A System for Printed Persian Documents

Z. Bahmani[1]

Received: 18, Apr. 2019
Accepted: 19, July 2019

Purpose: Introducing recognition systems and retrieval systems for Farsi printed document images and categorizing conducted researches with identifying strengths and weaknesses points of each category and presenting a retrieval system for Farsi printed document images in a new way.

Methodology: This paper is an applied research.An application designed and implied for Farsi printed document images retrieval. A new method in segmentation approach presented and implemented. A database including 50 Farsi documents scanned in 5 fonts provided for training and testing levels. Half of this database is used for training and other half for testing.

Findings: Persian printed document image recognition and retrieval systems follow one of three approaches: based on segmentation, based on sub-word shape and hybrid approach. The first approachhas received less attention dueto specific challenges. In this paper a system based on segmentation proposed and implemented and the results presented.

Conclusion: Results show that proposed system is anuseful method to Implement recognition systems and retrieval systems for Farsi printed document images.

DOI: 10.30484/nastinfo.2019.2173.1834

1. Lecturer, Computer Engineering, Behbahan Khatam Alanbia University of Technology, Behbahan, zahra. bahmani2009@gmail.com

Keywords:

Printed documents recognition of, Printed documents retrieval, Digital library, Sub-words, Sub-letters

NASTINFO

طراحی و پیاده‌سازی یک سیستم بازیابی اسناد چاپی فارسی

زهرا بهمنی[1]

هدف: معرفی، دسته‌بندی، و نقد پژوهش‌ها دربارهٔ سیستم‌های بازشناسی و بازیابی اسناد چاپی فارسی و پیشنهاد یک سیستم بازیابی اسناد چاپی با رویکردی نو.

روش‌شناسی: شیوه‌ای جدید با رویکرد جداسازی، طراحی و سپس پیاده‌سازی شده است. برای آموزش و آزمایش سیستم، پایگاه داده‌ای شامل تصویر ۵۰ صفحه متن فارسی در ۵ قلم مختلف فراهم آمد. از نیمی از این داده‌ها برای آموزش و نیمی دیگر برای آزمایش سیستم استفاده شد.

یافته‌ها: سیستم‌های بازشناسی یا بازیابی سند چاپی فارسی از یکی از سه رویکرد مبتنی بر جداسازی، مبتنی بر شکل کلی کلمه، و رویکرد ترکیبی پیروی می‌کنند. این پژوهش یک سیستم مبتنی بر جداسازی پیشنهاد و پیاده‌سازی و نتایج ارائه شده است.

نتیجه‌گیری: نتایج نشان می‌دهد نظام پیشنهادی، روش مناسبی برای پیاده‌سازی نظام‌های بازشناسی یا بازیابی اسناد فارسی است.

دریافت: ۹۸/۰۱/۳۰ **پذیرش:** ۹۸/۰۴/۲۹

۱. مربی، مهندسی کامپیوتر، دانشگاه صنعتی خاتم‌الانبیاء بهبهان
zahra.bahmani2009@gmail.com

کلیدواژه‌ها

بازشناسی اسناد چاپی، بازیابی اسناد چاپی، کتابخانه دیجیتالی، زیرکلمات، زیرحروف

مطالعات ملی کتابداری و سازماندهی اطلاعات،
دوره سی، شماره چهارم، زمستان ۱۳۹۸، ص ۴۶-۶۵.

DOI: 10.30484/nastinfo.2019.2173.1834

مقدمه

بازیابی محتوای متون تصویری که در کتابخانه‌های دیجیتالی، اتوماسیون‌های اداری، ادارات پست، و بانک‌ها رواج دارد چالش سیستم‌های بازیابی اطلاعات است. بازشناسی[1] و بازیابی[2] این متون زیرمجموعهٔ پردازش تصویر[3] است و با تکنولوژی بازشناسی نوری حروف[4] انجام می‌شود. اما در جاهایی مانند کتابخانه‌های دیجیتالی که با نسخه‌های قدیمی سروکار دارند، این روش به‌صرفه و عملی نیست. استفاده از سیستم‌های بازیابی به‌جای سیستم بازشناسی به‌سبب کم‌توانی تکنولوژی بازشناسی حروف، به‌ویژه در کار با تصاویر اسناد بی‌کیفیت متون قدیمی است. تصحیح دستی خروجی‌های بازشناسی نوری معمولاً امکان‌پذیر نیست.

در طراحی سیستم بازشناسی یا بازیابی اسناد چاپی سه رویکرد وجود دارد: رویکرد مبتنی بر جداسازی، رویکرد مبتنی بر شکل کلی کلمه، و رویکرد ترکیبی (عزمی، ۱۳۷۸). مشکل در رویکرد نخست، شکستن کلمه به حروف آن است؛ از این‌رو تمایل به استفاده از آن کم است. در رویکرد مبتنی بر شکل کلی کلمه، مشکل فراوانی فزایندهٔ زیرکلمات[5] فارسی و دشواری «آموزاندن» آنها به سیستم است. پژوهشهای انجام‌شده با این رویکرد بیش از رویکرد نخست بوده است. در رویکرد سوم چند پژوهشگر سعی کرده‌اند از مزایای هر دو روش به‌طور ترکیبی برای متون فارسی استفاده کنند.

تفاوت مهم دیگر سیستم‌های بازشناسی یا بازیابی اسناد چاپی در طرز کار آنها با نقطه‌های زیرکلمه‌هاست. اغلب سیستم‌ها در مرحله پردازش فقط بدنه زیرکلمات یا حروف را لحاظ می‌کنند و تشخیص نهایی به‌کمک نقطه‌ها را به‌مرحله پس از پردازش وامی‌گذارند. ولی راهکار دیگر بررسی هم‌زمان بدنه و نقطه‌هاست. یکی از مزایای روش اخیر در کار با اسناد کم‌کیفیت است. در این اسناد نقطه‌ها به بدنهٔ زیرکلمه متصل است و باعث افزایش خطا در بازشناسی می‌شود. ابراهیمی و کبیر (۱۳۸۴الف)، خسروی و کبیر (۱۳۸۸) و ابراهیمی و کبیر (۱۳۸۴ب) از معدود کارها به این روش هستند. در نظام‌آبادی و کبیر (۱۳۷۹) الگوریتمی برای جداکردن نقطه‌های چسبیده به بدنه ارائه شده است. در ادامه، کارهای انجام‌شده در هریک از سه رویکرد را مرور می‌کنیم.

1. Recognition
2. Retrieval
3. Image processing technology
4. Optical Character Recognition (OCR)
۵. Sub-Word بخشی از کلمه است که حرف قبلی به حرف بعدی نمی‌چسبد. حروف ا، د، ذ، ر، ز، ژ، و از این شمار هستند.

رویکرد جداسازی

همان‌طور که قبلاً اشاره شد چالش اصلی پیش روی سیستم‌های مبتنی بر جداسازی، شمار زیاد خطا در پیداکردن نقطه‌های جداسازی است. این خطا در سایر مراحل

ادامه می‌یابد و نتیجه نهایی را به‌شدت آسیب می‌زند. هرچند مراحل جداسازی و قطعه‌بندی بدنه کلمات به زیرکلمات و حروف، کار دشواری است؛ از سوی دیگر حل این مشکل، اندازه واژه‌نامه را نسبت به روش‌های مبتنی بر شکل کلی بسیار محدودتر خواهد کرد با روش‌های مبتنی بر شکل کلی کلمات روبه‌رو خواهند بود. در نتیجه این امر مراحل بعدی کار را بسیار راحت‌تر خواهد کرد.

