
بسم الله الرحمن الرحيم

دانشکده برق و رباتیک

گروه الکترونیک

پایان نامه کارشناسی ارشد مهندسی برق-الکترونیک

ناحیه بندی تصاویر اسناد پیچیده فارسی به بلوکهای متن، شکل و جدول

دانشجو :

مصطفی گلزاده حمزکانلو

استاد راهنما :

دکتر حسین خسروی

زمستان ۱۳۹۳

دانشگاه شاهرود

دانشکده : برق و رباتیک

گروه : الکترونیک

پایان نامه کارشناسی ارشد آقای مصطفی گلزاده

تحت عنوان: ناحیه بندی تصاویر اسناد پیچیده فارسی به بلوکهای متن، شکل و جدول

در تاریخ ..... توسط کمیته تخصصی زیر جهت اخذ مدرک کارشناسی ارشد مورد ارزیابی و با درجه ..... مورد پذیرش قرار گرفت.

اساتید راهنما	امضاء	اساتید مشاور	امضاء
نام و نام خانوادگی :		نام و نام خانوادگی :	
نام و نام خانوادگی :		نام و نام خانوادگی :	

اساتید داور	امضاء	نماینده تحصیلات تکمیلی	امضاء
نام و نام خانوادگی :		نام و نام خانوادگی :	
نام و نام خانوادگی :			
نام و نام خانوادگی :			
نام و نام خانوادگی :			

تقدیم

به پدر و مادر مهربانم

تشکر و قدردانی

سپاس خدای را که سخنوران، در ستودن او بمانند و شمارندگان، شمردن نعمت های او ندانند و کوشندگان، حق او را گزاردن نتوانند. و سلام و دورد بر محمّد و خاندان پاک او، طاهران معصوم، هم آنان که وجودمان وامدار وجودشان است؛ و نفرین پیوسته بر دشمنان ایشان تا روز رستاخیز...

بدون شک جایگاه و منزلت معلم، اجّل از آن است که در مقام قدردانی از زحمات بی شائبه‌ی او، با زبان قاصر و دست ناتوان، چیزی بنگاریم.

اما از آنجایی که تجلیل از معلم، سپاس از انسانی است که هدف و غایت آفرینش را تامین می کند و سلامت امانت‌هایی را که به دستش سپرده اند، تضمین؛ بر حسب وظیفه و از باب "من لم یشکر المنعم من المخلوقین لم یشکر اللّه عزّ و جلّ" : ازپدر و مادر عزیزم، این دو معلم بزرگوارم، که همواره بر کوتاهی و درشتی من، قلم عفو کشیده و کریمانه از کنار غفلت‌هایم گذشته‌اند و در تمام عرصه های زندگی یار و یاوری بی چشم داشت برای من بوده اند؛ از استاد با کمالات و شایسته؛ جناب آقای دکتر خسروی که در کمال سعه صدر، با حسن خلق و فروتنی، از هیچ کمکی در این عرصه بر من دریغ ننمودند و زحمت راهنمایی این پایان‌نامه را بر عهده گرفتند، کمال تشکر و قدردانی را دارم.

تعهد نامه

اینجانب **مصطفی گلزاده** دانشجوی دوره کارشناسی ارشد رشته **مهندسی برق الکترونیک گرایش دیجیتال** دانشکده برق و رباتیک دانشگاه شاهرود نویسنده پایان‌نامه **ناحیه بندی تصاویر اسناد پیچیده فارسی به بلوکهای متن، شکل و جدول** تحت راهنمائی **دکتر حسین خسروی** متعهد می شوم .

- تحقیقات در این پایان نامه توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است .
- در استفاده از نتایج پژوهشهای محققان دیگر به مرجع مورد استفاده استناد شده است .
- مطالب مندرج در پایان نامه تاکنون توسط خود یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ جا ارائه نشده است .
- کلیه حقوق این اثر به دانشگاه شاهرود می‌باشد و مقالات مستخرج با نام « دانشگاه شاهرود » و یا « Shahrood University» به چاپ خواهد رسید.

- حقوق معنوی تمام افرادی که در به دست آمدن نتایج اصلی پایان نامه تأثیرگذار بوده اند در مقالات مستخرج از پایان نامه رعایت می گردد.
- در کلیه مراحل انجام این پایان نامه ، در مواردی که از موجود زنده (یا بافتهای آنها) استفاده شده است ضوابط و اصول اخلاقی رعایت شده است.
- در کلیه مراحل انجام این پایان نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته یا استفاده شده است اصل رازداری، ضوابط و اصول اخلاق انسانی رعایت شده است.

تاریخ

امضای دانشجو

مالکیت نتایج و حق نشر

- کلیه حقوق معنوی این اثر و محصولات آن (مقالات مستخرج ، کتاب ، برنامه های رایانه ای ، نرم افزار ها و تجهیزاتی که ساخته شده است) متعلق به دانشگاه شاهرود می باشد . این مطلب باید به نحو مقتضی در تولیدات علمی مربوطه ذکر شود .
- استفاده از اطلاعات و نتایج موجود در پایان نامه بدون ذکر مرجع مجاز نمی باشد.

چکیده

سامانه‌های نویسه خوان نوری، OCR، نقش بزرگی در تحقق دولت الکترونیک و کاهش حجم بایگانیهای کاغذی و دیجیتال دارند. این سامانه ها از سه بخش اصلی پیش پردازش، شناسایی متن و پس پردازش تشکیل شده اند. طبیعی است که هر خطایی در مرحله پیش پردازش، بازگشت ناپذیر است، مثلا اگر زاویه چرخش سند اشتباه شناسایی شود، سبب خواهد شد که خطوط متن کج بشوند و فرایند شناسایی متن، به درستی صورت نگیرد. یکی از قسمتهای مهم در پیش پردازش، تحلیل پیکربندی اسناد است؛ به این معنا که مشخص کنیم کدام بخشها از تصویر سند، متن است، کدام بخشها جدول اند و چه نواحی ای شکل هستند. هر خطایی در این بخش، سبب تولید خطاهای بیشتر در فرایند OCR خواهد شد. در این پایان نامه به تحلیل ساختار اسناد فارسی چند ستونه می‌پردازیم. در زمینه تحلیل اسناد، سه رویکرد، متداول است، رویکرد پایین به بالا که از پیکسلها شروع می کند و با ادغام و رشد پیکسلها، به نواحی بزرگتر می رسد. رویکرد بالا به پایین مثل روش برش XY که ابتدا تصویر را با برشهایی به چند ناحیه تقسیم می کند و سپس با تکنیکهایی هر ناحیه، را به نواحی کوچکتر تجزیه می کند. ترکیب این دو روش هم با عنوان رویکرد ترکیبی شناخته می شود. ما یک رویکرد تقریبا ترکیبی که بیشتر مبتنی بر روش پایین به بالاست ارائه می دهیم. در این رویکرد از تکنیکهای آستانه گذاری وفقی، برچسب زنی مولفه ها، عملیات ریخت شناسی و تبدیل هاف استفاده شده و با یک الگوریتم مکاشفه ای و معرفی قوانین خاصی برای ترکیب نواحی کوچک بدون ادغام نواحی غیریکسان، سند را به ناحیه های متنی، جدول و شکل تقسیم می کنیم. روش معرفی شده روی اسناد متعدد چند ستونه و اسنادی که زمینه‌ی گرافیکی یا هنری دارند، آزمایش شده و عملکرد خوبی در مقایسه با نرم افزارهای پیشرو در حوزه OCR مثل OmniPage و FineReader ارائه می‌دهد. که نتایج به‌لحاظ‌عددی‌بدین‌شرح است که الگوریتم ما متن های فارسی را با ۷۲ ، شکل ها را با ۷۵ و جدول‌ها را۹۲ درصد درست تشخیص می دهد. و ۸۸ درصد اسناد فارسی را تقریبا درست ناحیه بندی می‌کند.

واژه‌های کلیدی: ناحیه بندی اسناد ، تحلیل ساختار اسناد ، قطعه بندی اسناد ، مستطیل محیطی، مؤلفه های پیوسته

فهرست مطالب

فصل اول: مقدمه..... ۱

۱-۱ مقدمه..... ۲

۱-۲ تعریف مسأله..... ۳

۱-۳ هدف پایان نامه..... ۶

۱-۴ ساختار پایان نامه..... ۷

فصل دوم: مروری بر کارهای انجام شده..... ۹

۲-۱ رویکرد پایین به بالا مبتنی برBAG..... ۱۰

۲-۱-۱ پیش پردازش..... ۱۰

۲-۱-۲تخمین جهت و تصحیح آن..... ۱۴

۲-۱-۳تشخیص نواحی..... ۱۶

۲-۲ تحلیل ساختار اسناد بر مبنای نواحی جداساز..... ۱۹

۲-۲-۱ آشکارسازی نواحی جداساز..... ۲۰

۲-۲-۲ آغشته سازی مشروط و یافتن ساختار سند..... ۲۲

۲-۳ برخی دیگر از کارهای انجام شده..... ۲۴

فصل سوم: مباحث تئوری..... ۳۱

۳-۱ دو سطحی سازی..... ۳۲

۳-۱-۱ دو سطحی سازی به روش سراسری..... ۳۲

۳-۱-۲ دو سطحی سازی به روش محلی..... ۳۳

۳-۲ روش های مبتنی بر تبدیل هاف.....۳۵

۳-۳ ریخت شناسی.....۳۹

۳-۳-۱عملگرهای مجموعه ای پایه.....۴۰

۳-۳-۲ عملگر گسترش.....۴۲

۳-۳-۳ عملگرفرسایش.....۴۳

فصل چهارم: روش پیشنهادی.....۴۵

۴-۱ مقدمه.....۴۶

۴-۲ چالش های تحلیل ساختار فیزیکی ۴۷

۴-۳ پیش پردازش و قطعه بندی.....۴۸

۴-۳-۱ دو سطحی سازی.....۴۸

۴-۳-۲ استخراج خطوط افقی و عمودی جهت تشخیص جدول.....۵۱

۴-۳-۳ به دست آوردن مؤلفه های پیوسته ۵۴

۴-۳-۴ حذف نویز ، شناسایی جدول های بزرگ ۵۶

۴-۴ ادغام اولیه ی مؤلفه ها در جهت افقی.....۵۷

۴-۴-۱ ادغام مستطیل های متداخل.....۵۷

۴-۴-۲ ادغام مستطیل های غیر متداخل.....۵۸

۴-۵ عملیات بر چسب زنی.....۶۰

۴-۵-۱ شناسایی جدول.....۶۱

۴-۵-۲ شناسایی متن.....۶۳

۴-۵-۳ شناسایی شکل..... ۶۴
۴-۶ ادغام نهایی بلوک های هم نوع در جهت افقی..... ۶۵
۴-۷ ادغام بلوک ها در جهت عمودی..... ۶۶
۴-۷-۱ ادغام در جهت عمودی: گام۱..... ۶۷
۴-۷-۲ ادغام در جهت عمودی: گام۲..... ۶۹
۴-۷-۳ ادغام در جهت عمودی: گام۳..... ۷۱
۴-۸ نتایج تجربی..... ۷۲
فصل پنجم: نتیجه‌گیری و پیشنهادات..... ۷۷
۵-۱ جمع بندی و نتیجه‌گیری..... ۷۸
۵-۲ پیشنهادات..... ۷۹
مراجع..... ۸۱

فهرست شکل‌ها

شکل۱-۱: نمایش مفهومی چرخه اسناد.....۲

شکل۱-۲:از سمت چپ به ترتیب عبارتند:سند اولیه،مؤلفه های پیوسته ، ادغام افقی،ادغام عمودی و تحلیل نهایی.....۵

شکل۱-۳:یک نمونه تصویر سند.....۷

شکل۱-۴: مراحل انجام تحلیل سند در این پایان‌نامه.....۷

شکل۲-۱:فلوچارت الگوریتم پیشنهادی[۱] ناحیه بندی اسناد.....۱۱

شکل۲-۲:نمایش BAG در تصویر،به ترتیب از چپ: تصویر خاکستری، تصویر دو سطحی، BAG(هر بلوک مستطیلBAG است).....۱۱

شکل۲-۳:الگوریتم تولیدBAG.....۱۲

شکل۲-۴:تعریف هندسی مربوط به اشیاء.....۱۴

شکل۲-۵:مختصات مستطیل محیطی مؤلفه های پیوسته با تصحیح جهت.....۱۶

شکل۲-۶:نتیجه خروجی الگوریتم [۱](متن:قرمز، جدول:آبی شکل:زرد).....۱۹

شکل۲-۷: نواحی جداساز در یک سند پیچیده.....۲۰

شکل۲-۸: تصویر سند پس از دو سطحی سازی و یافتن حاشیه ها.....۲۲

شکل۲-۹: نتیجه ی اعمال ادغام سازی افقی.....۲۳

شکل ۲-۱۰: شرایط نمونه ای که دو ناحیه ی مجاور ادغام نخواهند شد.....۲۴

شکل ۲-۱۱: نتیجه ی نهایی الگوریتم[۲]، تحلیل ساختار پیشنهادی.....۲۵

شکل۳-۱:سلول‌های‌فرضی (ρ_0,θ_0) و (ρ_1,θ_1)در فضای هاف هر قله به واحدهای به دقت صاف شده مرتبط است.....۳۷

شکل۳-۲:خطوط استخراج شده از یک دست نوشته..... ۳۸

شکل۳-۳: اعمال عملگر تفاضل دو تصویر..... ۴۱

شکل۳-۴: پس از گسترش(سمت راست)-تصویر اصلی(سمت چپ)..... ۴۳

شکل۳-۵: پس از فرسایش(سمت راست)-تصویر اصلی(سمت چپ)..... ۴۴

شکل ۴-۱ : مراحل انجام تشخیص و آشکار سازی بلوک های متن، شکل و جدول در اسناد..... ۴۷

شکل ۴-۲ : تعدادی از اسناد استفاده شده در این پایان نامه..... ۴۹

شکل۴-۳: نمونه از تصویر دو سطحی ایجاد شده(چپ:روش آتسو، راست: روش وفقی استفاده شده)..... ۵۱

شکل ۴-۴: (الف)تصویر دو سطحی اولیه، (ب)تصویر IM_H ترمیم خطوط افقی، (ج)تصویر IM_V ترمیم خطوط عمودی..... ۵۲

شکل ۴-۵: نمونه ای از آشکار سازی خطوط افقی و عمودی در اسناد(خطوط جهت دید بهتر ضخیم شده اند)..... ۵۴

شکل ۴-۶ : مستطیل های محیطی..... ۵۵

شکل ۴-۷:نمونه ای از ترسیم مستطیل های محیطی..... ۵۶

شکل۴-۸ :تشخیص جدول های بزرگ در سند..... ۵۷

شکل۴-۹ :حالت هایی از تداخل مستطیل ها..... ۵۹

شکل۴-۱۰ :ادغام افقی مستطیل ها..... ۶۰

شکل۴-۱۱ :پیدا کردن خطی منطبق بر یکی از اضلاع جدول..... ۶۲

شکل۴-۱۲:پیدا کردن ۲خط عمودی یا افقی از خط های جدول(جهت دید بهتر خطوط ضخیم شده اند)... ۶۲

شکل ۴-۱۳ :تشخیص اشتباه خط های شکل به عنوان خط های جدول(جهت دید بهتر خطوط ضخیم شده اند)..... ۶۳

شکل۴-۱۴ :نمونه ای از تشخیص بلوک های متنی.....۶۴

شکل۴-۱۵:تشخیص اشتباه متن های گرافیکی به عنوان شکل(موارد اشتباه شده با بیضی مشخص شده).....۶۵

شکل۴-۱۶ :ادغام اولیه ی افقی(سمت چپ)- ادغام نهایی افقی(سمت راست).....۶۶

شکل۴-۱۷ محاسبه ی فاصله ی عمودی بین دو بلوک(Dy).....۶۷

شکل۴-۱۸ محاسبه ی فاصله ی افقی بین بین خط های متناظر عمودی (Dx1,Dx2).....۶۸

شکل۴-۱۹ نمونه هایی از عدم ادغام عمودی نواحی(A با C وA با B وA با D)به دلیل بیش از حد بزرگ شدن مساحت ناحیه‌ی جدید.....۶۹

شکل۴-۲۰ خروجی گام۱ در ادغام عمودی.....۷۰

شکل۴-۲۱ خروجی گام۲ در ادغام عمودی.....۷۱

شکل۴-۲۲ خروجی گام۳ در ادغام عمودی.....۷۲

شکل۴-۲۳ نتایج الگوریتمهای تحلیل ساختار روی اسناد آزمایش .از چپ به راست:OmniPage18 و FineReader12و روش پیشنهادی ما.....۷۵

فهرست جداول

جدول۲-۱ مرور مختصری بر روش های مبتنی بر تحلیل منطقی.....۲۸

جدول۲-۲ مرور مختصری بر روش های مبتنی بر تحلیل هندسی اسناد.....۲۹

جدول۳-۱:الگوریتم های آستانه گذاری وفقی محلی.....۳۵

جدول۴-۱: ارزیابی عملکرد: نواحی که به تفکیک به طور صحیح تشخیص داده شده اند.....۷۳

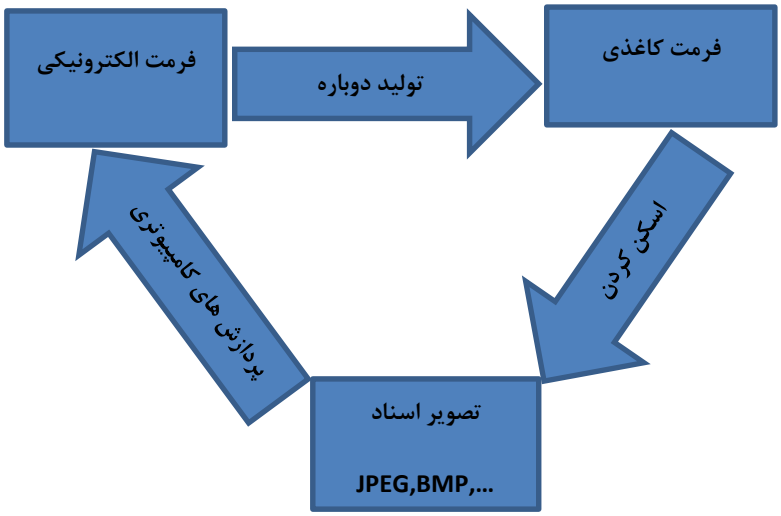
جدول۴-۲: ارزیابی عملکرد: اسنادی که کاملا بدون خطا ناحیه بندی شده اند.....۷۳

فصل اول:

مقدمه

۱-۱ مقدمه

سند[۱] **چیست** ؟منظور از سند،کلیه ی ارتباطات نوشتاری مانند گزارش های فنی،فایل های اداری،کتاب ها، روزنامه، مقالات و مجلات،نامه ها،چک بانکی و غیره است که حاوی اطلاعات می باشند. همواره اسناد به عنوان غالب ترین منبع در جوامع بشری بوده است که حاوی اطلاعات بوده و روشی را برای انتقال اطلاعات در زمان و مکان فراهم می کنند. اگر اسناد در گذشته بر پایه ی کاغذ بوده اند، اما امروزه اغلب به فرمت الکترونیکی می باشند. از مزایای اسناد الکترونیکی، آرشیو آسان و ماندگاری بالای آن ها است. ولی هنوز اسناد الکترونیکی برای مطالعه چاپ می شوند، علاوه بر آن در کتابخانه ها حجم زیادی از کتب قدیمی وجود دارد که می توان آن ها را به فرمت الکترونیکی تبدیل کرد و در کامپیوتر، دیسک فشرده نوری و دیسک سخت ذخیره نمود. چرخه ی زیر بیانگر این است که چگونه اسناد از فرمت الکترونیکی تبدیل به فرمت کاغذی و بالعکس می شوند[۱۴].



شکل۱-۱نمایش مفهومی چرخه اسناد

۱ Document

به موازات پیشرفت در تکنولوژی اطلاعات و نیاز روز افزون برای اطلاعات ، حجم اسناد شامل اطلاعات بیشتر و بیشتر شده است. با افزایش استفاده از اسناد الکترونیکی، حجم اسناد کاغذی هیچگاه کاهش پیدا نکرده است، زیرا که حجم انتشارات،روزنامه ها، مجلات،گزارش ها، کتاب ها و غیره ، به طور پیوسته رو به افزایش است و بیشتر افراد برای بایگانی کردن و مطالعه، اسناد کاغذی را ترجیح می دهند. ولی ذخیره و بازیابی این همه اسناد کاغذی که حجم آن ها به طور روز افزون زیاد می شود، مشکل و طاقت فرساست. در عوض اسناد الکترونیکی مزایای فراوانی دارند، مانند استفاده روز افزون از رایانه ها و اینترنت، ذخیره ی فشرده شده اسناد بدون اینکه به آنها آسیبی برسد ، بازیابی موثر، انتقال سریع و داشتن ساختار صریح. بنابراین تحقیقات گسترده ای روی تبدیل اسناد کاغذی به اسناد الکترونیکی صورت گرفته است. الگوریتم ها و تکنیک هایی که به تصاویر اسناد اعمال می شود تا یک توصیف قابل خواندن توسط رایانه ازاطلاعات پیکسل به دست آید، تحلیل اسناد نامیده می شود. هدف از تحلیل اسناد، بازشناسی متن، اجزای گرافیکی و جدول ها در تصویر اسناد و استخراج اطلاعات مورد نظر، آنگونه که انسان می خواهد، است. تحلیل ساختار اسناد مهم ترین گام در فرآیندهای تحلیل و بازیابی اسناد است و شامل جدا کردن متن،گرافیک ،جدول و سایر اجزای یک سند و سپس استخراج بلوک های متنی جدا مانند عناوین، سر صفحه ها، زیر نویس ها وشرح تصاویر می باشد[۱۵].

