JIPET

Journal of Intelligent Procedures in Electrical Technology
Vol. 13/ No. 49/ Spring 2022
P-ISSN: 2322-3871, E-ISSN: 2345-5594, http://jipet.iaun.ac.ir/

https://dorl.net/dor/20.1001.1.23223871.1401.13.49.5.4
Research Article

Printer Forensics Based on Identity Vectors of Image Texture Segmentation

Roozbeh Hamzehyan[1], *PhD Student*, Farbod Razzazi[1], *Associate Professor*, Alireza Behrad[2], *Professor*

[1]Department of Electrical and Computer Engineering- Science and Research Branch, Islamic Azad University, Tehran, Iran
[2]Electrical Engineering Department- Shahed University, Tehran, Iran
r.hamzehyan@srbiau.ac.ir, razzazi@srbiau.ac.ir, behrad@shahed.ac.ir

Abstract
Advances in the digital world are leading us to the development of digital forensic tools. The use of machine learning methods for source printer identification is one of the sub-fields of this area that is being developed. In this paper, a new method for extracting secondary features based on identity vector or i-vector to identify the print source is presented. In the proposed method, the classification process is accelerated only by extracting a low-dimension i-vector vector per page, without the use of optical character recognition (OCR) method, and by eliminating majority voting. Furthermore, the proposed method in extracting features is independent of the type and size of the font and the language of the text. Secondary features are obtained by splitting the document image into smaller patches and modeling the primary LBP features of the dark, border, and light areas in separate spaces. Modeling the primary features of different regions in separate total variability printer space makes it possible to extract class discriminator information from the remaining print texture in the bright area to increase classification accuracy. In this paper, the effect of using the texture of different regions and changing the patch dimensions using the SVM (Support Vector Machine) classifier through simulation has been carefully investigated. The simulation results show that only by refining the basic features of LBP we achieved 99.05% accuracy, which is more than the latest research in this field.

Keywords: printer forensics, i-vector, inter-class variation, printer source identification, total variability printer space

Received: 30 April 2021
Revised: 26 May 2021
Accepted: 28 July 2021

Corresponding Author: Dr. Farbod Razzazi

Citation: R. Hamzehyan, F. Razzazi, A. Behrad, "Printer forensics based on identity vectors of image texture segmentation", Journal of Intelligent Procedures in Electrical Technology, vol. 13, no. 49, pp. 67-82, June 2022 (in Persian).

https://dorl.net/dor/20.1001.1.23223871.1401.13.49.5.4

مقاله پژوهشی

بازجویی قانونی چاپگر مبتنی بر بردار هویت حاصل از ناحیه‌بندی بافت تصویر

روزبه حمزه‌ئیان[۱]، دانشجوی دکتری، فربد رزازی[۱]، دانشیار، علیرضا بهراد[۲]، استاد

۱- دانشکده مهندسی برق و کامپیوتر- واحد علوم و تحقیقات، دانشگاه آزاد اسلامی، تهران، ایران
۲- دانشکده فنی و مهندسی- دانشگاه شاهد، تهران، ایران
r.hamzehyan@srbiau.ac.ir, razzazi@srbiau.ac.ir, behrad@shahed.ac.ir

چکیده: پیشرفت در دنیای دیجیتال، ما را به سمت توسعه ابزار بازجویی قانونی دیجیتال سوق می‌دهد. استفاده از روش‌های یادگیری ماشین برای شناسایی منبع چاپ یکی از زیر مجموعه‌های این حوزه بوده که درحال توسعه است. در این مقاله، روش جدیدی برای استخراج ویژگی‌های ثانویه بر پایه بردار هویت (i-vector) برای شناسایی منبع چاپ ارائه شده است. در روش پیشنهادی تنها با استخراج یک بردار i-vector با بعد کم به‌ازای هر صفحه بدون استفاده از روش بازشناسی نوری نویسه‌ها (OCR) و با حذف رأی‌گیری اکثریت فرایند طبقه‌بندی تسریع شده است. به‌این ترتیب روش پیشنهادی در استخراج ویژگی‌ها مستقل از نوع و اندازه قلم نویسه‌ها و زبان متن است. ویژگی‌های ثانویه با افراز تصویر سند به تکه‌های کوچک‌تر و مدل‌سازی ویژگی‌های اولیه الگوی دودویی محلی (LBP) مربوط به ناحیه‌های تیره، مرز و روشن در فضاهای مجزا به‌دست می‌آید. مدل-سازی ویژگی‌های اولیه نواحی مختلف در فضاهای مجزای متغیر کل چاپگر، امکان استخراج اطلاعات جداکننده کلاس‌ها از بافت چاپ باقیمانده در ناحیه روشن را برای افزایش دقت و صحت طبقه‌بندی مهیا می‌کند. در این مقاله تأثیر استفاده از بافت نواحی مختلف و تغییر ابعاد تکه‌بندی با استفاده از طبقه‌بند ماشین بردار پشتیبان (SVM) از طریق شبیه‌سازی به‌دقت بررسی شده است. نتایج شبیه‌سازی، نشان می‌دهد که تنها با پالایش ویژگی‌های اولیه LBP به صحت ۹۹/۰۵ درصد دست یافته‌ایم که بیشتر از آخرین پژوهش‌های این حوزه است.

کلمات کلیدی: بازجویی قانونی چاپگر، بردار هویت، تغییرات درون کلاسی، شناسایی منبع چاپ، فضای متغیر کل چاپگر

تاریخ ارسال مقاله: ۱۴۰۰/۲/۱۰
تاریخ بازنگری مقاله: ۱۴۰۰/۳/۵
تاریخ پذیرش مقاله: ۱۴۰۰/۵/۶

نام نویسنده‌ی مسئول: دکتر فربد رزازی
نشانی نویسنده‌ی مسئول: تهران- دانشگاه آزاد اسلامی واحد واحد علوم و تحقیقات- گروه مهندسی برق مخابرت

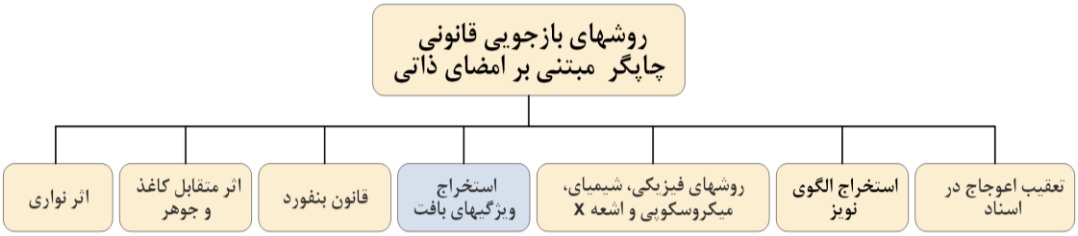
۱- مقدمه

امروزه در سیستم‌های اداری کاغذ ابزاری مستقیم در چاپ اسناد هویت، مالکیت، مالیاتی و تراکنش‌های مالی است. منبع بسیاری از این اسناد کاغذی، چاپگرهای لیزری یا نمونه‌های هم‌خانواده از این نوع چاپگرها است. دسترسی عمومی به چاپگرها و همچنین پیشرفت نرم‌افزارهای ویرایش تصویر، امکان ویرایش و بازتولید اسناد چاپی دیجیتال مشابه را به فرایندی در دسترس همگان تبدیل کرده‌است و باعث شده حوزه‌ی بالقوه‌ای برای وقوع جرم باشد. استفاده از متخصصان بازجویی قانونی در تشخیص یا تفکیک اسناد جعلی می‌تواند راهگشا باشد، ولی علاوه بر فرایند آموزش متخصصان، روش‌های مورد استفاده آن‌ها نیازمند تجهیزات و آزمایشگاه‌های خاص و پرهزینه است و یا بر پایه آزمایش‌هایی است که تخریب بخشی از سند را در پی دارد. گسترش روش‌های ماشینی بر پایه بینایی ماشین و پردازش تصویر که بتوانند به‌صورت غیر مخرب، سریع، ارزان و بدون نیاز به حضور متخصصان بازجویی قانونی اصالت اسناد را تعیین کنند، به‌عنوان نیاز احساس می‌شود. بازجویی قانونی تصویر[1] یا شناسایی منبع تصویر دیجیتال به معنی تعیین دستگاه منبع تولیدکننده تصویر با تکیه بر راه‌حل‌های ماشینی در سال‌های اخیر مورد توجه بوده و پیشرفت‌های زیادی داشته است [۱،۲]. حوزه مورد علاقه ما در این مقاله بازجویی قانونی چاپگر[2] یا شناسایی منبع چاپ در اسناد دیجیتال است که اگرچه به‌نوعی زیرمجموعه بازجویی قانونی تصویر است، ولی از نظر منبع تولید، نحوه تبدیل به تصویر دیجیتال، محتوا و بافت کاملاً متفاوت است. مسئله شناسایی منبع چاپ در اسناد پیچیده‌تر است، به این دلیل که در فرآیند چاپ، عناصر مکانیکی بیشتری درگیر و در هم تنیده‌اند. چنین عناصری در ایجاد یک امضای منحصر به فرد برای هر چاپگر نقش ایفا می‌کند. در حال حاضر راه‌حل مسئله شناسایی منبع چاپ به دو روش امضای خارجی و امضای ذاتی تقسیم می‌شود [۳]. در روش‌های امضای خارجی که در این مقاله مورد بحث نیست، چاپگر در هنگام چاپ، یک امضای قابل ردیابی را به سند اضافه می‌کند [۴]. برای اجرای مؤثر این روش غالباً بخش‌هایی به سند قبل از فرایند چاپ اضافه می‌شود یا نیاز به تغییراتی در سازوکار سخت‌افزاری چاپگر است. همه چاپگرها در زمان چاپ ویژگی‌هایی را به دلیل نقص ذاتی یا تفاوت‌های جزئی در مکانیسم عملکرد، به‌صورت یک اثر منحصر به فرد به سند چاپ شده منتقل می‌کنند. این اثر امضای ذاتی چاپگر شناخته می‌شود که معیار مناسبی برای شناسایی منبع چاپ یک سند است. روش‌های متعددی برای شناسایی منبع چاپ بر پایه به‌کارگیری امضای ذاتی به وجود آمده است. دسته‌بندی کلی این روش‌ها در شکل (۱) نمایش‌داده شده است. این روش‌ها شامل اثر نواری [۵]، شناسایی اثر متقابل کاغذ و جوهر [۶]، به‌کارگیری قانون بنفورد [۷]، استخراج ویژگی‌های بافت [۸-۱۳]، بررسی اعوجاج در اسناد [۱۴]، استخراج الگوی نویز [۱۵] و روش‌های فیزیکی، شیمیایی و میکروسکوپی است. در بین این روش‌ها، به‌کارگیری ویژگی‌های بافت چاپ نسبت به بقیه نتایج دقیق‌تری داشته و در سال‌های اخیر بیشتر موردعلاقه بوده است.

