ضعف‌های عمده سیستم‌های تشخیص حروف امروزی، یعنی عدم توانایی نرم‌افزارهای تشخیص حروف فعلی در خواندن متون چاپی با کیفیت پایین، استوار می‌باشد [۱۰]. بنابراین سعی شده است تا با پایین‌آوردن کیفیت حروف چاپ‌شده به طور مصنوعی، از فعالیت‌های نرم‌افزارهای رایانه‌ای مخرب جلوگیری به عمل آید. ولی این روش چندان در مقابل حملات مقاوم نبوده و ممکن است با اعمال معکوس تغییرات انجام‌شده، کلمات تغییریافته به حالت اولیه تبدیل شده و توسط سیستم‌های تشخیص حروف تشخیص داده شوند [۱].

البته روش‌های دیگری نیز برای تفکیک کاربران انسانی از نرم‌افزارهای رایانه‌ای وجود دارند که از تشخیص حروف استفاده نمی‌کنند، از این روش‌ها می‌توان روش‌های تفکیک ضمنی [۱۲]، استفاده از تصاویر [۴]، تبدیل متن به صدا [۱۳] و [۷] و رسم [۱۴] را نام برد.

به طور خلاصه می‌توان گفت که روش‌هایی که امروزه برای تفکیک کاربران انسانی از نرم‌افزارهای رایانه‌ای استفاده می‌شوند، معمولاً برای کاربران انسانی راحت نبوده و اکثر افراد با اکراه از این روش‌ها استفاده می‌کنند [۱۲].

۳- ویژگی‌های زبان‌های فارسی و عربی از نظر تشخیص حروف

برای استفاده از متون فارسی برای تفکیک کاربران فارسی و عربی زبان لازم است که ویژگی‌های زبان‌های فارسی و عربی را به خوبی شناخته و طراحی سیستم را متناسب با این ویژگی‌ها انجام داد. ویژگی‌های این زبان‌ها از نقطه نظر تشخیص حروف در این بخش بیان می‌شود [۱۵]. تفاوت اصلی زبان فارسی از زبان عربی وجود چهار حرف "گ، چ، پ، ژ" در زبان فارسی است.

در پایان این بخش نیز به بررسی خط نستعلیق می‌پردازیم.

۳-۱ نوشتن از راست به چپ

در زبان‌های فارسی و عربی نوشتن از راست به چپ انجام می‌شود، برخلاف زبان انگلیسی که در آن از چپ به راست و یا در زبان‌های دیگری که از بالا به پایین می‌نویسند. بنابراین تشخیص حروف فارسی نیز باید از راست به چپ انجام شود.

۳-۲ متصل‌بودن حروف

در زبان‌های فارسی و عربی، حروف هنگام نوشتن به یکدیگر متصل می‌شوند برخلاف زبان انگلیسی که در آن حروف به صورت مجزا نوشته می‌شوند. البته حروف دست‌نویس انگلیسی نیز به صورت متصل نوشته می‌شوند.

۳-۳ نقطه‌داربودن بعضی حروف

در زبان‌های فارسی و عربی، نقطه اهمیت زیادی داشته و نیمی از حروف الفبای آنها نقطه دارند. اهمیت نقطه از این نظر است که تعدادی از حروف فقط در تعداد یا محل نقاط با یکدیگر اختلاف دارند. در فارسی و عربی حروف می‌توانند بدون نقطه بوده یا یک تا سه نقطه داشته باشند 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 نقطه‌ها به یکدیگر می‌چسبند گاهی تشخیص بین دو نقطه یا سه نقطه امکان‌پذیر نیست.

۳-٤ علائم خاص

در زبان‌های فارسی و عربی از علائم "تشدید"، "تنوین"، "همزه" و "مد" نیز استفاده می‌شود. اگرچه بعضی از این علائم مختص زبان عربی می‌باشد ولی کمتر متن فارسی را می‌توان پیدا کرد که در آن این علائم موجود نباشد. این علائم روی حروف قرار می‌گیرند. مثال‌هایی برای این علائم، تشدید در کلمه "امّا"، تنوین در کلمه "حتماً"، مد در کلمه "قرآن" و همزه در کلمه "سؤال" می‌باشند. علامت دیگر، علامت "یای کوتاه‌شده" یا "یای میانجی" (ی) می‌باشد که روی حرف "ه" قرار می‌گیرد مثلاً در عبارت "خانهٔ دوست".

