بازشناسی دست‌نوشته برون‌خط عربی بر مبنای یک رویکرد تلفیقی جدید از مدل مخفی مارکوف و شبکه‌های عصبی ژرف

باقر باباعلی* **و بابک رکابدار**

دانشکده ریاضی، آمار و علوم رایانه، دانشگاه تهران، تهران، ایران

چکیده

مسأله مدل‌سازی و بازشناسی دست‌نوشته شباهت بسیار زیادی به مسأله مدل‌سازی و بازشناسی گفتار دارد. به همین علت می‌توان از رویکردهای به‌کار گرفته‌شده برای مسأله بازشناسی گفتار با اندکی تغییر در مراحل ابتدایی آن مانند استخراج ویژگی، برای بازشناسی دست‌نوشته نیز بهره برد. با گسترش رویکردهای ترکیبی HMM-DNN و استفاده از توابع هدف دنباله‌ای مانند MMI پیشرفت‌های قابل توجهی در حوزه بازشناسی گفتار حاصل شده است. این مقاله با استفاده از نرم‌افزار متن‌باز KALDI، که شهرت اصلی آن در حوزه بازشناسی گفتار و همچنین به‌کارگیری آخرین مدل‌های ترکیبی ارائه‌شده در آن، به‌کمک روش افزایش داده مدلی برای بازشناسی دست‌نوشته عربی ارائه داده است. این پژوهش بر روی دادگان KHATT انجام شده که نرخ خطای بازشناسی واژه را بر روی این دادگان به میزان ۷/۳۲ درصد مطلق کاهش داده است.

واژگان کلیدی: بازشناسی دست‌نوشته عربی، شبکه‌های عصبی ژرف، مدل مخفی مارکوف، نرم‌افزار متن‌باز KALDI

Off-line Arabic Handwritten Recognition Using a Novel Hybrid HMM-DNN Model

Bagher BabaAli* & Babak Rekabdar

School of Mathematics, Statistics and Computer Sciences, College of Science, University of Tehran, Tehran, Iran

Abstract

In order to facilitate the entry of data into the computer and its digitalization, automatic recognition of printed texts and manuscripts is one of the considerable aid to many applications. Research on automatic document recognition started decades ago with the recognition of isolated digits and letters, and today, due to advancements in machine learning methods, efforts are being made to identify a sequence of handwritten words. Generally, based on the type of text, document recognition is divided into two main categories: printed and handwritten. Due to the limited number of fonts relative to the diversity of handwriting of different writers, it is much easier to recognize printed texts than handwritten text; thus, the technology of recognizing printed texts has matured and has been marketed in the form of a product. Handwritting recognition task is usually done in two ways: online and offline; offline handwriting recognition involves the automated translation of text in image format to letters that can be used in computer and text-processing applications. Most of the research in the field of handwriting recognition has been conducted on Latin script, and a variety of tools and resources have been gathered for this script. This article focuses on the application of the latest methods in the field of speech recognition for the recognition of Arabic handwriting. The task of handwritten text modeling and recognizing is very similar to the task of speech modeling and recognition. For this reason, it is possible to apply the approaches used for the speech recognition with a slight change for the handwriting recognition. With the expansion of HMM-DNN hybrid approaches and

* Corresponding author

*نویسندهٔ عهده‌دار مکاتبات

• تاریخ ارسال مقاله: ۱۳۹۷/۱۱/۲۸ • تاریخ پذیرش: ۱۳۹۸/۰۶/۱۱ • تاریخ انتشار: ۱۳۹۹/۱۲/۰۴ • نوع مطالعه: کاربردی

the use of sequential objective functions such as MMI, significant improvements have been made in the accuracy of speech recognition system. This paper presents a pipeline for the offline Arabic handwritten text recognition using the open source KALDI toolkit, which is very well-known in the community of speech recognition, as well as the use of the latest hybrid models presented in it and data augmentation techniques. This research has been conducted on the Arabic KHATT database, which achieved 7.32% absolute reduction in word recognition error (WER) rate.

Keywords: Arabic Handwritten Recognition, Deep Neural Networks, Hidden Markov Model, Kaldi Toolkit

۱- مقدمه

در راستای تسهیل ورود اطلاعات به رایانه و دیجیتال‌سازی آن، بازشناسی خودکار متون چاپی و دست‌نوشته کمک بزرگی محسوب می‌شود که کاربردهای متعددی نظیر دسته‌بندی خودکار نامه‌ها، خواندن خودکار چک‌های بانکی و فرم‌های پرشده دستی، تبدیل متون دست‌نوشته تاریخی و صدها کاربرد دیگر برای آن قابل تصور است. پژوهش‌ها بر روی بازشناسی ماشینی متون از چند دهه قبل، ابتدا با بازشناسی ارقام و حروف محدود و مجزا آغاز شد و امروزه به لطف پیشرفت‌های علمی حاصل شده در حوزه یادگیری ماشین و بازشناسی الگو به تلاش در جهت بازشناسی دنباله‌ای از کلمات دست‌نوشته پیوسته منجر شده است. بازشناسی متون بر مبنای نوع متن به دو صورت چاپی[1] و دست‌نوشته[2] تقسیم‌بندی می‌شود. بنا به محدودبودن تعداد فونت‌های چاپی در مقایسه با تنوع دست‌خط افراد مختلف، بازشناسی متون چاپی در مقایسه با دست‌نوشته به مراتب راحت‌تر است؛ به نحوی که فناوری تشخیص متون چاپی به بلوغ رسیده و در قالب محصول به بازار عرضه شده است. تشخیص دست‌نوشته به‌طورمعمول به دو صورت برخط[3] و برون‌خط[4] انجام می‌شود. در حالت برخط، کاربر بر روی یک لوح فشرده شروع به نوشتن می‌کند و مختصات نوک قلم با سرعت ثابتی (به‌طورمعمول یکصد تا دویست هرتز) نمونه‌برداری[5] شده و بر مبنای این دنباله از مختصات نوک قلم، بازشناسی دست‌نوشته صورت می‌گیرد؛ ولی در حالت برون‌خط، ورودی تصویری است که از دست‌نوشته گرفته شده و از اطلاعات پویا حین نوشتن بی‌بهره است. به‌قدری این اطلاعات پویا در بهبود دقت بازشناسی مؤثر است که در [11] تلاش در جهت بازسازی آن از روی تصویر صورت گرفته است. در حالت برون‌خط تصویر دست‌نوشته ممکن است شامل خطوط متعددی (به‌عنوان مثال یک پاراگراف) باشد که در این صورت، به‌طورمعمول لازم است در ابتدا قطعه‌بندی خطوط[6] صورت گرفته و در ادامه بازشناسی تک‌تک خطوط به‌طور مجزا صورت گیرد. کیفیت تصاویر حاصل از پویش دست‌نوشته‌ها، تنوع بسیار بالای دست‌نوشته افراد مختلف، فاصله‌های ایجادشده بین کلمات و یا حتی بخش‌هایی از یک کلمه و همچنین تنوع در اندازه و نوع قلم را می‌توان از جمله عواملی دانست که بازشناسی دست‌نوشته برون‌خط را به یکی از چالش‌برانگیزترین پژوهش‌ها در حوزه تحلیل و بازشناسی متون تبدیل کرده است.

(شکل-۱): بازشناسی دست‌نوشته به دو صورت برخط و برون‌خط
(Figure-1): Offline and online handwritten recognition

خطوط نگارشی زنده و متداول در دنیا شامل لاتین، سیریلیک، عربی، هندی، گرجی و ارمنی است که متداول‌ترین آن لاتین است و از همین رو عمده پژوهش‌ها بیشتر بر روی خط لاتین صورت گرفته و منابع و دادگان‌های متعددی در این زمینه گردآوری شده است. در این بین، پژوهش‌های به‌مراتب کمتری بر روی دست‌نوشته عربی و به‌طور کلی زبان‌هایی که از خط عربی بهره می‌برند، از جمله فارسی رایج در ایران انجام شده است. این پژوهش با هدف پرداختن به بازشناسی دنباله‌ای از کلمات دست‌نوشته پیوسته فارسی بر مبنای رویکردها و روش‌های نوین ارائه‌شده در سالیان اخیر آغاز شد، ولی به‌دلیل عدم وجود یک دادگان دست‌نوشته فارسی مناسب برای انجام آزمایش‌ها و همچنین قرابت زیاد خط فارسی به عربی، آزمایش‌ها بر مبنای خط عربی و به‌طور مشخص دادگان KHATT انجام شد.

[1] Printed
[2] Handwritten
[3] Online
[4] Offline
[5] Sampling
[6] Line Segmentation

در ادامه این مقاله، در بخش بعد به مروری اجمالی بر روی دادگان‌ها و پژوهش‌های انجام‌شده در حوزه بازشناسی دست‌نوشته فارسی و عربی می‌پردازیم. بخش سوم به بررسی نحوه مدل‌سازی بازشناسی دست‌نوشته و رویکردهای رایج آن اختصاص داده شده است. در بخش چهارم به تشریح اجزای سامانهٔ به‌کار گرفته‌شده در این پژوهش پرداخته‌ایم. نتایج آزمایش‌ها و بررسی و تحلیل آنها در بخش پنجم آمده و درنهایت در بخش ششم جمع‌بندی ارائه شده است.