۱-۱- چاپگر لیزری

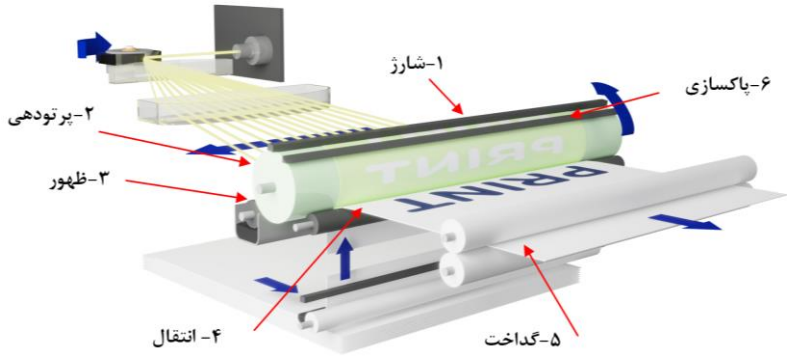
چاپگرهای لیزری به دلیل چاپ سریع و اقتصادی بودن امروزه بیشترین کاربرد را دارند و بیشتر اسناد موجود، توسط این نوع چاپگرها تولید می‌شوند. در تمامی چاپگرهای لیزری فرایند چاپ شامل دو مرحله پردازش و چاپ است. مرحله پردازش توسط پردازشگر چاپ[3] (RIP) و پس از ارسال دستور چاپ انجام می‌شود. پردازشگر چاپ، بخشی از نرم‌افزار چاپگر در رایانه است که داده‌ها را به زبان چاپگر ترجمه می‌کند. در این مرحله فونت‌ها و بردارها به ساختار یک ماتریس دودویی در ابعاد چاپ و تصاویر به ساختار ترام تبدیل می‌شوند [۱۶]. الگوریتم این تبدیل برای شرکت‌ها و انواع چاپگرها متفاوت است. در مرحله چاپ، داده-های چاپگر، طی یک فرایند شش مرحله‌ای مطابق شکل (۲) شامل شارژ، پرتودهی، ظهور، انتقال، گداخت و پاک‌سازی بر روی کاغذ چاپ می‌شوند. در انجام این فرایند اجزای مختلفی دخالت دارند و وجود اشکال و یا حتی تفاوت کوچک این اجزا در چاپگرهای مشابه، باعث به وجود آمدن عیوب منحصر به فرد در سطح تصویر چاپ می‌شود. مجموع این اثرات را به‌عنوان امضای ذاتی چاپگر می‌شناسیم. بافت چاپ را در چاپگرهای لیزری می‌توان به سه ناحیه تقسیم نمود. ناحیه ۱، بافت تیره چاپ، شامل بافت داخل نویسه‌ها و بخش‌هایی با تیرگی یکنواخت است و با توجه به ساختار مسطح و یکنواخت آن بیشترین وابستگی را به چاپگر دارد. ناحیه ۲، بافت مرزی، مرز ناحیه تیره و روشن را شامل می‌شود در این ناحیه ساختار بافت همانند ناحیه ۱

یکنواخت نیست و بسته به نوع کاغذ و چاپگر ساختار متفاوتی دارد. ناحیه ۳، شامل بافت باقیمانده از چاپ در ناحیه سفیدرنگ سند است. بافت این ناحیه کاملاً به نوع خاص چاپگر و کاغذ وابسته است و به دلیل اشکال در تیغه‌های پاک‌کننده یا اشکال در مراحل پرتودهی و شارژ به وجود می‌آید. در این ناحیه بافت چاپ غالباً به‌صورت تیره‌گی یکنواخت به کل فضای سفید روی سند چاپ شده منتقل می‌شود. با افزایش عمر چاپگر ممکن است به‌صورت خطوط تیره و روشن عمودی و یا چاپ شبح در صفحه دیده شود. در ادبیات پیشین کمتر به این ناحیه توجه شده چون تراکم کم آن ممکن است باعث شود ویژگی‌های استخراج‌شده به بافت کاغذ وابسته شود. در ادامه این بافت را بافت سایه/بافت روشن چاپ می‌نامیم و راهکاری برای استفاده از ویژگی‌های این بافت ارائه خواهیم داد. ناحیه‌های مختلف بافت چاپ در ادامه در بخش ۱-۴، شکل (۴) نمایش داده شده است.



شکل (۱): دسته‌بندی روش‌های شناسایی منبع چاپ بر پایه استخراج امضای ذاتی

Figure (1): Classification of printing source identification methods based on intrinsic signature



شکل (۲): نمایش مراحل مختلف چاپ در چاپگرهای لیزری

Figure (2): Printing steps in laser printers

۱-۲- رویکرد و ساختار مقاله

ما در این مقاله یک روش سریع و ساده و مقاوم در برابر تغییرات درون کلاسی را برای شناسایی منبع چاپ ارائه خواهیم داد. روش پیشنهادی بر اساس ناحیه‌بندی بافت چاپ موجود در تصویر سند عمل کرده و به‌ازای ویژگی‌های اولیه ناحیه‌های مشابه از تمامی تکه‌های یک سند یک بردار ویژگی ثانویه مبتنی بر بردار هویت[۴] (i-vector) استخراج می‌کند. در روش پیشنهادی از تکه‌بندی تصویر به‌عنوان جایگزین روش شناسایی نوری نویسه‌ها[۵] (OCR) استفاده شده است. در نتیجه روش پیشنهادی در مواجهه با اسناد دارای متن با زبان‌های متفاوت و با قلم‌های متنوع از نظر نوع و اندازه و حتی وجود تصویر در اسناد مقاوم است. در این مقاله تأثیر ویژگی‌های استخراج شده از بافت ناحیه‌های مختلف و تکه‌بندی با ابعاد متفاوت در کارایی الگوریتم شناسایی چاپگر منبع سند بررسی می‌شود. همچنین نشان می‌دهیم تکیه بر مدل‌سازی ویژگی‌های اولیه هر صفحه در فضای متغیر کل چاپگر مبتنی بر مدل مخلوط گوسی–مدل پس‌زمینه جهانی[۶] (GMM-UBM) و استخراج یک بردار ویژگی ثانویه i-vector به‌ازای هر صفحه، فرایند رأی‌گیری اکثریت را حذف و فرایند آموزش و ارزیابی طبقه‌بند را سرعت می‌بخشد. بخش‌های مقاله در ادامه شامل قسمت‌های زیر است. در بخش دوم روش‌های مرز دانش در شناسایی منبع چاپ را معرفی خواهیم کرد. روش مواجهه با تغییرات درون کلاسی با تکیه بر مدل‌سازی ویژگی‌های اولیه در فضای متغیر کل چاپگر را در بخش سوم مطالعه

می‌کنیم. در بخش چهارم الگوریتم پیشنهادی را بیان کرده و نحوه اجرای مراحل مختلف را شرح خواهیم داد. در بخش پنجم از طریق شبیه‌سازی پارامترهای تأثیرگذار بر روی الگوریتم پیشنهادی را در شرایط متفاوت ارزیابی کرده و نتایج با روش‌های مرز دانش مقایسه می‌شود. در پایان نتیجه‌گیری آمده است.

۲- مروری بر تحقیقات گذشته

در پژوهش‌های اخیر، روش‌های بر پایه استخراج بافت چاپ، به دلیل ارائه نتایج دقیق‌تر بیشتر مورد توجه بوده‌اند. روند کلیدر اغلب پژوهش‌ها تقریباً یکسان است و شامل استخراج ویژگی، انتخاب یا کاهش ویژگی و طبقه‌بندی می‌شود. پژوهش‌های مختلف با ایجاد تنوع در استخراج ویژگی سعی کرده‌اند اطلاعات بیشتری از بافت چاپ استخراج کنند و نتایج را بهبود دهند. این کار با افزایش بعد بردار ویژگی باعث زمان‌بر شدن پردازش‌ها در طبقه‌بند می‌شود. نوع بافت استخراج شده، ناحیه استخراج ویژگی، روش کاهش ویژگی و نوع طبقه‌بند مواردی است که پژوهش‌های مختلف را متمایز نموده است. در پژوهش [۱۳] برای شناسایی منبع چاپ اسناد تصویری و متنی با زبان‌های مختلف، توسط ۱۰ روش استخراج ویژگی متفاوت از نویسه‌های خاص در متن، یک مجموعه ۳۰۶ بعدی ویژگی استخراج شده است. با استفاده از مدل تصمیم‌گیری نظری ویژگی‌های برتر انتخاب و توسط دو روش ماشین بردار پشتیبان[۷] (SVM) و شبکه عصبی عمیق ۷ و ۹ لایه نتایج ارزیابی شده است. در پژوهش [۱۰] ویژگی‌های جداکننده چاپ بدون هیچ گونه پیش فرضی در مورد انتخاب ویژگی، مستقیماً از تصاویر نویسه‌های پرکاربرد متن توسط شبکه عصبی کانولوشنی موازی استخراج می‌شود. ویژگی‌های خروجی شبکه عصبی عمیق هر نویسه توسط SVM طبقه‌بندی شده و در نهایت رأی‌گیری اکثریت نتیجه انتساب چاپگر به ازای هر صفحه را مشخص می‌کند. اگرچه روش ارائه شده در مورد سایز نویسه‌ها مقاومت مناسبی دارد اما نیاز به فرایندی شبیه به OCR دارد که محدودیت‌هایی را ایجاد می‌کند. در مرجع [۱۱]، ناوارو و همکاران ویژگی‌های مهم انتساب چاپگر را بر روی سند به‌صورت یک نقشه دیداری باقابلیت تفسیر انسانی برجسته می‌کنند. این نقشه به متخصصان بازجویی قانونی چاپ، شواهد بیشتری درباره روند شناسایی منبع چاپ ارائه می‌کند. برای این منظور الگوریتمی به نام CTGF-map[۸] ارائه شده‌است. در این پژوهش پس از استخراج نویسه‌های درون تصویر سند، آنها را به شش ناحیه مختلف تقسیم کرده و تنها از سه ناحیه، ویژگی‌های CTGF–map استخراج می‌شود. برای انتخاب ویژگی و طبقه‌بندی از روش جنگل تصادفی استفاده شده‌است. در مرجع [۹] جوشی و همکاران برای حذف محدودیت-های ناشی از به‌کارگیری OCR، از آنالیز اجزای متصل برای استخراج نویسه‌ها استفاده کردند. از هر جزء پیوسته ویژگی‌های LTrP[۹] بر پایه الگوی باینری محلی از ناحیه مسطح و لبه‌ها، به‌صورت مجزا استخراج می‌شود تا برای طبقه‌بندی توسط طبقه‌بند SVM مورد استفاده قرار بگیرد. در مرحله طبقه‌بندی هر جزء متصل به‌صورت مجزا طبقه‌بندی شده و برای شناسایی منبع چاپ از رأی‌گیری اکثریت در هر صفحه استفاده می‌شود. صرف‌نظر از روش کار دقیق در این پژوهش‌ها، در مرحله آموزش طبقه‌بند مدل هر چاپگر یا مرز جداکننده کلاس‌ها فقط با توجه به ویژگی‌های استخراج شده از سندهای آموزشی همان چاپگر ساخته می‌شود، بنابراین تنها آن اسناد را می‌توانند مورد بررسی قرار دهند که شامل همان حرف یا گروهی از حروف است که برای تولید مدل بکار گرفته شده است. در مرحله استخراج ویژگی بیشتر روش‌هایی که بافت چاپ درون نویسه‌ها را معیار قرار می‌دهند از تکنیک‌هایی شبیه OCR برای استخراج این نویسه‌ها بهره می‌برند. این موضوع باعث ایجاد محدودیت در شناسایی منبع چاپ اسناد دارای قلم با نوع و اندازه مختلف و یا زبان‌های متفاوت می‌شود، این عوامل باعث تنوع درون کلاسی می‌شوند. استفاده از آنالیز اجزای متصل در مرجع [۹] کمی این پیچیدگی را کاهش داده ولی در نهایت برای تفکیک اطلاعات محتوایی و بافتی نیاز به بردارهای ویژگی با بعد بالا است که فرایند استخراج ویژگی و طبقه‌بندی را زمان‌بر می‌کند. نکته دیگر در این پژوهش‌ها در مورد مواجهه با اسناد شامل تصاویر چاپ شده با تکنیک ترام است. در این تحقیقات این بخش از اسناد به‌صورت ناگفته حذف می‌شود و یا فقط بخش‌های تیره‌تر مورد استفاده قرار می‌گیرد، درصورتی‌که می‌تواند شامل اطلاعات مفید بافت باشد. در تمامی پژوهش‌های گذشته طبقه‌بندی بر روی ویژگی‌هایی از نویسه‌ها یا بخش‌های کوچک‌تری که از هر صفحه استخراج شده است انجام می‌شود که باعث افزایش زمان پردازش در فرایند آموزش و ارزیابی طبقه بند می‌گردد. بعلاوه در این

روش‌ها برای رسیدن به تصمیم نهایی، بر روی تمام نتایج طبقه‌بندی در هر صفحه رأی‌گیری اکثریت به‌عنوان پس پردازش باید لحاظ شود.