در قطعه‌بندی زیرکلمات دو نوع رویکرد کلی وجود دارد: نخست سیستم‌هایی که سعی دارند محل دقیق انفصال حروف موجود در زیرکلمات را تشخیص دهند. دوم، سیستم‌هایی که زیرکلمات را به بخش‌های اغلب کوچک‌تر از حروف به‌نام زیرحرف[1] قطعه‌بندی می‌کنند. دلیل قطعه‌بندی به زیرحرف در دسته دوم، نامساوی‌بودن طول حروف و امکان خطا به‌خاطر شباهت در ساختار حروفی مانند «تیتر» و «سر» است. به‌سبب آنکه چنین کلماتی فقط به‌لحاظ نقطه باهم تفاوت دارند و نیز اغلب سیستم‌ها در فرایند پردازش، بدنه کلمات را بی‌توجه به نقطه‌های کلمه می‌شناسند، تشخیص نقطه انفصال حروف کار دشواری است. به‌همین دلیل، در این سیستم‌ها سعی شده است به‌جای مشخص‌کردن یک نقطه مشخص به‌عنوان نقطه انفصال دو حرف، ابتدا آن را زیرحروف تشخیص می‌دهند؛ سپس به‌کمک باقی ویژگی‌ها مانند نقطه‌ها، محل دقیق جداسازی را مشخص کنند.

با این رویکرد، عزمی (۱۳۷۸) الگوریتمی برای جداسازی حروف بدون توجه به نوع قلم ارائه کرده است. در کار او نقطه‌های جداسازی با اِعمال قواعدی در قالب یک نمودار حالت روی منحنی پیرامون کلمات تعیین شده است. برای بازشناسی نیز از دو الگوریتم استفاده شده است. الگوریتم اول از کدهای فریمن کانتور[2] حروف به‌عنوان ویژگی و از یک اتوماتون آماری برای طبقه‌بندی، و الگوریتم دوم از تبدیل هاف[3] با روش فازی به‌عنوان ویژگی و از شبکه پس‌انتشار خطا[4] به‌عنوان طبقه‌بندی‌کننده استفاده کرده است. نتایج این پژوهش برای جداسازی حروف به‌ازای متن چاپی شامل ۱۱۰۰۰ حرف با قلم‌های مختلف، ۹۹ درصد و برای بازشناسی روی مجموعه حروف شامل ۱۱۵۰۰ نمونه از ۱۰ قلم مختلف به‌ازای الگوریتم اول ۹۷/۱۳ و به‌ازای الگوریتم دوم ۹۸/۳۲ گزارش شده است.

شمسی، رسولی کناری، و شادروان (۱۳۸۸) برای جداسازی حروف، ابتدا با محاسبه هیستوگرام عمودی و پویش کلمه از راست به چپ، قسمت‌هایی را که برآمدگی کمتر از متوسط دارند به‌عنوان مرزهای بالقوه حروف علامت‌گذاری کرده‌اند. سپس این مرزها در صورت داشتن شرایطی مانند حداقل عرض لازم و وجود تغییر

1. Sub-Letter

۲. منحنی خطی تکه‌ای بسته است که از مرکز تمام پیکسل‌هایی عبور می‌کنند که از چهار طرف به پس‌زمینه خارجی و نه پیکسل دیگر متصل هستند.

۳. روشی برای استخراج ویژگی‌ها در آنالیز تصاویر، بینایی رایانه‌ای، و پردازش تصویر دیجیتال است.

۴. این الگوریتم جزء الگوریتم‌های یادگیری با نظارت است که اساساً از دو مسیر رفت‌وبرگشت تشکیل شده است.

محسوس در هیستوگرام عمودی، نقطه‌های جداسازی را مشخص می‌کنند. شکل ۱ نمونه‌ای از تشخیص مرزهای بالقوه و تشخیص مرزهای واقعی را نشان می‌دهد. در این پژوهش، نتیجه جداسازی حروف برای ۱۲ قلم مختلف در ۶ اندازه بیش از ۹۷ درصد گزارش شده است.

شکل ۱. تشخیص مرزهای بالقوه و بررسی شرایط تبدیل به مرزهای واقعی (شمسی و همکاران، ۱۳۸۸)

عزمی و کبیر (۱۳۷۸) روشی را برای جداسازی حروف ارائه داده‌اند. در این روش، نقطه‌های جداسازی اولیه با اِعمال قواعدی به‌صورت یک گرامر روی منحنی پیرامونی تعیین شده؛ سپس در مرحله پس‌پردازش، نقطه‌های نهایی جداسازی مشخص شده است. در این پژوهش الگوریتم روی حدود ۱۱۰۰۰ حرف از ۲۰ قلم مختلف انجام شد و نتایج صحیح حدود ۹۹ درصد گزارش شده است. نظام‌آبادی‌پور و کبیر (۱۳۸۳) با اصلاح الگوریتم عزمی، الگوریتمی برای متون با کیفیت پایین ارائه داده‌اند. در این پژوهش به اصلاح روش برچسب‌زنی کانتور بالایی و تکمیل قواعد جداسازی پراخته شده است.

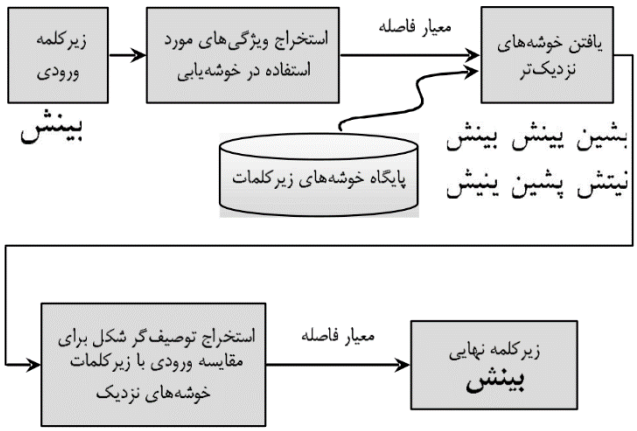
بهمنی و عزمی (۲۰۱۱الف) یک سیستم بازیابی سند مبتنی بر جداسازی ارائه داده‌اند و برای تشخیص زیرحروف از ویژگی‌های شکلی زیرحروف استفاده کرده‌اند که وابسته به قلم نیستند. هریک از چهار گروه زیرحروف، درخت تصمیم مربوط به خود دارند. در این درخت‌ها از ویژگی‌هایی مانند داشتن حفره در زیرحروفی مانند «ه»، وجود الف بالای زیرحرف مانند «ط»، و سایر ویژگی‌های ناوابسته به قلم استفاده شده است. همچنین، بهمنی و عزمی (۲۰۱۱ب) برای بهبود روش قبل، از شبکه عصبی RBF[1] استفاده کرده‌اند. در اینجا از ویژگی نمایه[2] در چهار جهت استفاده شده است. در این پژوهش، نرخ بازیابی صحیح حروف افزایش یافته است. کارهای دیگری نیز با این رویکرد انجام شده است (رفیعی کراچی، ۱۳۷۳؛ خسروی و کبیر، ۱۳۸۶؛ سرابی نوبخت، ۱۳۹۲؛ بهمنی، ۱۳۹۰).

