سال پنجم، شماره‌های ۱۷ و ۱۸، پاییز و زمستان ۱۳۹۲
صص: ۱۰۸- ۹۷

فصلنامه علمی- پژوهشی
فناوری اطلاعات و ارتباطات ایران

طراحی پایگاه داده کلمات دستنویس کردی سورانی برای سیستم‌های تشخیص نوری کلمات

فاطمه دانشفر* بصیر علاقه‌بند** وریا فتحی**

*کارشناس ارشد، دانشکدهٔ کامپیوتر، دانشگاه کردستان، کردستان

**کارشناس، فناوری اطلاعات، دانشکدهٔ کامپیوتر، دانشگاه کردستان، کردستان

تاریخ دریافت: ۱۳۹۲/۰۱/۱۵ تاریخ پذیرش: ۱۳۹۲/۰۳/۲۲

چکیده

یکی از اجزای زیربنایی سیستم‌های تشخیص تصویری کلمات (OCR) پایگاه داده‌ها هستند و سیستم‌هایی که در این زمینه طراحی می‌شوند بطور معمول از یک نوع پایگاه داده استفاده می‌کنند. بدیهی است از آنجا که موضوع مورد مطالعه در این سیستم ها شکل نوشتاری (رسم الخط) زبان‌های مختلف است بنابراین برای هر زبانی پایگاه دادهٔ بخصوصی لازم است. زبانی که این مقاله بر آن متمرکز شده کردی است و در این مقاله مراحل مختلف چگونگی طراحی اولین پایگاه دادهٔ حروف دستنویس برای زبان کردی شرح داده شده است. از آنجا که تاکنون هیچ پایگاه داده‌ای مخصوص تشخیص تصویری کلمات، مربوط به زبان کردی طراحی نشده است بنابراین زمینه‌ای بکر و مستعد برای انجام تحقیق محسوب می‌شود. همچنین با توجّه به اینکه زبان کردی دارای دو رسم‌الخط مختلف لاتین و آرامی است در این مقاله منحصراً به رسم الخط آرامی البته از نوع دستنویس آن پرداخته شده است.

واژه‌های کلیدی: سیستم‌های تشخیص تصویری کلمات، پایگاه داده‌ها، زبان کردی، دستنویس.

مقدمه

پایگاه داده‌های تصویری نقشی اساسی در سیستم‌های تشخیص تصویری کلمات دارند. تقریباً تمامی عملیات‌های آموزشی[1] و آزمایشی که در این سیستم‌ها انجام می‌پذیرد وابسته به پایگاه داده‌های حروف است. بنابراین پرواضح است نقش چنین جزئی از سیستم تا چه اندازه مهم و اساسی است. گاهی به دلیل حجم و تنوع داده‌ها در برخی پروژه‌ها، دیده شده که حتی بیش از یک پایگاه داده نیز استفاده شده است.زبانی که این پژوهش بر طراحی پایگاه دادهٔ تصویری آن متمرکز شده است زبان کردی است، این زبان دارای دو رسم‌الخط مختلف لاتین و آرامی است که این مقاله منحصرا به رسم الخط آرامی و به صورت دستنویس آن می‌پردازد هر چند که متدهایی که در این پژوهش ارائه شده‌اند قابلیت تعمیم به حوزه کارکرد متون تایپی را نیز دارند و به سهولت می‌توان این پایگاه داده را در خصوص کلمات تایپ شده نیز بکار برد.

1 .Training

نویسندهٔ عهده‌دار مکاتبات: فاطمه دانشفر f_daneshfar@yahoo.com

- **پیش زمینه**

همانطور که پیش از این اشاره شد شکل مورد نظر ما از رسم‌الخط زبان کردی همان رسم‌الخط آرامی یا به تعبیری کردی سورانی است. زبان‌هایی چون عربی، فارسی و کردی سورانی از الفبای موسوم به الفبای آرامی بهره می‌برند که یکی از مهم‌ترین و بزرگ‌ترین شاخه‌های زبان‌های دارای الفبای بهم چسبیده است، به همین دلیل شباهت‌های ظاهری فراوانی میان شکل نوشتاری این زبان‌ها (زبان‌های با الفبای آرامی) وجود دارد. بر این اساس چون تجربه‌های زبان کردی در زمینهٔ سیستم‌های مبتنی بر تشخیص تصویری کلمات، چه بصورت دستنویس و چه بصورت تایپی، بسیار اندک و ناچیز است، بنابراین برآن شدیم از تجاربی که در زبان‌های مشابه انجام گرفته‌اند به عنوان پایه و اساس کار خود بهره‌برداری کنیم. در این پژوهش سعی شده است تا بیش‌تر با استخراج قوانین و ضوابط متناسب با قواعد نوشتاری زبان کردی، روند کار به سمت یک فرآیند بومی سازی سوق داده شود. ضمن اینکه با توجّه به پیشرو بودن در این حوزه و عدم تجارب قبلی بر بستر زبان کردی، سعی شده است حتی‌الامکان قاعدهٔ سادگی را سرلوحهٔ کار خود قرار دهیم. بدین مفهوم که تأکید ما بیش‌تر بر انجام هر چه درست‌تر و اصولی‌تر کار بوده است، تا این که بر پیچیدگی و شاخ و برگ‌دادن به آن تأکید ورزیم. در ادامه به دو نمونه از پایگاه داده‌هایی که در دو زبان عربی و فارسی مطالعه شده است اشاره می‌شود.

