بررسی شیوه‌های متداول نگارش دست‌نوشته‌های برخط فارسی به منظور استفاده در بازشناسی آن‌ها

وحید قدس[۱] - احسان‌اله کبیر[۲]

[۱] دانشجوی دکتری، دانشگاه آزاد اسلامی واحد علوم و تحقیقات، تهران، ایران Email: v.ghods@semnaniau.ac.ir

[۲] دانشیار، دانشگاه تربیت مدرس، تهران، ایران Email: kabir@modares.ac.ir

چکیده: در بازشناسی دست‌نوشته‌ها به صورت برخط (یا برون‌خط)، آگاهی از انواع نوشتن حروف در کلمه‌ها ضروری است. در این مقاله، پایگاه داده برخط دانشگاه تربیت مدرس را انتخاب کرده‌ایم -که شامل حدود ۱۰،۰۰۰ نمونه از ۱،۰۰۰ زیرکلمه پرکاربرد فارسی است- و گروه‌بندی‌های متفاوت را از نظر بدنه اصلی و اجزای کوچک زیرکلمه‌ها معرفی کرده‌ایم. در بخش دیگر مقاله، به بررسی انواع دستخط می‌پردازیم و مهم‌ترین شکل‌های متفاوت نوشتن حروف را پیدا می‌کنیم. سپس با توجه به آن‌ها، زیرکلمه‌ها را گروه‌بندی می‌نماییم. بررسی فراوانی حروفی که به چند شکل نوشته می‌شوند و تحلیل آن‌ها از مباحث نهایی این مقاله است. نتایج اولیه بهبود موثر دقت بازشناسی را با انجام گروه‌بندی بر اساس انواع نوشتن حروف، نشان می‌دهد.

واژه‌های کلیدی: پایگاه داده، دست‌نوشته برخط، دستخط، فارسی، بازشناسی.

A study on usual styles of Farsi online handwriting in order to recognition

Vahid Ghods [1], Ehsanollah Kabir [2]

[1] Science and Research Branch, Islamic Azad University, Tehran, Iran. v.ghods@semnaniau.ac.ir

Tarbiat Modarres University, Tehran, Iran. kabir@modares.ac.ir

Abstract: Knowledge of the styles of writing letters in the word is necessary in recognition of online (or offline) handwriting. In this paper, we have chosen Tarbiat Modarres University (TMU) dataset which includes about 10,000 samples from the 1000 Farsi useful subwords. Different grouping based on the main body and the small signs of the subwords have been introduced. Diverse styles of Farsi handwriting were investigated and the most important forms of letters writing were found. Then, the subwords were classified according to their styles. Frequency of letters that are written in different forms was discussed and analyzed. The results showed effective improvement in recognition rate using grouping based on the writing styles of letters.

Keywords: Dataset, Online handwriting, Writing styles, Farsi, Recognition.

تاریخ ارسال مقاله: ۱۳۸۹.۰۸.۰۱

تاریخ اصلاح مقاله: ۱۳۹۱/۲/۱۰

تاریخ پذیرش مقاله: ۱۳۹۱/۲/۱۲

نام نویسنده‌ی مسئول : وحید قدس

نشانی نویسنده‌ی مسئول: ایران– تهران– حصارک– دانشگاه آزاد اسلامی واحد علوم و تحقیقات – دانشکده مهندسی برق و کامپیوتر

۱- مقدمه

با توجه به نحوه دریافت اطلاعات، بازشناسی دست‌نوشته به صورت برون‌خط، offline، و برخط، online، صورت می‌پذیرد. در زمینه بازشناسی برون‌خط دست‌نوشته عربی و فارسی تحقیقات زیادی صورت پذیرفته است [۱، ۲، ۱۳، ۱۴]، ولی در زمینه بازشناسی برخط دست‌نوشته در این دو زبان تحقیقات کمتری انجام شده است [۹-۳].

در بازشناسی برخط، مختصات نقاط مسیر حرکت قلم، تعداد حرکات قلم و فشار قلم در دسترس هستند. بازشناسی برخط نوشتار به دلیل راحت‌تر بودن نوشتن از تایپ کردن، عدم امکان تایپ در بعضی از موقعیت‌ها، عدم وجود یک صفحه کلید کامل روی کامپیوترهای کوچک و سخت بودن تایپ حروف در بعضی زبان‌ها به دلیل تعداد زیاد حروف آن‌ها، مورد توجه خاصی قرار گرفته است.

در [۱۵]، برای بازشناسی برخط کلمات دست نویس عربی از یک پایگاه داده ۴۰۰ کلمه‌ای استفاده شده است. نرخ بازشناسی در میان این ۴۰۰ کلمه ۸۰٪ گزارش شده است.

در تحقیق‌های دیگری از مدل مارکوف مخفی[۳، ۱۰]، برنامه نویسی پویا و استفاده از شبکه عصبی برای بازشناسی حروف و ارقام فارسی استفاده شده است. در [۱۶] با استفاده از بازشناسی فازی، بازشناسی کلمات فارسی برخط انجام شده است. در [۱۱]، دو روش ساده و کارا برای آشکارسازی دایره‌ی حروف به منظور گروه‌بندی آن‌ها ارائه می‌شود. این روش‌ها برای حروف تنها که به صورت بر خط نوشته شده باشند، مناسب هستند.

در [۱۷]، گروه‌بندی و بازشناسی حروف تنهای فارسی که به صورت برخط نوشته شده باشند، با استفاده از استخراج ویژگی‌های ساختاری آن‌ها ارائه شده است. در این روش، بدنه‌های اصلی حروف بر اساس شکل و ساختار نوشتاری آن‌ها به ۹ گروه تقسیم می‌شوند. پس از انتخاب و استخراج ویژگی‌ها، گروه‌بندی با استفاده از درخت تصمیم انجام می‌شود. در [۱۲]، حروف مجزای فارسی با یک الگوریتم سلسله مراتبی تکه بندی می‌شوند و با یک روش فازی بازشناسی می‌شوند. دقت سیستم برای نمونه‌های آموزشی ۹۳/۴٪ و برای نمونه‌های آزمایش ۹۰/۲٪ است. پایگاه مورد استفاده برای آموزش و آزمون، مرجع [۷] است. نمونه‌های ۷۰ نویسنده اول برای آموزش و نمونه‌های دیگر نویسنده‌ها (۵۴ نویسنده) برای آزمون به کار گرفته شده است. در یکی از آخرین تحقیقات در زبان فارسی ابتدا طبقه بندی بر اساس نقاط و سرکش‌ها و علائم کوچک انجام می‌شود و سپس در هر گروه با استفاده از بدنه اصلی، شناسایی حرف و زیرکلمه صورت می‌گیرد [۹-۶].