۳-۵ کشیدگی کلمات

در هنگام تایپ متون فارسی برای اینکه انتهای جملات در یک ستون قرار گیرند در کلمات آخر جمله از علامت "-" برای کشیده نوشتن کلمه استفاده می‌شود، که این علامت هیچ معنای خاصی نداشته و فقط برای زیبایی متن استفاده می‌شود. مانند کلمه "باشد" که به صورت "بـــــــــاشد" نیز نوشته می‌شود.

1. Segmentation
2. Recognition
3. Scatter Type
4. Pessimal Print

شکل ۴: نمونه‌ای از حروف مورد استفاده در روش پیشنهادی به منظور تفکیک کاربران انسانی از نرم‌افزارهای رایانه‌ای.

۳-۶ شکل‌های مختلف یک حرف

در زبان‌های فارسی و عربی یک حرف می‌تواند تا چهار شکل مختلف داشته باشد. شکل هر حرف متناسب با محل قرارگرفتن آن حرف در کلمه می‌باشد. مثلاً حرف "عین" در اول کلمه به صورت "عـ"، در وسط کلمه به صورت "ـعـ"، در آخر کلمه به صورت "ـع" و در حالت منفرد به صورت "ع" نوشته می‌شود.

۳-۷ اِعراب

در زبان‌های فارسی و عربی صداها موقع نوشتن کلمه نوشته نمی‌شوند ولی در محل‌هایی که احتمال اشتباه در موقع خواندن وجود دارد لازم است که برای بعضی از حروف، اعراب گذاشته شود. اعراب‌هایی که در زبان‌های فارسی و عربی استفاده می‌شوند مشابه زبان عربی بوده و عبارتند از فتحه (ـَـ)، ضمه (ـُـ) و کسره (ـِـ). اعراب در بالا و پایین حروف گذاشته شده و خواندن صحیح آن را امکان‌پذیر می‌کند. جداکردن اعراب از خود حروف در موقع تشخیص حروف کار آسانی نیست.

۳-۸ هم‌پوشانی حروف

در موقع حروف‌چینی یا تایپ متون فارسی و عربی بعضی از حروف روی حرف قبلی قرار می‌گیرند، مثلاً در هنگام چاپ کلمه "را" حرف "الف" روی حرف "ر" قرار می‌گیرد، این مسئله نیز کار تشخیص حروف را سخت‌تر می‌کند.

۳-۹ شکل خاص "لا"

در زبان‌های فارسی و عربی به خاطر زیبایی متن هرگاه حروف "لام" و "الف" پشت سر هم قرار گیرند به صورت "لا" نوشته می‌شوند، این مسئله در هنگام مرتب‌کردن یا جستجوی یک متن فارسی و عربی توسط کامپیوتر، مشکل ایجاد می‌کند. در تشخیص حروف می‌توان شکل "لا" را به عنوان یک حرف مستقل در نظر گرفت.

۳-۱۰ اندازه متفاوت حروف

در زبان‌های فارسی و عربی حروف از نظر اندازه یکسان نیستند، مثلاً حرف "ب" هنگام چاپ جای بیشتری از حرف "د" اشغال می‌کند. یکسان‌نبودن اندازه حروف به پیچیدگی تشخیص حروف فارسی و عربی کمک می‌کند.

۳-۱۱ نبودن فاصله بین کلمات

در زبان‌های فارسی و عربی بر خلاف زبان انگلیسی که هر کلمه با کلمه بعد از آن به وسیله فاصله جدا می‌شود، بین کلمات فاصله وجود ندارد. به همین علت جداکردن کلمات بدون توجه به تمامی جمله امکان‌پذیر نیست. در تشخیص حروف، نبودن فاصله بین کلمات باعث می‌شود که عمل تصحیح متن تشخیص داده شده با استفاده از فرهنگ لغت، مشکل شود.

۳-۱۲ خط نستعلیق

خط نستعلیق [۱۶] یکی از مشهورترین خطوط در زبان فارسی بوده و توسط میرعلی تبریزی ابداع شده است. این خط بسیار زیبا بوده و معمولاً برای نوشتن اشعار و متون ادبی استفاده می‌شود.