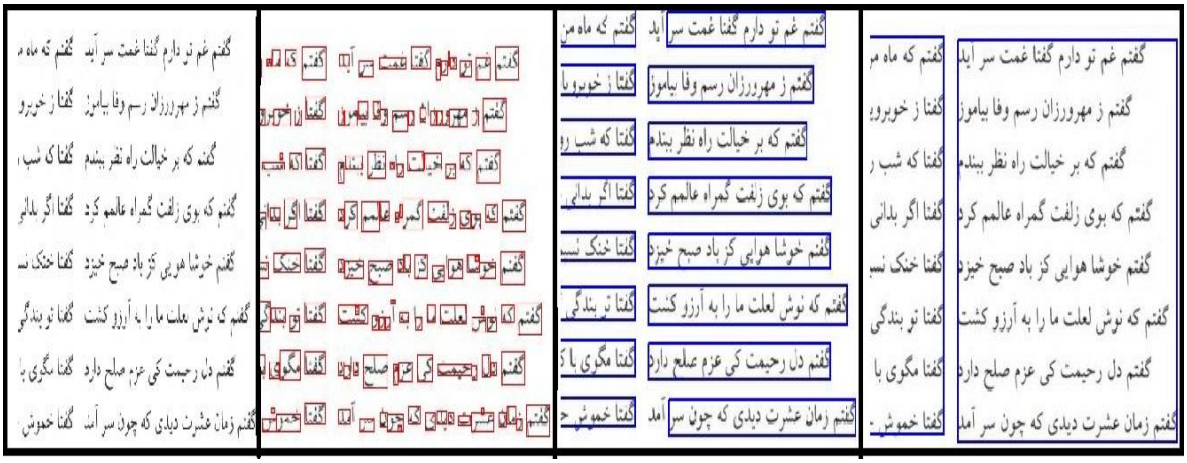
۱-۲ تعریف مسأله

داده ی ایده آل برای یک سیستم نویسه خوان[۱] ، یک تصویر متنی است که دارای تصویر و فرمول نباشد. متأسفانه در درصد بسیار کمی از تصاویر اسناد، داده ی اولیه ی ما ایده آل است. در نتیجه چالش های جدید برای نزدیک کردن داده ی اولیه ی نویسه خوان به داده ی ایده آل به وجود می آید. تحلیل ساختار اسناد متنی یک بخش مهم از سامانه های پردازش اسناد متنی و به ویژه نرم افزارهای نویسه خوان است.

۱ OCR

معمولا بعد از اعمال پیش پردازش روی سند و قبل از شناسایی متن آن ، ساختار سند باید تحلیل شده و نواحی مختلف متن، تصویر و جدول مشخص شود. الگوریتم ناحیه بندی یک مرحله مهم و مقدماتی در پردازش تصاویر مربوط به اسناد و تجزیه و تحلیل آن ها است. اهمیت این امر به این دلیل است که در عمل مقدم تر از دیگر فرآیندهای تشخیص یا طبقه بندی استفاده شده است. در صورت ناحیه بندی ضعیف تصاویر، سامانه طبقه بندی نویسه خوان ، با توجه به حضور عناصر غیر متنی نویسه های ناخواسته ای تولید می کند. محتوای تصاویر اسناد، جهت طبقه بندی به مجموعه ای دو کلاسه شامل متن و غیرمتن تقسیم شده است. کلاس غیر متن شامل اجزاء گرافیک، فرمول ریاضی، آرم ها، جداول و... است. سند ورودی ممکن است یک نوشته عادی یک ستونه و یا ترکیبی از تصاویر و نواحی متنی در یک ساختار پیچیده مثل روزنامه باشد. هدف الگوریتم های تحلیل ساختار این است که بلوک های مختلف متنی، تصویری و جدولی را آن گونه که انسان انتظار دارد شناسایی کنند. مثلا تمامی خطوط یک پاراگراف در یک ناحیه متنی قرار گیرند در حالی که متون موجود در ستون های مجاور یک سند چند ستونه باید جدا از هم باشند. روش های زیادی برای ناحیه بندی و تحلیل ساختار اسناد وجود دارد که بطور عمده می توان رویکرد آن ها را به سه گروه پائین به بالا ، بالا به پائین و ترکیبی طبقه بندی کرد. تکنیک های پایین به بالا که به صورت گسترده ای مورد استفاده قرار گرفته است ، از اجزای ریز تصویر مثل پیکسل ها شروع کرده و با به هم چسباندن آن ها بصورت متوالی بر مبنای همسایگی نزدیک ، مولفه های پیوسته شامل حروف و کلمات را تشکیل داده و در مرحله ی بعد با ادغام این نواحی بر مبنای قواعد مکاشفه ای، نواحی بزرگتر مانند ستون های متنی را تشکیل می دهند(شکل۱-۲ را مشاهده کنید). در رویکرد بالا به پایین از بلوکهای بزرگ مانند ستونهای متنی شروع کرده و شکستن متوالی بلوکها را تا رسیدن به اجزای ریز سند مانند کلمات و حروف ادامه می دهند. این روش ها نیازمند از پیش دانستن برخی مشخصات ساختار سند هستند. رویکرد ترکیبی ، ترکیبی از دو رویکرد گذشته اند و یا در هیچ یک از دو گروه فوق

نمی گنجند و به عنوان روشهای ترکیبی مشهورند. یکی از نواحی که در تصویر اسناد مشخص خواهیم کرد ناحیه ی متنی است و چون نحوه ی نگارش در فارسی و انگلیسی متفاوت است ، حروف فارسی به یکدیگر اتصال دارند و مؤلفه های پیوسته و به تبع آن مستطیل های محیطی بزرگتری نسبت به حروف انگلیسی دارند. که همین مسأله باعث ایجاد تداخل در بلوک ها می شود که باید سعی در حذف و یا ادغام این بلوک ها داشته باشیم. در متن فارسی همچنین تعداد نقطه های آن نیز زیاد است که باید مراقب باشیم آنها را نویز در نظر نگیریم و حذف نکنیم بلکه آنها را با حرف خودش ادغام کنیم. در واقع نقطه ها به تنهایی نباید یک مؤلفه ی پیوسته باشند.



شکل۱-۲:از سمت چپ به ترتیب عبارتند:سند اولیه،مؤلفه های پیوسته ،ادغام افقی،ادغام عمودی و تحلیل نهایی

روش های ناحیه بندی تصاویر اسناد را می توان به گروه های زیر طبقه بندی کرد:

۱) طبقه بندی بر مبنای پیکسل

۲) طبقه بندی بر مبنای مؤلفه های به هم پیوسته

۳) طبقه بندی بر مبنای بلوک و یا منطقه

۴) ناحیه بندی مبتنی بر ریخت شناسی چند درجه تفکیکی

۵) ناحیه بندی مبتنی بر مدل سازی تصاویر اسناد توسط نمودار (اطلاعات مکانی)

۱-۳ هدف پایان نامه

در این پایان نامه قصد داریم در تصاویر اسناد سه مورد را به صورت بلوک هایی مشخص کنیم: ۱-متن ۲ تصویر ۳-جدول ، استخراج اطلاعات متنی از اسناد با قالب بندی گوناگون یا پیچیده، به دلیل جایگذاری اشکال، عناوین و زیرنویس شکل ها، زمینه پیچیده، قالب بندی متن هنرمندانه، تنوع در اندازه قلم و فاصله خطوط، کاری بسیار مشکل و فراتر از توانایی سیستمهای تحلیل ساختار امروزی است .زمانی که با اسناد فارسی سر و کار داشته باشیم این امر سخت تر نیز میشود. چرا که در زبان فارسی هر واژه از یک یا چند زیرکلمه تشکیل شده است و این زیر کلمات اندازه و شکلهای کاملا متفاوتی دارند. همچنین هر حرف می تواند دارای یک، دو یا سه نقطه در بالا یا پایین و یا اجزای ناپیوسته دیگری (مانند" آ " یا "گ") باشد. برای بهبود این سیستم ها، آن ها باید بتوانند با ساختارهای سند پیچیده تر کار کنند و هم چنین انعطاف پذیری، سرعت و دقت بیشتری داشته باشند. الگوریتم پیشنهادی حاصل استفاده از چند تجربه یا تکنیک ابداعی است. در یک نگاه کلی میتوان الگوریتم را به پنج مرحله کلی تقسیم کرد: ۱- پیش پردازش، دو سطحی سازی و استخراج خطوط با استفاده از تبدیل هاف ۲-استخراج مؤلفه های پیوسته، حذف نویز و ترسیم مستطیل های محیطی ۳-حذف نویز، ادغام مستطیل ها در جهت افقی ۴- برچسب زنی بلوک های حاصل شده از ادغام و ادغام نهایی بلوک های هم نوع در جهت افقی ۵-ادغام بلوک های هم نوع در جهت عمودی طی۳گام. یک نمونه از اسناد استفاده شده در این پایان نامه در شکل(۱-۳) آمده است. همچنین کار انجام شده در این پایان نامه به طور خلاصه در شکل (۱-۴) نشان داده شده است.

شکل۱-۳:یک نمونه تصویر سند

شکل۱-۴: مراحل انجام تحلیل سند در این پایان‌نامه

۱-۴ ساختار پایان نامه

ساختار پایان نامه به این صورت است که فصل دوم ، مرور مختصری بر تحقیقات انجام شده در زمینه‌ی تحلیل تصویر اسناد را در بر دارد. در فصل سوم به شرح مختصر برخی از مباحث تئوری استفاده شده در

الگوریتم پیشنهادی می پردازیم. به عنوان مثال روش های دو سطحی سازی را بررسی می کنیم. همچنین در این فصل به روش های مبتنی بر تبدیل هاف جهت استخراج خط می پردازیم. در فصل چهارم، الگوریتم پیشنهادی ارائه می شود، می توان این فصل را اصلی ترین فصل متن حاضر دانست. در این فصل به بررسی جزئیات الگوریتم پیشنهادی پرداخته شده است. در پایان این فصل عملکرد الگوریتم پیشنهادی را با دو نرم افزار کاربردی مقایسه می کنیم. در فصل پنجم به جمع بندی ، نتیجه گیری و ارایه پیشنهادات پرداخته شده است.

فصل دوم:

مروری بر کارهای انجام شده

روش های زیادی برای ناحیه بندی صفحه و تحلیل ساختار وجود دارند که بطور عمده می توان آ نها را به سه گروه پائین به بالا، بالا به پائین و ترکیبی طبقه بندی کرد[۲]. تکنیک های پائین به بالا که به صورت گسترده ای مورد استفاده قرار گرفته است، از پیکسل ها شروع کرده و بر مبنای نوع همسایگی با به هم چسباندن آن ها ، مؤلفه های پیوسته را تشکیل داده و با قواعدی این نواحی ادغام می شوند. در این فصل در قسمت ۲-۱ نمونه ای از روش پائین به بالا و در قسمت ۲-۲ نیز نمونه ای از روش ترکیبی را توضیح می دهیم. و در بخش ۲-۳ برخی دیگر از کارهای انجام شده در زمینه ناحیه بندی و تحلیل تصاویر و ویژگی های هر کار را به طور مختصر بیان می کنیم.

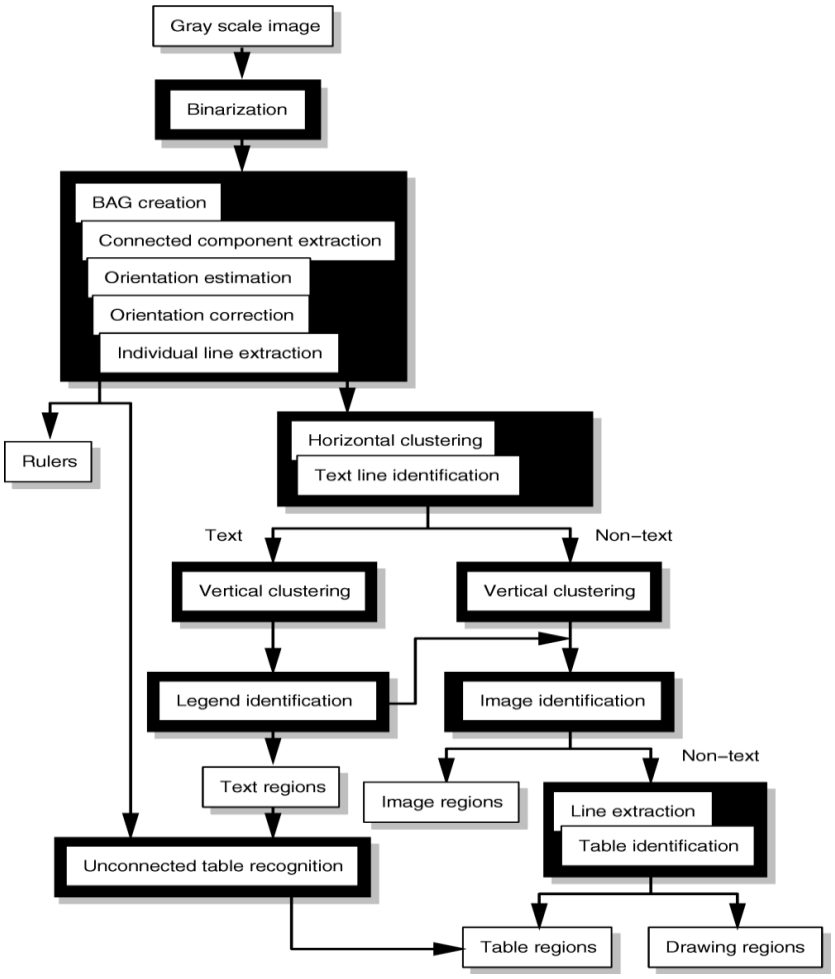
۲-۱ رویکرد پایین به بالا مبتنی برBAG

رویکرد سنتی پایین به بالا در [۱] از BAG[1] ها استفاده کرده است. این الگوریتم دقت بالایی دارد، در مدت تقریبی ۱/۴ ثانیه روی یک سند برای ایجاد مدل شامل تخمین جهت و تصحیح آن تقسیم بندی و برچسب گذاری برای یک تصویر که با کیفیت۳۰۰dpi اسکن شده است، انجام می دهد.

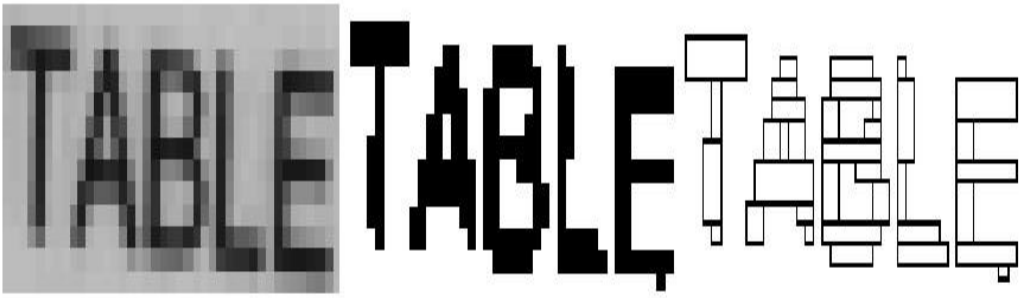
۲-۱-۱ پیش پردازش

در شکل۲-۱ فلوچارت مراحل عملیات تشخیص و ناحیه بندی اسناد ورودی را به صورت نمودار درختی مشاهده می کنیم. توسط الگوریتم پیشنهاد شده BAG ها در سند ورودی تولید شده است. در شکل ۲-۲ نیز یک BAG که در سند ورودی تولید شده است را می بینیم. الگوریتمی که در [۱] برای تولید BAG پیشنهاد شده است را در شکل۲-۳ می بینیم.

[1] Block Adjacency Graph



شکل۲-۱:فلوچارت الگوریتم[۱] ناحیه بندی اسناد



شکل۲-۲:نمایش BAG در تصویر،به ترتیب از چپ: تصویر خاکستری، تصویر دو سطحی، BAG(هر بلوک مستطیلBAG است)

Each run length, r, in the first row of the input image is regarded as a new block node n_i. For a new node, its geometric measure is the coordinates of its upper left and lower right vertices $((X_u, Y_u), (X_l, Y_l))$ initialized as $X_u = x_b(r), Y_u = y(r)$ and $X_l = x_e(r) + 1, Y_u = y(r) + 1$, where $x_b(r), x_e(r)$ and $y(r)$ denote the beginning and ending column numbers and row number of run length r.

For the successive rows in the image {

  For each run length r_c in the current row {

    If r_c is 8-connected to a run length in the preceding row {

      If r_c is 8-connected to only one run length r_l and r_l connects to only r_c below it and $|x_b(r_c) - X_u(n_i)| \le T_a$ and $|x_e(r_c) - X_l(n_i) + 1| \le T_a$ then r_c is merged into the block node n_i, where n_i is the block node involving r_l. $X_u = Average_{r_k \in n_i}\{x_b(r_k)\}, X_l = Average_{r_k \in n_i}\{x_e(r_k)\}$ and $Y_u = Y_u + 1$.

      Else, r_c is regarded as a new block node n_{i+1}, initialized with edges $e(n_{i+1}, n_j)$ to those block nodes $\{n_j\}$ in which there is a run length 8-connected to r_c.

    }

    Else, r_c is regarded as a new block node n_{i+1}.

  }

}

شکل۲-۳:الگوریتم تولیدBAG

بعد از ایجاد BAG ها مؤلفه‌های پیوسته را که شامل BAG های به هم چسبیده هستند تشکیل می دهیم

$C_i = \{n_j\}$ که باید شرایط زیر را داشته باشند(β_i ها همان BAG ها هستند):

۱- $C_i \subset \beta_i$

۲- به ازای هر n_j و n_k که متعلق است به C_i باید مسیر زیر وجود داشته باشد:

$$\left(n_{j1}, n_{j2}, n_{j3}, \dots n_k\right)$$

پس مؤلفه های پیوسته مبتنی بر BAG ها استخراج می شود. درجه ی محاسبات پیچیده در تشخیص مؤلفه های پیوسته از روش های ترتیبی کمتر است. مؤلفه های پیوسته که به یکدیگر نزدیک هستند را با همدیگر ادغام می شود و GTL[1] ها تشکیل می شود.

البته از مؤلفه های پیوسته خیلی کوچک و مؤلفه هایی که نزدیک به مرزهای تصویر و مشکوک به نویز هستند صرف نظر و حذف می شوند. اگر تمام مؤلفه های پیوسته به دست آمده، بطور کلی به عنوان اشیاء O_i در نظر گرفته شوند و پس از اینکه مختصات دو نقطه ی بالا چپ $(X_U(O_i),Y_U(O_i))$و پایین راست $(X_l(O_i),Y_l(O_i))$ را در مستطیل محیطی به دست آمد ، فاصله‌ی افقی و عمودی بین مؤلفه های پیوسته جهت ادغام یا عدم ادغام با استفاده از روابط زیر به دست می آید (در شکل ۲-۴ نیز نشان داده شده است).

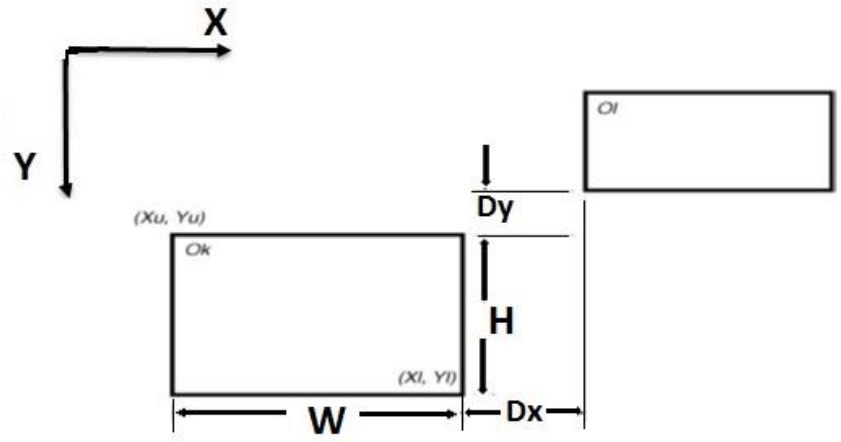
فاصله ی افقی

$$D_X(O_i,O_j) = \max[X_U(O_i), X_U(O_j)] - \min[X_l(O_i), X_l(O_j)] \qquad (۱-۲)$$

فاصله ی عمودی

$$D_y(O_i,O_j) = \max[Y_U(O_i), Y_U(O_j)] - \min[Y_l(O_i), Y_l(O_j)] \qquad (۲-۲)$$

۱ generalized text line



شکل۲-۴:تعریف هندسی مربوط به اشیاء

$$V_X(O_i,O_j)=\frac{-D_X(O_i,O_j)}{\min[W(O_i),W(O_j)]} \quad \text{و} \quad V_y(O_i,O_j)=\frac{-D_y(O_i,O_j)}{\min[H(O_i),H(O_j)]} \qquad (۳-۲)$$

$$W=X_L-X_U \quad \text{و} \quad H=Y_L-Y_U \qquad (۴-۲)$$

مستطیل های محیطی که شرایط زیر را از نظر فاصله داشته باشند ، با هم ادغام می کنیم.

$$D_x(O_i,O_j)<T_d \ , \ V_y(O_i,O_j)>0.5 \quad \text{و} \quad D_y(O_i,O_j)<T_d \ , \ V_x(O_i,O_j)>0.5 \qquad (۵-۲)$$

مقدار T_d بطور تجربی بدست می آید.

۲-۱-۲تخمین جهت و تصحیح آن

در [۱] با اعمال یک تبدیل هاف به مؤلفه های پیوسته یک الگوریتم تشخیص انحراف اسناد طراحی شده است. در مرحله اول زاویه های بزرگتری که استفاده شده، آشکار شده است. و در مرحله ی دوم زاویه های کوچکتر برای تخمین های دقیق تر استفاده استفاده شده است. از این الگوریتم برای تخمین زوایای تورب($\propto$)یک تصویر سند ورودی با زوایای داخلی $(-5^{\circ},5^{\circ})$ با رزولوشن 0.1° استفاده شده است. اگر مختصات (X_U,Y_U)و عرضW و ارتفاعH را در مستطیل محیطی در نظر گرفته شود. مختصات (X'_U,Y'_U)و به دنبال آن مختصات (X''_U,Y''_U)که با جهت تصحیح شده به دست آمده و در ادامه می آید.