در سال ۲۰۰۷ میلادی، یک پایگاه داده بزرگ از ارقام دست نویس برون‌خط فارسی ارائه شده است [۱۸]. در آن مقاله تنوع نوشتن ارقام فارسی بر روی بالغ بر ۱۰۰،۰۰۰ نمونه ارقام بررسی شده است. در ارتباط با پایگاه داده دست‌نوشته برخط فارسی اطلاعات محدودی وجود دارد. مهم‌ترین پایگاه داده ارائه شده مرجع [۷]، پایگاه داده برخط دانشگاه تربیت مدرس، است که در سال ۱۳۸۳ تولید شد.

در این مقاله به بررسی شیوه نگارش فارسی و تنوع دست‌نوشته‌های افراد مختلف در پایگاه داده [۷] می‌پردازیم تا از نتایج حاصل در بازشناسی دست‌نوشته‌های برخط استفاده کنیم. در بخش دوم مقاله به معرفی پایگاه داده پرداخته می‌شود. در بخش سوم گروه‌بندی بر اساس بدنه اصلی معرفی شده است. در بخش سوم، گروه‌بندی بر اساس اجزای کوچک ارائه شده است. بخش پنجم و ششم به معرفی انواع دستخط فارسی و بررسی و تحلیل آن می‌پردازد. در بخش هفتم از مقاله نتیجه‌گیری شده است.

۲- معرفی پایگاه داده

پایگاه داده [۷]، برای جمع‌آوری کلمات رایج فارسی از مطالب شش سال روزنامه همشهری و یک سال روزنامه کیهان استفاده کرده است. تعداد کل کلمات استخراج شده، ۳۱۳،۲۲۵ کلمه است. در ادامه، نویسنده ۲۹،۷۳۹ کلمه را که تعداد تکرار آن‌ها بیش از ۳۰ است انتخاب کرده است. تعداد کلماتی که تکرار آن‌ها کمتر از ۳۰ است حدود پنج میلیون و بیش از ۳۰، حدود یک میلیارد است (تکرارها به صورت جداگانه محاسبه شده است). در نتیجه، احتمال رخداد کلمه‌های انتخاب شده، بالای ۹۹ درصد است. هر کلمه از چند حرف تشکیل می‌شود. در زبان فارسی برخی از حروف یک کلمه به هم می‌چسبند. حروفی که در یک کلمه به هم می‌چسبند، زیرکلمه (Subword) را تشکیل می‌دهند. مثلاً، {فا، ر، سی} زیرکلمه‌های کلمه "فارسی" هستند. زیرکلمه‌ها جدا از هم نوشته می‌شوند؛ ولی ممکن است بعضی زیرکلمه‌ها در موقع نوشتن به هم بچسبند که در نوع برخط آن این مسئله وجود ندارد. از ۲۹،۷۳۹ کلمه انتخاب شده، ۷،۳۱۷ زیرکلمه استخراج شد. در مرحله جمع‌آوری داده، ۱،۰۰۰ زیرکلمه‌ای که تعداد تکرار بیشتری داشته‌اند، انتخاب شده است [۷] که بالای ۹۹ درصد از ۲۹،۷۳۹ کلمه و ۹۸ درصد از کل پایگاه داده، از آن‌ها ساخته می‌شود.. به طور خلاصه این پایگاه داده، دارای ۱،۰۰۰ زیرکلمه پر کاربرد در زبان فارسی است. هر زیرکلمه دارای چندین

نمونه دستخط است و در مجموع حدود ۱۰،۰۰۰ نمونه از این زیرکلمات وجود دارد که توسط افراد مختلف نوشته شده است. ۱۲۴ نفر در جمع‌آوری دست‌نوشته‌ها شرکت داشته‌اند که حدود نیمی از آن‌ها دانشجوی کاردانی، حدود نیمی دیگر دانشجوی کارشناسی و ۱۰ نفر دانشجوی کارشناسی ارشد و دکتری رشته مهندسی برق بوده‌اند. هر نفر حدود ۱۰۰ زیرکلمه از مجموعه ۱،۰۰۰ زیرکلمه پر تکرار را نوشته است. از هر زیرکلمه به طور متوسط ۱۰ نمونه جمع‌آوری شده است. برای جمع‌آوری داده‌ها از قلم و صفحه WACOM GRAPHIRE استفاده شده است. نویسندگان ۹۸ نفر مرد و ۲۶ نفر زن هستند. از این افراد فقط ۸ نفر چپ دست و بقیه راست دست هستند. در این مجموعه، اطلاعات فشار قلم وجود ندارد.

۳- گروه‌بندی بر اساس بدنه اصلی

یک روش که برای بازشناسی دست‌نوشته برخط پیشنهاد می‌شود این است که در گام اول گروه‌های دارای بدنه اصلی یکسان بازشناسی شوند. به این منظور فعالیت دیگری که بر روی پایگاه داده ۱،۰۰۰ زیرکلمه انجام شد، گروه‌بندی این زیرکلمات بر اساس بدنه اصلی است. بدین ترتیب زیرکلماتی که دارای بدنه اصلی یکسان هستند، بدون در نظر گرفتن سایر حرکات قلم آن‌ها و اجزای کوچک مربوط به آن‌ها در یک گروه قرار می‌گیرند. برای مثال زیرکلمات "تیر، تبر، نیر، پیر، ببر" در یک گروه هستند. نتیجه انجام این گروه‌بندی بر روی پایگاه داده، منجر به تشکیل ۶۴۰ گروه شد که در آن زیرکلمات هر گروه دارای بدنه‌های یکسان هستند. فراوانی این گروه‌بندی در جدول (۱) گزارش شده است.

جدول (۱): فراوانی گروه‌ها از نظر بدنه‌های یکسان

N	۱	۲	۳	۴	۵	۶	۷	۹	۱۲
فراوانی(M)	۴۷۵	۱۱۲	۲۴	۱۹	۸	۷	۳	۲	۲

(N: تعداد زیرکلمات یکسان در گروه، M: تعداد گروه‌هایی دارای N زیرکلمه یکسان)

برای مثال، منظور از تعداد زیرکلمات N=5، گروه‌هایی هستند که در آن‌ها ۵ کلاس زیرکلمه با بدنه اصلی یکسان وجود دارد که ۸ گروه از این نوع به دست آمده است و منظور از تعداد زیرکلمات N=1، گروه‌هایی هستند که زیرکلمات آن‌ها با هیچ زیرکلمه دیگری، در پایگاه داده ۱،۰۰۰ زیرکلمه‌ای، بدنه یکسان ندارند. لازم به ذکر است که از هر زیرکلمه در هر گروه چندین نمونه گرفته شده است.

بدیهی است زیرکلماتی که با یک حرکت قلم نوشته شده‌اند و دارای اجزای کوچک مانند نقطه یا سرکش نیستند، با هیچ بدنه اصلی دیگری یکسان نخواهند بود؛ زیرا، معنی آن تکراری بودن زیرکلمه در پایگاه داده است. تغییر جایگاه حرکات مربوط به اجزای کوچک زیرکلمات و یا نبود آن‌ها، موجب تشکیل یک گروه با بدنه‌های اصلی یکسان می‌شود.