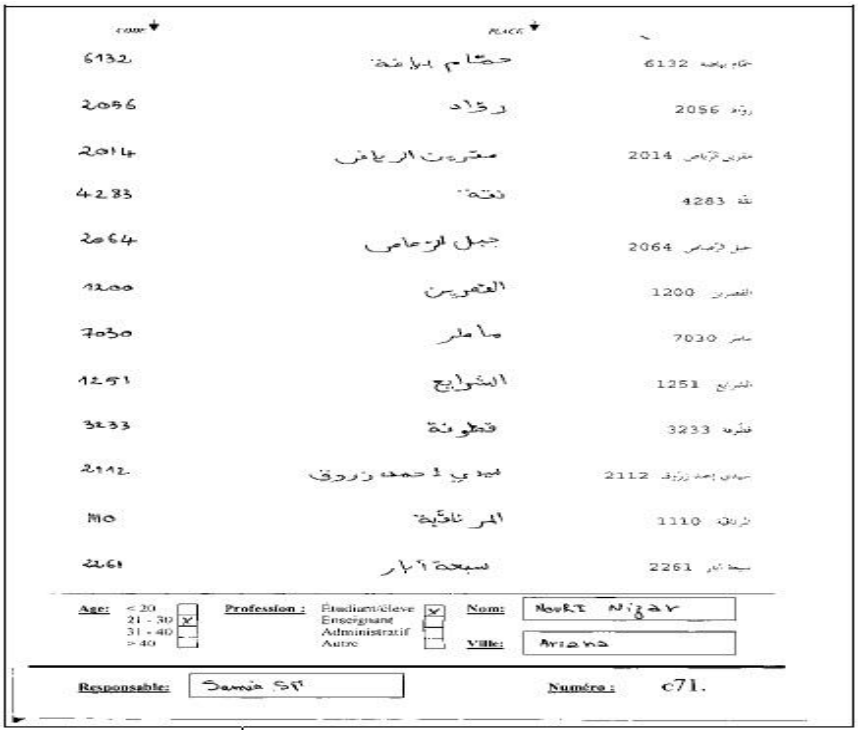
۱. پایگاه داده IFN/ENIT

پایگاه داده استانداردی است که حاوی متن کلمه، تصویر دستنویس آن، محل قرار گرفتن خط زمینه، تعداد کاراکتر هر کلمه و مانند آن است [۱]. این پایگاه داده تاکنون بیش‌ترین سهم را در پشتیبانی از پروژه‌های تشخیص تصویری کلمات زبان عربی داشته است و اکثر روش‌های ارائه شده بر بستر زبان عربی بر آن تست شده‌اند [۲]. خواص هر کلمه در پایگاه داده IFN/ENIT به دو صورت خودکار و یا با استفاده از بینایی انسان (استفاده از افراد برای تشخیص) برای هر کلمه تعیین شده است. خط زمینه از جمله ویژگی‌هایی است که با کمک گرفتن از عامل انسانی و نه به صورت خودکار، محل آن مشخص شده است. این پایگاه داده شامل نام ۹۴۶ شهر و روستا درکشور تونس است که با تکرار نام این شهرها توسط ۴۹۲ نفر از مردم نوشته شده و در حال حاضر شامل ۳۲۴۹۲ نمونه است. تمامی این نمونه‌ها هریک طی چندین مرحله امّاده‌سازی و پردازش در نهایت به تصویری دو سطحی (که در آن پیکسل‌های سیاه و سفید دارای مقادیر صفر و یک هستند) تبدیل و در پایگاه داده ذخیره شده‌اند. این پایگاه داده دارای دو نسخه متفاوت است که در این پروژه، آخرین نسخه آن مورد بررسی قرار گرفته است [۳]. شکل های ۱ و ۲ به ترتیب نمونه‌ای از تصاویر موجود در IFN/ENIT و فرم مربوطه که در این پایگاه داده برای جمع‌آوری تصاویر واژه‌ها از متن دستنویس مورد استفاده قرار گرفته است را نشان می‌دهند.