۲- مرور اجمالی دادگان‌ها و پژوهش‌ها

غالب پژوهش‌های انجام‌شده در حوزه بازشناسی دست‌نوشته فارسی به بازشناسی ارقام و حروف مجزا محدود می‌شود. در [34] و [31] و [36] با استفاده از روش‌های مختلف استخراج ویژگی و همچنین بهره‌گیری از روش‌های دسته‌بندی ماشین بردار پشتیبان[1] و K-NN[2] تلاش‌هایی برای بازشناسی ارقام دست‌نوشته فارسی انجام و در [24] همین هدف با استفاده از تجزیه ساختاری[3] دنبال شده است؛ همچنین پژوهش انجام‌شده در [18] برای بازشناسی ارقام دست‌نوشته فارسی از روش‌های بینایی رایانه[4] و ویژگی‌های توصیفی تصویر مانند HOG[5] بهره برده است. اما در کنار این پژوهش‌ها، موارد محدودی نیز برای بازشناسی کلمات مجزا در زبان فارسی نیز گزارش شده است. [8] و [9] تلاش‌هایی جهت بازشناسی کلمه توسط مدل مخفی مارکوف بوده‌اند؛ همچنین [6] با استخراج بردار ویژگی از تصاویر کلمات فارسی توسط تبدیل ویولت به‌دنبال بازشناسی کلمات فارسی از روی تصاویر بوده است. در [43] یک روش دومرحله‌ای برای بازشناسی کلمات دست‌نوشته فارسی به‌کمک بلوک‌بندی تطبیقی گرادیان تصویر ارائه شده است. درهمین‌اواخر نیز [30] از شبکه‌های عصبی ژرف[6] برای بازشناسی کلمات دست‌نوشته مجزا استفاده کرده است.

تا جایی که مطلع‌ایم، هیچ پژوهشی روی بازشناسی دنباله‌ای از کلمات فارسی به‌صورت یک خط دست‌نوشته گزارش نشده است. همان‌طور که درقبل گفته شد، مهم‌ترین علت این امر نبود دادگان دست‌نوشته فارسی مناسب برای این منظور بوده است. در زبان عربی نیز تعداد دادگان برای این منظور بسیار محدود است. کمبود دادگان مناسب باعث محدودیت پژوهش در این حوزه در زبان عربی و فارسی شده است. برای رفع این معضل، تلاش‌هایی به‌منظور جمع‌آوری دادگان استاندارد برای بازشناسی دست‌نوشته فارسی و عربی صورت گرفته است. هر چند در این پژوهش هدف بررسی بازشناسی دنباله‌ای از کلمات دست‌نوشته است، ولی در ادامه به بررسی مهم‌ترین دادگان‌های موجود عربی و فارسی که در بردارنده حروف مجزا، ارقام مجزا و رشته‌ای، تاریخ، واژه مجزا و درنهایت دنباله کلمات (متن) هستند، می‌پردازیم.

یکی از آخرین تلاش‌ها برای جمع‌آوری دادگان فارسی توسط [32] انجام شده است. این دادگان از تصاویر ارقام، حروف، واژگان و متون دست‌نوشته فارسی تشکیل شده است. تعداد نویسندگان این دادگان پانصد نفر است که از ۲۵۰ نفر مرد و ۲۵۰ نفر زن تشکیل شده است. مشکل عمده این دادگان نبود قطعه‌بندی خطی برای متون نوشته‌شده در قالب پاراگراف و همچنین نبود متن متناظر با هر خط از دست‌نوشته است که باعث شده قابل به‌کارگیری در پژوهش جاری نباشد.

یکی از مشهورترین دادگان دست‌نوشته فارسی دادگان HODA [19] است. دادگان HODA شامل ارقام دست‌نویس صفر تا نه است و از تصاویر دست‌نوشته بیش از ۱۱۹۴۲ نفر تشکیل شده است. تعداد تصاویر این دادگان بالغ بر هشتادهزار تصویر است که شصت‌هزار تعداد از آنها به‌عنوان مجموعه آموزش و بیست‌هزار تعداد از آنها به‌عنوان مجموعه آزمون در نظر گرفته شده است. یکی دیگر از مهمترین دادگان‌های فارسی، دادگان IFN/Farsi [22] است. این دادگان از تصاویر اسامی شهرها و روستاهای ایران تشکیل شده و برای مواردی که هدف بازشناسی کلمات مجزای فارسی است، مناسب است. تعداد نویسندگان آن ششصد نفر و تعداد تصاویر آن ۷۲۷۱ تصویر است.

تنها دادگان فارسی که شامل تصاویر خطوط دست‌نوشته فارسی است، و یا به عبارتی از قطعه‌بندی خطی برخوردار است دادگان HaFT [33] است. این دادگان با همکاری ششصد نویسنده تهیه شده است. هر نویسنده، سه برگه را که هر کدام شامل هشت خط بوده، نوشته است. در این دادگان تصاویر هر خط به‌صورت جداگانه در دسترس است؛ اما به‌دلیل این که متن هر تصویر موجود نیست مشابه دادگان [32] نمی‌توان از آن برای بازشناسی دنباله کلمات دست‌نوشته پیوسته فارسی که هدف این مقاله است، بهره برد.

[1] Support Vector Machine
[2] K-Nearest neighbor
[3] Structural Decomposition
[4] Computer Vision
[5] Histogram of Oriented Gradients
[6] Deep Neural Network

از مرسوم‌ترین دادگان‌های عربی می‌توان به IFN/Enit [25] و KHATT [21] اشاره کرد که اولی شامل کلمات مجزا و دومی تصاویر خطوط دست‌نوشته است که هر خط شامل دنباله‌ای از واژگان دست‌نوشته است. دادگان IFN/Enit شامل ۲۶۴۵۹ تصویر دست‌نوشته از ۹۳۷ اسم شهر و روستا در کشور تونس تهیه شده است. تعداد نویسندگان این دادگان ۴۱۱ نفر است. دادگان KHATT شامل ۱۳۳۵۷ تصویر از خطوط دست‌نوشته این دادگان شامل دوهزار پاراگراف یکسان و دوهزار پاراگراف منحصربه‌فرد است. این ویژگی باعث می‌شود که بتوان از این دادگان برای مسائل دیگری مانند تشخیص نویسنده نیز بهره برد. تعداد هزار نویسنده برای جمع‌آوری این دادگان مشارکت داشته‌اند. دادگان KHATT تنها دادگان عربی در دسترس است که برخلاف دو دادگان فارسی موجود، هم قطعه‌بندی خطی و هم متن متناظر با هر خط را دارد، به همین دلیل این دادگان به‌عنوان تنها گزینه موجود برای انجام آزمایش‌های این پژوهش انتخاب شد. تعداد بیشتری از دادگان‌های حوزه بازشناسی دست‌نوشته فارسی و عربی در جدول (۱) ادامه است.

(جدول-۱): دادگان‌های متداول برای بازشناسی دست‌نوشته فارسی و عربی

(Table-1): Popular datasets for Persian/Arabic handwritten recognition

نام دادگان	زبان	نوع داده	تعداد نویسنده	تعداد تصاویر
Sadri2016 [32]	فارسی	ارقام، حروف، کلمات و متون دست‌نوشته	500	250624
HODA [19]	فارسی	ارقام دست‌نوشته	11942	80000
IFN/Farsi [22]	فارسی	اسامی شهرها و روستاهای ایران	600	7271
IIaFT [33]	فارسی	یک خط دست‌نوشته	600	14400
FHT [42]	فارسی	پاراگراف دست‌نوشته	250	1000
IAUT/PIICN [5]	فارسی	اسامی شهرهای ایران	380	34200
IFHCDB [23]	فارسی	حروف و ارقام دست‌نوشته	–	70120
CENPARMI [15]	فارسی	ارقام، حروف، کلمات و متون دست‌نوشته	295	41004
KHATT [21]	عربی	یک خط دست‌نوشته	1000	13357
IFN/Enit [25]	عربی	اسامی شهرها و روستاهای تونس	411	26459
MADCAT Arabic [20]	عربی	یک خط دست‌نوشته	305	750000

۳- مدل‌سازی بازشناسی دست‌نوشته

مسأله مدل‌سازی و بازشناسی دست‌نوشته شباهت بسیار زیادی به مسأله بازشناسی گفتار دارد. از همین رو سامانه‌های متن‌باز موجود در حوزه بازشناسی گفتار نظیر KALDI [26]، HTK [41] و RASR [29] به‌منظور بازشناسی دست‌نوشته نیز به‌کار گرفته شده‌اند. اغلب ابتدا روش‌ها برای مسأله بازشناسی گفتار توسعه داده می‌شوند و سپس برای حوزه بازشناسی دست‌نوشته اختصاصی می‌شوند. جدول (۲) اجزای و مسائل مشابه این دو حوزه بازشناسی را در هر سطر جدول ارائه داده است.