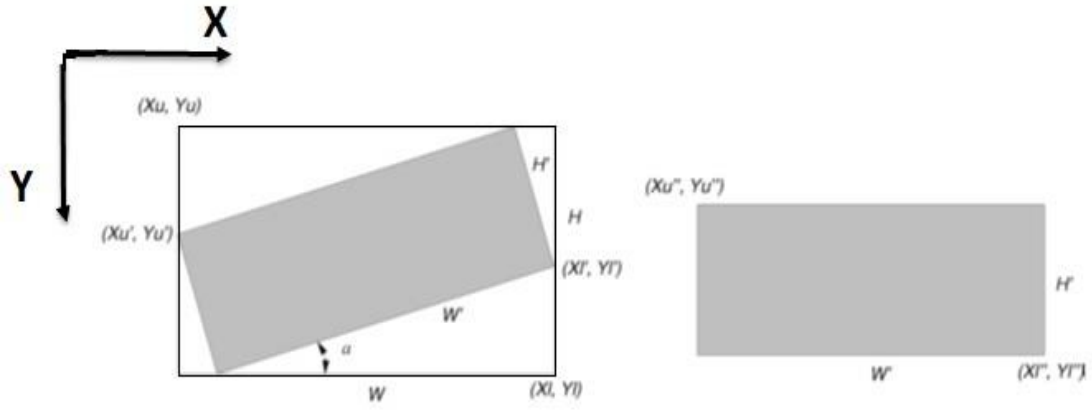
برای $\propto > 0$:

$$\begin{bmatrix} X'_u \\ Y'_u \\ W' \\ H' \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & \frac{\sin\alpha\cos\alpha}{\cos 2\alpha} & -\frac{\sin^2\alpha}{\cos 2\alpha} \\ 0 & 0 & \frac{\cos\alpha}{\cos 2\alpha} & -\frac{\sin\alpha}{\cos 2\alpha} \\ 0 & 0 & -\frac{\sin\alpha}{\cos 2\alpha} & \frac{\cos\alpha}{\cos 2\alpha} \end{bmatrix} \begin{bmatrix} X_u \\ Y_u \\ W \\ H \end{bmatrix} \qquad (۶-۲)$$

برای $\propto < 0$:

$$\begin{bmatrix} X'_u \\ Y'_u \\ W' \\ H' \end{bmatrix} = \begin{bmatrix} 1 & 0 & -\frac{\sin^2\alpha}{\cos 2\alpha} & -\frac{\sin\alpha\cos\alpha}{\cos 2\alpha} \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{\cos\alpha}{\cos 2\alpha} & \frac{\sin\alpha}{\cos 2\alpha} \\ 0 & 0 & \frac{\sin\alpha}{\cos 2\alpha} & \frac{\cos\alpha}{\cos 2\alpha} \end{bmatrix} \begin{bmatrix} X_u \\ Y_u \\ W \\ H \end{bmatrix} \qquad (۷-۲)$$

$$\begin{bmatrix} X_u^{''} \\ Y_u^{''} \end{bmatrix} = \begin{bmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{bmatrix} \begin{bmatrix} X_u^{'} \\ Y_u^{'} \end{bmatrix} \quad \text{و} \quad Y_l^{''} = (Y_u^{''} + H^{'}) \quad ، \quad X_l^{''} = (X_u^{''} + W^{'}) \qquad (۸-۲)$$



شکل۲-۵:مختصات مستطیل محیطی مؤلفه های پیوسته با تصحیح جهت

۲-۱-۳تشخیص نواحی

ویژگی های استخراج شده جهت تشخیص نواحی مورد نظر استفاده می شود.

-تشخیص متن:

یک GTL ، t_i برچسب متن زده می شود اگر شرایط زیر را داشته باشد:

۱- ارتفاع آن $H(t_i) < T_{th}$ باشد و مؤلفه های پیوسته در آن به طور افقی در یک ردیف قرار داشته باشند. به عنوان مثال، انحراف معیار لبه های پایین مؤلفه های پیوسته در آن کمتر از T_{La} است .

۲- عرض آن $W(t_i) > T_{th}$ و مؤلفه های پیوسته در آن تقریبا ارتفاع یکسانی داشته باشند. به عنوان مثال نسبت بین انحراف معیار و مقدار میانگین ارتفاع مؤلفه های پیوسته کمتر از ۰/۳ باشد. در [۱] از مؤلفه های پیوسته $H(c_j) < 0.3H(t_i)$ صرفنظر شده است.

$$T_{La} = 0.05^{''} \quad \text{و} \quad T_{th} = 0.3^{''}$$

اگر یک ناحیه ی متنی کوچک b_t باشد ، که توسط یک ناحیه ی غیر متنی b_n محاصره شده است و فاصله ی آنها با یکدیگر از یک مقدار آستانه ای کمتر باشد ، کل ناحیه تصویر در نظر گرفته شده است. و از ناحیه ی متنی صرفنظر می شود.

-تشخیص تصویر:

بر اساس مدل سند، می توان تعداد $A(I_i)$ پیکسل های سیاه در یک ناحیه ی تصویر I_i را محاسبه کرد. که برابر است با مجموع نواحی نودهای BAG ، که در ناحیه دخالت و نقش دارند.به عبارت دیگر:

$$A(I_i) = \sum_{t_j \in I_i} \sum_{C_k \in t_j} \sum_{n_l \in c_k} W(n_L).H(n_L) \qquad (۹-۲)$$

*یک ناحیه تصویری b_i ناحیه ی غیرمتنی نامیده می شود اگر:

۱) ناحیه ای بزرگ باشد $\min[W(b_i), H(b_i)] > T_{is}$

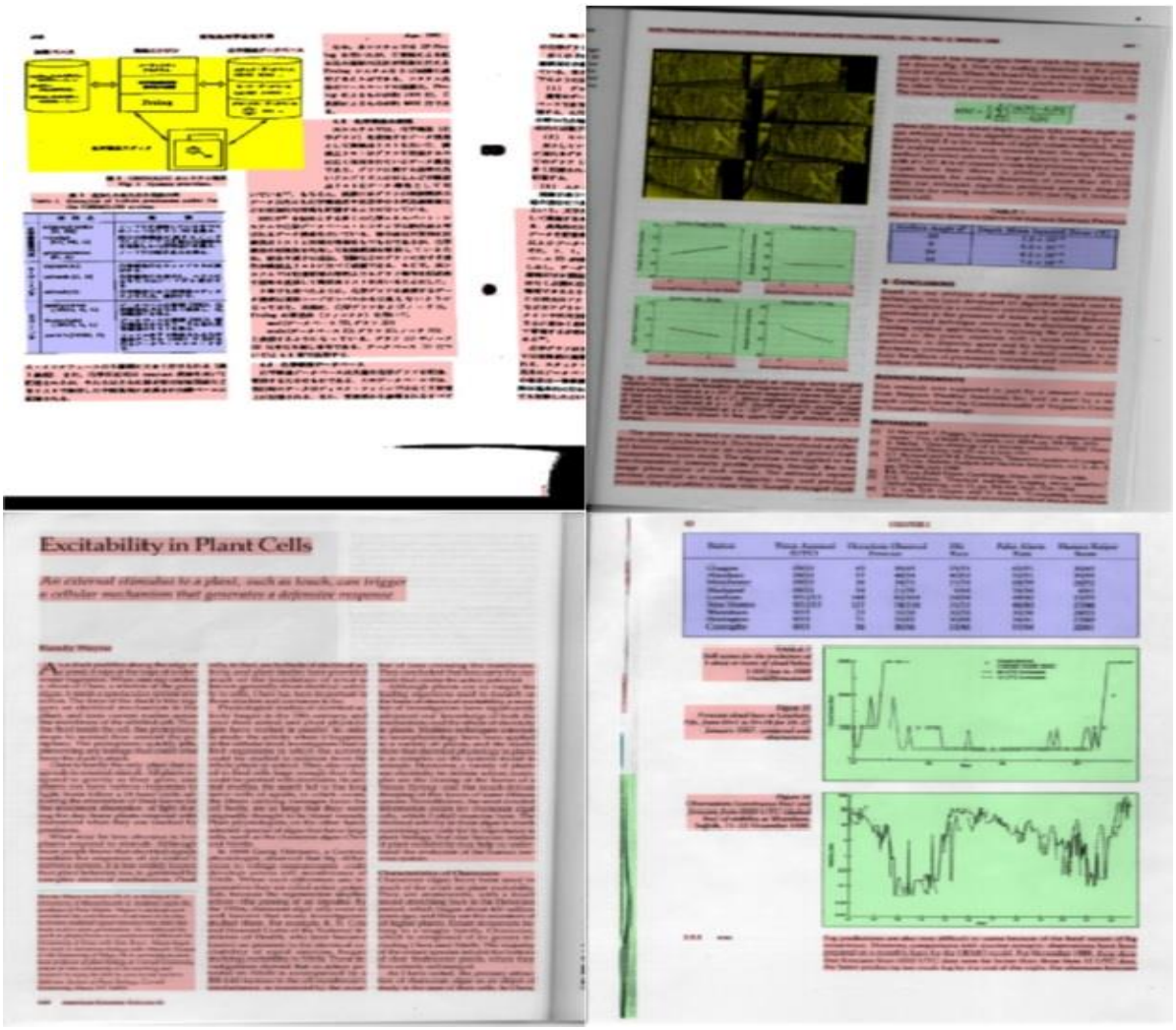
۲) نسبت تعداد پیکسلهای سیاه به تعداد کل ناحیه $\frac{A(I_i)}{W(I_i).H(I_i)}$ بزرگتر از T_{ia}(مقدار آستانه) باشد که $T_{ia} = 0.4$و $T_{is} = 0.5^{"}$.

-تشخیص جدول:

برای نواحی غیر متنی باقیمانده، ابتدا خطوط افقی و عمودی استخراج می شود. اگر آنها وجود داشته باشند خطوط بالا و پایین باید طول مشابه داشته باشند و تفاوت آنها کمتر از $T_{tLw} = 0.2^{"}$ باشد. میانگین و انحراف معیار جهت هر دو خط افقی و عمودی باید کمتر از یک درجه باشد. سرانجام میانگین ارتفاع و تغییرات مؤلفه های پیوسته در یک ناحیه جدول باید کوچک باشد(میانگین باید کمتر از T_{thm} و انحراف معیار باید کمتر از T_{ths} باشد، که در [۱] از $T_{thm} = 0.3^{"}$ و $T_{ths} = 0.07^{"}$ استفاده شده است). هر ناحیه‌ی غیر متنی باقیمانده به عنوان ناحیه ی اشکال ترسیمی در نظر گرفته می شود. نتیجه ی اعمال الگوریتم [۱] در اسناد ورودی را در شکل۲-۶ می بینیم. الگوریتم های متعددی بر مبنای تکنیک پایین به بالا به وجود آمده است مثلا (Mitchell and Yan 2000) از برچسب زنی مؤلفه ها و ادغام الگوهای مجاور

بر مبنای نزدیکی مکانی و طبقه بندی نوع الگو برای تحلیل ساختار استفاده کرده است. (whal,wong et al. 1982) و (Shi and Govindaraju 2005) از آغشته سازی[1] برای به هم چسباندن کلمات یک خط استفاده کرده اند، (Tan and Zhang 2001) از شبکه عصبی و (Liang ,Ha et al.1996) از ایده ی رشد ناحیه[2] بر مبنای مستطیل های محیطی مولفه های پیوسته تصویر استفاده کرده است. (Yuan and Tan 2005) از یک الگوریتم سلسله مراتبی برای ادغام نواحی استفاده کرده اند. آنها یک تابع تجربی را پیاده سازی کرده و با استفاده از آن میزان همسایگی عناصر مجاور را به دست آورده و عمل ادغام را انجام داده اند. (Gupta ,Niranjan et al.2006) از تحیل مستطیل های محیطی مولفه ها استفاده کرده و پس از ادغام مولفه ها به صورت افقی، به منظور یافتن خطوط متنی، یک مرحله ادغام عمودی برای یافتن پاراگرافها انجام داده است.

[1] Smearing
[2] Region Growing



شکل۲-۶:نتیجه خروجی الگوریتم [۱](متن:قرمز، جدول:آبی شکل:زرد)

۲-۲ تحلیل ساختار اسناد بر مبنای نواحی جداساز

با بررسی رفتار انسان در تشخیص نواحی مختلف یک سند در [۲]، متوجه شده اند که انسان معمولا مکان اجزای یک سند را بر اساس حاشیه های اطراف آن ها تشخیص می دهد. در [۲] الگوریتم بر مبنای آشکارسازی نواحی جداساز و آغشته سازی مقید استوار است. یعنی ما آغشته سازی ها و ادغام ها را فقط بین دو ناحیه‌ی جدا ساز انجام می‌دهیم. نواحی جداساز مستطیل های سفیدی هستند که بین نواحی

مجاور، مانند دو ستون متنی، قرار دارند. با استفاده از این ایده مشکل عمده روش های مبتنی بر ادغام را که همان یکی شدن نواحی مجاور است، برطرف شده است. در روش پیشنهاد شده در[۲]پس از شناسایی جداسازها، ابتدا آغشته سازی و پس از آن برچسب زنی مؤلفه ها روی تصویر سند اعمال می شود و به این ترتیب نواحی اولیه بدست می آیند. با در دست داشتن نواحی اولیه، چند مرحله ادغام مشروط نواحی انجام شده تا ساختار نهایی سند به دست آید. مشکل عمده روشهای مبتنی بر آغشته سازی این است که اگر دو بلوک مجاور در سند، نزدیک به هم باشند به گونه ای که فاصله‌ی آنها کمتر از آستانه‌ی آغشته سازی باشد، آن دو بلوک به صورت ناخواسته به هم خواهند چسبید.

شکل۲-۷: نواحی جداساز در یک سند پیچیده

هر چند تغییر آستانه‌ی آغشته سازی می تواند به صورت موردی مشکل را برطرف کند، اما مسلما راه حل خوبی برای این مشکل نیست.

۲-۲-۱ آشکارسازی نواحی جداساز

برای یافتن نواحی جداساز، ابتدا از تصویر سند یک نسخه سیاه و سفید تهیه و سپس نقاط حاشیه صفحه شناسایی می شود. این پیکسل ها نباید در فرآیند آشکارسازی نواحی جداساز، دخالت داشته باشند،چون که ممکن است سبب یافتن جداسازهای نادرست شوند.

الگوریتم به کار گرفته شده در[۲] را برای شناسایی این نقاط می‌توان به صورت زیر توصیف کرد:

۱. افکنش تجمعی عمودی مربوط به پیکسلهای زمینه را بیاب

$$vp(x,y)=vp(x,y-1)+f(x,y)$$
$$vp(0,0)=f(0,0)$$

در این رابطه $f(x,y)$ مقدار پیکسل در نقطه ی(x,y) است : ۱ برای پیکسل زمینه و . برای پیکسل تصویر. Vp نیز ماتریس افکنش تجمعی عمودی است .

۲. هر سطر را از چپ به راست جاروب کن و پیکسلهای حاشیه را به صورت شبه کد زیر پیدا کن.

در این شبه کد به ازای هر ردیف (H) تمام ستون ها (W)جاروب می شود و برای هر پیکسل این شرط ها بررسی می شوند: اگر پیکسل از یک آستانه ای جلوتر بود و مقدار روشنایی آن صفر بود یا پیکسل قبلی آن حاشیه نباشد ، آن پیکسل حاشیه است و از الگوریتم خارج می شود.

(border:یک‌ماتریس بولین به اندازه‌ی‌تصویراست که هر جا‌پیکسل‌حاشیه‌باشد‌مقدار‌آن‌true می شود.)

```
for y = 0 to H
   for x = 0 to W
      if(x ≥ Th && (f(x,y)=0 || !border(x-1,y)))
         break;
      elseif(vp(x,y) ≥ y+1-Th || vp(x,H-Th)≥H-y-Th)
         border(x,y) = true;
```

Th یک آستانه ی کوچک، مثلا ۵ است.

۳. گام ۲ را از سمت راست به چپ تکرار کن.

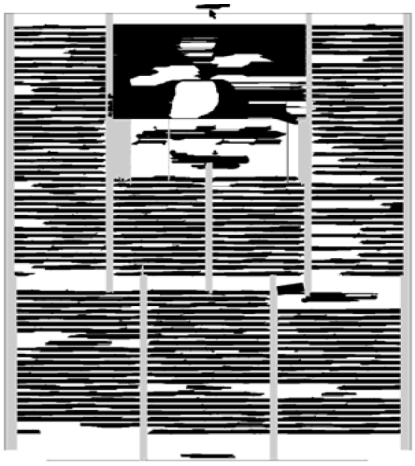
حال باید نواحی جداساز را پیدا کرد ، باید تنها نقاطی را یافت که متعلق به نواحی جداساز باشد و کمک کنند که فرایند ادغام سازی به نواحی اطراف این جداسازها مقید شوند. با این کار سرعت یافتن این نواحی به طور قابل توجهی افزایش خواهد یافت. در روش[۲] این شناسایی به سهولت انجام می شود. تمامی نقاطی که مقدار افکنش تجمعی زمینه در آنجا از یک حد آستانه مثلا ۱۰۰ پیکسل بیشتر باشد، به عنوان ناحیه جداساز برچسب می‌خورد. تنها پارامتر تقریبا مهم این الگوریتم ، همین حد آستانه است که بیشتر به درجه ی تفکیک وابسته است.

شکل۲-۸: تصویر سند پس از دو سطحی سازی و یافتن حاشیه ها

۲-۲-۲ آغشته سازی مشروط و یافتن ساختار سند

عمل ادغام سازی افقی با آستانه‌ی بزرگ ، به نحوی که مطمئن شویم کلمات یک سطر متن همه به هم خواهند چسبید، بر تصویر اعمال می شود. این آغشته سازی به محض رسیدن به نواحی که برچسب جداساز دارند، متوقف می شود. در [۲] ۲۵ پیکسل به عنوان حد آستانه انتخاب شده است. ادغام سازی عمودی ممکن است اشتباها دو بلوک متفاوت، مثلا متن و تصویر، را به یکدیگر بچسباند. از این رو تنها در جهت افقی، ادغام سازی اعمال شده است. نتیجه این عمل در شکل۲-۹ نشان داده شده است. پس از

این مرحله، تمام مولفه های پیوسته‌ی تصویر با استفاده از تکنیک برچسب زنی پیدا می شود. سپس طی چند مرحله مولفه‌های نزدیک به هم بر اساس نوعشان (تصویر یا متن) با هم ادغام می شوند تا ساختارنهایی سند به دست آید.



شکل۲-۹: نتیجه ی اعمال ادغام سازی افقی

در [۲] دو ناحیه ی مجاور تنها در صورتی با هم ادغام می شوند که شرایط زیر ارضا شود:

۱. نوع آن ها یکی باشد(تصویر یا متن).

۲. دو ناحیه حداقل در یکی از راستاهای افقی و عمودی همپوشانی داشته و در راستای دیگر نیز فاصله کمی از هم داشته باشند.

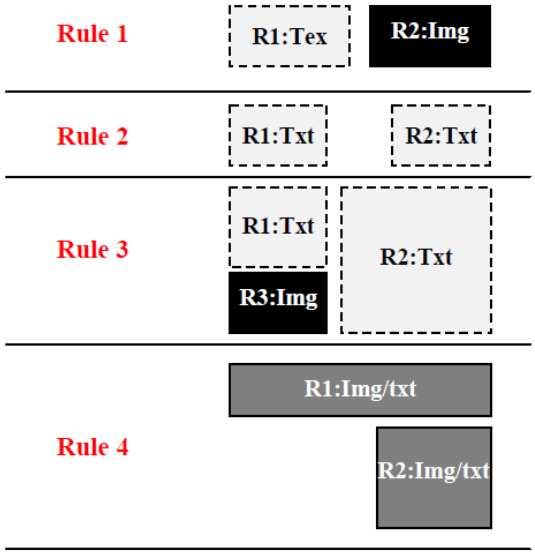
۳. ترکیب آن ها در بر گیرنده ناحیه ای با نوع متفاوت از آن ها نباشد.

۴. مساحت ناحیه جدید از ۱/۵ برابر مجموع مساحت های دو ناحیه‌ی قدیم بزرگتر نشود(نمایش مصور این قوانین در شکل۲-۱۰ آمده است).

- عمل ادغام نواحی در ۳ گام انجام می شود:

۱. ادغام مولفه های تقریبا هم عرض که بصورت عمودی مرتب باشند. با این کار از سطرهای زیر هم بلوک های متنی تشکیل می شود.

۲. ادغام مؤلفه های هم ردیف افقی، این کار برای غلبه بر خطاهای احتمالی مرحله‌ی تشخیص نواحی جداساز است. مثلا اگر کلمه ای مربوط به یک خط باشد ولی پس از آغشته سازی به دلیل حضور یک جداساز اشتباه، به سایر مؤلفه های سطر نچسبیده، در این مرحله اصلاح می شود.



شکل ۲-۱۰: شرایط نمونه ای که دو ناحیه ی مجاور ادغام نخواهند شد

۳. در مرحله آخر، نواحی که با هم همپوشانی دارند، اگر هم نوع باشند با یکدیگر ادغام می شوند و اگر از یک نوع نباشند، همپوشانی آنها اصلاح می شود.

شکل ۲-۱۱ خروجی نهایی الگوریتم را نمایش می دهد. چنانکه دیده می شود، ساختار سند به خوبی و همانگونه که مورد انتظار است استخراج شده است. در این شکل رنگ آبی برای متن و رنگ زرد برای نواحی تصویری استفاده شده است.

۲-۳ برخی دیگر از کارهای انجام شده

از سوی دیگر، روش های بالا به پایین از بلوکهای بزرگ مانند ستونهای متنی شروع کرده و شکستن متوالی بلوکها را تا رسیدن به اجزای ریز سند مانند کلمات و حروف ادامه می‌دهند. این روش ها نیازمند از

پیش دانستن برخی مشخصات ساختار سند هستند. الگوریتم های زیادی هستند که از این روش بهره می گیرند.

(Nagy,Seth et al.1992) از برش های متوالی افقی/عمودی (Recursive x-y cut) استفاده کرده است. (wang and Srihari 1989) از پروفایل افکنش تصویر و (Hose and Hoshino 2002) از تبدیل فوریه

شکل ۲-۱۱: نتیجه ی نهایی الگوریتم[۲]، تحلیل ساختار پیشنهادی

برای این امر بهره گرفته اند. روش های زیاد دیگری وجود دارند که ترکیبی از دو رویکرد گذشته اند و یا در هیچ یک از دو گروه فوق نمی‌گنجند و به عنوان روشهای ترکیبی (Hybrid Methods) مشهورند. در میان این روش ها می توان از روش های مبتنی بر بافت (Chetverikov , Liang et al .1996) ، فیلتر گابور (jain and Bhattacharjee 1992) روش های آماری (Laven,Leishman et al.2005)و برنامه ریزی خطی(Gao,wang et al.2007) نام برد.

- ✓ Bukhari و دیگران در روش ناحیه بندی و طبقه بندی مبتنی بر مؤلفه های به هم پیوسته ، نشان دادند که نیاز به یک ناحیه بندی بلوکی که قبلا معرفی شده اند نیست. آنها یک شبکه ی

MLP خود تنظیم را آموزش دادند که طبقه بندی به متن و غیر متن را با استفاده از مؤلفه های به هم پیوسته ، شکل و محتوای آن (به عنوان یک بردار ویژگی) انجام دادند[۳].