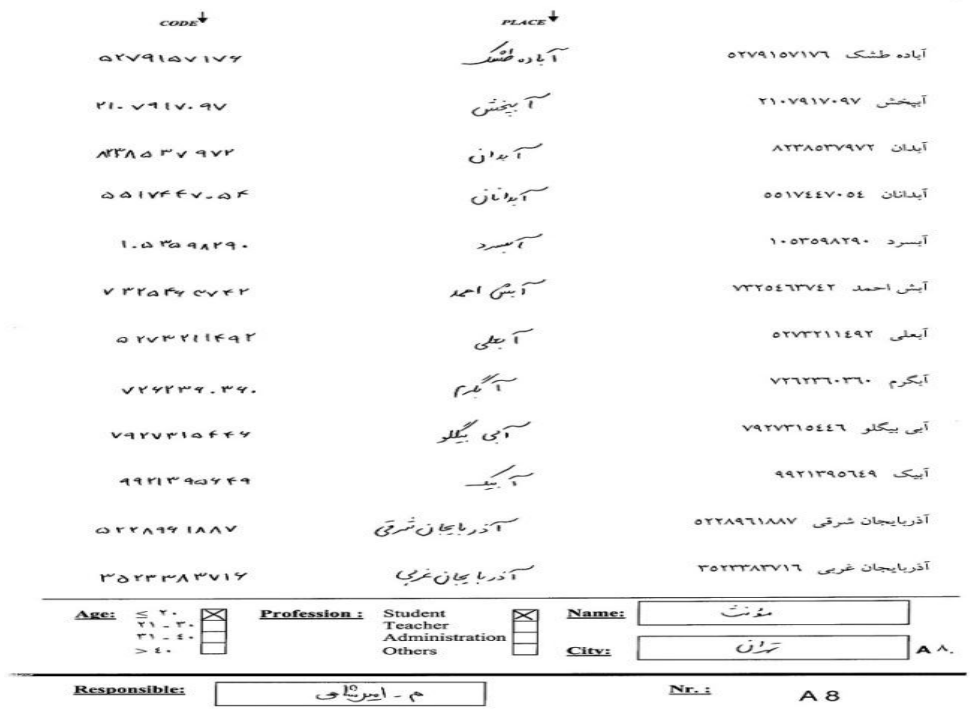
شکل ۱. نمونه‌ای از تصاویر موجود در پایگاه داده IFN/ENIT [۱]

۲. پایگاه داده IfN/Farsi

این پایگاه داده نیز اقتباسی است از IFN/ENIT با این تفاوت که بر بستر زبان فارسی پیاده شده است و طبیعتاً بدلیل الگوبرداری از آن وجوه تشابه نسبتاً زیادی میان آن‌ها وجود دارد[۴]. دلیل این اقتباس شباهت فراوان رسم الخط زبان فارسی نسبت به عربی است بنابراین با توجّه به موفقیت و مقبولیت گسترده‌ای که IFN/ENIT کسب کرده است، راه ساده‌تر آن بوده که از ساختار کلّی آن برای ساخت یک پایگاه داده فارسی برای تشخیص تصویری کلمات استفاده شود.

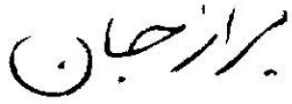


شکل ۲. فرمی (بصورت پرشده) که در پایگاه داده IFN/ENIT برای گردآوری تصاویر واژه‌های دستنویس استفاده شده است [۱]



شکل ۳. فرمی (بصورت پرشده) که در پایگاه داده IfN/Farsi برای گردآوری تصاویر واژه‌های دستنویس استفاده شده است [۱]

امّا پایگاه داده IfN/Farsi متشکل از ۷۲۷۱ تصویر باینری از اسامی ۱۰۸۰ شهر و استان کشور ایران است که توسط ۶۰۰ نفر در شرایط سنی و جنسی مختلف به رشته تحریر در آمده است. همچنین برای هر تصویر خصوصیاتی از جمله فایل تصویری آن، ZIP code، کد کلمه، خط کرسی، تعداد حروف و زیرواژه‌ها[۲] و همچنین در صورت وجود داشتن نقطه، تعداد نقطه‌ها درنظر گرفته شده است [۵] [۶]. در شکل ۴ نمونه‌ای از تصویر یک واژه موجود در پایگاه داده IfN/Farsi و در شکل ۳ نمونه‌ای از فرم پر شده‌ای را می‌بینید که در این پایگاه داده برای جمع‌آوری تصاویر واژه‌ها از متن دستنویس مورد استفاده قرار گرفته است. در حقیقت تصویر شکل ۴ از فرم شکل ۳ استخراج گردیده است.