۳- استخراج بردار هویت i-vector و مدل‌سازی ویژگی‌ها در فضای متغیر کل چاپگر

یکی از چالش‌های اصلی در شناسایی منبع چاپ، وجود تشابه در بافت چاپ چاپگرهای مختلف است. در نتیجه ممکن است نمونه‌هایی از کلاس یکسان از دید طبقه‌بند به‌اشتباه به‌عنوان کلاس متفاوت انتخاب شود. به‌صورت کلی، هرگونه تنوع در یک کلاس که باعث شود نمونه‌های آن به‌عنوان کلاس‌های مختلف طبقه‌بندی شوند، به‌عنوان تغییرات درون کلاسی شناخته می‌شود. این تغییرات یکی از دلایل کاهش عملکرد در این رویکرد است. مواردی مانند تغییر در نوع و اندازه قلم، زبان نوشتاری، سطح تونر، میزان روشنایی، فشرده‌سازی تصویر، سیستم تصویربرداری و تغییرات بافت کاغذ می‌تواند به‌عنوان عوامل تشدیدکننده تغییرات درون کلاسی در نظر گرفته شود. در این حوزه، وابستگی بافت چاپ به بافت کاغذ باعث می‌شود بافت استخراج شده از کلاس‌های متفاوت به یکدیگر نزدیک شود و باعث کاهش فاصله کلاس‌های متفاوت گردد. این موضوع در ناحیه مرز و مخصوصاً بافت سایه چاپ به دلیل تنک بودن بافت چاپ بیشتر اهمیت پیدا می‌کند و به همین دلیل در بسیاری از پژوهش‌ها ویژگی‌های این ناحیه‌ها برای استفاده در فرایند شناسایی چاپگر استفاده نمی‌شود [۹،۱۱،۱۲]. تحلیل عوامل مشترک[۱۰] (JFA) [۱۷،۱۸]، ابزاری قدرتمند در رویکرد مدل سازی تغییرات درون کلاسی است که با موفقیت در دو حوزه تشخیص گوینده و شناسایی چهره، بر پایه مدل‌سازی ویژگی‌ها در فضای GMM-UBM، استفاده شده است. در این مدل‌سازی، UBM به‌صورت یک مدل گوسی مخلوط از نمونه‌های تمامی کلاس‌ها، به‌عنوان مدل پس‌زمینه ساخته می‌شود. این مدل ویژگی‌ها را مستقل از منبع خاص به‌صورت نقاط کلیدی تجمع داده و یا ساختار عمومی توزیع آماری ویژگی‌ها بیان می‌کند. مدل هر کلاس با وفق‌دهی پارامترهای UBM با داده‌های آموزشی همان کلاس، به‌روز می‌شود. بنابراین ویژگی‌هایی که توسط نمونه‌های آموزشی قابل مدل‌سازی نیستند از طریق UBM کپی می‌گردد. مدل JFA در شناسایی منبع چاپ را می‌توان به‌صورت یک توزیع گوسی روی ابربردارهای وابسته به چاپگر و وابسته به تغییرات درون کلاسی تصویر کرد. در این مدل‌سازی، فرض بر این است که می‌توان با تعداد کمی از متغیر پنهان مستقل از هم با نام عوامل چاپگر و عوامل کانال (بیان کننده تغییرات درون کلاسی)، واریانس توزیع ابربردارهای وابسته به چاپگر و وابسته به تغییرات درون کلاسی را بیان کرد. اما باتوجه‌به نتایجی که دهک در مرجع [۱۹] نمایش داد، عوامل کانال حاوی اطلاعاتی درباره مدل چاپگر هستند و حذف آن‌ها باعث از بین رفتن بخشی از اطلاعات متمایز کننده بین کلاس‌ها خواهد شد. بر این اساس، یک سیستم شناسایی چاپگر منبع بر مبنای استفاده از آنالیز عاملی به‌عنوان استخراج ویژگی با ساختار جدید مبتنی بر مرجع [۲۰] ارائه شده که شامل فضای جدید با بعد کم است و حاوی اطلاعات چاپگر و تغییرات درون کلاسی است. این فضا به نام متغیر کل چاپگر شناخته و با ماتریس T تعریف می‌شود، که حاوی بردارهای ویژه متناسب با بزرگ‌ترین مقادیر ویژه ماتریس کوواریانس متغیر کل چاپگر است و به فرم رابطه (۱) تعریف می‌شود:

$$M = m + Tw \qquad (۱)$$

برای تصویر سند، ابربردار GMM وابسته به چاپگر و تغییرات درون کلاسی با M نمایش داده می‌شود. M دارای توزیع نرمال با میانگین m و کوواریانس 'TT است. در این رابطه m ابربردار میانگین UBM و مستقل از چاپگر و تغییرات درون کلاسی است و T یک ماتریس مستطیلی با رتبه پایین است که فضای متغیر کل چاپگر را بیان می‌کند و w عوامل این فضا و بردار تصادفی با توزیع نرمال N(0,1) است. مؤلفه w عوامل کل چاپگر است و بردار هویت یا i-vector نام‌گذاری می‌شود. مدل‌سازی i-vector یک روش مدل‌سازی تغییرات درون و بین‌کلاسی در یک فضای مشترک با ابعاد پایین است. این روش اجازه می‌دهد ویژگی‌های اولیه بافت چاپ به فضای جدید متغیر کل چاپگر با بعد کمتر نگاشته شود. در واقع i-vector یک متغیر پنهان است که با توزیع خلفی مشروط بر آمار بام-ولش برای تصویر سند تعریف می‌شود. در برآورد i-vector ها، ابتدا UBM به‌عنوان یک ساختار GMM با استفاده از داده‌های غیر هدف محاسبه می‌شود. اگر فرض کنیم تصویر سند u به L تکه تقسیم شود، به‌ازای هر تصویر دنباله‌ای به طول L از ویژگی‌های بافت تکه‌ها به فرم $\{y_1,y_2,\ldots y_L\}$ استخراج می‌شود تا برای به‌دست آوردن آمار مرتبه صفر و

یک بام-ولش به ترتیب با رابطه‌های (۲) و (۳) استفاده شود. تعداد تکه‌ها یکی از پارامترهایی است که در مدل‌سازی ویژگی‌های ثانویه مؤثر است.

$$N_c = \sum_{t=1}^{L} P(c \mid y_t, \Omega) \quad (۲)$$

$$F_c = \sum_{t=1}^{L} P(c|y_t,\Omega)\, y_t \quad (۳)$$

در این روابط Ω از مخلوط C جزء گوسی به فرم GMM تشکیل شده (c=1,2,…,C اندیس گوسی‌ها است) و UBM را می‌سازد. توسط $P(c|y_t,\Omega)$ احتمال خلفی جزء c مخلوط تولید شده توسط بردار y_t بیان می‌شود. آمار متمرکز مرتبه اول بام-ولش توسط رابطه (۴) به دست می‌آید،

$$\tilde{F}_c = \sum_{t=1}^{L} P(c|y_t,\Omega)(y_t - m_c)\, y_t \quad (۴)$$

دراین رابطه m_c میانگین جزء c از مخلوط گوسی در UBM است. مقدار i-vector برای تصویر سند معین با استفاده از رابطه (۵) محاسبه می‌شود. دراین‌رابطه N(u) ماتریس قطری از آمار مرتبه صفر با ابعاد CF×CF است که اجزای قطری را بلوک‌های N_cI تشکیل داده‌اند.

$$w = \left(I + T^t \Sigma^{-1} N(u) T\right)^{-1} . T^t \Sigma^{-1} \tilde{F}(u) \quad (۵)$$

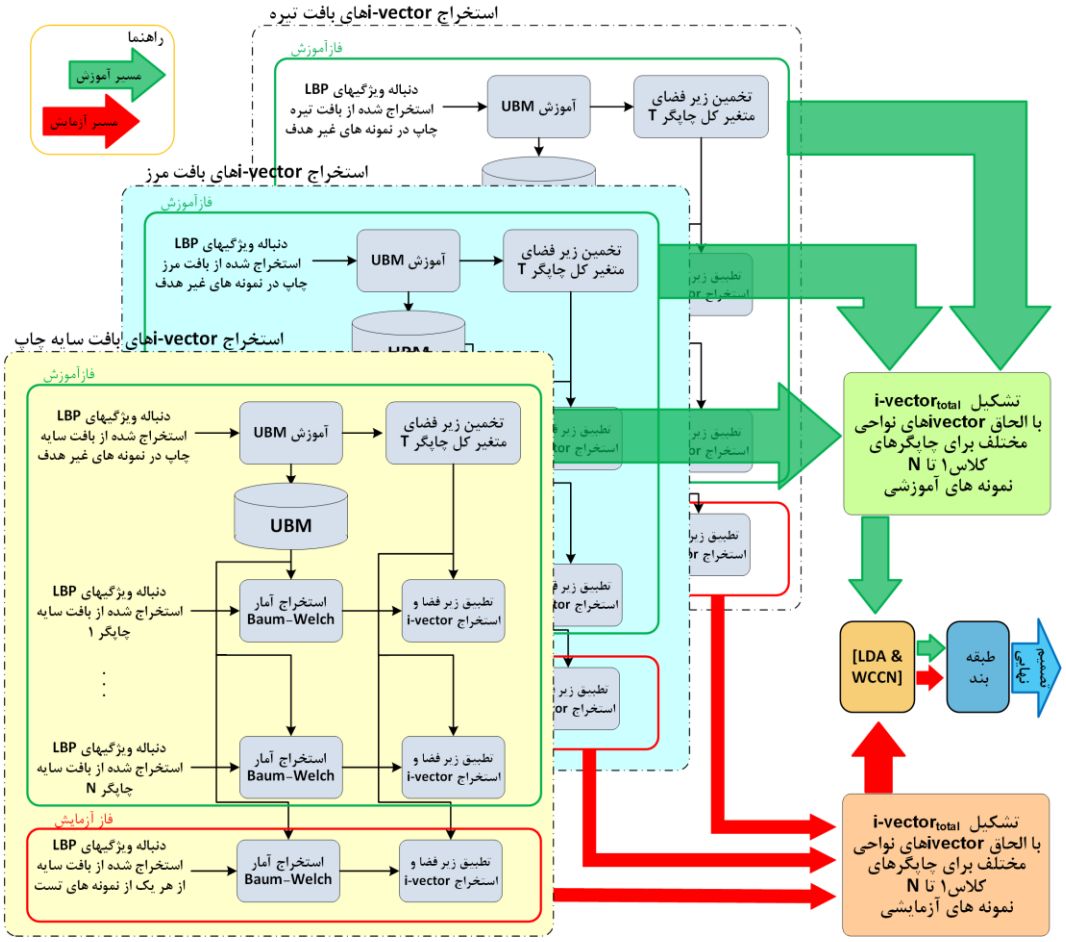
در این معادله $\tilde{F}(u)$ به فرم $[\tilde{F}_1,\ldots,\tilde{F}_c]^T$ با ابعاد CF×1 است. Σ یک ماتریس کوواریانس قطری با ابعاد CF×CF است که در طول آموزش آنالیز عاملی تخمین زده می‌شود و متغیرهای باقیمانده را که توسط ماتریس T در نظر گرفته نشده، مدل می‌کند [۱۹].