یک آزمایش اولیه برای بازشناسی گروه‌های با بدنه یکسان انجام دادیم و از طبقه‌بند مدل مخفی مارکوف (HMM) استفاده کردیم. از آنجایی که مدل مخفی مارکوف نیاز به داده‌های زیاد برای آموزش هر مدل دارد از گروه‌هایی که دارای N>=۴ است، استفاده کردیم که شامل ۴۲ گروه شد. پس از آموزش با استفاده از ۷۰٪ داده‌ها، مدل HMM پیوسته تمام بدنه‌های اصلی مربوط به زیر کلمه‌ها به دست آمد. همه مدل‌های HMM با تعداد حالات (State) برابر ۱۰ و تعداد مخلوط گاوسی (Mixture) برابر ۱۰ در نظر گرفته شد. در مرحله آزمون با استفاده از ۳۰٪ باقی مانده داده‌ها، دقت بازشناسی گروه‌ها ۷۱٪ به دست آمد.

۴- گروه‌بندی بر اساس اجزای کوچک

از نکات مهم تمایز دستخط فارسی با لاتین، وجود علائم و نقاط در حرکت‌های قلم بعد از نوشتن بدنه اصلی زیرکلمات است. در این تحقیق به بخش‌هایی از حرف که به بدنه اصلی اضافه می‌شود؛ یعنی بخش‌هایی که در حرکت دوم قلم و حرکت‌های بعد از آن شکل می‌گیرد، اجزای کوچک حرف گوییم (مانند نقاط و سرکش). این وجه تمایز، ایده‌های جدید را در گروه‌بندی و یا بازشناسی زیرکلمات و کلمات فارسی ایجاد می‌کند. در [۹]، دقت بازشناسی با رعایت چند قانون در نوشتن اجزای کوچک به بالای ۷۰ درصد در زیرکلمات رسیده است. ۱۱ علامت مهم در فارسی وجود دارد که در جدول (۲) آورده شده است. با توجه به حروف به کار برده شده در ۱،۰۰۰ زیرکلمه پایگاه داده، فراوانی ۱۱ علامت فارسی در این پایگاه داده را به دست آوردیم (جدول (۳)).

جدول (۲): علائم مهم در فارسی

ردیف (کد علامت)	علامت	حروف مرتبط
۱	یک نقطه بالا	خ – ذ – ز – ض – ظ – غ – ف - ن
۲	یک نقطه پایین	ب – ج
۳	دو نقطه بالا	ت – ق
۴	دو نقطه پایین	ی (فقط در جایگاه اول و وسط)-(ی-ـی)
۵	سه نقطه بالا	ث – ژ – ش
۶	سه نقطه پایین	پ – چ
۷	سرکش	ک
۸	دو سرکش	گ
۹	دسته "ط" و "ظ"	ط – ظ
۱۰	همزه	أ – (ئ، ـئ) - ؤ
۱۱	کلاه "الف"	آ

جدول (۳): فراوانی علائم فارسی در پایگاه داده

کد علامت	فراوانی	درصد(٪)
۱	۴۲۷	۲۷/۸
۲	۱۸۱	۱۱/۸
۳	۳۰۳	۱۹/۷
۴	۲۴۰	۱۵/۶
۵	۱۲۲	۷/۹۶
۶	۵۶	۳/۶۶
۷	۹۷	۶/۳۳
۸	۴۴	۲/۸۷
۹	۴۸	۳/۱۳
۱۰	۱۳	۰/۸۵
۱۱	۱	۰/۰۶

برای نمونه در ۱،۰۰۰ زیرکلمه مورد نظر، ۱۲۲ حرف که سه نقطه بالا دارند به دست آمد که ۷/۹۶ درصد علائم را در پایگاه داده تشکیل داده‌اند. ضمناً علائم پایین حروف، فقط یک نقطه، دو نقطه و سه نقطه هستند و بقیه علائم بالای حروف هستند. نکته دیگری که در این ۱،۰۰۰ زیرکلمه بررسی شد و از نظر گروه‌بندی حائز اهمیت است، تقسیم این زیرکلمات به ۴ گروه بدون علامت، تمام علائم در بالا، تمام علائم در پایین و علائم در بالا و پایین است. نتایج این گروه‌بندی در جدول (۴) است.

جدول (۴): فراوانی توزیع علائم فارسی در پایگاه داده

ردیف	گروه	فراوانی
۱	بدون علامت	۱۳۴
۲	تمام علائم در بالا	۴۵۳
۳	تمام علائم در پایین	۱۵۱
۴	علائم در بالا و پایین	۲۶۲

اگر پایگاه داده را به نوعی گروه‌بندی کنیم که زیرکلمات دارای علائم یکسان در یک گروه قرار گیرند، ۱،۰۰۰ زیرکلمه، ۱۷۰ گروه می‌شوند. در هر گروه زیرکلماتی قرار می‌گیرند که علائم آن‌ها از نظر ترتیب و شکل یکسان هستند ؛ مثلاً، زیرکلمات "بیمه، بیر، بیما، بیا، بسیا" یک گروه می‌شوند و طبق جدول (۲) کد علامت (۴-۲) را می‌گیرند. تمام زیرکلمات بدون علامت را در یک گروه قرار داده‌ایم که تعداد آن‌ها ۱۳۴ است. بالاترین تعداد اعضای این ۱۷۰ گروه، ۱۰۵ است که مربوط به زیرکلمات با "یک نقطه بالا" است. سایر گروه‌های پر عضو در جدول (۵) دیده می‌شود. همچنین در جدول (۶) فراوانی گروه‌های کم عضو آورده شده است. زیرکلماتی نظیر "نچه" با کد علامت (۶-۱) یا "کمیته" با کد علامت (۳-۴-۷) با هیچ زیرکلمه دیگری مشابهت علائم ندارند و تک عضوی هستند.

جدول (۵): گروه‌های پر عضو از نظر علائم

ردیف	گروه (با علائم)	تعداد عضو
۱	یک نقطه بالا (۱)	۱۰۵
۲	دو نقطه بالا (۳)	۷۱
۳	یک نقطه پایین (۲)	۵۹
۴	دو نقطه پایین (۴)	۴۵
۵	یک نقطه بالا- دو نقطه بالا (۳-۱)	۲۹

جدول (۶): فراوانی گروه‌های کم عضو از نظر علائم

گروه	فراوانی
تک عضوی	۸۰
دو عضوی	۳۵
سه عضوی	۱۰

۵-انواع دستخط فارسی

دستخط فارسی در قرون متمادی تغییرات عمده‌ای کرده است. عموم این تغییرات، دستخط را به سمت ساده‌تر و سریع‌تر نوشتن پیش برده است. اما تمام این تغییرات از قوانین اصلی نگارش الفبای فارسی خارج نشده است و تنها ترتیب نوشتن اجزای حروف و یا بخشی از شکل حروف به صورت با قاعده تغییر کرده است. بیشتر این تغییرات در حروفی که در وسط بدنه اصلی زیرکلمه هستند حادث شده است و علت آن را می‌توان میل به نوشتن بدنه اصلی زیرکلمات در یک حرکت قلم و در نتیجه سرعت بیشتر نگارش دانست.