(جدول-۲): اجزای و مسائل مشابه در بازشناسی گفتار و بازشناسی دست‌نوشته

(Table-2): Similar issues in speech recognition and handwritten recognition

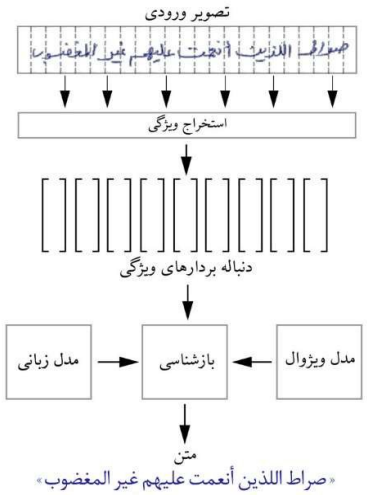
حوزه بازشناسی گفتار	حوزه بازشناسی دست‌نوشته
مدل‌سازی آکوستیک	مدل‌سازی ویژوال
مدل‌سازی زبانی	مدل‌سازی زبانی
مدل‌سازی طول واج	مدل‌سازی طول شناسه
مدل‌سازی بافت واج	مدل‌سازی بافت شناسه
تطبیق گوینده	تطبیق نویسنده

مشابه بازشناسی گفتار، در رویکرد آماری بازشناسی دست‌نوشته، با بیشینه‌سازی احتمال پسین رابطه (۱) و بهره‌گیری از قاعده بیز محتمل‌ترین دنباله واژگان (متن) به‌ازای تصویر دست‌نوشته ورودی به‌دست می‌آید:

$$W^* = \underset{W}{\operatorname{argmax}} P(W|X) = \underset{W}{\operatorname{argmax}} P(X|W)P(W) \quad (۱)$$

که $X = x_1 x_2 x_3 \dots x_T$ دنباله بردارهای ویژگی استخراج‌شده از دست‌نوشته و $W = w_1 w_2 \dots w_T$ دنباله واژگان متن است. در این رابطه $P(X|W)$ مدل ویژوال است که توزیع احتمالی دست‌نوشته به شرط متن را تخمین می‌زند و

P(W) مدل زبانی است که احتمال رخداد دنباله واژگان را تخمین می‌زند. در شکل (۲) نمای کلی یک سامانه بازشناسی دست‌نوشته پیوسته آمده است.



(شکل-۲): نمای کلی از سامانه بازشناسی دست‌نوشته
(Figure-2): Generic block diagram of the handwritten recognition systems

در مدل‌سازی ویژوال(P(X|W))، هدف برچسب‌زنی دنباله بردارهای ویژگی به طول T با یک دنباله به طول N از شناسه‌ها یا واژگان است که به‌طورمعمول N بسیار کوچک‌تر از T است. مدل مخفی مارکوف به‌دلیل سادگی و توان بالایی که در مدل‌سازی الگوهای متوالی[1] دارد، یکی از متداول‌ترین روش‌های مدل‌سازی ویژوال محسوب می‌شود. در این روش، به‌طورمعمول از مدل مخلوط گوسی[2] برای مدل‌سازی احتمال مشاهده بردارهای ویژگی در حالات مدل مخفی مارکوف استفاده می‌شود که دراصطلاح رویکرد HMM-GMM نامیده می‌شود.

از مهم‌ترین پژوهش‌هایی که با استفاده از رویکرد HMM-GMM انجام شده است، می‌توان به [16] اشاره کرد. در این پژوهش ابتدا کلمات متن با تجزیه مورفولوژیکی[3] به زیرکلمات[4] شکسته می‌شوند؛ بنابراین مجموعه‌واژگان[5] این سامانه ترکیبی از کلمات و زیرکلمات است که این سبب می‌شود تا با ترکیب زیرکلمات موجود در مجموعه‌واژگان بتوان هر کلمه‌ای را شناسایی کرد. این پژوهش از یک مدل زبانی n-gram که بر روی متن تجزیه‌شده آموزش داده شده است استفاده می‌کند. همچنین در [2] با استفاده از بردارهای ویژگی استخراج‌شده بر مبنای پیکسل از تصاویر دودویی دست‌نوشته و همچنین استفاده از ایده ترکیب[6] نتایج خروجی سه سامانه بازشناسی مبتنی بر HMM سامانه کارایی ارائه داده شده است. در [3] که هدف بازشناسی کلمات مجزا است دو نوع ویژگی از تصویر واژه استخراج می‌کند. ویژگی نخست بر مبنای پنجره‌ای که بر روی تصویر لغزانیده داده می‌شود، استخراج می‌شود و از آن برای آموزش مدل‌های HMM استفاده می‌کند. ویژگی دوم شامل ویژگی‌های ساختاری مانند تعداد زیرواژه و تعداد علامت‌ها از جمله نقطه است. در این سامانه ابتدا بازشناسی به‌وسیلهٔ مدل‌های HMM انجام و در ادامه نتایج به‌دست‌آمده به‌وسیلهٔ ویژگی‌های ساختاری امتیازدهی مجدد[7] می‌شود.

برای سالیان زیادی در عرصه بازشناسی دست‌نوشته روش HMM-GMM به‌عنوان یک رویکرد متداول مطرح بوده است. با مطرح‌شدن یادگیری ژرف در سالیان اخیر، رویکرد معروف به ترکیبی که مبتنی بر تلفیق مدل مخفی مارکوف با شبکه‌های عصبی ژرف (HMM-DNN) است، توانست از رویکرد متداول HMM-GMM پیشی بگیرد و نظر پژوهش‌گران را به خود جلب کند. در رویکرد HMM-GMM برای مدل‌سازی هر شناسه یک مدل مخفی مارکوف چپ به راست آموزش داده می‌شود. برای مدل‌سازی احتمال درست‌نمایی[8] هر بردار ویژگی از دنباله بردارهای ویژگی استخراج‌شده از دست‌نوشته در هر حالت[9]، از یک مدل مخلوط گاوسی -که یک مدل تولیدی[10] محسوب می‌شود- بهره گرفته می‌شود. هر مدل مخلوط گاوسی شامل تعدادی گاوسی چندبعدی (هم بعد با بردار ویژگی) است که اوزان، میانگین و ماتریس کوواریانس (به‌طورمعمول قطری در نظر گرفته می‌شود) آنها حین فرآیند آموزش بر مبنای داده‌های آموزشی مربوطه تخمین زده می‌شود.

در رویکرد ترکیبی (HMM-DNN) نحوه به‌کارگیری مدل مخفی مارکوف مشابه رویکرد HMM-GMM است با این تفاوت که برای تخمین احتمال درست‌نمایی بردارهای ویژگی در حالت‌ها، یک شبکه عصبی - که یک روش تمایزی[11] محسوب می‌شود- به‌کار گرفته می‌شود. ورودی این شبکه یک بردار ویژگی به‌همراه تعدادی از بردارهای ویژگی مجاور آن (از هر دو طرف) و خروجی آن احتمال

[1] Sequential Patterns
[2] Gaussian Mixture Model
[3] Morphological Decomposition
[4] Sub-word
[5] Lexicon

[6] Fusion
[7] Rescoring
[8] Likelihood
[9] State
[10] Generative
[11] Discriminative

پسین است. تعداد نورون‌های خروجی شبکه برابر تعداد کل حالات مدل‌های مخفی مارکوف به‌کار گرفته‌شده برای مدل‌سازی شناسه‌ها است که خروجی هر یک از این نورون‌ها احتمال پسین حالت مربوطه به ازای بردار ویژگی ورودی آن است ($P(s|x)$). از آنجا که نیاز به احتمال درست‌نمایی حالات است، احتمال پسین هر حالت بر احتمال پیشین آن تقسیم می‌شود. احتمال پیشین حالات بر مبنای داده‌های آموزشی و به‌کمک الگوریتم هم‌ترازی اجباری[1] به‌دست می‌آید. در رویکرد ترکیبی، شبکه‌های عصبی ژرف مختلفی نظیر MLP، TDNN و LSTM استفاده می‌شود. در این رویکرد برای شروع فرایند آموزش نیاز به وجود یک هم‌ترازی اولیه بین بردارهای ویژگی و حالات HMM است که برای این منظور به‌طورمعمول از یک سامانه مبتنی بر رویکرد HMM-GMM برای ایجاد این هم‌ترازی اولیه بهره گرفته که این وابستگی به نوعی ضعف این رویکرد محسوب می‌شود. از پژوهش‌های اخیر مبتنی بر رویکرد ترکیبی می‌توان از [38] نام برد. این پژوهش که تمرکز اصلی آن بر روی بهبود نرمال‌سازی تصویر و همچنین بررسی روش‌های مختلف استخراج ویژگی است، از شبکه‌های عصبی موجود در نرم‌افزار متن‌باز KALDI برای پیاده‌سازی مدل ترکیبی بهره می‌برد.