شکل ۴. نمونه‌ای از تصاویر موجود در پایگاه داده IfN/Farsi [۱]

۳. پایگاه داده کردی

تعداد واژه‌هایی که از لغت‌نامه‌های موجود کردی برای پایگاه داده استخراج شده‌اند بیش از ۵۵۰۰۰ کلمه است، امّا به دلیل محدودیت عملیاتی در واقع تنها از این میان ۲۱۰۰ کلمه به عنوان جامعه، نمونه برای نسخه‌برداری دستنویس انتخاب شده‌اند. فرم‌های استانداردی شبیه آنچه که در پایگاه داده‌های IFN/ENIT و IfN/Farsi بکار رفته‌اند برای این کار تدارک یافته‌اند. در هر فرم ۹ کلمه بصورت سرمشق نوشته شده است و به هر فرد بطور متوسط ۵ فرم تحویل داده شده تا دستخط خود را مقابل کلمات تایپ شده بنویسد. شکل ۵ نمونه‌ای از تصویر یک فرم استاندارد پرشده است.بدین ترتیب دو نوع کلّی از داده‌ها را خواهیم داشت: ابتدا، متن کلمات و سپس عکس‌های اسکن شده از فرم‌هایی که توسط افراد مختلف پر شده است. به دلیل حجم نسبتاً پایین داده‌ها تصمیم گرفته شد که پیاده سازی پایگاه داده به وسیلهٔ نرم‌افزار مایکروسافت اکسس[۳] انجام گیرد. همچنین پیش از هر چیز برنامه کاربردی[۴] برای تولید و مدیریت داده‌ها طراحی شد تا به کمک ابزار اکسس، ساختمان اصلی پایگاه داده‌ها را تشکیل دهد. قسمت اصلی وظایف این برنامهٔ کاربردی در مرحلهٔ اول گردآوری و واردکردن اطلاعات به پایگاه داده‌ها و همچنین ویرایش آن‌ها است.

در کل، هر رکورد این پایگاه داده تنها دربردارندهٔ دو موجودیت[۵] است: یکی موجودیت کلمه و دیگری عکس اسکن شده که متناظر با کلمات هستند. هر یک از عکس‌های مذکور شامل تصویر اسکن شده یک کلمه دستنویس است که توسط افراد گوناگون نوشته شده است. بدین ترتیب بدلیل اینکه برای هر کلمه چند عکس اسکن شده مربوط به دستخط‌های متفاوت وجود دارد پس رابطهٔ مابین موجودیت‌ها یک به چند و از سمت کلمات به عکس‌ها است (شکل ۶). در ادامه به تفصیل به دو موجودیت مذکور، جدول‌ها[۶] و صفت‌های[۷] مربوط به آن‌ها پرداخته خواهد شد.

- **موجودیت کلمه**

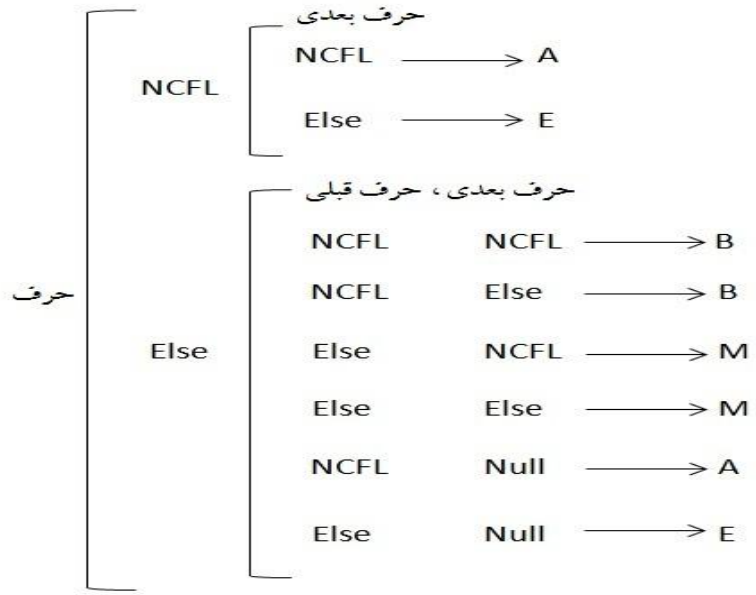
محوریت این موجودیت، کلمات و صفات آن‌ها هستند. یک واژه را به تنهایی در نظر بگیرید، از آن‌رو که سیستم‌های تشخیص تصویری کلمات به اطلاعات جامعی نیاز دارند تا بتوانند در سطحی مطلوب عمل تشخیص را انجام دهند بنابراین باید ویژگی‌های مهم ساختار ظاهری واژه‌ها استخراج شوند. صفاتی که در اینجا برای کلمات در نظر گرفته شده است و فیلدهای جدول موجودیت کلمه را تشکیل می‌دهند عبارتند از: نام کلمه، تعداد کلمه، کد کلمه، تعداد زیرواژه، تعداد حروف یا کاراکتر. فیلد نام کلمه، خود واژه مورد نظر را بصورت متنی[۸] دربردارد. کلّید اصلی[۹] جدول حاضر نام کلمه است و تعداد کل کلمات منحصر بفردی که در پایگاه‌داده‌ها وجود دارند ۵۵۹۳۷ کلمه است.