جهت بازشناسی دستخط فارسی خصوصاً به صورت برخط، ناگزیر به بررسی انواع شیوه‌های نگارش فارسی زبانان و شناخت تنوع آن‌ها هستیم. برای نمونه، این تغییر در شکل نوشتن حروف "ح" و "ک" اتفاق

می‌افتد، به طوری که نویسنده زمانی که می‌خواهد در یک زیرکلمه یک حرف را به "ح" یا گروه هم‌شکل آن بچسباند، این حرف به این شکل () در می‌آید. تقریباً شیوه نگارش غالب، زمانی که "ح" با یک حرکت قلم در وسط یا آخر زیرکلمه می‌آید، به این شکل است. همچنین اگر نویسنده بخواهد در یک زیرکلمه یک حرف را به حرف "ک" بچسباند و بعد از "ک"، حرفی باشد که حرکت به سمت بالا دارد(یعنی حروف "ا"، "ل"، "ک")، از آنجایی که در"ک" حرکت عمودی به سمت بالا و سپس به سمت پایین کمی مشکل و وقت گیر است برای راحتی و تسریع در نوشتن، نویسنده حرکت رفت و برگشت عمودی در "ک" را به یک حرکت دایره در جهت ساعت‌گرد تبدیل می‌کند که به این شکل ( ) در می‌آید.

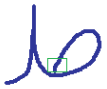
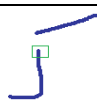
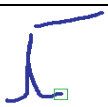
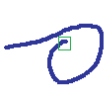
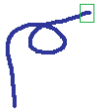
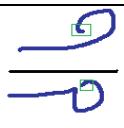
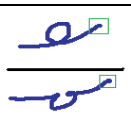
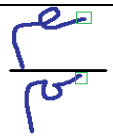
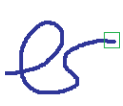
در حرف "ط"، عرف نوشتن بر این است که دسته "ط" در حرکت دیگری از قلم نوشته شود. برخی از افراد ترتیب نوشتن اجزای این حرف را تغییر می‌دهند. در یک حالت اگر "ط" در ابتدای زیرکلمه باشد، از دسته آن شروع به نوشتن کرده و سپس بیضی آن را در همان حرکت قرار می‌دهند (). در حالت دیگر در هر مکانی از زیرکلمه ابتدا بیضی آن را نوشته و سپس در همان حرکت قلم دسته آن را در یک حرکت رفت و برگشت عمودی قرار می‌دهند( ).

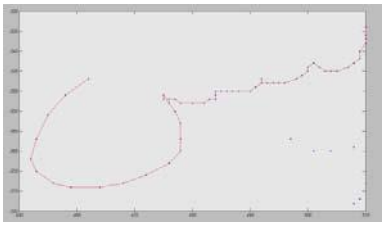
در حروف دیگر مانند "م"، "ه"، "س"، "ع" و خانواده‌های هم‌شکل آن‌ها نیز با توجه به جایگاه این حروف در زیرکلمه (اول، وسط، آخر) ترتیب اجزای این حروف و یا شکل نوشتن آن‌ها در افراد مختلف، متفاوت است. در جدول (۷)، مهم‌ترین حروفی که دارای تنوع نگارش هستند، آورده شده است. نشانه □، نقطه شروع را نشان می‌دهد.

لازم به ذکر است می‌توان برای دقت بیشتر برای حرف "م" سه نوع نوشتن در نظر گرفت. معیار ما از در نظر گرفتن دو نوع، حرکت پادساعت‌گرد (نوع ۱) و ساعت‌گرد (نوع ۲) دایره حرف "م" است. در این مرحله، به پالایش و اصلاح اطلاعات ۱،۰۰۰ کلاس زیرکلمه در پایگاه داده پرداختیم. این اصلاحات به قرار زیر است:

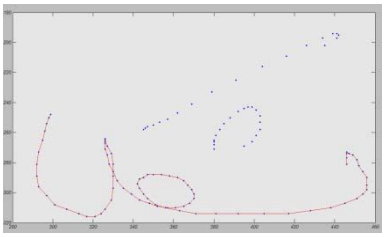
در برخی از کلاس‌ها، نمونه‌هایی از کلاس‌های دیگر یا نمونه‌های نامربوط به اشتباه وجود داشت که حذف شدند. در برخی دیگر از کلاس‌ها، نمونه‌هایی وجود داشتند که از نظر املایی یا ساختاری غلط بودند که این نمونه‌ها نیز از مجموعه داده‌ها حذف شدند. در شکل (۱) چند نمونه از این اشکال‌ها آورده شده است.

جدول (۷): تنوع نگارش در دستخط فارسی

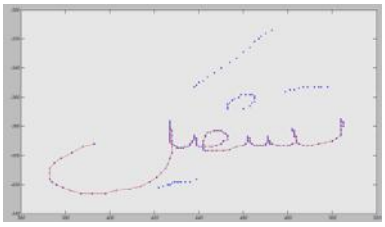
حرف	اول زیرکلمه	وسط زیرکلمه	آخر زیرکلمه
س(ش) - نوع ۱			
س(ش) - نوع ۲			
ط(ظ) - نوع ۱			
ط(ظ) - نوع ۲		-	-
ط(ظ) - نوع ۳			
ع(غ) - نوع ۱			
ع(غ) - نوع ۲	-		
ک(گ) - نوع ۱			
ک(گ) - نوع ۲			-
م - نوع ۱			
م - نوع ۲			
ه - نوع ۱			
ه - نوع ۲	-		



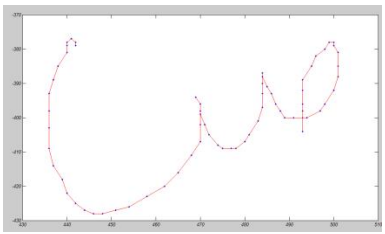
الف) "پس" - یک دندانه اضافه دارد.



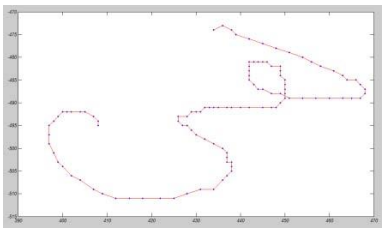
ب) "شکا" - به اشتباه "شکلا" نوشته شده است.



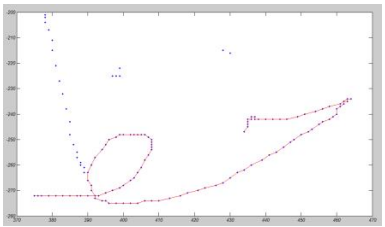
ج) "تشکیل" - از "ک" نوع (۲) استفاده شده است، در حالی که پس از "ک"، حروف ("ا"، "ل") نیامده است.