بردار i-vector در فضایی استخراج می‌شود که هیچ تمایزی بین تغییرات کانال و چاپگر وجود ندارد و این تغییرات در فضای متغیر کل چاپگر جبران می‌شود. جبران اثرات کانال یا تغییرات درون کلاسی در فضای متغیر کل چاپگر سریع‌تر و آسان‌تر انجام می‌شود، زیرا ابعاد i-vector در مقایسه با بردارهای موجود در GMM کمتر است. برآورد i-vector ها در فرایند آموزش بدون نظارت انجام می‌شود [۲۰] به همین دلیل استفاده از آنالیز افتراقی خطی[۱۱] (LDA) بر روی بردارهای خروجی به منظور افزایش فاصله بین کلاس‌های متفاوت، منجر به افزایش کارایی الگوریتم می‌گردد. بر اساس نتایج در حوزه‌های دیگر انتظار داریم اطلاعات غیر چاپگر مانند تغییرات درون کلاسی، بر میزان اندازه i-vectorها تأثیر گذارد، بنابراین حذف اثر اندازه در فرآیند تصمیم‌گیری باعث افزایش توانایی سیستم در تصمیم‌گیری خواهدشد. اعمال روش‌های نرمال‌سازی آماری مانند نرمال-سازی کوواریانس درون کلاسی[۱۲] (WCCN) در بهبود نتایج مؤثر است [۲۱].

۴- معماری پیشنهادی بازجویی قانونی چاپگر بر اساس i-vector

در این مقاله یک روش جدید بر پایه استفاده از روش‌های GMM-UBM و تحلیل عوامل مشترک، برای شناسایی منبع چاپ سندها ارائه شده است. مرحله ابتدایی روش پیشنهادی با پیش‌پردازش شروع می‌شود و سپس تصویر اسناد در ساختار شبکهٔ منظم به تکه‌های کوچک‌تر غیر همپوشان تقسیم شده و پس از ناحیه بندی، بافت چاپ از نواحی مختلف در تکه‌ها استخراج می‌گردد. ویژگی‌هایی که از ناحیه‌های مختلف در تکه‌های هر سند استخراج شده را به‌عنوان ویژگی‌های اولیه در نظر گرفته و آن‌ها را در فضای متغیر کل چاپگر (به‌صورت مجزا) مدل کرده تا ویژگی ثانویه به شکل i-vector برای طبقه‌بندی بر اساس آن‌ها استخراج شود. در الگوریتم پیشنهادی به‌ازای نواحی مشابه از تکه‌های مختلف تصویر هر سند به‌صورت موازی یک i-vector محاسبه می‌شود. برای رسیدن به کارایی بیشتر در طبقه‌بندی، i-vectorهای استخراج شده از نواحی مختلف به هم الحاق می‌شوند. بررسی سهم i-vectorهای استخراج شده از ناحیه‌های مختلف تصویر سند یکی از مواردی است که در این تحقیق به آن پرداخته شده است. در واقع ویژگی‌های ناحیه مرز و به‌ویژه بافت سایه چاپ به دلیل تنک بودن به بافت کاغذ وابسته هستند. در پایگاه داده‌های استاندارد، از کاغذ یکسان برای چاپ تمام اسناد استفاده شده است، به همین دلیل در پژوهش‌های پیشین برای این‌که این موضوع باعث کاهش فاصله بین کلاس‌ها و انحراف در تصمیم‌گیری طبقه‌بند نشود،

ویژگی‌های برخی از این نواحی، از تصمیم‌گیری‌ها حذف می‌شوند. معماری روش پیشنهادی با پالایش ویژگی‌های اولیه و با ساخت یک مدل پس‌زمینه مجزا به‌ازای ویژگی‌های اولیه استخراج‌شدهٔ هر یک از نواحی، روشی برای تفکیک اطلاعات مشترک ناشی از بافت کاغذ و اطلاعات تنک جداکننده ناشی از بافت چاپ، ارائه می‌دهد. در روندنمای شکل (۳) معماری روش پیشنهادی برای شناسایی منبع چاپ نمایش‌داده‌شده است. در ادامه مراحل مختلف را شرح می‌دهیم.



شکل (۳): نمودار گردشی الگوریتم پیشنهادی

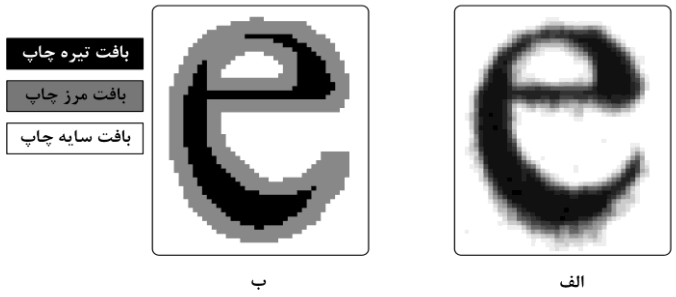
Figure (3): flowchart of the proposed algorithm

۴-۱- پیش‌پردازش و استخراج ویژگی

در اولین مرحله توسط پویشگر نسخه دیجیتالی از اسناد فیزیکی چاپی آماده می‌شود. برای پیشگیری از انحراف در تصمیم-گیری‌ها، در تبدیل تمامی اسناد از پویشگر یکسان استفاده می‌شود. عواملی همچون سایه‌های ایجاد شده در لبه‌های کاغذ ناشی از نور خارجی در زمان پویش و تاخوردگی احتمالی لبه‌ها با برش درصد کمی از پیکسل‌های مرزی در چهار طرف سند حذف می‌شود تا به‌اشتباه به‌عنوان بافت چاپ تفسیر نشود. در روش پیشنهادی ویژگی‌های اولیه به‌صورت مجزا از بافت چاپ نواحی مختلف استخراج می‌شود. برای تفکیک ویژگی‌های این نواحی، یک ماسک باینری در ابعاد تصویر با استفاده از فیلتر میانه، آستانه‌گذاری اوتسو (Otsu) و عملیات ریخت‌شناسی ایجاد می‌کنیم. بافت تیره، بافت سایه چاپ استخراج شده و مرز این دو ناحیه با استفاده از روش کَنی (Canny) و عملیات ریخت‌شناسی به دست می‌آید. این ماسک امکان جداکردن این سه ناحیه را فراهم می‌کند، در شکل (۴) تصویر نویسه e و نمونه ماسک حاصل نمایش‌داده‌شده است. در روش پیشنهادی از ویژگی‌های الگوی باینری محلی LBP به‌عنوان ویژگی‌های اولیه بافت استفاده می‌شود. توسعه زیادی برای الگوریتم LBP ارائه شده است. ما برای نمایش قابلیت روش پیشنهادی از ساختار پایه LBP استفاده کرده‌ایم [۲۲]. به دلیل تغییر ناپذیر بودن LBP با تغییرات

خطی در شدت سطح خاکستری تصویر، ویژگی‌های استخراج شده توسط آن نسبت به روشن شدن سند چاپ شده به دلیل تغییر در سطح پودر تونر و تغییرات حساسیت OPC، مقاوم است. روند استخراج ویژگی‌های اولیه LBP از نواحی سه‌گانه در هر سند در دو مرحله انجام می‌شود. در مرحله اول از تصویر هر سند، تصویر LBP محاسبه می‌شود. در مرحله دوم، تصویر LBP در یک ساختار شبکه‌ای منظم به‌صورت غیر همپوشان افراز شده و با انطباق هر تکه با بخش متناظر در ماسک سه وضعیتی، بافت منطبق با ناحیه تیره، مرز و بافت سایه چاپ تفکیک شده و هیستوگرام آن‌ها به‌عنوان ویژگی استخراج می‌شود. اگر هر یک از n تصویر به L تکه تقسیم شده باشد، هیستوگرام‌های ناحیه تیره، مرز و بافت سایه چاپ از تکه‌های مختلف برای استخراج i-vector ها به ترتیب در دنباله‌های مجزای LBP_D ، LBP_E و LBP_L با طول L به فرم رابطه (۶) جای‌گذاری می‌شود تا برای استخراج i-vector های هر ناحیه بکار رود.

$$LBP_D=\left\{\begin{bmatrix}y_{1,1}\\ \vdots \\ y_{n,1}\end{bmatrix}\cdots\begin{bmatrix}y_{1,L}\\ \vdots \\ y_{n,L}\end{bmatrix}\right\},\ LBP_E=\left\{\begin{bmatrix}y_{1,1}\\ \vdots \\ y_{n,1}\end{bmatrix}\cdots\begin{bmatrix}y_{1,L}\\ \vdots \\ y_{n,L}\end{bmatrix}\right\},\quad LBP_L=\left\{\begin{bmatrix}y_{1,1}\\ \vdots \\ y_{n,1}\end{bmatrix}\cdots\begin{bmatrix}y_{1,L}\\ \vdots \\ y_{n,L}\end{bmatrix}\right\} \qquad (۶)$$



شکل (۴): الف) نمونه نویسه e برداشت شده از تصاویر پایگاه‌داده ب) ماسک حاصل برای جداکردن ناحیه بافت تیره، مرز و سایه
Figure (3): a) Sample of character e, taken from database images. b) The resulting mask for separate the dark, borders, and shadows areas

۴-۲- استخراج i-vector و طبقه‌بندی

در روش پیشنهادی فرایند استخراج i-vector از ویژگی‌های اولیه ناحیه‌های مختلف به‌صورت مجزا و موازی انجام می‌گیرد. برای ساخت UBM و استخراج زیر فضای متغیر کل چاپگر به‌ازای هر ناحیه، ابتدا تصاویر موجود در پایگاه‌داده به دو بخش آموزشی و آزمایشی تفکیک می‌شود. دنباله ویژگی‌های اولیه هر ناحیه از تصاویر آموزشی، برای ساخت UBM و تخمین فضای متغیر کل چاپگر در آن ناحیه استفاده می‌شود. تعداد مخلوط‌های گوسی سازنده UBM و بُعد فضای متغیر کل چاپگر با تعداد و تنوع بردارهای ویژگی که برای آموزش این دو فضا استفاده می‌شود ارتباط دارد. به‌ازای دنباله ویژگی اولیه هر ناحیه در هر صفحه، یک i-vector به‌عنوان ویژگی ثانویه استخراج می‌شود. بردارهای i-vector استخراج شده از ناحیه‌های متفاوت در هر تصویر به‌هم الحاق شده و بردار حاصل که آن را $i\text{-vector}_{total}$ می‌نامیم در طبقه‌بند بسته به نوع تصویر برای آموزش یا ارزیابی استفاده می‌شود. با استخراج یک بردار ویژگی ثانویه به‌ازای تصویر هر صفحه، فرایند پس‌پردازش رأی‌گیری اکثریت که بخش ثابت تمامی پژوهش‌های پیشین بود، حذف شده است. با کاهش تعداد و ابعاد کلی بردارهای ویژگی فرایند طبقه‌بندی سریع‌تر انجام می‌شود. استخراج i-vectorها در ساختاری بدون نظارت انجام می‌شود؛ بنابراین بخشی از فرایند جبران اثر تغییرات درون کلاسی باید قبل یا درون طبقه‌بند انجام شود. در فضای متغیر کل چاپگر، تغییرات درون کلاسی در اندازه i-vector و اطلاعات جداکننده کلاس‌ها در زاویه i-vector مشاهده می‌شود [۲۰]. روش‌های نرمال‌سازی که بر اساس کاهش کوواریانس درون کلاسی عمل می‌کنند مانند WCCN و یا استفاده از LDA برای نگاشت ویژگی‌های ثانویه در جهتی که بیشترین جداسازی بین کلاس‌ها را ایجاد کند، می‌تواند در کاهش این اثرات مؤثر باشد. ما تأثیر اعمال LDA و WCCN بر روی i-vectorها قبل از طبقه‌بندی را در بخش آزمایش‌ها بررسی می‌کنیم. انتخاب طبقه‌بند وابسته به توزیع و تعداد نمونه‌ها است. استفاده از طبقه‌بند PLDA[۱۳] (تجزیه و تحلیل افتراقی خطی احتمالی) و SVM رویکرد غالب در روش‌های بر پایه مدل‌سازی ویژگی‌ها در فضای UBM است. با این تفاوت که روش PLDA برای همگرا شدن به نتایج مطلوب، نیاز به نمونه‌های بیشتری نسبت به SVM دارد.