۱. شبکه‌هایی متشکل از سه لایه پیشرو با اتصالات کامل هستند که با دریافت یک بردار ورودی در لایه اول و انتشار نتایج به لایه میانی، خروجی شبکه در لایه آخر (لایه خروجی) دریافت می‌شود.

۲. به‌ازای هر سطر یا ستون از اطلاعات دودویی تصویر حرف یا زیرکلمه، فاصله مرز مشخص‌شده برای تصویر مذکور تا نزدیک‌ترین نقطه سیاه تصویر محاسبه می‌شود.

رویکرد شکل کلی کلمه

در رویکرد مبتنی بر شکل کلی کلمه، کلمه یا زیرکلمه به‌حروف تشکیل‌دهنده آن شکسته نمی‌شود، بلکه سعی بر آن است که زیرکلمه به‌صورت یکپارچه تشخیص داده شود. در این رویکرد هر زیرکلمه یک دسته منحصربه‌فرد است. محاسبه تعداد کل زیرکلمات موجود در زبان فارسی کار دشواری است؛ ولی پژوهش‌ها وجود بیش از ۱۰۰۰۰ زیرکلمه را گزارش داده‌اند (خسروی و کبیر، ۱۳۸۸). یعنی تصویر یک زیرکلمه باید با مشخصات بیش از ۱۰۰۰۰ زیرکلمه دیگر مقایسه شود. این کار بسیار زمان‌بر و احتمال خطا در آن به‌شدت بالاست. چاره کار آن بوده است که زیرکلمات به‌شاخه‌های متعدد دسته‌بندی شوند به‌شکلی که هر شاخه دربرگیرنده زیرکلمات شبیه به‌هم با ویژگی‌های یکسان باشد. برای مرحله تشخیص یا بازیابی، ابتدا برای هر کلمه شاخه یا شاخه‌های مربوط تشخیص داده می‌شود؛ سپس از طریق استخراج ویژگی‌هایی که تفاوت زیرکلمات را به‌شکل بارزتر نشان می‌دهند، زیرکلمه مدنظر تشخیص داده می‌شود. در شکل ۲ فرایند کلی بازشناسی زیرکلمات، با استفاده از شکل کلی زیرکلمه آمده است.



شکل ۲. فرایند کلی بازشناسی با استفاده از شکل کلی (خسروی و کبیر، ۱۳۸۸)

پیشینه رویکرد شکل کلی کلمه

عزمی و کبیر (۱۳۸۳) الگوریتمی مبتنی بر شکل کلی زیرکلمات ارائه داده‌اند. آنها سه واژه‌نامه با استفاده از ویژگی‌های مکان مشخصه[1]، توصیف‌کننده‌های فوریه[2]، و برچسپ کانتور بالایی طراحی کرده‌اند. برای هریک از این سه ویژگی با استفاده از

۱. به هر نقطه از پس‌زمینه تصویر کلمه یا حرف مدنظر یک کد اختصاص می‌دهد که متناظر با تعداد برخورد یک خط عمودی و یک خط افقی ترسیم‌شده از آن نقطه با بدنه تصویر مدنظر است.
۲. یک ابزار ریاضی، برای تبدیل و توصیف مکانی تصویر به اجزای فرکانسی آن است.

همه زیرکلمات موجود واژه‌نامه طراحی شده است. هریک از مدخل‌های واژه‌نامه متناظر با یک همسایگی (شاخه) است. برای استخراج ویژگی، ابتدا نقطه‌ها و علائم از بدنه زیرکلمات جدا شده است. در مرحله بازشناسیِ زیرکلمه مجهول، به واژه‌نامه مراجعه شده و مدخل و درنتیجه همسایگی متناظر با آن مشخص می‌شود. در این پژوهش انتخاب درست مدخل به‌معنای صحت بازشناسی بوده است. درنهایت، نتایج بازشناسی برای قلم‌های آموزش‌داده‌شده برای ۳ واژه‌نامه ذکرشده به‌ترتیب ۹۸/۲، ۹۸/۶۵، و ۹۹/۴ درصد گزارش شده است.

ابراهیمی (۱۳۸۴) از شکل کلی زیرکلمات، برای بازیابی تصاویر نامه‌های اداری استفاده کرده است. در همین پژوهش برای بازشناسی متن از خوشه‌بندی تصاویر[1] ۱۲۷۰۰ زیرکلمه به ۴۰۰ خوشه و انتخاب ۱۰ خوشه نزدیک‌تر استفاده شده است. پژوهش ذکرشده روی دو مجموعه مختلف یکی شامل ۵ تصویرِ تولیدشده با ۵ قلم متفاوت و دیگری شامل ۴ تصویرِ پویش‌شده از منابع مختلف آزمایش شده است. درصد بازشناسی به‌ازای مجموعه اول ۹۶ درصد و به‌ازای مجموعه دوم ۹۰ درصد گزارش شده است. از سایر کارهای انجام‌شده با این رویکرد می‌توان به ابراهیمی و کبیری (۱۳۸۵؛ ۱۳۸۴) اشاره کرد.

خسروی و کبیر (۱۳۸۸) سیستم بازشناسی متون فارسی با دو رویکرد مختلف ارائه داده‌اند: رویکرد اول مبتنی بر شکل کلی زیرکلمات به‌همراه نقطه‌ها و علائم و رویکرد دوم مبتنی بر شکل کلی زیرکلمات با حذف علائم و نقطه‌هاست. در این پژوهش از الگوریتم ISODATA[2] برای خوشه‌بندی استفاده و مراکز خوشه‌ها از طریق یک الگوریتم خوشه‌بندی سلسله‌مراتبی محاسبه شده است. در هر دو الگوریتم مانند اغلب سیستم‌های مبتنی بر شکل کلی زیرکلمه، بازشناسی در دو مرحله انتخاب خوشه‌های نزدیک به‌زیرکلمه ورودی و سپس گزینش نزدیک‌ترین زیرکلمه از میان زیرکلمات خوشه‌های انتخابی انجام شده است. نامور (۱۳۹۵) چهار روش برای بهبود بازشناسی تصاویر متون فارسی ارائه داده است. روش اول، بهبود بازشناسی با استفاده از اطلاعات در سطح زیرکلمه و الگوریتم ویتربی[3] است. در روش دوم، تغییری در الگوریتم ویتربی اعمال شده و در روش‌های سوم و چهارم از اطلاعات در سطح کلمه به‌لحاظ لغوی و نحوی استفاده شده است. در این پژوهش نسخه ۳,۰۴ تسراکت[4] استفاده شده است. کارهای مشابه دیگری نیز با این رویکرد انجام شده است (Nasrollahi & Ebrahimi, 2013; Pourasad, Hassibi, & Ghorbani, 2012; 2013؛ داودی و کبیر، ۱۳۹۳؛ نامور و عزمی، ۱۳۹۶).

۱. فرایند طبقه‌بندی به‌گونه‌ای که نمونه‌های قرارگرفته در یک گروه شبیه‌تر از نمونه‌های قرارگرفته در سایر گروه‌ها باشند.
۲. نسخه بهبودیافته الگوریتم K-means است.
۳. برای پیداکردن محتمل‌ترین مسیر از حالت‌های پنهان.
۴. OCR موتور منبع باز است که بین سال‌های ۱۹۸۴ و ۱۹۹۴ در شرکت HP ایجاد شد.