در سال ۲۰۱۴ رویکرد انتهابه‌انتها[2] ارائه شد که مدل مخفی مارکوف به‌طورکامل کنار گذاشته شده و از یک شبکه عصبی بازگشتی ژرف که به‌طورمعمول از نوع LSTM است به‌همراه روش CTC[3] [13] برای مدل‌سازی بهره می‌برد. ورودی شبکه، بردارهای ویژگی است و خروجی آن شناسه‌های زبان به‌همراه شناسه فاصله. بنابراین تعداد نورون‌های خروجی تعداد شناسه‌های نوشتاری زبان به‌علاوهٔ یک است. از آنجا که در این رویکرد، شبکه می‌بایست هر دو مدل ویژوال و زبانی را به‌همراه هم یاد بگیرد، بنابراین باید به‌اندازهٔ کافی بزرگ باشد و دادهٔ آموزشی کافی در اختیار داشته باشیم. به همین خاطر به‌کارگیری این رویکرد تنها در مواقعی توصیه می‌شود که داده آموزشی زیاد و بستر محاسباتی قوی در اختیار است. از مزایای این رویکرد سادگی ساختار آن و عدم نیاز به مجموعه‌واژگان است که امکان بازشناسی کلمات خارج از مجموعه‌واژگان[4] را فراهم می‌کند. گفتنی است که عدم نیاز رویکرد انتهابه‌انتها به وجود مجموعه‌واژگان در کنار مزیتی که دارد یک عیب نیز دارد. در دو رویکرد گفته‌شدهٔ قبلی که مبتنی بر وجود مجموعه‌واژگان هستند، فضای جستجو در هنگام رمزگشایی[5] به کلمات مجموعه‌واژگان محدود شده است و همچنین استفاده از مدل زبانی حین رمزگشایی باعث محدودتر فضای جستجو می‌شود که درنهایت به بهبود دقت بازشناسی منجر می‌شود.

در سال ۲۰۱۷ پژوهش [1] از رویکرد انتهابه‌انتها به‌کمک روش CTC برای بازشناسی واژگان پیوسته در زبان عربی بهره برد. این پژوهش بر روی دادگان KHATT انجام شده و بازشناسی در سطح شناسه انجام می‌شود. برای این منظور از یک شبکه عصبی بازگشتی LSTM چندجهته[6] استفاده شده است. پژوهش [17] که درسال ۲۰۱۸ انجام شده است، با بررسی حالات مختلف برای ترکیب چند مدل مبتنی بر شبکه‌های عصبی بازگشتی BLSTM-CTC تلاش می‌کند تا بهترین شیوهٔ ترکیب مدل‌ها را ارائه کند. در این کار ترکیب مدل‌ها در سه سطح پایین، سطح میانی و سطح بالا بررسی می‌شود که نتایج نشان می‌دهد ترکیب سطح بالای مدل‌ها بیشترین بهبود را در نرخ خطا به‌همراه داشته است. ترکیب سطح بالا به این صورت است که چند مدل BLSTM-CTC به‌صورت مجزا آموزش داده و خروجی رمزگشایی همه آنها با هم ترکیب می‌شوند.

همان‌طور که در ابتدای این فصل بیان شد، مسأله مدل‌سازی و بازشناسی دست‌نوشته شباهت بسیار زیادی به مسأله مدل‌سازی و بازشناسی گفتار دارد. هر دو مسأله ذاتاً دنباله‌ای هستند و تنها تفاوت آنها فقط در جنس سیگنال ورودی است. به همین علت انتظار می‌رود مدل‌هایی که در مسأله بازشناسی گفتار عملکرد مناسبی داشته‌اند، در مسأله بازشناسی دست‌نوشته نیز عملکرد قابل قبولی داشته باشند. امروزه روش‌های مبتنی بر رویکرد ترکیبی به‌عنوان مرز دانش در حوزه بازشناسی گفتار شناخته می‌شوند که عمده نقطه‌ضعف آنها وابستگی به یک هم‌ترازی شناسه‌ای اولیه در شروع روند آموزش است. خوشبختانه مدل‌هایی که درهمین‌اواخر بر مبنای رویکرد ترکیبی ارائه شده‌اند، تلاش کرده‌اند تا وابستگی به این هم‌ترازی اولیه را از بین ببرند که تا حد زیادی نیز موفق بوده‌اند. با توجه به عملکرد مرز دانشی مدل‌های مبتنی بر رویکرد ترکیبی و همچنین سرعت به‌نسبه بالای آموزش آنها، در این پژوهش این دسته از مدل‌ها را برای مدل‌سازی انتخاب کرده‌ایم.

[1] Force Alignment
[2] End-to-End
[3] Connectionist Temporal Classification
[4] Out Of Vocabulary
[5] Decoding
[6] Multi-Directional Long-Short-Term-Memory

۴- اجزای سامانه

این مقاله به‌کمک روش‌های پیش‌پردازش و نرمال‌سازی تصاویر که در [38] ارائه شده است و همچنین بهره‌گیری از مدل‌سازی‌های به‌روز مبتنی بر رویکرد ترکیبی، سامانه‌ای را ارائه می‌دهد که بهبود چشم‌گیری در میزان دقت بازشناسی کلمات بر روی دادگان عربی KHATT را به همراه داشته است. در ادامه به بررسی جزئیات روش‌های پیش‌پردازش و نرمال‌سازی تصاویر ورودی و روش‌های مدل‌سازی ترکیبی به‌کار گرفته‌شده خواهیم پرداخت.

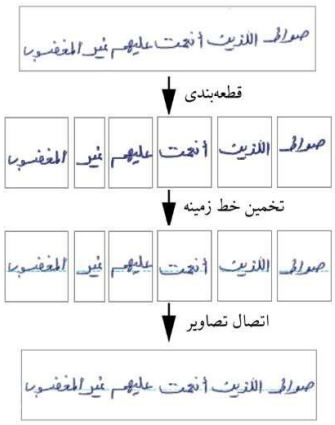
۴-۱- پیش‌پردازش و نرمال‌سازی تصاویر

پیش‌پردازش و نرمال‌سازی تصاویر ورودی در حوزه بازشناسی دست‌نوشته به‌خصوص دست‌نوشته‌های عربی اهمیت بالایی دارد. پژوهش [38] با تکیه بر نرمال‌سازی تصاویر قبل از آموزش مدل توانسته است بهبود قابل توجهی در میزان دقت به‌دست‌آمده بر روی دادگان عربی KHATT گزارش کند. در ادامه به بررسی جزئیات نرمال‌سازی‌های انجام‌شده در این پژوهش می‌پردازیم.

نخستین مرحله در نرمال‌سازی تصویر ورودی، تخمین خط زمینه در محدوده کلمه است. در این جا، این کار توسط روشی که در [37] ارائه شده است، انجام می‌شود. به این صورت که باریک‌ترین نواری که بیشترین بخش از پیکسل‌های یک واژه را در تصویر در بر بگیرد به‌عنوان خط زمینهٔ بالا و پایین انتخاب و درنهایت خط زمینه پایین به‌عنوان خط زمینه لحاظ می‌شود.

(شکل-۳): تخمین خط زمینه بالا و پایین
(Figure-3): Upper and lower baseline estimation

برای تخمین خط زمینه یک خط دست‌نوشته، ابتدا تصویر به‌صورت عمودی به بخش‌هایی که هر کدام شامل یک کلمه هستند تقسیم می‌شود؛ سپس برای هر بخش که شامل یک کلمه است به روش گفته‌شده در بالا، خط زمینه تخمین زده می‌شود. بعد از تخمین خط زمینه، تصویر را حول این خط به میزانی می‌چرخانیم که خط زمینه به‌طورکامل افقی شود؛ سپس همه بخش‌ها را به هم متصل می‌کنیم تا تصویر نرمال‌شدهٔ کل خط به‌دست آید. این کار کجی تصاویر خطوط را اصلاح می‌کند.



(شکل-۴): مراحل تخمین و تصحیح خط زمینه
(Figure-4): Steps of baseline estimation and correction

یکی دیگر از ویژگی‌های تصاویر دست‌نوشته‌های عربی این است که حروف به‌طورمعمول به‌طور کامل در راستای عمودی قرار ندارند و با توجه به سبک دست‌خط نویسنده قدری به چپ و یا راست متمایل هستند. مرحله بعدی در نرمال‌سازی تصویر، ازبین‌بردن این اریبی کلمات است. یکی از روش‌ها برای ازبین‌بردن این اریبی استفاده از تبدیل Shear است. برای پیداکردن زاویه مناسب در این تبدیل از تبدیل Hough [10] مطابق آنچه در [37] آمده استفاده شده است.

(شکل-۵): اریبی‌های متفاوت در دست‌نوشته فارسی
(Figure-5): Slants in Persian Handwritten

در این پژوهش از پیکسل‌های خام تصویر به‌عنوان بردار ویژگی استفاده شده است. به این صورت که یک پنجره به عرض یک پیکسل و همچنین با طول گام یک پیکسل بر روی تصویر لغزانیده می‌شود و یک بردار ویژگی را به‌اندازهٔ عرض تصویر که مقادیر خام پیکسل‌ها هستند، تولید می‌کند. به این دلیل که ضروری است، تعداد ویژگی‌های تمام بردارهای ویژگی با یکدیگر برابر باشد؛ به این منظور بعد از مراحل نرمال‌سازی، تمام تصاویر را با حفظ نسبت طول به عرض آنها، به تصاویری با عرض پنجاه پیکسل تغییر اندازه می‌دهیم.