این در حالی است که SVM به دلیل عملکرد مبتنی بر پیدا کردن مرز بین کلاس‌ها می‌تواند در پایگاه داده با نمونه کم به نتایج مناسب‌تری برسد. به‌دلیل ساختار پایگاه داده استاندارد مورد استفاده در این مقاله که در بخش بعدی معرفی می‌شود، برای طبقه‌بندی از SVM با هسته غیر خطی گوسی[۱۴] (RBF) استفاده می‌کنیم.

۵- شبیه‌سازی و نتایج

در این بخش پایگاه‌داده، معیارهای ارزیابی، روش تعیین پارامترها و آزمایش‌ها را ارائه می‌دهیم. همچنین پس از ارزیابی روش پیشنهادی در سناریوهای مختلف، نتایج به‌دست‌آمده را تحلیل و با روش‌های مرز دانش در این حوزه مقایسه و بر روی نتایج بحث می‌کنیم.

۵-۱- پایگاه‌داده و روش ارزیابی

برای ارزیابی کارایی روش پیشنهادی از پایگاه‌داده مرجع [۱۲] استفاده شده است. این تنها پایگاه‌داده مناسب ارزیابی روش ما است که به‌صورت عمومی در دسترس قرار دارد. این پایگاه‌داده شامل ۱۲۰ سند از ویکی‌پدیا به دو زبان انگلیسی و پرتغالی، با قلم و سایز متفاوت است که توسط ۱۰ نوع چاپگر لیزری چاپ شده است. حدود ۵۰ درصد این اسناد علاوه بر متن شامل تصویر نیز است. تمام این اسناد بر روی کاغذ ۷۵ gr/m^2 چاپ شدند و توسط یک پویشگر مشابه با دقت 600dpi به نسخه دیجیتال در قالب فشرده نشده، ذخیره شدند. معیار ارزیابی ما برای سنجش میزان کارایی الگوریتم، صحت، دقت و معیار F1-score است که به کمک ماتریس درهم‌ریختگی CM که یک معیار شناخته شده در ارزیابی کارایی الگوریتم‌های طبقه‌بندی است، محاسبه می‌گردد. معیارهای صحت میانگین، صحت هر کلاس، دقت و F1 مطابق با روابط ۷ تا ۱۰ محاسبه می‌شوند.

$$\text{Accuracy}=\frac{\sum_{i=1}^{10}\text{CM}(i,i)}{\sum_{i=1}^{10}\sum_{j=1}^{10}\text{CM}(i,j)} \quad (۷)$$

$$\text{Precision}(i)=\frac{\text{CM}(i,i)}{\sum_{j=1}^{10}\text{CM}(j,i)} \quad (۸)$$

$$\text{Recall}(i)=\frac{\text{CM}(i,i)}{\sum_{j=1}^{10}\text{CM}(i,j)} \quad (۹)$$

$$\text{F1-score}(i)=2\times\frac{\text{Precision}(i)\times\text{Recall}(i)}{\text{Precision}(i)+\text{Recall}(i)} \quad (۱۰)$$

برای ارزیابی این معیارها از اعتبارسنجی متقابل ۵ لایه بر روی ویژگی‌های اولیه بهره بردیم. نتایج آزمایش با میانگین‌گیری از پنج مرتبه تکرار اعتبارسنجی متقابل ۵ لایه روی برچسب‌های تصادفی حاصل می‌شود. این موضوع سابقه دارد و در مراجع [۸]، [۹] و [۱۲] از پنج مرتبه تکرار اعتبار سنجی متقابل ۲ لایه استفاده شده است. جهت تعیین پارامترهای مناسب برای مدل‌سازی ویژگی‌های اولیه LBP در فضای متغیر کل چاپگر و استخراج i-vectorها از یک جستجوی شبکه‌ای بر روی بازه‌ای از مقادیر C و T_D (C (تعداد مخلوط‌های گوسی سازنده UBM)و T_D (بعد فضای متغیر کل چاپگر)) استفاده می‌شود. برای طبقه بندی i-vectorها، طبقه بند SVM با هسته RBF با ساختار یک در برابر یک، بکار گرفته شده است. همچنین تأثیر به‌کارگیری روش‌های LDA و WCCN بر روی i-vectorها قبل از طبقه‌بندی بررسی می‌شود.

۵-۲- ارزیابی مدل‌سازی بافت موجود در ناحیه‌های مختلف در فضای متغیر کل چاپگر

یکی از پیشنهادهای این مقاله، استفاده از بافت چاپ موجود در نواحی تیره، مرز و سایه چاپ با مدل‌سازی مجزای ویژگی‌های اولیه بافت LBP هر یک از این ناحیه‌ها و استخراج ویژگی‌های ثانویه به‌ازای هر ناحیه است. برای بررسی این موضوع و برای

داشتن معیاری جهت سنجش میزان کارایی روش پیشنهادی در آزمایش‌های مشابهی، ابتدا ویژگی‌های اولیه LBP_D، LBP_E و LBP_L استخراج شده از بافت چاپ نواحی مختلف هر یک از تکه‌های تصویر را بدون اعمال روش پیشنهادی به‌صورت مجزا و به‌صورت الحاق شده (LBP_{total}=[LBP_D , LBP_E, LBP_L]) توسط SVM طبقه‌بندی و با رأی‌گیری اکثریت نتیجه انتساب چاپگر بر روی تصاویر آزمایشی را مشخص کردیم. در مرحله دوم بر اساس روش پیشنهادی با مدل‌سازی مجزای ویژگی‌های اولیه هر ناحیه در فضای متغیر کل چاپگر، به‌ازای تصویر هر صفحه، ویژگی‌های ثانویه $i\text{-vector}_D$، $i\text{-vector}_E$ و $i\text{-vector}_L$ به ترتیب به‌ازای ناحیه تیره، مرز و سایه چاپ به دست آورده و ابتدا به‌صورت مجزا، سپس دوبه‌دو و در نهایت با الحاق به هم و ساخت $i\text{-vector}_{total}$ فرایند طبقه‌بندی SVM انجام می‌شود. در آخرین مرحله از اولین ارزیابی، ویژگی‌های اولیه LBP_{total} در یک فضای مشترک پالایش شده تنها یک i-vector به‌ازای هر سه ناحیه برای طبقه‌بندی SVM استخراج و ارزیابی می‌شود. نتایج این ارزیابی‌ها بر اساس تکه بندی ۴۴۲ عددی در جدول (۱) نمایش‌داده‌شده است. در تمامی مراحل این آزمایش مناسب‌ترین مقادیر برای پارامترهای مدل‌سازی فضای متغیر کل چاپگر بر اساس جستجوی شبکه‌ای به‌دست‌آمده است. همچنین برچسب نمونه‌های آموزشی و تست و تعداد تکرار الگوریتم در تمامی آزمایش‌ها کاملاً مشابه است. بر اساس نتایج جدول (۱)، طبقه‌بندی مستقیم ویژگی‌های LBP بر روی بافت سایه چاپ مقدار صحت ۳۴/۵۹ درصد را نتیجه می‌دهد. با مقایسه طبقه‌بندی $i\text{-vector}_L$ در روش پیشنهادی با صحت ۸۲/۲۳ درصد بهبود کارایی به‌وضوح مشخص است. همین مسئله در طبقه‌بندی بافت تیره و مرز در روش پیشنهادی به ترتیب باعث ۱۰/۲۹ درصد و ۲۶/۴۴ درصد افزایش در صحت طبقه‌بندی شده است؛ بنابراین روش پیشنهادی به‌خوبی اطلاعات جداکننده کلاس‌ها در نواحی مختلف را از بافت چاپ استخراج می‌کند. این موضوع به‌خصوص در بافت سایه و مرز که تنک و پراکنده هستند بیشتر قابل‌توجه است. مقایسه نتایج طبقه‌بندی بردار ویژگی اولیه حاصل از الحاق LBP هر سه ناحیه یعنی LBP_{total} و طبقه‌بندی $i\text{-vector}_{total}$ (حاصل از الحاق i-vector سه ناحیه) نتایج با افزایش کارایی ۱۱/۶۷ درصد به نفع روش پیشنهادی است. به‌علاوه مقایسه نتایج طبقه‌بندی $i\text{-vector}_D$ و $i\text{-vector}_{total}$ نشان می‌دهد الحاق ویژگی‌های ثانویه بافت سایه و مرز باعث افزایش کارایی ۱/۷ درصد شده است. بااین‌وجود ممکن است احساس شود به‌کارگیری بافت تیره و مرز برای رسیدن به صحت مناسب کفایت می‌کند و بکارگیری بافت سایه چاپ تأثیری به‌جز افزایش بار محاسباتی نخواهد داشت. به این منظور ما در آزمایش مشابهی i-vectorهای استخراج شده از بافت تیره، مرز و روشن را دوبه‌دو با هم الحاق کرده و نتایج طبقه‌بندی را در جدول (۱) جهت مقایسه قرار داده‌ایم.

Table (1): Performance evaluation of the proposed algorithm for different areas of the image in 442-patches

جدول (۱): ارزیابی کارایی الگوریتم پیشنهادی برای ناحیه‌های متفاوت تصویر در حالت ۴۴۲ تکه‌ای

ویژگی ورودی طبقه‌بند SVM با کرنل RBF	صحت میانگین	C	T_D	بعد بردار ویژگی
*ویژگی‌های اولیه LBP_L بافت سایه چاپ	۳۴/۵۹	-	-	۵۹
*ویژگی‌های اولیه LBP_E بافت مرز چاپ	۶۷/۲۷	-	-	۵۹
*ویژگی‌های اولیه LBP_D بافت تیره چاپ	۸۵/۷۴	-	-	۵۹
*ویژگی‌های اولیه LBP_{total}=[LBP_D , LBP_E , LBP_L]	۸۶/۰۳	-	-	۱۷۷
ویژگی‌های ثانویه $i\text{-vector}_L$ با مدل‌سازی LBP_L	۸۲/۲۳	۱۶	۲۴	۱۶
ویژگی‌های ثانویه $i\text{-vector}_E$ با مدل‌سازی LBP_E	۹۳/۷۱	۱۶	۲۴	۱۶
ویژگی‌های ثانویه $i\text{-vector}_D$ با مدل‌سازی LBP_D	۹۶/۰۳	۱۶	۲۴	۱۶
ویژگی‌های ثانویه $i\text{-vector}_{total}$= [$i\text{-vector}_D$, $i\text{-vector}_E$, $i\text{-vector}_L$]	**۹۷/۷۰**	**۱۶**	**۲۴**	**۴۸**
ویژگی‌های ثانویه i-vector با پالایش بردار ویژگی LBP_{total}	۹۵/۷۴	۸	۲۸	۸
ویژگی‌های ثانویه [$i\text{-vector}_D$, $i\text{-vector}_L$]	۹۷/۱۲	۱۶	۲۴	۳۲
ویژگی‌های ثانویه [$i\text{-vector}_D$, $i\text{-vector}_E$]	۹۷/۳۱	۱۶	۲۴	۳۲
ویژگی‌های ثانویه [$i\text{-vector}_E$, $i\text{-vector}_L$]	۹۴/۹۸	۱۶	۲۴	۳۲