رویکرد ترکیبی و پیشینهٔ آن

در رویکرد ترکیبی از مزایای هر دو رویکرد پیشین برای بهبود نتایج استفاده می‌شود. برای این منظور معمولاً سعی بر آن است که زیرکلمات حتی‌المقدور به حروف یا زیرحروف جداسازی شود. در جاهایی که امکان جداسازی وجود ندارد یا خطا زیاد است، از شکل کلی زیرکلمات استفاده می‌شود. ازجمله کارهای انجام‌شده با این رویکرد پایان‌نامه دکترای عزمی (۱۳۷۸) است. در این سیستم حروف شاخص زیرکلمه جداسازی و بازشناسی شده است. همچنین یک واژه‌نامه تصویری برای بدنهٔ زیرکلمات طراحی شده است. در بازشناسی زیرکلمات از موقعیت و نوع نقطه‌ها و علائم استفاده شده است. درنهایت، به‌کمک مدل مخفی مارکوف[۱] و یک الگوریتم ویتربی تغییر یافته و با استفاده از اطلاعات آماری احتمال رخداد متوالی حروف شاخص و امتیاز حاصل از بازشناسی آنها، بازشناسی زیرکلمات انجام شده است.

در مرتضوی طباطبائی (۱۳۹۱)، ابتدا برخی حروف که امکان جداسازی و بازشناسی با دقت زیاد را داشته‌اند به‌عنوان حروف شاخص مشخص شده‌اند. سپس با استفاده از الگوریتم ژنتیک[۲] و عملگرهای مورفولوژی[۳]، مجموعه‌ای از عملگرهای مورفولوژی به‌همراه ساختار بهینه‌ای که قادر به توصیف گروه‌ها باشند مشخص شده است. زیرکلمات این گروه‌ها براساس حروف شاخص کدگذاری شده‌اند. یک واژه‌نامه تصویری نیز ایجاد و بازشناسی با استفاده از RBF و ویژگی موجک[۴] انجام شد.

روش‌شناسی

در این پژوهش یک سیستم بازیابی اسناد چاپی فارسی طراحی شده و به‌وسیله نرم‌افزار متلب پیاده‌سازی شده است. برای آموزش و ارزیابی سیستم، ۵۰ سند فارسی در ۵ قلم پرکاربرد چاپ‌شده پویش شد. نیمی از این اسناد در مرحله آموزش به سیستم ارائه شد و از بقیه در مرحله آزمایش برای ارزیابی استفاده شد. در سیستم‌های بازشناسی و بازیابی معمولاً کلمات به بخش بدنه و نقطه‌ها تقسیم می‌شوند تا کار با آن راحت‌تر شود. در این پژوهش نیز هر کلمه به دو بخش بدنه و نقطه‌ها تقسیم شده است. ازآنجاکه حروف جدا در ساختار زبان فارسی زیرکلمات را شکل می‌دهد از آن برای بررسی بدنه کلمه از طریق بررسی بدنهٔ زیرکلمات تشکیل‌دهنده آن استفاده شد.

از رویکرد مبتنی بر جداسازی به روشی تازه در این پژوهش استفاده شده است که نیاز به تخمین نقطه دقیق جداسازی ندارد. در این روش، عناصر اصلی تشکیل‌دهنده

۱. مدلی آماری است که در آن سیستم مدل‌شده به‌صورت یک فرایند مارکوف با حالت‌های مشاهده‌نشده (پنهان) فرض می‌شود.
۲. در این الگوریتم نحوه تکامل ژنتیکی موجودات زنده شبیه‌سازی می‌شود.
۳. ابزار قدرتمند برای استخراج اطلاعات ساختاری از تصویر است.
۴. موجک در پردازش تصویر، طبقه‌بندی تصاویر، حذف اختلالات تصاویر، و فشرده‌سازی تصاویر استفاده می‌شود.

حروف به‌عنوان بخش‌های اصلی از بدنه استخراج شد و سایر بخش‌های بدنه به‌عنوان اتصال‌دهنده عناصر اصلی حذف شده است؛ زیرا این بخش‌ها نقشی در تعیین هویت حروف ندارند و صرفاً به‌دلیل پیوسته‌نویسی در زبان فارسی به بدنه کلمه اضافه شده‌اند. برای تشخیص این عناصر اتصال‌دهنده از روش تشخیص خط زمینه استفاده شد. بدین‌گونه، شکل بخش‌هایی به‌عنوان نامزد برای عناصر اتصال‌دهنده مشخص و با ارزیابی دقیق‌تر نقطه‌های اضافه حذف شده است. پس از استخراج عناصر اصلی، این عناصر که در این پژوهش زیرحرف نامیده شدند، به چهار دستهٔ زیرحرف ابتدایی، میانی، انتهایی، و مجزا تقسیم شده‌اند. بدین ترتیب، چهار واژه‌نامه طراحی شد. هر زیرحرف با توجه به موقعیت قرارگیری در ساختار بدنه زیرکلمه، به یکی از این واژه‌نامه‌ها تعلق دارد. در مرحله تشخیص زیرحرف‌ها، ویژگی‌های نمایه در چهار جهت اصلی از این عناصر استخراج شد. با کنار هم قراردادن این چهار نمایه، بُردارِ ویژگی تشکیل شده است. در مرحله تشخیص از شبکه عصبی RBF استفاده شد. شبکه‌های عصبی یکی از روش‌های پرکاربرد در بازشناسی نوری حروف است (Reynaldo Phangtriastu, Harefa, & Felita Tanoto, 2017). به‌ازای هر دسته از زیرحرف‌های یک شبکه عصبی آموزش داده شده و در مرحله تشخیص از نتایج این شبکه‌ها استفاده شده است.

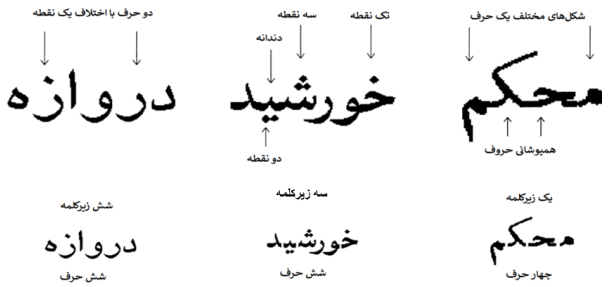
در هر دسته به‌ازای هر زیرحرف یک شبکه عصبی نهایی برای تأیید نهایی زیرحرف به سیستم آموزش داده و از آن استفاده شد. بدین ترتیب، بدنه هر زیرکلمه به‌وسیله کد مربوط به زیرحرف‌های تشکیل‌دهنده آن در واژه‌نامه مربوط به‌زیرحرف کد شده است. با کنار هم قرارگرفتن کد بدنه زیرکلمات، کد بدنه کلمه تشکیل شد. برای اضافه‌کردن نقطه‌ها به بدنهٔ زیرکلمه، فاصله نقطه‌ها تا زیرحرف‌ها محاسبه و هر نقطه به نزدیک‌ترین زیرحرف منتسب شد. برای ارزیابی سیستم ذکرشده، نیمی از صفحات پویش‌شده که از آنها در مرحله آموزش استفاده نشد به‌وسیله سیستم پیشنهادی به‌ازای حروف الفبای فارسی بازیابی شد. برای بازیابی حروف، ابتدا حرف مدنظر به زیرحروف تشکیل‌دهنده حرف کد؛ سپس کد تولیدشده با کد مربوط به‌تصاویر اسناد اسکن‌شده تولید سیستم مقایسه شد. در جاهایی که کد انطباق داشت در تصویر سند مربوط محل حرف نشان داده شد. نتایج ارزیابی سیستم در بخش نتایج آمده است.