یکی از ویژگی‌های این خط کشیده نوشتن حروف در آن می‌باشد. مثلاً "ب" می‌تواند اندازه‌ای برابر ۳ نقطه تا ۱۲ نقطه داشته باشد، که این کشیدگی کار تشخیص حروف نستعلیق را برای برنامه‌های تشخیص حروف مشکل می‌کند. ویژگی عمده دیگر این خط وجود "س" و "ش" کشیده و بدون دندانه در آن می‌باشد که این مسأله نیز تشخیص حروف "س" و "ش" را مشکل می‌کند. مسأله دیگر شکل معکوس برای حرف "ی" می‌باشد که می‌تواند تا ابتدای کلمه ادامه یابد؛ مانند کلمه "شکستگی" در شکل ۴. ویژگی دیگر خط نستعلیق فشرده نوشتن حروف در آن می‌باشد که حتی در بعضی از موارد جای نقطه‌ها هم تنگ می‌گردد.

با توجه به موارد ذکرشده، در این پروژه از خط نستعلیق برای شناسایی کاربران انسانی استفاده شده است چون این خط به راحتی توسط کاربران فارسی و عربی زبان خوانده شده ولی برنامه‌های تشخیص حروف امروزی قادر به خواندن آن نیستند.

٤- الگوریتم پیشنهادی

در این مقاله به ارائه روشی برای تفکیک کاربران انسانی فارسی و عربی زبان از نرم‌افزارهای رایانه‌ای با استفاده از ویژگی‌های زبان‌های فارسی و عربی پرداخته شده است.

به منظور شرح الگوریتم پیشنهادی و ویژگی‌های آن، در هر مرحله به شرح عمل انجام‌شده در این حوزه در زبان انگلیسی پرداخته و سپس به مقایسه آن با الگوریتم پیشنهادی در زبان فارسی و عربی می‌پردازیم.

٤-۱ افزودن نویز به تصویر

یکی از مؤثرترین روش‌ها در جلوگیری از تشخیص حروف به وسیله نرم‌افزارهای رایانه‌ای، افزودن نویز به تصویر می‌باشد. بدین منظور با انجام کارهایی همچون افزودن پس‌زمینه‌هایی به تصویر موجب می‌شوند که نرم‌افزارهای تشخیص حروف با مشکل مواجه گردند. البته امروزه روش‌هایی به منظور حذف این نویزها در تصاویر کلمات انگلیسی ابداع شده است، اما در زبان‌های فارسی و عربی به علت وجود سه عامل نقطه، علائم خاص و اعراب، جداسازی نویز از تصویر مشکل‌تر می‌باشد. زیرا حذف نویز باعث حذف نقاط و خطوط ریز اطراف متن می‌شود، حال آنکه در زبان‌های فارسی و عربی نیمی از حروف دارای نقطه بوده، همچنین بعضی از کلمات دارای علائمی چون تشدید و همزه بوده و در نهایت اینکه بعضی مواقع و به خصوص در محل‌هایی که احتمال اشتباه در خواندن وجود دارد، از اعراب‌گذاری استفاده می‌شود. این نقاط و علائم ممکن است در هنگام حذف نویز از تصویر، حذف شده و لذا تشخیص حروف توسط نرم‌افزار رایانه‌ای به درستی صورت نمی‌گیرد.

٤-۲ استفاده از حروف مشابه

به منظور پیچیده‌ساختن عمل تشخیص حروف، معمولاً از کلماتی که دارای حروف با شکل مشابه، مانند حروف "m" و "w" یا "g" و "q" یا "i" و "j" هستند، استفاده می‌شود. در زبان‌های فارسی و عربی تعداد اینگونه حروف مشابه بسیار بیشتر است. این تشابه زیاد حروف، معمولاً از تفاوت حروف در تعداد نقاط ناشی می‌شود، مانند حروف "ب، پ، ت و ث" یا "س و ش" و غیره که در تعداد یا محل نقاط اختلاف دارند.

حروفی نیز مانند "ک، گ و ل" نیز از نظر شکل ظاهری مشابه هستند. به طور کلی تعداد حروف مشابه در زبان‌های فارسی و عربی بسیار زیاد بوده لذا تفکیک این حروف از یکدیگر برای نرم‌افزارهای تشخیص حروف دشوار می‌باشد.