- ✓ Okun و دیگران از ویژگی های مؤلفه های پیوسته و آماری استفاده کرده است[۴].
- ✓ Wang و دیگران از سیستم طبقه بندی مبتنی بر بلوک استفاده کرده اند که هر بلوک با یک بردار ویژگی ۲۵ بعدی معرفی شده است و با استفاده از یک طبقه بند درختی، یک تصمیم بهینه شده جهت طبقه بندی بلوک ها به یکی از کلاس های مطلوب گرفتند[۵].
- ✓ Keyers و دیگران نشان دادند که طبقه بندی بلوکی را می توان به تنهایی با استفاده از بردار ویژگی طول هیستوگرام انجام داد. روش طبقه بندی بلوکی به نتایج ناحیه بندی متن به بلوک ها بستگی دارد، استفاده از روش غلط در ناحیه بندی بلوکی به عدم طبقه بندی منجر می شود[۶].
- ✓ Bloomberg و دیگران از روشی مبتنی بر ریخت شناسی چند درجه تفکیکی[۱] برای ناحیه بندی متن استفاده کردند ، این روش به طور ویژه برای جدا کردن سایه ها از تصاویر متن استفاده می شود. این روش برای جدا سازی سایه از متن خوب است اما برای جدا کردن دیگر عناصر غیر متن مانند جداول و تصاویر و ... از متن خوب نیست[۷].

یک روش ساده ناحیه بندی نیز مبتنی بر این فرضیه است، که اندازه عناصر غیر متنی بزرگتر از عناصر متنی در تصاویر متنی است.

- ✓ Oliveira و دیگران یک الگوریتم با بازدهی بالا برای ناحیه بندی تصاویر اسنادی که دارای خطوط متنی منحرف و پیچ دار می‌باشد پیشنهاد دادند. خطوط متنی پیچ دار اغلب زمانی که از اسکنرهای مسطح و یا دوربین های دیجیتال برای رقومی سازی اسناد استفاده شود به وجود می

۱ Multi resolution

آید. جبران سازی خسارت ناشی از اعوجاج یک گام مهم قبل از پردازش اسناد از طریق OCR است. ورودی این الگوریتم یک تصویر دو سطحی است که مراحل زیر را روی آن انجام می دهد:

۱) به دست آوردن مستطیل محیطی هر جزء

۲) اضافه کردن یک جعبه به یک درخت K بعدی

۳) به دست آوردن تمام نزدیک ترین همسایگی ها (K-NN)

۴) ایجاد صف اولویت همسایه ها

۵) برای هر همسایگی در صف اولویت ، یک گروه متنی و خط شروع از آنها ایجاد می شود[۸]

در بین کارهایی که برای ناحیه بندی تصاویر اسناد فارسی انجام شده است:

- ✓ آقای مهدی عباسی و دیگران روشی برای استخراج متن فارسی از تصاویررنگی پیشنهاد کرده اند که درالگوریتم آن از میان موجک های موجود، موجک گسسته Haar به علت داشتن سرعت پردازش بالا انتخاب شده است این موجک برروی اجزای رنگ (R,G,B) اعمال می شود تا لبه های متن شناسایی شوند. لبه‌های‌قوی که اکثر لبه های متن را در خود دارند با اعمال آشکار ساز لبه رنگی آشکار سازی می شوند با دو سطحی کردن جزئیات و اعمال عملگر منطقی AND نواحی نامزد متن شناسایی می شوند[۹].
- ✓ خانم محدثه قیاسی و دیگران یک روش ناحیه بندی بدون نظارت پیشنهاد می کنند ، الگوریتم پیشنهادی شامل سه مرحله ی پیش ناحیه بندی تصویر به داخل ابرپیکسل ها، استخراج ویژگیها و ادغام ابر پیکسل ها می باشد. در این روش ابتدا تصویر به ناحیه هایی به نام ابر پیکسل تجزیه می شود سپس برای استخراج ویژگیهای بافت درهرابرپیکسل ازیک توصیفگر بافت استفاده میشود. درمرحله بعد ابرپیکسلهای همسایه با درنظر گرفتن دو معیار اهمیت ادغام و توقف ، ادغام می شوند. آزمایشها نشان می دهند که با اجرای این الگوریتم پیشنهادی تصویر به نواحی متمایز

در بافت تجزیه می شود. همچنین مرز ایجاد شده بین نواحی با بافت های مختلف مرز نرم و همواری می باشد[۱۰].

جدول۲-۱: مرور مختصری بر روش های مبتنی بر تحلیل منطقی

ردیف	نویسنده	سال	رویکرد	ویژگی ها
۱	Akiyama et al	۱۹۹۰	افکنش پروفایل و مبتنی بر طرح	اسناد ژاپنی یا انگلیسی که شامل سرخط ها ، خطوط متنی و گرافیکی
۲	Esposito et al	۱۹۹۰	طرح خوشه بندی	پردازش روزنامه، با دقت ۸۹ درصد
۳	Chenevoy et al.	۱۹۹۱	بالا به پایین	مدیریت فرضیه ها در LISP
۴	Kreich et al	۱۹۹۱	استخراج مولفه های پیوسته	کلاس های زیادی از نمونه ها، مبتنی بر دانش
۵	Krishnamoorthy et al	۱۹۹۳	درختی های افقی و عمودی وبلوک های گرامر با برنامه های UNIX	تصحیح خطا با ردیابی بازگشتی، محاسبات پیچیده
۶	Sivaramakrishnan et al	۱۹۹۵	بردار ویژگی آماری و طبقه بند های درختی	چندین کلاس ، مبتنی بر پیکسل

جدول۲-۲: مرور مختصری بر روش های مبتنی بر تحلیل هندسی اسناد

ردیف	نویسنده	سال	رویکرد	ویژگی ها
۱	Wahl et al	۱۹۸۲	یکنواخت سازی طول اجرا	تشخیص انحراف-مدت اجرا
۲	Nagy et al	۱۹۸۴	برش های متوالی افقی و عمودی	تشخیص انحراف، بلوک های مستطیل
۳	Wang et al	۱۹۸۹	یکنواخت سازی طول اجرا- برش های متوالی افقی و عمودی	تشخیص انحراف،تحلیل روزنامه
۴	Fujisawa et al	۱۹۹۰	بالا به پایین	اسناد ژاپنی
۵	Fisher et al	۱۹۹۰	یکنواخت سازی طول اجرا و استخراج مولفه های پیوسته	تشخیص نواحی متنی وغیر متنی
۶	Pavlidis et al.	۱۹۹۱	افکنش در جهت عمودی	تشخیص نواحی متنی وغیر متنی
۷	Baird	۱۹۹۲	استراتژی عمومی به محلی	ظرفیت زبان های مختلف، تشخیص انحراف
۸	Jain et al	۱۹۹۲	فیلتر گابور	ویژگی های چند کاناله از تصویر خاکستری
۹	Lebourgeois et al	۱۹۹۲	فیلتر با پنجره های 8X3	اسناد نامحدود وعدم تشخیص انحراف
۱۰	Pavlidis et al.	۱۹۹۲	آغشته سازی اافقی و پایین به بالا	تشخیص انحراف کوچک،پارامترهای ثابت
۱۱	Akindele et al	۱۹۹۳	ردیابی فضای سفید	بلوک های چند منظوره، در نظر گرفتن نواحی متنی
۱۲	Amamoto et al	۱۹۹۳	عملیات ریخت شناسی در فضای سفید	تشخیص نوشته های افقی و عمودی، عدم تشخیص انحراف
۱۳	Zlatopolsky	۱۹۹۴	استخراج مولفه های پیوسته	چند سند اریب و انحراف دار، پارامترهای حساس
۱۴	Sylwester et al	۱۹۹۵	قابل آموزش با برش های افقی و عمودی	نسبتا قوی ، انحراف و نویزهای آزاد
۱۵	Liu et al.	۱۹۹۶	بالا به پایین وپایین به بالا تطبیقی	نواحی غیر مربعی و تشخیص انحراف

فصل سوم:

مباحث تئوری

۳-۱ دو سطحی سازی

در این قسمت به تعدادی از روش های دو سطحی سازی و مقایسه ی آنها با یکدیگر می پردازیم.

۳-۱-۱ دو سطحی سازی به روش سراسری

در دو سطحی سازی به روش سراسری، یک و تنها یک آستانه برای کل تصویر به دست می آید. اگر سطح خاکستری یک پیکسل، بزرگتر از آستانه باشد مقدار پیکسل، یک و غیر این صورت پیکسل مقدار صفر را اختیار می کند.

دو سطحی سازی به روش آتسو[۱]

با توجه به تحقیقاتی که بر روی روشهای مختلف دو سطحی سازی انجام شد، محققان به این نتیجه رسیدند که روش آتسو جامعیت بیشتری نسبت به سایر روش های دو سطحی سازی به روش سراسری دارد. در این روش مقدار بهینه‌ی آستانه طوری تعیین می‌شود که مقدار واریانس درون‌کلاسی سفید وسیاه ، کمینه و واریانس بین کلاسی بیشینه گردد. اگرچه روش آتسو به دلیل بدست آوردن تنها یک آستانه، از لحاظ محاسباتی بی هزینه است و برای بیشتر اسناد اسکن شده جواب قابل قبولی را ارائه می دهد، ولی برای اسنادی مانند کتب قطور، اسناد تاریخی و تصاویر گرفته شده به وسیله دوربین دیجیتال که دارای روشنایی غیر یکنواخت می باشند، باعث ایجاد نویزهای فلفلی می گردد.

در حالت کلی دو سطحی سازی به روش سراسری در ۵ گروه دسته بندی می شود:

۱-روش های مبتنی بر شکل هیستوگرام: در این روش ها قله ها ، دره ها و انحنای هیستوگرام نرم[۲]تصاویر مورد بررسی قرار می گیرد.

---

۱ Otsu

۲ Smoothed histogram

۲-روش های مبتنی بر خوشه بندی :سطوح خاکستری موجود در تصاویر در دو گروه پیش زمینه و پس زمینه خوشه بندی شده و یا به عبارتی به صورت تلفیقی از دو توزیع گوسی مدل می شوند.

۳-روش های مبتنی بر آنتروپی :روش هایی‌که در آن ها ، از آنتروپی نواحی پیش زمینه پس زمینه و اختلاف آنتروپی بین تصویر اصلی و تصویر دو سطحی شده استفاده می شود.

۴-روش های مبتنی بر شئ گرایی: روش هایی که در آن ها ، از مقدار تشابه میان تصویر خاکستری و تصویر دو سطحی شده مانند تشابه شکل فازی ، انطباق لبه و ... استفاده می شود.

۵-روش های فضایی: روش هایی که در آن ها از توزیع های احتمال مرتبه بالا و یا همبستگی بین نقاط تصویر استفاده می شود.

۳-۱-۲ دو سطحی سازی به روش محلی

وقتی‌پس‌زمینه‌دارای روشنایی غیر یکنواخت باشد ،روش‌های آستانه گیری سراسری با شکست مواجه می‌شوند. در دو سطحی سازی به روش محلی، برای هر کدام از پیکسل های تصویر یک آستانه اختیار می گردد. روش های گسترده ای در این حوزه وجود دارد به عنوان مثال روش هایی بر پایه ی استفاده از واریانس و میانگین محلی نقاط اطراف هر پیکسل و روش هایی بر پایه‌ی مقایسه‌ی سطح خاکستری هر پیکسل با سطح خاکستری پیکسل های همسایه‌ی آن می باشد. بر پایه ی تحقیقات صورت گرفته دو سطحی سازی به روش محلی جواب بهتری را نسبت به سایر روش های دو سطحی سازی دارد، برای همین توجه خود را روی این دسته از دو سطحی سازی معطوف می کنیم.

آستانه گذاری وفقی محلی

در این دسته از الگوریتم ها یک آستانه در هر پیکسل محاسبه می شود که به بعضی از اطلاعات آماری مانند گستره واریانس یا پارامترهای انطباق سطح همسایگی پیکسل بستگی دارد. در ادامه آستانه ی

th(i,j) به عنوان تابعی از مختصات های (i,j) در هر پیکسل نشان داده شده است. یا اگر این ممکن نیست تصمیمات شئ پس زمینه بودن به وسیله‌ی متغیر B(i,j) نشان داده می‌شوند. روش‌های متعددی در این دسته از الگوریتم ها پیشنهاد شده است.

روش کنتراست محلی

مقدار ارزش خاکستری پیکسل را با میانگین ارزش های خاکستری همسایگی حول آن پیکسل مقایسه می کند. اگر پیکسل به طور مشخصی از میانگین تیره تر است به عنوان نویسه و در غیر این صورت به عنوان پس زمینه نمایش داده می شود. آستانه در مقدار وسط گستره تنظیم می شود که میانگین مینیمم و ماکزیمم ارزش های خاکستری در پنجره محلی پیشنهاد شده است. با این وجود اگر کنتراست C(i,j) از یک آستانه مشخص کمتر باشد گفته می شود همسایگی تنها شامل یک کلاس است ، به عبارتی پس زمینه یا پیش نما بودن به مقدار th(i,j) وابسته است. در گستره‌ی دینامیکی، تصویرگسترده می شود سپس نرم کنندگی غیر خطی در آن اتفاق می افتد که در آن لبه ها ی تیز حفظ می شود. نرم کنندگی شامل جایگزین کردن هر پیکسل با میانگین یک پنجره از همسایه اش (۳×۳، ۵×۵ و .....). جدول ۳-۱ خلاصه ای از این روش ها را نشان می دهد.

جدول۳-۱:الگوریتم های آستانه گذاری وفقی محلی[۲۰]

Local_Niblack[110]	$T(i,j)=m(i,j)+k.\sigma(i,j)$ where $k=-0.2$ and local window size is $b=15$
Local_Sauvola[111]	$T(i,j)=m(i,j)+\left\{1+k.\left[\frac{\sigma(i,j)}{R}-1\right]\right\}$ where $k=0.5$ and $R=128$
Local_White[112]	$B(i,j)=\begin{cases}1 & \text{if } m_{w\times w}(i,j)<I(i,j)*\text{bias}\\ 0 & \text{otherwise}\end{cases}$ where $m_{w\times w}(i,j)$ is the local mean over a $w=15$-sized window and bias=2.
Local_Bernsen[113]	$T(i,j)=0.5\{\max_w[I(i+m,j+n)]+\min_w[I(i+m,j+n)]\}$ where $w=31$, provided contrast $C(i,j)=I_{high}(i,j)-I_{low}(i,j)\geqslant 15$.
Local_Palumbo[21]	$B(i,j)=1$ if $I(i,j)\leqslant T_1$ or $m_{neigh}T_3+T_5>m_{center}T_4$ where $T_1=20$, $T_2=20$, $T_3=0.85$, $T_4=1.0$, $T_5=0$, neighborhood size is 3×3.
Local_Yanowitz[115]	$\lim_{n\to\infty} T_n(i,j)=T_{n-1}(i,j)+R_n(i,j)/4$ where $R_n(i,j)$ is the thinned Laplacian of the image.
Local_Kamel[1]	$B(i,j)=1$ if $\{[L(i+b,j)\wedge L(i-b,j)]\vee[L(i,j+b)\wedge L(i,j-b)]\}$ $\{[L(i+b,j+b)\wedge L(i-b,j-b)]\vee[L(i+b,j-b)\wedge L(i-b,j+b)]\}$ where $L(i,j)=\begin{cases}1 & \text{if } [m_{w\times w}(i,j)-I(i,j)]\geqslant T_0\\ 0 & \text{otherwise}\end{cases}$, $w=17$, $T_0=40$
Local_Oh[13]	Define the optimal threshold value (T_{opt}) by using a global thresholding method, such as the Kapur[53] method, then locally fine tune the pixels between [T_0-T_1] considering local covariance ($T_0<T_{opt}<T_1$).
Local_Yasuda[114]	$B(i,j)=1$ if $m_{w\times w}(i,j)<T_3$ or $\sigma_{w\times w}(i,j)>T_4$ where $w=3$, $T_1=50$, $b=16$, $T_2=16$, T_3 128, $T_4=35$

۳-۲ روش های مبتنی بر تبدیل هاف[1]

تبدیل هاف یک تکنیک بسیار محبوب برای پیدا کردن خطوط صاف در تصاویر است. صف بندی های بالقوه در دامنه ی هاف فرض شده و در دامنه‌ی تصویر معتبر می شوند. بنابراین هیچ فرضی در مورد

[1] Hough

مسیرهای خط متن ساخته نمی شود(ممکن است چند تا از آنها در یک صفحه وجود داشته باشند). مراکز مؤلفه های پیوسته متصل به هم ، واحدهای تبدیل هاف هستند.

$$x\cos\theta + y\sin\theta = \rho \qquad (۱-۳)$$

در تفسیر پارامترهای ρ, θ خط افقی دارای $\theta = 0^{\circ}$ است، به طوری که ρ برابر با مثبت x است. به طور مشابه خط عمودی دارای $\theta = 90^{\circ}$ است ، به طوری که ρ برابر y مثبت است. جذابیت محاسباتی تبدیل هاف ناشی از تقسیم فضای پارامتر $\rho\theta$ به سلول های انباشتگر[۱] است ، که $[\theta_{\min}, \theta_{\max}]$ و $[\rho_{\min}, \rho_{\max}]$ بازه های مورد انتظار مقادیر پارامتر هستند. معمولا بازه ی ماکزیمم مقادیر برابر با $-D \le \rho \le D$ و $-90 \le \theta \le 90$ است، که D دورترین فاصله ی بین گوشه های مقابل در تصویر است. سلول موجود در مختصات (i, j) با مقدار انباشتگر $A(i, j)$ متناظر با مربع مربوط به مختصات (ρ_i, θ_j) فضای پارامتر می باشد. در آغاز ، این سلول ها برابر با صفر هستند. سپس ، برای هر نقطه ی همسایه (x_k, y_k) در صفحه تصویر (یعنی صفحه xy) ، θ را برابر با مقادیر زیر تقسیم های مجاز در محور θ قرار می دهیم ، و آن را برای ρ متناظر ، با استفاده از معادله $\rho = x_k \cos\theta + y_k \sin\theta$ حل می کنیم.

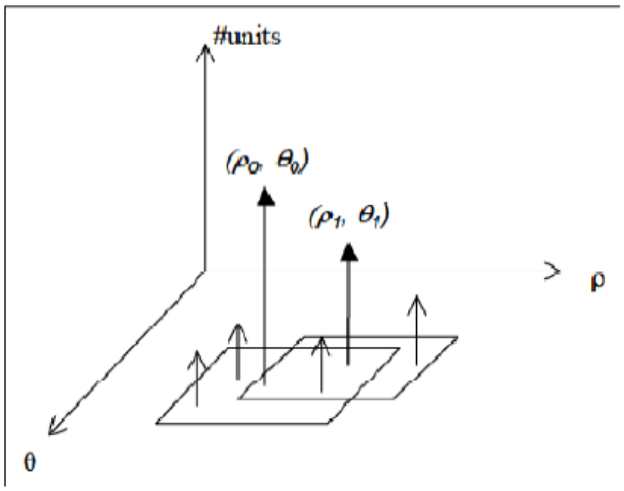
سپس مقادیر ρ که به دست آمدند ، به نزدیکترین مقدار سلول در امتداد محور ρ گرد می شود. سپس سلول انباشتگر متناظر افزایش می یابد. در انتهای این رویه ، مقدار Q در سلول $A(i, j)$ به معنای این است که Q نقطه در صفحه xy ، روی خط $x\cos\theta_j + y\sin\theta_j = \rho_i$ قرار دارند. تعداد زیر تقسیم ها در صفحه، $\rho\theta$ دقت هم راستایی این نقاط را تعیین می کند. مجموعه ای از واحدهای صف بندی شده در تصویر ، در طول یک خط با پارامترهای (ρ, θ) در سلول مرتبط (ρ, θ) در دامنه‌ی هاف قرار دارند. صف بندی های شامل واحدهای زیاد، با سلول هایی با قله‌ی بالا در دامنه‌ی هاف مرتبط هستند. برای در نظر گرفتن نوسانات خطوط متن دست نویس یعنی اینکه واحدهایی که در یک خط متن هستند به طور دقیق

---

۱ accumulator

صف بندی نشده اند. دو فرض برای هر صف بندی مد نظر قرار می‌گیرد و یک صف بندی از واحدهای ساختار سلولی یک سلول ابتدایی تشکیل شده است. فرض اول ساختار سلولی از یک سلول (ρ_0,θ_0) ، شامل همه‌ی سلول های قرار گرفته در یک خوشه با مراکز (ρ,θ) می باشد. سلول (ρ,θ) شامل بیشترین تعداد واحدها را در نظر بگیرید.

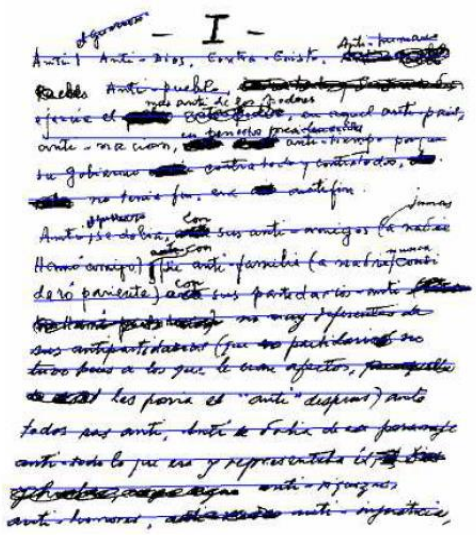
فرض دوم (ρ_1,θ_1) در ساختار سلول (ρ_0,θ_0) جستجو می شود. صف بندی انتخاب شده میان این دو فرضیه قوی‌ترین آن هاست یعنی آن صف بندی که شامل بیشترین تعداد واحد در ساختار سلولش است. و سلول مرتبط (ρ_0,θ_0) یا (ρ_1,θ_1) سلول ابتدایی است.



شکل۳-۱:سلول های فرضی (ρ_0,θ_0) و (ρ_1,θ_1) در فضای هاف هر قله به واحدهای به دقت صاف شده مرتبط است.

با این وجود خطوط متن دقیق به ندرت به صف بندی هایی با بیشترین تعداد واحد مربوط می شوند. مانند صف بندی های متقاطع باید شامل واحدهای بیشتری از خطوط متن باشند. یک صف بندی بالقوه با استفاده از اطلاعات متنی معتبر می شود یعنی با در نظر گرفتن همسایه های داخلی و خارجی. یک همسایه درونی یک واحد j یک همسایه‌ی صف بندی درون هاف است. یک همسایه‌ی خارجی یک همسایه ی خارج صف بندی هاف است که در دایره ای به شعاع δ_j از واحدj قرار می‌گیرد. فاصله‌ی δ_j ، فاصله

ی میانگین فواصل همسایه‌ی داخلی از واحدj است. برای معتبر شدن ، یک صف بندی بالقوه می‌تواند شامل واحدهای خارجی کمتری نسبت به واحدهای داخلی باشد. این صفر رد کردن صف بندی هایی که ربط ادراکی ندارند را باعث می گردد. این روش می تواند خطوط متن جهت دار شده و حاشیه نویسی های شیب دار را تحت این فرض که چنین خط هایی تقریبا صاف هستند را استخراج کند.