ویژگی‌های نوشتاری زبان فارسی

خط فارسی از راست به چپ نوشته می‌شود و مانند خیلی از زبان‌های دیگر واحد آن کلمه است. در کلمات فارسی، بیشتر حروف از دو طرف به حروف مجاور خود

می‌چسبند. ولی حروف {ا، د، ذ، ر، ز، ژ، و} به حرف بعد خود نمی‌چسبند. بنابراین هر کلمه می‌تواند شامل یک یا چند بخش باشد که به آنها زیرکلمه می‌گویند. هر کلمه از یک یا چند زیرکلمه تشکیل می‌شود. یک زیرکلمه از یک یا چند حرف چسبیده به‌هم تشکیل می‌شود (کارگروه خط و زبان فارسی در محیط رایانه‌ای، ۱۳۷۸). موقعیت حروف در زیرکلمه، حروف موجود در آن را به چهار دستهٔ حروف ابتدایی، میانی، انتهایی، و مجزا تقسیم می‌کند. شکل حروف یکسان در موقعیت‌های متفاوت، با هم تفاوت دارد. ۱۷ حرف از ۳۲ حرف فارسی یک، دو، و سه نقطه دارند. این نقطه‌ها در بالا یا پایین حرف قرار می‌گیرند. تفاوت برخی حروف در زبان فارسی فقط در تعداد یا محل نقطه‌هاست مانند {س و ش}. در نوشتار فارسی به‌ندرت از اِعراب {ً، ، ، ، ، ُ، ، ء} استفاده می‌شود و ممکن است این علائم به‌اشتباه نقطه تشخیص داده شوند.

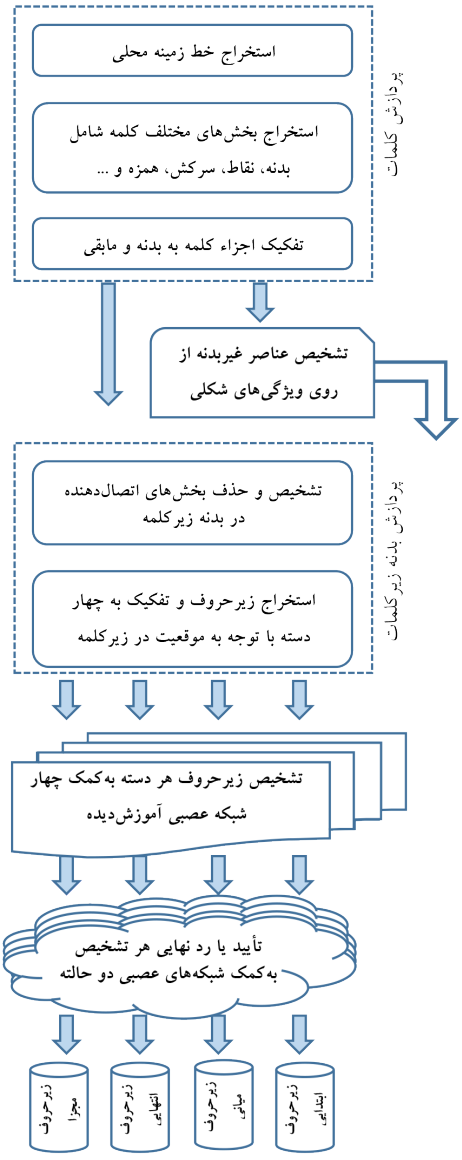
مهم‌ترین ویژگی دشواری کار با اسناد فارسی و زبان‌های مشابه، پیوستگی حروف در زیرکلمات است. این وضع کارها در این زمینه را به سه شاخه پیش‌گفته تقسیم کرده است. تلاش همه آنها دست‌یافتن به روشی برای غلبه بر این ویژگی در زبان فارسی است. نکته دیگر در به‌هم‌چسبیدگی حروف که کار را برای رویکرد مبتنی بر جداسازی دشوارتر می‌کند، یکسان‌نبودن طول حروف است. برای مثال، حروف {ا، س} از کوتاه‌ترین و بلندترین حروف الفبای فارسی هستند. تفاوت طولی آنها بسیار زیاد است. در نوشتار فارسی، برخی اعداد ساختاری شبیه حروف دارند و گاه تفکیک آنها برای سیستم مشکل‌ساز است. مانند عدد ۱ و حرف {ا}. درنهایت، تغییر شکل بعضی حروف در برخی از کلمات در اغلب قلم‌های فارسی مانند {لا}، باعث ایجاد حالاتی استثنایی می‌شود. دیگر ویژگی زبان فارسی نسبت حروف به خط زمینه است: مانند {ا، ک، ت} بالای خط زمینه و برخی {ی، ر، و} زیر خط زمینه قرار می‌گیرند. برخی نیز {س، ش، ح} بخشی بالا و بخشی پایین خط زمینه قرار می‌گیرند (شکل ۳).



شکل ۳. برخی ویژگی‌های نگارش فارسی (Azmi & Kabir, 2001)

ساختار سیستم پیشنهادی

شکل ۴ بخش‌های سیستم را نشان می‌دهد. زیرحرف‌ها براساس محل آن در بدنه زیرکلمه چهار دسته‌اند: زیرحرف آغازین، میانی، پایانی، و جدا. هر دسته به‌کمک یک شبکه عصبی RBF آموزش و تشخیص داده شده‌اند.



شکل ۴. ساختار کلی سیستم پیشنهادی

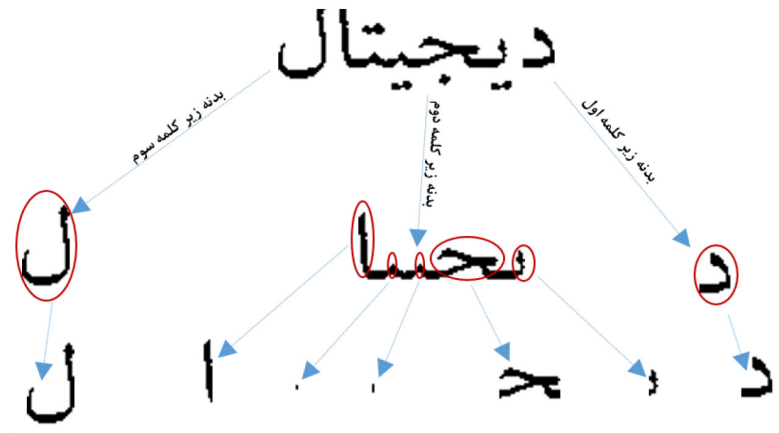
بخش‌های زیرکلمات در روش پیشنهادی ما

در سیستم‌های جداساز، بدنهٔ زیرکلمه به حرف یا زیرحرف شکسته می‌شود. ما در اینجا روشی را برای بخش‌بندی بدنه زیرکلمات ارائه کرده‌ایم که نیازی به تخمین نقطه انفصال حروف ندارد. در این روش و با نگاه نو به ساختار نوشتار فارسی، از یک‌سو، بار محاسباتی برای تشخیص نقطه دقیق انفصال را کاسته‌ایم و از سوی دیگر، خطای ناشی از تشخیص غلط را حذف کرده‌ایم.