٤-٣ متصل‌بودن حروف به یکدیگر

یکی از مشکل‌ترین کارها برای نرم‌افزارهای تشخیص حروف، جداسازی حروف از یکدیگر می‌باشد. با توجه به جدابودن حروف از یکدیگر در زبان انگلیسی، در سیستم‌های تفکیک کاربران انسانی از نرم‌افزارهای رایانه‌ای سعی می‌شود تا با روش‌هایی همچون کم‌کردن فاصله حروف، اتصال حروف به یکدیگر به وسیله خطوط و کمان (مانند روش ۲-۶) و غیره، حروف را به یکدیگر متصل نمایند.

در زبان‌های فارسی و عربی بر خلاف زبان انگلیسی، حروف به هنگام نوشتن به یکدیگر متصل می‌شوند. لذا نیازی به اعمال روش‌های فوق برای اتصال حروف به یکدیگر نمی‌باشد. از آنجایی که این امر به طور طبیعی صورت می‌گیرد، در نتیجه با مشکلاتی که در زبان انگلیسی در هنگام متصل‌کردن حروف به طور مصنوعی وجود دارد؛ همچون نادرست خواندن حروف توسط انسان‌ها و ایجاد حروف اضافه، وجود نخواهد داشت.

از طرف دیگر در زبان‌های فارسی و عربی علاوه بر فاصله میان‌کلمه‌ای، در بعضی کلمات مانند "می‌باشد"، "بی‌خوابی" و غیره فاصله میان‌حرفی؛ یعنی فاصله کوتاهی که میان حروف کلمات وجود دارد؛ مانند فاصله کوتاهی که میان دو قسمت "می" و "باشد" در کلمه "می‌باشد" قرار دارد، نیز وجود دارد که این امر عمل تشخیص حروف را برای نرم‌افزارهای رایانه‌ای مشکل می‌سازد. حتی در مواردی فاصله میان‌کلمه‌ای نیز در عبارات فارسی و عربی رعایت نمی‌شود، به ویژه در خط نستعلیق، که این امر بر دشواری تشخیص حروف می‌افزاید.

علاوه بر دو مورد فوق، در زبان‌های فارسی و عربی با مسئله هم‌پوشانی حروف، یعنی قرارگرفتن بعضی از حروف روی حرف قبلی؛ نیز مواجه هستیم که این امر نیز جداسازی حروف از یکدیگر را دشوارتر می‌سازد.

٤-٤ تفاوت اندازه حروف

در زبان انگلیسی حروف دارای اندازه ثابتی هستند، در حالی که در زبان‌های فارسی و عربی به خصوص در خط نستعلیق به منظور زیبایی کلمه، یک یا چند حرف از کلمه به صورت کشیده نوشته می‌شوند. مثلاً حرف "ب" در کلمه "بـــاخت" می‌تواند به صورت کشیده نوشته شود. این امر موجب می‌شود که نرم‌افزارهای تشخیص حروف نتوانند به درستی حروف کلمه را از یکدیگر جدا ساخته و تشخیص دهند.

همچنین در زبان‌های فارسی و عربی اندازه همه حروف یکسان نیست و به عنوان مثال حرف "پ" بیشتر از حرف "د" فضا اشغال می‌کند. لذا کار برای نرم‌افزارهای تشخیص حروف دشوارتر می‌شود.

٤-٥ شکل‌های مختلف حروف

در زبان انگلیسی حروف تنها دارای دو حالت بزرگ و کوچک هستند، اما در زبان‌های فارسی و عربی به علت اتصال حروف به یکدیگر، حروف در مکان‌های مختلف دارای شکل‌های متفاوتی هستند. به طوری که هر حرف با توجه به مکانی که در آن قرار دارد، ممکن است دارای چهار شکل گوناگون باشد. در خط نستعلیق بعضی حروف همچون حروف "س" و "ی"، در حالت‌های یکسان نیز دارای اشکال گوناگونی می‌باشند. این شکل‌های گوناگون موجب پیچیدگی بیشتر تشخیص حروف در زبان‌های فارسی و عربی می‌شود.