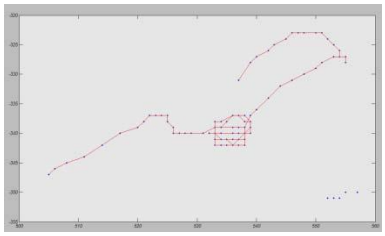
د) "صا" - یک دندانه اضافه دارد.



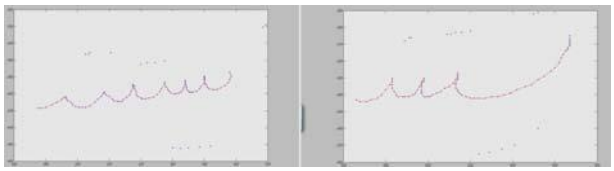
ه) "هن" - نقطه "ن" در همان حرکت اول قلم نوشته شده است.



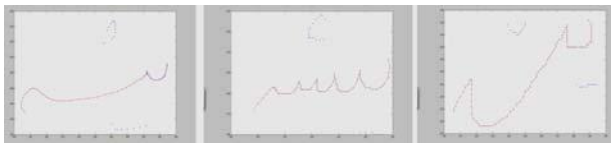
و) "خط" - به اشتباه "خظ" نوشته شده است.



ز) "جمه" - حرف "م" چند دور چرخیده است.



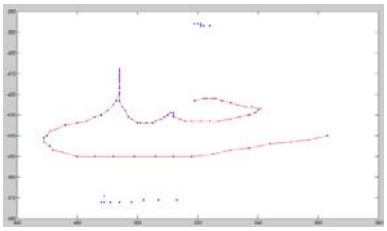
ح) "نیستند" - یک دندانه کم دارد. (هر دو مورد)



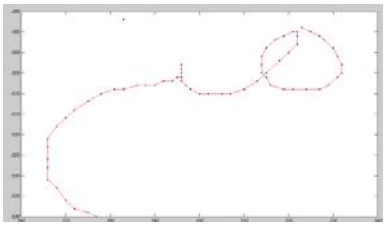
ط) "یشه" - یک دندانه اضافه دارد. (هر سه مورد)

شکل (۱): نمونه‌های اشتباه از نظر املایی یا ساختاری

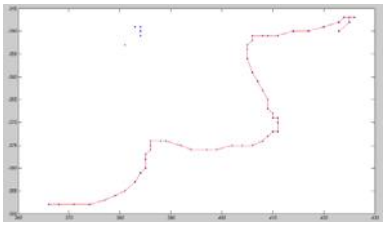
همچنین چند مورد از نمونه‌هایی را که حروف در کلمه به شکلی متفاوت از آنچه در جدول (۷) آمده است، در شکل (۲) نشان داده‌ایم. به دلیل استفاده اندک در دستخط فارسی، این شکل‌های نوشتن را در نظر نگرفته‌ایم و از مجموعه داده‌ها حذف شده است. لازم به ذکر است که برای مواردی نظیر شکل (۱-ه) و شکل (۲-الف، ب، ج، د، ه، و) می‌توان نوع جدید نوشتن حرف را در جدول (۷) تعریف و اضافه کرد. مثلاً در شکل (۲-الف)، حرف "ی" آخر که برگشت به عقب دارد را به عنوان "ی" آخر نوع (۲) تعریف کرد.



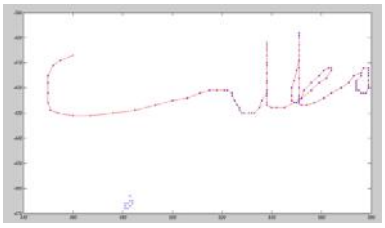
الف) "خیلی" - حرف "ی" آخر، برگشت به عقب دارد.



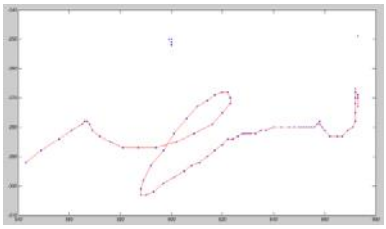
ب) "هن" – حرف "ن" آخر، برگشت به عقب دارد.



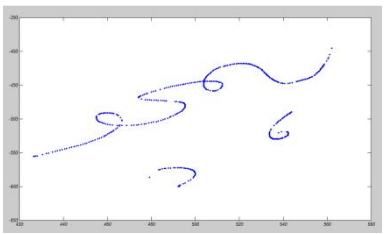
ج) "کز" – سرکش "ک" اول، در همان حرکت اول قلم نوشته شده است.



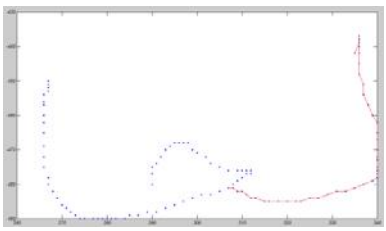
د) "مطلب" – از "ط" نوع (۲) در وسط کلمه استفاده شده است. یعنی در یک حرکت بالا و سپس پایین قلم ابتدا دسته "ط" و سپس بیضی آن نوشته شده است.



ه) "نسخه" – حرف "خ" بطور پیوسته و به سمت بالا نوشته شده است.



و) "همچو" – از "ه" به شکل مد استفاده شده است.



ز) "لحا" – "حا" در حرکت دوم قلم نوشته شده است.

شکل (۲): نمونه‌های نوشتن حروف به شکلی متفاوت با جدول (۷)

از آنجایی که نویسندگان این مجموعه داده، زیرکلمات را بدون هیچ قیدی نوشته‌اند، در اکثر کلاس‌هایی که زیرکلمه مربوط به آن‌ها، دارای حداقل یکی از حروف جدول (۷) است، مخلوطی از انواع نوشتن آن حرف وجود دارد. برای مثال کلمه "پسر" دارای دو مدل دستخط یکی با "س" دندانه‌دار (نوع ۱) و دیگری با "س" کشیده (نوع ۲) است. حتی در برخی از زیرکلمات مانند "سعه" پنج مدل نوشتن دیده می‌شود که علت آن ترکیب انواع مختلف نوشتن حروف تشکیل دهنده آن است که در جدول (۸) قابل مشاهده است.

جدول (۸): انواع مختلف نوشتن "سعه"

مدل نوشتن	س	ع	ه
۱	نوع۱	نوع۱	نوع۱
۲	نوع۲	نوع۱	نوع۱
۳	نوع۱	نوع۲	نوع۱
۴	نوع۱	نوع۲	نوع۲
۵	نوع۲	نوع۲	نوع۱

در نتیجه در مرحله بعد به اصلاح کلاس‌های پایگاه داده و افزایش مدل دستخط زیرکلمات پرداختیم و به دلایلی که در فوق ذکر شد مجبور به انجام دستی این کار شدیم.