شکل۳-۲:خطوط استخراج شده از یک دست نوشته

تبدیل هاف می‌تواند به خطوط نوسانی دست نویس ها هم اعمال شود. تبدیل هاف ابتدا به نقاط مینیمم (واحدها)در یک نوار عمودی روی سمت چپ تصویر اعمال می شود. صف بندی ها داخل دامنه هاف با شروع از یک مسیر اصلی جستجو می شوند و بوسیله‌ی گروهی کردن سلول ها در یک جستجوی کامل در ۶ جهت این کار انجام می شود. پس یک پنجره‌ی متحرک که به طرح خوشه بندی در دامنه تصویر مرتبط است واحدهای باقیمانده را به صف بندی ها اختصاص می دهد. طرح خوشه بندی تولید خطوط جدید شروع شونده در وسط صفحه را ممکن می کند.

تبدیل هاف سلسله مراتبی

پس از استخراج مؤلفه های پیوسته در تصویر سند ، در تعدادی از خطوط متن به جای هر مؤلفه ی پیوسته یک نقطه نماینده از آن را انتخاب و سپس با استفاده از تبدیل هاف سلسله مراتبی ، تشخیص زاویه در دو رزولوشن زاویه ای درشت و ریز با یک الگوریتم تبدیل هاف انجام می پذیرد. الگوریتم کلی تشخیص کجی به قرار زیر می‌باشد:

۱- دو سطحی سازی ، ۲- یافتن مؤلفه های پیوسته در تصویر سند ، ۳- حذف مؤلفه های پیوسته کوچک (نویز ، نقاط حروف) و بزرگ (تصاویر و گرافیک) از تصویر سند ۴- نگاشت هر جزء مؤلفه های پیوسته به یک نقطه درون آن جزء ، ۵- اعمال تبدیل هاف با رزولوشن زاویه ای درشت ۶- اعمال تبدیل هاف با رزولوشن زاویه ای ریز در محدوده ای از زاویه تخمینی قبلی

۳-۳ ریخت شناسی[۱]

لغت ریخت شناسی عموما در مورد شاخه ای از علم که درباره شکل و ساختار اجسام صحبت می کند، به کار می رود. در این جا از این لغت با محتوای ریخت شناسی ریاضیاتی و به عنوان ابزاری برای استخراج اجزای تصویر استفاده می شود، این ابزار در ارائه و توصیف شکل نواحی و خصوصیاتی مانند مرزها، اسکلت و تحدب بدنه بسیار مفید می باشد، همچنین ما بسیار علاقه مند به تکنیکهای ریخت شناسی درراستای پیش پردازش و پس پردازش تصاویر و انجام عملیات هایی مانند فیلترینگ، نازک سازی و هرس کردن ریخت شناسی هستیم. تکنیک های مورد استفاده برای تشخیص اندازه توزیع ذرات هر یک تصویر بخش مهمی از زمینه علمی ذره شناسی را تشکیل می دهند، از تکنیک های ریخت شناسی می توان برای اندازه گیری اندازه توزیع ذرات به صورت غیر مستقیم استفاده نمود، این کار بدون تشخیص انحصاری و اندازه گیری هر یک از ذرات انجام می‌پذیرد، برای ذراتی‌که شکل منظم می باشند و روشن‌تر

۱ morphological

از پس زمینه هستند. توابع ریخت شناسی بر پایه ی استفاده از نظریه‌ی مجموعه ها در زمینه‌ی پردازش تصویر می‌باشند، به این دلیل توابع ریخت شناسی روش هایی نسبتا قدرتمند و یکتا درشماری از مسائل پردازش تصویر محسوب می شوند به طور کلی از توابع ریخت شناسی به عنوان ابزاری برای شناسایی مرزها، حذف جزئیات در تصاویر، فیلترینگ و ...استفاده می شود. یکی از مهمترین مراحل در پیش پردازش تصویر ، پردازش شکل شناسی می باشد. در این قسمت تنها به بررسی شکل شناسی برای تصاویر دودویی خواهیم پرداخت. منظور از تصاویر دودویی، تصاویر با دو سطح روشنایی ۰ یا ۱ می باشد که در آن منظور از ۰ رنگ سیاه و منظور از ۱ رنگ سفید می باشد. پردازش شکل شناسی عموما از عملگرهای مجموعه ای استفاده می کند. از شکل شناسی نیز بیشتر برای استخراج نقاط کلیدی تصویر ، حذف نقاط غیر مفید تصویر و موارد مشابه دیگر استفاده می کنیم. در این قسمت ابتدا عملگرهای مجموعه ای پایه را برای پردازش شکل شناسی بررسی می کنیم. سپس به بررسی مهمترین عملگرهای مجموعه ای که ویژه پردازش شکل شناسی می باشند ، خواهیم پرداخت.

۳-۳-۱عملگرهای مجموعه‌ای پایه

این عملگر ها شامل عملگر اجتماع ، اشتراک ، تفاضل دو مجموعه می باشد. اگر هر تصویر دو سطحی را یک مجموعه در نظر بگیریم ، اجتماع دو تصویر باینری هم اندازه ، تصویری خواهد بود که در این تصویر هر پیکسلی که در تصویر اول یا دوم مقدار ۱ داشته باشد ، مقدار ۱ خواهد داشت. برای پیاده سازی عملگر اجتماع برای دو تصویر باینری ، می توانیم پیکسل های متناظر را در دو تصویر باهم OR بیتی کنیم.

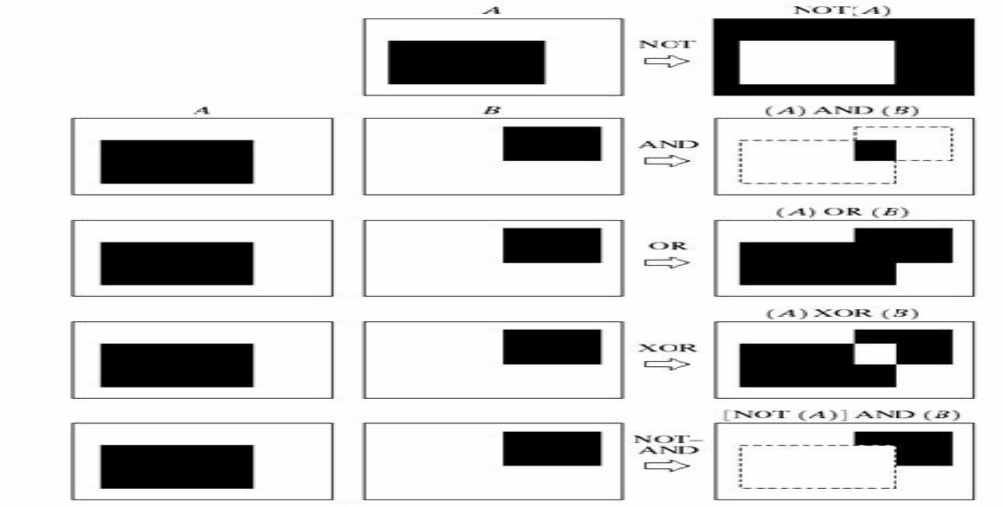
- **اشتراک دو تصویر دوسطحی** هم اندازه ، تصویری خواهد بود که در این تصویر هر پیکسلی که در تصویر اول و دوم مقدار ۱ داشته باشد ، مقدار ۱ خواهد داشت. برای پیاده سازی عملگر

اشتراک برای دو تصویر باینری ، می توانیم پیکسل های متناظر را در دو تصویر باهم AND بیتی کنیم.

- **تفاضل دو تصویر دوسطحی** هم اندازه ، تصویری خواهد بود که در این تصویر پیکسل هایی از تصویر اول با مقدار ۱ که در تصویر دوم مقدارشان ۱ نباشد ، مقدار ۱ خواهند داشت. همچنین عملگر تک عملوندی مکمل نیز عملگری است که پس از اعمال آن بر روی تصویر دوسطحی ، در تصویر حاصل شده، مقادیر ۱ به . و مقادیر . به ۱ تغییر می یابند. برای پیاده سازی عملگر مکمل می توانیم هر پیکسل از تصویر دوسطحی را NOT بیتی کنیم. شکل ۳-۳ زیر نتیجه اعمال عملگر مذکور را بر روی دو تصویر دوسطحی نشان می دهد. حال ازعملگر مکمل استفاده کرده و عملگر تفاضل را می توانیم به شکل دیگری پیاده سازی کنیم. تفاضل مجموعه A از مجموعه B را می توانیم به صورت زیر تعریف کنیم:

$$A - B = A \cap B' \qquad (۲-۳)$$

یعنی تفاضل مجموعه A از مجموعه B برابر است با اشتراک مجموعه A با مکمل مجموعهB.



شکل۳-۳: اعمال عملگر تفاضل دو تصویر

- **عملگر قرینه** : عملگر قرینه برای تصویر A بدین معنی است که تمام پیکسل ها نسبت به مرکز تصویر قرینه شوند(W اعضای مجموعه‌ای است که تولید می کنیم مثلا مجموعه reflection).

$$reflection(A) = \{W \mid W = -b, b \in B\} \qquad (۳-۳)$$

۳-۳-۲ عملگر گسترش[۱]

همانطور که از نام عملگر پیداست ، این عملگر باعث گسترش نقاط ۱ در تصویر می شود. در عملگر گسترش

نیز از یک نقاب همانند آنچه که در" ارتقا تصویر در حوزه مکانی "بود، استفاده می کنیم. در اینجا به

جای نقاب به آن عنصر ساختاری[۲] می گوییم که مقادیر عنصر ساختاری ۱ یا صفر می تواند باشد. گسترش تصویر A با عنصر ساختاری B به صورت زیر تعریف می شود :

$$A \oplus B = \{W \mid reflection(B) \cap A \neq NULL, w \in A\} \qquad (۴-۳)$$

که در اینجا reflection عنصر ساختاری B را حول مرکز خود قرینه می کند. به عبارت دیگر گسترش A با عنصر ساختاری B بدین معنی است که اگر عنصر ساختاری B را بر روی پیکسل های A حرکت دهیم ، و در هربار حرکت اشتراک عنصر ساختاری با محدوده زیر عنصر ساختاری در تصویر A تهی نباشد، مقدار پیکسل مرکزی که عنصر ساختاری بر روی آن قرار گرفته است ، برابر ۱ خواهد شد. شکل ۳-۴ خروجی تصویر را پس از گسترش تصویر با یک عنصر ساختاری۳×۳ که همه عناصر آن ۱ است را نشان می دهند:

---

۱ Dilation

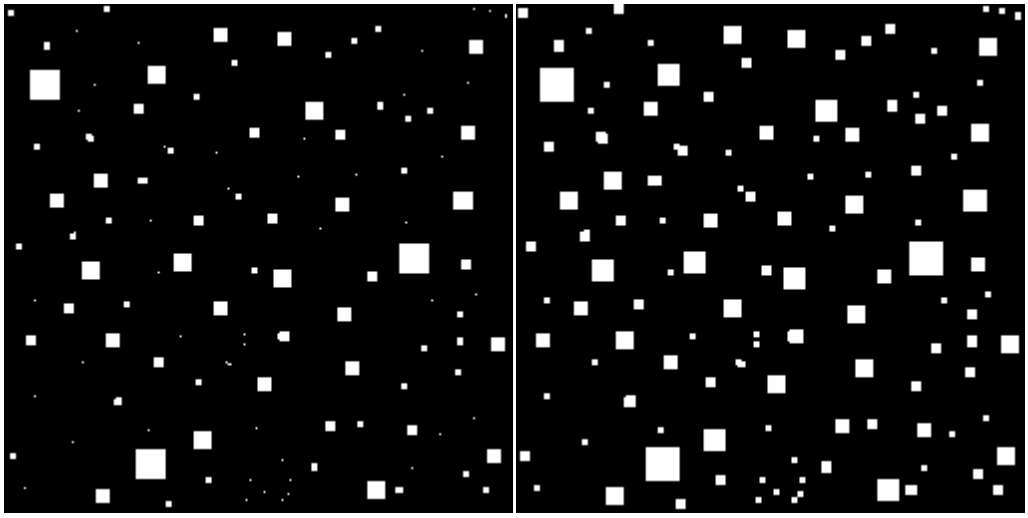
۲ Structure Element(SE)

عنصر ساختاری

1 1 1

1 1 1

1 1 1



شکل۳-۴: پس از گسترش(سمت راست)-تصویر اصلی(سمت چپ)

۳-۳-۳ عملگرفرسایش[1]

همانطور که از نام عملگر پیداست ، این عملگر باعث سایش نقاط ۱ در تصویر می شود. همانند عملگر گسترش ، درعملگر فرسایش نیز از یک عنصر ساختاری استفاده می کنیم که مقادیر عنصر ساختاری ۱ یا صفر می تواند باشد. به ازای هر پیکسل مرکز عنصر ساختاری را روی پیکسل قرار داده و عملگر فرسایش را با توجه به مقادیر عنصر ساختاری در مورد آن پیکسل اعمال می‌کنیم. سایش تصویر A با عنصر ساختمانی B به صورت زیر تعریف می شود:

$$A\Theta B = \{W \mid (B) \subseteq A, w \in A\} \qquad (۵-۳)$$

[1]Erosion

به عبارت دیگر گسترشA با عنصر ساختمانی B بدین معنی است که اگر عنصر ساختاری B را بر روی پیکسل های A حرکت دهیم ، و در هر بار حرکت همه نقاطی که در زیر مقادیر ۱ از عنصر ساختاری قرار گرفته اند نیز ، مقداز یک داشته باشند ، مقدار پیکسل حاصل نیز ۱ خواهد بود.

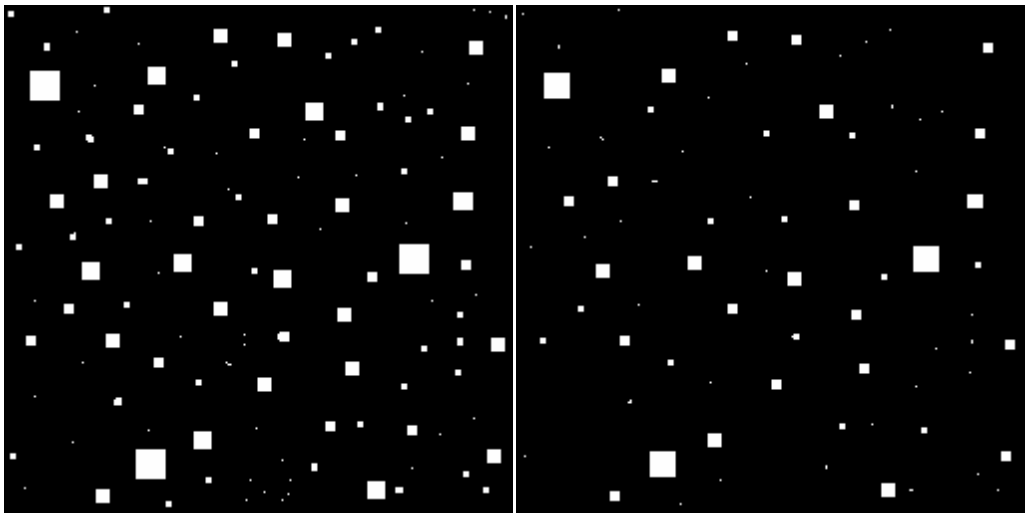
شکل زیر خروجی تصویر را پس از سایش تصویر با یک عنصر ساختاری ۳×۳ که همه عناصر آن ۱ است را نشان می دهد:

عنصر ساختاری

1 1 1

1 1 1

1 1 1



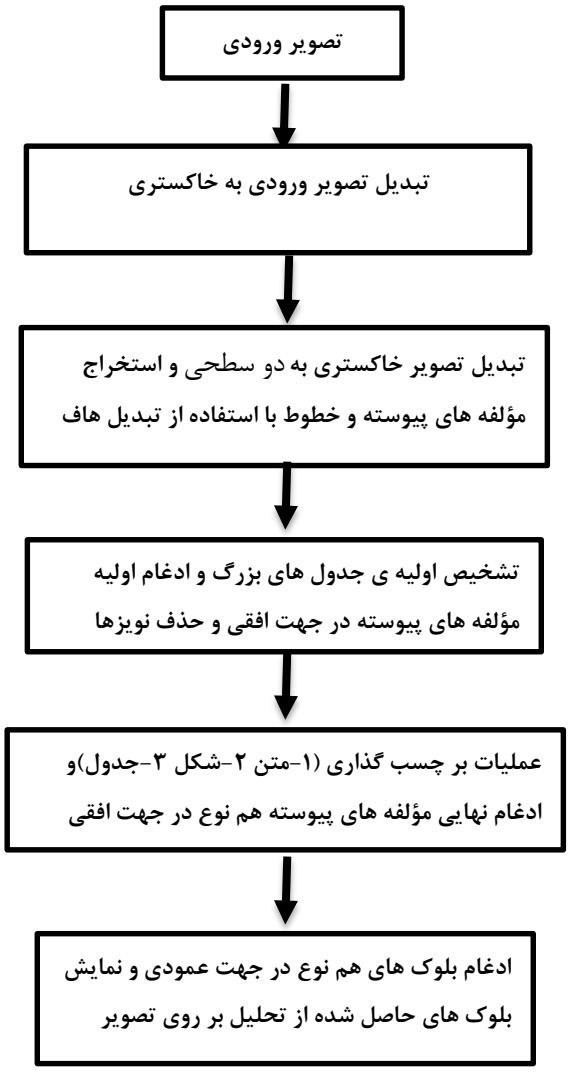
شکل۳-۵: پس از فرسایش(سمت راست)-تصویر اصلی(سمت چپ)

فصل چهارم:

روش پیشنهادی

۴-۱ مقدمه

هدف از انجام این پایان نامه ، تحلیل ساختار اسناد پیچیده و ناحیه بندی تصاویر اسناد به بلوکهای متن، شکل و جدول است. در تصویر خروجی اطراف بلوک های متنی را با کادر آبی رنگ و اطراف شکل ها را با کادر قرمز و اطراف جداول را با کادر سبز رنگ مشخص می کنیم. استخراج اطلاعات متنی از اسناد به دلیل جایگذاری شکل ها، عناوین و زیرنویس شکل ها، زمینه پیچیده، تنوع در اندازه قلم و فاصله خطوط کاری بسیار مشکل و فراتر از توانایی سیستمهای تحلیل ساختار امروزی است. تصویر یک سند می تواند شامل نواحی گوناگونی مانند بلوک های متنی، پاراگراف ها، خطوط، واژه ها، شکل ها، جداول و زمینه باشد که آن ها را برچسب گذاری می کنیم. هم چنین می توان یک سند را متشکل از برچسب های کارکردی دانست(مانند عناوین، شرح تصویرها، یک یا چند تصویر برای سند) که به برخی از نواحی اختصاص داده می شوند. ساختار یک سند ، تقسیم و زیر تقسیم های متوالی محتوای یک سند به اجزای کوچکتری به نام شئ است که این شئ ها خود به دو دسته تقسیم می شوند: ساده و مر کب. ساختار اسناد را ازیک دیدگاه می توان به دو نوع تقسیم کرد: ۱-ساختار فیزیکی ۲-ساختار منطقی ساختار فیزیکی نتیجه تقسیم و زیر تقسیم های محتوای ساختار یک سند به بخش های کوچکتر، بر مبنای نمایش است : مانند متن، تصویر، جدول و .... ساختار منطقی نتیجه تقسیم و زیر تقسیم های محتوای ساختار یک سند به بخش های کوچکتر، بر مبنای معنای قابل درک انسان است : مانند عنوان مقاله ، نام نویسنده، زیرنویس، شرح تصویر و... مراحل تحلیل ، شناسایی مؤلفه های پیوسته، تشکیل بلوک ها و ناحیه بندی اسناد که در این پایان نامه انجام شده است به طور خلاصه در شکل (۴-۱) نشان داده شده است.



شکل ۴-۱ : مراحل انجام تشخیص و آشکار سازی بلوک های متن، شکل و جدول در اسناد

۴-۲ چالش های تحلیل ساختار فیزیکی

- **قطعه بندی نواحی**، در روش های پایین به بالا، هنگامی که تلاش می کنیم اجزاء کوچکتر را به نواحی گروه کنیم، باید تصمیم بگیریم که کجا معیار همگنی برآورده شده است. همگنی معمولا بر اساس یک پارامترهایی تعریف می‌شود که مقادیر آن نقش مهمی در موفقیت الگوریتم

بازی می کنند محاسبه پارامترهایی که برای تمام اسناد ممکن خوب کار کند، اغلب امکان پذیر نیست.

- **طبقه بندی نواحی،** چالش بعدی تعیین نوع ناحیه قطعه بندی شده است. مشکلات نوعی طبقه بندی بین متن ، تصویر و شناسایی جدول ها هستند. طبقه بندی ناحیه برای بسیاری از عملیات مراحل بعدی پیش نیاز است.

۴-۳ پیش پردازش و قطعه بندی

به طور خلاصه در مرحله پیش پردازش و قطعه بندی کارهای زیر صورت می گیرد:

دو سطحی سازی ، استخراج مؤلفه های پیوسته و قطعه بندی کردن اولیه‌ی تصویر به وسیله ی مستطیل های محیطی ، حذف کردن نویزها ، شناسایی خطوط عمودی و افقی و شناسایی جدول های بزرگ.

۴-۳-۱ دو سطحی سازی

ابتدا سند ورودی را با توجه به کیفیت و رزولوشنی که دارد تغییر اندازه می دهیم تا مؤلفه های پیوسته غیر ضروری آشکار سازی نشوند و تعداد مؤلفه های پیوسته بیش از حد نشوند مدت زمان اجرای برنامه طولانی نشود. چون در تعداد زیادی از حلقه هایی که مقایسه بین تمام مؤلفه های پیوسته انجام می شود ، موجب کاهش شدید سرعت برنامه می شود. کیفیت تصاویر اسنادی که ما در اختیار داریم حداکثرdpi۳۰۰ است. با استفاده از روابط زیر ابتدا کیفیت تصویر ورودی را به حداکثر dpi ۱۰۰می رسانیم و طول و عرض تصویر جدید را به دست می آوریم. تعدادی از اسناد مورد استفاده در این پایان نامه را در شکل(۴-۲) می بینیم. (dpiکیفیت تصویر اولیه است که ما آن را ۳۰۰ در نظر گرفته ایم).

$$H2 = \frac{H1 \times 100}{dpi} \qquad W2 = \frac{W1 \times 100}{dpi} \qquad (۱-۴)$$

بعد از تغییر اندازه ی تصویر ورودی آن را به فضای خاکستری تبدیل می کنیم. برای ساخت تصویر دو سطحی از آستانه گیری با روش وفقی محلی استفاده می کنیم. تصویر خاکستری شده را با استفاده از یک نقاب میانگیری ۱۵×۱۵هموار سازی می کنیم .