۲-۴- مدل‌سازی

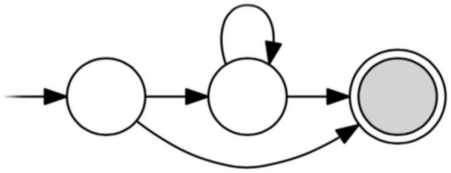
در روند آموزش مدل‌های مبتنی بر رویکرد ترکیبی (HMM-DNN) که هدف این پژوهش است، می‌توان از توابع هدف مختلفی استفاده کرد. این توابع هدف را می‌توان به دو دسته توابع هدف در سطح قاب و توابع هدف در سطح دنباله تقسیم کرد. در مدل‌هایی که توابع هدف آنها در سطح قاب تعریف شده است، به‌روزرسانی پارامترهای مدل بعد از مشاهده هر بردار ویژگی از ورودی انجام می‌شود؛ درحالی‌که در مدل‌هایی که توابع هدف آنها دنباله‌ای است، ابتدا باید تمام بردارهای ویژگی داده آموزشی مشاهده و سپس پارامترهای مدل به‌روزرسانی شوند. از آن‌جا که بازشناسی دست‌نوشته یک مسأله ذاتاً دنباله‌ای است، انتظار می‌رود توابع هدف دنباله‌ای عملکرد بهتری داشته باشند. از جمله مهم‌ترین توابع دنباله‌ای معرفی‌شده در حوزهٔ بازشناسی گفتار که می‌توان در بازشناسی دست‌نوشته نیز از آنها بهره برد، MMI[1] [4]، sMBR[2] [12]، MWE[3] و MPE[4] [28] هستند که شباهت‌های بسیار زیادی به یکدیگر دارند. پژوهش‌های اخیر در حوزه بازشناسی گفتار نشان داده [40,39,35] که با استفاده از این توابع بهبود در خور توجهی قابل حصول است که انتظار می‌رود به‌کارگیری آن‌ها در بازشناسی دست‌نوشته نیز مؤثر واقع شود.

پایه توابع هدف تمایزی بر مبنای بیشینه‌سازی اطلاعات متقابل بین $\mathbb{M}_w$ به‌عنوان گراف آموزشی و $x^{(u)}$ به‌عنوان دنباله بردارهای ویژگی ورودی uام به‌ازای تمام ورودی‌ها است که به‌صورت زیر تعریف می‌شود:

$$\mathcal{F}_{MMI} = \sum_{u=1}^{U} I(x^{(u)}, \mathbb{M}_{w^{(u)}}) \quad = \quad \sum_{u=1}^{U} \log \frac{p_\lambda(x^{(u)}|\mathbb{M}_{w^{(u)}})}{p_\lambda(x^{(u)})} \qquad (۲)$$

که در آن I تابع اطلاعات مشترک و U تعداد کل نمونه‌های آموزشی است؛ بنابراین این تابع هدف که تابع هدف MMI نامیده می‌شود، علاوه‌براین که در سطح دنباله است، برخلاف تابع هدف بیشینه-شباهت یک تابع هدف تمایزی نیز است. یک رویکرد ترکیبی که از این تابع هدف بهره برده است LF-MMI[5] [27] نامیده می‌شود و در حال حاضر مرز دانش در بازشناسی گفتار به حساب می‌آید. روش LF-MMI در [14] به‌نحوی تغییر داده شده است که نیاز به هم‌ترازی قبلی ندارد. این روش جدید LF-MMI تخت‌آغاز[6] نامیده شده است.

روش LF-MMI در تئوری تفاوتی با روش MMI ندارد. محاسبه بر روی گراف مخرج در روش MMI مسأله چالش‌برانگیزی است. به‌دلیل این‌که گراف مخرج، گرافی است متشکل از تمام حالات ممکن، بنابراین محاسبات بر روی آن بسیار پر هزینه است؛ ازاین‌رو به‌طورمعمول محاسبه گراف مخرج به‌وسیلهٔ n-تا بهترین[7] و یا لاتیس تقریب زده می‌شود. به این معنی که به جای محاسبه بر روی تمام حالات ممکن، محاسبات تنها بر روی محتمل‌ترین حالات انجام می‌شود. در روش MMI همچنین برای مقداردهی اولیه شبکه عصبی از پیش‌آموزش شبکه به‌وسیلهٔ شبکه‌های باور عمیق[8] استفاده می‌شود [7].



(شکل-۶): توپولوژی شناسه در روش LF-MMI
(Figure-6): HMM topology for characters in LF-MMI

تفاوت روش LF-MMI با روش آموزش تمایزی MMI در این است که گراف مخرج به جای n-تا بهترین و یا لاتیس به‌صورت دقیق محاسبه و همچنین آموزش شبکه نیز با مقداردهی اولیه تصادفی شروع می‌شود. برای کاهش هزینه محاسبات گراف مخرج سه روش پیشنهاد شده است:

۱- در گراف مخرج به جای استفاده از مدل زبانی در سطح واژه، از مدل زبانی در سطح شناسه استفاده می‌شود. این کار باعث می‌شود اندازهٔ گراف دو تا سه مرتبه کوچک‌تر شود.

۲- جهت انجام الگوریتم‌های پیش‌رو-پس‌رو[9] برای گراف مخرج از پردازنده‌های گرافیکی استفاده می‌شود.

۳- نرخ قاب در خروجی شبکه به یک سوم کاهش داده می‌شود. این کار با استفاده از شبکه‌های TDNN به‌سادگی قابل انجام است. این کار باعث می‌شود تا هم

[1] Maximum Mutual Information
[2] State-level Minimum Bayes Risk
[3] Minimum Word Error
[4] Minimum Phone Error
[5] Lattice-Free Maximum Mutual Information
[6] Flat start
[7] N-best
[8] Deep Belief Network (DBN)
[9] Forward-Backward algorithms

پیچیدگی محاسبات در خود شبکه کاهش و همچنین تعداد دفعاتی که پارامترهای شبکه را به‌روزرسانی می‌کنیم، کاهش یابد. کاهش نرخ قاب در خروجی باعث می‌شود تا به‌ازای هر سه قاب از ورودی، یک بردار را از احتمال‌های پسین در خروجی داشته باشیم. به همین علت باید توپولوژی مدل مارکوف شناسه‌ها را به‌نحوی تغییر دهیم که با یک گذر بتوان از هر شناسه عبور کرد؛ در غیر این صورت کوتاه‌ترین طول برای شناسه سه برابر حالت عادی می‌شود که مسلماً تأثیر منفی بر دقت بازشناسی شناسه‌ها خواهد داشت. توپولوژی HMM استفاده‌شده در روش LF-MMI در شکل (۶) مشاهده می‌شود.

ضعف عمدهٔ مدل LF-MMI مشابه سایر رویکردهای ترکیبی این است که آموزش آن نیاز به یک هم‌ترازی اولیه دارد. هدف اصلی از ارائه مدل LF-MMI تخت‌آغاز مرتفع‌کردن این مشکل است. به عبارتی این روش در تلاش است که تمام فرایند آموزش در یک مرحله انجام شود و نیازی به وجود هم‌ترازی اولیه نباشد. روش LF-MMI در دو مرحله از یک هم‌ترازی اولیه استفاده می‌کند. نخستین مرحله برای ساخت گراف صورت است. گراف صورت در این روش بر مبنای لاتیس ساخته‌شده به‌وسیلهٔ یک مدل آموزش‌دیده‌شده، طول شناسه‌ها را تخمین می‌زند و گرافی تولید می‌کند که در آن هیچ طوقه‌ای[1] وجود نداد. وابستگی بعدی در مرحله آموزش درخت تصمیم بر مبنای هم‌ترازی اولیه است که به‌وسیلهٔ آن حالت‌های HMM به هم گره زده می‌شوند. در روش LF-MMI تخت‌آغاز برای حذف این وابستگی در مرحلهٔ نخست، از یک گراف ترکیبی که برای آموزش مدل‌های HMM-GMM استفاده می‌شود، به‌عنوان گراف صورت استفاده می‌کنیم. این گراف هیچ محدودیتی برای داشتن طوقه ندارد؛ ازاین‌رو نیازی به هیچ هم‌ترازی اولیه‌ای نیز ندارد. برای حذف وابستگی درخت تصمیم مورد استفاده به هم‌ترازی اولیه در روش LF-MMI تخت‌آغاز، از تمام ترکیب‌های دوشناسه‌ای استفاده شده است؛ درواقع هیچ کدام از حالت‌های HMM به هم گره زده نمی‌شوند. در آزمایش‌های [27] نشان داده شده است که استفاده از ترکیب‌های دوشناسه‌ای عملکردی معادل ترکیب‌های سه‌شناسه‌ای دارند؛ همچنین در آزمایش‌های اخیر [14] نشان داده شده است که اگر هیچ کدام از حالت‌ها را به هم گره نزنیم، دقت به‌دست آمده تغییر چندانی نمی‌کند؛ زیرا شبکه یاد می‌گیرد تا خروجی حالت‌های بدون استفاده را همیشه صفر تولید کند و این باعث می‌شود تا این حالت‌ها در هیچ‌کدام از محاسبات گراف صورت و مخرج اثری نداشته باشند.