C: تعداد مخلوط‌های گوسی سازنده UBM

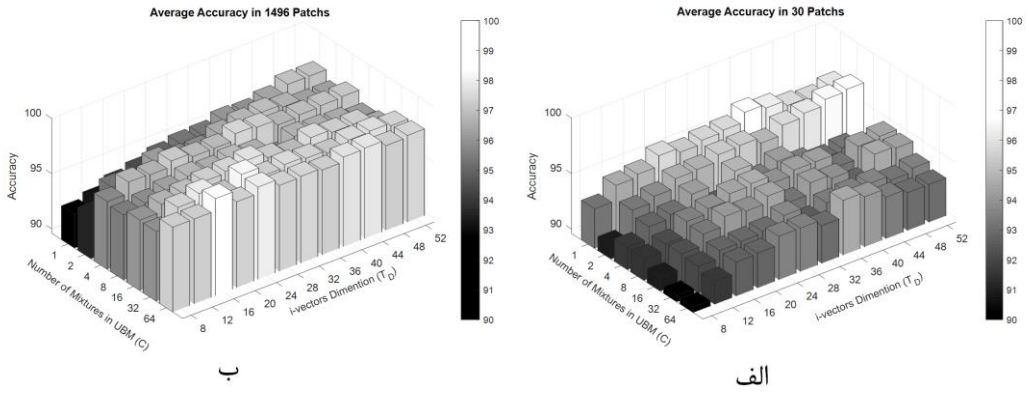
T_D : بعد فضای متغیر کل چاپگر

نتایج این آزمایش تأثیر مثبت بافت سایه چاپ در تصویر سند برای افزایش صحت طبقه‌بندی بر اساس الگوریتم پیشنهادی را تأیید می‌کند. نتایج طبقه‌بندی در آخرین ارزیابی که در آن از ویژگی‌های اولیه LBP_{total} تنها یک i-vector مشترک استخراج شده دارای صحت ۹۵/۷۴ درصد است که حدود ۲ درصد نسبت به مدل‌سازی ویژگی‌های اولیه در فضاهای مجزا، افت کارایی را نشان می‌دهد. در واقع در این حالت نتیجه طبقه‌بندی حتی نسبت به طبقه‌بندی i-vector بافت تیره نیز تا حدی کمتر شده است؛ بنابراین مدل‌سازی و پالایش ویژگی‌های اولیه هر ناحیه در فضای مجزا روش مناسب‌تری است.

۵-۳- بررسی ابعاد تکه‌ها در کارایی الگوریتم پیشنهادی

اندازه تکه‌ها و به‌تناسب آن تعداد تکه‌ها یکی از پارامترهای تأثیرگذار در نتایج الگوریتم پیشنهادی است که در این ارزیابی بررسی می‌شود. تکه‌بندی تصویر و استخراج ویژگی‌های اولیه از بافت ناحیه‌های مختلف در هر یک از تکه‌ها، روش جایگزین OCR در معماری پیشنهادی این مقاله است. به این منظور ما کارایی الگوریتم پیشنهادی را با استفاده از طبقه‌بند SVM با کرنل RBF بر روی $i\text{-vector}_{total}$، توسط تکه بندی‌هایی با تعداد ۳۰، ۱۹۲، ۴۴۲، ۸۸۴ و ۱۴۹۶ ارزیابی می‌کنیم. به‌ازای هر یک از حالت‌ها برای یافتن بیشترین مقدار صحت از یک جستجوی شبکه‌ای بر روی C و T_D استفاده می‌شود. برای درک این موضوع در شکل(۵) دو نمودار ستونی سه‌بعدی به‌ازای تغییر مقادیر C و T_D برای حالت ۳۰ و ۱۴۹۶ تکه‌ای نمایش‌داده‌شده است. ارتفاع هر یک از ستون‌ها در شکل میزان میانگین صحت به‌ازای C و T_D متناظر را بیان می‌کند. به‌صورت کلی در استخراج ویژگی‌های اولیه LBP، هر چه ابعاد تکه‌ها بزرگ‌تر باشد ویژگی‌ها کلی‌تر و کوچک‌تر شدن ابعاد جزئیات بیشتری را استخراج می‌کند. با مقایسه نمودارهای (الف) و (ب) در شکل (۵) واضح است تعداد تکه‌ها در انتخاب پارامترهای مؤثر برای مدل‌سازی فضای پس‌زمینه و استخراج i-vector به‌عنوان ویژگی‌های ثانویه تأثیر می‌گذارد. در روش پیشنهادی به‌ازای تمام تکه‌های هر صفحه، فقط یک i-vector به‌عنوان بردار ویژگی ثانویه استخراج می‌شود و در فرایند استخراج i-vectorها فرض این است که هر مجموعه ویژگی اولیه از یک سند، مربوط به یک کلاس مستقل است. پس با افزایش تعداد تکه‌ها برای مدل‌سازی فضای پس‌زمینه، امکان آموزش کامل‌تر یک GMM با بعد بیشتر وجود دارد. این موضوع به‌وضوح با بررسی نمودارهای شکل (۵) مشاهده می‌شود.

در این ارزیابی به‌ازای تکه بندی‌های مختلف، هیچ‌گونه تفکیکی بین اسناد متنی و اسناد داری تصویر و متن در نظر گرفته نشده و ویژگی‌های اولیه از همه بخش‌های اسناد استخراج می‌شود. وجود متن با قلم و اندازه متفاوت و وجود هم‌زمان تصویر در بین ۵۰ درصد اسناد و عدم حذف آن باعث تغییرات بین کلاسی در بین ویژگی‌ها خواهد شد. به همین خاطر در هر حالت تأثیر به‌کارگیری LDA و WCCN بر روی i-vectorها برای کاهش اثرات تغییرات درون کلاسی نیز بررسی می‌شود. نتایج این ارزیابی در جدول (۲) گزارش شده است. در جدول (۲) به‌ازای تکه بندی‌های مختلف، نمونه یک‌تکه تصویر، برای داشتن تصوّر دیداری از اندازه و میزان بافت چاپ درون آن، نمایش‌داده‌شده است. در این آزمایش‌ها با افزایش تعداد تکه‌ها به‌ازای هر صفحه و بزرگ‌تر شدن GMMها ابعاد فضای متغیر کل چاپگر و در نتیجه ابعاد i-vectorها کاهش‌یافته است.



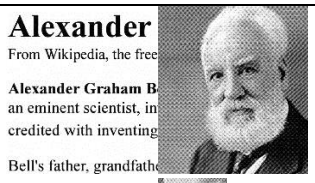
شکل (۵): جستجوی شبکه‌ای بر روی صحت میانگین، برای تعیین پارامترها تخمین i-vector در الف)۳۰ و ب) ۱۴۹۶ تکه‌ای

Figure (5): The grid search on average accuracy, to determine i-vector estimation parameters a) in 30 and b) in 1496 paches

افزایش تعداد تکه‌ها با تأثیر بر انتخاب پارامترهای T_D و C باعث مدل‌سازی دقیق‌تر i-vector هر سند خواهد شد. ایجاد یک مصالحه مناسب بین انتخاب اندازه تکه‌ها و پارامترهای استخراج i-vector می‌تواند ما را به رسیدن به نتیجه بهتر هدایت کند.
در جدول (۲)، ستون a نمایش‌دهنده صحت میانگین در خروجی الگوریتم پیشنهادی با طبقه‌بند SVM بدون به‌کارگیری روش‌های جبران‌سازی تغییرات درون کلاسی است. در این ستون افزایش صحت میانگین و کاهش واریانس با افزایش تعداد تکه ها در هر صفحه بر اساس نتایج شبیه‌سازی تأیید می‌شود. با افزایش تعداد تکه‌ها به‌ازای هر صفحه، به دلیل افزایش تعداد بردارهای ویژگی هر صفحه فرایند تطبیق مدل پس‌زمینه به‌ازای ویژگی‌های اولیه دقیق‌تر/کامل‌تر انجام می‌شود و باعث شده i-vector معادل دقیق‌تر استخراج شود.

Table (2): Evaluation of the proposed algorithm in different number of patches

جدول (۲): ارزیابی الگوریتم پیشنهادی در تکه‌بندی‌های متفاوت

تعداد تکه‌ها در هر صفحه	ابعاد تکه (پیکسل)	نمایی از هر تکه	a	b	c
۳۰	۹۶۰×۸۶۰		۹۶/۸۷±۰/۸۶ T_D=۴۸, C=۲	۹۷/۳۶±۰/۶۵	۹۷/۳۶±۰/۶۱
۱۹۲	۳۶۰×۳۶۰		۹۷/۲۱±۰/۶۴ T_D=۴۴, C=۱۶	۹۸/۱۰±۰/۴۹	۹۸/۲۸±۰/۵۲
۴۴۲	۲۲۰×۲۶۰		۹۷/۷۰±۰/۵۱ T_D=۲۴, C=۱۶	۹۷/۸۷±۰/۷۸	۹۸/۶۴±۰/۵۰
۸۸۴	۱۷۰×۱۷۰		۹۷/۹۸±۰/۵۷ T_D=۲۰, C=۳۲	۹۸/۵۵±۰/۶۷	۹۸/۸۱±۰/۴۹
۱۴۹۶	۱۳۰×۱۳۰		۹۸/۱۲±۰/۳۷ T_D=۱۲, C=۶۴	۹۸/۴۲±۰/۴۱	۹۹/۰۵±۰/۳۴

a) طبقه‌بندی $i\text{-vector}_{total}$ توسط SVM با کرنل RBF

b) اعمال LDA بر روی $i\text{-vector}_{total}$ و طبقه‌بندی توسط SVM با کرنل RBF

c) اعمال LDA و WCCN بر روی $i\text{-vector}_{total}$ و طبقه‌بندی توسط SVM با کرنل RBF

توجه: بُعد بردارهای ویژگی باتوجه‌به تعداد کلاس‌ها در b و c پس از اعمال LDA به ۹ تقلیل می‌یابد.