٤-٦ بهره‌گیری از لغت‌نامه

در زبان انگلیسی به منظور تکمیل تشخیص کلمات، از لغت‌نامه‌ها استفاده می‌کنند. متأسفانه برای زبان‌های فارسی و عربی لغت‌نامه‌های مناسبی وجود ندارد، لذا استفاده از این تکنیک برای تشخیص حروف در متون فارسی و عربی امکان‌پذیر نیست. در نتیجه قدرت نرم‌افزارهای رایانه‌ای تشخیص حروف برای خواندن متون فارسی و عربی کاهش می‌یابد.

با توجه به عوامل فوق، در این طرح برای تفکیک کاربران انسانی فارسی یا عربی زبان از نرم‌افزارهای رایانه‌ای، از تصویر کلمات فارسی و عربی استفاده شده است. بدین منظور ابتدا یک کلمه فارسی یا عربی با طول ۳ تا ۸ حرف انتخاب می‌شود. گرچه انتخاب کلمات طولانی‌تر عمل تشخیص حروف توسط نرم‌افزارهای رایانه‌ای را مشکل‌تر می‌سازد، اما تایپ کلمات طولانی موجب ناراحتی کاربر انسانی می‌شود.

لغت انتخاب‌شده به صورت یک پرونده تصویری می‌باشد. سپس تصویر با یک پس‌زمینه سیاه و سفید ترکیب شده و بر روی این تصویر مقداری نویز قرار داده می‌شود تا همان‌طور که در قسمت ۴-۱ ذکر شده تشخیص حروف توسط نرم‌افزارهای رایانه‌ای مشکل شود. البته پس‌زمینه انتخابی و مقدار نویز افزوده‌شده به گونه‌ای می‌باشند که خواندن کلمه برای کاربر انسانی به آسانی امکان‌پذیر باشد.

در نهایت تصویر به کاربر نشان داده شده و از او خواسته می‌شود تا آن کلمه را تایپ کند. در صورت تطابق کلمه تایپ‌شده با کلمه نشان داده شده، اجازه انجام عملیات مورد نظر به کاربر داده می‌شود.

٥- نتایج عملی

در این مقاله، به ارائه روشی برای تفکیک کاربران فارسی یا عربی‌زبان از نرم‌افزارهای رایانه‌ای به کمک متون فارسی و عربی پرداخته شده است. روش پیشنهادی به صورت عملی و توسط زبان برنامه‌نویسی جاوا پیاده‌سازی شده است. این نرم‌افزار در قالب یک اپلت[1] جاوا درون یک صفحه وب جایگذاری شده و پس از قرارگرفتن بر روی یک پایگاه اینترنتی مورد آزمایش قرار گرفت. اپلت‌های جاوا نرم‌افزارهایی به زبان جاوا هستند که بر روی صفحات وب و از طریق شبکه جهانی اینترنت قابل اجرا می‌باشند.

در پیاده‌سازی این روش، بانکی از تصاویر کلمات نوشته‌شده به شیوه نستعلیق در قالب تصویری PNG[2] تهیه شد. نمونه‌ای از این کلمات در شکل ۴ آورده شده‌اند. نرم‌افزار مذکور به طور تصادفی یکی از این تصاویر را انتخاب کرده و به کاربر نشان می‌دهد. سپس از وی درخواست می‌کند تا کلمه نشان داده شده را تایپ کند. پس از تایپ کلمه توسط کاربر، در صورت تطابق کلمه تایپ‌شده با تصویر کلمه نمایش داده شده، نرم‌افزار با پیغام مناسب کاربر را مطلع می‌سازد.

کلمه نسبت داده شده به هر تصویر در پرونده‌ای جداگانه نگهداری شده و نرم‌افزار فوق برای تشخیص صحت کلمه تایپ‌شده توسط کاربر به آن پرونده مراجعه می‌کند. با توجه به پیچیدگی‌های خط نستعلیق و عدم قدرت نرم‌افزارهای تشخیص حروف در بازشناسی متون فارسی و عربی

1. Applet
2. Portable Network Graphics

حتی در معمولی‌ترین حالت، لذا در اینجا از تصاویر معمولی کلمات بهره گرفته شده و بر خلاف روش‌های به کار رفته در زبان انگلیسی، هیچ‌گونه اعوجاجی به تصویر داده نشده است. این امر موجب می‌شود تا تشخیص کلمات برای کاربران راحت‌تر باشد. هم‌اکنون نمونه آزمایشی این نرم‌افزار بر روی پایگاه اینترنتی www.shirali.ir/projects/nastaliqcaptcha موجود می‌باشد.