این مقاله به‌کمک روش‌های پیش‌پردازش و نرمال‌سازی تصاویر که در [38] ارائه شده است و همچنین بهره‌گیری از رویکردهای ترکیبی LF-MMI و LF-MMI تخت‌آغاز جهت مدل‌سازی، سامانه‌ای را ارائه می‌دهد که بهبود چشم‌گیری در میزان دقت بازشناسی واژگان بر روی دادگان عربیKHATT را به‌همراه داشته است. شبکه عصبی استفاده‌شده در کلیه آزمایش‌های انجام‌شده در این مقاله شامل هفت لایه پیچشی[2] در ابتدا و پنج لایه تأخیر زمانی[3] در انتهاست.

۵- نتایج و آزمایش‌ها

آزمایش‌ها بر روی دادگان عربی KHATT انجام شده است. دادگان KHATT شامل تصاویر خطوط دست‌نوشته است که از فرم‌های تکمیل شده توسط هزار نویسنده استخراج شده است. این دادگان شامل دو بخش مجزا است. بخش نخست شامل جملاتی ثابتی است که توسط تمام نویسنده‌ها نوشته شده و شامل تمام شناسه‌ها و حالات مختلف آنها است. بخش دوم، شامل جملاتی یکتاست که از ۴۶ منبع مختلف جمع‌آوری شده و برای هر نویسنده متفاوت است. این دادگان درمجموع شامل ۱۳۳۵۷ تصویر خطوط دست‌نوشته است که از بین آنها ۹۴۶۲ خط برای مجموعه آموزش، ۱۸۹۹ خط برای مجموعه توسعه و ۱۹۹۶ خط برای مجموعه آزمون در نظر گرفته شده است. در این پژوهش برای تقسیم‌بندی تصاویر به مجموعه‌های آموزش، توسعه و آزمون مشابه [16] و [38] عمل شده است تا نتایج قابل مقایسه باشد. در انجام آزمایش‌ها، از یک مدل زبانی 3-gram استفاده شده که از متون داده آموزشی استخراج شده است. نرخ کلمات خارج از مجموعه‌واژگان برای این دادگان برابر ۱۱/۴ درصد است. عملکرد سامانه بر مبنای درصد خطای بازشناسی واژه ارزیابی شده است که مطابق رابطه زیر محاسبه می‌شود:

$$\frac{\text{تعداد جایگزینی+تعداد حذف+تعداد درج}}{\text{تعداد کل کلمات}} * 100 \qquad (۳)$$

[1] Loop

[2] Convolutional
[3] Time Delay

(جدول-۳): درصد خطای بازشناسی کلمه برای مجموعه توسعه و آزمون تعریف شده بر روی دادگان khatt به‌ازای روش‌های مختلف استخراج ویژگی و مدل‌سازی در نرم‌افزارKALDI

(Table-3): Word error rate of recognition on the test set and development set of KHATT data set for various feature extraction and modeling techniques in KALDI

شماره	ویژگی	واحد مدل‌سازی	مدل‌سازی	خطای بازشناسی کلمه (٪)	
				مجموعه توسعه	مجموعه آزمون
1	Raw pixels + delta	Mono-Character	HMM-GMM	103.87	103.41
2	Raw pixels + delta	Tri-Character	HMM-GMM	78.05	76.46
3	Raw pixels + LDA + MLLT	Tri-Character	HMM-GMM	56.08	54.14
4	Raw pixels + LDA – MLLT + WAT	Tri-Character	HMM-GMM	55.46	53.21
5	Raw pixels	Bi-Character	HMM-DNN (LF-MMI)	29.31	27.96

در جدول (۳) نتایج بازشناسی بر روی مجموعه توسعه و آزمون دادگان KHATT به‌ازای ویژگی‌ها و مدل‌های مختلف آمده است. سطر نخست، نتایج را برای ساده‌ترین حالت، یعنی استفاده از پیکسل‌های خام و مشتق نخست آنها به‌عنوان ویژگی و همچنین بهره‌بردن از مدل مبتنی بر تک‌شناسه نشان می‌دهد. در سطر دوم تفاوت نسبت به سطر نخست تنها در روش مدل‌سازی اتفاق افتاده است که مدل‌سازی تک‌شناسه‌ای به سه‌شناسه‌ای تغییر یافته است و همان‌طورکه مشاهده می‌شود، بهبود قابل توجهی به همراه داشته است. در سطر سوم، بر روی هر بردار ویژگی که درواقع یک ستون از پیکسل‌های تصویر اصلی است، به همراه بردارهای ویژگی قبل و بعد، دو تبدیل خطی LDA[1] و MLLT[2] به ترتیب اعمال شده و درنهایت یک بردار چهل مؤلفه‌ای به‌دست می‌آید که مدل‌سازی سه‌شناسه‌ای بر مبنای آن صورت می‌گیرد. همان‌طور که در جدول نیز مشاهده می‌شود، استفاده از تبدیل‌های LDA و MLLT به جای مشتق نخست، باعث بهبود مطلق خطای بازشناسی به میزان ۲۲/۳۲ درصد شده است. با توجه به این بهبود قابل توجه، از این مرحله به بعد در تمام مراحل از این دو تبدیل استفاده شده است. در سطر چهارم، آموزش تطبیقی به نویسنده[3] صورت گرفته است. همچنین در مرحله بازشناسی، ویژگی‌های هر گوینده به‌کمک روش fMLLR تطبیق داده شده که درنهایت منجر به کاهش ۰/۹۳ درصدی خطای بازشناسی شده است. در سطر پنجم تنها از پیکسل‌های خام تصویر به‌عنوان بردار ویژگی استفاده شده است. به این علت که مدل‌های ترکیبی نیاز به یک هم‌ترازی شناسه‌ای اولیه دارند، در این مدل، ابتدا به‌کمک مدل آموزش داده‌شده در سطر چهارم، یک هم‌ترازی اولیه برای بردارهای ویژگی داده‌های آموزشی به‌دست می‌آوریم و از آن در روند آموزش شبکه عصبی رویکرد ترکیبی LF-MMI استفاده می‌کنیم. نتایج نشان می‌دهند که استفاده از این مدل باعث بهبود مطلق ۲۵/۲۵ درصدی در خطای بازشناسی کلمات شده است.

۱-۵- بررسی تأثیر هم‌ترازی شناسه‌ای اولیه

همان‌طور که در بخش ۳ بیان شد، مدل LF-MMI تخت‌آغاز گسترشی از روش LF-MMI است که نیاز به هم‌ترازی شناسه‌ای اولیه ندارد. در سطر نخست جدول (۴) نتایج آزمایش با استفاده از مدل LF-MMI تخت‌آغاز قابل مشاهده است. همچنین با توجه به این که مدل LF-MMI نیازمند یک هم‌ترازی اولیه است، این مدل را دوباره و با استفاده از هم‌ترازی‌های اولیه متفاوت مورد آزمایش قرار دادیم. در سطر دوم جدول (۴)، از هم‌ترازی به‌دست‌آمده از مدل سطر چهارم جدول (۳) استفاده شده است. در سطر سوم، هم‌ترازی اولیه به‌دست‌آمده به‌وسیلهٔ مدل سطر پنجم جدول (۳) که خود که یک مدل ترکیبی است، استفاده شده است. در انتها و در سطر چهارم، هم‌ترازی به‌دست‌آمده به‌وسیلهٔ مدل LF-MMI تخت‌آغاز، به‌عنوان هم‌ترازی شناسه‌ای اولیه مورد استفاده قرار گرفته است. با توجه به نتایج به‌دست‌آمده، می‌توان مشاهده کرد که هرچه هم‌ترازی اولیه دقیق‌تر باشد، دقت روش LF-MMI نیز بیش‌تر است؛ همچنین دقتی که روش LF-MMI تخت‌آغاز به همراه داشته با توجه به بی‌نیازبودن از هم‌ترازی اولیه، دقت درخور توجهی است.

[1] Linear Discriminant Analysis
[2] Maximum Likelihood Linear Transform
[3] Writer Adaptive Training (WAT)

(جدول-۴): نتایج به‌دست‌آمده از دو هم‌ترازی اولیه متفاوت
(Table-4): Results of two different initial alignments

مدل‌سازی	هم‌ترازی اولیه	خطای بازشناسی واژه (٪)	
		توسعه	آزمون
HMM-DNN (Flatstart LF-MMI)	-	28.87	26.83
HMM-DNN (LF-MMI)	HMM-GMM	29.31	27.96
HMM-DNN (LF-MMI)	HMM-DNN (LF-MMI)	27.66	26.26
HMM-DNN (LF-MMI)	HMM-DNN (Flatstart LF-MMI)	26.21	25.32

۵-۲- بررسی اثر افزایش داده[1]

جهت بررسی اثر افزایش داده بر نرخ خطای بازشناسی کلمه، با انجام تبدیلات مختلفی از جمله افزودن نویز گاوسی، ایجاد اریبی[2]، چرخش[3]، حذف تصادفی[4] پیکسل‌ها، مات‌کردن[5] و تغییر مقیاس[6] بر روی تصاویر آموزشی، تلاش شد تا داده‌های آموزشی را به‌صورت مصنوعی افزایش دهیم. تبدیل‌های انجام‌شده باعث می‌شود تا مدل در هنگام آموزش تصاویر متنوعی ببیند که این باعث کاهش بیش‌برازش[7] و افزایش قابلیت تعمیم‌پذیری مدل می‌شود.