جداسازی

در این روش، حروف موجود در هر زیرکلمه به دو بخش تقسیم می‌شود: عناصر اصلی حرف و قسمت‌های اتصال‌دهنده. در ساختار نوشتار فارسی، قسمت‌های اتصال‌دهنده نقشی در تعیین هویت حرف ندارند و فقط وظیفه اتصال عناصر اصلی را دارند. در سیستم پیشنهادی، این قسمت‌ها از بدنه زیرکلمه حذف می‌شود تا عناصر اصلی که معرفی‌کننده هویت حرف‌اند استخراج شود (شکل ۵).

عناصر اصلی حروف براساس موقعیت قرارگیری آنها در زیرکلمات به‌چهار دستهٔ عناصر آغازین، میانی، پایانی، و جدا تقسیم شده‌اند. چهار واژه‌نامه برای هریک از این چهار دسته تعریف شده است.



شکل ۵. بخش‌بندی بدنه زیرکلمات، کلمه دیجیتال و حذف قسمت‌های اتصال‌دهنده و استخراج عناصر اصلی حروف

برای مشخص‌کردن بخش‌های اتصال‌دهنده زیرحروف از سه ویژگیِ خط زمینه محلی، نمایهٔ بالای زیرکلمه، و هیستوگرام عمودی استفاده شده است. بعد از مشخص‌کردن نمایهٔ بالای زیرکلمه، نقطه‌هایی از زیرکلمه که نمودار آن در فاصله کمتر از عرض قلم با خط زمینه قرار دارند به‌عنوان نامزد عنصر اتصال‌دهنده زیرحروف انتخاب شد. از میان نقطه‌های نامزد برای بخش اتصال‌دهنده، قسمت‌هایی که با عناصر اصلی حرف هم‌پوشانی دارند حذف شده است. برای مشخص‌کردن این قسمت‌ها از هیستوگرام عمودی زیرکلمه استفاده شده است. در اینجا نقطه‌های نامزدی که ارتفاع هیستوگرام عمودی آنها از ۱/۵ برابر عرض قلم بیشتر بوده رد شده و باقی نقطه‌ها به‌عنوان نقطه‌های اتصال‌دهنده زیرحروف معرفی شده‌اند.

استثناها

با توجه به اینکه ما برای تشخیص بخش‌های اتصال‌دهنده از خط زمینه محلی به‌ازای هر کلمه استفاده کرده‌ایم گاهی خط زمینه محلی اشتباه تشخیص داده می‌شود. این مسئله زمانی اتفاق می‌افتد که کلمه فقط از حروفی که منفصل هستند تشکیل شده باشد. مانند کلمات ری، وی، آن. برای اصلاح و پیشگیری از این قبیل، خط زمینهٔ محلی با خط زمینهٔ سراسری مقایسه و در صورت اختلاف بیشتر از حدِ مجاز، خط زمینه سراسری به‌جای خط زمینه محلی جایگزین می‌شود.

واژه‌نامه طراحی‌شده

در سیستم‌های مبتنی بر جداسازی معمولاً اندازه واژه‌نامه در مقایسه با رویکرد شکل کلی بسیار کوچک‌تر تعریف می‌شود و بیشتر متناسب با تعداد حروف الفباست. در سیستم پیشنهادی ما چهار واژه‌نامه تعریف شده است. با توجه به اینکه در مرحله تشخیص عناصر اصلی، هر عنصر با توجه به موقعیت قرارگیری آن در زیرکلمه به یکی از چهار واژه‌نامه تعلق خواهد داشت، عملاً اندازه واژه‌نامه نهایی برابر با تعداد مدخل‌های مشترک واژه‌نامه‌ها، به‌علاوه حالت‌های خاص در هریک از واژه‌نامه‌هاست. در شکل ۶، تصویر چهار واژه‌نامه و واژه‌نامه نهایی آمده است. امتیاز خاص این واژه‌نامه علاوه بر ناوابستگی به تشخیص نقطه‌ای خاص برای جداسازی تعداد محدود مدخل‌ها و تمایز میان اعضای آن است. این با ابزارهایی مانند شبکه‌های عصبی به‌راحتی تشخیص داده می‌شود.

مدخل	دیکشنری نهایی	دیکشنری یک	دیکشنری دو	دیکشنری سه	دیکشنری چهار
۱	ا	أ	إ	إ	ا
۲	د	د	ـد	ـد	...
۳	ح	حـ	ـحـ	ـح	ح
٤	د	د	ـد	ـد	د
٥	ر	...	...	ـر	ر
٦	ص	صـ	ـصـ	...	...
۷	ط	طـ	ـطـ	ـط	ط
۸	ع	عـ	ـعـ	ـع	ع
۹	ق	قـ	ـقـ	ـق	ق
۱۰	ک	کـ	ـکـ	...	...
۱۱	ل	...	...	ـل	ل
۱۲	م	مـ	ـمـ	ـم	م
۱۳	ن	...	...	ـن	ن
۱٤	و	...	...	ـو	و
۱٥	ه	هـ	ـهـ	ـه	ه
۱٦	ی	...	...	ـی	ی
۱۷	لا		...	ـلا	لا

شکل ۶. واژه‌نامه‌های طراحی‌شده مربوط به چهارگروه زیرحروف و واژه‌نامه نهایی

حذف بعضی از خطاهای جداسازی از راه گسترش واژه‌نامه

برخی اشتباهات در مرحله جداسازی شامل شکستن حرف «د» به دو قسمت است، که به‌خاطر نوشتن این حرف در برخی از قلم‌ها در راستای خط زمینه اتفاق می‌افتد. برای رفع این قبیل اشتباهات در اسناد چاپی، در واژه‌نامه عناصر ابتدایی، مدخلی برای «د»

به‌صورت عنصری ابتدایی تعریف شده است. یکی دیگر از این مشکلات، زوج‌حرف «لا» است، که در بیشتر قلم‌های فارسی بدین شکل نوشته می‌شود و در بیشتر سیستم‌های مبتنی بر جداسازی به‌صورت استثنا بررسی می‌شود. در سیستم پیشنهادی ما برای غلبه بر این مشکل، زوج‌حرف «لا» یک عنصر رفتار شده است. این عنصر در واژه‌نامه عناصر پایانی و عناصر مجزا یک مدخل جدا درنظر گرفته شده است.

کدگذاری

در سیستم پیشنهادی، ما برای کدگذاری نهایی از موقعیت زیرحروف چشم‌پوشی کردیم؛ زیرا عملاً این مسئله به ساختار نوشتاری زبان فارسی برمی‌گردد و در مفهوم کلمه بی‌تأثیر است. یک واژه‌نامه نهایی تعریف و به‌ازای زیرحروف با معنای یکسان یک مدخل تعریف کرده‌ایم. بدین ترتیب، برای جستجوی کلمه کلیدی در کد تولیدشده، فرایند جستجو در کل کلمه بدون توجه به انفصال در قسمت‌هایی انجام می‌شود که حروف غیرمتصل وجود دارد. این امر جستجوی کلمه کلیدی را به‌عنوان بخشی از یک کلمه دیگر آسان می‌کند.