با انجام گروه‌بندی جدید، ۱،۰۰۰ کلاس قبلی به ۱۷۱۱ کلاس اصلاح شده جدید تبدیل شد. سه آزمایش با مجموعه داده اصلاح شده (TMUII) انجام شد که نتایج آن‌ها با مجموعه داده قبلی (TMUI) در جدول (۹) مقایسه شده است. در آزمایش اول، گروه‌بندی را بر اساس بدنه اصلی بر روی ۴۲ گروه منتخب با شرایطی که در بخش ۳ ذکر شد، اجرا کردیم و دقت بازشناسی از ۷۱٪ به ۸۰٪ ارتقا یافت. ویژگی‌های در نظر گرفته شده عبارتند از:

x(i), y(i), Theta, Delta_Theta, Sin(Theta), Cos(Theta), Sin(Delta_Theta), Cos(Delta_Theta), Vx, Vy.

در آزمایش دیگری، از شبکه عصبی MLP سه لایه با الگوریتم یادگیری پس انتشار خطا برای گروه‌بندی ۴۲ گروه منتخب استفاده گردید. برای این شبکه‌ها، ۳۱ ورودی که شامل ۳۰ ویژگی و یک ورودی بایاس است در نظر گرفته شد. چارچوب هر زیرکلمه به اندازه ۱۵×۱۵ نرمالیزه شده است و در طول مسیر حرکت قلم ۱۵ نقطه متساوی‌الفاصله در نظر گرفته شد. مختصات این نقاط به عنوان ۳۰ ویژگی فرض شده است. آموزش شبکه‌های عصبی با مجموعه داده قبلی و اصلاح شده انجام شد. دقت بازشناسی برای مجموعه آزمون از ۶۹٪ در مجموعه قبلی به ۸۳٪ در مجموعه اصلاح‌شده افزایش یافت.

در یک آزمایش دیگر، گروه‌بندی مجموعه داده کامل را با مدل مخفی مارکوف و روش آموزش جاسازی شده (Embedded Training) [19]، قبل و بعد از اصلاح مجموعه داده انجام دادیم. دقت گروه‌بندی بر اساس بدنه اصلی از ۴۲٪ به ۶۰٪ در مجموعه داده اصلاح شده افزایش

یافت. در تمام آزمایش‌ها، ۷۰٪ داده‌ها برای آموزش و مابقی برای مرحله آزمون به کار گرفته شده است.

۶- بحث و بررسی در انواع دستخط

کار دیگری که بر روی حدود ۱۰،۰۰۰ دستخط از پایگاه داده ۱،۰۰۰ زیرکلمه‌ای انجام شد، بررسی انواع دستخط حروفی است که به شکل‌های متفاوت با توجه به جدول (۷) نوشته می‌شوند که در جدول (۱۰) آمده است.

جدول (۹): مقایسه دقت گروه بندی بین TMU I و TMU II

مجموعه داده	روش	TMU I (%)	TMU II (%)
۴۲ گروه	شبکه عصبی	۶۹	۸۳
۴۲ گروه	HMM	۷۱	۸۰
کامل	HMM (Embedded Training)	۴۲	۶۰

جدول (۱۰): مقایسه انواع دستخط حروف هم‌خانواده

ردیف	حرف	نوع	فراوانی در اول	فراوانی در وسط	فراوانی در آخر	مقایسه۱	مقایسه۲
						تعداد کل (درصد)	
۱	س (ش)	۱	۴۱۰ (۳۰۳)	۳۲۹ (۳۳۱)	۸۷ (۱۰۹)	۱۵۶۹ (٪۷۳/۷)	
۲	س (ش)	۲	۲۰۳ (۱۲۱)	۱۲۹ (۷۵)	۱۳ (۲۰)	۵۶۱ (٪۲۶/۳)	
						تعداد کل در جایگاه اول (درصد)	**تعداد کل در نوع ۱ و ۳ (درصد)**
۳	ط (ظ)	۱	۶۱ (۳۰)	۷۲ (۳۹)	۴۸ (۱۶)	۹۱ (٪۶۲/۳)	۲۶۶ (٪۷۳/۷)
۴	ط (ظ)	۲	۱۲ (۶)	تعریف نشده	تعریف نشده	۱۸ (٪۱۲/۳)	-
۵	ط (ظ)	۳	۲۸ (۹)	۴۵ (۱۰)	۰ (۳)	۳۷ (٪۲۵/۴)	۹۵ (٪۲۶/۳)
						تعداد کل در جایگاه وسط و آخر (درصد)	
۶	ع (غ)	۱	۳۹۷ (۶۸)	۱۲۷ (۸)	۳۸ (۰)	۱۷۳ (۳۶/۸)	
۷	ع (غ)	۲	تعریف نشده	۲۰۹ (۲۶)	۶۲ (۰)	۲۹۷ (۶۳/۲)	
						تعداد قبل از (ا،ل) در جایگاه اول و وسط (درصد)	
۸	ک (گ)	۱	۴۴۷ (۱۵۵)	۳۳۰ (۲۰۴)	۷۵ (۳۰)	-	
	ک (گ) [قبل از (ا،ل)]	۱	۲۰ (۹)	۶۲ (۴۹)	-	۱۴۰ (٪۵۲/۴)	
۹	ک (گ)	۲	۲۴ (۱۳)	۶۰ (۳۰)	تعریف نشده	۱۲۷ (٪۴۷/۶)	
						تعداد کل (درصد)	
۱۰	م	۱	۷۲۴	۵۴۱	۱۰۵	۱۳۷۰ (٪۶۸/۲)	
۱۱	م	۲	۳۶۲	۱۳۳	۱۴۴	۶۳۹ (٪۳۱/۸)	
						درصد در جایگاه وسط	**درصد در جایگاه آخر**
۱۲	ه	۱	۳۸۰	۱۵۱	۶۷۷	٪۳۳/۳	٪۷۷/۶
۱۳	ه	۲	تعریف نشده	۳۰۲	۱۹۵	٪۶۶/۷	٪۲۲/۴