با آموزش مدل ترکیبی LF-MMI با استفاده از هم‌ترازی به‌دست‌آمده توسط مدل LF-MMI تخت‌آغاز به‌عنوان هم‌ترازی اولیه، بر روی داده‌های آموزشی افزایش داده‌شده، نتایجی که در جدول (۵) مشاهده می‌شود به‌دست آمد. نتایج نشان می‌دهند که نرخ خطای بازشناسی کلمه بر روی داده توسعه به میزان ۰/۸۷ درصد و بر روی دادهٔ آزمون به میزان ۱/۱۸ درصد بهبود مطلق داشته است.

(جدول-۵): نتایج به‌دست‌آمده از آموزش مدل بر روی داده آموزشی افزایش یافته
(Table-5): Results of model training using augmented training data

مدل	نرخ خطای واژه (٪)	
	توسعه	آزمون
LF-MMI (without augmentation)	26.21	25.32
LF-MMI (with augmentation)	25.34	24.14

۵-۳- بررسی اثر مدل زبانی

از دیگر آزمایش‌های انجام‌شده در این پژوهش، بررسی اثر انواع مختلف مدل زبانی بر روی نتایج به‌دست‌آمده است. مدل‌های زبانی مختلف به‌وسیلهٔ مدل LF-MMI با هم‌ترازی اولیه به‌دست‌آمده از مدل تخت‌آغاز مورد آزمایش قرار گرفتند. این مدل‌های شامل 2-gram و 3-gram بودند که هر کدام به‌وسیلهٔ روش‌های هموارسازی kneser-ney،good-turing و بیشینه بی‌نظمی[8] هموار شدند. جدول (۶) نتایج به‌دست‌آمده از این آزمایش‌ها را نشان می‌دهد.

(جدول-۶): نتایج به‌دست‌آمده از انواع مدل زبانی و روش‌های مختلف هموارسازی
(Table-6): Results of various language model types and smoothing techniques

مدل	هموارسازی	خطای بازشناسی کلمه (٪)	
		توسعه	آزمون
2-gram	Good-Turing	25.20	24.00
2-gram	Kneser-Ney	24.57	23.19
2-gram	MaxEnt	24.96	23.68
3-gram	Good-Turing	25.34	24.14
3-gram	Kneser-Ney	24.88	23.52
3-gram	MaxEnt	24.60	23.18

همان‌طور که در جدول مشاهده می‌شود، استفاده از مدل زبانی 2-gram باعث کاهش نرخ خطا شده است. این کاهش به ازای استفاده از روش هموارسازی Kneser-Ney به بیشترین حد خود رسیده و همچنین استفاده از روش‌های هموارسازی Kneser-Ney و بیشینه بی‌نظمی در مدل 3-gram نیز باعث کاهش نرخ خطا شده است. در مدل 3-gram، روش هموارسازی بیشینه بی‌نظمی بیشترین بهبود را داشته است.

۵-۴- مقایسه نتایج با سایر پژوهش‌ها

در جدول (۷) نتایج به‌دست‌آمده در این پژوهش با سایر پژوهش‌ها مقایسه شده است. تا جایی که مطلع هستیم بهترین دقت بر روی دادگان خط با استفاده از مدل زبانی استخراج‌شده از متن داده آموزشی توسط [38] ارائه شده است. هر دو پژوهش [16] و [38] از مدل زبانی 3-gram استخراج‌شده از متون داده آموزشی استفاده می‌کنند. به همین علت ما نیز بهترین دقتی را که به‌وسیلهٔ یک مدل زبانی 3-gram حاصل شده با این پژوهش‌ها مقایسه می‌کنیم. همان‌طور که در جدول (۷) قابل مشاهده است، دقت بازشناسی واژه به‌دست‌آمده در این پژوهش بهترین دقت گزارش‌شده را به میزان ۷/۳۲ درصد مطلق بهبود بخشیده است.

[1] Data Augmentation
[2] Skewing
[3] Rotating
[4] Dropout
[5] Blurring
[6] Scaling
[7] Over-fitting
[8] Maximum Entropy

(جدول-۷): مقایسه نتایج به‌دست‌آمده با سایر نتایج گزارش‌شده بر روی دادگان KHATT

(Table-7): Comparison of our results with other reported results on KHATT dataset

پژوهش	نرخ خطای واژه (٪)	
	توسعه	آزمون
Hamdani et al. [16]	33.60	34.10
QCRI [38]	29.40	30.50
Jemni et al. [17]	31.60	29.44
پژوهش جاری	24.60	23.18

۶- جمع‌بندی

در این مقاله بعد از مروری اجمالی بر روی پژوهش‌های صورت‌گرفته و منابع موجود در حوزه بازشناسی دست‌نوشته عربی و فارسی، به بررسی شباهت‌های دو مسأله بازشناسی گفتار و بازشناسی دست‌نوشته پرداختیم؛ سپس با مروری کامل بر روی رویکردهای مرسوم برای مدل‌سازی ویژوال دست‌نوشته، با توجه به نتایج قابل قبول رویکردهای مبتنی بر HMM-DNN در حوزه بازشناسی گفتار و همچنین بی‌نیازی این رویکرد از انجام استخراج ویژگی صریح و مهندسی‌شده از تصاویر ورودی به‌دلیل وجود شبکه عصبی، این رویکرد را برای آزمایش بر روی دادگان KHATT انتخاب کردیم. با به‌کارگیری مدل‌های LF-MMI و LF-MMI تخت‌آغاز که از جدیدترین مدل‌های مبتنی بر HMM-DNN هستند، و همچنین بهره‌بردن از روش افزایش داده موفق به بهبود نرخ خطای کلمه بر روی این دادگان شدیم. بهترین نرخ خطای واژه بر روی مجموعه توسعه برابر با ۲۴/۶۰ و بر روی مجموعه آزمون برابر با ۲۳/۱۸ به‌دست آمد که بهبود درخور توجهی نسبت به نتایج گزارش‌شده قبلی بر روی دادگان KHATT است. تمرکز اصلی این پژوهش بر روی به‌کارگیری مدل‌های مرسوم در حوزه بازشناسی گفتار برای حل مسأله بازشناسی دست‌نوشته عربی بوده است.

سپاس‌گزاری

نویسندگان مقاله برخود لازم می‌دانند از همکاری و حمایت آقای *دنیل پووی* پژوهش‌گر ارشد مرکز پردازش زبان و گفتار (CLSP) دانشگاه جانز هاپکینز و دکتر *حسین هادیان* فارغ‌التحصیل مقطع دکترای هوش مصنوعی دانشگاه صنعتی شریف به‌خاطر راهنمایی‌های بی‌دریغ‌شان و *فهیم عمادالدین* پژوهش‌گر مؤسسه پژوهشی QCRI کشور قطر به‌خاطر به‌اشتراک‌گذاشتن برخی منابع و کدها صمیمانه سپاس‌گزاری کند.

۷- مراجع — 7- References

[1] R. Ahmad, S. Naz, M. Z. Afzal, S. F. Rashid, M. Liwicki, A. Dengel, "KHATT: A Deep Learning Benchmark on Arabic Script.", *In 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR),* pp. 10-14, 2017.

[2] R. Al-Hajj, C. Mokbel, C. Mokbel, L. Likforman-Sulem and L. Likforman-Sulem, "Combination of HMM-Based Classifiers for the Recognition of Arabic Handwritten Words", Ninth International Conference on Document Analysis and Recognition (ICDAR 2007) Vol 2, 2007.

[3] J. AlKhateeb, J. Ren, J. Jiang and H. Al-Muhtaseb, "Offline handwritten Arabic cursive text recognition using Hidden Markov Models and re-ranking", *Pattern Recognition Letters*, vol. 32, no. 8, pp. 1081-1088, 2011.

[4] L. Bahl, P. Brown, P. de Souza and R. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition", *in International Conference on Acoustics, Speech, and Signal Processing*, vol. 11, pp. 49-52, 1986.

[5] A. M. Bidgoli, M. Sarhadi, "IAUT/PHCN: Islamic Azad University of Tehran/Persian handwritten city names, a very large database of handwritten Persian word.'', *11th International Conference on Frontiers in Handwriting Recognition*, pp. 192-197, 2008.

[6] A. Broumandnia, J. Shanbehzadeh and M. Rezakhah Varnoosfaderani, "Persian/arabic handwritten word recognition using M-band packet wavelet transform", *Image and Vision Computing,* vol. 26, no. 6, pp. 829-842, 2008.

[7] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large- vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, 2012.

[8] M. Dehghan, K. Faez, M. Ahmadi and M. Shridhar, "Unconstrained Farsi handwritten word recognition using fuzzy vector quantization and hidden Markov models", *Pattern Recognition Letters*, vol. 22, no. 2, pp. 209-214, 2001.

[9] M. Dehghan, K. Faez, M. Ahmadi, M. Shridhar, "Handwritten Farsi (Arabic) word recognition: a holistic approach using discrete HMM," *Pattern Recognition*, vol. 34, no. 5, pp. 1057-1065, 2001.

[10] R. Duda and P. Hart, "Use of the Hough transformation to detect lines and curves in pictures", *Communications of the ACM*, vol. 15, no. 1, pp. 11-15, 1972.