بازیابی

برای بازیابی ابتدا کلمه کلیدی از کاربر دریافت و حروف موجود در آن مشخص می‌شود. زیرحرف‌های موجود در هر حرف مشخص و کد مربوط به آن با استفاده از واژه‌نامه استخراج می‌شود. بدین ترتیب، با کنار هم گذاشتن کد زیرحروف، حروف موجود در کلمه کلیدی، کد مربوط به کلمه تولید می‌شود. سپس با کد مربوط به تصویر صفحات پویش‌شده تطبیق داده می‌شود.

برای انتساب نقطه‌ها و علائم به حروف مربوط به آنها، موقعیت دقیق نقطه‌ها و علائم و همچنین زیرحروف در زیرکلمه تعیین شده است. سپس هر نقطه به زیرحرفی که موقعیت ابتدای آن با موقعیت ابتدای نقطه یا علامت مدنظر نزدیک‌تر است منتسب شده است.

یافته‌ها

برای ارزیابی سیستم پیشنهادی ۵۰ سند در ۵ قلم مختلف پویش و برای ارزیابی به سیستم ارائه شده است. از نیمی از اسناد برای آموزش به سیستم و نیمی دیگر برای آزمون سیستم استفاده شده است.

نتایج دسته‌بندی

در سیستم پیشنهادی در مرحله تشخیص عناصر از شبکه عصبی RBF استفاده کردیم. ویژگی این سیستم نمایه در چهار جهت بالا، پایین، چپ، و راست است. مرحله دسته‌بندی به‌سبب محدودشدن تعداد دسته‌ها و تمایز ظاهری عناصر اصلی به‌راحتی انجام می‌شود. در پیاده‌سازی سیستم پیشنهادی شبکه‌های آموزش‌دیده توانستند همه زیرحرف‌ها را با خطای کمتر از ۱ درصد و صحت بالای ۹۹ درصد تشخیص دهند.

نتایج بازیابی

برای ارزیابی نهایی سیستم پیشنهادی، ۳۲ حرف الفبای فارسی به‌وسیله سیستم بازیابی شد که نتایج آن در جدول ۱ آمده است.

جدول ۱. نتایج بازیابی حروف الفبا

ردیف	حرف	درصد	ردیف	حرف	درصد	ردیف	حرف	درصد	ردیف	حرف	درصد
۱	الف	۹۹/۷۸	۹	خ	۱۰۰	۱۷	ص	۹۹/۱۳	۲۵	ک	۹۹/۷۰
۲	ب	۹۷/۶۰	۱۰	د	۹۹/۰۵	۱۸	ض	۱۰۰	۲۶	گ	۱۰۰
۳	پ	۹۷/۷۲	۱۱	ذ	۱۰۰	۱۹	ط	۱۰۰	۲۷	ل	۹۹/۶۴
۴	ت	۹۸/۹۴	۱۲	ر	۹۹/۶۲	۲۰	ظ	۱۰۰	۲۸	م	۹۹/۴۶
۵	ث	۹۷/۴۷	۱۳	ز	۹۸/۱۱	۲۱	ع	۹۸/۱۵	۲۹	ن	۹۹/۲۳
۶	ج	۹۸/۶۱	۱۴	ژ	۱۰۰	۲۲	غ	۱۰۰	۳۰	و	۹۹/۴۲
۷	چ	۹۸/۳۳	۱۵	س	۹۸/۳۶	۲۳	ف	۹۸/۹۷	۳۱	ه	۹۹/۲۱
۸	ح	۹۹/۲۴	۱۶	ش	۹۷/۳۵	۲۴	ق	۹۷/۵۹	۳۲	ی	۹۹/۳۴

نتیجه‌گیری

در سیستمی که ما پیشنهاد کرده‌ایم شیوه جدیدی از بخش‌بندی زیرکلمات فارسی برای بازشناسی یا بازیابی اسناد چاپی فارسی ارائه شده است. مهم‌ترین مزیت سیستم پیشنهادی نیازنبودن به تشخیص نقطه دقیق انفصال حروف و پرهیز از خطای ناشی از آن است.

در سیستم پیشنهادی ما تعداد زیرحروف در هریک از دسته‌های زیرحروف آغازین، میانی، پایانی، و جدا کمتر از ۱۵ زیرحرف است. ازآنجایی‌که هر دسته از زیرحرف‌ها فقط در دسته خود مقایسه می‌شوند، تعداد دسته‌ها کمتر از ۱۵شده است،

که این تعداد حتی از تعداد حروف الفبای فارسی کمتر است؛ بنابراین دسته‌بندی با خطای بسیار کمتر و سرعت بیشتر اتفاق می‌افتد. در سیستم پیشنهادی ما از شبکه عصبی RBF برای تشخیص زیرحروف استفاده شده است. با توجه به اندک‌بودن دسته‌ها و قدرت شبکه عصبی، کار تشخیص با دقت بالای ۹۹ درصد به‌ازای هر زیرحرف ممکن شده است.

درنهایت به‌منظور آزمون سیستم، از صفحات چاپ‌شده و پویش‌شده استفاده کردیم و آن را با ۳۲ حرف الفبای فارسی ارزیابی کردیم. همان‌طور که در جدول ۱ آمده است، بازیابی اسناد مطلوب بود. خطایی که در برخی حروف رخ داد ناشی از چرخش صفحه در هنگام پویش بود. این مشکل هرگاه برای تشخیص خط زمینه به چرخش صفحه نیاز نباشد نتیجه می‌تواند بسیار بهبود یابد (Boukharouba, 2017).

از روش پیشنهادی ما می‌توان برای اسناد دست‌نویس نیز استفاده کرد؛ زیرا معضل مقایسه بدنه زیرکلمه‌ها در دست‌نویس‌ها به تفاوت بسیار بزرگ‌تر و تشخیص نقطه‌های جداسازی نیز دشوارتر است. روش پیشنهادی ما می‌تواند در کار بر اسناد دست‌نویس نیز این مشکلات را حل کند و حاصل کار سیستم‌های بازشناسی و بازیابی آنها را بهبود ببخشد.

مآخذ

ابراهیمی، افشین (۱۳۸۴). *استفاده از شکل کلی زیر- کلمات چاپی در بازیابی تصویر مستندات و بازشناسی متون فارسی*. پایان‌نامه دکتری، دانشگاه تربیت مدرس، تهران.

ابراهیمی، افشین؛ کبیر، احسان‌اله (۱۳۸۴الف، ۴-۶ بهمن). بازشناسی زیر-کلمات چاپی با درنظرگرفتن نقاط آنها. مقاله ارائه‌شده در *یازدهمین کنفرانس بین‌المللی سالانه انجمن کامپیوتر ایران، تهران*. بازیابی ۲۶ دی ۱۳۹۸، از https://www.civilica.com/Paper-ACCSI11-ACCSI11_077.html

ابراهیمی، افشین؛ کبیر، احسان‌اله (۱۳۸۴ب، ۴-۶ بهمن). طراحی یک دیکشنری تصویری برای زیر-کلمات چاپی با درنظرگرفتن نقاط آنها. مقاله ارائه‌شده در *یازدهمین کنفرانس بین‌المللی سالانه انجمن کامپیوتر ایران*. تهران. بازیابی ۲۶ دی ۱۳۹۸، از https://www.civilica.com/Paper-ACCSI11-ACCSI11_137.html

ابراهیمی، افشین؛ کبیر، احسان‌اله (۱۳۸۵). خوشه‌بندی تصاویر زیر- کلمات چاپی فارسی با استفاده از ویژگی‌های مکان مشخصه و الگوریتم K- میانگین. *دانشکده فنی دانشگاه تبریز*، ۳۳ (۱)، ۱-۱۱.