---

2 .Sub-Word

3 .Microsoft Access
4 .Application Program
5 .Entity
6 .Table
7 .Property
8 .Text
9 .Primary Key

شکل ۵. نمونهٔ فرم بکار رفته شده که در پایگاه داده کردی برای جمع‌آوری تصاویر واژه‌های دستنویس استفاده شده است



شکل ۶. نمودار کدگذاری واژه‌ها: در این نمودار خط سیر کدگذاری هر حرف با هر نوع قرارگیری در ساختمان واژه‌ها مشخص شده است

فیلد نام کلمه، خود واژه مورد نظر را بصورت متنی[10] دربردارد. کلّید اصلی[11] جدول حاضر نام کلمه است و تعداد کل کلمات منحصر بفردی که در پایگاه‌داده‌ها وجود دارند ۵۵۹۳۷ کلمه است.

فیلد تعداد کلمه عملاً برای کلمات مرکب یا عباراتی کاربرد دارد که از ترکیب چند واژه درست شده باشند. تعداد کلمات با شمارش تعداد فاصله‌ها[12] محاسبه می‌شود، یعنی،

تعداد کلمه = تعداد فاصله + ۱

کد کلمه نشان‌دهندهٔ کاراکترهای کلمه و نیز جایگاه و نوع قرارگیری هر یک از آن‌ها در ساختمان کلمه است. بطور مثال کد واژهٔ «مباره‌ک» بدین‌گونه بیان می‌شود:

|B_م|M_ب|I_ا|A_ر|A_ه|A_ک. همانطور که مشاهده می‌کنید کاراکترها به ترتیب و براساس نحوهٔ قرارگرفتن کدگذاری شده‌اند. حرف «م» در آغاز کلمه قرار دارد و با کد «B[13]» یا آغاز نوشته می‌شود، حرف «ب» از دو طرف به حروف دیگری چسبیده و اصطلاحاً «ب» وسط نامیده شده و با کد «M[14]» نوشته می‌شود، حرف «ا» تنها از طرف راست به حرف دیگری چسبیده و اصطلاحاً «آ»ی آخر گفته می‌شود و با کد «E[15]» نمایش داده می‌شود، همچنین حروف «ر»، «ه» و «ک» چون به حروف دیگری نچسبیده‌اند و تنها هستند با کد «A[16]» نشان داده می‌شوند. نحوهٔ تولید این کد از این قرار است، حروف الفبای کردی را به دو نوع NCFL و Else تقسیم می‌کنیم. حروف « ا - د - ر - ز - ژ - و - ه (دو چشم) - Null » در نوشتار کردی هیچ‌گاه از سمت چپ به حروف دیگر نمی‌چسبند به همین دلیل اسم این گروه از حروف غیرچسبان از سمت چپ نامیده شده است، و باقی حروف نیز در گروه Else قرار می‌گیرند. بنابراین تعاریف، نمودار کدگذاری واژه‌ها را می‌توان بصورت شکل ۶ طراحی کرد.

همانطور که می بینید برای تعیین کدی که همراه هر کاراکتر می‌آید ابتدا می‌بایست نوع خود کاراکتر مشخص شود، سپس با توجّه به این پارامتر، از حروف بعدی و قبلی کمک گرفته و کد آن بدست آورده شود. نمودار فوق سلسله مراتب بدست آمدن تمامی حالات در فرآیند کدگذاری را تعیین می‌کند. البته باید توجّه داشت، فرض بر آن است که حرف اول هر کلمه کد B یا شروع، خواهد گرفت حال آنکه برای حرف آخر لزومی ندارد که همراه کد E یا آخر بیاید چون امکان دارد یک کاراکتر منفرد باشد و در آن صورت کد A معرف آن خواهد بود. به عنوان مثال در واژه «کوردستان» حرف «ن» با کد A همراه می‌شود.