[11] A. Elbaati, H. Boubaker, M. Kherallah, A. Ennaji, H. Abed and A. Alimi, "Arabic Handwriting Recognition Using Restored Stroke

of Farsi handwritten city names." In International Conference on Frontiers in Handwriting Recognition. 2008.

[23] S. Mozaffari, K. Faez, F. Faradji, M. Ziaratban, S. M. Golzan, "A comprehensive isolated Farsi/Arabic character database for handwritten OCR research," In Tenth International Workshop on Frontiers in Handwriting Recognition, Suvisoft, 2006.

[24] S. Mozaffari, K. Faez and M. Ziaratban, "Structural decomposition and statistical description of Farsi/Arabic handwritten numeric characters", Eighth International Conference on Document Analysis and Recognition (ICDAR'05), 2005.

[25] M. Pechwitz, S. S. Maddouri, V. Märgner, N. Ellouze, H. Amiri. "IFN/ENIT-database of handwritten Arabic words," *In Proc. of CIFED*, vol. 2, pp. 127-136. 2002.

[26] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, K. Vesely, "The Kaldi speech recognition toolkit," *In IEEE 2011 workshop on automatic speech recognition and understanding, IEEE Signal Processing Society*, 2011.

[27] D. Povey, V. Peddinti, D. Galvez, P. Ghahrmani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for asr based on lattice-free mmi," in Interspeech, 2016.

[28] D. Povey and P. C. Woodland, "Minimum phone error and i-smoothing for improved discriminative train- ing," *in IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP, vol. 1. IEEE*, pp. I–105, 2002.

[29] D. Rybach, S. Hahn, P. Lehnen, D. Nolden, M. Sundermeyer, Z. Tüske, S. Wiesler, R. Schlüter, H Ney, "Rasr-the rwth aachen university open source speech recognition toolkit," In Proc. IEEE Automatic Speech Recognition and Understanding Workshop. 2011.

[30] R. Sabzi, Z. Fotoohinya, A. Khalili, S. Golzari, Z. Salkhorde, S. Behravesh, S. Akbarpour, "Recognizing Persian handwritten words using deep convolutional networks," *in Artificial Intelligence and Signal Processing Conference (AISP)*, pp. 85-90, 2017.

[31] J. Sadri, C. Y. Suen, and T. D. Bui, "Application of support vector machines for recognition of handwritten Arabic/Persian digits," *In Proceedings of Second Iranian Conference on Machine Vision and Image Processing*, vol. 1, pp. 300-307. 2003.

Chronology", 2009 10th International Conference on Document Analysis and Recognition, 2009.

[12] V. Goel and W. Byrne, "Minimum Bayes-risk automatic speech recognition", *Computer Speech & Language,* vol. 14, no. 2, pp. 115-135, 2000.

[13] A. Graves, S. Fernández, F. Gomez and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," *In Proceedings of the 23rd international conference on Machine learning*, pp. 369-376, 2006.

[14] H. Hadian, H. Sameti, D. Povey and S. Khudanpur, "Flat-Start Single-Stage Discriminatively Trained HMM-Based Models for ASR", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 1949-1961, 2018.

[15] P. Haghighi, N. Nobile, C. He and C. Suen, "A New Large-Scale Multi-purpose Handwritten Farsi Database", *Lecture Notes in Computer Science*, pp. 278-286, 2009.

[16] M. Hamdani, A. Mousa and H. Ney, "Open Vocabulary Arabic Handwriting Recognition Using Morphological Decomposition", 2013 12th International Conference on Document Analysis and Recognition, 2013.

[17] S. K. Jemni, Y. Kessentini, S. Kanoun, J. Ogier, "Offline Arabic Handwriting Recognition Using BLSTMs Combination.", *In 2018 13th IAPR International Workshop on Document Analysis Systems (DAS),* pp. 31-36, 2018.

[18] S. Khorashadizadeh, A. Latif, "Arabic/Farsi Handwritten Digit Recognition usin Histogra of Oriented Gradient and Chain Code Histogram", *International Arab Journal of Information Technology (IAJIT),* vol. 13, no. 4, 2016.

[19] H. Khosravi, E. Kabir, "Introducing a very large dataset of handwritten Farsi digits and a study on their varieties.", *Pattern recognition letters,* vol. 28, no. 10, pp. 1133-1141. 2007.

[20] D. Lee, S. Ismael, S. Grimes, D. Doermann, S. Strassel, Z. Song, "MADCAT Phase 1 Training Set", LDC2012T15. DVD. Philadelphia: Linguistic Data Consortium, 2012.

[21] S. A. Mahmoud, I. Ahmad, M. Alshayeb, W. G. Al-Khatib, M. T. Parvez, G. A. Fink, V. Margner, H. E. Abed, "KHATT: Arabic offline handwritten text database," *In 2012 International Conference on Frontiers in Handwriting Recognition (ICFHR 2012)*, pp. 449-454, 2012.

[22] S. Mozaffari, H. E. Abed, V. Märgner, K. Faez, A. Amirshahi. "IfN/Farsi-Database: a database

[32] J. Sadri, M. R. Yeganehzad, J. Saghi, "A novel comprehensive database for offline Persian handwriting recognition.", *Pattern Recognition*, vol. 60, pp. 378-393, 2016.

[33] R. Safabaksh, A. Ghanbarian and G. Ghiasi, "HaFT: A handwritten Farsi text database", 2013 8th Iranian Conference on Machine Vision and Image Processing (MVIP), 2013.

[34] H. Sajedi, "Handwriting recognition of digits, signs, and numerical strings in Persian", *Computers & Electrical Engineering*, vol. 49, pp. 52-65, 2016.

[35] H. Sak, O. Vinyals, G. Heigold, A. Senior, E. McDermott, R. Monga, and M. Mao, "Sequence discriminative distributed training of long short-term memory recurrent neural networks," in Interspeech, 2014.

[36] H. Soltanzadeh, M Rahmati, "Recognition of Persian handwritten digits using image profiles of multiple orientations," *Pattern Recognition Letters*, vol. 25, no. 14, pp. 1569-1576, 2004.

[37] F. Stahlberg and S. Vogel, "Detecting dense foreground stripes in Arabic handwriting for accurate baseline positioning", 2015 13th International Conference on Document Analysis and Recognition (ICDAR), 2015.

[38] F. Stahlberg and S. Vogel, "The QCRI Recognition System for Handwritten Arabic", *Image Analysis and Processing*, pp. 276-286, 2015.

[39] P. Voigtlaender, P. Doetsch, S. Wiesler, R. Schlüter, and H. Ney, "Sequence-discriminative training of re- current neural networks," *in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE*, pp. 2100–2104, 2015.

[40] K. Veselỳ, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural net- works," *in INTERSPEECH*, pp. 2345–2349, 2013.

[41] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. Woodland, The HTK Book (for version 3.4). Cambridge Univ. Eng. Dept., 2009.

[42] M. Ziaratban, K. Faez and F. Bagheri, "FHT: An Unconstraint Farsi Handwritten Text Database", *2009 10th International Conference on Document Analysis and Recognition*, pp. 281-285, 2009.

[43] بایسته، تاشک الهام، احمدی فرد، علیرضا و خسروی، حسین "یک روش دو مرحله‌ای برای بازشناسی کلمات دست‌نوشته فارسی به کمک بلوک‌بندی تطبیقی گرادیان تصویر" مجله علمی و پژوهشی پردازش علائم و داده‌ها. ۱۳۹۴; ۱۲ (۳) :۱۵-۲۹

[43] E. Bayesteh Tashk, A. Ahmadifard and H. khosravi, ''A two step method for offline handwritten Farsi word recognition using adaptive division of gradient image'', *JSDP*, Vol.12 (3), pp.15-29, 2015.

باقر باباعلی مدرک کارشناسی خود را در سال ۱۳۷۹ در رشتهٔ مهندسی رایانه (گرایش سخت‌افزار) از دانشگاه شیراز و همچنین مدرک کارشناسی ارشد و دکترای خود را به‌ترتیب در سال‌های ۱۳۸۲ و ۱۳۸۹ در درشتهٔ مهندسی رایانه (گرایش هوش مصنوعی) از دانشگاه صنعتی شریف دریافت کرد. زمینه‌های پژوهشی مورد علاقهٔ ایشان یادگیری ماشین و بازشناسی الگوی آماری، بازشناسی گفتار، بازشناسی الگوهای دنباله‌ای، یادگیری ژرف و کاربردهای آن بوده و در حال حاضر عضو هیأت علمی در دانشکدهٔ ریاضی، آمار و علوم رایانه دانشگاه تهران است.

نشانی رایانامهٔ ایشان عبارت است از:

babaali@ut.ac.ir

بابک رکابدار مدرک کارشناسی خود را در سال ۱۳۹۵ در رشتهٔ علوم رایانه از دانشگاه امیرکبیر دریافت کرد. همچنین مهر ماه سال ۱۳۹۵ به‌منظور دریافت مدرک کارشناسی ارشد (گرایش هوش مصنوعی) وارد دانشگاه تهران شد. زمینه‌های پژوهشی مورد علاقهٔ ایشان پردازش زبان طبیعی، یادگیری ماشین، بازشناسی دست‌نوشته و یادگیری ژرف است.

نشانی رایانامهٔ ایشان عبارت است از:

Babak.rekabdar@ut.ac.ir