روش ارائه‌شده در این مقاله توسط تعدادی کاربر با بازه سنی بین ۱۲ تا ۵۵ سال آزمایش گردید. در ۹۷ درصد مواقع، کاربران در اولین مرتبه قادر به تشخیص صحیح کلمات نمایش داده شده بودند و تنها در ۳ درصد مواقع که معمولاً ناشی از خطا در تایپ‌کردن کلمات بود، در تلاش‌های بعدی توانستند کلمه را به درستی تشخیص دهند. کاربران از نظر راحتی استفاده، از این روش اظهار رضایت کردند.

٦- نتیجه‌گیری

در این مقاله، روشی برای تفکیک کاربران انسانی فارسی یا عربی‌زبان از نرم‌افزارهای رایانه‌ای به وسیله متون فارسی و عربی ارائه شده است. زبان عربی زبان مذهبی همه مسلمانان جهان است. همچنین خط نستعلیق در دیگر کشورها مانند پاکستان و مصر نیز استفاده می‌شود. لذا این روش طیف وسیعی از کاربران اینترنتی را پوشش می‌دهد.

در زبان فارسی علاوه بر خط نستعلیق، انواع دیگری از خطوط از جمله خط شکسته نستعلیق و خط شکسته وجود دارد که بسیار پیچیده‌تر از خط نستعلیق می‌باشند، به طوری که خواندن بعضی از این خطوط توسط انسان نیز مشکل می‌باشد. لذا می‌توان از این خطوط برای تفکیک کاربران انسانی از نرم‌افزارهای رایانه‌ای بهره برد.

از آنجا که روش تفکیک کاربران انسانی از نرم‌افزارهای رایانه‌ای علی‌الخصوص روش استفاده از متون فارسی و عربی، به تازگی مطرح شده است، لذا همچنان زمینه‌های بسیاری برای گسترش و بهبود در این روش‌ها وجود دارد که بایستی با تحقیقات بیشتر اقدام به توسعه آن نمود.

مراجع

[1] G. Mori and J. Malik, "Recognizing objects in adversarial clutter: breaking a visual CAPTCHA," in *Proc. of the IEEE CS Society Conf. on Computer Vision and Pattern Recognition (CVPR'03),* pp. 134-141, Madison, US, Jun. 2003.

[2] G. Moy, N. Jones, C. Harkless, and R. Potter, "Distortion estimation techniques in solving visual CAPTCHAs," in *Proc. of the 2004 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR'04)*, vol. 2, pp. 23-28, Jun./Jul. 2004.

[3] H. S. Baird and K. Popat, "Human interactive proofs and document image analysis," in *Proc. of the 5th IAPR International Workshop on Document Analysis Systems*, pp. 507-518, Princeton, US, Aug. 2002.

[4] M. Blum, L. A. Von Ahn, and J. Langford, *Completely Automatic Public Turing Test to Tell Computers and Humans Apart*, The CAPTCHA Project, www.captcha.net, Deptartment of Computer Science, Carnegie-Mellon University, 2000.

[5] M. Chew and H. S. Baird, "Baffletext: a human interactive proof," in *Proc. of the 10th SPIE/IS&T Conf. on Document Recognition and Retrieval (DR&R2003),* pp. 305-316, Santa Clara, US, Jan. 2003.

[6] A. Rusu and V. Govindaraju, "Handwritten CAPTCHA: using the difference in the abilities of humans and machines in reading handwritten words," in *Proc. of the 9th International Workshop on Frontiers in Handwriting Recognition (IWFHR-9)*, pp. 226-231, Oct. 2004.

[7] *PayPal Registration*, in URL: http://www.paypal.com/

[8] W. Liao and C. Chang, "Embedding information within dynamic visual patterns," in *Proc. of the IEEE International Conf. on Multimedia and Expo 2004 (ICME'04)*, vol. 2, pp. 895-898, Jun. 2004.