تغییرات درون کلاسی از طریق طبقه‌بند SVM با کرنل RBF تااندازه‌ای جبران می‌شود، بااین‌وجود در این جدول در ستون b و c به ترتیب نتیجه اعمال LDA و LDA+WCCN بر روی i-vectorها برای جبران اثرات تغییرات درون کلاسی نیز بررسی شده است. باتوجه‌به نتایج ارزیابی‌ها به‌کارگیری LDA و WCCN به همراه طبقه‌بندی SVM در تمامی حالت‌ها باعث افزایش حدود ۱ درصد تا ۲ درصد کارایی شده و باعث می‌شود نتایج با روش‌های مطرح در این حوزه قابل قیاس شود. در واقع انگیزه اصلی در استفاده از LDA و WCCN کاهش ابعاد بردارهای ویژگی نیست، بلکه این روش‌ها کوواریانس درون کلاسی را کاهش می‌دهند و داده‌ها را در جهتی که بهترین جداسازی کلاس‌ها ممکن باشد می‌نگارند. این موضوع به کاهش تغییرات درون کلاسی و افزایش کارایی الگوریتم در شناسایی چاپگر منبع سند کمک می‌کند. البته اثر کاهش ابعاد بردارهای ویژگی، بَعد از استفاده از LDA بر سرعت طبقه‌بندی مسئله مهم و قابل‌توجه‌ای است. زمان پردازش یکی از پارامترهایی است که در کاربردهایی با تعداد نمونه‌های بالا اهمیت پیدا می‌کند و می‌توان آن را متناسب با بار محاسباتی الگوریتم در نظر گرفت. البته اشاره به این نکته ضروری است که اصولاً شناسایی چاپگر یک فرایند برون خط است و در نتیجه پیچیدگی محاسباتی و زمان پردازش الگوریتم‌ها کمتر برای پژوهشگران اهمیت داشته است. زمان پردازش برای استخراج i-vectorها با بزرگ‌تر شدن بعد GMM و بعد فضای متغیر کل چاپگر افزایش می‌یابد. ایجاد یک مصالحه بین انتخاب پارامترها و صحت طبقه‌بندی باتوجه‌به زمان پردازش ممکن است در برخی کاربردها اهمیت داشته باشد. ما برای اینکه درک بهتری از این موضوع داشته باشیم، توسط کامپیوتر دسکتاپ با پردازنده Intel i5-4670@3.4GHz و ۸ گیگابایت حافظه، زمان استخراج i-vectorها و طبقه‌بندی را در

اعتبارسنجی متقابل ۵ لایه در دو حالت ۳۰ و ۱۴۹۶ تکه‌ای (بدون اعمال LDA و WCCN) اندازه‌گیری کردیم. زمان استخراج i-vectorها در حالت ۳۰ تکه‌ای با انتخاب C=۲ و T_D=۵۲ برابر ۹۷/۵۷ ثانیه و ۴/۲۱ ثانیه برای طبقه‌بندی است. این زمان برای حالت ۱۴۹۶ تکه‌ای با انتخاب C=۶۴ و T_D=۱۲ برابر ۲۸۴۱/۵۶ ثانیه برای i-vector و ۱/۲۱ ثانیه برای طبقه‌بندی است. در مجموع فرایند استخراج ویژگی‌های ثانویه بسته به انتخاب پارامترهای C و T_D و تعداد تکه‌ها در محدوده ۳۰ ثانیه تا ۵۵ دقیقه است. در روش پیشنهادی به‌ازای تمامی ویژگی‌های استخراج شده از یک صفحه سند فقط یک بردار ویژگی ثانویه با ابعاد کم استخراج می‌شود. ازاین‌رو زمان طبقه‌بندی ویژگی‌ها نسبت به روش‌های مشابه کاهش می‌یابد. این موضوع امکان اجرای سریع‌تر طبقه‌بندی SVM با هسته غیرخطی را فراهم می‌کند. با مقایسه حداکثر زمان پردازش روش پیشنهادی با مقدار گزارش شده در پژوهش [۱۱] که ۱۴/۵ساعت است، سریع‌تر بودن روش پیشنهادی مشاهده می‌شود که می‌تواند یک مزیت باشد. برای داشتن نتیجه قابل قیاس در شرایط برابر، زمان طبقه‌بندی مستقیم ویژگی‌های LBP در حالت ۴۴۲ تکه‌ای مربوط به آزمایش اول را در کامپیوتر با مشخصات ذکر شده بدون استخراج i-vector، اندازه گرفتیم که حدود ۸ ساعت است که نسبت به طبقه‌بندی i-vector ها بسیار بیشتر است.

۵-۴- آنالیز خطا و مقایسه نتایج با روش‌های مرز دانش

برای آنالیز خطا ماتریس درهم‌ریختگی به‌ازای بهترین نتیجه، در جدول (۳) نمایش‌داده‌شده است. همان‌طور که مشاهده می‌شود نتایج طبقه‌بندی برای شش چاپگر B4070، C1150، H1518، LE260، OC330 و SC315 صحت ۱۰۰ درصد را نشان می‌دهد و مقدار خطا برای چاپگر C4370 و C3240 بسیار ناچیز است. مقدار حداقل صحت مربوط به کلاس H225A و سپس کلاس H225B، به ترتیب با ۹۴/۲۴ درصد و ۹۶/۸۸ درصد است. ازآنجایی‌که این دو چاپگر مشابه هستند، امضای ذاتی این دو وابستگی بیشتری به یکدیگر دارد و باعث ایجاد خطا می‌شود. علاوه بر این کاهش صحت شناسایی در این دو کلاس می‌تواند به دلیل عملکرد روش پیشنهادی در کاهش فاصله‌های درون کلاسی نیز باشد. رفع این مشکل می‌تواند به کارایی بیشتر الگوریتم منجر شود. بررسی بیشتر بر روی معیار دقت و F1-score نیز نتایج مشابه جدول صحت را به‌ازای کلاس‌های مختلف تأیید می‌کند. در جدول (۴) نتایج ارزیابی صحت و F1-score روش پیشنهادی با روش‌های مرز دانش در این حوزه مقایسه شده است. نتایج تمامی روش‌های گزارش شده در این جدول بر روی یک پایگاه‌داده یکسان به‌دست‌آمده است. همان‌طور که مشخص است روش پیشنهادی در حالت ۱۴۹۶ تکه با صحت ۹۹/۰۵ درصد و میزان F1-score ۰/۹۹۰۵ از تمامی روش‌های به‌روز در این حوزه بیشتر است. در روش مرجع [۸] که با صحت ۹۸/۹۲ درصد نتایج قابل رقابتی را نمایش می‌دهد از بردار ویژگی با ابعاد ۱۰۸۵۶ استفاده شده که قطعاً فرایند طبقه‌بندی در این روش در مقایسه با روش پیشنهادی با بردارهای ویژگی ۲۷ بعدی بسیار متفاوت است.

Table (3): Confusion matrix

جدول (۳): ماتریس درهم‌ریختگی

مدل چاپگر	طبقه‌بندی صحیح	طبقه‌بندی غلط	دقت هر کلاس	F1-score
B4070	۱۰۰	۰	۱۰۰	۱
C1150	۱۰۰	۰	۱۰۰	۱
C3240	۹۹/۸۳	۰/۱۷ (C4370)	۹۹/۳۳	۰/۹۹۵۸
C4370	۹۹/۳۳	۰/۶۷ (C3240)	۹۹/۸۳	۰/۹۹۵۸
H1518	۱۰۰	۰	۱۰۰	۱
H225A	۹۴/۲۴	۵/۷۶ (H225B)	۹۷/۰۳	۰/۹۵۶۱
H225B	۹۶/۸۸	۳/۱۲ (H225B)	۹۳/۹۵	۰/۹۵۳۹
LE260	۱۰۰	۰	۱۰۰	۱
OC330	۱۰۰	۰	۱۰۰	۱
SC315	۱۰۰	۰	۱۰۰	۱
صحت میانگین	۹۹/۰۵			

از نتایج جدول (۴) مشخص است که روش پیشنهادی توانسته از ویژگی‌های اولیه LBP ناحیه‌های مختلف اطلاعات تفکیک کننده مناسبی را استخراج کند و این اطلاعات مفید را در قالب i-vector با بعد کم بنگارد. قابل ذکر است که هدف اصلی این پژوهش نشان دادن قابلیت نگاشت ویژگی‌های اولیه استخراج شده از نواحی متفاوت با مدل‌سازی فضای متغیر کل چاپگر به‌صورت تفکیک شده و استخراج i-vector ها به‌عنوان ویژگی ثانویه از هر ناحیه و نمایش قابلیت‌های ایجاد شده از این طریق در حوزه شناسایی منبع چاپ بوده است. برخلاف برخی از پژوهش‌های جدول (۴)، که سعی کردند با ترکیب مجموعه‌ای از ویژگی‌های مختلف و یا استخراج ویژگی چند مقیاسی و چندجهتی، اطلاعات بیشتری را برای افزایش صحت طبقه‌بندی بدست آوردند نتایج روش پیشنهادی، فقط با پالایش ویژگی‌های اولیه که از نسخه اولیه LBP بدست می‌آید، حاصل شده است. در روش پیشنهادی به‌دلیل استفاده از بافت باقیمانده چاپ در تمامی نواحی، تعداد اسنادی که به دلیل کمبود بافت چاپ در چاپگرهای مختلف از فرایند ارزیابی حذف می‌شوند نسبت به روش‌های دیگر جدول (۴) کمتر است به همین دلیل مقدار F1-score کلی که بر اساس وزن کلاس‌ها محاسبه شده بسیار به مقدار صحت نزدیک است. این نشان می‌دهد با وجود اسناد شامل بافت تیره کم، روش پیشنهادی در تصمیم گیری دچار خطا نمی‌شود. نتایج، رسیدن به صحت بیشتر در طبقه‌بندی به‌واسطه روش پیشنهادی و تأثیر مدل‌سازی ویژگی‌های اولیه در فضای متغیر کل چاپ و استخراج ویژگی‌های ثانویه را نشان می‌دهد. روش پیشنهادی بر اساس مدل‌سازی ویژگی‌های ثانویه از طریق تحلیل عوامل مشترک انجام می‌شود که قبل از این در حوزه شناسایی گوینده برای استخراج ویژگی‌های مستقل از محتوا بکار گرفته می‌شد. در واقع باتکیه‌بر روش‌های بر پایه تحلیل عوامل مشترک و تطبیق فضای متغیر کل چاپ به مسئله شناسایی چاپگر از طریق الگوریتم پیشنهادی، ویژگی‌های ثانویهٔ استخراج شده وابستگی کمتری به شکل متن (به‌عنوان محتوا) دارند و بیشتر به بافت چاپ وابسته هستند. به‌کارگیری این روش به همراه تفکیک بافت چاپ در ناحیه‌های مختلف و مدل‌سازی مجزای آن‌ها در فضاهای موازی، به دلیل تراکم و ساختار متفاوت آن‌ها، کمک کرده به صحت مناسبی در شناسایی چاپگر منبع توسط روش پیشنهادی برسیم. یکی دیگر از دلایل بهبود کارایی روش پیشنهادی کاهش تغییرات درون کلاسی است. ازآنجایی‌که روش پیشنهادی مبتنی بر مدل‌سازی ویژگی‌های اولیه در مدل UBM-GMM است توانسته تأثیر این تغییرات در ویژگی‌های ثانویه را کاهش دهد و استفاده از روش‌های LDA و WCCN به کاهش باقیمانده این تغییرات در i-vector ها کمک می‌کند. همین موضوع باعث شده بتوانیم از ویژگی‌های استخراج شده در بافت باقیمانده چاپ در ناحیه روشن برای افزایش کارایی الگوریتم پیشنهادی بهره ببریم. روش‌های CTGF3×3 [۱۱]، CNN$\{S^{raw},S^{med},S^{avg}\}_{a,e}$ [۱۰] و CC-RS-LTrP-PoEP [۹] ویژگی‌ها را از بافت موجود در نویسه‌های خاص یا تمامی نویسه‌ها، استخراج می‌کنند. این موضوع، فرایند استخراج ویژگی را نیازمند OCR یا فرایندهای مشابه می‌کند و باعث ایجاد پیچیدگی می‌شود، همچنین ممکن است محدودیت استفاده از زبان خاص و نوع و یا سایز قلم را ایجاد کند. بعلاوه در اسناد دارای تصویر نیز بخشی از اطلاعات موجود در تصاویر بدون استفاده می‌ماند. عملاً روش پیشنهادی با افراز تصویر به تکه‌های کوچک‌تر در برابر OCR این محدودیت‌ها را ندارد.