جدول (۱۱): کدهای توصیفی حروف

حرف	جدا	اول	وسط	آخر
آ	۱	-	-	-
ا	۲	-	-	۳
ب	۴	۵	۶	۷
پ	۸	۹	۱۰	۱۱
ت	۱۲	۱۳	۱۴	۱۵
ث	۱۶	۱۷	۱۸	۱۹
ج	۲۰	۲۱	۲۲	۲۳
چ	۲۴	۲۵	۲۶	۲۷
ح	۲۸	۲۹	۳۰	۳۱
خ	۳۲	۳۳	۳۴	۳۵
د	۳۶	-	-	۳۷
ذ	۳۸	-	-	۳۹
ر	۴۰	-	-	۴۱
ز	۴۲	-	-	۴۳
ژ	۴۴	-	-	۴۵
س	۴۶	۴۷	۴۸	۴۹
ش	۵۰	۵۱	۵۲	۵۳
ص	۵۴	۵۵	۵۶	۵۷
ض	۵۸	۵۹	۶۰	۶۱
ط	۶۲	۶۳	۶۴	۶۵
ظ	۶۶	۶۷	۶۸	۶۹
ع	۷۰	۷۱	۷۲	۷۳
غ	۷۴	۷۵	۷۶	۷۷
ف	۷۸	۷۹	۸۰	۸۱
ق	۸۲	۸۳	۸۴	۸۵
ک	۸۶	۸۷	۸۸	۸۹
گ	۹۰	۹۱	۹۲	۹۳
ل	۹۴	۹۵	۹۶	۹۷
م	۹۸	۹۹	۱۰۰	۱۰۱
ن	۱۰۲	۱۰۳	۱۰۴	۱۰۵
و	۱۰۶	۱۰۶	۱۰۷	۱۰۷
ه	۱۰۸	۱۰۹	۱۱۰	۱۱۱
ی	۱۱۲	۱۱۳	۱۱۴	۱۱۵
همزه(ء، ئ، ؤ، أ)	۱۱۶	۱۱۷	۱۱۸	۱۱۹
س (نوع ۲)	۱۲۰	۱۲۱	۱۲۲	۱۲۳
ش (نوع ۲)	۱۲۴	۱۲۵	۱۲۶	۱۲۷
ط (نوع ۲)	۱۲۸	۱۲۹	-	-
ظ (نوع ۲)	۱۳۰	۱۳۱	-	-
ط (نوع ۳)	۱۳۲	۱۳۳	۱۳۴	۱۳۵
ظ (نوع ۳)	۱۳۶	۱۳۷	۱۳۸	۱۳۹
ع (نوع ۲)	۱۴۰	۱۴۱	۱۴۲	۱۴۳
غ (نوع ۲)	۱۴۴	۱۴۵	۱۴۶	۱۴۷
ک (نوع ۲)	۱۴۸	۱۴۹	۱۵۰	۱۵۱
گ (نوع ۲)	۱۵۲	۱۵۳	۱۵۴	۱۵۵
م (نوع ۲)	۱۵۶	۱۵۷	۱۵۸	۱۵۹
ه (نوع ۲)	۱۶۰	۱۶۱	۱۶۲	۱۶۳

در جدول (۱۰)، اعداد داخل پرانتز ستون‌های "فراوانی" مربوط به حرف دومی است که در ستون "حرف" و داخل پرانتز قرار دارند. نوع مقایسه در حروف مختلف با توجه به بدنه اصلی و جایگاه حرف تفاوت دارد. در نتیجه ستون‌های "مقایسه" متغیر است. این ستون‌ها به صورت تعداد رخداد و درصد آن (که داخل پرانتز است) نشان داده شده است.

در پایگاه داده هر کلاس دارای یک توصیفگر مربوط به خود است. توصیفگر زیرکلمات از کنار هم قرار گرفتن کد مربوط به حروف آن تشکیل شده است. توصیفگرها تنها بیانگر ترتیب جایگاه حروف در زیرکلمات هستند و حاوی اطلاعات دیگری مانند نقاط جداسازی حروف در زیرکلمه نیستند. اکثر حروف فارسی دارای ۴ کد توصیفی متفاوت هستند که مربوط به جایگاه آن حرف در زیرکلمه هستند. بدین ترتیب تعداد کدهای توصیفگر ۱۱۹عدد به دست آمد. سپس، برای حروفی که در جدول (۷) آمده و به صورتی غیر از دستخط نوع (۱) نوشته شده است، کدهای توصیفی جدید اضافه شد که به این ترتیب تعداد کل کدهای توصیفی به ۱۶۳ رسید. در جدول (۱۱) نحوه تخصیص کدهای توصیفی نشان داده شده است. کدهای توصیفی ۱۲۰ و بعد از آن مربوط به دستخط‌های غیر از نوع (۱) است. برای حرف "ح" و خانواده آن برای حالت "جدا" و "اول" شکل رسمی ولی برای حالت "وسط" و "آخر" شکل ( ) در نظر گرفته شده است.

۶-۱- تحلیل جدول (۱۰)

در تمام تحلیل‌های زیر، مقایسه بر روی بدنه‌های اصلی حروف صورت گرفته است. در ردیف ۱ تعداد حروف "س" و "ش" از دستخط نوع (۱) به تفکیک در جایگاه‌های اول، دوم و سوم آورده شده است و در ردیف ۲ تعداد این حروف از نوع دستخط (۲) آمده است. در مجموع، در مورد بدنه اصلی "س" و "ش" ۷۳/۷٪ افراد دارای دستخط نوع (۱) و ۲۶/۳ درصد آن‌ها دارای دستخط نوع (۲) بوده‌اند (با توجه به جدول (۷)).

از آنجایی که حروف "ط" و "ظ" از نوع دستخط (۲) فقط در جایگاه اول قرار دارد، دو مقایسه بین ردیف‌های ۳، ۴ و ۵ (انواع دستخط حروف "ط" و "ظ") انجام شده است. در جایگاه اول نوشتن حروف "ط" و "ظ"،

۶۲/۳ درصد نویسندگان از دستخط نوع (۱)، ۱۲/۳ درصد از دستخط نوع (۲) و ۲۵/۴ درصد، از دستخط نوع (۳) استفاده می‌کنند. مقایسه دیگر، بین ردیف ۳ و ۵ است و در مجموع جایگاه‌های اول، وسط و آخر، ۷۳/۷ درصد نویسندگان دارای دستخط نوع (۱) و ۲۶/۳ درصد دارای دستخط نوع (۳) بوده‌اند

حرف "ع" در جایگاه اول کلمه دارای یک نوع دستخط است. بنابراین در ردیف ۶ و ۷ (مربوط به حروف "ع" و "غ") جایگاه‌های وسط و آخر را مقایسه کرده‌ایم. قابل توجه است که ۶۳/۲ درصد نویسندگان این مجموعه داده دستخط نوع (۲) را به کار می‌برند () و تنها ۳۶/۸ درصد از دستخط نوع (۱) استفاده کرده‌اند. به نظر می‌رسد علت این امر میل به سرعت زیاد و حرکت به جلوی قلم باشد.

در مورد ردیف ۸ و ۹ یعنی بدنه اصلی حروف "ک" و "گ" همان‌طوری که قبلاً ذکر شد دستخط نوع (۲) این حروف تنها در صورتی که پس از آن‌ها حروف (ا ، ـا ، ـل) باشند، احتمال نوشتن دارد و همچنین در اول یا وسط زیرکلمه رخ می‌دهد. به همین منظور و برای مقایسه درست، تعداد حروف ردیف ۸ را که پس از آن‌ها حروف (ا ، ـا ، ـل) بودند و در جایگاه اول و یا وسط قرار داشتند، جدا کردیم. نتیجه آن که ۴۷/۶ درصد افراد از دستخط نوع (۲) (به شکل شبه‌دایره) استفاده می‌کنند.