شکل ۴-۲ : تعدادی از اسناد استفاده شده در این پایان نامه

بعد از اینکه هموارسازی تصویر خاکستری شده را انجام دادیم ، برای اینکه خط ها ی گرافیکی و لبه ها یی که در برخی تصاویر با زمینه های گرافیکی وجود دارند، در تصویر دو سطحی به عنوان خط های جدول آشکار نشوند و همچنین قطعه ای از تصویر را به عنوان یک مؤلفه‌ی پیوسته بزرگ و غیر واقعی در نظر نگیرد ، ما قبل از ایجاد تصویر دو سطحی اقدام به حذف و نرم کردن لبه های سند خاکستری کرده ایم. این کار را به این ترتیب انجام داده ایم که با کم کردن تصویر هموار شده ($I_{blurred}$) از تصویر خاکستری(I_{gray}) یک تصویر با لبه های تیز شده (I_d) ایجاد می کنیم. و در نهایت با کم کردن ضریبی از تصویر لبه ها از تصویر خاکستری یک تصویر خاکستری با لبه های نرم و کم رنگ شده به نام (I_2) به دست می آوریم و این تصویر را دو سطحی کنیم. به دلیل اینکه روشنایی اسناد در نقاط مختلف آن دارای مقدارهای مختلفی هستند. برای هر پیکسل یک مقدار آستانه متفاوت در نظر می گیریم که این مقدار آستانه را به طور تجربی و با آزمایش مقادیر مختلف به دست آورده ایم. به ازای هر پیکسل تصویر (I_2) ، اگر روشنایی پیکسل مورد نظر کمتر از ۰/۸۴روشنایی پیکسل متناظر در تصویر هموار شده باشد را صفر در نظر می گیریم و در صورتیکه بیشتر از آن باشد را یک در نظر می گیریم . تصویر دو سطحی (I_{bin}) به دست آمده را در شکل (۴-۳) می بینیم. عملیات ریاضی انجام شده جهت دو سطحی سازی اسناد ، و تصویر خروجی که دو سطحی شده را در ادامه می بینیم.

$$I_d = I_{gray} - I_{blurred} \longrightarrow I_2 = I_{gray} - 1.5 \times I_d \longrightarrow I_{bin} = (I_2 < I_{blurred} \times 0.84) \qquad (۲-۴)$$

شکل۴-۳: نمونه از تصویر دو سطحی ایجاد شده(چپ:روش آتسو، راست: روش وفقی استفاده شده)

۴-۳-۲ استخراج خطوط افقی و عمودی جهت تشخیص جدول

در گام نخست ما تخمینی از قطعات افقی و عمودی از خطوط می زنیم. تخمین نهایی از خطوط زمانی انجام خواهد شد که خطوط با طول خیلی کوچک و خطوط مربوط به نواحی متنی و شکل ها را حذف کرده باشیم. الگوریتم پیشنهاد شده جهت آشکار سازی خطوط مبتنی بر یک سری عملیات ریخت شناسی با عناصر ساختاری مناسب است. از عملیات ریخت شناسی به این دلیل استفاده می شود که اگر بین خطوط شکستگی وجود داشته باشد با بهبود و ترمیم آنها خطوطی واقعی تر را آشکار سازی کند. تمام پارامترهای مورد استفاده در این گام ها به کمیتی که مقدار میانگین ارتفاع کاراکترها است بستگی دارد که ما به صورت تجربی آن را (AH=20) در نظر می گیریم. با تصویر دو سطحی شده آغاز می کنیم:

ما اقدام به عملیات ریخت شناسی با عناصر ساختاری مناسب بر روی تصویر دو سطحی می کنیم. تا محل های شکستگی خطوط ترمیم شوند اما به کاراکتر های همسایه‌ی خود متصل نشوند. ما تصاویر IM_H و IM_V را که به ترتیب بهبود یافته‌ی خطوط افقی و عمودی را نشان می دهند را به دست می آوریم، که

روابط ریاضی آنها در ادامه می آید($\Theta,\oplus$ به ترتیب اعمال گسترش و فرسایش است). IM تصویر دو سطحی اولیه است (شکل ۴-۴ را ببینید).

$$IM_H = IM \cup (((IM \Theta B_{HR}) \cup (IM \Theta B_{HL})) \oplus B_H) \qquad (۳-۴)$$

$$B_{HR} = [111\ldots\boxed{1}],\ B_{HL} = [\boxed{1}\ldots111],\ B_H = \begin{bmatrix} 1 & . & . & . & 1 \\ . & . & \boxed{1} & . & . \\ 1 & . & . & . & 1 \end{bmatrix}$$

(AH بالای B_{HR} و B_{HL}؛ $0.5AH+1$ بالای B_H؛ $0.2AH+1$ کنار B_H)

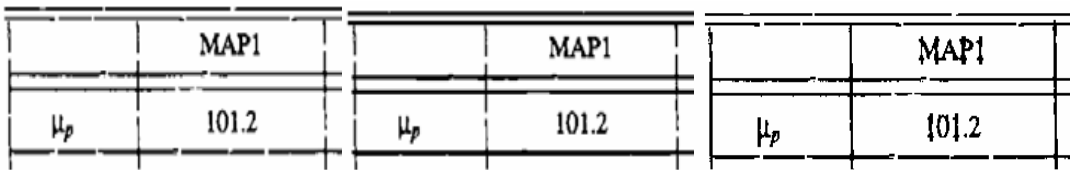
(۴-۴)

$$IM_V = IM \cup (((IM \Theta B_{VD}) \cup (IM \Theta B_{VU})) \oplus B_V) \qquad (۵-۴)$$

$$B_{VD} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ . \\ . \\ \boxed{1} \end{bmatrix},\ B_{VU} = \begin{bmatrix} \boxed{1} \\ . \\ . \\ 1 \\ 1 \\ 1 \end{bmatrix},\ B_V = \begin{bmatrix} 1 & . & 1 \\ . & . & . \\ . & \boxed{1} & . \\ . & . & . \\ 1 & . & 1 \end{bmatrix}$$

($1.2\,AH$ کنار B_{VD} و B_{VU}؛ $0.2AH+1$ بالای B_V؛ $0.5AH+1$ کنار B_V)

(۶-۴)



(الف) (ب) (ج)

شکل ۴-۴: (الف)تصویر دو سطحی اولیه، (ب)تصویر IM_H ترمیم خطوط افقی، (ج)تصویر IM_V ترمیم خطوط عمودی

در نهایت تصاویر IM_H و IM_V را با هم جمع می کنیم که حاصل آن تصویر دو سطحی اولیه است که خطوط افقی و عمودی آن ترمیم شده اند. حالا برای اینکه خطوط افقی و عمودی مناسب برای تشخیص جدول را آشکار سازی کنیم و موقعیت آنها و مختصات ابتدا و انتها و طول آنها و زاویه ی خطوط را به دست آوریم از تبدیل هاف استفاده می کنیم. به این ترتیب که ابتدا ماتریس هاف (H) را از روی بردارهایی در فضای $\rho\theta$ برای تصویری که خروجی مرحله‌ی قبل بود، به دست می آوریم. ماتریس هاف ، مقادیر سلول های انباشتگر را در فضای $\rho\theta$ را نشان می دهد. اولین مرحله در استفاده از تبدیل هاف برای تشخیص خط و پیوند دادن ، یافتن سلول های انباشتگر با تعداد زیاد است. به دلیل کمی سازی در فضای پارامتر تبدیل هاف ، و این حقیقت که لبه ها در تصاویر ، کاملا مستقیم نیستند، قله های تبدیل هاف معمولا در بیش از یک سلول تبدیل هاف قرار می‌گیرند. بنابراین باید تعداد معینی از قله ها را بیابیم. وقتی تعدادی از قله ها در تبدیل هاف مشخص شدند ، باید تعیین کنیم که آیا قطعات خط معنا داری از آن قله ها وجود دارد یا خیر ، علاوه بر این باید تعیین کنیم که شروع و پایان خطوط در کجاست. در ادامه قسمتی از کد و توابعی که با کمک آنها خطوط اسناد را استخراج کرده ایم ، را مشاهده می کنیم.

```
[H, theta, rho]=Hough (g4,'thetaResolution', 0.1);
Peaks=houghpeaks (h, 500, 'Threshold', fact*max (h (:)));
Lines=houghlines (g4, theta, rho, peaks);
```

خروجی تابع houghlines یک ساختار است که طول آن برابر با تعداد خطوط پیدا شده است و هر عنصر این ساختمان ،یک خط را مشخص می‌کندکه دارای فیلدهای [x1,y1] و [x2,y2] که مختصات ابتدا و انتهای هر خط است. حالا با داشتن مختصات ابتدا و انتهای هر خط طول(L) ، و زاویه ای که خط های پیدا شده با محور xها می‌سازند را به راحتی پیدا می‌کنیم و برای تمام خطوط در متغیری ذخیره می کنیم.

$$L = \sqrt{(x2 - x1)^2 + (y2 - y1)^2} \qquad \text{و} \qquad \theta = Arc\tan(\frac{|y2 - y1|}{|x2 - x1|}) \qquad (۷-۴)$$

با توجه به اینکه حالا هم طول خط ها و هم زاویه‌ی آنها را داریم با درنظرگرفتن آستانه هایی برای زاویه و طول خط‌ها فقط خط های عمودی وافقی را که طول آنها از یک حدی بزرگتر است را حفظ می کنیم و بقیه‌ی خطوط را حذف می‌کنیم. در شکل(۴-۵) نمونه ای از استخراج خطوط افقی و عمودی را مشاهده می‌کنیم.

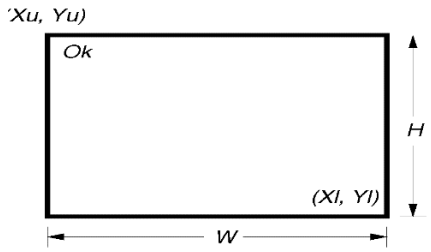
۴-۳-۳ به دست آوردن مؤلفه های پیوسته

به نظر می‌رسد بعد از به دست آورن تصویر دو سطحی مناسب، مهمترین مرحله‌ی تحلیل ساختار اسناد به دست آوردن مؤلفه های پیوسته و قطعه بندی کردن اولیه‌ی تصویر به وسیله‌ی ترسیم مستطیل های محیطی در اطراف مؤلفه های پیوسته است. با استفاده از توابعی مؤلفه های پیوسته تصویر دو سطحی را به دست می آوریم.

شکل ۴-۵: نمونه ای از آشکار سازی خطوط افقی و عمودی در اسناد(خطوط جهت دید بهتر ضخیم شده اند)

:Bwconncomp

این تابع تصویر دو سطحی و نوع همسایگی را به عنوان ورودی می‌گیرد و درخروجی مؤلفه های پیوسته را به همراه پیکسل‌های هر مؤلفه‌ی پیوسته ، تعداد مؤلفه ها، اندازه‌ی تصویر و نوع همسایگی که انتخاب کرده ایم را به ما می‌دهد. حالا برای ترسیم مستطیل های محیطی در اطراف مؤلفه های پیوسته یا حداقل مختصات دو نقطه از مستطیل را باید داشته باشیم یا اینکه مختصات یک نقطه به همراه طول وعرض مستطیل را داشته باشیم. با توجه به اینکه مختصات تمام پیکسل ها به تفکیک هر مؤلفه را داریم می توانیم به راحتی مستطیل محیطی را ترسیم کنیم.



شکل ۴-۶ : مستطیل های محیطی

البته ما به دلیل دقیق تر شدن ترسیم مستطیل ها از تابع Regionprops استفاده می کنیم. مؤلفه‌های پیوسته را به عنوان ورودی به تابع می دهیم و یکی از فیلدهای تابع که BoundingBox است را انتخاب می کنیم تا خروجی تابع به ما مشخصات مستطیل های محیطی را به ازای هر مؤلفه ی پیوسته بدهد . خروجی تابع یک ساختاری است که مختصات رأس بالا سمت چپ مستطیل و طول و عرض آن را به ما می دهد. نمونه ای از ترسیم مستطیل های محیطی سند ورودی را در شکل ۴-۷ می بینیم.

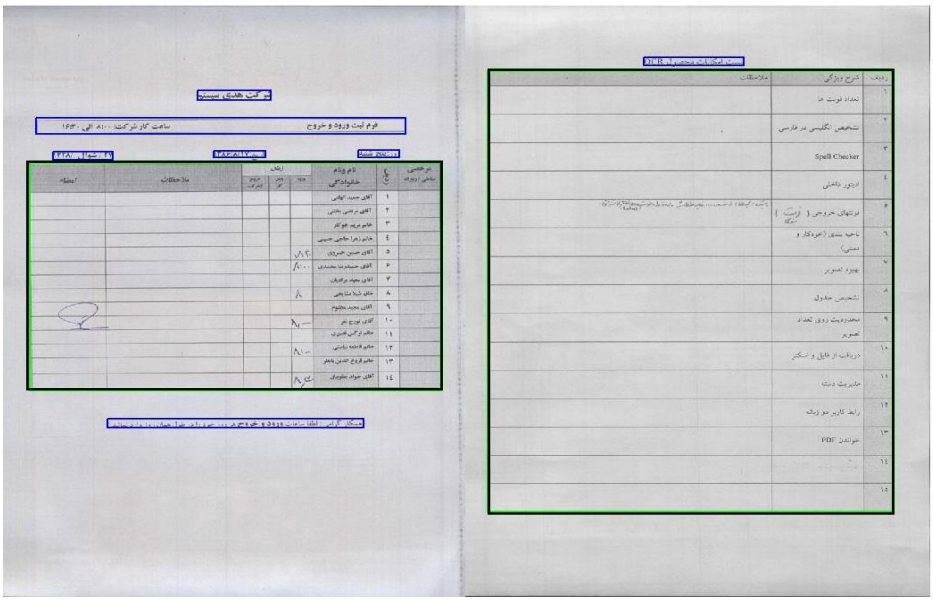
شکل ۴-۷:نمونه ای از ترسیم مستطیل های محیطی

۴-۳-۴ حذف نویز ، شناسایی جدول های بزرگ

حالا با داشتن مستطیل های محیطی ، از این به بعد ما فقط با این مستطیل ها برای ناحیه بندی کار می کنیم. دراین مرحله‌ابتدا اقدام به حذف قسمتی از نویزها می کنیم ، مواردی از مصداق های نویز عبارتند از:مؤلفه هایی را که خیلی کوچک هستند مثلا مستطیل هایی با عرض (ارتفاع فونت) کوچکتر از ۲/۵ ، مستطیل هایی که نسبت عرض به طول آن بزرگتر از ۱۰ باشند ، مساحت مستطیل ها کمتر از ۹ باشد ، مستطیل هایی که عرض کوچکتر مساوی۳ داشته باشند و طول خیلی بزرگی نسبت به عرض داشته باشند ، مستطیل هایی که طول یا عرض آنها بزرگتر از ۰/۹ طول یا عرض سند باشد ...

برای حذف هر نویز طول و عرض مستطیل متناظر با آن را صفر می کنیم و بعد از این قبل از هر اقدامی روی مؤلفه های پیوسته صفر نبودن طول و عرض مستطیل آن را بررسی می‌کنیم. در برخی اسناد جدول های بزرگی وجود دارد که شاید در روال طبیعی الگوریتم آنها به اشتباه شکل یا اینکه به عنوان نویز در نظر گرفته شوند برای اینکه این اشتباه انجام نشود در این مرحله مؤلفه های پیوسته‌ی بزرگ را با خط هایی را که در مرحله‌ی قبل آشکار شده اند مقایسه ای انجام می دهیم تا اگر جدول بزرگ در سند وجود

دارد به درستی تشخیص داده شود. در این مرحله فقط برای مؤلفه های پیوسته‌ی بزرگ که فاصله‌ی کمی با حداقل یکی از خط های آشکار سازی شده دارند یعنی مشکوک به جدول است یک ویژگی دیگر را محاسبه می کنیم. در این موارد نسبت پیکسل های سیاه به کل پیکسل های مؤلفه‌ی پیوسته را محاسبه می‌کنیم اگر این نسبت از یک مقدار آستانه کمتر بود به طور قطع آن را جدول در نظر می گیریم. (شکل ۴-۸ را ببینید).



شکل۴-۸ :تشخیص جدول های بزرگ در سند

۴-۴ ادغام اولیه‌ی مؤلفه ها در جهت افقی

در این مرحله هر مؤلفه‌ی پیوسته با تمام مؤلفه ها مقایسه می شود که البته مستطیل های مؤلفه ها با یکدیگر مقایسه می شوند.

۴-۴-۱ ادغام مستطیل های متداخل

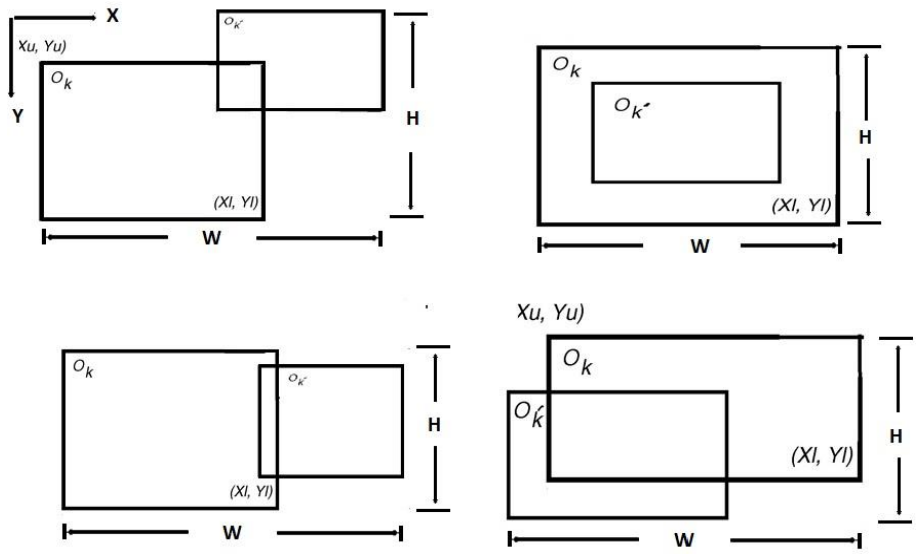
مستطیل های محیطی که اطراف مؤلفه های پیوسته را فرا گرفته اند، به دلیل چهار گوش و قائم الزاویه بودن هر چهار زاویه‌ی آن به ناچار حالت هایی به وجود می آید که این مستطیل ها با یکدیگر تداخل

دارند. در این مرحله که قبل از برچسب زنی است و می خواهیم در جهت افقی ادغام اولیه ای برای مستطیل های نزدیک به هم انجام دهیم. در حین عملیات ادغام باید تمام حالت های تداخل مستطیل ها را در نظر بگیریم. حالت هایی از تداخل مستطیل ها را در شکل ۴-۹ مشاهده می کنیم.

۴-۴-۲ ادغام مستطیل های غیر متداخل

اگر دو مستطیل که مورد مقایسه با هم قرار می گیرند با هم تداخل نداشتند، با داشتن یک سری شرایط با هم ادغام می شوند. ابتدا فاصله ی افقی دو مستطیل را به دست می آوریم، سپس مختصات مراکز مستطیل ها را به دست می آوریم و فاصله‌ی عمودی مراکز مستطیل ها را به دست می آوریم ، و همچنین فاصله ی عمودی یکی از اضلاع افقی در دو مستطیل را نسبت به هم به دست می آوریم اگر این فاصله هایی که به دست آوردیم کمتر

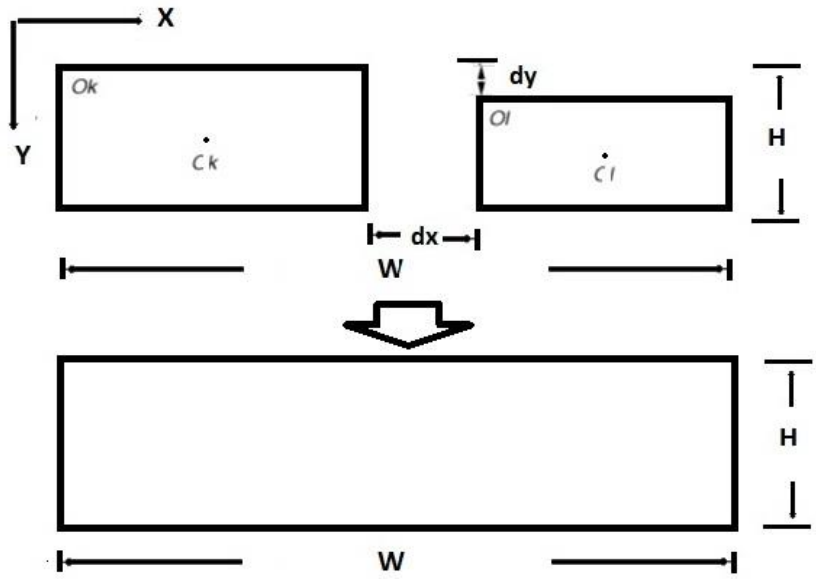
از یک مقدار آستانه‌ای مثلا ۱۰ بود این دو مستطیل با یکدیگر ادغام می شوند و تبدیل به یک مستطیل می شوند. که حالا باید طول و عرض و مختصات رأس بالا چپ مستطیل جدید را به دست آوریم. و این مقادیر جدید را جایگزین مقادیر قدیمی یکی از مستطیل ها کنیم و مستطیل دیگر را با صفر کردن طول و عرض آن حذف کنیم.(شکل ۴-۱۰ را ببینید.)



شکل۴-۹ :حالت هایی از تداخل مستطیل ها

بعد از این مرحله که ادغام اولیه در جهت افقی البته انجام شده است مستطیل های جدید و بزرگتری به وجود می آید. قبل از انجام کار دیگری بر روی این مؤلفه ها، باید ماهیت آنها و اینکه به کدام دسته از خروجی مورد نظر ما (جدول،متن، شکل) تعلق دارد را مشخص کنیم. قبل از انجام عملیات برچسب زنی بلوک های به دست آمده یک گام دیگر اقدام به حذف نویز می کنیم. چون خطوط متنی که کوچکترین ارتفاع را دارند مشخص شده اند و ارتفاع آنها حداقل ۹ یا ۱۰ است مستطیل هایی که ارتفاعی کمتر از مثلا ۶ دارند را حذف می کنیم. همچنین بلوک هایی که ارتفاع بزرگی دارند(مثلا ۶۰۰) و نسبت ارتفاع به طول مستطیل بزرگ باشد (مثلا بزرگتر از ۳) و جدول هم نباشد(در مرحله ی قبل جدول بزرگ تشخیص داده نشده باشد) آن را حذف می کنیم.