Table (4): Comparison of the proposed method with knowledge boundary methods

جدول (۴): مقایسه روش پیشنهادی با روش‌های مطرح

روش و ویژگی	طبقه بند	صحت میانگین(%)	انحراف معیار	F1-score	بعد بردار ویژگی
CTGF-[۱۲]	SVM	۷۲/۴۶	۰/۰۳۷۷	n.a.	۲۲۹۵
GLCM-MD-[۱۲]	SVM	۹۱/۰۸	۰/۰۰۸۹	n.a.	۱۷۶
GLCM-MDMS-[۱۲]	SVM	۹۴/۳۰	۰/۰۱۱	n.a.	۷۰۴
CTGF3×3-[۱۱]	Random Forest	۹۵/۶۰	۳/۸۶	۰/۹۱۹۵	۵۴۷
CTGF_GLCM_MDMS-[۱۲]	SVM	۹۶/۲۶	۰/۰۰۵۴	۰/۹۶۲۵	۹۹۷۰+۷۰۴
CC-RS-LTrP-PoEP-[۹]	SVM	۹۷/۱۲	۰/۶۷	۰/۹۷۱۷	۱۵۳۴
CNN$\{S^{raw},S^{med},S^{avg}\}_{a,e}$-[۱۰]	SVM	۹۷/۳۳	۰/۰۰۶۵	۰/۹۷۳۷	۱۵۰۰+۱۵۰۰
PLSTD-[۸]	SVM	۹۸/۹۲	۰/۴۳	۰/۹۸۹۴	۱۰۸۵۶
روش پیشنهادی	**SVM**	**۹۹/۰۵**	**۰/۳۴**	**۰/۹۹۰۵**	**۲۷=۹+۹+۹**

در تمامی روش‌های موجود در جدول (۴)، طبقه‌بندی بر روی ویژگی‌های استخراج شده از اجزای کوچک‌تر در هر صفحه انجام می‌شود. رأی‌گیری اکثریت، بر روی نتایج طبقه‌بندی اجزای درون یک صفحه، بخش پایانی تمامی این روش‌ها است. ما با پیشنهاد استخراج تنها یک بردار ویژگی ثانویه به‌ازای هر صفحه فرایند رأی‌گیری را حذف و سرعت طبقه‌بندی را افزایش دادیم. در روش پیشنهادی با کاهش این پیچیدگی‌ها نتایج قابل قبولی را در مقایسه با روش‌های مرز دانش به‌دست آورده‌ایم.

۶- نتیجه‌گیری

در این مقاله یک روش جدید بر مبنای ناحیه‌بندی بافت چاپ و استخراج ویژگی‌های ثانویه i-vector از هر ناحیه در فضای متغیر کل چاپگر، برای شناسایی منبع چاپ سند ارائه شده است. روش پیشنهادی بر پایه استفاده از روش‌های GMM-UBM و مدل‌سازی ویژگی‌های اولیه در فضای متغیر کل چاپگر، اطلاعات تفکیک کننده‌ای را از نواحی مختلف سند مخصوصاً بافت باقیمانده در ناحیه روشن، نتیجه می‌دهد. ویژگی‌های ثانویه i-vector با افراز تصویر اسناد به تکه‌های کوچک‌تر و مدل‌سازی ویژگی‌های اولیه این تکه‌ها و با حذف فرایند OCR به‌دست می‌آید. با این روش نسبت به تغییرات نوع و اندازه قلم اسناد استقلال ایجاد شده است. با این وجود نتایج آزمایش‌های به‌دست آمده صحت مناسب‌تری نسبت به روش‌های مرز دانش در این حوزه را نمایش می‌دهد. ما در این مقاله، تأثیر استخراج i-vector از ناحیه‌های مختلف بافت تیره، مرز و سایه چاپ را به‌دقت بررسی کردیم و از طریق نتایج شبیه‌سازی نشان دادیم که روش پیشنهادی با صحت و دقت بیشتری نسبت به روش‌های مشابه می‌تواند مدل چاپگر منبع سند را مشخص کند. همچنین تأثیر اعمال LDA و WCCN بر روی ویژگی‌های ثانویه با کاهش تغییرات درون کلاسی و ایجاد تفکیک‌پذیری بهتر بین کلاس‌ها، در طبقه‌بند SVM با کرنل RBF بررسی شد که منجر به افزایش کارایی الگوریتم گردید. با بررسی و مقایسه زمان استخراج ویژگی‌های i-vector و آموزش و ارزیابی طبقه‌بند، نشان دادیم روش پیشنهادی به دلیل استخراج یک بردار ویژگی با بعد کم به‌ازای تصویر هر صفحه و حذف فرایند رأی‌گیری اکثریت می‌تواند فرایند شناسایی را سریع انجام دهد. تأثیر اندازه و تعداد تکه‌ها را در کارایی الگوریتم پیشنهادی و اثر آن بر پارامترهای تأثیرگذار در استخراج ویژگی‌های ثانویه یکی دیگر از مواردی بود که در این مقاله به‌دقت بررسی شد و مشخص شد که با افزایش تعداد تکه‌ها صحت میانگین افزایش می‌یابد. به‌عنوان پیشنهاد برای ادامه کار ارزیابی مقاومت روش پیشنهادی در برابر انواع حمله‌های تصویری مانند تغییر میزان روشنایی، تغییر ابعاد ، فشرده‌سازی و موارد مشابه برای کارهای آینده پیشنهاد می‌شود. همچنین ساخت یک پایگاه داده مناسب که شامل اسناد چاپ شده به زبان‌های مختلف و بر روی کاغذهای مختلف و پویش این اسناد با چند نوع پویشگر، می‌تواند به تحقیقات در این زمینه برای پیشرفت این نوع الگوریتم‌ها در شرایط واقعی کمک کند.

References

مراجع

[1] P. Yang, D. Baracchi, R. Ni, Y. Zhao, F. Argenti, A. Piva, "A survey of deep learning-based source image forensics", Journal Imaging, vol. 6, no. 3, p. 9, Mar. 2020 (doi: 10.3390/jimaging6030009).

[2] V. Itier, O. Strauss, L. Morel, W. Puech, "Color noise correlation-based splicing detection for image forensics", Multimedia Tools and Applications, pp. 1–19, Jan. 2021 (doi: 10.1007/s11042-020-10326-5).

[3] A.T.S. Ho, S. Li, Handbook of digital forensics of multimedia data and devices, Chichester, UK: John Wiley & Sons, Ltd, 2015.

[4] P.-J. Chiang, J.P. Allebach, G.T.-C. Chiu, "Extrinsic signature embedding and detection in electrophotographic halftoned images through exposure modulation", IEEE Trans. on Information Forensics and Security, vol. 6, no. 3, pp. 946–959, Sept. 2011 (doi: 10.1109/TIFS.2011.2156789).

[5] P.J. Chiang, G.N. Ali, A.K. Mikkilineni, E.J. Delp, J.P. Allebach, G.T.C. Chiu, "Extrinsic signatures embedding and detection for information hiding and secure printing in electrophotography", Proceeding of the IEEE/ACC, pp. 1-6, Minneapolis, MN, USA, June 2006 (doi: 10.1109/ACC.2006.1656604).

[6] G. Adams, S. Pollard, S. Simske, "A study of the interaction of paper substrates on printed forensic imaging", Proceedings of the ACM, pp. 263-266, Limerick , Ireland, Sept. 2011 (doi: 10.1145/2034691.20-34743).

[7] W. Jiang, A.T.S.S. Ho, H. Treharne, Y.Q. Shi, "A novel multi-size block Benford's law scheme for printer

identification", Pacific-Rim Conference on Multimedia, vol. 6297 LNCS, no. PART 1, pp. 643–652, Springer, 2010.
[8] S. Joshi, N. Khanna, "Source printer classification using printer specific local texture descriptor", IEEE Trans. on Information Forensics and Security, vol. 15, no. 1, pp. 160–171, 2020 (doi: 10.1109/TIFS.2019.2-919869).
[9] S. Joshi, N. Khanna, "Single classifier-based passive system for source printer classification using local texture features", IEEE Trans. on Information Forensics and Security, vol. 13, no. 7, pp. 1603–1614, July 2018 (doi: 10.1109/TIFS.2017.2779441).
[10] A. Ferreira, L. Bondi, L. Baroffio, P. Bestagini, J. Huang, J.A. Santos, S. Tubaro, A. Rocha, "Data-driven feature characterization techniques for laser printer attribution", IEEE Trans. on Information Forensics and Security, vol. 12, no. 8, pp. 1860–1873, Aug. 2017 (doi: 10.1109/TIFS.2017.2692722).
[11] L.C. Navarro, A.K.W. Navarro, A. Rocha, R. Dahab, "Connecting the dots: Toward accountable machine-learning printer attribution methods", Journal of Visual Communication and Image Representation, vol. 53, pp. 257–272, May 2018 (doi: 10.1016/j.jvcir.2018.04.002).
[12] A. Ferreira, L.C. Navarro, G. Pinheiro, J.A. Santos, A. Rocha, "Laser printer attribution: Exploring new features and beyond", Forensic Science International, vol. 247, pp. 105–125, Feb. 2015 (doi: 10.1016/j.fo-rsciint.2014.11.030).
[13] M.J. Tsai, I. Yuadi, Y.H. Tao, "Decision-theoretic model to identify printed sources", Multimedia Tools and Applications, vol. 77, no. 20, pp. 27543–27587, Oct. 2018 (doi: 10.1007/s11042-018-5938-0).
[14] J. Hao, X. Kong, S. Shang, "Printer identification using page geometric distortion on text lines", Proceeding of the IEEE/ChinaSIP, pp. 856–860, Chengdu, China, July 2015 (doi: 10.1109/ChinaSIP.2015.7230526).
[15] P.J. Chiang, N. Khanna, A.K. Mikkilineni, M.V.O. Segovia, J.P. Allebach, G.T.C. Chiu, E.J. Delp, "Printer and scanner forensics: Models and methods", Intelligent Multimedia Analysis for Security Applications, vol. 282, pp. 145–187, March 2010 (doi: 10.1007/978-3-642-11756-5_7).
[16] S. Escher, T. Strafe, "Robustness analysis of a passive printer identification scheme for halftone images", Proceeding of the IEEE/ICIP, pp. 4357–4361, Beijing, China, Sept. 2017 (doi: 10.1109/ICIP.2017.829710-5).
[17] P. Kenny, G. Boulianne, P. Ouellet, P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition", IEEE Trans. on Audio, Speech and Language Processing, vol. 15, no. 4, pp. 1435–1447, May 2007 (doi: 10.1109/TASL.2006.881693).
[18] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, P. Dumouchel, "A study of interspeaker variability in speaker verification", IEEE Trans. on Audio, Speech and Language Processing, vol. 16, no. 5, pp. 980–988, July 2008 (doi: 10.1109/TASL.2008.925147).
[19] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, P. Ouellet, "Front-end factor analysis for speaker verification", IEEE Trans. on Audio, Speech and Language Processing, vol. 19, no. 4, pp. 788–798, May 2011 (doi: 10.1109/TASL.2010.2064307).
[20] P. Verma, P.K. Das, "I-Vectors in speech processing applications: a survey", International Journal of Speech Technology, vol. 18, no. 4, pp. 529–546, Dec. 2015 (doi: 10.1007/s10772-015-9295-3).
[21] Y. Xing, P. Tan, C. Zhang, "Improved i-vector speaker verification based on WCCN and ZT-norm", Chinese Conference on Biometric Recognition, pp. 424-431, Springer, Cham, Oct. 2016 (doi: 10.1007/978-3-319-46654-5_47).
[22] T. Ojala, M. Pietikäinen, D. Harwood, "A comparative study of texture measures with classification based on featured distributions", Pattern Recognition, vol. 29, no. 1, pp. 51–59, Jan. 1996 (doi: 10.1016/0031-3203(95)00067-4).

زیرنویس‌ها

1. Image forensics
2. Printer forensics
3. Raster image processor
4. Identity vector
5. Optical character recognition
6. Gaussian mixture model- universal background model
7. Support vector machine
8. Convolution texture gradient filter-map
9. Local tetra patterns
10. Joint factor analysis
11. Linear discriminant analysis
12. Within-class covariance normalization
13. Probabilistic linear discriminant analysis
14. Radial basis function