[9] *Microsoft Hotmail*, in URL: http://www.hotmail.com/

[10] A. L. Coates, H. S. Baird, and R. J. Fateman, "Pessimal print: a reverse turing test," in *Proc. of the 6th International Conf. on Document Analysis and Recognition,* pp. 1154-1158, Seattle, US, Sep. 2001.

[11] H. S. Baird and T. Riopka, "Scattertype: a reading CAPTCHA resistant to segmentation attack," in *Proc. of the IS&T/SPIE Conf. on Document Recognition & Retrieval XII (DR&R2005),* pp. 197-207, San Jose, US, Jan. 2005.

[12] H. S. Baird and J. L. Bentley, "Implicit CAPTCHAs," in *Proc. of the SPIE/IS&T Conf. on Document Recognition and Retrieval XII (DR&R2005),* pp. 191-196, San Jose, US, Jan. 2005.

[13] T. Y. Chan, "Using a text-to-speech synthesizer to generate a reverse turing test," in *Proc. of the 15th IEEE Int. Conf. on Tools with Artificial Intelligence*, pp. 226-232, Nov. 2003.

[14] M. Shirali-Shahreza and S. Shirali-Shahreza, "Drawing CAPTCHA," in *Proc. of the 28th Int. Conf. Information Technology Interfaces (ITI 2006),* pp. 475-480, *Cavtat, Croatia*, Jun. 2006.

[۱۵] م. شیرعلی شهرضا، *تشخیص ارقام و کلمات دست‌نویس فارسی به کمک شبکه‌های عصبی*، پایان‌نامه دوره دکترا، دانشکده مهندسی برق، دانشگاه صنعتی امیرکبیر، ۱۳۷۴.

[۱۶] ح. فضائلی، *اطلس خط*، انتشارات سروش، چاپ ششم، تهران، ۱۳۷۰.

محمدحسن شیرعلی شهرضا تحصیلات خود را در رشته مهندسی کامپیوتر (سخت‌افزار) در مقاطع کارشناسی و کارشناسی ارشد به ترتیب در سال‌های ۱۳۶۴ و ۱۳۶۷ در دانشگاه‌های صنعتی اصفهان و صنعتی شریف به پایان رسانده است. وی در سال ۱۳۷۴ به عنوان اولین دانش‌آموخته دکترای مهندسی کامپیوتر در دانشگاه‌های ایران از دانشگاه صنعتی امیرکبیر (پلی‌تکنیک تهران) فارغ‌التحصیل شد. نامبرده یک سال از دوره دکترای خود را در سال ۱۳۷۳ در دانشگاه متدیست جنوب (SMU) در آمریکا بوده است.

دکتر شیرعلی شهرضا از سال ۱۳۷۵ عضو هیئت علمی دانشکده مهندسی کامپیوتر دانشگاه یزد می‌باشد. زمینه‌های تحقیقاتی مورد علاقه وی شامل تشخیص حروف فارسی، نهان‌نگاری داده، تفکیک کاربران انسانی از رایانه و شبکه‌های عصبی می‌باشد.

محمد شیرعلی شهرضا دانشجوی سال دوم کارشناسی رشته علوم کامپیوتر در دانشگاه صنعتی شریف می‌باشد. وی فارغ‌التحصیل دبیرستان علامه حلی تهران (تیزهوشان) است. او موفق به کسب رتبه اول پنجمین جشنواره جوان خوارزمی با ارائه طرح "نهان‌نگاری داده در تصویر" شد. او در سال ۱۳۸۵ به عنوان پژوهشگر جوان ممتاز انجمن رمز ایران در مقطع کارشناسی برگزیده شد.

وی در دومین کنفرانس بین‌المللی ICTTA 2006 و همچنین در یازدهمین کنفرانس بین‌المللی انجمن کامپیوتر ایران به عنوان جوان‌ترین محقق برگزیده شد.

وی تاکنون ۲۹ مقاله علمی در کنفرانس‌های بین‌المللی ارائه کرده و دارای هشت مقاله چاپ‌شده در مجلات معتبر علمی است.

زمینه‌های تحقیقاتی مورد علاقه او نهان‌نگاری اطلاعات، برنامه‌نویسی تلفن همراه و سیستم‌های تفکیک کاربران انسانی از رایانه می‌باشد.