نتیجه دیگر اینکه ۶۸/۲ درصد نویسندگان برای نوشتن حرف "م" در جهت پاد ساعت‌گرد و ۳۱/۸ درصد از افراد از حرکت ساعت‌گرد برای نوشتن گردی "م" استفاده کرده‌اند.

با توجه به تنوع نوشتن حرف "ه" در جایگاه‌های مختلف کلمه، دستخط‌های مختلف آن را در جایگاه‌های اول، وسط و آخر به صورت جداگانه بررسی کرده‌ایم. در اول زیرکلمه، غالب نویسندگان مجموعه داده از دستخط نوع (۱) استفاده می‌کنند. در جایگاه وسط ۶۶/۷ درصد از دستخط نوع (۲) بهره می‌برند و در جایگاه آخر ۷۷/۶ درصد افراد از دستخط نوع (۱) در نوشتن حرف "ه" استفاده می‌کنند.

۷- نتیجه‌گیری

در این مقاله به معرفی و بررسی گروه‌بندی‌های مفید جهت استفاده در بازشناسی دست‌نوشته‌های فارسی پرداختیم و نتیجه‌های اولیه بازشناسی با استفاده از گروه‌بندی‌ها ذکر شد که بهبود موثر را نشان می‌دهد. همچنین بررسی شیوه نگارش و انواع نوشتن زیرکلمه‌های فارسی را انجام دادیم. جهت بازشناسی دست‌نوشته‌های فارسی مخصوصاً به صورت برخط، ناگزیر به بررسی انواع نوشتن حروف در کلمه‌ها هستیم. مهم‌ترین حروفی را که در کلمه‌ها دارای بیش از یک نوع نوشتن هستند، بیان کردیم و فراوانی آن‌ها را در پایگاه داده مورد نظر بررسی کردیم. سپس اصلاح و افزایش گروه‌های پایگاه داده برای شکل‌های جدید حروف زیرکلمه‌ها انجام شد. محققان محترم می‌توانند با ارسال نامه الکترونیکی به آدرس نویسندگان از جزئیات بخش‌های این مقاله و دریافت فایل گروه‌بندی‌های مختلف پایگاه داده بهره‌مند شوند.

۸- مراجع

[۱] ر .عزمی، بازشناسی متون چاپی فارسی، رساله دکتری مهندسی برق، دانشگاه تربیت مدرس، تابستان ۱۳۷۸.

[۲] م .شیرعلی شهرضا، تشخیص کلمات و ارقام دست‌نویس فارسی بوسیلهٔ شبکه‌های عصبی، رساله دکتری، دانشکده برق، دانشگاه صنعتی امیرکبیر، ۱۳۷۴.

[۳] ه. ساجدی، م. جم زاد ، ح. ثامتی و ب. باباعلی، "ارائه یک روش مبتنی بر گروه‌بندی برای بازشناسی حروف مجزای برخط فارسی به کمک مدل مخفی مارکوف"، دوازدهمین کنفرانس بین‌المللی انجمن کامپیوتر ایران، صص ۴۲۵-۴۱۹، ۱۳۸۵.

[۴] ه .مظفری، ف .رهگذر و ش .شریف ، "تشخیص برخط ارقام دست‌نویس فارسی"، دومین کنفرانس سالانه انجمن کامپیوتر ایران، دانشگاه صنعتی امیرکبیر، تهران، صص ۲۰۴-۱۹۶، ۱۳۷۵.

[۵] ک .عباسیان و ا .کبیر، "بازشناسی برخط نویسه های فارسی" ، ششمین کنفرانس مهندسی برق ایران، دانشگاه خواجه نصیرالدین طوسی ، تهران، صص ۱۴۶-۱۴۱، ۱۳۷۷.

[۶] س.م. رضوی، بازشناسی دست نوشته برخط فارسی، رساله دکتری مهندسی برق ,دانشگاه تربیت مدرس، ۱۳۸۵.

[۷] س .م .رضوی و ا .کبیر، "یک پایگاه داده برای بازشناسی دست‌نوشته های برخط فارسی"، ششمین کنفرانس سیستم‌های هوشمند، کرمان، ۱۳۸۳.

[۸] س .م .رضوی و ا .کبیر، "روشی ساده برای بازشناسی برخط حروف مجزای فارسی"، مجله دانشکده مهندسی دانشگاه فردوسی مشهد، ۱۳۸۴.

[۹] س .م .رضوی و ا .کبیر، "روشی ساده برای بازشناسی برخط زیر - کلمات فارسی" ، نشریه مهندسی برق و مهندسی کامپیوتر ایران، سال ۲، شماره۲، صص ۷۲-۶۳، ۱۳۸۴.

[۱۰] م. ا. مهرعلیان، ک. فولادی، "بازشناسی برخط حروف مجزای دست‌نویس فارسی بر اساس تشخیص گروه اصلی بدنه با استفاده از مدل مخفی مارکوف"، پانزدهمین کنفرانس سالانه انجمن کامپیوتر ایران، تهران، ۱۳۸۸.

[۱۱] و. قدس، ا. کبیر، "گروه‌بندی حروف برخط با آشکارسازی کاسه آن‌ها"، پانزدهمین کنفرانس سالانه انجمن کامپیوتر ایران، تهران، ۱۳۸۸.

[۱۲] م. سلیمانی باغشاه، س. باقری شورکی و ش. کسائی، "بازشناسی و یادگیری حروف مجزای برخط فارسی به روش فازی"، چهاردهمین کنفرانس بین‌المللی مهندسی برق ایران، دانشگاه صنعتی امیرکبیر ، تهران، ۱۳۸۵.

[13] M. Dehghan, K. Faez, M. Ahmadi and M. Shridhar, "Handwritten Farsi (Arabic) Word Recognition: A Holistic Approach Using Discrete HMM", Pattern Recognition, Vol. 34, pp. 1057-1065, 2001.

[14] R. Plamondon, N. Srihari, "On-line and Off-line Handwriting Recognition: A Comprehensive Survey", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No. 1, pp. 63-84, 2000.

[15] A. Amin, "Machine Recognition of Handwritten Arabic Words by the IRAC II System", Proc. 6th Int. Conf. on Pattern Recognition. Munich, Germany, pp. 34-36, 1982.

[16] R. Halavati and S. Bagheri Shouraki, "Recognition of Persian Online Handwriting Using Elastic Fuzzy Pattern Recognition", International Journal of Pattern Recognition and Artificial Intelligence, Vol. 21, No. 12, pp. 491-513, 2007.

[17] V. Ghods, E. Kabir, "Feature Extraction for Online Farsi Characters", 12th Int. Conf. on Frontiers in Handwriting Recognition (ICFHR), India, 2010.

[18] H. Khosravi, E. Kabir, "Introducing a very large dataset of handwritten Farsi digits and a study on their varieties", Pattern Recognition Letters, 28 (10), pp. 1133-1141, 2007.

[19] L R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition", Proceedings of the IEEE, 77: 257–285, 1989.