فیلد تعداد زیرواژه‌ها به شمارهٔ اجزای هم‌بند هر کلمه اشاره دارد با این شرط که اعراب، نقطه‌گذاری‌ها و سرکش جزء زیرواژه‌ها محسوب نمی‌شوند؛ به عنوان مثال در واژهٔ «مباره‌ک»، «مبا»، «ر»، «ه» و «ک» زیرواژه هستند. تعداد زیرواژه‌های هر واژه توسط معادلات (۱) یا (۲) محاسبه می‌شود.

$$subwords = \sum(E\_X) + \sum(A\_X) \quad (1)$$

$$subwords = \sum(B\_X) + \sum(A\_X) \quad (2)$$

معادلات (۱) و (۲) دو شکل متفاوت از یک فرمول هستند. در واقع معادله (۱) بدین معناست که جمع تمام حروفی که با کد E همراه هستند (یعنی حرف آخر، زیرواژه است) به علاوهٔ حروفی که در کدگذاری با کد A می‌آیند (یعنی تنها هستند) بدون اینکه به حروف دیگر کلمه وصل باشند خود یک زیرواژه محسوب می‌شوند. امّا در معادلهٔ (۲) تفاوت فقط در این است که به جای تعداد حروف با کد E از حروف با کد B استفاده شده است، یعنی حرف اول زیرواژه‌ها. این بدان معنا است که برای شمارش تعداد زیرواژه‌ها کافی است تعداد حروف اول زیرواژه‌ها (با توجّه به

10 .Text
11 .Primary Key
12 .Space
13 .Beginning
14 .Middle
15 .End
16 .Alone

جدول ۱. نمونه‌ای از جدول موجودیت کلمه

نام کلمه	تعداد کلمه	کد کلمه	تعداد زیرواژه
ئەوراز	۱	ئ-B|ە-E|و-A|ر-A|ا-A|ز-A	۵
ئاسپایی	۱	ئ-B|ا-E|س-B|پ-M|ا-E|ی-B|ی-E	۳
ئاستش جیا	۲	ئ-B| ا-E|س-B|ت-M|ش-E|ج-B|ی-M|ا-E	۳
ئاگرخانە	۱	ئ-B|ا-E|گ-B|ر-E|خ-B| ا-E|ن-B|ە-E	۴
ئابرا	۱	ئ-B|ا-E|ب-B|ر-E| ا-A	۳
باوە قورەت	۲	ب-B| ا-E|و-A|ە-A|ق-B|و-E|ر-A|ە-A|ە-A	۷
بازوبەن	۱	ب-B| ا-E|ز-A|و-A|ب-B|ە-E|ن-A	۵

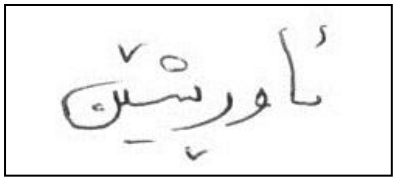
به خاصیت کدگذاری) به علاوه حروفِ تنها یا به عکس، تعداد حروف آخر زیرواژه‌ها به اضافه حروف تنها را محاسبه کنیم. توجّه داشته باشید به این دلیل از حروف وسط زیرواژه‌ها نمی‌توان استفاده کرد که امکان دارد در یک زیرواژه بیش از یک حرف وسط وجود داشته باشد. جدول ۱ نمونه‌ای از جدول موجودیت کلمه است که خصوصیات واژه‌ها و مقادیر آن‌ها را نمایش می‌دهد.

- **موجودیت عکس**

عکس‌های اسکن شده در واقع تصویر دستنویس واژه‌هایی هستند که توسط افراد مختلف نوشته شده‌اند. عکس‌ها از فرم‌های کاغذی استانداردی تهیّه می‌شوند که نویسندگان در مقابل هر واژه‌ای که تایپ شده دو مرتبه آن را با دستخط خود بازنویسی می‌کنند. شکل ۷ نمونه‌ای از این فرم‌ها را نشان می‌دهد.

امّا صفات موجودیت عکس یا همان فیلدهای جدول متناظر آن شامل موارد زیر هستند: نام کلمه، فایل عکس، خط کرسی[۱۷]، کیفیت خط کرسی، کیفیت عکس و نام نویسنده.

فیلد اول همان فیلدی است که در جدول کلمه نیز با همین نام وجود دارد. فیلد فایل عکس در واقع شامل آدرس فایل عکس اسکن شده می‌باشد و اینجا به عنوان کلّید اصلی بشمار می‌آید. در شکل ۷ عکس اسکن شده یک واژه نمایش داده شده است. این عکس‌ها از تصویر فرم‌های استانداردِ پر شده (شکل ۵) استخراج می‌گردند. بدیهی است به دلیل بالا بودن تعداد کلمات موجود در پایگاه داده نمی‌توان انتظار داشت هر فردی که فرم های استاندارد را پر می‌کند به ازای تمامی واژه‌ها دستخط خود را بنویسد پس برای هر نفر تعداد محدودی فرم در نظر گرفته می‌شود.



شکل ۷. نمونه‌ای از عکس اسکن شده یک کلمه

امّا صفت خط کرسی اساسا یک صفت مرکب است ولی چون در مدل رابطه‌ای (جدولی) نمی‌توان صفت مرکب را نمایش داد، آن را به اجزای تشکیل دهنده‌اش یعنی خط کرسی طرف راست و طرف چپ تقسیم می‌کنیم.