بلوک هایی که طول بزرگی دارند (مثلا ۷۰۰) و نسبت طول به ارتفاع مستطیل باشد (مثلا کوچکتر از ۱/۲) و جدول هم نباشد، حذف می شود. اگر طول بلوک از یک مقدار آستانه ای(مثلا ۱۵) کوچکتر باشد حذف می شود.



شکل۴-۱۰ :ادغام افقی مستطیل ها

۴-۵ عملیات بر چسب زنی

یکی از ویژگی هایی که در هر سه گروه اشیاء(متن ، جدول ، شکل) برای شناسایی مورد استفاده قرار می گیرد و ویژگی خوب و قابل استنادی هست نسبت پیکسل های سیاه به کل پیکسل ها برای هر کدام از مستطیل های محیطی است .اینکه چرا در همان مرحله ی نخست ما این نسبت را به دست نیاوردیم به این دلیل است که تعداد مستطیل های محیطی اولیه خیلی زیاد است واگر بخواهیم برای هر کدام یک حلقه به تعداد کل پیکسل های آن بزنیم مدت زمان اجرای برنامه خیلی طولانی می شود. پس ویژگی اول که برای تمام مستطیل های محیطی مؤلفه های پیوسته به دست می آوریم، نسبت پیکسل های سیاه به کل پیکسل های همان مستطیل است. ویژگی بعدی که به ما برای شناسایی کمک می کند، استفاده از طول ، عرض و نسبت طول به عرض مستطیل ها است . در متن ها که هنوز فقط به صورت افقی ادغام شده اند باید مستطیل هایی با طول نسبتا بزرگی باشند (یا نسبت طول به عرض آن ها عدد بزرگی باشد).

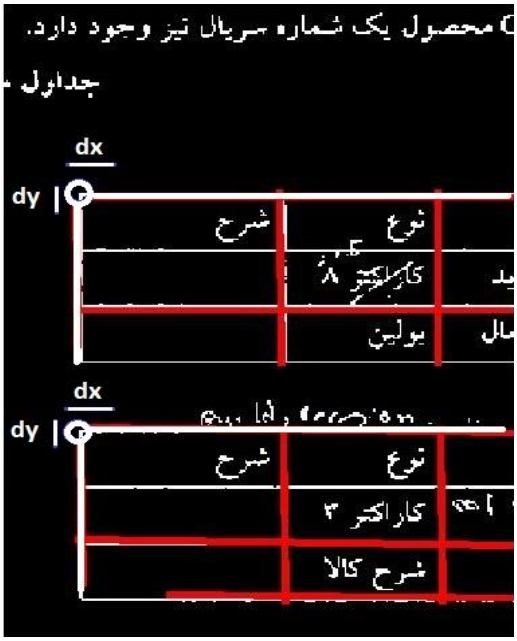
اگر این ویژگی را قبل از ادغام اولیه ی افقی به دست آوریم چون مستطیل ها هنوز کوچک هستند و زیاد شکل متنی نگرفته اند ممکن است در تشخیص ماهیت مستطیل ها اشتباه زیاد صورت گیرد.

حالا با استفاده از ویژگی های ذکر شده و ترکیب آنها (AND , OR منطقی) عملیات برچسب زنی را آغاز می کنیم. در این پایان نامه ما جدول ها را با رنگ سبز(برچسب۳) متن ها را با رنگ آبی(برچسب۱) و شکل ها را با رنگ قرمز(برچسب۲) مشخص می‌کنیم. قبل از شروع برچسب زنی تمام بلوک ها را برچسب صفر زده ایم ، (البته به جزء جدول بزرگ که قبلا تشخیص داده شده اند و برچسب۳ دارند)یعنی اگر بلوکی در هیچ کدام از دسته‌های زیرکه بررسی می شوند قرار نگیرند با همان برچسب صفرکه مفهوم مبهم بودن را دارند باقی می ماند.

۴-۵-۱ شناسایی جدول

جهت شناسایی جدول از خطوط افقی و عمودی که در مراحل قبلی آشکار سازی شده است، استفاده می کنیم. هر کدام از بلوک های تشکیل شده را با تمام خط ها مقایسه می‌کنیم، البته اگر کوچکترین جدول را دو ردیفه درنظر بگیریم یک مقدار حداقل برای ارتفاع مستطیل هایی که ممکن است جدول باشند در نظر می گیریم بنابراین برای شناسایی جدول ها فقط مستطیل هایی که ارتفاع آنها از یک مقدار آستانه ای (مثلا ۴۰) بزرگتر است را در نظر می گیریم. برای جدول بودن مستطیل ها حتما باید یکی از خط های آشکار شده منطبق بر طول یا عرض مستطیل مورد بررسی باشد یا اینکه خط قسمتی از طول یا عرض مستطیل باشد. واین مسأله را اینگونه مشخص می کنیم که به ازای هر مستطیل فاصله ی راس بالا چپ مستطیل را از دو سر تمام خط ها به دست می آوریم اگر یکی از این فاصله ها از یک حدی (مثلا ۱۰) کمتر باشد(شکل۴-۱۱) ، شرط بعدی را بررسی می کنیم و آن پیدا کردن حداقل دو تا خط عمودی یا افقی جدول است، که در بین خط های قبلا آشکار شده باشند (شکل۴-۱۲) البته برای اینکه این دو خط

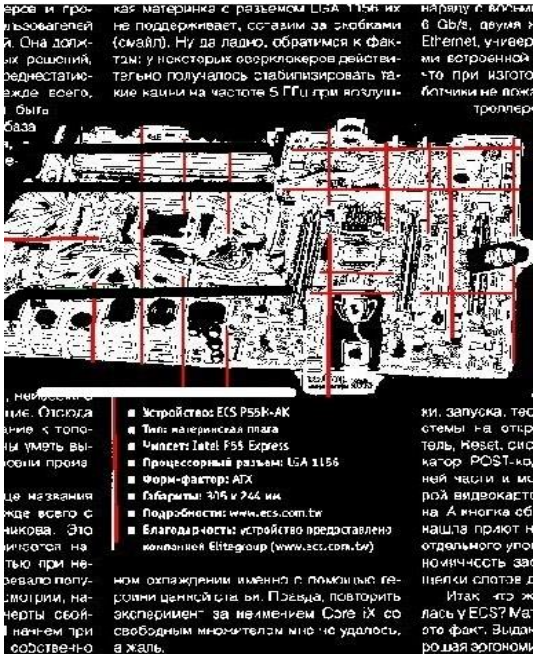
را با خطوطی که مربوط به قسمت دیگری از سند مثل خط های شکل ها هستند اشتباه نشوند یک شرایط خاصی نیز برای آنها قرار می دهیم.



شکل۴-۱۱ :پیدا کردن خطی منطبق بر یکی از اضلاع جدول

شکل۴-۱۲:پیدا کردن ۲خط عمودی یا افقی از خط های جدول(جهت دید بهتر خطوط ضخیم شده اند)

(شکل۴-۱۳)مثلا اگر خط پیدا شده ای‌که منطبق بر ضلع مستطیل است عمود باشد آن دو خط آشکار شده‌ی جدول باید حتما سمت راست‌آن‌باشند. واگر خط منطبق بر ضلع مستطیل افقی باشد ،آن دو خط آشکارشده‌ی جدول باید حتما پایین تر از آن باشند. بعد از بررسی کردن برقراری این شرایط، شرط نسبت



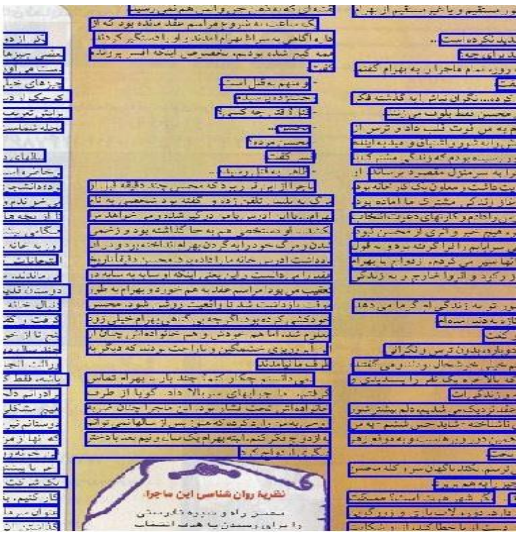
شکل ۴-۱۳ :تشخیص اشتباه خط های شکل به عنوان خط های جدول(جهت دید بهتر خطوط ضخیم شده اند)

تعداد پیکسل های سیاه به تمام پیکسل های مستطیل را بررسی می کنیم، که برای جدول این نسبت باید از یک مقدار آستانه (مثلا ۰/۳) کمتر باشد. البته موارد خاصی از جدول هایی که زمینه ی گرافیکی دارند هستند که این نسبت برای آنها مقداری بزرگ است.

۴-۵-۲ شناسایی متن

در این مرحله شناسایی متن ها خیلی آسان تر از مراحل قبل است چون در مراحل قبل که هنوز ادغام افقی صورت نگرفته است و مستطیل های اولیه خیلی کوچک هستند (مثلا مستطیل های اطراف حروف مجزا یا اطراف نقاط حروف ...) و ممکن است نوع آنها اشتباه در نظر گرفته شود. همچنین بسیاری از مستطیل ها قبل از ادغام با هم تداخل دارند. و خیلی آسان تر از مراحل بعد است، چون در مراحل بعد

ادغام در جهت عمودی صورت می گیرد و ارتفاع بلوک ها نیز افزایش می یابد، و تمایز آنها از شکل ها و جدول ها خیلی سخت تر می شود. در این مرحله فقط طول مستطیل ها بزرگ است و ارتفاع نسبتا کوتاهی دارند (چون در این مرحله هر خط از متن ها به تنهایی در نظر گرفته شده). پس یکی از شرایط متن بودن بلوک ها این است که ارتفاع آنها از یک مقدار آستانه‌ای(مثلا ۴۰) کمتر باشد، نسبت طول به عرض باید از یک مقدار (مثلا ۲) بیشتر باشد ، طول مستطیل باید از یک مقدار (مثلا ۱۷) بیشتر باشد. شرط مهمی که باید برقرار باشد نسبت پیکسل های سیاه در مستطیل مورد نظر است که باید از یک مقداری (مثلا ۰/۳) کمتر باشد. این شرایط باید AND منطقی شوند. حالت های استثنایی در متن ها نیز وجود دارد اگر مثلا نسبت پیکسل های سیاه از آستانه‌ی ما برای متن بودن بیشتر باشد ما برای جبران این کمبود در شرط نسبت طول به عرض مستطیل مقدار آستانه‌ی آن را افزایش می دهیم که این شرایط جدید باید با تمام شرایط قبلی OR منطقی شوند(شکل۴-۱۴).



شکل۴-۱۴ :نمونه ای از تشخیص بلوک های متنی

۴-۵-۳ شناسایی شکل

در بررسی بلوک های ایجاد شده اگر بلوک مورد نظر در دو قسمت قبلی برچسب نخورد و همچنان برچسب صفر داشته باشد، در این قسمت بررسی می شود. اگر شرایط مدنظر ما را داشته باشد برچسب ۲

می خورد یعنی شکل تشخیص داده می شود. شرطی که برای شکل بودن مهم است نسبت پیکسل های سیاه بلوک ها است، به دلیل اینکه معمولا شکل ها بزرگ هستند طول مستطیل باید از یک مقداری بزرگتر باشد ، به همین دلیل نیز باید ارتفاع بزرگتری از یک مقدار آستانه (مثلا ۳۰) داشته باشند. در حالتی که نسبت پیکسل های سیاه مقدار تقریبا بزرگی است می توانیم مقدار آستانه‌ی طول و ارتفاع مستطیل را کوچکتر در نظر بگیریم.

با اینکه ویژگی نسبت پیکسل های سیاه ویژگی خوبی است نقطه‌ی ضعف هم دارد که ما را در تشخیص ، دچار اشتباه می کند. مثلا در مواردی که متن ما دارای فونت بزرگ یا ضخیم یا نوع هنرمندانه است (عنوان ها)، این نسبت مقدار زیادی دارد. و چون معمولا دارای طول کوچک و نسبت طول به عرض کوچکی است به اشتباه شکل در نظر گرفته می شود(شکل۴-۱۵).

شکل۴-۱۵:تشخیص اشتباه متن های گرافیکی به عنوان شکل(موارد اشتباه شده با بیضی مشخص شده)

۴-۶ ادغام نهایی بلوک های هم نوع در جهت افقی

با اینکه ادغام افقی در یک مرحله انجام شده است، باز هم درصد بالایی از مستطیل هایی در مجاورت یکدیگر وجود دارند که باید با هم ادغام می شده اند ولی نشده اند. بنابراین برای کامل شدن ادغام در جهت افقی در یک مرحله‌ی دیگر نیز تمام بلوک ها با یکدیگر مقایسه می شوند. در این مرحله چون یک

بار قبلا عمل ادغام صورت گرفته است تعداد بلوک ها به طور چشمگیری کاهش می یابد (مثلا در یک مورد از ۴۷۸۵ به ۱۴۸ کاهش یافته است). بنابراین مدت زمان زیادی برای اجراء به خود اختصاص نمی دهد(شکل۴-۱۶). ادغام در این مرحله با قبل تفاوت دارد، اینجا مقدار آستانه ای را که برای فاصله‌ی افقی جهت ادغام شدن در نظر می‌گیریم را کمی افزایش داده ایم تا احتمال ادغام در جهت افقی بالاتر برود. تفاوت دیگری که دارد ما در اینجا شرط هم نوع بودن بلوک ها را به شرط های ادغام افزوده ایم یعنی دو بلوک در صورتی با هم ادغام می شوند که هر دو مثلا متن یا شکل باشند. البته در این مرحله قبل از ادغام دو بلوک ما این مسآله را بررسی می کنیم که اگر یکی از این بلوک ها داخل بلوک دیگر باشد و هر دو نیز از یک نوع باشند بلوک داخلی را نیز حذف می کنیم.

ج

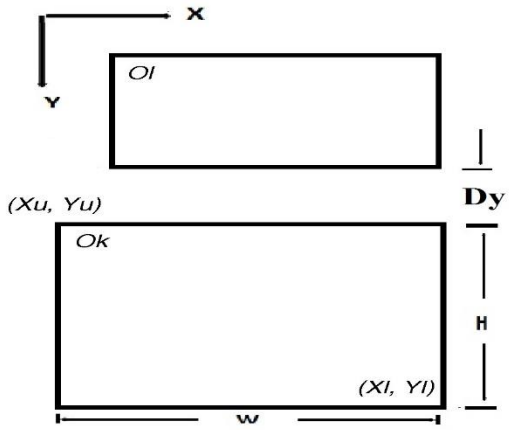
شکل۴-۱۶ :ادغام اولیه ی افقی(سمت چپ)- ادغام نهایی افقی(سمت راست)

۴-۷ ادغام بلوک ها در جهت عمودی

قصد داریم در این مرحله بلوک ها را در جهت عمودی ادغام کنیم. البته واضح است که بلوک های مورد بررسی ما که تعداد آنها خیلی کمتر شده است ، جهت ادغام باید هم نوع باشند. این مرحله خود به ۳ گام تقسیم می شوند ،که در ادامه شرح داده خواهد شد.

۴-۷-۱ ادغام در جهت عمودی: گام۱

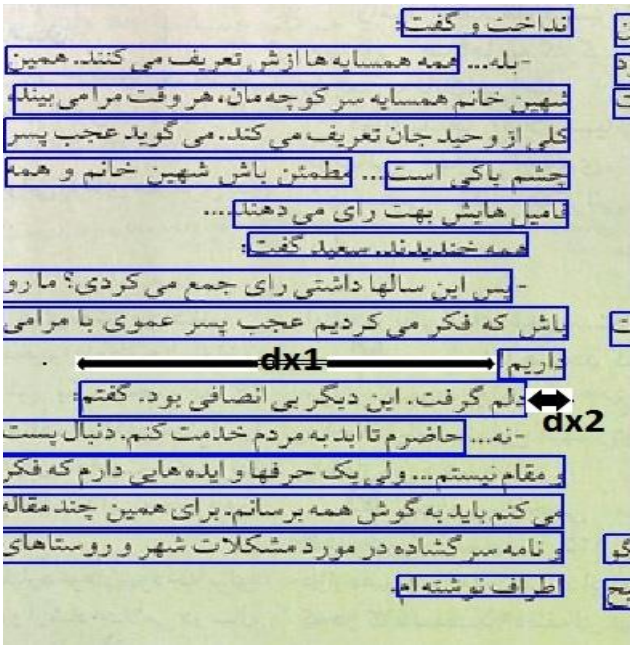
در این گام نیز باید تمام بلوک ها را با هم مقایسه کرده و در صورتیکه شرایط در نظر گرفته شده را داشته باشند دو بلوک مورد بررسی را با هم ادغام می کنیم. با شروع مقایسه قبل از بررسی شرایط ادغام ابتدا یک سری شرایط احتیاطی را بررسی می کنیم. اول از همه باید مطمئن شویم که این دو بلوک داخل همدیگر نیستند. دوم اینکه دقیقا مانند قسمت **۴-۴-۱** اگر بلوک هایی باشند که با یکدیگر تداخل داشته باشند (شکل۴-۹)، به همان روش آنها را در جهت افقی ادغام می کنیم. بعد از در نظر گرفتن این موارد شروع به بررسی شرایط ادغام دو بلوک می کنیم. ابتدا فاصله ی عمودی بین دو بلوک را به دست می آوریم(شکل۴-۱۷).



شکل۴-۱۷: محاسبه ی فاصله ی عمودی بین دو بلوک(Dy)

جهت انجام ادغام، فاصله عمودی دو بلوک هم نوع باید کمتر از مقدار آستانه (مثلا ۱۴) باشد. اگر دو بلوک از نوع شکل باشند این مقدار آستانه بزرگتر (مثلا ۲۰) می شود. شرط دیگری که برای ادغام باید برآورده شود کمتر بودن فاصله ی افقی بین خط های متناظر عمودی در دو بلوک از یک مقدار آستانه است (منظور از فاصله ی افقی همان dx2وdx1 در شکل۴-۱۸است). این شرط و مقدارآستانه‌اش در تحلیل ساختاری درست اسناد عامل خیلی مهمی است. البته در این گام چون اندازه‌ی فاصله عمودی برای ادغام

را کوچک گرفتیم ، مقدار آستانه را عدد بزرگی (مقدار ماکزیمم طول دو بلوک)در نظرگرفتیم. مفهوم آن این است که اگر فاصله ی عمودی دو تا بلوک از هم خیلی کم باشد آنگاه اندازه ی dx1 و dx2 حداکثر می توانند تا مقدار max(w1,w2) باشد. به دلیل پرهیز از چسبیدن دو جدول مجزا در این گام فرض کردیم که بلوک های مورد بررسی جدول نباشند. شرط مهم دیگری که در این گام جهت ادغام عمودی دو بلوک بررسی می شود ، این است که مساحت بلوک جدید که از ادغام دو بلوک حاصل می شود باید کمتر از دو برابر مجموع مساحت های دو بلوک قدیم باشد. نمونه ای از عدم ادغام را در شکل۴-۱۹ می بینید که اگر دو بلوک مشخص شده با هم ادغام می شدند، مساحت ناحیه ی جدید از دو برابر مجموع مساحت های دو ناحیه ی فدیم بزرگتر می شد.



شکل۴-۱۸: محاسبه ی فاصله ی افقی بین بین خط های متناظر عمودی (Dx1,Dx2)

شکل۴-۱۹: نمونه هایی‌از عدم ادغام عمودی نواحی(A با Cو A با B و A با D)به‌دلیل بیش‌از حد بزرگ شدن مساحت ناحیه‌ی جدید

حالا در پایان بررسی شرایط گام۱، نتیجه را تا این مرحله در شکل۴-۲۰ می‌بینید.

۴-۷-۲ ادغام در جهت عمودی: گام۲

با اینکه ما یک گام در ادغام عمودی پیش رفته ایم، ولی چون آستانه‌ی فاصله‌ی عمودی بین دو بلوک جهت ادغام را عددی کوچک گرفته بودیم تعدادی از ادغام های عمودی که باید صورت می گرفت انجام نشده است. به این دلیل کوچک گرفتیم که آستانه‌ی فاصله‌ی افقیdx1 وdx2 را عدد بزرگی گرفته بودیم. (شکل ۴-۱۸ را ببینید) حالا ادامه‌ی ادغام عمودی را در این گام انجام می‌دهیم. همان طور که می دانیم با ادغام عمودی که در گام قبل داشتیم، ارتفاع بلوک های سند بلندتر شده است، به همین دلیل ممکن است مواردی در سند پیش بیاید که یک بلوک داخل دیگری باشد یا اینکه بین دو بلوک تداخل ایجاد شود. مانند مراحل قبل تمام بلوک ها باید با همدیگر مقایسه شوند، تا اگر دو بلوک هم نوع شرایط ادغام را داشتند با هم ادغام شوند، قبل از بررسی شرایط ادغام اگر یکی از بلوک ها داخل دیگری باشد بلوک داخلی را حذف می‌کنیم(با شرط هم نوع بودن) همچنین اگر دو بلوک با هم تداخل داشته باشند با رعایت شرایطی که در مراحل قبل ذکر شد ادغام در جهت افقی انجام می شود. فاصله ی عمودی دو بلوک

شکل۴-۲۰: خروجی گام۱ در ادغام عمودی

را مانند شکل۴-۱۷ به دست می آوریم، جهت ادغام این فاصله باید کمتر از مقدار آستانه‌ای باشد که نسبت به گام قبل بیشتر شده است(مثلا ۳۶). این آستانه را به این دلیل چند برابر کردیم که بلوک های باقیمانده ادغام شوند و دیگر بلوکی باقی نماند. اما نکته ای که اینجا هست باید مراقب باشیم که بلوک ها دارای طولی تقریبا برابر باشند و فواصل dx1 و dx2 زیاد بزرگ نباشند(مثلا ۱۵). شرط دیگری که در این گام باید با احتیاط بیشتری از آن استفاده کرد مساحت ناحیه‌ی جدید است که نباید از ۱/۵ برابر مجموع مساحت های دو ناحیه‌ی فدیم بزرگتر شود. دلیل احتیاط بیشتر در استفاده از این شرط این است که در متن های چند ستونه احتمال ادغام بلوک هایی از دو ستون مختلف وجود دارد. البته این مسأله را با آستانه ی فواصل dx1 و dx2 کنترل می کنیم. در ضمن جهت جلوگیری از چسبیدن دو جدول ، بلوک های مورد بررسی نباید از نوع جدول باشند. نمونه ای از خروجی این گام را در شکل۴-۲۱ می بینیم.

