نخستین کنفرانس بین المللی پردازش خط و زبان فارسی
۱۵ و ۱۶ شهریور ۱۳۹۱
دانشگاه سمنان - دانشکده مهندسی برق و کامپیوتر

مجموعه داده‌های برخط حروف تنهای کردی و فارسی

بشیر فتوحی[۱]، احسان‌اله کبیر[۲]
[۱] دانشکده مهندسی برق و کامپیوتر، دانشگاه تربیت مدرس bashir.futouhi@modares.ac.ir
[۲] دانشکده مهندسی برق و کامپیوتر، دانشگاه تربیت مدرس kabir@modares.ac.ir

چکیده

در این مقاله دو مجموعه داده برخط حروف الفبای کردی و فارسی ارائه می شود. تا کنون هیچ مجموعه داده مناسبی برای نوشتار کردی ارائه نشده است. این مقاله مجموعه داده‌ای ۱۰۰ نفره با تنوع بالا برای حروف کردی و مجموعه داده‌ای ۲۰۰ نفره برای حروف فارسی در اختیار قرار می دهد. در آینده تعداد مجموعه نخست نیز به ۲۰۰ افزایش خواهد یافت. این مجموعه داده‌ها دارای نرخ نمونه‌برداری بالا و تنوع سنی-جنسی مناسب هستند. تفاوت دیگر این مجموعه‌ها با موارد پیشین، استفاده از سیستم پایش و جمع‌آوری داده Labview است که منجر به مشاهده دقیق‌تر داده‌های معیوب و حذف آنها شده است. در کنار موارد فوق، استفاده از سیستم‌های مدرن نوشتاری منجر به ساخت مجموعه داده‌ای دقیق‌تر و طبیعی‌تر شده است. این دو مجموعه‌داده نخستین گام کارآمد در راستای ایجاد سیستم‌های بازشناسی برخط نوشتار کردی و فارسی خواهند بود. محاسبات آماری بر روی این دو مجموعه داده اطلاعات مناسبی را مانند درصد به‌کارگیری نوع خاصی از نقطه گذاری، جهت نوشتن و تشابه گروه حروف با بدنه یکسان در اختیار قرار می‌دهد.

کلمات کلیدی

مجموعه داده، برخط، حروف، فارسی، کردی، بازشناسی،

۱- مقدمه

بازشناسی دستنوشته یکی از زمینه‌های مورد توجه و بسیار کاربردی در گستره سیستم‌های بازشناسی الگو است. بازشناسی دست‌نوشته منجر به سهولت کاربری ادوات چند رسانه‌ای، امکان کاهش مصرف بی رویه کاغذ و... می‌شود. راحت‌تر بودن نوشتن از تایپ کردن، عدم امکان تایپ در ادوات کوچک، تعداد بالای حروف در برخی زبان‌ها و... از مشکلاتی هستند که به کمک این روش بازشناسی برطرف خواهند شد.

در زمینه بازشناسی برخط دست‌نوشته فعالیت‌های گسترده‌ای صورت گرفته است. با توجه به نحوه دریافت داده‌ها دو راهکار برخط (Online) و برون‌خط (Offline) جهت بازشناسی دستنوشته‌ها وجود دارد. وجود اطلاعاتی از نقاط حرکتی ، توالی و فشار قلم منجر شده است تا در کنار استفاده از روش بازشناسی برون‌خط، روش برخط نیز کاربرد گسترده‌ای داشته باشد. گاهی ممکن است تلفیق این دو روش منجر به بهبود کارایی سیستم بازشناسی شود. در این میان روش بازشناسی برخط از مدت‌ها پیش بر روی نوشتار و حروف زبان‌های گوناگون بررسی و امکان سنجی شده است. به دلیل تنوع کمتر و گستردگی کاربرد و نیاز، بهترین نتایج بازشناسی بر حروف لاتین مشاهده می شود [۱-۲]. تحقیقاتی در زمینه بازشناسی حروف چینی، عربی و فارسی نیز انجام شده است [۳-۱۲]. روش‌های ساده‌ای برای دسته‌بندی و تقسیم حروف و زیرکلمات فارسی به گروه‌های گوناگون نیز مشاهده می شود [۱۳-۱٤]. در این میان اثری از بازشناسی حروف کردی مشاهده نمی شود. یکی از دلایل این موضوع فقدان مجموعه‌داده برخط مناسب است.

جمع آوری داده‌های اولیه همواره یکی از گام‌های بنیادی و مهم در طراحی سیستم‌های بازشناسی است. نوشتار به عنوان یک ویژگی متنوع در بین انسان‌ها، تابعی از خلقیات، سن، جنسیت و... است. بر این اساس وجود مجموعه‌داده مناسب در این مورد اهمیت بیشتری دارد. در مواردی، گوناگونی نوشتار چندان بازشناسی را تحت تاثیر قرار می‌دهد که مجموعه حروف خاصی به عنوان مجموعه داده انتخاب می شود؛ گاهی شکل این حروف از شکل معمول آنها بسیار دور است [۱۵].

تا کنون هیچ مجموعه داده برخطی از حروف کردی در اختیار نبوده است. در زمینه نوشتار فارسی نیز دو مجموعه داده برخط با عنوان " حروف برخط دانشگاه تربیت مدرس" و" زیر-کلمات برخط دانشگاه تربیت مدرس " وجود دارد. این دو مجموعه‌داده در کنار تمام مزایای خود به دلیل نا مناسب بودن سیستم نوشتار