17 .Baseline

نام نویسنده:	لطفا هر کلمه را داخل کادر کمرنگ و خوانا بنویسید	صفحه : 6142
کیلدان		
کیل کردن		
کیلان		
کیلانه		
کیلانی		
کیلب		
کیلبه		
کیلبو		
کیلپه‌نی		

شکل ۸. تصویر فوق، فرم نمونهٔ تصاویر دستنویس واژه‌ها را نشان می‌دهد

خط کرسی با استفاده از یک روش متاهیوریستیک بدست می‌آید [۷]. روشی که در آن تصویر به دو قسمت مساوی چپ و راست تقسیم می‌شود و برای هریک خط کرسی جداگانه استخراج می‌گردد. دو فیلد خط کرسی طرف راست و چپ مقدار عددی را نگهداری می‌کنند که فاصله یا تعداد پیکسلی را نشان می‌دهد که خط کرسی در منتهی‌الیه سمت راست و چپ عکس با قسمت بالای عکس دارد. فیلد کیفیت خط کرسی و کیفیت عکس هر دو مقادیری کیفی می‌گیرند امّا نحوهٔ اختصاص مقادیر مزبور بصورت دستی است. بدین ترتیب که عاملی انسانی کیفیت خط کرسی و عکس را مورد قضاوت قرار می‌دهد و از میان امتیازهای درست، خوب و خطا یکی را برمی‌گزیند. این امتیازها می‌تواند به ترتیب معادل: ۱۰۰ ، ۵۰ و ۰ درصد میزان درستی خط کرسی تفسیر شود.

فیلد نام نویسنده نیز برای ثبت و ضبط نام شخصی که دستخط خود را به ما هدیه کرده و در این راه کمکی به ما نموده در نظر گرفته شده است.

در جدول شماره ۲ نمونه‌ای از جدول موجودیت عکس به نمایش گذاشته شده است.

همان‌گونه که قبلاً نیز اشاره شد از آنجا که به ازای هر واژه منحصر بفرد چند عکس (حداقل دو عکس) وجود دارد پس رابطهٔ میان موجودیت‌های کلمه و عکس یک به چند است، در نتیجه صفت نام کلمه در موجودیت کلمه و صفت فایل عکس در موجودیت عکس، کلّید خارجی[1] خواهند بود. این رابطه را می‌شود رابطهٔ تعلق نام‌گذاری کرد زیرا هر عکس تنها به یک کلمه تعلق دارد.

نتیجه گیری

تاکنون در رابطه با تشخیص تصویری نوشتار زبان کردی و تدوین یک پایگاه داده تصویری که یکی از اجزای اساسی هر سیستم تشخیصی تصویری کلمات است تحقیقات ناچیزی انجام شده است. در این مقاله تلاش بر آن بوده تا گامی جدّی برای جبران این کاستی‌ها برداشته شود و پایگاه داده اختصاصی برای زبان کردی طراحی شود.

بدین ترتیب پس از جمع‌آوری تعداد قابل قبولی از لغات کردی در حدود بیش از ۵۵۰۰۰ لغت براساس فرهنگ لغت‌های معتبر و جداکردن بیش از ۲۰۰۰ کلمه از میان آن‌ها این کلمات بر روی فرم‌های استاندارد پیاده‌سازی شد تا برای نسخه‌برداری لغات دستنویس آماده گردند. پس از اتمام کار نسخه‌برداری، توسط افراد متعددی که هر یک تعداد معینی فرم دریافت کرده بودند، با بهره‌گیری از یک برنامه کاربردی که از پیش برای استخراج تصویر کلمات دستنویس از فرم‌های استاندارد طراحی شده بود، تصاویر تفکیک شده واژه‌هااستخراج شده و وارد پایگاه داده‌ها شدند. این برنامه کاربردی تمامی خصوصیات موجودیت عکس را پوشش داده و از طریق بود، تصاویر تفکیک شده واژه‌ها استخراج آن‌ها از فرم‌های استاندارد فراهم می‌آورد. بدین ترتیب با تطبیق و ایجاد رابطهٔ یک به چند موسوم به تعلّق مابین موجودیت کلمه و موجودیت عکس کل پایگاه داده‌ها شکل گرفت. این پایگاه داده کمک مؤثری در زمینهٔ طراحی سیستم‌های تشخیص تصویری واژه‌های زبان کردی است که تاکنون مشکلات فراوانی در طراحی آن‌ها وجود داشته است. در حال حاضر از این پایگاه داده در طراحی سیستم تشخیص نوری کلمات و در مقالات [۱۲] و [۱۳] استفاده شده است.