شکل۴-۲۱ :خروجی گام۲ در ادغام عمودی

۴-۷-۳ ادغام در جهت عمودی: گام۳

این گام بیشتر جنبه‌ی احتیاطی دارد تا اگر بلوک هایی باقیمانده باشند که ادغام نشده اند دیگر در این مرحله کار ادغام عمودی وکل تحلیل سختاری سند به اتمام برسد و خروجی نهایی را در این مرحله به دست می‌آوریم. دقیقا شرایط گام های قبل را در این مرحله اعمال می کنیم ، با این تفاوت که مقادیر آستانه‌ی قبلی را خیلی کوچکتر می گیریم ، به این دلیل که بلوک ها نسبت به قبل خیلی بزرگتر شده اند و اگر اشتباه ادغام شوند کل ناحیه بندی سند و تحلیل ساختاری آن دچار مشکل می شود. خروجی نهایی را در شکل ۴-۲۲ می بینیم.

شکل۴-۲۲ :خروجی گام۳ در ادغام عمودی

۴-۸ نتایج تجربی

این الگوریتم برای تحلیل ساختار اسناد پیچیده ی چند ستونه طراحی شده و برای این نوع متون به خوبی عمل می کند. ضمن اینکه عملکرد آن روی اسناد ساده نیز بسیار خوب است. برای مقایسه با روش های نوین

تحلیل اسناد ، الگوریتم خود را با ۲ نرم افزار تجاری روز مقایسه کردیم:

عملکرد الگوریتم پیشنهادی را با دو نرم افزار OminiPage(نسخه ۱۸) از شرکت Nuance و FineReader

(نسخه ۱۲) از شرکت Abby مقایسه می‌کنیم. ما با استفاده از یک پایگاه داده‌ی مشترک برای هر سه الگوریتم میزان موفقیت هر روش را به تفکیک در سه شاخه‌ی مورد نظر(شکل، جدول و متن) را به صورت کمی به دست آورده ایم که آنها را در جداول ۴-۱ و ۴-۲ ذکر می کنیم. همان طور که از این آمار و ارقام مشخص است الگوریتم پیشنهادی ما در اسناد فارسی که اغلب چند ستونه هم هستند برتری خوبی نسبت به این دو نرم افزار دارد. الگوریتم ما متن های فارسی را با ۷۲ درصد و شکل ها را با ۷۵ درصد و جدول ها

را با ۹۲ درصد به درستی تشخیص می دهد. و همچنین ۸۸ درصد اسناد چند ستونه‌ی فارسی را به درستی ناحیه بندی می‌کند.

جدول ۴-۱ : ارزیابی عملکرد: نواحی که به تفکیک به طور صحیح تشخیص داده شده اند

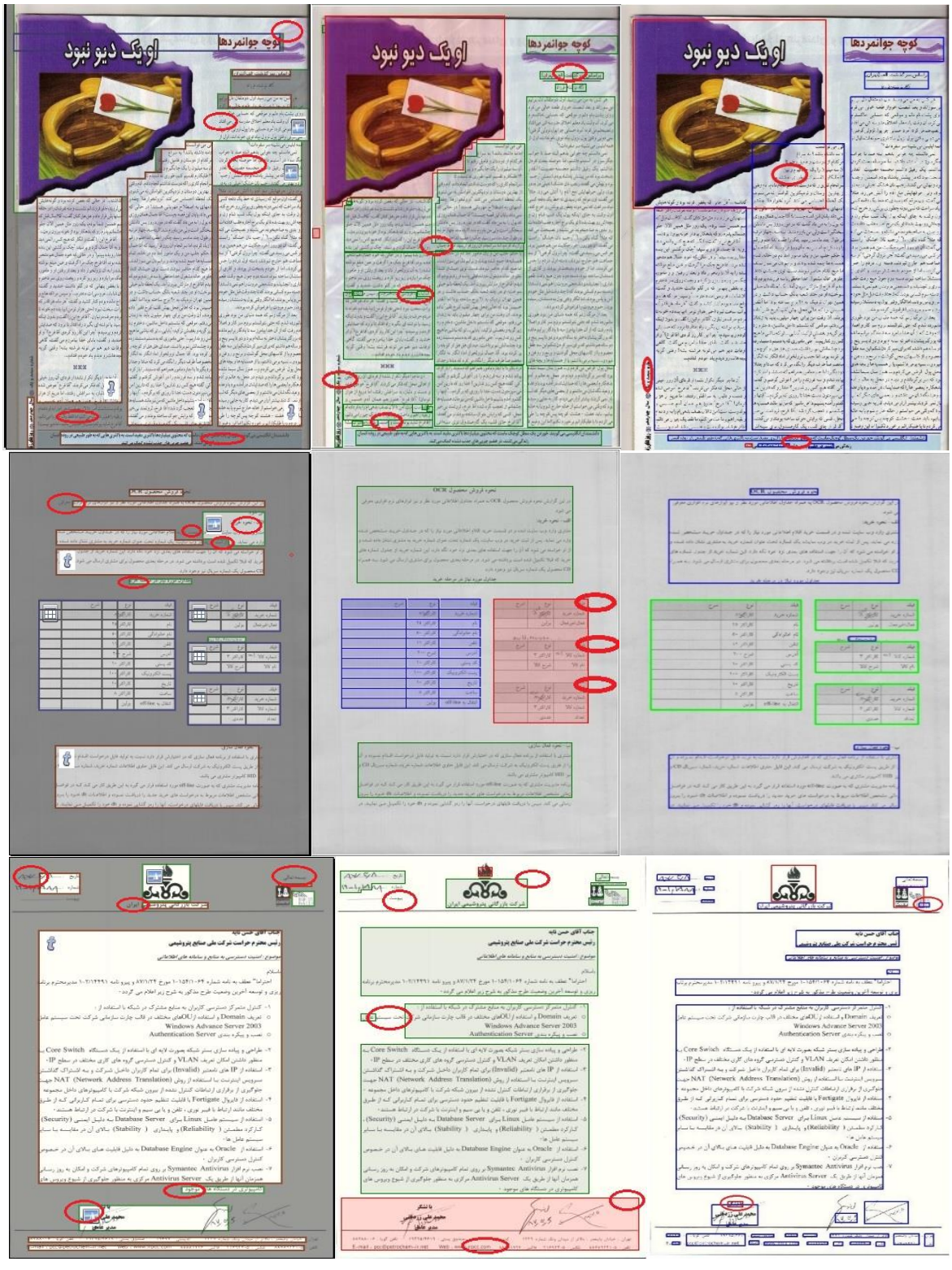
	الگوریتم پیشنهادی			Omnipage18			Finereader12		
	متن	شکل	جدول	متن	شکل	جدول	متن	شکل	جدول
اسناد فارسی(٪)	۷۲	۷۵	۹۲	۲۳	۳۴	۷۷	۴۸	۶۰	۷۷
اسناد لاتین(٪)	۷۶	۷۶	۶۶	۸۵	۸۰	۷۳	۹۱	۸۹	۹۰

جدول ۴-۲ : ارزیابی عملکرد: اسنادی که کاملا بدون خطا ناحیه بندی شده اند

	الگوریتم پیشنهادی	Omnipage18	Finereader12
اسناد فارسی(٪)	۸۸	۴۴	۵۶
اسناد لاتین(٪)	۵۲	۸۴	۸۸

برای مقایسه‌ی الگوریتم ها در جزئیات و دیدن تفاوت ها و نقاط ضعف و قوت الگوریتم ما نسبت به دو نرم افزار قوی ، ۵ سند مختلف را از ۵ منبع مختلف انتخاب کردیم. دو صفحه‌ی ۳ ستونه از یک مجله با زمینه و متن های گرافیکی، دو نامه‌ی اداری که در میان متن آنها آرم و جدول نیز وجو دارد و یک سند ۳ ستونه با شکل هایی در میان متن ها از کتابی غیر فارسی. شکل۴-۲۳ نتایج اعمال ۳ الگوریتم ذکر شده را نشان می دهد. خطاها، یعنی بلوک های آشکار نشده، به غلط آشکار شده و یا ادغام هایی که

نباید صورت می گرفت، توسط یک بیضی قرمز رنگ روی هر تصویر مشخص شده است. همان گونه که در این شکل مشاهده می شود، الگوریتم پیشنهادی در مواردی از دو الگوریتم دیگر بهتر عمل کرده است. FineReader ساختار کلی اسناد پیچیده را بهتر شناسایی می کند ولی مشکلاتی هم دارد: در برخی صفحات قسمتی از بلوک ها اصلا شناسایی نشده اند و در برخی موارد دو ستون جداگانه متنی با هم ادغام شده است یا با بلوکهای تصویر و جدول‌های جداگانه با هم ادغام شده اند. OmniPage هم مشکلاتی دارد: ادغام کردن نواحی متن و تصویر و ادغام دو ستون متنی جداگانه که نباید با هم ادغام شوند و متن ها را در اسنادی که زمینه‌ی گرافیکی دارند اشتباهی شکل یا جدول در نظر گرفته است. الگوریتم ما در مورد سند۱ قسمتی از بلوک متنی را با بلوک شکل ادغام کرده است. و در مورد متن های حاشیه ی سند که به صورت عمودی و فونتی کوچک نوشته است را به عنوان نویز حذف می کند .



شکل۴-۲۳ : نتایج الگوریتمهای تحلیل ساختار روی اسناد آزمایش .از چپ به راست:OmniPage18 و FineReader12و روش پیشنهادی ما (ادامه دارد)

ادامه ی شکل۴-۲۳ : نتایج الگوریتمهای تحلیل ساختار روی اسناد آزمایش .از چپ به راست:OmniPage18 و FineReader12و روش پیشنهادی ما

فصل پنجم:

نتیجه گیری و پیشنهادات

۵-۱ جمع بندی و نتیجه‌گیری

در این پایان نامه به بررسی روش های تحلیل اسناد پیچیده پرداختیم و یک روش برای تحلیل اسناد پیچیده فارسی ارائه دادیم. در روش پیشنهاد شده سند ورودی را با توجه به کیفیت تصویر و رزولوشن تغییر اندازه دادیم. پیش پردازش و دو سطحی سازی روی تصویر انجام دادیم. ابتدا خط ها و ویژگی های آنها (زاویه،طول،مختصات نقاط دو سر آنها) را از تصویر دو سطحی توسط عملگرهای ریخت شناسی و تبدیل هاف استخراج کردیم. سپس مؤلفه های پیوسته را از تصویر دو سطحی به دست آوردیم. ادغام در جهت افقی را در یک مرحله انجام داده ، سپس برچسب زنی مؤلفه ها انجام شده و بعد از آن ادغام نهایی در جهت افقی انجام شد ، در ادامه طی سه گام بلوک های نزدیک به یکدیگر و هم نوع را در جهت عمودی با هم ادغام کردیم تا ساختار نهایی سند تولید شود. این نکته نیز قابل ذکر است که در هر کدام از مراحل بیان شده قبل از ادغام ، اقدام به حذف نویز کردیم. در این پایان نامه جهت شناسایی جدول ها از روش جدیدی استفاده کردیم. این روش بر شناسایی خطوط افقی و عمودی استوار است. ترمیم خطوط افقی و عمودی سند در تصویر دو سطحی با کمک عملیات ریخت شناسی و شناسایی تمام خطوط به همراه ویژگی های آنها (نقاط ابتدا و انتهایی ، طول خط ،زاویه‌ی خط ها) و سپس حذف خطوطی‌که طول آنها کمتر از یک حد آستانه است ، و به دست آوردن خط های افقی و عمودی از بین خط های باقیمانده ، حالا جهت شناسایی جداول اسناد دو ضلع متصل به رأس بالای چپ مؤلفه های پیوسته را با تمام خطوط افقی و عمودی به دست آمده مقایسه می کنیم. این روش شناسایی جدول ها در تمام سندهایی که در اختیار داشتیم به درستی جواب داد.

در مقایسه ای که بین خروجی های روش پیشنهاد شده و دو نرم افزارتجاری پیشرو در زمینه‌ی OCR یعنی OmniPage و FineReader که البته سالهاست در این حوزه مشغول فعالیت اند ، انجام دادیم. روش ارائه شده عملکرد قابل قبولی از خود نشان داد. از نقاط قوت و مزایای الگوریتم پیشنهادی نسبت به

این دو نرم افزار تجاری می توان به چند مورد اشاره کرد: الگوریتم پیشنهادی عملکرد به نسبت بهتری در اسناد چند ستونه و پیچیده دارد و در تعداد کمتری از اسناد چند ستونه می توان یافت که این الگوریتم ستون های جدا را با یکدیگر ادغام کرده باشد. ولی این مسأله در نرم افزارهای مورد مقایسه در تعداد زیادی از اسناد به اشتباه دو ستون مجاور متنی را با هم ادغام کرده اند. مزیت بعدی الگوریتم پیشنهادی که به وضوح در موارد مقایسه شده آشکار است ، در اسنادی است که زمینه ی یکنواخت ندارند و بعضا دارای پس زمینه هستند. در این موارد الگوریتم پیشنهادی مثلا متن ها را به خوبی تشخیص می دهد این در حالی است که دو نرم افزار ذکر شده در تعداد زیادی از اسناد به اشتباه متن ها را تصویر تشخیص داده است. مزیت بعدی الگوریتم پیشنهادی در تشخیص جدول ها است. در اسنادی که در اختیار داشتیم که محتوی جدول بودند الگوریتم ما در تمام موارد جدول ها را به درستی تشخیص داده است. در صورتیکه همین اسناد را با دو نرم افزار تجاری مورد تحلیل قرار دادیم تعدادی از جدول ها را به درستی تشخیص نداده اند ، و یا به اشتباه جدول ها را به عنوان شکل تشخیص داده اند. به نظر می رسد که دلیل این مزیت یعنی تشخیص خوب جدول ها استفاده مناسب از عملیات ریخت شناسی و تبدیل هاف و در نظر گرفتن شرایطی متعدد و نسبتا خوب در قسمت برچسب زنی جهت تشخیص جدول ها می باشد. یکی از ایراداتی که روش ارائه شده دارد،مقادیر آستانه‌ای است که در برنامه در نظر گرفتیم. با تغییر سند ورودی با ساختار جدید مقادیر آستانه ی جدیدی باید در نظر گرفت ، با مقادیر قبلی ممکن است خروجی الگوریتم جواب مناسبی ندهد ، هر چند با تغییر مقادیر آستانه ممکن است بصورت موردی جواب قابل قبولی دهد.

۵-۲ پیشنهادات

برای حل مشکلات و بهبود الگوریتم، چند پیشنهاد ارائه می شود:

- اگر بتوانیم پارامترهای افقی و عمودی الگوریتم قطعه بندی و ادغام ، مانند آستانه هایی که برای ادغام بلوک ها در جهت افقی و عمودی در نظر می گیریم را به صورت وفقی تعیین کنیم، احتمالا خطای قطعه بندی و ادغام کاهش خواهد یافت.
- با توجه به این که پردازشهای بعدی برای متون دستنویس و چاپی می تواند متفاوت باشد ، خوب است که دو کلاس مجزا برای آنها در نظر گرفت.
- شناسایی متن با رنگ معکوس ، یکی از مشکلاتی که بسیاری از الگوریتم های شناسایی متن با آن روبرو هستند ، این است که نمی توانند متنهای با رنگ معکوس (مانند متن سفید در زمینه سیاه) را شناسایی کنند. برای حل این مشکل می توان لبه های تصویر را استخراج کرد و از ویژگیهای لبه های نویسه ها و زیر کلمات مانند حضور لبه های موازی و یکنواختی رنگ بین آنها برای شناسایی متن استفاده کرد[۱۳].
- بررسی نواحی که برچسب مبهم خورده اند و استفاده از الگوریتم دیگری جهت رفع ابهام آنها.
- شناسایی نواحی جداساز ، نواحی جداساز توسط یک الگوریتم سریع و با پیچیدگی محاسباتی کم از روی افکنش تجمعی عمودی تصویر سیاه و سفید سند و پس از حذف حاشیه های تصویر به دست آیند. آغشته سازی در مرز بین نواحی جداساز اعمال شود تا بلوکهای مستقل مجاور یکدیگر با هم ادغام نشوند.

مراجع

[1]Anil K.Jain, Bin Yu,*" Document Representation and Its Application to Page Decomposition"* , IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 20, NO. 3, MARCH 1998

[۲] ح.خسروی "رویکرد یکپارچه برای بازشناسی متون چاپی فارسی" رساله دکتری ، دانشگاه تربیت مدرس ۱۳۸۷

[3] Bukhari, S. S., Shafait, F., and Breuel, T. M., "*Document image segmentation using discriminative learning over connected components*", in Proc. 9th IAPR Workshop on Document Analysis Systems, pp. 183-190, 2010.

[4] Okun, O., Doermann, D., and Pietikainen, M. "*Page Segmentation and Zone Classification: The State of the Art*". Technical Report LAMP-TR-036, CAR-TR927, CS-TR-4079, University of Maryland, College Park, 1999.

[5] Wang, Y., Phillips, I. T., and Haralick, R. M. "*Document zone content classification and its performance evaluation*". Pattern Recognition, 39: pp. 57–73, 2006.

[6] Keysers, D., Shafait, F., and Breuel, T. M., "*Document image zone classification- a simple high-performance approach*", in Proc. 2nd Int. Conf. Computer Vision Theory and Applications, pp.44-51, Mar. 2007.

[7] Bloomberg, D. S., "*Multiresolution morphological approach to document image analysis*", in Proc. Int. Conf. Document Analysis and Recognition (ICDAR), pp. 963-971, 1991.

[8] Daniel Oliveira, Rafael Lins, "*An Efficient Algorithm for Segmenting Warped Text-lines in Document Images*" Gabriel Torreão Universidade Federal de Pernambuco, Recife, IEEE 2013 12th International Conference on Document Analysis and Recognition

[۹] م عباسی ، ح نظام آبادی پور "روشی جدید برای استخراج متن فارسی از تصاویر بر اساس آشکار سازی لبه های رنگی در حوزه موجک" دانشکده فنی مهندسی دانشگاه باهنر کرمان۱۳۹۱.

[۱۰] م قیاسی ،م رضاییان،ر امیر فتاحی"بخش بندی بدون نظارت بافت تصویر مبتنی بر تجزیه تصویر به ابر پیکسل و توصیف بافت" دانشگاه صنعتی اصفهان ۱۳۹۲.

[11] A. M. Namboordiri, A. K. Jain., "*Document Structure and Layout Analysis.*" Digital Document Processing. s.l.: Springer London, 2007, pp. 29-48.

[12] M. Valizadeh, E. Kabir S. Jalili., "*An Efficient Hierarchical Method for Text Region Extraction in Degraded Document Images*." 2008. Iranian Conference on Machine Vision and Image Processing.

[13] P. Parodi, R. Fontana., "*Efficient and Flexible Text Extraction from Document Pages*." International Journal on Document Analysis and Recognition (IJDAR), s.l.: Springer-Verlag, 1999.

[۱۴] م شامقلی،"رفع اعوجاج و بهبود کیفیت تصاویر اسکن شده از کتب فارسی" پایان نامه ی کارشناسی ارشد، دانشگاه شاهرود ۱۳۹۲.

[۱۵] ح بهین،"تحلیل ساختار اسناد برای استخراج نواحی متنی" پایان نامه ی کارشناسی ارشد، دانشگاه صنعتی سهند ۱۳۸۹.

[16] B.Gatos, D.Danatsas"*Automatic Table Detection in Document Images*" Computational Intelligence Laboratory, Institute of Informatics and Telecommunications National Center for Scientific Research "Demokritos" GR 15310 Athens, Greece (2005)

[17] T.Dhiran, M.Tech Scholar "*Table Detection and Extraction from Image Document*" International Journal of Computer & Organization Trends –Volume 3 Issue 7 – August 2013.

[18] Shafait and Smith (2010), "*Table Detection in Heterogeneous Documents*", DAS '10 Proceedings of the 9th IAPR International Workshop on Document Analysis Systems.

[19] Gatos, B., Pratikakis, I., Perantonis, S.J (2004): "*An adaptive binarisation technique for low quality historical documents*". IARP Workshop on Document Analysis Systems (DAS2004), Lecture Notes in Computer Science (3163), pp.102-113

[۲۰] س خسروی راد ،"رفع اعوجاجات غیر خطی در تصاویر اسناد فارسی" پایان نامه ی کارشناسی ارشد، دانشگاه شاهرود ۱۳۹۱.

Abstract

OCR Systems play a major role in the realization of e-government and to reduce the volume of paper and digital archives. These system use three main parts: preprocessing, recognition of text and post-processing. It is natural that any error in the preprocessing stage, is irreversible. For example, if the skew angle of the document incorrectly identified, will cause text lines to be crooked and make the identification process, to be failed. One of the important parts in the processing, is layout analysis, in the sense that we identify which parts of a document are text, which parts are table, and what areas are image. Any error in this section, will introduce more errors in the OCR process. In this thesis we introduce an algorithm for analysis of multi-column Persian documents. In this field, three approaches are common, bottom-up approach that starts with the integration and development of pixels to create larger areas. Top-down approach, such as XY cut method, first divides the image into several sections and then break down each region into smaller regions. The combination of these two methods are known as hybrid approach. We use a hybrid approach which most of it, are based on bottom up approach. In this approach, we use adaptive thresholding, component labeling, morphological operations and Hough transform in a heuristic algorithm and introduce specific rules for combining small areas without the integration of non-similar areas, to decompose document into text, table and image parts. The proposed method is tested on several multi-column documents with artistic or graphical background and it outperformed the leading OCR software like OmniPage and FineReader. That Numerical results are as follows Our algorithm Persian text with 72 figures 75 and tables 92 percent true diagnose. And 88 percent of Persian documents is almost correct segmentation.

Keywords: Document layout analysis, Page segmentation, bounding box, connected components

Shahrood University

Faculty of Electrical and Robotics Engineering

Segmentation of Complex Persian Documents into Text, Graphics and Table Blocks

Thesis

Submitted in Partial Fulfillment of the

Requirements for the Degree of Master of Science (M.Sc.)

In Electronic Engineering

By:

Mostafa Golzadeh Hamzkanloo

Supervisor:

Dr.Hossein Khosravi

Winter 2015