بهمنی، زهرا (۱۳۹۰). *بازیابی براساس محتوای اسناد چاپی فارسی*. پایان‌نامه کارشناسی ارشد، دانشگاه الزهرا، تهران.

خسروی، حسین؛ کبیر، احسان‌اله (۱۳۸۶، ۶–۸ آذر). بازشناسی متن چاپی فارسی برمبنای جداسازی هوشمند. مقاله ارائه‌شده در *سومین کنفرانس بین‌المللی فناوری اطلاعات و دانش، مشهد*. بازیابی ۲۶ دی ۱۳۹۸، از https://www.civilica.com/Paper-ICIKT03-ICIKT03_037.html

خسروی، حسین؛ کبیر، احسان‌اله (۱۳۸۸). ارزیابی روش‌های بازشناسی متون فارسی برمبنای شکل کلی زیرکلمات. *نشریه مهندسی برق و مهندسی کامپیوتر ایران، ۷* (۴)، ۲۶۷–۲۸۰.

داودی، هما؛ کبیر، احسان‌اله (۱۳۹۳). استفاده از مناطق شاخص زیر– کلمات چاپی فارسی برای کاهش فضای جستجو در بازشناسی آنها. *مهندسی برق و مهندسی کامپیوتر ایران*، ۱۲ (۱)، ۱–۱۱.

رفیعی کراچی، شعبانعلی (۱۳۷۳). شکستن *کلمات تایپ‌شده به حروف در رسم‌الخط‌های مختلف*. پایان‌نامه کارشناسی ارشد، دانشگاه تربیت مدرس، تهران.

سرابی نوبخت، سعید (۱۳۹۲). *بازشناسی مستقل از اندازه متون چاپی فارسی با استفاده از توصیفگرهای مستقل از مقیاس و روش‌های انتخاب ویژگی*. پایان‌نامه کارشناسی ارشد، دانشگاه خوارزمی، تهران.

شمسی، محبوبه؛ رسولی کناری، عبدالرضا؛ و شادروان، سوده (۱۳۸۸). روشی نو در تشخیص حروف در متون چاپی عربی و فارسی با استفاده از پویش خط زمینه. *مهندسی برق مجلسی*، ۳ (۳)، ۵۱–۵۸.

عزمی، رضا (۱۳۷۸). *بازشناسی متون چاپی فارسی*. پایان‌نامه دکتری، دانشگاه تربیت مدرس، تهران.

عزمی، رضا؛ کبیر، احسان‌اله (۱۳۷۸). معرفی روش جدیدی برای جداسازی حروف در متون چاپی بدون توجه به نوع قلم. *استقلال، ۱۸* (۲)، ۱–۱۰.

عزمی، رضا؛ کبیر، احسان‌اله (۱۳۸۳). طراحی سه دیکشنری تصویری برای بازشناسی زیرکلمات چاپی. *امیرکبیر*، ۱۵ (آ۵۹–)، ۲۹ –۴۳.

مرتضوی طباطبائی، زهراسادات (۱۳۹۱). *بازشناسی متون فارسی مبتنی بر کدگذاری شکل و اطلاعات معنایی زمینه*. پایان‌نامه کارشناسی ارشد، دانشگاه الزهرا، تهران.

نامور، بی‌تا (۱۳۹۵). *بهبود بازشناسی متون فارسی با استفاده از اطلاعات در سطح زیرکلمه و کلمه*. پایان‌نامه کارشناسی ارشد، دانشگاه الزهرا، تهران.

نظام‌آبادی، حسین؛ کبیر، احسان‌اله (۱۳۷۹، ۱۷ اسفند). جداسازی نقاط چسبیده به بدنه حروف

چاپی. مقاله ارائه‌شده در *اولین کنفرانس ماشین بینایی و پردازش تصویر ایران، بیرجند*. بازیابی ۲۸ دی ۱۳۹۸، از https://www.civilica.com/Paper-ICMVIP01-ICMVIP01_037.html

نظام‌آبادی‌پور، حسین؛ کبیر، احسان‌اله (۱۳۸۳). الگوریتم اصلاح شده جداسازی حروف در متون چاپی با برچسب‌زدن به کانتور بالایی کلمات. *استقلال*، ۲۳ (۱)،۴۸ –۳۳.

نامور، بی‌تا؛ عزمی، رضا (۱۳۹۶، ۳۰–۳۱ فروردین). بهبود بازشناسی متن فارسی با استفاده از اطلاعات در سطح کلمات. مقاله ارائه‌شده در *سومین کنفرانس بین‌المللی بازشناسی الگو و تحلیل تصویر ایران. شهرکرد*. بازیابی ۲۹ دی ۱۳۹۸، از https://www.civilica.com/Paper-IPRIA03-IPRIA03_058.html

کارگروه خط و زبان فارسی در محیط رایانه‌ای (۱۳۸۷). *پژوهشنامه نویسه‌خوان نوری (OCR) فارسی*. تهران: شورای عالی اطلاع‌رسانی.

Azmi, R., & Kabir, E. (2001). A new segmentation technique for omnifont Farsi text. *Pattern Recognition Letters, 22* (2), 97-104.

Bahmani, Z., & Azmi, R. (2011a). Farsi/Arabic Document Image Retrieval Through Sub Letter Shape Coding. In *International Conference on Networks and Information ICNI Chengdu*. New York: ASME.

Bahmani, Z., & Azmi, R. (2011b). Farsi/Arabic document image retrieval through sub -letter shape coding for mixed Farsi/Arabic and English text. *International Journal of Computer Science Issues, 8* (5), 166-172.

Nasrollahi, S., & Ebrahimi, A. (2013). Printed Persian subword recognition using wavelet packet descriptors. *Hindawi Publishing Corporation Journal of Engineering*, 1-12. Retrieved Januery 19, 2020, from https://pdfs.semanticscholar.org/19dc/c02e406669291fbc9406fb862d5732e71a90.pdf

Pourasad, Y., Hassibi, H., & Ghorbani, A. (2012). A Farsi/Arabic word spotting approach for printed document images. *Internatıonal Journal of Natural and Engineering Sciences, 6* (1), 15-18.

Pourasad, Y., Hassibi, H., & Ghorbani, A. (2013). A word spotting method for Farsi machine-printed document images. *Turkish Journal of Electrical Engineering & Computer Sciences, 21* (3), 734-746.

Boukharouba, A. (2017). A new algorithm for skew correction and baseline detection

based on the randomized Hough Transform. *Journal of King Saud University - Computer and Information Sciences, 29* (1),29-38.

Reynaldo Phangtriastu, M., Harefa, J., & Felita Tanoto, D. (2017). Comparison between neural network and support vector machine in optical character recognition. *Procedia Computer Science, 116,* 351-357.

استناد به این مقاله:

بهمنی، زهرا (۱۳۹۸). طراحی و پیاده‌سازی یک سیستم بازیابی اسناد چاپی فارسی. *مطالعات ملی کتابداری و سازماندهی اطلاعات*، ۳۰ (۴)، ۴۶-۶۵.