پیوست

در تصاویر ۹ و ۱۰ برنامه کاربردی که برای استنساخ کلمات دستنویس طراحی شده است و نحوهٔ کارکرد آن به نمایش گذاشته شده است.

جدول ۲. نمونه‌ای از جدول موجودیت عکس

نام کلمه	فایل عکس	خط کرسی	کیفیت خط کرسی	کیفیت عکس	نام نویسنده
ئەوراز	آدرس فایل	۶۰[1]، ۴۵[1]	درست	درست	X
ئاسپایی	"	۶۲ ، ۵۶	خوب	درست	X
ئاستش جیا	"	۵۹ ، ۴۲	خطا	خوب	X
ئاگرخانه	"	۵۵ ، ۵۰	خوب	درست	X
ئابرا	"	۶۵ ، ۳۹	خوب	درست	X
باوه قورەت	"	۶۱ ، ۵۲	خطا	درست	X
بازوبەن	"	۶۷ ، ۵۷	درست	درست	X

شکل ۹. صفحه اول برنامه کاربردی

شکل ۱۰. در این قسمت از برنامهٔ کاربردی، ویرایش‌های مربوط به خصوصیات موجودیت عکس انجام می‌گیرد.

منابع

1.Pechwitz, M., Maddouri, S. S., Märgner, V., Ellouze, N., & Amiri, H. (2002). IFN/ENIT-database of handwritten Arabic words. In Proc. of CIFED (Vol. 2, pp. 127-136).

2.AlKhateeb, J. H., Ren, J., Jiang, J., & Al-Muhtaseb, H. (2011). Offline handwritten Arabic cursive text recognition using Hidden Markov Models and re-ranking.Pattern Recognition Letters, 32(8), 1081-1088.

3.AlKhateeb, J. H., Pauplin, O., Ren, J., & Jiang, J. (2011). Performance of hidden Markov model and dynamic Bayesian network classifiers on handwritten Arabic word recognition. Knowledge-Based Systems, 24(5), 680-688.

4.Mozaffari, S., El Abed, H., Margner, V., Faez, K., & Amirshahi, A. (2008). IfN/Farsi-database: A database of farsi handwritten city names. In International

5.Conference on Frontiers in Handwriting Recognition.

6.Mozaffari, S., Faez, K., Faradji, F., Ziaratban, M., & Golzan, S. M. (2006). A comprehensive isolated Farsi/Arabic character database for handwritten OCR research. In Tenth International Workshop on Frontiers in Handwriting Recognition.

7.Solimanpour, F., Sadri, J., & Suen, C. Y. (2006). Standard databases for recognition of handwritten digits, numerical strings, legal amounts, letters and dates in Farsi language.

In Tenth International Workshop on Frontiers in Handwriting Recognition.

۸.بصیر علاقه بند. فاطمه دانشفر و وریا فتحی، راهکار متاهیوریستیک برای تشخیص خط زمینه در سیستمهای تشخیص نوری حروف در زبآن‌های دارای رسم الخط بهم چسبیده، یازدهمین کنفرانس سیستم‌های هوشمند ایران، ۱۳۹۱

9.El-Hajj, R., & likforman-Sulem, L., & Mokbe, C., (2005). Arabic Handwriting Recognition Using Baseline Dependant Features and Hidden Markov Modeling, Eight International Conference on Document Analysis and Recognition, IEEE: 20(5). 893-897.

10.Latfi, F. Nader, F & Mouldi, B., (2006). Arabic word recognition by using fuzzy classifier, Journal of Applied Scinces. 3, 650-617.

11.Nawaz, S.N., & Sarfraz, M., & Zidouri, A., & Al-Khatib, W.G., (2003). An approach to offline Arabic character recognition using neural networks, Paper presented at the 10th IEEE International Conference on Electronics, Circuits and Systems.

12.AlKhateeb, J., & Ren, J., & Jiang, J., & Al-Muhtaseb, H., (2011), Offline handwritten Arabic cursive text recognition using Hidden Markov Models and re-ranking, Pattern Recognition Letters (Elsevier).32, 1081-1088.

۱۳.فاطمه دانشفر، بصیر علاقه بند، وریا فتحی، مروری بر روش‌های موجودِ تخمین خط زمینه در زبآن‌های با رسم¬الخط پیوسته و ارائه الگوریتمی جدید، مجله علمی پژوهشی رایانش نرم و فناوری اطلاعات، جلد ۳ شماره ۱، ۴۴-۳۴.

14.F. Daneshfar, W. Fathy and B. Alaqeband, (2015), A Metaheuristic Algorithm for OCR Baseline Detection of Arabic Languages, accepted to be published as a book chapter at, Artificial Intelligent Algorithms and Techniques for Handling Uncertainties: Theory and Practice